Université catholique de Louvain
Faculté des bioingénieurs
Earth and Life Institute

Harnessing the Data Revolution for Food Security and Poverty Mapping

Synergies between Mobile Phone Data, Earth Observation and Official Statistics in Senegal

Damien C. Jacques

May 2018

Thèse présentée en vue de l'obtention
du grade de docteur en sciences agronomiques
et ingénierie biologique

Promoteur: Pierre Defourny

Composition du jury:

Promoteur:	Pr Pierre Defourny	(UCL, Belgique)
Président:	Pr Bruno Delvaux	(UCL, Belgique)
Lecteurs:	Pr Patrick Bogaert	(UCL, Belgique)
	Pr Catherine Linard	(UNamur, Belgique)
	Dr François Kayitakire	(EU-JRC, Italie)

To N.

Acknowledgements

Merci à Pierre, mon promoteur, pour m'avoir accordé sa confiance, ouvert les portes de son labo et aidé à obtenir une bourse FRIA. Du projet initial (LAI, évapotranspiration, campagnes de terrain, Belgique), il ne restera finalement que les surfaces! Le voyage fut parfois mouvementé mais nous voilà à destination. Merci également aux membres de mon jury, Catherine Linard, Patrick Bogaert et François Kayitakire, pour avoir accepté de consacrer du temps à relire et commenter ce manuscript. Leurs remarques judicieuses ont grandement contribué à améliorer la qualité de ce travail. Merci également à Bruno Delvaux d'avoir accepté de présider le jury.

Merci à l'ensemble de mes collègues ENGE, une joyeuse bande de gais lurons sans qui toutes ses années n'auraient pas eu la même saveur (soupers de Noël, launch party, pause cafet, pendaison, sport tournament (le maillot gendarme), weddings...)

First things first (l'élégance à la liégeoise), une mention spéciale pour les deux cow-boys, gurus, galactic problem solvers, disciples de Jah et de Shiva, D4D winners, princes des Balkans et bien sûr glocroppers avec qui j'ai partagé mes plus grandes "échelles" et mes plus longs "serpents" (François et Juan-Carlos en savent quelque chose) dans l'antre du C.385. C'est une véritable chance d'avoir pu rencontrer des gars comme vous, exceptionnels à tout point de vue. J'ai un peu traîné des pieds mais me voilà enfin sorti du nid, la dépendaison peut débuter (il y a d'ailleurs toujours ce qu'il faut dans les armoires). Les globcroppers vont enfin faire honneur à leur "glob" (un peu moins au "crop") et stratégiquement occupé 3 continents. Une page se tourne mais le livre reste ouvert ($\mu\epsilon\tau\alpha\phi o\rho\alpha$).

Merci aux plus "anciens" (partis ou toujours là): Guadalupe (on va pas en faire tout un queso), Wilfred (mes pensées à Gisèle), Farid, Sophie, Xavier, JPK (président), Céline (go thesis ;)), Doriane, Catherine (how high), Thomas, Sarah, Jolan (la main verte), Aurélie, Astrid (godmother), Aline (ragots/campagnes de terrain/friterie), Manu (latin smile), Eric, Inès (indicateur coloré), Yannick, Olivier, Mayo (mais qu'on lui érige une statue!), Alexandre, radouxju (les discussion techniques super deep / 36,926 au compteur, le top 3 est faisable, je commence tout doucement à me faire piéger par SO pour ma part), Emilie, François S. (les bons plans lunch, prudence avec l'aloe vera). Merci aussi à Emmanuel pour son esprit festif des premiers jours. Merci à Pascale pour le support logistique et administratif dans la bonne humeur.

Merci à notre génération des “nouveaux“, Benjamin, Maité, Quentin, Renaud, Catherine et Coco qui a donné un bon coup dans la fourmilière à son arrivée en 2012. On se souviendra des 'soupers nouveaux' pour le meilleur et pour le pire ;). Mention spéciale à Coco qui bien qu'outsider non-agro, a su *imposer son style* comme il se doit.

Merci à la génération “post - nouveaux“ (partis ou toujours là): Olivia, Nicolas M., Nicolas N., Nicolas B., Cindy, Julie, Guillaume, Christophe, Florent, Antoine pour avoir su apporter un vent de fraicheur sur le labo tout en nous poussant gentiment vers la catégorie des “anciens”. Merci en particulier à MJ, sur qui j'ai pu toujours compter, good luck pour les last miles :).

Merci à toute l'équipe du D4D: mes camarades espagnols (Pedro et David), italiens (Renato), américains (Sveta), indiens (Neeti), allemands (Till et Fabian) et français (Nicolas, Stéphanie). Remerciement spécial à Eduardo pour m'avoir initié à l'agricultural economics. Merci à mes amis Sénégalais, Vieux et Bamba, pour leur accueil exceptionnel. Merci aussi à Frédéric Gaspart pour son introduction à l'économétrie et pour les (meta-)meta-discussion de fin de journée.

Merci aussi aux copains de PhD Hub, Focus Research, et d'OpenCon. Un tout grand merci à mes amis toulousains Pierre et Laurent (+ Françoise et Cécile). J'admire votre passion pour la recherche et votre manière de partager vos connaissances aux plus jeunes et moins expérimentés, j'ai beaucoup appris à vos cotés. J'espère un jour pouvoir susciter le même enthousiasme chez d'autres.

Merci aux camarades de l'ULB qui m'ont accueilli chaleureusement ces derniers mois. Merci en particulier à Eléonore pour m'avoir permis de jongler entre thèse et projet. J'espère que tu ne te mords pas trop les doigts d'avoir engagé un doctorant en “fin“ de thèse.

Merci Papa pour m'avoir donner le goût pour la science et initier aux rouages de l'Academia Obscura. Merci Maman pour ton soutien inconditionnel et ta tolérance pour les chercheurs (c'était la dernière étape, on peut souffler :)). Merci à tous les deux pour votre amour, l'éducation et les valeurs que vous m'avez inculqués, c'est ma plus grande richesse. Merci aussi à mes deux soeurs que j'adore (et Soso l'impératrice). Big up à tous mes amis “hors-académique” qui m'ont permis de garder un pied sur terre. Parce qu'après tout une thèse... ça n'est qu'une thèse.

Het beste voor het laatste, dank je wel Isa. Si on devait ne garder qu'une chose de ces 5 ans, ça serait notre rencontre. Merci pour ton amour et ton soutien indéfectible, sans hésitation le meilleur *drive* pour achever ce travail. Depuis qu'on se connaît, je suis en “fin de thèse”; c'est dire si la finalisation a pris du temps. Mais on y est, occupons nous maintenant de la tienne pour que la post thesis era, qui s'annonce haute en couleur, puisse commencer :) Le meilleur reste à venir.

Louvain-la-Neuve, Mai 2018

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANSD	Agence Nationale de Statistique et de la Démographie
BSC	Base Station Controller
BTS	Base Transceiver Station
CD	Census District
CDMA	Code-Division Multiple Access
CDR	Call Data Record or Call Detail Record
CILSS	Comité inter-Etats de lutte contre la sécheresse au Sahel
CSA	Commissariat à la Sécurité Alimentaire
CSE	Centre de Suivi Ecologique
CV	Coefficient of Variation
D4D	Data For Development
DAPSA	Direction de l'Analyse, de la Prévision et des Statistiques
DHS	Demographic and Health Surveys
EO	Earth Observation
ESA	European Space Agency
FAO	Food and Agriculture Organization
GDPR	General Data Protection Regulation
GIS	Geographic Information System
GOANA	Grande Offensive Agricole pour la Nourriture et l'Abondance
GP	Gaussian Process
GPS	Global Positioning System
GSM	Global System for Mobile Communications
GSMA	GSM Association
ICT	Information and Communications Technology
INSEE	Institut National de la Statistique et des Etudes Economiques
LA	Location Area
LOOCV	Leave-One-Out Cross Validation
LTE	Long Term Evolution
M2M	Machine to Machine
MDG	Millennium Development Goal
MICS	Multiple Indicator Cluster Survey

MIT	Massachusetts Institute of Technology
MNO	Mobile Network Operator
MODIS	MODerate Resolution Imaging Spectroradiometer
MPI	Multidimensional Poverty Index
MSC	Mobile Switching Centre
NASA	National Aeronautics and Space Administration
NDA	Non-Disclosure Agreement
NDVI	Normalized Difference Vegetation Index
NIR	Near-InfraRed
NSO	National Statistics Office
OCEAN	Openness Conscientiousness Extraversion Agreeableness Neuroticism
OD	Origin Destination
OPAL	Open Algorithm
OPHI	Oxford Poverty & Human Development Initiative
PPP	Purchasing Power Parity
PNAR	Programme National d'Autosuffisance en Riz
PSRFMS	Programme Spécial de Relance de la Filière Manioc au Sénégal
RGPHAE	Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage
RMSE	Root Mean Square Error
SAED	Société Nationale d'Aménagement et d'Exploitation des Terres du Delta du fleuve Sénégal et des vallées du fleuve Sénégal et de la Falémé
SDG	Sustainable Development Goal
SIM	Subscriber Identification Module
SODAGRI	Société de Développement Agricole et Industriel du Sénégal
SODEFITEX	Société de Développement et des Fibres Textiles du Sénégal
SPOT	Satellite Pour l'Observation de la Terre
SMS	Short Message Service
UN	United Nations
USAID	United States Agency for International Development
VAM	Vulnerability Analysis and Mapping
VLR	Visitor Location Register
WFP	World Food Programme

Introduction

In the world,
767 million people live in extreme poverty,
815 million people are undernourished.
1 in 10 people.

These numbers paint a frightening and appalling picture of today's world. These are the most recent estimates made available by internationally renowned institutions (FAO *et al.*, 2017; World Bank, 2016), widely spread by the media and development agencies. They are used to set the target of the two first Sustainable Developments Goals adopted by the United Nations in 2015: "no poverty" and "zero hunger" by 2030. However, there is every reason to believe that they are highly inaccurate; because they are based on outdated, incomplete and unreliable data (World Bank, 2017; Lucci *et al.*, 2018; Carr-Hill, 2013).

Both of these estimations are the result of the aggregation of hundred of household sample surveys carried out in each country[1]. The statistical representativeness of such surveys varies widely and differs according to the period and the regions where the data were collected. In particular, remote, unstable or war zones are often inaccessible for data collection. As an illustration, poverty estimation presented in the preamble is impacted by a severe bias as all data from the Middle East and North Africa, covering 400 million people, were simply omitted because of data coverage and quality issues (World Bank, 2016). They were replaced by a rough estimate of $\sim$ 6 million poor people for the region. However, according to a recent report published by the UN Economic and Social Commission for Western Asia, in the Syrian Arab Republic alone more than 11 million people were pushed into extreme poverty due to five years of conflict (Abu-Ismail *et al.*, 2016).

Because of budgetary constraints, household sample surveys are typically carried out once every five years[2]. The latest comprehensive data on global poverty dates back to 2013 and measures poverty using the USD1.90-a-day 2011

[1] e.g. Demographic and Health Surveys (DHS) funded by the United States Agency for International Development (USAID) or Living Standards Measurement Surveys (LSMS) funded by the World Bank.

[2] Regarding historical data, the picture is even worse. Between 2002 and 2011, among the 155 countries for which the World Bank monitors poverty data, 57 have only one or zero poverty data point available (Serajuddin *et al.*, 2015).

purchasing power parity[3] poverty line, i.e., a 7 years old dataset. A similar flaw is attested in the FAO's estimation on undernourishment, as dozens of countries (including the Democratic Republic of Congo, Burundi, Somalia, the Syrian Arab Republic, Libya, South Sudan, Palestine, and Eritrea) were not reported due to insufficient or not reliable data (FAO *et al.*, 2017).

On the other hand, when scaling up sampling estimations to the national level, accurate estimations of the population are required. Due to very high costs, even a country such as the United States performs a complete census of its population only once every ten years. Although this census is carried out according to the highest professional standards, 16 million people were omitted in the last estimation in 2010 (Mule, 2012). In view of the U.S. experience, it is expected that population estimation in developing countries may be seriously biased leading to potential strong inaccuracy of any estimation derived from household surveys. For example, for political reasons, no population census was conducted in Lebanon since 1932 (Gordon, 2016). In Somalia, the most recent census was held between 1985 and 1986 but was never publicly released (UNFPA, 2014). In Afghanistan, the only national census was conducted in 1979 and only covered 67% of districts due to insecurity (Pinney, 2011). In a remarkable paper, Carr-Hill (2013) shows how household surveys under-represent particular citizens (e.g., slum populations, populations living in areas at risk, etc.) and omit by design certain demographical groups (e.g. homeless, refugees, mobile population, etc.) from the target population. Due to the resulting population undercounting, he estimates that between 300 and 350 million of the poorest ($\sim$ 50%) are likely to be missing in global estimation. Finally, non-sampling errors, i.e., errors that can occur during data collection and the processing of survey data (coverage, response, coding errors, etc.), make things even worse. If not properly controlled, they can be more damaging than sampling errors for large-scale household surveys (Banda, 2003).

As illustrated above, **accurate**, **relevant**, **timely**, and **accessible** data are strongly needed to assess progress towards the development goals. Good quality data are a fundamental prerequisite to explain which policies work and which do not work in eradicating poverty and other social deprivations (Serajuddin *et al.*, 2015). Data are needed to make the situation of the most deprived visible to policymakers. The poor and the undernourished, who often lack a political voice, may remain invisible unless data reveal where and who they are. As a recent UN report stated, "one of the most fundamental inequalities is between those who are counted and those who are not" (Independent Expert Advisory Group on a Data Revolution for Sustainable Development, 2014). Accurate data are not only required for monitoring but also for identifying the *root causes* of social deprivation to avoid getting stuck in managing the symptoms. Finally, when causes are known, early warning systems should be set up using the best

[3]Purchasing Power Parities (PPPs) provide the ratio of the prices of the same good or service in national currencies in different economies. In simple terms, this means that the poverty line applied in India is not obtained by converting U.S. dollars into rupees at the exchange rate, but there is an additional adjustment to allow for the fact that the purchasing power in India is different.

up to date data available. They should alert in case of rapid deterioration of current conditions, such as poor harvests or price increases, allowing authorities to implement appropriate strategies to prevent crises.

According to Stuart *et al.* (2015), to improve the data ecosystem, solutions are threefold:

> "(i) increasing investments in the capacity of national statistical offices, thereby potentially improving the scope and frequency of household surveys; (ii) using alternative sources of data to fill gaps; and (iii) making better use of the data we already have."

This thesis aims to contribute to the two last points by mobilizing the 'data revolution' for sustainable development.

A Data Revolution towards the Sustainable Development Goals

The Millennium Developments Goals (MDGs) were eight international development goals for 2015 about poverty, hunger, health, education, gender equality and environment adopted by 189 United Nations member states during the UN Millennium Summit in September 2000 (UN General Assembly, 2000). They represent an unprecedented commitment made by the nations of the world to implement and achieve internationally agreed development goals, most of which are also basic human rights, as pledged in the 1948 Universal Declaration of Human Rights (UN General Assembly, 1948). While substantial gains have been made, all MDGs were not completely fulfilled at the end of 2015 (United Nations, 2015).

A post-2015 agenda was then adopted in September 2015 at a historic UN Summit, and on 1 January 2016, the 17 Sustainable Development Goals (SDGs) of the 2030 Agenda for Sustainable Development officially came into force (Figure 1). To overcome the limitations of the MDGs, the 17 SDGs were designed to be broader and more ambitious. An important concern was to ensure that 'no one is left behind' by giving the highest priority to people who are the hardest to reach. While global development goals (MDGs, SDGs) are not a panacea (Fehling *et al.*, 2013; Friedman, 2013; Reddy and Kvangraven, 2015; Easterly, 2009), they constitute the only effort today of a world commitment towards improving the human condition at a global scale. US economist Jeffrey Sachs, a longstanding advocate of the MDGs, concedes that the new goals will not be easy to implement. But he argues:

> "The SDGs are a very broad and complex agenda. Whether it can work out is an open question. But there is now an amazing amount of discussion. There is a sense that this is a sensible framework. I'm not saying a new dawn has broken, but at least governments are saying we need to try."

Figure 1: The 17 Sustainable Development Goals (SDGs) adopted in 2015 by the United Nations. The first two SDGs, 'no poverty' and 'zero hunger' by 2030, are covered in this thesis.

It was made clear that progressing on SDGs and planning relevant actions to achieve them, would necessarily entail collecting accurate, reliable and comprehensive quality data to measure them. As Donald Kaberuka, the former African Development Bank President, stated "if it cannot be measured, it cannot be done" ('Delivering on the Data Revolution in Sub-Saharan Africa' event, October 2014). Important efforts have already been made for the MDGs but as mentioned before, huge knowledge and data gaps remain regarding some of the biggest challenges that we face (Boerma and Stansfield, 2007; Devarajan, 2013; Murray, 2007).

During the last decades, the data ecosystem has evolved a lot. The emergence of new technologies (mobile phones, 'internet of things'...) and the growing part of citizen-generated data result in the amount of digital data more than doubling every two years (Gantz and Reinsel, 2011). The actual digital landscape provides the Sustainable Development Goals with a strategic momentum that should be used to guide innovative research towards development applications. It also raises important questions on meta-issues such as data privacy, data literacy[4], or data sovereignty[5].

Aware of these new opportunities and challenges, in August 2014, UN Secretary-General Ban Ki-moon asked an Independent Expert Advisory Group to make concrete recommendations on bringing about a data revolution in sustainable development (Independent Expert Advisory Group on a Data Revolution for Sustainable Development, 2014; Kindornay *et al.*, 2016; Fletcher *et al.*,

[4]Data literacy is the ability to read, create and communicate data as information.
[5]Data sovereignty is the concept that data are subject to the laws and governance structures within the nation they are collected.

2015; Stuart *et al.*, 2015; Glassman and Ezeh, 2014). According to the resulting report, the **data revolution** for sustainable development is (Independent Expert Advisory Group on a Data Revolution for Sustainable Development, 2014):

> - "The integration of the new data with traditional data to produce high-quality information that is more detailed, timely and relevant for many purposes and users, especially to foster and monitor sustainable development;
> - The increase in the usefulness of data through a much greater degree of openness and transparency, avoiding invasion of privacy and abuse of human rights from misuse of data on individuals and groups, and minimising inequality in production, access to and use of data;
> - Ultimately, more empowered people, better policies, better decisions and greater participation and accountability, leading to better outcomes for people and the planet."

In this context, the **Global Partnership for Sustainable Development Data** was established to help stakeholders across countries and sectors fully harness the data revolution for sustainable development. The Global Partnership is a growing network of more than 280 members, including governments, the private sector, civil society, international organizations, academic institutions, foundations, statistics agencies, and other data communities. Among the key partners is **Global Pulse**, a United Nations initiative, launched by the Secretary-General in 2009, to leverage innovations in digital data, rapid data collection, and analysis to help decision-makers gain a real-time understanding of how crises impact vulnerable populations. Global Pulse functions as an innovation lab, bringing together expertise from inside and outside the UN to harness the new world of digital data and real-time analytics for global development. Another initiative is the **Data-Pop Alliance**, 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. More focused on real case studies, the **Flowminder Foundation** is a non-profit organization based in Sweden acting as another key player in the Global Partnership. Its mission is to improve public health and welfare in low- and middle-income countries. The Flowminder Foundation works with governments, inter-governmental organizations and NGOs on extensive datasets, composed of billions of data points including anonymous mobile operator data, satellite, and household survey data. Their analyses allow mapping the distributions and characteristics of vulnerable populations in low- and middle-income countries. Finally, older organizations are also involved such as **Partnership in Statistics for Development in the 21st Century** (Paris 21) that aims to improve governance in developing countries by promoting the integration of statistics and reliable data in the decision-making process. Paris 21 was

established by the United Nations, the European Commission, the Organisation for Economic Co-operation and Development, the International Monetary Fund, and the World Bank in 1999. Those are only a few examples, an exhaustive and updated members list of the Global Partnership for Sustainable Development Data can be found online (www.data4sdgs.org/partner-listing).

The New Data Landscape

The change in the data landscape due to the 'information explosion'[6] of the last decades, has been accounted for in many ways. Reference is generally made to the emergence of the concept of *Big Data*.

The term Big Data, in its modern sense, first appeared in the introduction of a paper of Michael Cox and David Ellsworth, published in the Proceedings of the IEEE 8th conference on Visualization (Cox and Ellsworth, 1997):

> "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources."

Later on, Laney (2001), an analyst with the Meta Group, published a research note titled "3D Data Management: Controlling Data Volume, Velocity, and Variety" introducing the "3Vs" that are now regarded as the key dimensions of Big Data (although the term itself does not appear in Laney's note). **Volume** means that data scale becomes increasingly bigger; **Velocity** relates to the rate at which new data is created; and **Variety** indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data.

In 2011, an International Data Corporation (IDC) report introduced a fourth 'V' for **Value** by defining Big Data as a concept describing (Gantz and Reinsel, 2011):

> "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis."

The '4Vs' model is supposed to highlight the most critical challenge of Big Data, which is how to discover values from unstructured datasets with an enormous scale and rapid generation (Chen *et al.*, 2014).

[6] term first used in 1941, according to the Oxford English Dictionary.

Provocatively, danah boyd and Kate Crawford minimized the revolutionary aspect of Big Data by presenting it as a cultural phenomenon and criticizing the widespread belief that bigger data are always better data (boyd and Crawford, 2012):

> "We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:
>
> (1) *Technology*: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
>
> (2) *Analysis*: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
>
> (3) *Mythology*: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."

In this thesis, we make use of two different data sources that fall under the umbrella of Big Data: mobile phone data and Earth Observation (EO) data. The Copernicus program of the European Commission with its Sentinel satellites produces approximately 10 TB of EO data per day (Kempeneers and Soille, 2017). While mobile phone dataset can include several million of subscribers making billions of interactions (calls, SMS...) (Naboulsi *et al.*, 2016).

The End of Poverty and Hunger

In this thesis, the data revolution is harnessed for applications related to the two first SDGs: to "end poverty in all its forms everywhere" and to "end hunger, achieve food security and improved nutrition and promote sustainable agriculture" by 2030. The popular belief is that one comes with the other; poor people are often hungry, and hungry people are often poor. To such an extent that some have introduced the concept of food poverty defined as the inability to afford or access healthy food. A "poor" person was primarily defined as someone without enough to eat (Banerjee and Duflo, 2012). Going further, some have argued that a nutrition-based poverty trap exists. To feed themselves, people have to work in the fields and grow their food (subsistence farming) or make money with another job and buy the food produced by others. Yet, to work they need sufficient calorific energy coming from food (Dasgupta and Ray, 1986). This vicious circle makes it difficult to assess which is the direction of the causation. In any case, improvement in one of the two SDGs should, in one way or another, benefit the other.

Poverty is Multidimensional

Poverty is an abstract word for which the definition is highly subjective. What poverty is taken to mean mainly depends on who is asking the question, how

it is understood, and ultimately, who responds (International Poverty Centre, 2006). The only point of agreement about poverty is that we all want to see it disappear. Misturelli and Heffernan (2010) performed a synchronic analysis of 159 definitions of poverty offered by different development actors over a 30-year period (1970-2000). The results illustrated that key terms were associated with different connotation challenging the various actors to share a common understanding of poverty. Consequently, any attempt to reach a universal definition of poverty may be doomed from the start.

Is it confined to material aspects of life, or does it also include social, cultural and political aspects? Is it about what may be achieved, given the resources available and the prevailing environment, or what is actually achieved? Should specificity of each country be taken into account or should definitions and measurement methods be applied in the same way everywhere and used for comparisons? What is the rationale for defining a poverty line? Should it be absolute as in the Sustainable Development Goals and in most developing countries, or relative[7] as in the rich OECD countries? Is poverty more useful when measured at the individual, household or administrative level? Finally, the dynamic, particularly the seasonality, of poverty status should be considered. Therefore, the time period over which poverty is identified must be specified.

As demonstrated, a universally accepted definition of poverty does not exist. Lacking viable alternatives, the best approach is likely to give voice to those primarily concerned, the poor. To that end, the largest ever participatory poverty assessment, the World Bank's research programme, *Voices of the Poor*, offers an outstanding approach (Narayan-Parker and Patel, 2000). Over 60,000 poor women and men from 60 countries were convened in small groups to express their experiences regarding poverty. One of them stated:

> "Poverty is pain; it feels like a disease. It attacks a person not only materially but also morally. It eats away one's dignity and drives one into total despair".

In Guatemala, poverty was defined by poor people as having inadequate food and housing and having to rely on charity. In Cameroon, the poor distinguish themselves from the non-poor in five main ways: (i) the presence of hunger in their households, (ii) fewer meals a day and nutritionally inadequate diets, (iii) a higher percentage of their meager and irregular income spent on food, (iv) non-existent or low sources of cash income and (v) feelings of powerlessness and an inability to make themselves heard. In Moldova, most poor people said that the worst aspects of poverty were hunger, poor health, lack of adequate clothing and poor housing conditions.

In light of these testimonials, it should be clear that poverty is a complex and multidimensional phenomenon. Hence one should be cautious to understand which aspects of poverty are taken into account, when poverty is discussed. Most

[7] The inability to reach a minimum acceptable standard of living in a particular society.

countries of the world define poverty in a one-dimensional way, using income or consumption levels (Alkire *et al.*, 2015). However, income is only one aspect of poverty and when defining their experience, poor people often include a lack of education, health, housing, empowerment, employment, personal security, etc. It is therefore unlikely to embrace all facets of poverty by addressing only one perspective. In this thesis, we used the Multidimensional Poverty Index (MPI), developed by OPHI, that takes into account health, education and living standard to define the poverty status of household (Alkire and Santos, 2010). Wang *et al.* (2016) showed that rural household income per capita and other household characteristic variables explain only 7.9% of the variation in the MPI, demonstrating the real multidimensional nature of MPI. Covering multiple dimensions requires multiple sources of data. To that end, the new data sources offered by the data revolution are a key opportunity.

Food Security: Early Warning Systems

Similarly to poverty, food security suffers from conceptual limitations. Even a decade ago, there were about 200 definitions in published writings (Barrett, 2010). The reader is referred to Burchi and De Muro (2016) for a comprehensive review of different approaches to analyzing food security. Here, we will stick to the prevailing definition agreed upon at the 1996 World Food Summit (FAO, 1996):

> "Food security exists when all people, at all times, have physical, social and economic access to sufficient, safe and nutritious food which meets their dietary needs and food preferences for an active and healthy life."

From this definition, four main dimensions of food security can be identified: (i) physical **availability** of food, (ii) economic and physical **access** to food, (iii) food **utilisation** and (iv) **stability** of the other three dimensions over time (Figure 2). These concepts are intrinsically hierarchical, with availability necessary but not sufficient to ensure access, which is, in turn, necessary but not sufficient for effective utilization and stability (Webb *et al.*, 2006).

At least concerning food availability, today's world produces enough food to feed everyone. For the world as a whole, per capita food availability[8] rose from about 2200 kcal/person/day in the early 1960s to 2900 kcal/person/day in 2013 as a result of diet change and innovations in agriculture (www.fao.org/faostat). In comparison, the minimum dietary energy requirement is around 1850 kcal/person/day (FAO, 2008) and the recommendations in the United States are 2600 and 2000 kcal/day for moderately active men and women respectively, between 31 and 35 (DeSalvo *et al.*, 2016). Hic *et al.* (2016) estimated a food surplus of 20-50% of the required calories in most OECD and transition countries in 2010, resulting in a global food surplus of 1200 trillion kcal/yr (20%). This amount of food is enough to feed around 1.4 billion people

[8] After allowing for waste, animal-feed and non-food uses such as biofuel.

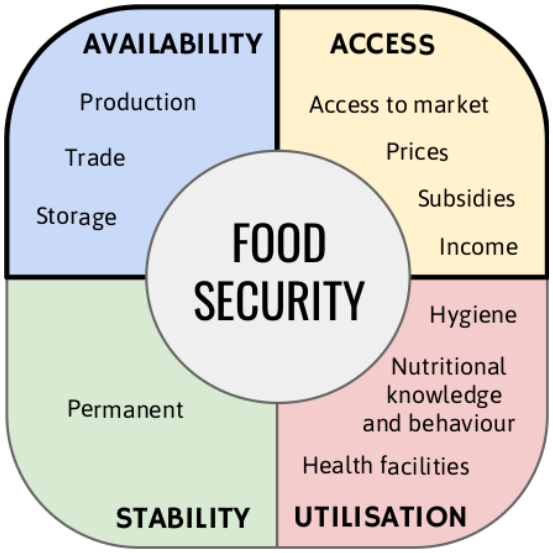


Figure 2: Food security components as defined by FAO (FAO, 1996). In this thesis, the availability and access components are covered in Chapter 1 and 3.

with a daily diet of 2370 kcal/person/day. Furthermore, it is worth noticing that around one-third of global food production (about 1.3 billion tons per year) is lost or wasted (FAO, 2013; Godfray *et al.*, 2010), about one-third of global cereal production is fed to animals (Alexandratos *et al.*, 2012; Shepon *et al.*, 2018), and part of agricultural land are used for biofuels production (Pickett *et al.*, 2008). As Nobel Laureate Amartya Sen wrote more than 30 years ago (Sen, 1981):

> "Starvation is the characteristic of some people not having enough food to eat. It is not the characteristic of there being not enough food to eat. While the latter can be a cause of the former, it is but one of many possible causes"

In view of this, the Malthusian specter[9] of absolute scarcity does not exist in today's world (Alexandratos *et al.*, 2012). There is starvation, but only as a result of the way the food gets shared among us. At the international scale, inequality in food availability is mainly a result of the distribution of agricultural land and water resources relative to the distribution of human populations (Kummu and Varis, 2011; Seekell *et al.*, 2011). This inequality can be enhanced or redressed by human action such as international migration or international trade (Reuveny, 2007; MacDonald *et al.*, 2015; Carr *et al.*, 2016). At the intra-national scale, disparity in access to food is influenced by patterns of poverty, conflict, and the accessibility of distribution networks

[9]Thomas Malthus (1766-1834) argued that there were fixed quantities of certain resources, such as agricultural land, and that this placed strict environmental limits of food production and, by extension, human population.

(Barrett, 2010). Frelat *et al.* (2016) suggested that targeting poverty through improving market access and off-farm opportunities is a better strategy to increase food security than focusing on agricultural production and closing yield gaps. Having said that, with population increase, diet change, land and water scarcity, and climate change, it was estimated that food production should roughly double by 2050 to meet the increasing demand (Rockström *et al.*, 2009; Thornton, 2010; Foley *et al.*, 2011; Hertel, 2011; Hanjra and Qureshi, 2010).

Due to the unequal distribution of world crop production, countries in the less productive areas tend to be more at risk regarding food availability. Especially in regions where subsistence farming is still the norm and where demographic pressure continues to rise[10]. In sub-Saharan Africa, production failure is frequent due to, among others, natural disasters, poor rainy seasons or pest ravage. In such countries, early warning systems play a crucial role in providing timely information on the risk of food crises (Genesio *et al.*, 2011; Davies *et al.*, 1991). To that end, Earth Observation data, as well as agro-meteorological models, can deliver precious information on expected levels of production (Lambert *et al.*, 2016; Atzberger, 2013; Waldner *et al.*, 2015a; Sultan *et al.*, 2013). A good quality of the estimations provided by early warning systems is primordial as information failure might exacerbate an on-going food crisis. According to Devereux (2009), weaknesses in official famine early warning due to problems of data accuracy and credibility lead governments to underrate the food crisis in Ethiopia (1999-2000), Malawi (2001-2002) and Niger (2005).

Food prices are a key element determining food accessibility. A high price prevents food access to consumers, but at the same time, it allows producers to increase their income. The ultimate impact of prices on food security and poverty depends upon the balance between these two effects. Studies generally found that most poor households in developing countries are net buyers/consumers of food such that higher food prices are expected to foster food insecurity and increase poverty in the short run (Ivanic and Martin, 2008; Wodon and Zaman, 2009; De Hoyos and Medvedev, 2011). However, this picture is nuanced by other observation showing that net food sellers would be disproportionately represented among the poor challenging the idea that higher food prices unambiguously deteriorate the income of the poor (Aksoy and Isik-Dikmelik, 2008). In any case, the effect is likely to be context-specific and, as a result, it is difficult to draw any general conclusion about this question. Understanding the formation of food prices is, therefore, vital to make policy recommendations based on empirical evidence.

The greatest barrier to market efficiency lies in the transaction costs. In African food markets, these are often associated with infrastructure deficit (roads, storage facilities, etc.), but the impact of limited information and mistrust can also be significant. Trust and information access are fostered within the social network of each trader. In economy, this asset refers to the concept of social capital. Several studies have shown its impact on transaction costs, but they are

[10] a 3.5-fold increase in the population of Africa is projected for 2100 (Gerland *et al.*, 2014).

generally based on a small sample of traders. The new sources of data brought by the data revolution makes it now possible to approximate social capital of the market traders of an entire country.

Senegal as a Case Study

Senegal serves as a case study for all research presented in this thesis. Senegal, officially *la République du Sénégal*, is the westernmost country of continental Africa and owes its name to the Senegal River, which borders it to the North and East (Figure 3). The name 'Senegal' comes from the Wolof 'Sunuu Gaal', which means 'Our Boat'. The country is located in the Sudano-Sahelian ecoclimatic zone making the transition between the Sahara to the north and the Sudanian Savanna to the south. It is bordered by Mauritania in the North, Mali to the East, and Guinea and Guinea-Bissau to the South. The Gambia forms a virtual enclave within Senegal, penetrating more than 300 km inland along the banks of the Gambia River. The Gambia separates Senegal's southern region of Casamance from the rest of the country. Senegal covers a land area of almost 197,000 square kilometers and has an estimated population of about 15 million (56 percent rural) (Jalloh *et al.*, 2013).

Agriculture employs 77% of the economically active population and accounts for roughly one-sixth of the gross domestic product, down from nearly one-quarter in the mid-1980s (D'Alessandro *et al.*, 2015). Although sector output has expanded by 70% over the past 30 years, population growth has quadrupled so that supply is not meeting demand and the country is a net importer of cereal crops (mainly rice and wheat). Agriculture is mainly rain-fed and depends heavily upon the seasonal rainfall amounts and distribution. The average rainfall varies from over 1000 mm in the South to less than 300 mm in the North. The distribution and kinds of crops are closely tied to the amount, distribution, and timing of rainfall. Crops in the northern half of the country are particularly prone to the effects of erratic rainfall and drought. This region is the most affected by drought, and its population is often exposed to food insecurity. Food staples — millet, sorghum, maize, cassava, and rice — are grown for domestic consumption, while cotton and groundnuts are produced for export. In addition to drought, crop production is subject to threats from pests such as desert locusts.

Senegal ranks 162 with a Human Development Index of 0.49 (UNDP, 2017). As one of the poorest country in the world, it has 75% of population living in multidimensional poverty (OPHI, 2013). Progressing towards the SDGs in such a country is therefore crucial and more challenging than in most part of the world.

Regarding the implementation of the data revolution, Senegal has several assets compared to other African countries. First, statistical capacities are rather good compared to the African average. According to the sta-

Figure 3: Geographical location of Senegal (Intercarto, 2004).

tistical capacity score[11] computed by the World Bank, Senegal is classified third among all sub-Saharan Africa countries (datatopics.worldbank.org/statisticalcapacity/). Consequently, the country serves as a perfect case study to demonstrate innovative approaches to get the most out of traditional data (official statistics), one of the spearheads of the data revolution, in an African context.

Regarding the use of alternative data sources to fill the knowledge gap, mobile phones are probably the most promising data source in Africa as the technology is widely adopted across the continent. In 2016, there were 99%

[11]The Statistical Capacity Indicator is a composite score assessing the capacity of a country's statistical system. It is based on a diagnostic framework evaluating the following areas: methodology; data sources; and periodicity and timeliness. Countries are scored against 25 criteria in these areas, using publicly available information and/or country input. The overall Statistical Capacity score is then calculated as a simple average of all three area scores on a scale of 0-100.

of mobile phone subscriptions per 100 inhabitants in Senegal which implies that most of the population own cell phones. Sonatel-Orange leads the telecom market in Senegal with a market share of 52.2% followed by Tigo and Expresso. The company has pioneered the use of mobile phone data for development by providing several datasets in African countries (Ivory Coast, Senegal) to the research community in the frame of the Data For Development (D4D) challenges (Blondel *et al.*, 2012; de Montjoye *et al.*, 2014). These data represent a unique opportunity to study innovative ways to tackle development issues using the digital traces of millions of subscribers. The D4D dataset is used throughout this thesis with applications related to the first two SDGs.

Scope and Objectives

As the world's governments have pledged to reach the Sustainable Development Goals by 2030, there is a pressing need to mobilize the data revolution to bridge the current knowledge gap, hold political leaders accountable and promote sustainable development in Africa. Now more than ever, we need **relevant**, **accurate** and **timely** information to support better decision-making and track progress towards the development goals. This means making better use of existing data but also harness the potential of alternative data sources produced by the fast-moving and growing digital industry. This new era of *Big Data* fosters a strategic momentum to meet the development goals but also comes with great challenges such as data **access** and **privacy**.

Official statistics are the trusted source of information, generally delivered by national statistics offices (NSO), and used to support policy making. They are assumed to provide unbiased and accurate information thanks to data collected with the highest quality standard (e.g. household surveys or censuses). However, the collection of such data is expensive and difficult to set up. Most of African NSOs are underfunded, lack infrastructures and training and are, therefore, not able to guarantee a sufficient level of quality in the data they gather (Jerven, 2013b). Furthermore, the frequency with which the data are collected is low (e.g. every $\sim$10 years for censuses in best cases) and, in case of surveys, they are limited to a small sample of the population.

This critical situation creates a knowledge gap as fundamental information are missing to track progress towards the SDGs in developing countries. It can be filled by increasing funding and better training of NSO but also by exploring the potential of alternative data sources (e.g. mobile phone data and EO data) to complement traditional data in providing official statistics (i.e. household surveys or censuses). However, due to their unconventional nature, these data are not necessarily always the most appropriate for this task. To adequately supplement official statistics, alternative data sources should at least meet one of the following criteria:

1. **Relevance**: the data provide new type of information than currently available.
2. **Accuracy**: the data provide more accurate information than currently available.

3. **Timeliness**: the data provide earlier or more frequent information than currently available.

In this thesis, we explore how two types of alternative data sources – falling under the umbrella of Big Data – Earth observation and mobile phone data, can meet these criteria to supplement official statistics related to the two first SDGs, no poverty and zero hunger by 2030.

EO data are collected from space by satellite or from sky by airplane or drone. EO satellites can provide daily images of any given point on the Earth's surface (though at different level of precision depending on the satellite[12]). Furthermore, the data collected by public agencies (ESA and NASA) are generally freely available online[13]. Among other, EO data can be used to describe the biophysical material at the surface of the earth (land cover) and how people utilize it (land use) (Green *et al.*, 1994). Furthermore, the frequency of image acquisition (up to twice daily) allows to study temporal dynamic of different land cover (e.g. the evolution of a vegetation index over a growing season). EO data used in this thesis include nighttime lights and vegetation index that serve as a proxy of economic activity and crop productivity respectively.

The use of mobile phone data for development is more recent (Blondel *et al.*, 2012). These data are particularly valuable to estimate metrics at individual level. In this thesis, call detail records[14] (CDRs) are used to capture the individualistic, spatial and temporal patterns of million of subscribers to map poverty at fine scale. Furthermore, the call network between food market areas is used to approximate the social capital of millet traders. Unlike EO data, CDRs are never openly shared with researchers due to their sensitive nature. To efficiently supplement official statistics, there are, therefore, two additional requirements specific to this kind of data:

4. **Access**: the data must be easily accessible, free of charge or at low cost.

5. **Protection**: the data must not impair individual and business privacy.

This thesis focuses on mobile phone metadata but most of the issues discussed in the chapters applied equally to similar data sources that capture geographic digital footprints and are held by private companies (social networks data, credit cards data, emails etc.).

[12]There is a trade-off between the spatial resolution ($\sim$ the pixel size) and the temporal resolution (the frequency of acquisition) of images taken by EO satellites depending on their distance from the Earth's surface. For instance, Sentinel-2 satellites (ESA) provide an image at 10 m of spatial resolution every 5 days while MODIS (NASA) provide images at 250 m twice a day. This limitation can be overcome by deploying a constellation of low Earth orbit microsatellites that can get high resolution images every day (see for instance Planet and its 175+ satellites).

[13]EO data used in this thesis, collected by NASA or ESA, are freely accessible online. High resolution images are sold by private companies like Airbus or Digital Globe.

[14]Mobile phone metadata collected by telecoms companies for billing purpose. These data provide information on when, how, from where and with whom we communicate.

In this context, the overarching objective of this thesis can be formulated as follows:

Harnessing the unique features of Earth Observation and mobile phone data to develop and test new methods contributing to bridge the knowledge gap in food security and poverty mapping in Africa.

This objective is more specifically structured by the following main research questions covering the five requirements for Big Data sources (outline above) to adequately supplement traditional data for official statistics:

*1. What innovative use of CDRs can deliver **relevant** information to support the achievement of main development challenges such as poverty and food security?*

Over the last fifteen years, the field of research dedicated to CDRs analysis has progressively emerged with a recent focus towards societal applications in developing countries. In this thesis, some existing applications of CDRs for development are first briefly reviewed. Then, original and innovative ways to take advantage of this rich data source for food security and poverty monitoring in Sub-Saharan Africa are explored. In particular, it is shown how calls between market areas can be used to approximate the social capital of traders in food markets and how the individual patterns of phone activity allow mapping poverty with an unprecedented level of detail and frequency.

*2. What are the **accuracy** requirements of alternative data sources (CDRs, EO data) to adequately supplement official statistics?*

CDRs hold a tremendous amount of information on mobility, social networks, and socio-demographics of people, with the potential of providing **near real-time** information. The data are massive in the sense that they represent a large part of the population, however it would be erroneous to presume that they are statistically representative of the whole nation. Indeed, a significant bias might arise from, among others, the penetration rate of the technology and the market share of the company. This could ultimately impact the accuracy of any information extracted from the data.

On the other hand, EO data can be used to provide **timely** estimation of production to support early warning systems. To be useful, the new data source should have an accuracy providing lower production error than existing historical crop statistics. In this thesis, we use historical crop statistics to define the accuracy requirements of early estimators of cropland area, crop area and crop yield based on EO data.

*3. Can we secure **access** to sensitive data such as CDRs while **protecting** individual and business privacy?*

Access to CDRs data is generally limited because of risks for individual and business privacy and rely upon the goodwill of the private companies that manage them. This issue is developed in the thesis including a discussion about the trade-off between usability and privacy. These comments may have implications beyond the field of mobile phone data because several of the discussed considerations apply equally to similar data sources (social network data, credit cards data, etc.).

Relying on these data in an operational context requires the implementation of pragmatic solution ensuring the sustained access to the data. Keeping the sensitive data safely stored behind the firewall of the data providers is one approach discussed in this thesis. It is practically illustrated with an application in poverty mapping.

Outline

Each Chapter is introduced below by stating its research goals and by outlining its relationship with the overarching research questions stated before.

Chapter 1 outlines a methodological framework to define the accuracy requirements of early estimators of cropland area, crop area and crop yield, along the season, according to (i) the inter-annual variability and the trend of historical data, (ii) the calendar of official statistics data collection, and (iii) the time at which the early estimators can theoretically be available. This directly answers the second research question. This work contributes to SDG 2, "zero hunger", by focusing on food availability, the first component of food security.

Chapter 2 introduces mobile phone metadata (with a focus on Call Detail Records) and their application for global development issues. After a description of their specific features (network, temporal and geospatial dimensions), a brief review of their current use in developing countries is given. Challenges for operational use are also discussed. In particular, limitations due to potential bias, data access, and privacy are highlighted. This Chapter provides some answers to the three research questions.

Chapter 3 investigates the impact of social capital on food prices and market functioning in Africa. Social capital can lower transactions costs by, e.g., reducing the information and search costs, increasing trust or cutting down the administrative burden. Social capital is modeled using a unique data set of mobile phone communications between 9 million people integrated in a spatial equilibrium model. This innovative approach contributes to meeting the objective defined in the first research question. This work contributes to SDG 2, "zero hunger", by focusing on food access, the second component of food security.

Chapter 4 outlines a computational framework to efficiently combine disparate data sources, such as mobile phone and environmental data, to provide more accurate predictions of poverty and its determinants, for finest spatial micro-regions in Senegal. This approach can be used to generate poverty maps more frequently and assist policy makers in designing better interventions for poverty eradication. This work seeks to answer both second and third research question by proposing a framework to safely integrate several data sources. This work contributes to SDG 1, "no poverty".

The main findings are summarized and discussed in the conclusions of this document.

Chapter 1

Accuracy Requirements for Early Estimation of Crop Production in Senegal

Your hunger is never satiated, your thirst is never quenched; you can never sleep until you are no longer tired.
A poor man, Senegal 1995 (Narayan-Parker and Patel, 2000)

Highlights

- Efficient early warning systems for food security are needed to prevent food crisis (**relevance**).
- Early estimators of production are generally assessed regardless of the existing historical crop statistics.
- This Chapter outlines a methodological framework to define the accuracy requirements of early warning estimators of cropland area, crop area and crop yield, along the season, according to (i) the inter-annual variability and the trend of historical data, (ii) the calendar of official statistics data collection, and (iii) the time at which the early estimators can theoretically be available (**accuracy**).
- The inter-annual variability of crop yield is the main factor limiting the accuracy of pre-harvest production estimates.
- Estimators of cropland area were useful to improve production prediction of the main crops in Senegal stressing the value of the cropland mapping for food security (**relevance**).
- While applied to Senegal, this study can be reproduced in any place where reliable agricultural statistics are available.

Abstract

Early warning systems for food security rely on timely and accurate estimations of crop production. Several approaches have been developed to get early estimations of area and yield, the two components of crop production. The most common methods, based on Earth observation data, are image classification for crop area and correlation with vegetation index for crop yield. Regardless of the approach used, early estimators of cropland area, crop area or crop yield should have an accuracy providing lower production error than existing historical crop statistics. The objective of this Chapter is to develop a methodological framework to define the accuracy requirements for early estimators of cropland area, crop area and crop yield in Senegal. These requirements are made according to (i) the inter-annual variability and the trend of historical data, (ii) the calendar of official statistics data collection, and (iii) the time at which early estimations of cropland area, crop area and crop yield can theoretically be available. This framework is applied to the seven main crops in Senegal using 20 years of crop production data. Results show that the inter-annual variability of crop yield is the main factor limiting the accuracy of pre-harvest production forecast. Estimators of cropland area can be used to improve production prediction of groundnuts, millet and rice, the three main crops in Senegal stressing the value of cropland mapping for food security. While applied to Senegal, this study could easily be reproduced in any country where reliable agricultural statistics are available.

1.1 Introduction

Early warning systems for food security rely on timely and accurate estimations of crop production (Genesio *et al.*, 2011; Hutchinson, 1991). Their role is to inform on risks of food crises to efficiently put in place adequate response strategies (Davies *et al.*, 1991). Traditionally, crop statistics data are produced by national statistical offices (NSO) using adequate sampling strategies and statistical inference (Keita *et al.*, 2012). Final production estimates are typically available a few months after the harvest (around November in Senegal, see Figure 1.1). Yet, earlier estimation (ideally before the harvest) would substantially improve the efficiency of the political responses to food shortages. This particularly applies to countries with highly variable agricultural production such as Senegal.

Several approaches have been developed to get early estimations of area and yield, the two components of crop production. For crop area estimation, the most common method relies on the classification of Earth Observation data (pixel counting) following a two-step procedure: (i) an early estimate of the total cropland area and (ii) later, the discrimination of each crop within the cropland (Xiong *et al.*, 2017; Lambert *et al.*, 2016; Atzberger, 2013). The bias of this method mainly comes from two sources of error: the presence of mixed (border) pixels and the misclassification of pure pixels (Delincé *et al.*, 2017). How significant are these errors depends predominantly on the spatial resolution of the images, the spectral separability of the different crops, the classification algorithm and the quality of the training data (Duveiller and Defourny, 2010; Gómez *et al.*, 2016; Waldner *et al.*, 2017b; Inglada *et al.*, 2015b). As for crop

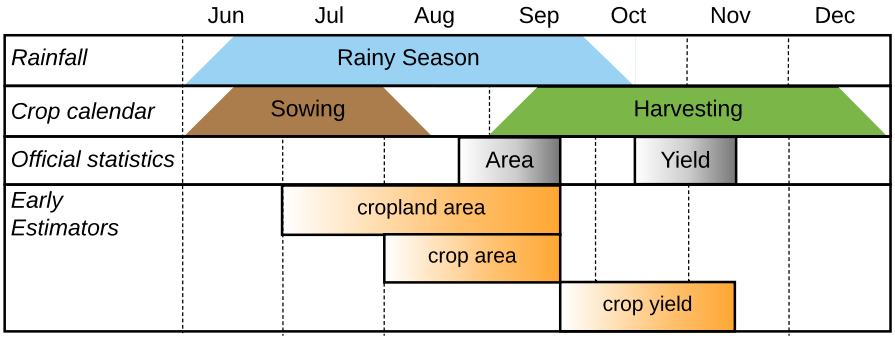


Figure 1.1: Calendar of the critical periods regarding early warning systems for food security in Senegal. The rainy season lasts from June to October. Sowing starts with the first rain, and harvesting follows the end of the rainy season. Official statistics of crop area are collected from late August to the end of September, and official statistics of crop yield are collected from mid-October to mid-November. Using traditional methods such as the classification of Earth Observation data, the probability to get a good estimation of cropland area and crop area increases over the season until September when the actual value is known. On the other hand, crop yield estimations are hard to predict before the peak of the growing season in mid-September.

yield, it was shown to be correlated with vegetation indices derived from Earth Observation data (Tucker *et al.*, 1985; Groten, 1993; Rembold *et al.*, 2013; Burke and Lobell, 2017; Lambert *et al.*, 2017). Agro-meteorological models such as SARRA-H (Sultan *et al.*, 2005, 2013) or GLAM (Challinor *et al.*, 2004) are also used and generally considered more rigorous. However, these models require setting several parameters for which data are often lacking (e.g. the fertilization rate).

These approaches are typically assessed irrespectively of the existing historical crop statistics. Yet, to be useful, early estimators of cropland area, crop area or crop yield should have an accuracy providing lower production error than existing information. For instance, having early estimators of crop area and yield providing final production estimate with an error of 25% would not be useful if it is possible to achieve an error of only 20% by simply using the historical average and the trend of production (information available at the beginning of the growing season). The lower the inter-annual variability of production (and the easier it is predictable by a trend), the higher would be the accuracy requirements for early estimators of cropland area, crop area and crop yield. In the Sudano-Sahelian zone, this variability is generally significant due to the high inter-annual change in rainfall pattern specific to the region (Graef and Haigis, 2001; Grist and Nicholson, 2001; Haarsma *et al.*, 2005). Several factors can also play a role such as agronomic conditions (soil quality), crop management practices (fertilization rate, phytosanitary protection), pest outbreaks (locust, birds), infrastructure development (storage facilities) as well as market conditions (price, demand, competition) (Sghir *et al.*, 2015; D'Alessandro *et al.*, 2015).

From an early warning perspective, a critical factor that should be taken into account is the date on which ground truth data are available. In Senegal, official statistics of crop areas are known at the end of September and crop yield at the end of November (Figure 1.1 - Official statistics). It is unlikely to get accurate estimations of yield earlier than September as the end of the vegetation growth, a critical period in determining the final yield, occurs at the same time (Figure 1.2 - NDVI zones). Indeed, most of the study estimating crop yield from EO data either use the maximum of a vegetation index such as the Normalized Difference Vegetation Index (NDVI) (Groten, 1993; Lambert *et al.*, 2017) or its integration over the growing cycle (Tucker *et al.*, 1985). Practically, it means that early estimators of cropland and crop area are useful before September and estimators of yield, between September and November (Figure 1.1 - Early Estimators).

The overall objective of this Chapter is to develop a methodological framework defining the accuracy requirements for early estimators of cropland area, crop area and crop yield in Senegal. These requirements are made according to (i) the inter-annual variability and the trend of historical data, (ii) the calendar of official statistics data collection, and (iii) the time at which early estimations of cropland area, crop area and crop yield can theoretically be available (based on EO data). This framework is applied to the seven main crops in Senegal

using 20 years of crop production data. While applied to Senegal, this study could easily be reproduced in any country where reliable agricultural statistics are available.

1.2 Agriculture in Senegal

Today, the agriculture sector accounts for one-sixth of the Senegalese gross domestic product (D'Alessandro *et al.*, 2015). Rice is the main staple food consumed by Senegal's population, but production is only able to meet half of the demand (Centre de Gestion et d'Économie Rurale de la Vallée du fleuve Sénégal, 2014). Because of its importance in Senegalese food, rice self-sufficiency has been the target of public authorities for many years. Following the 2007-2008 food crisis, several public initiatives such as the *Programme National d'Autosuffisance en Riz* (PNAR) or the *Grande Offensive Agricole pour la Nourriture et l'Abondance* (GOANA), helped to increase the production over the years. Most of the rice is irrigated and produced in Senegal river valley by the *Société Nationale d'Aménagement et d'Exploitation des Terres du Delta du fleuve Sénégal et des vallées du fleuve Sénégal et de la Falémé* (SAED) and the *Société de Développement agricole et industriel* (SODAGRI) (Figure 1.2).

Millet, and to a less extent, sorghum, serves as the main local subsistence food crops (Dong, 2011). Millet is the most drought-resistant crop in the country, and it covers a third of Senegal's cultivated areas (Figure 1.3). Most of the millet is grown in the regions of Kaolack, Kaffrine and Fatick where it is interchanged with groundnuts (Figure 1.2). This crop rotation is crucial because groundnuts fix nitrogen in the soil.

Groundnuts are the main cash crop of the country. While the sector is going through a crisis since the nineties due to several factors such as the different agricultural policies, the market fluctuation or the deterioration of farm equipment, it is still the first production in Senegal (Figure 1.4) and it plays a vital role in rural economy (Noba *et al.*, 2014). Groundnuts are primarily cultivated in the Groundnuts Basin located in the region of Fatick, Kaolack, and Kaffrine.

Cotton is the second cash crop in the country (Figure 1.4). The production is low and almost entirely exported (Guiro *et al.*, 2005). In Senegal, most of the cotton is produced by the *Société de Développement et des Fibres Textiles* (SODEFITEX) in the region of Kolda.

Cassava and maize have raised public policies interest as an interesting source of diversification for the subsistence agriculture in Senegal. In 2004, showing an aggressive agricultural policy and revived interest, the Senegalese government launched a major program, the *Programme Spécial de Relance de la Filière Manioc au Sénégal* (PSRFMS), for intensifying the production of cassava for food security purposes (Diallo *et al.*, 2013). The cassava and maize sector have also benefited from the GOANA support in 2008. Cassava is mainly

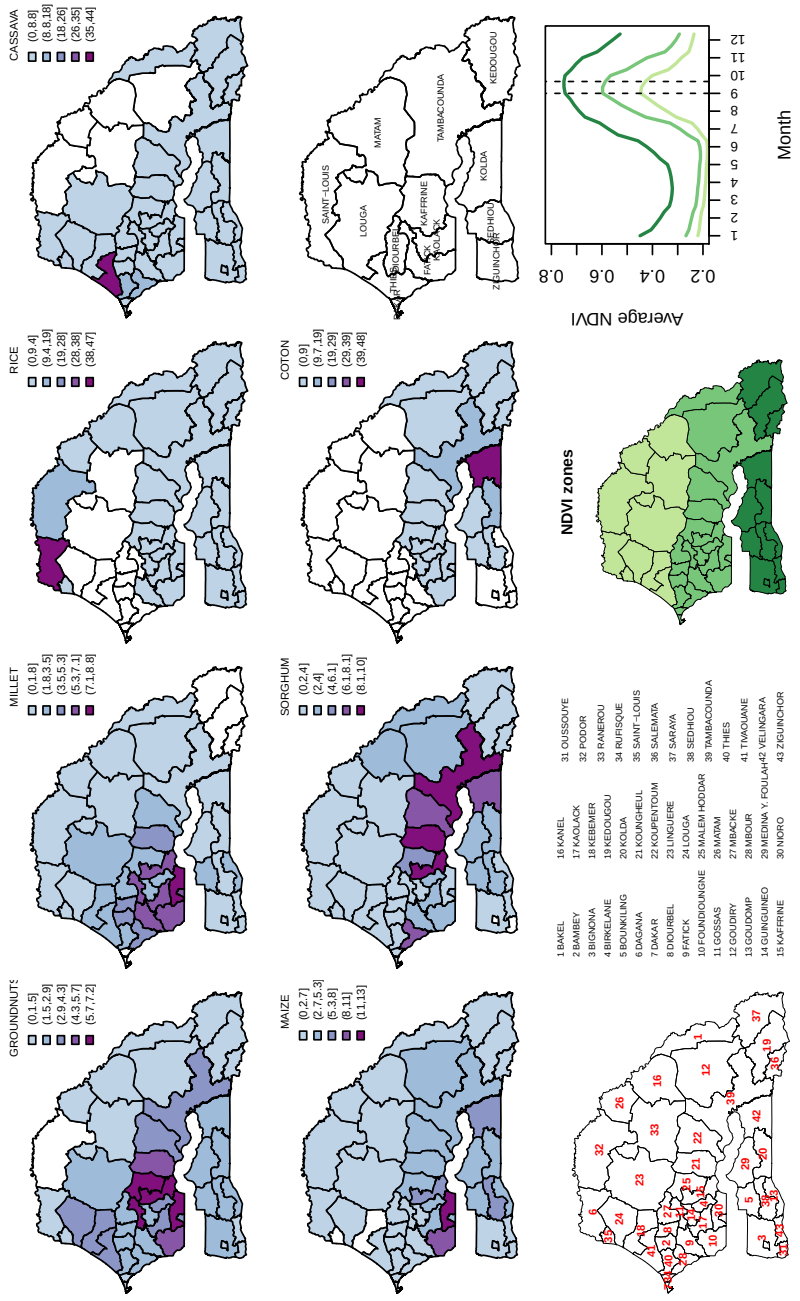


Figure 1.2: Average production (2010-2016, source: DAPSA) by department for each crop in percentage of the total production (white areas correspond to null production). A map of the 14 regions in Senegal is also shown. The two last plots show the average NDVI (from MODIS images) in 3 zones following a South-North gradient. The dashed lines show the period of maximum NDVI (end of September and beginning of October).

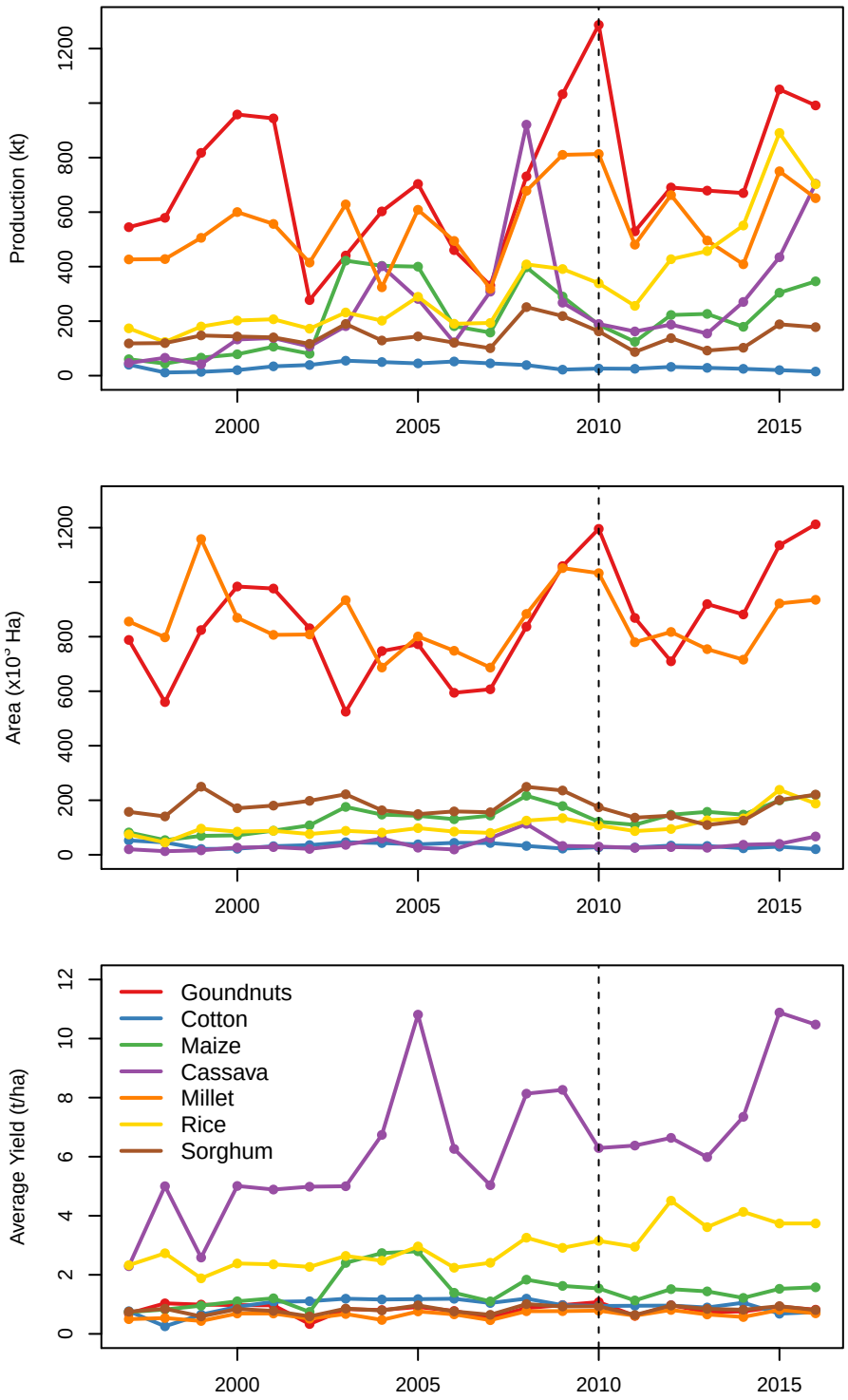


Figure 1.3: Production, area and yield time-series of each crop. The vertical dotted line shows the date upon which departments level data were available.

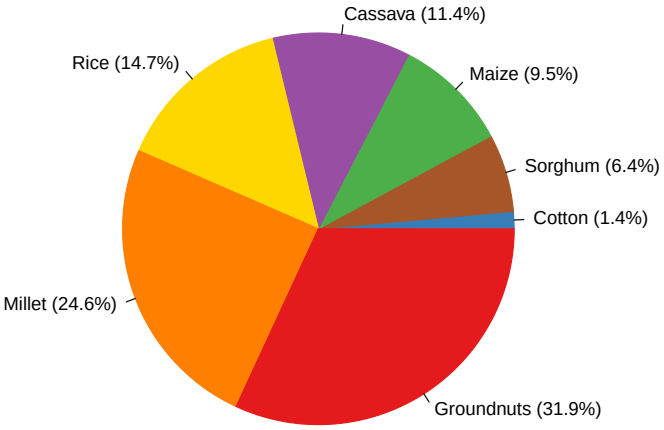


Figure 1.4: Pie chart of the average national production (1997-2016) of the seven main crops in Senegal.

cultivated in Thiès region, and maize, in Fatick and Kaolack region (Figure 1.2).

Other principal crops in Senegal include sesame (12 kt in 2016), watermelon (285 kt in 2016) and cowpea (100 kt in 2016). These are not taken into account in this study.

1.3 Data

Official agricultural statistics delivered by the *Direction de l'Analyse, de la Prévision et des Statistiques Agricoles* (DAPSA) are used in this analysis (Direction de l'Analyse, de la Prévision et des Statistiques Agricoles, 2013). In this study, these are regarded as the ground truth and serve as the reference to evaluate the inter-annual variability of production. It makes senses in Senegal as the statistical capacities are rather good compared to the African average (see Introduction of the thesis).

	Production (kt or %)			
	min	mean	max	CV
Groundnuts	277.3	715.9	1286.9	36.8
Millet	318.8	552.3	813.3	26.8
Rice	123.5	329.4	890.4	60.5
Cassava	42.1	255.6	920.9	86.2
Maize	44.3	213.8	422.0	60.2
Sorghum	86.9	144.5	251.5	29.7
Cotton	11.6	32.0	55.0	41.3

	Area (10^3 ha or %)			
	min	mean	max	CV
Groundnuts	524.8	850.8	1211.0	23.7
Millet	686.9	852.3	1158.2	14.6
Rice	45.2	106.6	238.2	40.6
Cassava	13.1	36.3	113.2	64.2
Maize	53.7	135.2	218.5	35.5
Sorghum	108.8	177.0	249.7	23.4
Cotton	20.6	33.6	52.6	28.3

	Yield (t/ha or %)			
	min	mean	max	CV
Groundnuts	0.33	0.83	1.08	21.7
Millet	0.44	0.64	0.81	18.8
Rice	1.88	2.93	4.51	24.2
Cassava	2.29	6.46	10.97	37.0
Maize	0.74	1.47	2.80	40.1
Sorghum	0.59	0.82	1.01	15.9
Cotton	0.26	0.95	1.19	25.3

Table 1.1: Descriptive statistics of national production, area, and yield for each crop over 20 years (1997-2016). CV stands for Coefficient of Variation computed as the standard deviation divided by the mean (expressed in %) which is a measure of the inter-annual variability (without taking into account for the trend).

The estimates are based on a two-stage stratified sample of 6300 agricultural households in accordance with an harmonized approach intended to be applied similarly in all countries of the *Comité inter-États de lutte contre la sécheresse au Sahel* (CILSS). The sample size has slightly changed over the years, but the method remains the same. As a first stage, between 18 and 30 census districts (CDs) are randomly drawn in each of the 42 departments of Senegal (Dakar is not included) in proportion to their population for a total of 900 sampled CDs out of 17,165. Then, within each CD, 7 agricultural households are randomly drawn which gives a final sample of 126 to 210 households by department.

The data collection is divided into two phases (Figure 1.1). During Phase 1, which starts at the end of August, the crop area is estimated. The crop type and the location of all fields of each sampled household are recorded. Phase 2 begins in October at the end of the growing season. At that time, the yield is estimated by averaging measurements taken within 60 plots for each crop and department among the monitored fields (dimension specific for each crop, for instance, 5m x 5m for groundnuts and 1m x 1m for rice). Both phases last 30 days meaning that the area is known in September and the yield, in November. Some crops are not monitored by the DAPSA and their statistics are provided by other institutions. These are the rainy season irrigated rice in SAED (St-Louis and Matam Region as well as Bakel Department) and SODAGRI (Anambé, Velingara Department) zone, and the cotton in SODEFITEX zone.

The seven main crops of Senegal are considered in this study: groundnuts, millet, rice, cassava, maize, sorghum and cotton (Figure 1.4, Table 1.1). Rice referred to paddy rice, when husked, it loses about 30% of its weight.

National production data were available from 1997 to 2016 and local production (at department level) were available from 2010 to 2016. Both data were directly obtained from the DAPSA.

1.4 Method

The objective of this Chapter can be further defined in specific research questions listed hereafter.

First, based on twenty years of national agricultural production:

1. What part of the inter-annual variability of crop production is predictable by the trend?

Second, based on seven years of departmental agricultural production:

2. How the inter-annual variability of crop production is spatially distributed?

Finally, according to (i) the inter-annual variability of production and the trend, of historical data (first research question) (ii) the calendar of official statistics data collection (Figure 1.1 - Official Statistics), and (iii) the time at which early estimations of cropland area, crop area and crop yield can theoretically be available based on EO data (Figure 1.1 - Early Estimators):

3. What is the lowest error of crop production estimation achievable along the season?

4. What are the accuracy requirements of early estimators of cropland area, crop area and crop yield?

The following sections detail the methodological background used to answer these research questions.

1.4.1 Production Estimator and Error Measurement

For a crop i and a year t, the error between the actual production p_{it} and the production estimation $\hat{p}_{it}$ is given by:

$$\epsilon_{it} = \hat{p}_{it} - p_{it} \tag{1.1}$$

A good estimator should provide low error $|\epsilon_{it}|$ with $\epsilon_{it} = 0$ for a perfect estimator. An estimator can give good estimates for a year t but poorly estimates the year $t + 1$. For this reason, the quality of an estimator should ideally be assessed over several years. From a vector of predictions ϵ_{it}, ...,ϵ_{iT} (where T is the number of years in the time series), several criteria can be chosen as indicator of quality. The most obvious is the average error $\bar{\epsilon}_i$ of each year t:

$$\bar{\epsilon}_i = \frac{\sum_{t=1}^{T} |\epsilon_{it}|}{T} \tag{1.2}$$

To assess the quality of an estimator, one could also define a threshold value that $|\epsilon_{it}|$ should never exceed. For instance, an estimator would be considered insufficient if one year is estimated with an error exceeding 30%, even if the average error of all years $\bar{\epsilon}$ is below this threshold. In this study, we opted

for the Root Mean Square Error (RMSE) as the indicator of quality of a crop production estimator:

$$\text{RMSE}_i = \sqrt{\frac{\sum_{t=1}^{T} \epsilon^2}{T}} \tag{1.3}$$

Compared to the simple average $\bar{\epsilon}_i$ (Eq. 1.2), RMSE penalizes large errors. To be able to compare RMSE of different crop, RMSE_i was normalized by the mean of production $\bar{p}_i$ which gives the coefficient of variation $\text{CV}(\text{RMSE}_i)$:

$$\text{CV}(\text{RMSE}_i) = \frac{\text{RMSE}_i}{\bar{p}_i} \tag{1.4}$$

A good estimator will have a low CV(RMSE) while a poor estimator will have a large CV(RMSE).

1.4.2 Average of Past Data

Without an apparent trend, the production (at country level) of a crop i and a year t can be estimated by the average of past production data:

$$\hat{p}_{may,it|T} = \bar{p}_i = (p_{i1} + ... + p_{iT})/T \tag{1.5}$$

where $p_{i1}, ..., p_{iT}$ are the past production of crop i. Because it only uses past data, this estimator is available at the beginning of the growing season (in May in Senegal).

To get the CV(RMSE) of this estimator, we used historical data spanning twenty years from 1997 to 2016. We computed the error of estimation for each year t using the average of all other years T (19 years) as predictor. This approach corresponds to a leave-one-out cross validation (LOOCV). The CV(RMSE) will be large if the inter-annual variability of production (the CV of the time series given in Table 1.1) of the crop is high. Conversely, if the inter-annual variability of the crop production is low, the CV(RMSE) will be low and the accuracy requirements of early estimators of cropland, crop area and crop yield will be high.

Similarly to Eq. 1.5, at the beginning of the season, we can estimate the area a_{it} and the yield y_{it} using $\hat{a}_{it|T}$ and $\hat{y}_{it|T}$ respectively. Furthermore, an estimator of the proportion of crop area in total cropland area $\widehat{(a/c)}_{it|T}$ can also be computed using historical average of a_{it}/c_t where c_t is the total cropland area of year t. This last estimator can be interesting when the total cropland area of the predicted year is the only information available (Figure 1.1 – Early Estimators).

These four estimators $\hat{p}_{it|T}$, $\hat{a}_{it|T}$, $\hat{y}_{it|T}$ and $\widehat{(a/c)}_{it|T}$ are available at the beginning of a new growing season (in May in Senegal) and form the baseline

for the accuracy requirements of early estimators of cropland area, crop area and crop yield computed from, for instance, EO data.

1.4.3 Trend of Past Data

In temperate climate and developed countries, most of the production variability can be explained by the trend because long-term policies, market change as well as technology and practices improvements are the driving force of production growth (Defourny *et al.*, 2007). In the Sudano-Sahelian zone, it is not the case as unpredictable events such as droughts, floods, pest outbreaks (locust, birds), limited access to inputs or conflicts have a large effect on production variability and long-term policies are not always efficient (D'Alessandro *et al.*, 2015). Yet, supporting programs and economic investments in a specific sector may still result in a sustained increase in production. To identify such a trend, simple linear (Eq. 1.6) and exponential (Eq. 1.7) regression models (ordinary least square) were run with national production, area, and yield against time for each crop:

$$u_{it} = \alpha_i + \beta_i t + \epsilon_{it} \tag{1.6}$$

$$log(u_{it}) = \alpha_i + \beta_i t + \epsilon_{it} \tag{1.7}$$

where u_{it} is the production p_{it}, the area a_{it} or the yield y_{it} of crop i and year t, and ϵ is the regression residual. We assumed that there was an actual trend in a time series when the β_i coefficient was statistically significant (p-value<0.01). If the linear and the exponential model had both a significant β_i, the model with the highest coefficient of determination R^2 was selected.

As theoretically predictable and, therefore, known at the beginning of the growing season, significant trends were subtracted from the historical time series:

$$z_{it} = u_{it} - \text{trend}_{it} \quad \text{with} \quad \text{trend}_{it} = \begin{cases} \alpha_i + \beta_i t & \text{if the trend is linear,} \\ e^{\alpha_i + \beta_i t} & \text{if the trend is exponential} \end{cases} \tag{1.8}$$

where z_{it} is the detrended production, the detrended area or the detrended yield of crop i and year t.

For the crops with a significant trend, Eq. 1.5 can then be rewritten as:

$$\hat{u}_{may,it|T} = \bar{z}_i + \text{trend}_{it} \tag{1.9}$$

where $\bar{z}_i = (z_{i1} + ... + z_{iT})/T$ with $z_{i1}, ..., z_{iT}$ the detrended past production, area or yield of crop i as defined in Eq. 1.8.

1.4.4 Spatial Variability

Production variability is a function of time but also space. Some departments have a stronger variability than others. Only seven years of historical data were available at the department level (2010-2016) which was not enough for a rigorous analysis. However, a general picture can be derived from $\hat{p}_{may,it|T}$ accuracy for each department and each crop. For this analysis, the trend was not taken into account as hard to accurately estimate on short time series. The spatial analysis of production variability has interesting operational implication because it allows to broadly stratifying the country in priority zones. Departments that contribute the most to the national variability should be targeted (sampled) first to get accurate estimations of production.

1.4.5 Early Estimators

Along the season, early estimators of cropland area $\hat{c}_{it}$, crop area $\hat{a}_{it}$ and crop yield $\hat{y}_{it}$ become available thanks to EO data (see Figure 1.1 – Early Estimators).

In July/August, crops start growing. Using EO data, it is impossible to make the distinction between each crop but the entire area covered by the cropland can be estimated. The production is estimated as:

$$\hat{p}_{jul,it} = \hat{c}_t \widehat{(a/c)}_{it|T} \hat{y}_{it|T} \tag{1.10}$$

In August/September, crops have now sufficiently grown to be identified. Using EO data, the area of each crop can be estimated. However, it is still too early to predict the crop yield. The production is estimated as:

$$\hat{p}_{aug,it} = \hat{a}_{it} \hat{y}_{it|T} \tag{1.11}$$

In September, the growing season has reached its peak and the official statistics of crop area a_{it} are known. The production is estimated as:

$$\hat{p}_{sep,it} = a_{it} \hat{y}_{it|T} \tag{1.12}$$

After September, the growing season is over and the crop yield can be estimated using EO data. The production is estimated as:

$$\hat{p}_{oct,it} = a_{it} \hat{y}_{it} \tag{1.13}$$

Finally in November, official statistics of crop yield y_{it} are known and the final estimate of production (considered the ground truth) is available:

$$p_{nov,it} = a_{it} y_{it} \tag{1.14}$$

The dates indicated for p_{may}, p_{sep} and p_{nov} are set by the calendar of data collection carried out by the DAPSA (Figure 1.1 – Official Statistics). On the other hand, the dates for p_{jul}, p_{aug} and p_{oct} reflect the probability to have, at this time, a good estimator of cropland area $\hat{c}$, crop area $\hat{a}$ and crop yield $\hat{y}$, using EO data. These are given for clarity and ease of interpretation. The important point is the temporal sequence at which each estimator is available.

1.4.6 Accuracy Requirements

During the growing season, a new estimator should be used if it increases the accuracy of production prediction. Therefore, some rules can be set to define the accuracy requirements of the early estimators $\hat{c}_t$, $\hat{a}_{it}$, $\hat{y}_{it}$:

$$\begin{aligned} \text{CV(RMSE}_{may,\ i}) \geq \text{CV(RMSE}_{jul,\ i}) \geq \text{CV(RMSE}_{aug,\ i}) \geq \\ \text{CV(RMSE}_{oct,\ i}) \geq \text{CV(RMSE}_{sep,\ i}) \geq 0 \end{aligned} \quad (1.15)$$

From this temporal sequence, it may be inferred that the CV(RMSE) of pre-harvest production estimation (i.e. before September in Senegal) will never be lower than CV(RMSE$_{sep}$) and should not be higher than CV(RMSE$_{may}$). It means that the main factor limiting the accuracy of pre-harvest production estimates is the accuracy of $\hat{y}_{it|T}$ which itself only depends on the inter-annual variability of the yield and its predictability by a trend.

CV(RMSE$_{may,\ i}$) and CV(RMSE$_{sep,\ i}$) are independent from the accuracy of the estimators $\hat{c}$, $\hat{a}$ or $\hat{y}$ and only depend on historical data. However, CV(RMSE$_{jul,\ i}$) and CV(RMSE$_{aug,\ i}$) depend on the accuracy of $\hat{c}$ and $\hat{a}$ respectively. It means that when both estimators are available (from August), the best estimator between $\hat{p}_{jul}$ or $\hat{p}_{aug}$ depend on the combined accuracy of $\hat{c}$ and $\hat{a}$. The impact of this case on the accuracy requirements is explored in the results.

1.5 Results

1.5.1 Trend Analysis

As expected, the trend (exponential) is highly significant for rice and explains most of the variance of its production, area, and yield (Table 1.2). It is a long-standing objective of the government to become self-sufficient in rice production to decrease the importation costs (see Data section). Several supporting programs and projects have been dedicated to this crop over the years leading to an increase in area and yield and, therefore, production. While the impact is less important, a significant exponential trend is also observed for cassava and maize. The trend in cassava production seems to be more explained by a yield increase while for maize, the trend is driven by a rise in cultivated areas. These two crops have also been supported by public policies to diversify the subsistence agriculture offering.

1.5.2 Lowest Error of Early Estimation of Production

Figure 1.5 shows the lowest error of production estimation (expressed in CV(RMSE)) achievable for each crop according to the increasing data availability along the season (considering a perfect estimator of cropland area, $\hat{c} = c$)).

At the beginning of the season in May, production estimates are given by the historical average (and the trend, if any) of past production data ($\hat{p}_{may}$). The CV(RMSE$_{may}$) at this time depends on the inter-annual variability of production and its predictability by a trend. Despite their significant trend, cassava and maize have the highest CV(RMSE$_{may}$) (77% and 59% respectively) due to the high inter-annual variability of their production (CV of 86.2% and 60.2% respectively, see Table 1.1). All others variables have a CV(RMSE$_{may}$) in line with their CV except for rice due to its significant trend (CV(RMSE) = 30% and CV = 60%). The crop with the lowest CV(RMSE$_{may}$)

	Production		Area		Yield	
	β	R^2	β	R^2	β	R^2
Groundnuts	13.64	0.09	16.2	0.23	-0.86	0.00
Millet	0.02	0.12	-1.46	0.00	10.71	0.26
Rice	**0.08**	**0.80**	**0.05**	**0.61**	**0.03**	**0.69**
Cassava	**0.09**	**0.47**	0.04	0.23	**0.05**	**0.55**
Maize	**0.07**	**0.39**	**0.05**	**0.58**	0.02	0.13
Sorghum	0.91	0.02	-0.01	0.02	7.35	0.12
Cotton	-0.43	0.04	-0.76	0.22	5.15	0.02

Table 1.2: Results of trend analysis using OLS regressions (linear or exponential) performed over historical production, area and yield (1997-2016). Slope (β) with a p-value < 0.01, marked in bold, were considered significant.

is millet (CV(RMSE$_{may}$) = 27%) because of its relatively lower inter-annual variability (CV = 27%).

Having a perfect estimator of cropland decreases the error of production estimation for all crops apart from cotton for which an estimator of cropland, even perfect, is virtually useless. It means that the cropland area alone is already a valuable information to improve production estimation. Groundnuts, sorghum, rice and millet have all a CV(RMSE$_{aug}$) in the 20-26% range. However, for highly variable crops (cassava and maize), the error of estimation is still large. Note that in Figure 1.5, we arbitrary set the date at which a perfect cropland would be known (in July). However, the actual cropland is only known in September at the same time than crop area (when it becomes useless). It is, therefore, just an indication of what error is expected before September, if a perfect estimation of cropland area could be obtained. Practically, as a perfect early estimator of cropland does not exist, the CV(RMSE$_{aug}$) achievable with an estimation of cropland will be in-between the one of $\hat{p}_{may}$ and $\hat{p}_{jul}$ depending on the accuracy of the cropland (see section 1.5.4).

In September, the official statistics of crop area are known (Figure 1.1 – Official Statistics). The CV(RMSE$_{may}$) and CV(RMSE$_{sep}$) define the range of error of pre-harvest production estimation. The error should not be larger than CV(RMSE$_{\text{may}}$) and cannot be smaller than CV(RMSE$_{\text{sep}}$) because the yield cannot theoretically be computed before the peak of the growing season in September (Figure 1.1 – Early Estimators). All crops have a CV(RMSE$_{sep}$) below 26% apart from cotton and maize (representing less than 12% of the total agricultural production in Senegal, see Figure 1.4). This result is completely independent from the accuracy of early estimators of cropland area or crop area.

Finally, there is a period between September and November when area is known and yield can be estimated with EO data. A good estimator of crop yield can substantially decrease the error of production estimates. During this period, the best prediction of production can be achieved.

1.5.3 Spatial Variability

Figure 1.6 shows a stratification of Senegal based on $\hat{p}_{may}$ error in each department (using 2010-2016 data). It illustrates the distribution of the production variability in the country. If the production was estimated using $\hat{p}_{may}$ in 35% of all departments (stratum [0-10]), the expected error of total production would be, in average, less than 10% for each crop. For a maximal average error of 20%, the number of departments jumps to 58%. It means that around 40% of the departments explained most of the national inter-annual variability of production for all crops. These departments are also the most productive (pearson correlation of 0.79). (Figure 1.2). This stratification of the country could be used to reallocate the sampling effort to departments with high variability.

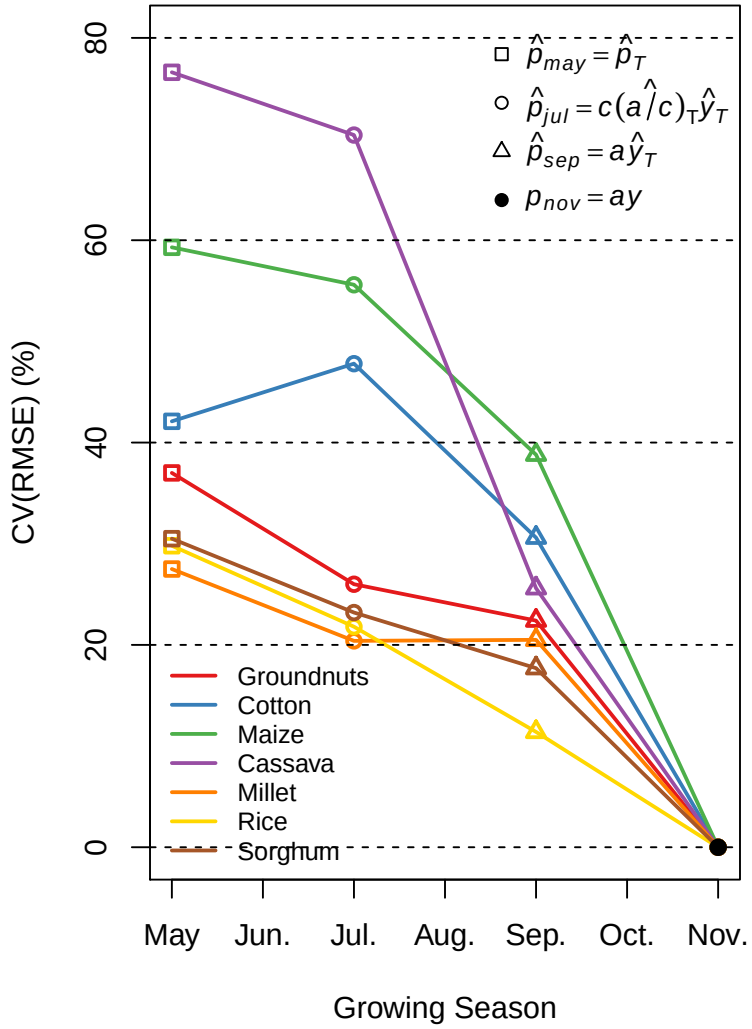


Figure 1.5: Lowest error of production estimation, expressed in CV(RMSE), achievable for each crop according to the increasing data availability along the season (considering perfect estimator of cropland area, $\hat{c} = c$).

1.5.4 Accuracy Requirements of Cropland Area, Crop Area and Crop Yield

Table 1.3 presents the accuracy requirements of cropland area, crop area and crop yield computed using the rules defined in Eq. 1.15. As already mentioned in the previous sections, knowing the cropland area is useless for cotton as $\hat{p}_{jul}$ gives higher errors of production estimation than $\hat{p}_{may}$. On the other hand, the accuracy requirement for cassava cropland is low but, as shown in Figure 1.5, the error of $\hat{p}_{jul}$ is still high. For the other crops, the error should be lower than

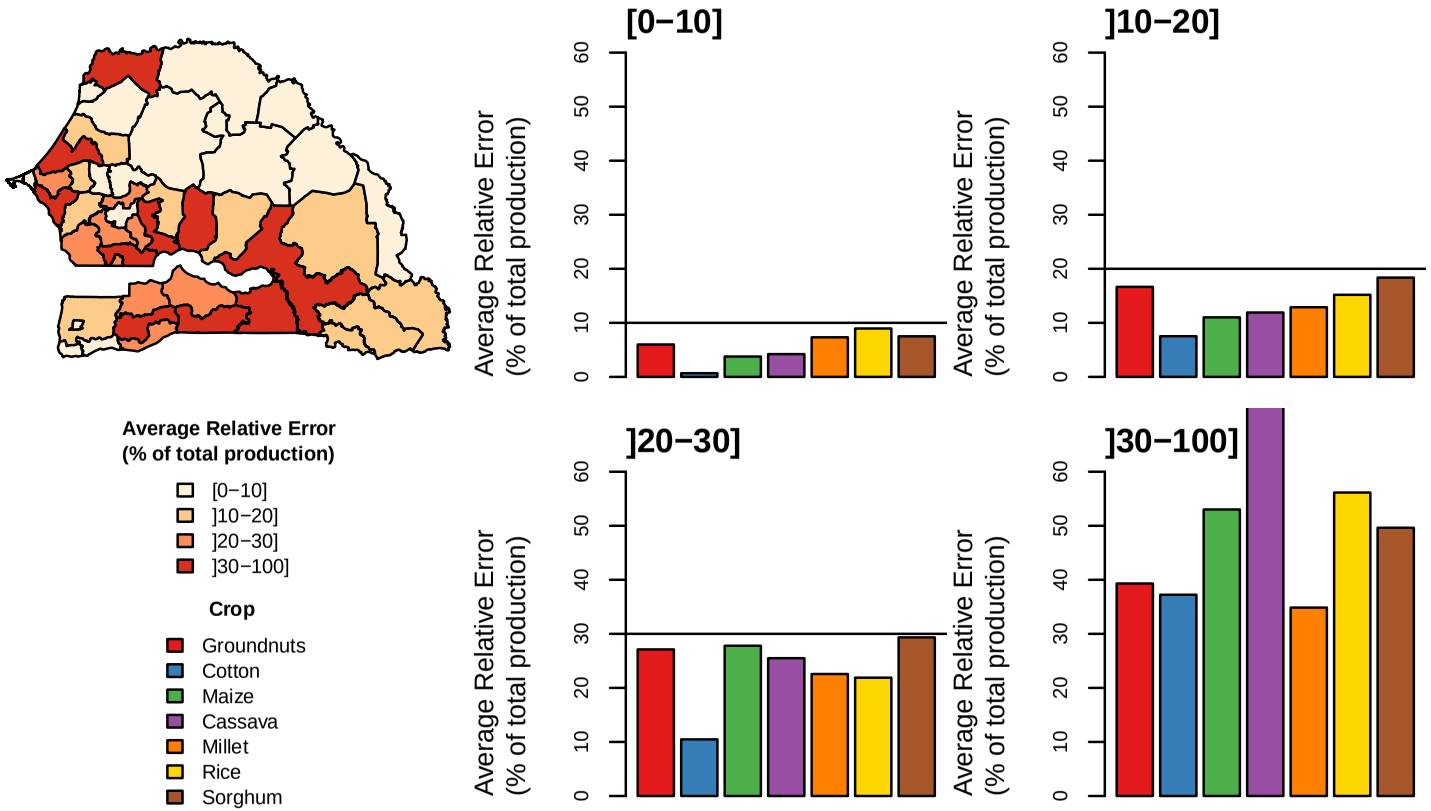


Figure 1.6: Stratification of Senegal based on average $\hat{p}_{may}$ error by department (expressed in percentage of national production) in each department (data from 2010 to 2016). Strata are shown on the map while the average production error per crop and for each stratum is depicted by the bar plots. The number of departments per stratum are 15 in [0-10], 10 in]10-20], 8 in]20-30], and 10 in]30-100].

	Groundnuts	Millet	Rice	Cassava	Maize	Sorghum	Cotton
cropland area, $\hat{c}$	24%	17%	20%	$> 50\%$	17%	17%	-
crop area, $\hat{a}$	27%	17%	23%	50%	35%	24%	27%
crop yield, $\hat{y}$	21%	20%	10%	20%	34%	13%	28%

Table 1.3: Accuracy requirements of cropland area, crop area and crop yield for each crop, expressed in maximum error of estimation (%). The requirements takes the worst case scenario of underestimation and overestimation.

17-24%. As the cropland area is the same for each crop, a cropland classification based on EO data should have at least an overall accuracy of 83% to provide estimation of production more accurate than the historical average and trend ($\hat{p}_{may}$) for all crops. Below 76% of accuracy, such an estimator would be useless as it would not improve the production prediction of any crops compared to $\hat{p}_{may}$ (excluding cassava).

For crop area, the maximum error allowed is a bit higher ranging from 17 to 50%. Excluding cassava, the minimum accuracy of a crop area classification is 65% (and 73% by excluding maize and cassava). And the minimum accuracy to improve each crop is again 83%. Cassava has lower requirements because the CV(RMSE_{may}) was very high for this crop. It is therefore easy to improve its production prediction even with a poor estimator of crop area.

Finally for crop yield, the requirements ranges from an error of 10 to 34%. The lower the CV(RMSE_{sep}), the higher is the requirement. A very accurate estimator of crop yield is needed to improve production estimation of rice and sorghum after September (maximum error of 10 and 13% respectively) while estimates for maize and cotton can be improved easier (maximum error of 34 and 28% respectively).

1.5.5 Combined Requirements for Cropland Area and Crop Area Estimators

Figure 1.7 shows which estimator gives the lowest CV(RMSE) between $\hat{p}_{may}$, $\hat{p}_{jul}$ and $\hat{p}_{aug}$ according to the error of the estimators of cropland area $\hat{c}$ and crop area $\hat{a}$. In other words, this figure indicates which estimator, between $\hat{p}_{jul}$ and $\hat{p}_{aug}$, should primarily be used to estimate production before September. As both $\hat{p}_{jul}$ and $\hat{p}_{aug}$ are available, this case can be observed from August. However, it could be before August if an accurate enough estimator of crop area would be available at that time.

The boundaries of the white areas indicate the accuracy requirements for $\hat{p}_{jul}$ and $\hat{p}_{aug}$ reported in Table 1.3. The whiter a plot, the harder it is to get better estimation of production than $\hat{p}_{may}$. It reflects the low variability of production and/or its predictability by a trend.

Again, the figure shows that cropland is useless for cotton and virtually useless for cassava as even poor estimates of crop area (error $> 50\%$) already give better results than $\hat{p}_{jul}$ no matter the accuracy of $\hat{c}$. For the same error, it is always better to use an estimator of crop area rather than of cropland to predict the production, i.e., CV(RMSE_{jul}) is always lower than CV(RMSE_{aug}) for the same accuracy of $\hat{c}$ and $\hat{a}$. Millet is an exception as for $\hat{c}$ and $\hat{a}$ with identical accuracy, the two estimators $\hat{p}_{jul}$ and $\hat{p}_{aug}$ provide the same results as shown by the 1:1 diagonal boundary between blue and green areas. For groundnuts (and millet), if a very accurate estimator of $\hat{c}$ is available, it might be better to estimate the production with $\hat{p}_{jul}$ rather than $\hat{p}_{aug}$. The effect is less strong for rice and sorghum.

Production is computed from the multiplication of two separated estimators ($\hat{a}$ and $\hat{y}$). The errors of $\hat{a}$ and $\hat{y}$ can either compensate for each other (underestimation times overestimation), or multiply each other (underestimation times underestimation or overestimation times overestimation). This fact means that if, for instance, $\hat{y}$ overestimates the yield then the error of $\hat{p}$ will be lower if $\hat{a}$ underestimates the area. This inevitably leads to asymmetric accuracy requirements depending on the potential compensation of errors. As shown on Figure 1.7, this asymmetry has not a strong effect except for maize ($\hat{p}_{aug}$). This is explained by the yields in 2003 to 2005 that are much higher than the average and are therefore underestimated by $\hat{y_{t|T}}$. For this crop, for the same absolute error, an overestimation of $\hat{a}$ gives better production estimation than an underestimation.

1.5.6 Combined Requirements for Crop Yield and Crop Area Estimators

It has been already mentioned that it was very unlikely to get accurate estimations of yield before the end of the growing season (based on EO data). However, an exact estimator is not necessarily needed. Indeed, any estimator that would provide more precise estimation than $\hat{y}_{it|T}$ can help achieve a more accurate estimation of production than $\hat{p}_{sep}$. Figure 1.8 shows the accuracy requirements of crop yield to get lower production error than $\hat{p}_{may}$ (dotted lines) and $\hat{p}_{sep}$ (full lines) depending on the accuracy of crop area, and vice versa. For the same accuracy, two underestimated components always give lower errors of production than two overestimated (due to the multiplication of errors). On the other hand, the compensation of errors (when one component is underestimated and the other overestimated) can highly decrease the accuracy requirements of $\hat{y}$ and $\hat{a}$. For early warning, only the worst case scenario, when the two components were overestimated (top right of the Figure 1.8), should be considered. When the area is known in September, the accuracy requirements of crop yield are given by the intersection of $\hat{p}_{sep}$ isoline with the vertical line corresponding to an error of 0% for crop area (the values are reported in Table 1.3).

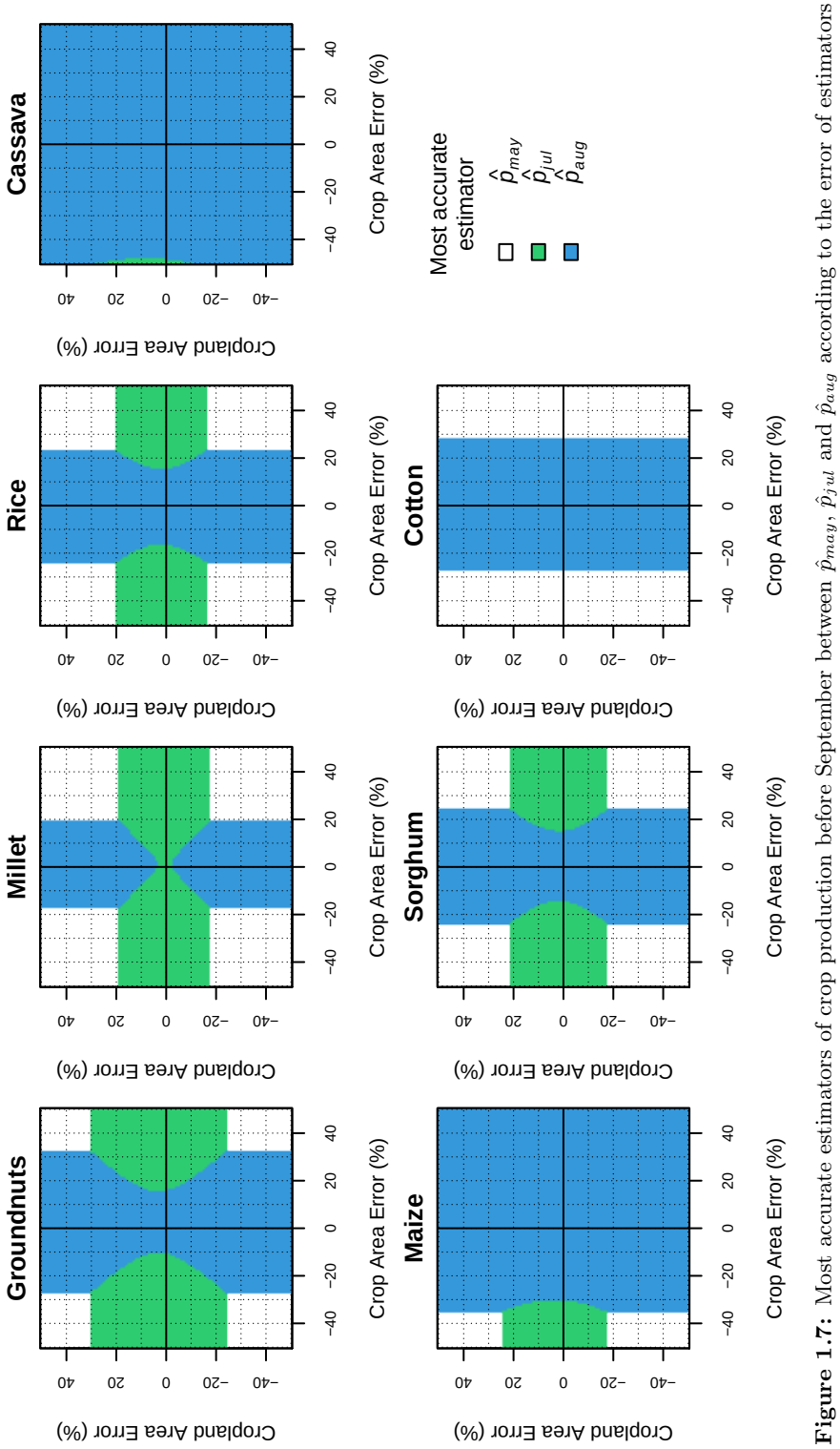


Figure 1.7: Most accurate estimators of crop production before September between $\hat{p}_{may}$, $\hat{p}_{jul}$ and $\hat{p}_{aug}$ according to the error of estimators of cropland $\hat{c}$ and crop area $\hat{a}$.

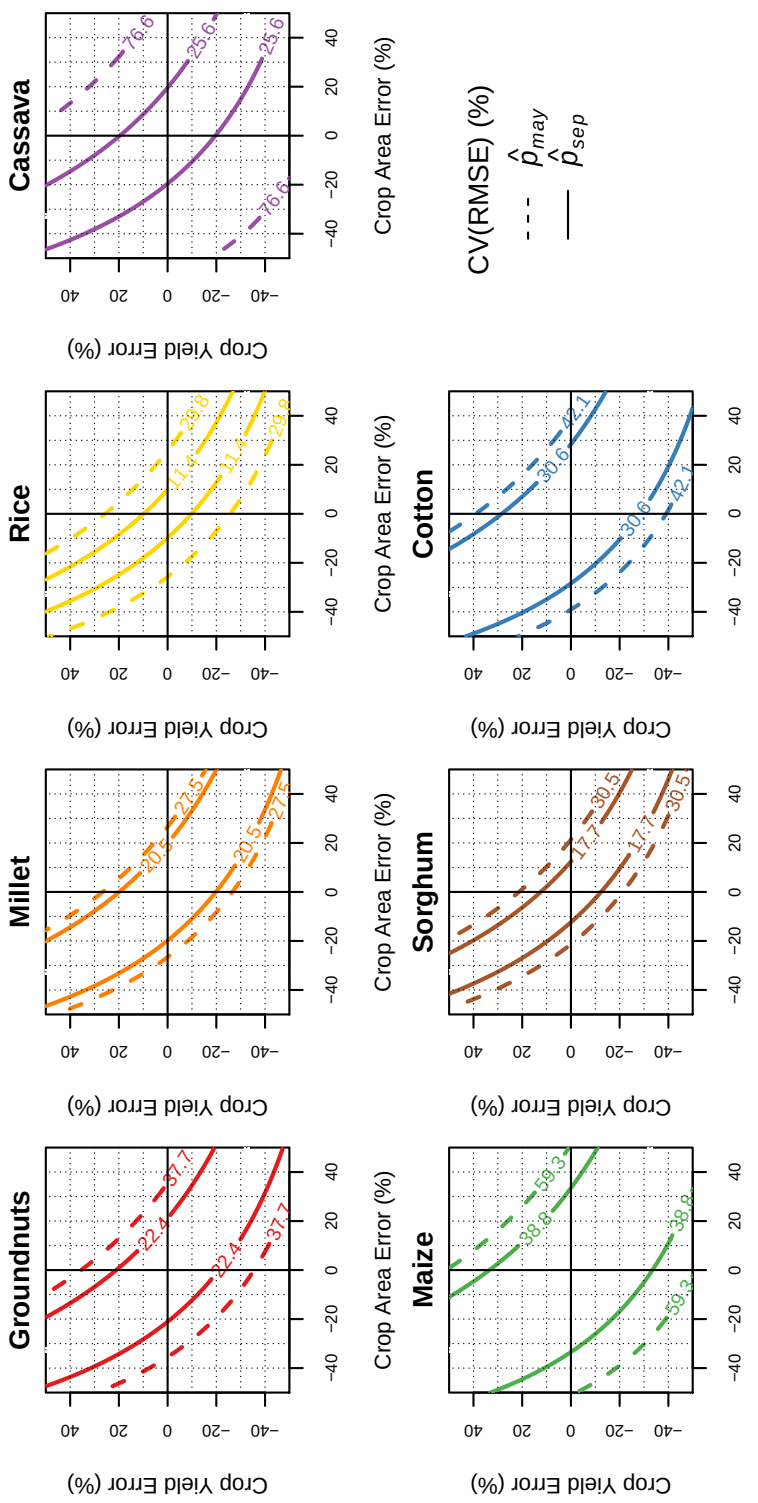


Figure 1.8: Production accuracy, expressed in CV(RMSE) (%), achievable for a combination of yield and area error for each crop. The isolines show the combination of yield and area error giving the same CV(RMSE) than $\hat{p}_{may}$ (dotted lines) and $\hat{p}_{june}$ (full lines). To be useful the combination of errors of the estimators should give a CV(RMSE) that falls within the limits of the dotted lines before September and within the limits of the full lines after September with crop area error = 0% at this date.

1.6 Discussion

Due to supportive programs, significant trends were observed in rice, cassava and maize production (Table 1.2). The trend was substantial for rice by explaining 80% of the production variance. While trend can explain most of the variability of some crops, they generally result from specific policies that can abruptly change from year to year (e.g. due to specific policy). Furthermore, the trend cannot increase indefinitely as biophysical factors limit the yield and the extent of the cultivated areas. Consequently, accounting for the trend for early warning should always be made in light of the current political context and the environmental constraints. More extended time series would also strengthen the analysis. In particular, a long-term trend coming from progressive technical improvement (new equipment, new crop varieties, etc.) would have been easier to model.

From the results, accuracy requirements for early estimators of cropland area, crop area and crop yield can be defined by two key values: the maximum error to improve the estimation of production of all crops (17%, 17%, 10% respectively) and the maximum error to improve the estimation of production of at least one crop (24%, 27%, 34% respectively). We exclude cassava for cropland area and cassava and maize for crop area because the error of production, CV(RMSE), remain large at lower requirements. Not that for Cassava, early estimates of production might not really make sense because the seeding and harvesting of this crop do not necessarily occurred the same year. The number given here for this crop should, therefore, be interpreted with caution. Estimators of cropland area can be used to improve the accuracy of early production estimates of groundnuts, millet and rice, the three main crops in Senegal stressing the value of cropland mapping for food security. To be able to use these requirements operationally, the expected accuracy of $\hat{c}$, $\hat{a}$ and $\hat{y}$ should be known before the beginning of the season. This requires a preliminary benchmarking of each estimator using historical data.

Because the actual area is known at the beginning of the harvest (Figure 1.1), the inter-annual variability of crop yield is the main factor limiting the accuracy of pre-harvest production estimates. The underlying assumption is that the crop yield is not predictable before the harvest. In the Sudano-Sahelian zone, production variability is generally associated with rainfall variability because agriculture is mostly water-limited (D'Alessandro *et al.*, 2015). The yield of rain-fed crops has been shown many times to be correlated with accumulated rainfall (Dennett *et al.*, 1981; Sultan *et al.*, 2013). In Figure A1 in Appendix, we show the correlation between cumulative rainfall over the rainy season and the crop yield. Groundnuts, sorghum, and millet were poorly correlated before September while maize, cotton and cassava showed higher correlation for July. This early correlation for maize, cotton and cassava could be used to improve pre-harvest production estimation.

Early estimations of cropland and crop area can be obtained from the classification of Earth Observation images. The downside of this approach is that the best methods of classification tend to rely on supervised algorithm calibrated with training data. To be efficient, these data should be collected using a sound statistical sampling, directly on the ground (or by photo-interpretation, but this approach requires very high resolution due to the fragmentation of the cropland landscape in Africa). In the context of underfunded NSO, there is little argument to support the organization of two field campaigns (for calibration and official statistics) collecting the exact same information (even tough a smaller sample is needed to calibrate classification algorithms). On the other hand, some classification methods do not require unbiased calibration sample (e.g. maximum likelihood). Furthermore, several methods have been developed to automatically derive a calibration sample using past data, such as the crop map of the previous year, as reference (Matton *et al.*, 2015; Waldner *et al.*, 2017a). Note that for crop yield, simulated data from agro-meteorological models have also been used as calibration set (Burke and Lobell, 2017; Lobell *et al.*, 2015). Therefore, emphasis should be put on the development of such methods rather than on the one relying on ground data. On the other hand, exhaustive mapping has an added value for early warning because it allows to precisely locate the areas with potential crop failures. Furthermore, if major events occur such as a drought or locust invasion, timeliness can be more critical than very accurate estimates. In these cases, photo-interpretation or classification of Earth Observation images can provide precious information for an early diagnosis of the situation (Renier *et al.*, 2015; Rhee *et al.*, 2010).

In this Chapter, crop statistics provided by national offices were considered as the ground truth. However, it only provides another estimation of production impacted by a sampling bias and variance (Stehman, 2005). In Senegal, the quality of the statistical capacities are rather good compared to the African average[1]. The two-stage stratified sampling limits the sampling variance and if a bias exists, it is likely to be similar from year to year since the sampling method is always the same.

We used the CV(RMSE) as the error measurement of production estimators. However, the distribution of errors can be very different for the same value of CV(RMSE) (see Figure A2 in Appendix). The interpretation of such an indicator is, therefore, not necessarily straightforward and prevents the appropriate management of early warning information by decision makers. In a paper on early warning systems for food security in West Africa, Genesio *et al.* (2011) concluded that a better tailoring of information is needed as the interpretation of forecasts indicators is not well understood by the decision makers. In reporting statistical indicators, it is essential to consider the intended audience and their understanding of their statistical meaning as a misinterpretation of information can subsequently impact on decision making (Wallach *et al.*, 2015; Budescu *et al.*, 2009; Morton *et al.*, 2011). Genesio *et al.* (2011) also stressed the fact

[1]Senegal is classified third among all sub-Saharan countries according to the statistical score computed by the World Bank (see Thesis Introduction)

that politically, it is easier to deal with the effect of a disaster than to take action on the basis of uncertain information.

Finally, food availability is only one component of food security. For early warning system, monitoring access to food is capital as it is often regarded as more important than food availability (Sen, 1981). Access is conditioned by several factors including staple prices and household income (Jacques *et al.*, 2018; Pokhriyal and Jacques, 2017). Chapter 3 and 4 of this thesis are specifically tackling these issues thanks to extensive geospatial analysis.

1.7 Conclusions

This Chapter proposed a methodological framework to define accuracy requirement for early warning estimators of crop production according to the inter-annual variability of historical data. The analysis was applied to the seven main crops in Senegal, a country with highly variable agricultural production, using time series of twenty years of production data.

We showed that the inter-annual variability of crop yield was the main factor limiting the accuracy of pre-harvest production estimates because the actual crop area is always known earlier than the crop yield in Senegal. The lower the yearly variability of the yield, or the easier this variability is predictable (e.g. by a trend), the more accurate could be the pre-harvest estimation of production for a specific crop. Interestingly, estimators of cropland area were useful to improve production predictions of the main crops in Senegal stressing the value of cropland mapping for food security.

That being said, apart from rice (mainly predict by the trend), the lowest CV(RMSE) achievable before the harvest (considering a perfect estimator of crop area) were rather high ranging from 18% to 40%. Therefore, quantitatively estimate production for early warning using E0 data might be a futile attempt. On the other hand, EO data are still highly relevant as a source of qualitative information for early warning by, for instance, identifying strong anomalies in vegetation index or rainfall pattern. They also provide precise and exhaustive geospatial information at high frequency which is not given by simple crop statistics. Furthermore, they constitute sometimes the only source of information available in remote or unsecured areas. In another vein, alternative data sources such as citizen science/crowdsourcing might be a promising bottom up approach to get early indicators of food security directly from the ground (Minet *et al.*, 2017).

This analysis was based on the premise that estimates from official statistics were robust and could be considered as the ground truth data. This assumption might be strong as sampling, and non-sampling errors can have a great impact on the accuracy of such statistics. Therefore, the inter-annual variability of production and consequently, the accuracy requirements, could be wrongly estimated. A depth assessment of the quality of the estimation provided by official statistics should, therefore, be carried out first before firmly relying on the outputs of such analysis. More robust results are also expected from more extended time series and analyses at the local scale.

The framework applied in this study is the first step to further research on the inter-annual variability of production in the context of early warning. It stresses the importance of better tailoring early estimators of production in line with the variability of historical data and the calendar of official statistics data collection.

Chapter 2

Mobile Phone Metadata for Development

More Africans have access to cell phone service than piped water.
(CNN, 2016)

Highlights

- Call Detail Records (CDRs), data collected by the telecom companies for billing purpose, have been recently used to tackle development issues thanks to initiatives such as the Data For Development (D4D) Challenge organized by Orange.
- CDRs have been used to model the spread of infectious diseases, study road traffic, support electrification planning strategies or map socio-economic level of population (**relevance**).
- These data can be conceptually described by a geospatial, dynamic, weighted and directed network.
- While massive, CDRs are not statistically representative of the whole population due to several sources of bias (market, usage, spatial and temporal resolution) (**accuracy**).
- Personal information can be extracted from CDRs causing a significant threat to privacy (**protection**).
- The potential of such data might exceed their limitations. Compared to traditional data collected to compute official statistics, CDRs are cost-effective and have the potential to provide near real-time insights about the situation of million of individuals (**timeliness, access**).

Abstract

Mobile phones are now widely adopted by most of the world population. Each time a call is made (or a SMS sent), a Call Detail Record (CDR) is generated by the telecom companies for billing purpose. These metadata provide information on when, how, from where and with whom we communicate. Conceptually, they can be described as a geospatial, dynamic, weighted and directed network. Applications of CDRs for development are numerous. They have been used to model the spread of infectious diseases, study road traffic, support electrification planning strategies or map socio-economic level of population. While massive, CDRs are not statistically representative of the whole population due to several sources of bias (market, usage, spatial and temporal resolution). Furthermore, mobile phone metadata are held by telecom companies. Consequently, their access is not necessarily straightforward and can seriously hamper any operational application. Finally, a trade-off exists between privacy and utility when using sensitive data like CDRs. New initiatives such as Open Algorithm might help to deal with these fundamental questions by allowing researchers to run algorithms on the data that remain safely stored behind the firewall of the providers.

2.1 Introduction

Over the past two decades, access to telecommunication services has seen exponential growth. From around 100 million in 1995, the number of mobile cellular subscriptions has risen to 7.4 billion worldwide in 2016 – the equivalent of the entire world population (ITU World Telecommunication, 2016b). No technology has ever spread faster around the world (The Economist, 2008). This growth was primarily driven by wireless technologies and liberalization of telecommunication markets, along with new financing and technology, which have enabled faster and less costly network rollout.

However, this does not mean that every person in the world has subscribed to a mobile service. Because many individuals own several handsets or have multiple subscriber identity module (SIM) cards, the number of subscribers, estimated to 4.7 billion worldwide, is substantially lower than the number of subscriptions (GSMA Intelligence, 2016). This is because the number of subscriptions tend to exaggerate the mobile phone penetration rate[1] in developed economies. On the other hand, in many developing countries, mobile phone access is higher than subscription numbers would suggest. Access is indeed fostered in countries where sharing mobile phones is a common practice, especially within large households (Aker and Mbiti, 2010). In a world bank report, the practical impact of the difference between subscription and household penetration[2] is clearly explained (World Bank, 2012):

> "Take Senegal, where the subscription penetration was 57 per 100 people in 2009, but household penetration was estimated to be 30 points higher at 87. This larger household size can dramatically extend access to mobile phones, considering that on average nine persons are in each Senegalese household." It results that "several low-income nations have higher mobile phone home penetration than some developed economies. For example, Senegal, along with some other low- and middle-income economies, has a higher proportion of homes with mobile phones than either Canada or the United States".

The economic potential of the mobile phone is tremendous. 2015 has been a year of continued growth in the mobile industry, with operator revenues exceeding $1 trillion. The mobile ecosystem generated 4.2% of Gross World Product and directly support 17 million jobs (GSMA Intelligence, 2016). The mobile telephony is increasingly recognized as an essential tool of development by improving the flow of information and providing a platform for financial services (Aker and Blumenstock, 2014). The more striking example lies in the mobile money service which is now widely established and brings financial inclusion to previously unbanked and underbanked populations across the developing world (1.9 billion people globally). Mobile phones are also a key platform to bring internet access to people across the globe, particularly in

[1] The mobile phone penetration rate is the number of active mobile phone users per 100 people within a specific population.

[2] Portion of total households having access to mobile phone within a specific population.

developing regions where fixed broadband services are prohibitively expensive, and fixed-line infrastructure is limited. At the end of 2015, 2.5 billion individuals from the developing world were accessing the internet through mobile devices (GSMA Intelligence, 2016).

Modern mobile phones (smartphones) are now integrated computers with dozens of embedded sensors, such as accelerometer, digital compass, gyroscope, GPS, microphone, and camera, which enable the emergence of several research applications based on personal sensing (Lane *et al.*, 2010). These are promising avenues that are believed to revolutionize many sectors of the economy, but the penetration of smartphones is still low in Africa because of their higher cost. For example, only 19% of the Senegalese population reported owning a smartphone in 2015 (Poushter, 2016).

On the other hand, even the most basic handset passively generates a vast amount of metadata leaving behind a digital trace[3] of the activity of its user. These metadata provide information on when, how, from where and with whom we communicate (Blondel *et al.*, 2015). At first, researchers realized the potential of such data by uploading tracking software into consenting subjects' phones through the Reality Mining project of the MIT[4] (Eagle and Pentland, 2006). They later gained access to actual metadata directly from mobile network providers, leading to larger-scale research and greater analytical power (e.g., Gonzalez *et al.*, 2008). Until then, more and more datasets were opened up to the scientific community, and mobile phone metadata are now seen as a typical example of empirical data used in network science (Barabási, 2016). The applications are tremendous, particularly for studies related to mobility, social network and socio-demographics of people (Blondel *et al.*, 2015; Saramäki and Moro, 2015; Naboulsi *et al.*, 2016). Several initiatives have emerged, such as the Data For Development (D4D) challenge organized by Orange, that provided datasets to the research community for projects related to development. In a recent survey carried out by the World Bank, mobile phone data appeared at the top position, just before satellite imagery, in the Big Data sources used in SDG-related projects (Ballivian, 2014).

The objective of this Chapter is to introduce mobile phone metadata, in particular call data records used by the companies for billing purpose, and their potential for the SDGs. We first present some elements of mobile network infrastructure, a prerequisite to understand the characteristics of data collected by a mobile network operator. We then describe the specific features that make CDRs unique and how they can be used to help achieving the SDGs. Finally, we discuss the statistical limitations of such data and the risks associated with their use (in particular, data access and privacy).

[3]Also called digital shadow, digital footprint or data exhaust.

[4]Note that as early as 1999, it was already demonstrated that OD matrix could be obtained from the localisation of mobile phones (White and Wells, 2002). But the potential of CDRs for computational social science was discovered later.

2.2 Elements of Mobile Network Operator Infrastructure

This section has been written based on material developed in Swenson *et al.* (2006); Tiru (2014); Ricciato *et al.* (2015); Janecek *et al.* (2015); Ricciato *et al.* (2017).

Almost all Mobile Network Operators (MNO) in the world use two main mobile technologies – GSM (Global System for Mobiles) and CDMA (Code Division Multiple Access)[5]. The market share of subscribers using CDMA worldwide is 15-25% (mostly in North-America and some Asian countries) and 75-85% for GSM (the rest of the world). The main difference between these technologies is the radio signaling technology. The practical implication is that a mobile device is tied to a particular network within CDMA network, while a Subscriber Identity Module (SIM) card is tied to a specific network within GSM network. It is, therefore, easier to switch mobile devices within GSM networks thanks to SIM card's portability. GSM phones without CDMA support cannot run within CDMA network (and conversely). Throughout this thesis, data from GSM network are used but most of the subsequent observations are universal and apply equally independently of the technology.

A GSM network is a radio network of individual cells, known as Base Transceiver Station (BTS). The BTS is responsible for transmitting and emitting radio communications between the network and the mobile devices which on the ground can be identified by the antenna tower and equipment (Figure 2.1, left). To improve the network efficiency, BTSs are hierarchically grouped in Location Area (LA) and controlled by a Base Station Controller (BSC) (Figure 2.1 and 2.2). The BSC is responsible for handover procedures[6] within a single LA between one BTS to another. BSCs are controlled by Mobile Switching Centres (MSCs) that accommodates the Visitor Location Register (VLR) – the registry for holding the information about the LA in which the mobile devices is located (Figure 2.2). Finally, MSCs report to the Network Management System (NMS) where all administrative and central procedures reside. Usually, only data that are transmitted from MSCs to NMS are stored for different purposes while lower level data traffic are deleted. NMS accommodates different registries and databases that are important for the network functioning, in particular, billing databases (Call Data Records[7]). Components of Mobile Positioning System, the system used for pinpointing users' location for emergency and security services (e.g., E112 and E911 directives in EU and US), also resides in NMS and MSCs.

Different MNOs may share some of the network equipment. For example, it is not uncommon that different MNOs share the same BTS. There are also

[5]CDMA may disappear in favor of GSM network due to the spread of the fastest and high-quality Long Term Evolution (LTE) technology that uses a similar technology as GSM.
[6]i.e. BTS switch when a call in progress moves from one base station to another.
[7]Call Detail Records and Call Data Records are used interchangeably in this thesis.

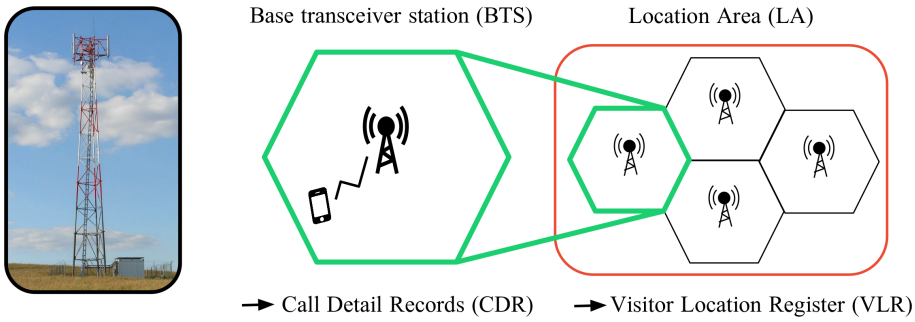


Figure 2.1: The two main sources of location data collected by Mobile Network Operators: the Base Transceiver Station (shown in green and picture on the left) is stored in the Call Detail Records and the Location Area (shown in red) is stored in the Visitor Location Register (VLR).

special types of MNOs (virtual MNOs) that do not possess any network infrastructure, but instead rent it from other MNOs. In such case usually the virtual MNOs do not have access to all operational data.

Mobile phones that are switched on are either in idle or in active state. In idle state a mobile phone is not allocated any radio resource, but it constantly evaluate if it needs to switch to another cell with a better signal strength (it listens but does not transmit). Thus, in idle state a mobile phone receives passively, which implies that the network is unable to identify cell changes of idle mobile phones, except when such a switch is explicitly requested by the mobile phone. Cell switches within one LA (from one BTS to another) are not reported, but cell shifts from one LA to another are. It turns out that the LA of any switched-on mobile phone is known at any time by the MNO.

A mobile phone remains most of the time in idle state and only become active during a call or a data transaction. When a mobile phone user dials a number to make a call, a call initiation request is sent to the MSC. The MSC validates the request by checking the user's identity and airtime balance in the records of its database. If valid, a connexion is initiated with the third party, and the MSC requests the base station to move the mobile phone to an unused voice channel so that the call can begin. Once a call is in progress, the MSC adjusts the power transmitted by the mobile phone as it moves in and out of the coverage area of each base station. When a mobile phone with a call in progress moves from one BTS to another, handover procedure are automatically managed by the network.

In summary, the state of the mobile phone (idle or active) determines the temporal and spatial accuracy of the user location within the network data system.

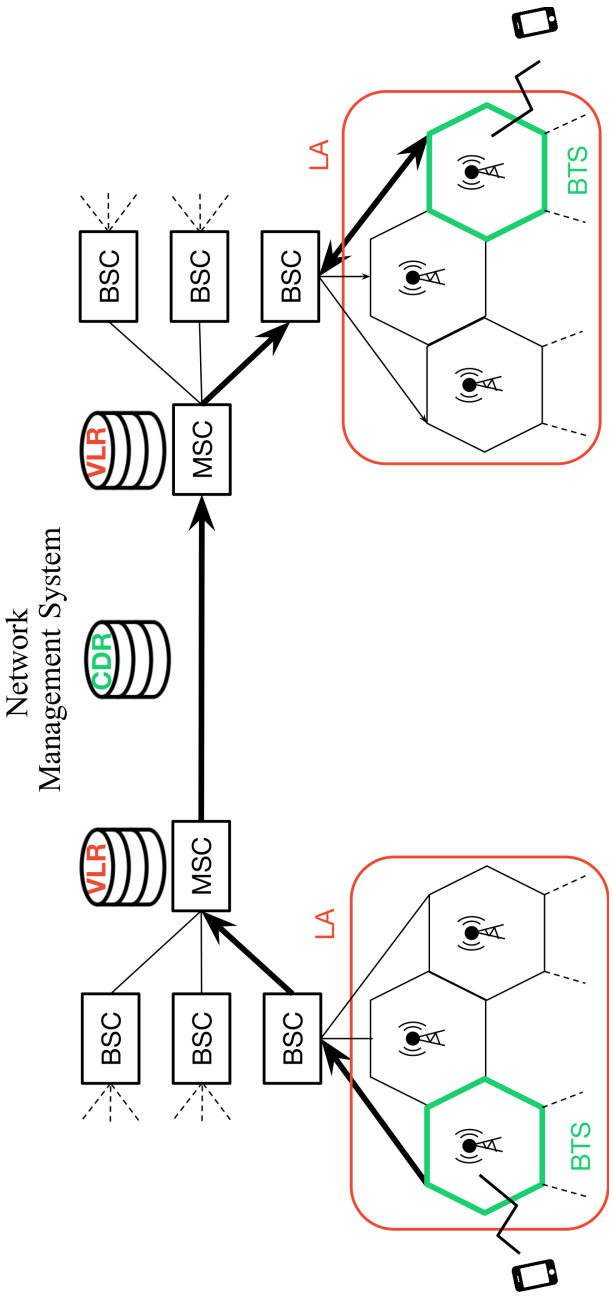


Figure 2.2: Simplified structure of a GSM network. MSC stands for Mobile Switching Center, BSC for Base Station Controller, LA for Location Area, BTS for Base Transceiver Station, VLR for Visitor Location Register and CDR for Call Detail Records. Arrows indicate the propagation of the signal needed to locate the recipient when a call is initiated.

2.3 Mobile Phone Metadata

2.3.1 Call Detail Records (CDR)

CDRs typically include BTS location information of the caller and recipient (starting cell), as well as time stamp and call duration (Table 2.1). However, the information stored in CDRs really depends on each MNO as no standardised structure exists (e.g., whether long calls are chunked into multiple CDRs) (Tartarelli *et al.*, 2010). Contrary to what the term suggests, CDRs are also used for SMS and data connection (sometimes stored in separate data files). CDR data contain cell-level locations, but only for *active* mobile phone engaged in voice call, SMS or data connections. Their use for billing purpose implies an archived and a constant update of the data without changing nor deleting old records. This means that long time series are easily accessible for data analytics.

2.3.2 Visitor Location Register (VLR)

The VLR is a dynamic database supporting the operation of the Mobile Switching Center (MSC). Principally, the VLR caches temporary data about the current LA location of all mobiles, both active and idle. Due to the completeness of VLR data, VLR records provide an instantaneous description of the location of all mobile phones, at LA level. VLR data are highly dynamic because LA locations are updated continuously. As a result, bulk VLR reading can only be conducted in real-time in the background of network operations.

2.3.3 Passive Monitoring Systems

Some MNOs track signaling and traffic exchange (e.g. for handover) in the network through passive monitoring systems. These systems aim to assess and resolve network operation and troubleshooting (Tartarelli *et al.*, 2010). Using the network data, passive monitoring systems are able to locate every mobile phone with great accuracy, both in terms of time and space. Locations at cell and LA level are provided, both for active and for idle mobile phones.

Table 2.1: Sample of typical call data records.

Caller SIM	Callee SIM	Outgoing BTS	Incoming BTS	Timestamp	Call duration (sec)
0458685984	0488595496	12	365	2018-01-18 15:22:12	456
0458685984	0458685984	12	25	2018-01-18 22:24:12	35
0469875254	0498563201	879	567	2018-01-19 08:47:10	125
(...)	(...)	(...)	(...)	(...)	(...)

2.3.4 Selecting a dataset

Selecting a data source for analysis is a trade-off between spatial and temporal resolution, and data accessibility (Ricciato *et al.*, 2017). CDRs provide information at the highest spatial resolution (cell level, see Figure 2.4) but are event-based. Therefore, data are only available when the user makes a call (or send an SMS/data). This can be a limitation for mobility studies that require regular location update (see Statistical limitations section). On the other hand, the mobile phone usage (number of calls, when and where, etc.) might provide precious information on the user's socio-demographic profile. CDRs are stored offline and therefore easily accessible. Because of that, these are, by far, the data most frequently used in the literature (Blondel *et al.*, 2015).

VLR provide data at the finest temporal resolution but at LA level. The spatial resolution of LA is much lower than BTS. For instance, the Ile-de-France region has almost 10,000 BTS grouped in only 32 LAs, each of which has between 150 and 500 BTS (Bonnel *et al.*, 2015).

Lastly, passive monitoring systems leveraging signaling data combine the best of the two worlds. However, the systematic acquisition of such data are generally based on proprietary system and not available to all operators. As a result, very little research has been conducted using these data (Janecek *et al.*, 2015; Valerio *et al.*, 2009).

Other data collected by MNO that are relevant for research applications include **airtime credit purchases** and **customer client profiles**. Airtime credit purchases are useful to predict socio-economic status of population (Gutierrez *et al.*, 2013a; Decuyper *et al.*, 2014; Blumenstock, 2015) while customer client profile enrich mobile phone metadata with personnal information such as gender and age. However, while mobile phone operators have access to all the information filed by their customers and their operational data, they generally limit the access to only a sample of what is available depending on their own privacy policies and the regulation on privacy protection of each country.

In this thesis, we use one year of CDR data provided by Orange-Sonatel in the frame of the D4D challenge and covering 9 million of customers (all Orange customer) in Senegal (de Montjoye *et al.*, 2014). Only users with an average of more than 1000 interactions per week (presumed to be machines or shared phones) were removed from the database.

The following sections further detail the specific characteristics of CDRs.

2.4 Data Features

Conceptually, CDRs can be described as a geospatial, dynamic, weighted and directed network (Figure 2.3 – A). In the following sections, each of these dimensions is further developed.

2.4.1 Network Dimension

A network is defined by its nodes (or vertex) and links (or edges, ties). In CDRs, nodes are SIM cards ($\sim$ mobile phones) and links are the calls (or SMS) exchanged between two SIMs. Furthermore, the links are directed from the caller to the callee and the duration of calls weights each tie (Figure 2.3 – A). By neglecting the geospatial dimension and aggregating (e.g., by taking the sum) the total call duration between each pair of users over a given period, a static representation of the CDRs network can be obtained (Figure 2.3 – B). After temporal aggregation, the link weights of the network might also represent the number of calls or SMS exchanged between two users.

According to graph theory, such network can be represented through its adjacency matrix A_{ij} (Eq. 2.1).

$$A_{ij} = \begin{pmatrix} 0 & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & 0 & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & 0 \end{pmatrix} \tag{2.1}$$

The values of A_{ij} represent the weight of the links (call duration, number of calls or number of SMS) between node i and j. Because the network is directed, the matrix is asymmetric and $A_{ij} \neq A_{ji}$. Note that $A_{ii} = 0$ because one user cannot call himself.

The topological structure (degree, density, connectivity, etc.) of CDRs network inform on the characteristics of people's social networks. Such analysis lead to the emergence of a new field of research called computational social science (Lazer *et al.*, 2009). One interesting possibility offered by CDR network is the detection of social communities by identifying group of users that interact more with each other than with the rest of the population. This requires efficient algorithm capable to handle large dataset (see for instance the 'Louvain method' in Blondel *et al.*, 2008).

2.4.2 Geospatial Dimension

As previously explained, each call in CDRs is geolocated at the BTS level (Figure 2.4 – A) so that a mobile phone user can be located within the coverage of this BTS. One BTS generally accommodate more than one antenna (typically three). Therefore, the location of the user can be more precisely defined within the coverage of one antenna (Figure 2.4 – B). However, MNOs might only provide CDRs at BTS scale to preserve anonymity (e.g., the D4D dataset).

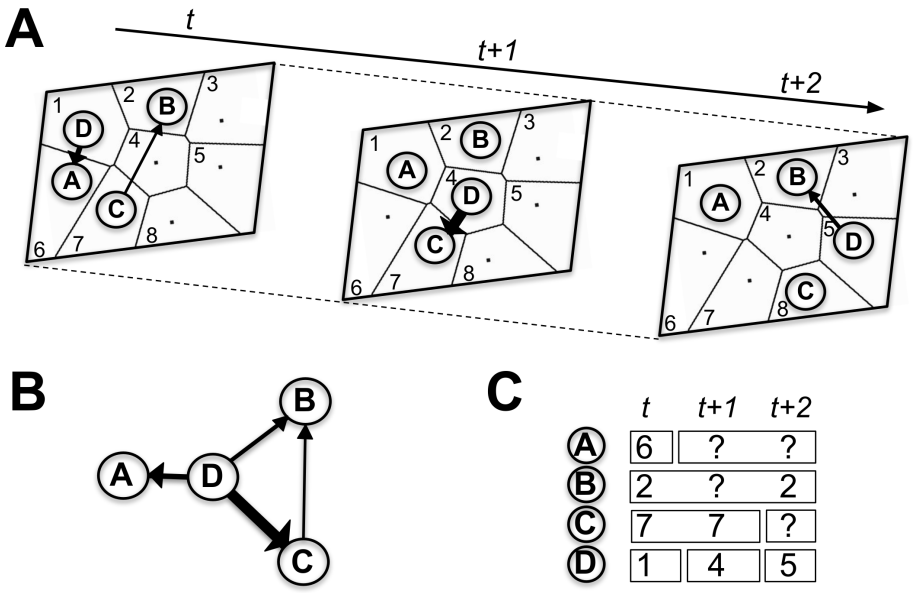


Figure 2.3: Schematic representations of CDRs data. Letters (A-D) represent SIM cards ($\sim$ individuals), numbers (1-8) represent antenna coverage approximated by a Voronoï tesselation, and arrows represent call direction (head) and duration (width). (A) geo-spatial, dynamic, directed weighted network (here weights are call duration), (B) static, directed weighted network (over t to $t+2$ period), and (C) dynamic trajectories of SIM cards.

The coverage of one particular antenna depends on its technical characteristics (power, technology, etc.) and the BTS density (Figure 2.5). More antennas are required in areas with more traffic which means that in urban and sub-urban areas, cell areas typically span between hundreds of meters (micro-cells) and a few kilometers of diameter, while sparsely populated areas are covered by few macro-cells. Smaller cells (pico-cell and femto-cells) can also be deployed in highly crowded areas, such as shopping malls, train stations, or airports. The antenna density is still considerably lower in developing countries than in developed areas (see for instance Figure 2.5). It is worth mentioning that analyzing the characteristics of the signal exchanged between the phones and the BTS allow to reduce spatial uncertainty by inferring the user-antenna distance (using response-delay and strength) (Figure 2.4 – C). Finally, triangulation can further increase the spatial accuracy of the estimation of the user's location (Figure 2.4 – D). Such procedure usually requires authorization and approval from the user except for emergency response such as E112 and E911 directives in EU and US.

2.4.3 Temporal Dimension

The temporal resolution of CDRs depends on the activity of mobile phone users. The best case is a user multiplying short event (SMS or short calls). It is worth

noting that a long call does not increase the temporal resolution as most of the time, only the starting cell (where the call was initiated) is recorded in CDRs. While it is a limitation for mobility analysis, the temporal patterns of mobile phone usage is interesting to study the socio-economic profiles of population (Blumenstock *et al.*, 2015a; Soto *et al.*, 2011).

On the other hand, the temporal dimension of CDRs allows studying the dynamic of social networks. Saramäki and Moro (2015) showed the great value of this approach for different scales of analysis (Figure 2.6). For instance, one can explore the resilience of social ties for an individual or a community after a shock such as a loss (at individual level) or a disaster (at community level).

Finally, one key element of CDRs is their potential for near real time applications. As the data are collected on the fly, they are virtually instantly available for analysis. This aspect has significant implications for post-disaster monitoring and early warning systems.

2.4.4 Spatio-Temporal Dimension

The combination of both spatial and temporal dimension provide valuable information for dynamic population mapping and mobility analysis as well as for land use classification.

The most straightforward application of CDRs is the estimation of the number of people at specific place and time. This implies modeling the relationship between the number of active mobile phone users with the actual population over time and space. Using census as calibration data, accurate estimations of population can be obtained during night-time (Deville *et al.*, 2014b). However, the fact that the relationship between active mobile phone users and the actual population is not independent of time and space makes it harder to accurately extrapolate these models during day-time. This would require calibration data for daytime population which are rarely available.

From the CDRs, a spatial trajectory can be computed for each user and use in mobility analyses (Figure 2.3 – C). The quality of the inference of such an approach depends on the frequency of the user activity (e.g., in Figure 2.3 – C, mobility of user D will be better predicted than user A). In particular, mobile phone activity is known to be bursty (Barabasi, 2005). Users tend

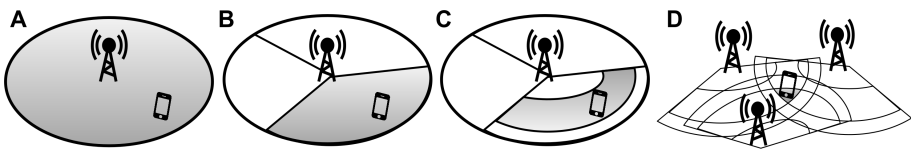


Figure 2.4: Location at (A) the base station level, (B) the sector level, (C) the sector level knowing signal characteristics, (D) triangulation.

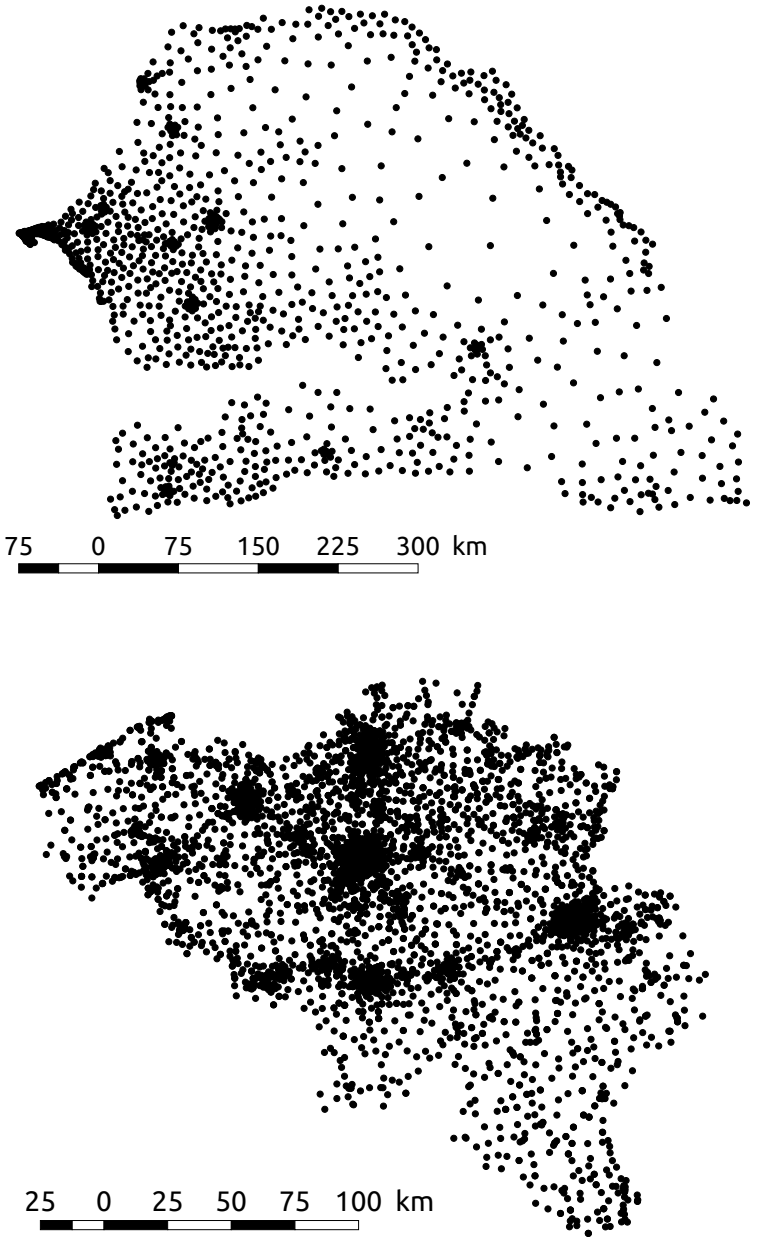


Figure 2.5: BTS maps of Orange Sonatel in Senegal (top) and Orange Mobistar in Belgium (bottom).

to place most of their calls in short bursts, followed by long periods with no

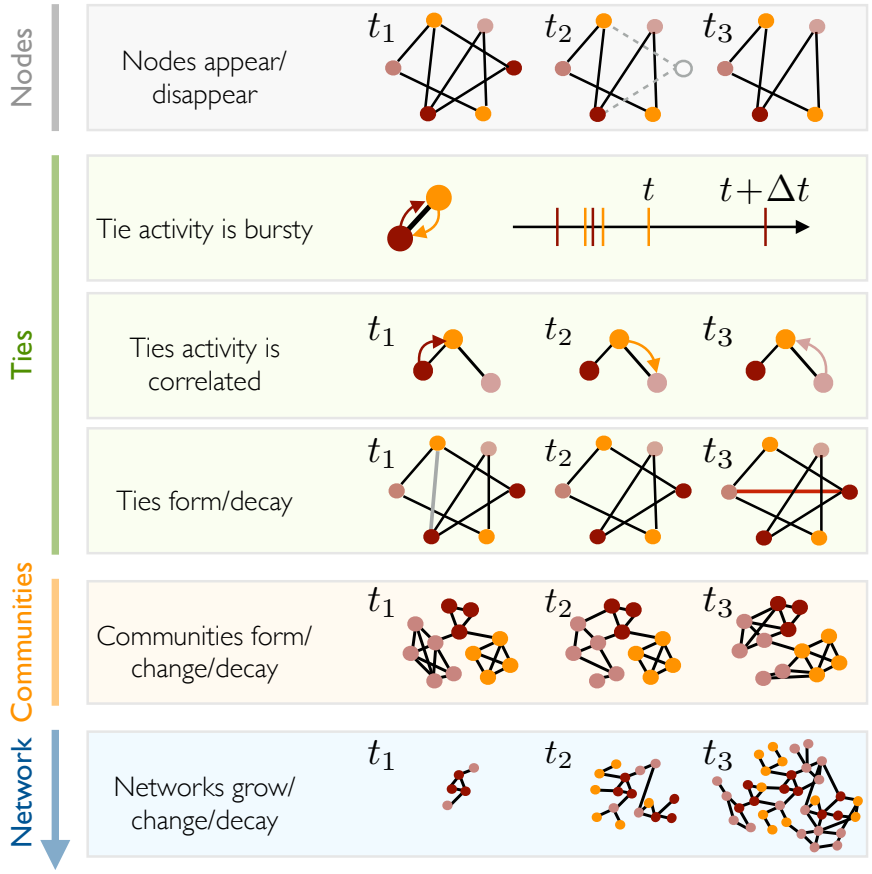


Figure 2.6: Temporal features of network at different structural and spatial scales. Figure from Saramäki and Moro (2015) reproduced with permission of the authors.

call activity, during which information about the user's location is lacking[8]. However, despite their temporal sparseness and spatial coarseness, CDRs still offer great insights into the movement patterns of individuals and communities (Becker *et al.*, 2013). Furthermore, Song *et al.* (2010a) showed that human mobility was highly predictable, regardless of the distance traveled, due to the regularity of our daily mobility. A typical application of mobility analysis with CDRs is the computation of origin-destination matrices at different temporal scales. It allows identifying different mobility dynamics such as daily commuting and long-term migration.

On the other hand, the temporal signature of BTS activity can be used to define land use patterns. This method has been applied in urban areas to make

[8]This observation also applies to several other human activities such as web browsing, stock trading or library visits (Vázquez *et al.*, 2006; Barabási, 2010).

the distinction, between residential, business, industrial and leisure areas among others (Lenormand *et al.*, 2015).

2.5 Applications of Mobile Phone Metadata for Development

The use of CDRs for data for development is relatively new. The growing interest in the field has been triggered by telecom companies opening large datasets to the research community. The Data For Development (D4D) challenge, launched by Orange in 2013, was the first release of an extensive CDR database from an African country (Ivory Coast) to the international research community (Blondel *et al.*, 2012). It was also the first CDRs-related project to be labeled as 'development', and gained huge publicity after the United Nations, the World Economic Forum and several high-profile academic institutions (including MIT and Cambridge University) endorsed it (Taylor, 2015). The initiative was very successful and resulted in dozens of innovative projects developed by research labs from around the world. Encouraged by this success, a second D4D challenge, with data from Senegal, was organized in 2014. Once again, it resulted in several creative projects (de Montjoye *et al.*, 2014). The pioneer work of the Big Data for Good team of Telefonica[9] in the development of algorithms for inferring socio-economic welfare from mobile phone use patterns should also be acknowledged.

Hereafter, we briefly present five contrasted applications of CDRs analysis for development in five different developing countries (Kenya, Haiti, Rwanda, Ivory Coast, Senegal). The aim here is not to give an exhaustive review of all the possible applications of CDRs but rather a quick overview of what can been done. For a detailed and broad review of CDRs data analysis, the reader is referred to the works of Blondel *et al.* (2015), Naboulsi *et al.* (2016) and Saramäki and Moro (2015).

2.5.1 Health

One of the most common (and successful) use of CDRs data for development is for epidemiological studies of human infectious diseases. For instance, Wesolowski *et al.* (2012b) used CDRs from Kenya to identify the dynamics of human carriers that drive parasite importation between regions. They analyzed the regional travel patterns of nearly 15 million individuals over the course of a year and characterized the degree of connectivity among different areas in Kenya. Using a simple transmission model and malaria infection prevalence data, they were then able to map the importation routes that contribute the most to malaria epidemiology on regional spatial scales.

[9]Telefonica is a Spanish multinational broadband and telecommunications provider serving over 200 million users in Latin America, Europe, and the United States.

2.5.2 Post-Disaster Management

Another application that greatly benefits from mobility data derived from CDRs is crisis management following a disaster. Lu *et al.* (2012) analyzed the movements of nearly two million SIM card holders before and after the 2010 earthquake in Haiti, finding that one-fifth of Port-au-Prince's residents left the city by three weeks after the disaster. They also show that the trajectory of people fleeing from regions hit by the earthquake was highly correlated to their mobility patterns during normal times. Such findings suggested that population movements during disasters are significantly more predictable than previously thought and highly influenced by people's social support structures.

2.5.3 Poverty and Socio-Economics Level

Assessing socio-economics levels, in particular poverty prevalence, is another recent development of mobile phone metadata. Blumenstock *et al.* (2015a) showed how the individual's past history of mobile phone use can inferred his/her socio-economic status using records of billions of interactions on Rwanda's largest mobile phone network. They validated their approach with a phone surveys of a geographically stratified random sample of 856 individual subscribers and using a DHS composite wealth index at micro-region level. Chapter 4 of this thesis explores this topic further.

2.5.4 Transportation

Using CDR data from the first D4D for the city of Abidjan (Ivory Coast), Berlingerio *et al.* (2013) evaluated which new routes would best improve the existing transit network to increase ridership and user satisfaction, both in terms of reduced travel and wait time. Four new routes have been proposed by the optimization system (called AllAboard), resulting in an expected reduction of 10% city-wide travel times.

2.5.5 Energy

The first prize of the D4D Senegal challenge was awarded to a research project which assessed the contribution of mobile phone data for the development of bottom-up energy demand models in Senegal (Martinez-Cesena *et al.*, 2015). Specifically, the research team introduced a framework that combines mobile phone data analysis (mobile phone activity was used as a proxy of the energy consumption), socio-economic, geo-referenced data analysis, and state-of-the-art energy infrastructure engineering techniques to assess the techno-economic feasibility of different centralized and decentralized electrification options for rural areas in a developing country. The result was a country map of electrification recommended option between (i) extensions of the existing medium voltage grid, (ii) diesel engine-based community-level Microgrids, and (iii) individual household-level solar photovoltaic systems.

2.6 Statistical Limitations

CDRs are a good example of Big Data source that can be diverted from their primary purpose to approximate socio-economic variables and population mobility. As they are not designed for this purpose, this means that an unavoidable bias will always impact any application based on these data. If not properly understood, this could lead to serious misinterpretation of the results and ultimately, have harmful impacts in misleading policy-makers. This section reviews some of the sources of inaccuracy inherent to mobile phone metadata.

2.6.1 Technical issues

MNOs suffer occasional down-time during which data are not recorded (missing data). Furthermore, cells can also be deactivated for maintenance or resource optimization (e.g., during low activity period such as nighttime). On the other hand, incorrect data can arise at different level of the data collection due to encoding or other technical issues (e.g. duplicated records, records with incorrect time values, etc.).

2.6.2 Selection bias

People generating CDR data have selected themselves as data generators through their activity. This is called a 'selection bias'. First, while the penetration of mobile phone is very high in the developing world, some sociodemographic groups (typically young children and senior people) are still left out of the analysis when considering mobile phone metadata. The adoption base in Africa has been more traditionally skewed towards a wealthier, educated, urban and predominantly male population (Aker and Mbiti, 2010; Blumenstock and Eagle, 2010). Additionally, one SIM card does not necessarily correspond to one person. Figure 2.7 illustrates all the possible association schemes between SIM and persons. In developing countries, this is frequent that someone owns different SIM cards to be able to switch between mobile carrier's network depending on promotional campaigns. Phone sharing is also a common practice among the poorest. On the other hand, data access is most often limited to only one provider in the country. This could be problematic if the choice of a MNO is correlated with the socio-economic profile of individuals.

However, depending on the application, the impact of ownership bias might not be as strong as expected. For instance, Wesolowski *et al.* (2013) show that mobility estimates are surprisingly robust to the substantial biases in phone ownership across different geographical and socio-economic groups using 1-year data of 15 million individuals in Kenya.

2.6.3 Spatial bias

Lacking data on antenna power and orientation, their coverages are generally approximated by means of a Voronoï tessellation (Figure 2.3 – A). It assumes that

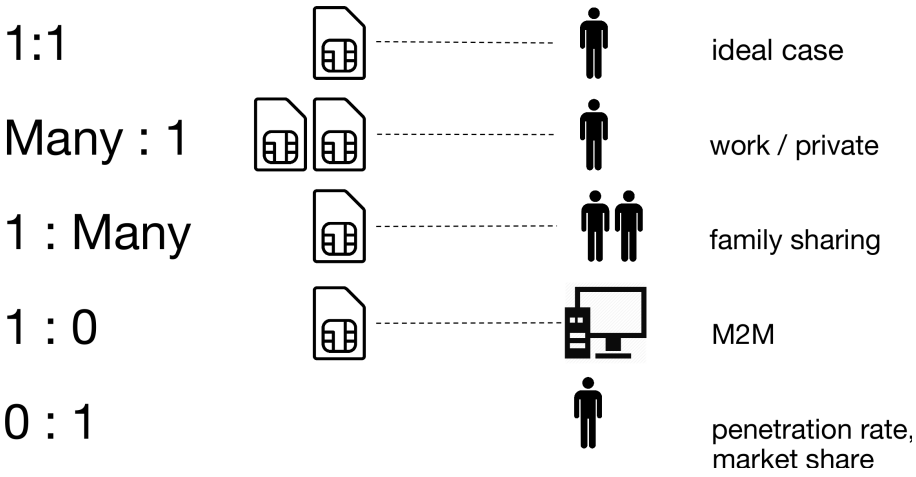


Figure 2.7: Possible association schemes between SIM and persons. Figure adapted from Ricciato *et al.* (2015). M2M stands for Machine to Machine communication.

mobile phones always connect to the closest antenna. However, dozens of factors play a role in the decision of the system to assign a specific cell to a phone (e.g., signal strength, atmospheric conditions, traffic overload, maintenance schedule). It is, therefore, possible that a person at the same location, making five successive phone calls, will connect to five different antennas. Designed for business and not tracking, CDRs provide information that helps companies manage their operations, not track phones. On the other hand, as it has been already mentioned, the spatial resolution of CDRs depends on the BTS density. It means that remote and unpopulated areas, where populations at risk (such as poor and food insecure population) are generally found, have lower spatial resolution than urban areas due to a lower antenna density. Finally, CDR data are always limited to one country and due to technical challenge, the cross-border movements are difficult to capture. This is an important limitation for large scale epidemiological studies because the spread of a disease does not stop at the border. Other similar data sources capturing geographic digital footprints (e.g. tweets) may be used to overcome this limitation (Blanford *et al.*, 2015).

2.7 Data Access

Because private companies hold the data, there is no guarantee of access. It requires an agreement between researchers and the telecom company. The companies might be reluctant to provide access due the threat to subscribers privacy that can result in a loss of customers. Therefore, their interest to open/sell datasets is somewhat limited. Yet, it is generally possible to get access to a particular data set for testing/research purposes, but still far harder (for

legal or commercial reasons) to extend this access for production purposes.

Research teams and institutions have learned the hard way that even in case of emergency situations, being granted access to CDRs can still remain an unsurmountable obstacle. When the Ebola epidemic broke out in Sierra Leone, Liberia, and Guinea in 2014, a group of academics and international development actors began to call for the use of aggregated location data from mobile phone networks in order to facilitate the response effort (The Economist, 2014). After dozens of conference calls over many months involving over fifty participating organizations (including several UN agencies), permission was finally granted by the relevant local authorities – except for Liberia. Despite having the highest death toll of any country that experienced the Ebola epidemic, Liberia never released CDRs, in part due to concern about their ability to enforce privacy protection (McDonald, 2016). This experience raises fundamental questions on the trade-off between privacy and utility and how it can be adjusted according to the level of emergency of a situation.

In the aim to respond to these concerns, the mobile phone industry association – the GSMA – has developed a 'Mobile for Development Intelligence' programme to persuade mobile providers from developing countries to share data with researchers, industries and development organisations. However, their focus is primarily on commercial outputs as the goal of their open data portal – Mobile for Development Products and Services Trackers – is worded as:

> "offer the industry access to high quality data to help improve business decision making, increase total investment from both the commercial mobile industry and the development sector as well as to accelerate economic, environmental and social impact from mobile solutions." (Metcalfe, 2013)

2.8 Data Privacy

A lot of personal information can be extracted from CDRs. Using such data, it is easy to know where people live and work as well as tracking most of their movement (Calabrese *et al.*, 2013; Ahas *et al.*, 2010). Their social network can be characterized allowing, for instance, the examination of the evolution of relationships over time (Eagle *et al.*, 2009b). The way people use their phone is also a good indicator of their personality. For instance, de Montjoye *et al.* (2013b) showed how CDRs could be used to infer five main traits of personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism (a socio-psychological model known as OCEAN). Based on facebook data, the same model was used by Cambridge Analytica to micro-target of campaign material to US voters with the purpose of influencing the 2016 presidential campaign (Youyou *et al.*, 2015; Grassegger and Krogerus, 2017).

To protect people's privacy, mobile phone data are always anonymized, i.e, all personal data such as name, address, phone number, etc., are either removed

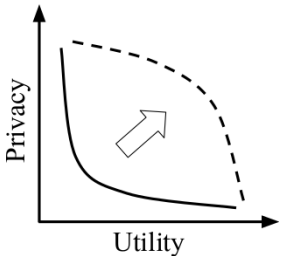


Figure 2.8: Schematic representation of the trade-off between privacy and utility of personal data. Full line is the actual relationship and the dotted line shows the ideal relationship. The figure is adapted from an OPAL presentation (www.opalproject.org).

from the database or replaced by a randomly generated number to avoid identification[10]. Data are then provided to a third party after a non-disclosure agreement (NDA) was signed with the MNO. The purpose of the agreement is to prevent CDRs to be shared to another party, and to define the scope of research questions that will be explored with the data. Both the anonymization procedure and the NDA are supposed to preserve the safety of users privacy.

However, if individual patterns are unique enough, additional information can be used to link the data back to an individual. Using fifteen months of human mobility data derived from CDRs, de Montjoye *et al.* (2013a) showed that four randomly chosen points (i.e., four places where a user was at a specific time) are enough to uniquely characterize 95% of the users, whereas two randomly chosen points still uniquely characterize more than half of the users.

Data aggregation allows to further strengthen privacy. In the case of CDRs, several approach can be used. First, users can be aggregated by BTS. With this aggregation, it is no more possible to track one specific user and mobility analyses cannot be performed anymore. On the other hand, it is still relevant for dynamic population mapping as it only requires the number of a users at a specific place and time. This is also still useful to study spatial network based on antenna-to-antenna traffic. To keep mobility analysis feasible, temporal and/or spatial aggregation can be used. However, decreasing the resolution comes with a loss in data utility so that a trade-off exists between privacy protection and the preservation of data value (Figure 2.8). Furthermore, de Montjoye *et al.* (2013a) showed that blurring the spatial and temporal resolution does not significantly impact the number of points needed to re-identify a user in the database. Finally, to preserve privacy, noise can also be added to some variables of the database (e.g. random spatial reallocation of BTS).

On the other hand, in today's digital world, increasing privacy is only useful if it is done for all sources at the same time. This issue was defined as the 'secondhand smoke problem' by Lane *et al.* (2010):

[10]This process is known as pseudo-anonymization.

> "[...] the secondhand smoke problem of mobile sensing creates new privacy challenges, such as:
>
> • How can the privacy of third parties be effectively protected when other people wearing sensors are nearby?
>
> • How can mismatched privacy policies be managed when two different people are close enough to each other for their sensors to collect information from the other party?"

With the aim to solve some of these issues, the Open Algorithm (OPAL) project was set up by Orange, MIT Media Lab, Data-Pop Alliance, the World Economic Forum and the Imperial College London (Hardjono *et al.*, 2016). The idea behind OPAL is to bring about a paradigm shift in mobile metadata analysis by moving the code to the data rather than data to the code. It means that instead of providing data directly to researchers through NDA, CDRs (or other sensitive data) remain behind the firewall of each provider. Only certified algorithms, meeting predetermined privacy standards, can be run on the data in this secure environment and only aggregated results are shared with the user (www.opalproject.org). It allows facilitating data access while preserving business and individual privacy. In Chapter 4, we illustrate a similar approach where two models of poverty mapping based on disparate data sources can be combined without the need to share the raw data.

2.9 Conclusions

In this Chapter, the specific features of mobile phone metadata were discussed with a focus on applications for development. The amount of information held in these data is fantastic. Among other, they have been used to model the spread of infectious diseases, study road traffic, support electrification planning strategies or map the socio-economic level of population. While massive, CDRs are not statistically representative of the whole population due to several sources of bias. Furthermore, data access and privacy are significant challenges that are not necessarily straightforward to resolve.

While the challenges exist, the potential of such data might exceed the limitations. Compared to traditional data collected to compute official statistics, they are cost-effective and can provide faster or even near real-time insights. They might also be used to test concepts and define future research questions. On the other hand, the Reality Mining project demonstrated that observed behavior using mobile phone metadata strongly differs from what was self-reported by the same individuals (Eagle and Pentland, 2006). This suggests that the subjectivity of the subjects' perception produces a significant bias in traditional surveys. The objectivity coming from their exogeneity is, therefore, another strength of mobile phone metadata.

Chapter 3

Social Capital and Transaction Costs in Millet Markets[1]

The most important asset is ... an extended and well-placed family network from which one can derive jobs, credit, and financial assistance.
A poor man, Senegal 1995 (Narayan-Parker and Patel, 2000)

Highlights

- Monitoring of staple prices are crucial for formulating policies that improve food security in Africa (**relevance**).
- Transaction costs are the most important barrier to efficient market functioning. Social capital can lower these costs by reducing the information and search costs, increasing trust or cutting down the administrative burden.
- A spatial equilibrium model is used that integrates disparate set of data to evaluate the impact of social capital, approximated with calls between market areas, on millet market in Senegal.
- Results demonstrate that accounting for the social capital in the transaction costs explained 1-9% of the price variance depending on the year.
- The year-specific effect remains challenging to assess but could be related to a strengthening of risk aversion following a poor harvest.

[1] Adapted from **Jacques, D. C.**, Marinho, E., d'Andrimont, R., Waldner, F., Radoux, J., Gaspart F. and Defourny, P. 2018. Social capital and transaction costs in millet markets. Heliyon, 4(1).

Abstract

In sub-Saharan Africa, transaction costs are believed to be the most significant barrier that prevents smallholders and farmers from gaining access to markets and productive assets. In this Chapter, we explore the impact of social capital on millet prices for three contrasted years in Senegal. Social capital is approximated using a unique data set on mobile phone communications between 9 million people allowing to simulate the business network between economic agents. Our approach is a spatial equilibrium model that integrates a diversified set of data. Local supply and demand were respectively derived from remotely sensed imagery and population density maps. The road network was used to establish market catchment areas, and transportation costs were derived from distances between markets. Results demonstrate that accounting for the social capital in the transaction costs explained 1-9% of the price variance depending on the year. The year-specific effect remains challenging to assess but could be related to a strengthening of risk aversion following a poor harvest.

3.1 Introduction

In sub-Saharan Africa, the functioning of food markets is jeopardized by several barriers that prevent smallholders and farmers from gaining access to markets and productive assets. The most significant of these barriers is believed to be the transaction costs, the observable and hidden costs associated with arranging and carrying out a transaction. The role that social capital might play in shaping these costs is a research question that has captured the attention of many during the last two decades (Durlauf and Fafchamps, 2005). In this Chapter, we refer to the concept of social capital as introduced by Putnam *et al.* (1994), that is, the "features of social organization, such as trust, norms, and networks that can improve the efficiency of society" (p.167).

Social capital can lower transaction costs by, e.g., reducing the information and search costs, increasing trust or cutting down the administrative burden (Fafchamps and Minten, 2001; Knack and Keefer, 1997; Fukuyama, 1995; Woolcock and Narayan, 2000; Fafchamps, 2006). If agents are not well informed about price differences across markets, time periods or buyers and sellers of different types, or if such information is asymmetric, they cannot engage in optimal arbitrage (Tollens, 2006). On the other hand, trust helps to mitigate the abuse that can occur during the purchase and sale of commodities (non-delivery, late payment, deficient quality, incorrect quantity...) (Bigsten *et al.*, 2000). As they can more easily find and screen each other, well-connected agents will also be more likely to trade together (Barr, 2000). However, the effect is not necessarily positive as overreliance in the activities and decisions of relatives can lead to overpricing due to traders' errors (Levine *et al.*, 2014; Portes, 2014). Nevertheless, limited information and mistrust generally results in inefficient transmission of prices due to local surpluses or scarcities, which ultimately affects both producers and consumers.

According to Durlauf and Fafchamps (2005), the literature on the effects of social capital can be divided into individual and aggregate studies. On the one hand, individual studies explore the effect of social capital on some individual outcomes. For instance, Fafchamps and Minten (2002) found a significant effect of social capital on total sales of food traders in Madagascar, Mawejje and Terje Holden (2014) highlighted that social capital can help Ugandan household to receive higher prices for coffee and Grootaert (1999) demonstrated the effect of social capital on household expenditure in Indonesia. On the other hand, aggregate studies mainly focused on the relationship between social capital and per capita output growth at a high level of aggregation, e.g., a country or a region (Beugelsdijk and Schaik, 2003; Guiso *et al.*, 2004). The standard approach of all these studies is generally to run linear regressions on cross-sectional data with some outcome of interest against empirical proxies for social capital and a set of controls. The significance of the coefficients of the social capital variables allows to conclude on their effect on the outcome. One challenge of empirical work on social capital is therefore to identify observable variables that can be used as proxies for social capital (Portes, 2000). An array of variables have

been proposed in empirical papers and include the number of known traders, the number of relatives involved in agricultural trade, the number of languages the trader speaks, or some measures of ethnic homogeneity for organizations formed by households (Fafchamps and Minten, 2001, 2002; Grootaert, 1999; Isham, 2002; Gabre-Madhin, 2001).

In this study, we explored the effect of transaction costs generated by social capital on millet retail prices in Senegalese food markets for three contrasted years. Millet serves as the main local subsistence food crops in many Sahelian countries, including Senegal. Millet prices are therefore an important indicator of food security as they directly impact farmers' income and their ability to access staple foods (Pokhriyal and Jacques, 2017). Furthermore, millet is an interesting case study as most trade is local (little cross-border trade). We modeled social capital using a unique data set of mobile phone communications between 9 million people. Our assumption is that the intermarket calls reflects the business network of economic agents from different markets. Other things being equal, well-connected agents are more likely to trade with one another because the transaction costs are reduced between them. To evaluate the effect of social capital on millet prices and market functioning, we focused on intermarket trades as traders are the economic agents most exposed to the effect of transaction costs (Fafchamps and Minten, 2001). To that end, we adopted an original approach in the form of a spatial equilibrium model which enabled us to compare different market functioning scenarios, i.e., with and without transaction costs accounting for social capital. Local supply and demand were respectively derived from remotely sensed imagery and population density maps. The road network was used to establish market catchment areas, and transportation costs were derived from distances between markets. The emphasis was on the parsimony of the model by making use of available data sets without seeking to determine the eventual causal links of the mechanisms involved.

We showed that taking into account the impact of social capital on transaction costs explained between 1 and 9 percent of the price variance depending on the year. The year-specific effect remains challenging to assess but could be related to a strengthening of risk aversion following a poor harvest. In any case, the high difference between the years suggests that the effect of social capital in agricultural markets is very dynamic and context specific.

The remainder of the Chapter is organized as follows: Section 3.2 describes the empirical framework. Section 3.3 presents the data, Section 3.4 provides the findings and discussions as well as the limitations of the approach and some policy implications. Lastly, Section 3.5 provides the main conclusions of the analysis.

3.2 Model

To assess the effect of social capital on transaction costs, we used a simple spatial equilibrium model. Specifically, a point-location model consisting of a network with markets located at network nodes, and network links that serve only for commodity transportation flows (Enke, 1951; Takayama and Judge, 1971; McNew and Fackler, 1997), which differs from the agents-on-links models (Hotelling, 1990). In a given market, i, the price is a function of local supply, S, demand, D and time, t:

$$p_i = f(S_i, D_i, t) \tag{3.1}$$

Each pair of market nodes is linked by trade. Without transaction costs, price differences between markets are only depending on the transportation cost, r (assumed to be identical throughout the country), and the distance between market i and j, d_{ij}:

$$p_i - p_j \leq d_{ij} r \tag{3.2}$$

with equality if trading actually occurs.

We introduced a transaction cost, associated with social capital, as the multiplicative parameter $\frac{1}{s}$. For each pair of markets ij and depending on the level of social capital, the transaction cost was either null ($s_{ij} = 1$) or infinite ($s_{ij} = 0$). It follows that the arbitrage condition becomes:

$$p_i - p_j \leq \frac{d_{ij} r}{s_{ij}} \quad \text{with} \quad s_{ij} \in \{0, 1\} \tag{3.3}$$

Through an optimization procedure, the two unknown parameters, r and s, were estimated. The objective was to assess if Eq. 3.3 allowed to explain more of the price variance than Eq. 3.2, i.e., if taking into account the impact of social capital on transaction costs allowed to better explain price variance.

In a nutshell, three scenarios were compared: (i) $r = \infty$ (Scenario I - segregated markets), (ii) $0 < r < \infty$ and $\bar{s} = 1$ (Scenario II - trade without transaction costs) and (iii) $0 < r < \infty$ and $0 < \bar{s} \leq 1$ (Scenario III - trade with transaction costs) with $\bar{s}$, the mean of all s_{ij}.

Our approach (Figure 3.1) involved estimating the retail price for millet in each market from the estimation of local demand and supply (Eq. 3.1). Population was used as a proxy of demand. Local food production was approximated by a vegetation index derived from satellite image time series combined with national statistics and used as the supply input. Only local production was considered as millet is little affected by international trade. Both types of data were spatially aggregated in the catchment area of each market, which equates to the area that minimizes road distance between each market (Figure 3.2).

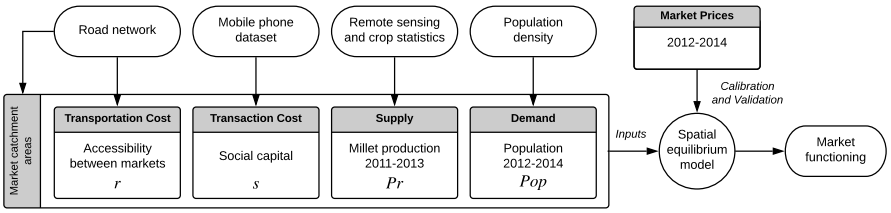


Figure 3.1: Overview of the data inputs and their associated variables.

3.2.1 Scenario I (segregated markets)

For segregated markets, the price was estimated using the following multiple regression model (ordinary least squares) fitted separately for each year (Eq. 3.4):

$$p_{it} = \alpha_0 + \alpha_1 log\left(\frac{Pop_i}{Pr_i^{harv}}\right) + \alpha_2 t + \epsilon_{it} \tag{3.4}$$

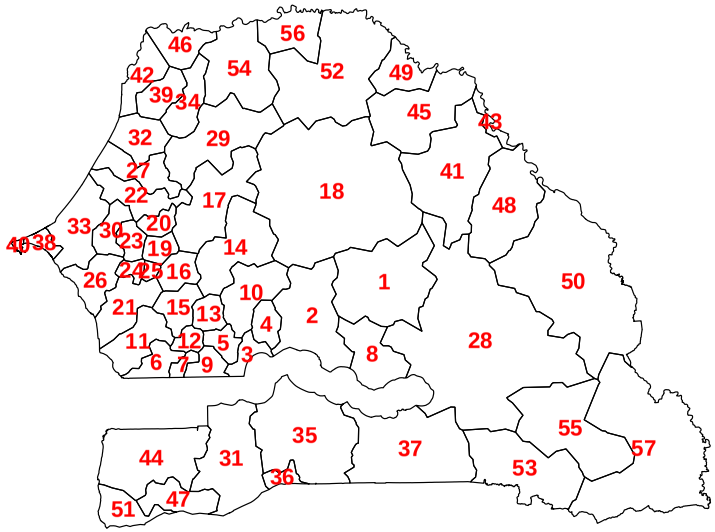
where Pop_i and Pr_i^{harv} are the population and the millet production (in tons) harvested in the catchment area of the market i, t, the month and ϵ_{it}, the regression residual. We restricted the analysis from January to August. This corresponds to the period following harvest up to the lean season, and is regarded as the most critical period affecting price evolution. This scenario is conceptual because no markets are actually isolated. Therefore, the production is transferred from producing to consuming areas to balance supply and demand. The actual price in the markets depends on the production (supply) after the transfers have taken place (Scenario II and III).

3.2.2 Scenario II and Scenario III (Spatial equilibrium model)

One could use Eq. 3.4 to estimate the market prices within the spatial equilibrium model. However, there is no reason to believe that a model fitted on unrealistic distribution of production, i.e., production that would not be traded (Pr^{harv}), would be appropriate to simulate trading. To overcome this limitation, we introduced pseudo prices, proportional to actual prices, defined as Eq. 3.5:

$$p_{it_0} \propto \tilde{p}_{it_0} = log\left(\frac{Pop_i}{Pr_i}\right) \tag{3.5}$$

The initial conditions of the spatial equilibrium model were the market prices estimated by Eq. 3.5 with $Pr_i = Pr_i^{harv}$. All production transfers were assumed to occur in t_0 so that the month variable t was not used in the spatial equilibrium model. The arbitrage condition defined in Eq. 3.3 was then applied to each pair of markets at once, with the transportation cost r set as the transportation pseudo cost $\tilde{r} \propto r$. If an opportunity of arbitrage was possible between market i and j, i.e., if the condition defined in Eq. 3.3 was not satisfied, a ton of millet was transferred between the two markets. The new level of



1 KOUTHIABA	20 NDINDY	39 MPAL
2 KOUNGUEUL	21 FATICK	40 DAKAR
3 DIAMA GADIO	22 THILMAKHA	41 OUROSSOGUI
4 DIOLI MANDAKH	23 BAMBEY	42 SAINT−LOUIS
5 MABO	24 MBAFAYE	43 MATAM
6 NDRAME ESCALE	25 DIAKHAO	44 BIGNONA
7 MISSIRAH	26 SANDJARA	45 THIODAYE
8 MERETO	27 SAGATTA	46 ROSSO
9 POROKHANE	28 TAMBACOUNDA	47 ZIGUINCHOR
10 KAFFRINE	29 NDIAGNE	48 ORKODIERE
11 PASSY	30 TOUBA TOUL	49 GALOYA
12 N'DOFFANE	31 SEDHIOU	50 BAKEL
13 BIRKELANE	32 LOUGA	51 OUSSOUYE
14 MBAR	33 THIES	52 DODEL
15 KAOLACK	34 GOUILLE MBEUTH	53 SALEMATA
16 GOSSAS	35 KOLDA	54 DAGANA
17 TOUBA	36 SARE YOBA	55 MAKO
18 VELINGARA	37 DIAOBE	56 THILLE BOUBACAR
19 DIOURBEL	38 THIAROYE	57 KEDOUGOU

Figure 3.2: Catchment areas of each market.

production in each market then allowed estimation of new prices (using Eq. 3.5) on which the condition of transfer was again applied. Through this approach, the model iterates until an equilibrium was reached, i.e., when all profiTable transfers of production between markets had occurred and Eq. 3.3 was satisfied for all pair of markets. The only effect of the model was the reallocation of production between the markets. The total production remained unchanged.

Actual market prices were then estimated by fitting the following model at the equilibrium:

$$p_{it} = \alpha_0 + \alpha_1 log\left(\frac{Pop_i}{Pr_i^{eq}}\right) + \alpha_2 t + \epsilon_{it} \tag{3.6}$$

where Pr_i^{eq} is the production at the equilibrium of market i obtained for a specific pair of parameters $\tilde{r}$ and $\bar{s}$ (mean of all s_{ij}), and it is the only variable that changes compared to Eq. 3.4. This approach was repeated with several values of $\tilde{r} \in \left[min\left(\frac{\Delta\tilde{p}_{ij}}{d_{ij}}\right), max\left(\frac{\Delta\tilde{p}_{ij}}{d_{ij}}\right)\right]$ (with $\Delta\tilde{p}_{ij} = \tilde{p}_i - \tilde{p}_j$) and $\bar{s} \in]0, 1]$ for Scenario III or $\bar{s} = 1$ for Scenario II.

Based on the maximum R^2, the coefficients and parameters $\alpha_{0,opt}$, $\alpha_{1,opt}$, $\alpha_{2,opt}$, $\tilde{r}_{opt}$, $\bar{s}_{opt}$ of the optimal model (Eq. 3.6) were estimated. These values allowed us to define the relationship between pseudo prices $\tilde{p}$ and actual prices p and consequently, $\tilde{r}_{opt}$ and r_{opt}:

$$p_{it_0} = \alpha_{0,opt} + \alpha_{1,opt} log(\tilde{p}_{it_0}) \tag{3.7}$$

$$r_{opt} = \alpha_{1,opt}\tilde{r}_{opt} \tag{3.8}$$

Finally, the accuracy of r_{opt} and $\bar{s}_{opt}$ estimates was assessed by investigating the performance of the optimal model for values of r and $\bar{s}$ close to r_{opt} and $\bar{s}_{opt}$.

It is worth mentioning that in Scenario III, the number of market pairs that can trade with each other decreases (i.e. with null transaction costs or $s = 1$) as $\bar{s}$ is closer to 0. Therefore, the difference of performance between the model of Scenario II and III is unlikely to be explained by a random effect and can reliably be associated with an actual impact of the transaction costs, shaped by the social capital, on market functioning. For confirmation of this, the model of Scenario III was tested using 10 different random allocations of s (drawn from the same normal distribution than the original data) to each market pair. None increase in model performance is expected from these random simulations compared to Scenario II.

3.3 Data

3.3.1 Market Prices

Domestic price data were sourced from the Vulnerability Analysis and Mapping (VAM) Food and Commodity Prices Data Store of the UN World Food Program (Vulnerability Analysis and Mapping unit, 2000-2014). In Senegal, VAM collects its data from the Commissariat pour la Sécurité Alimentaire. The data set consists of monthly retail prices (when available) from 60 markets distributed across the 14 regions of Senegal between 2012 and 2014. Of the four markets located in Dakar (Dakar, Tilène, Gueule Tapee and Castors), only the market with the least missing data (Tilène) was retained and used in the model.

Figure 3.3 shows the temporal evolution of average millet prices in Senegal. A typical price decrease after harvest (around September) is observable, followed by a gradual increase until the lean season. This trend is highly correlated with the supply trend. For instance, the impact of a poor and late harvest such as in 2011 (480 kt) was accompanied by a delay in peak price in October, compared with a relatively good year such as 2010 (810 kt) that is characterized by an early price decrease in July.

A small price decrease at the beginning of some years (February 2008, February 2009, February 2010, March 2013) can be observed in Figure 3.3. It could be explained by rice substitution (harvested around October-December) resulting in a decrease of the demand for millet. Price data from May 2013 were discarded because they clearly exhibited errors in encoding or sampling (correlation with other months of 2013 ranging from 0.05 to 0.35). Additionally, data for May 2012 were retained even though the sudden drop at this date. This drop is most likely explained by food assistance provided by the WFP at this period (WFP, 2012b). As food aid is not taken into account in the model, poor price estimations were expected for this month. The spatial distribution of the price clearly shows a correlation with the production-population ratio (Figure 3.4).

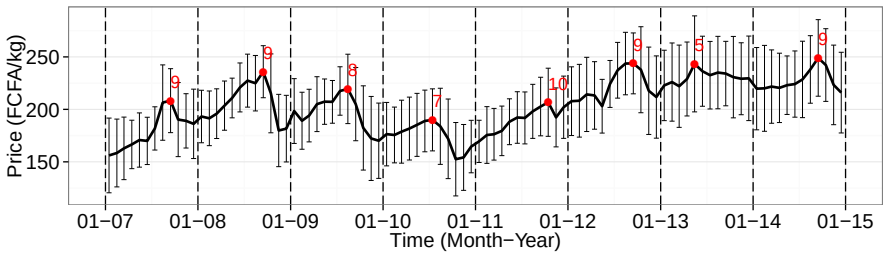


Figure 3.3: Average monthly millet prices (FCFA/kg) and standard deviation for each month from 2007 to 2014 in Senegal. Red annotations indicate the maximum value of each year and the corresponding month.

Although not all markets had price data for the period of interest, We decided to keep them all (apart from those in Dakar, see above) to better simulate the spatial dynamics of market trade. Consequently, several price predictions were not validated due to missing data.

3.3.2 Catchment Areas and Transportation Cost

Most of the food transportation (>95%) in Senegal relies on the road network (Bertholet *et al.*, 2004). The distance by road was therefore used to approximate the transportation cost and define the catchment areas. A topological network was derived from the Global Insight data (manually edited by visual assessment in areas where important roads were missing) and minimum traveling times were computed using Dijkstra's algorithm (Dijkstra, 1959).

The transportation cost was defined as being directly proportional to the distance between markets (Eq. 3.2). Transportation through the Gambia was assumed to be null, i.e., the costs of crossing the border based on custom duties and other costs was assumed to be equivalent or more expensive than going around Gambia via the road. This assumption is consistent with the experience of local people.

The catchment areas for each market were estimated based on the area that minimizes the road distance between each market (Figure 3.2). In the absence of secondary traveling directions, the underlying assumption was that traders travel to the nearest main road and then to the nearest market using the road network in order to sell their products. Each point in space was therefore assigned to a single market based on the shortest traveling distance by the nearest road.

3.3.3 Demand and Population

The demand was estimated using population distribution maps acquired from the Worldpop project (Linard *et al.*, 2012). Worldpop maps provide an estimate of the number of inhabitants in a given grid square (0.00083 decimal degrees, ∼100 m at the equator). The estimation was based on a random forest model trained on official 2009 population estimates at the commune level (method described in Stevens *et al.* (2015)). The number of communes (113) was twice the number of catchment areas, ensuring accurate aggregation at the catchment level. The map for 2010 (not available for the years of interest) was used as a proxy of the spatial distribution of the population. The World Bank estimation of the total population for each year (2012-2014) was then used to adjust this distribution. The unequal birth rate throughout the country may affect the accuracy of this extrapolation.

3.3.4 Supply and Production

In general, the production of cereal food crops is unable to meet the needs of the Senegalese population. Only in years with abundant rainfall does the

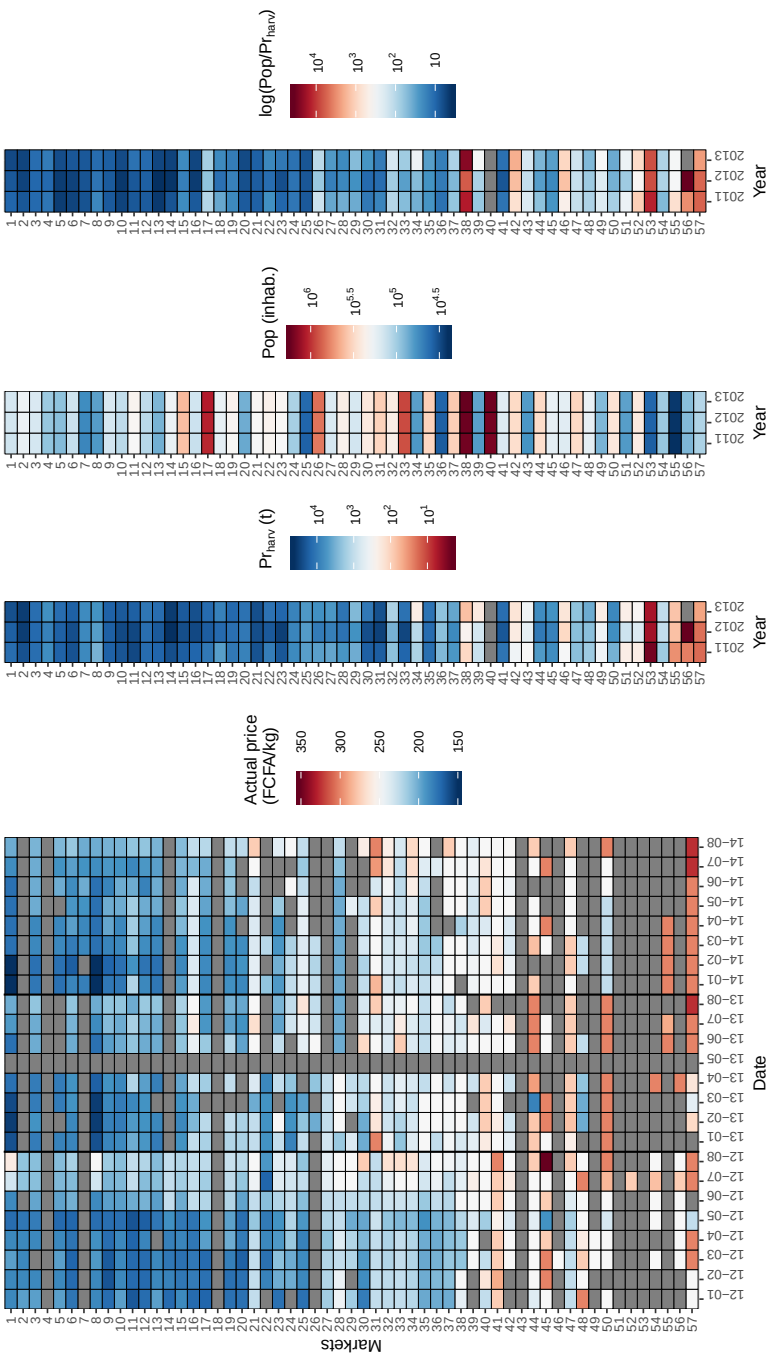


Figure 3.4: Millet prices, millet production, population and the ratio of population and production by market for 2012 to 2014 (January to August). Market numbers refer to the number-name matching in Figure 3.2. No data are shown in grey.

country approach self-sufficiency in staple food crops in rural areas. Conversely, in times of poor harvest, millet is scarce due to the limited trade of this crop in the region. This deficiency is overcome by an increase of the rice imports, leading to a shift from millet to rice consumption in households that can afford it (Ndiaye, 2007). Most of the millet is produced in the regions of Kaolack, Kaffrine and Fatick (see Chapter 1, Figure 1.2) in rotation with groundnuts, the main cash crop (Commissariat de la Sécurité Alimentaire, 2000-2014). This rotation is crucial because groundnuts, being a legume crop, fix nitrogen in the soil.

Millet production estimates from the Direction de l'Analyse, de la Prévision et des Statistiques Agricoles (DAPSA) were selected as a proxy of the supply (Direction de l'Analyse, de la Prévision et des Statistiques Agricoles, 2013). These estimates are based on a two-stage stratified sample of around 6000 households and can be considered sufficiently accurate (see Chapter 1). However, because the granularity of these data is at the department level, 10-day temporal synthesis of 1-km SPOT-VEGETATION satellite images were used to convert them to the market catchment area level. The millet area was assumed to be uniformly distributed within the cultivated areas of a department, and the millet yield was spatially allocated to each 1-km pixel according to the distribution of Normalized Difference Vegetation Index (NDVI) data accumulated during the growing season. The NDVI, defined as the difference between near-infrared and red reflectances normalized by their sum, serves as a useful yield proxy when yield is mainly driven by plant vegetative growth, as occurs in regions where water or soil fertility are the main limiting factors, such as the Sahel (Samaké *et al.*, 2005; Rockström and De Rouw, 1997).

In practical terms, cultivated areas were masked using the Land Cover Map produced by the Global Land Cover Network (2005; 1:100.000 scale; Leonardi (2008)) based on GlobCover 2005 map (Defourny *et al.*, 2009), the most accurate map of the country (Waldner *et al.*, 2015b). Since we lacked reliable information on the spatial distribution of millet, we assumed that it was evenly grown across the cultivated area of a specific department. Then, for each pixel within cultivated areas, NDVI values above 0.2 during the millet growing season (from July to November) were integrated, which limited the contribution of bare soil to the signal. The actual millet production observed at the department level was then spatially allocated at the pixel level using the NDVI as an indicator of productivity. Thus, a proportionally higher production was allocated to pixels with high NDVI within a department. Finally, using the market catchment area boundaries, pixel production values were aggregated to generate millet production, Pr_{harv}, by market.

3.3.5 Mobile Phone Calls, Social Capital and Transaction Costs

The last decade has seen a drastic increase of mobile phone users in Africa (Castells *et al.*, 2009). In Senegal, the number of mobile cellular subscriptions was 25% in 2006 but reached 100% in 2014 (International Telecommunica-

tion Union, 2015). Several studies have demonstrated the impact of mobile phone access on price dispersion in food markets by, among others, reducing information costs (Aker, 2010; Jensen, 2007). Each time a call is made, a Call Data Record (CDR) is generated by the telecom companies for billing purposes. These metadata provide information on when, how and with whom one communicates. The communication itself is not recorded. After anonymization, some of these metadata were made available to the scientific community. As a result, the past few years have seen a rise in research projects such as the Data for Development challenge (www.d4d.orange.com) that was set up by Orange in 2013 (Ivory Coast) and 2015 (Senegal) to foster the use of CDRs for societal developments (de Montjoye *et al.*, 2014; Blondel *et al.*, 2012). In particular, results have shown that the call intensity between people is a good indicator of social networks (Blondel *et al.*, 2015; Candia *et al.*, 2008b; Eagle *et al.*, 2008, 2009b). This property was used to approximate social capital assuming that business and social network were correlated.

The Call Data Records were provided by Sonatel Orange through the D4D challenge (Senegal) framework. The original data set of phone calls between more than 9 million Orange customers in Senegal between January 1^{st}, 2013 and December 31^{st}, 2013 was processed to remove presumed machine-based calls (see de Montjoye *et al.* (2014) for a description of the method).

From the mobile phone data, a contingency Table with the yearly average sum of calls from, and to, each market (i.e., from antenna within a buffer range of 10 km) was generated. This resulted in an origin-destination matrix containing the average number of calls between all market pairs. The transaction cost s, associated with social capital, was defined from this matrix following a two-step procedure.

First, the average number of calls made from and to each market pair was normalized by the calls made within the destination market, which was assumed to be proportional to the population living or working in the market area (Eq. 3.9). This normalization was performed to ensure that small but close markets were considered well connected and, therefore, the social capital.

$$Ncalls_{ij} = log\left(\frac{Calls_{ij} + Calls_{ji}}{Calls_{jj}}\right) \neq Ncalls_{ji} \tag{3.9}$$

where $Calls_{ij}$ is the yearly average sum of calls from market i to market j. $Ncalls$ provides an approximation of the strength of social relationships between traders of two markets. $Ncalls$ is asymmetric (i.e. $Ncalls_{ij} \neq Ncalls_{ji}$) as the denominator change according to the trade direction. Practically, it means that big cities have always a large $Ncalls$ value with small cities and vice versa. The assumption is that trade is easier from large to small cities than the other way around.

Second, in order to link $Ncalls$, the proxy of social capital, with s, the transaction costs, we tested several limit values, $Ncalls_{lim}$, above which it was

assumed that traders from two markets traded without transaction costs ($s = 1$), and below which the traders could not trade ($s = 0$), as follows (Eq. 3.10):

$$s_{ij} = \begin{cases} 1 & \text{if } Ncalls_{ij} \geqslant Ncalls_{lim}, \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

A value of $Ncalls_{lim}$ corresponds to a specific $\bar{s}$ value for all markets.

Data were only available for 2013, we therefore assumed that the call pattern was similar for 2012 and 2014, since these are the two closest years.

Table 3.1: Multiple regression analysis results.

Panel I - Eq. 3.4 for Scenario I with the month (w/ t) and without the month (w/o t) variable

	2012 (n=354)		2013 (n=270)		2014 (n=299)	
	w/o t	w/ t	w/o t	w/ t	w/o t	w/ t
α_0[a]	195.7±3.0***	173.0±4.0***	205.2±3.4***	196.1±4.7***	196.9±3.1***	186.5±4.4***
α_1[a] $(log(Pop/Pr^{harv}))$	6.3±0.7***	6.2±0.6***	6.4±0.8***	6.3±0.8***	7.0±0.7***	7.0±0.7***
α_2[a] (t)	-	5.0±0.6***	-	2.1±0.8**	-	2.4±0.7**
R^2	0.19	0.34	0.19	0.22	0.27	0.29

Panel II - Eq. 3.6 for Scenario II and III

	2012 (n=354)		2013 (n=270)		2014 (n=299)	
	Scenario II	Scenario III	Scenario II	Scenario III	Scenario II	Scenario III
$\alpha_{0,opt}$[a]	77.9±7.0***	-6.9±10.0	-32.2±14.0*	-71.9±13.4***	72.1±6.5**	63.6±6.7***
$\alpha_{1,opt}$[a] $(log(Pop/Pr^{eq}))$	36.5±2.0***	62.1±3.0***	84.4±4.6***	97.0±4.4***	43.2±1.9***	46.0±1.9***
$\alpha_{2,opt}$[a] (t)	4.8±0.5***	4.7±0.5***	2.0±0.6***	2.3±0.5***	2.8±0.5***	2.8±0.5***
R^2	0.55	0.61	0.57	0.66	0.66	0.67
RMSE (FCFA/kg)	22.2	20.7	22.7	20.1	18.7	18.2

Panel III - r_{opt} and $\bar{s}_{opt}$ estimates

	2012 (n=354)		2013 (n=270)		2014 (n=299)	
	Scenario II	Scenario III	Scenario II	Scenario III	Scenario II	Scenario III
r_{opt}[b](FCFA/kg.100 km)	27.6±4.1	23.2±2.9	39.5±3.3	31.8±3.1	35.7±3.8	34.6± 4.2
$\bar{s}_{opt}$[b]	-	0.43±0.09	-	0.43±0.07	-	0.74±0.08

[a]± standard error; [b] ± the error corresponding to the range of parameter values for which R^2 is not lower than 1% from the maximum R^2; $^*p < 0.1$, $^{**}p < 0.01$, $^{***}p < 0.001$

3.4 Results and Discussion

3.4.1 Scenario I ($r = \infty$)

Panel I in Table 3.1 presents the results obtained by applying Eq. 3.4, i.e., where all markets are independent. As expected, before transfers from surplus to deficit areas begin, the coefficient of determination between actual and estimated millet prices was low ($R^2 = 0.19 - 0.27$). This allowed us to reject the perfect market segregation scenario in Senegal.

Adding the Month (t) variable improved the results for each year ($R^2 = 0.22 - 0.34$), particularly in 2012. This is explained by the higher temporal variation in prices for this year. The average price difference between January and August 2012 was 34.6 FCFA/kg, compared with 17.5 FCFA/kg in 2013 and 21.0 FCFA/kg in 2014. All coefficients were significant ($p < 0.01$) and their expected sign was observed.

Highly productive areas with low population tended to have lower prices than less productive areas with high population (positive coefficient) while the price tended to be higher in August than in January (positive coefficient). The coefficient of the Pop/Pr ratio was not significantly different from year to year.

3.4.2 Scenario II ($0 < r < \infty$ and $\bar{s} = 1$)

The situation with null transaction costs was studied thanks to the arbitrage condition defined in Eq. 3.2. The dashed lines on Figure 3.5 (left) show the R^2 of the model for several values of $\tilde{r}$. Compared to the segregated markets situation (dotted lines on Figure 3.5, left), r alone was able to explain between 21 and 38% of the price variance between markets. The model converged to an optimal value, based on the maximum R^2, relatively similar from year to year (28-40 FCFA/kg.100km, see panel III in Table 3.1).

These values of transportation costs define the trade flows based on differentials of p_{t_0} which are not prices actually observed and should therefore not be directly interpreted. By comparison, transportation prices negotiated by smaller operators are ∼3-4.2 FCFA/kg.100km (Hamilton, 2010). The value, r, takes into account not only the actual transportation cost, but also all costs that increase proportionally with the inter-market distance. For instance, it is more risky to deal with remote traders due to, for instance, a lack of trust, or a more expensive time investment. It is therefore very likely that social capital also explains part of the r value.

The smoothness of the spatial equilibrium model convergence using $\alpha_{0,opt}$,$\alpha_{1,opt}$ and $\alpha_{2,opt}$ is shown on Figure 3.6 (left).

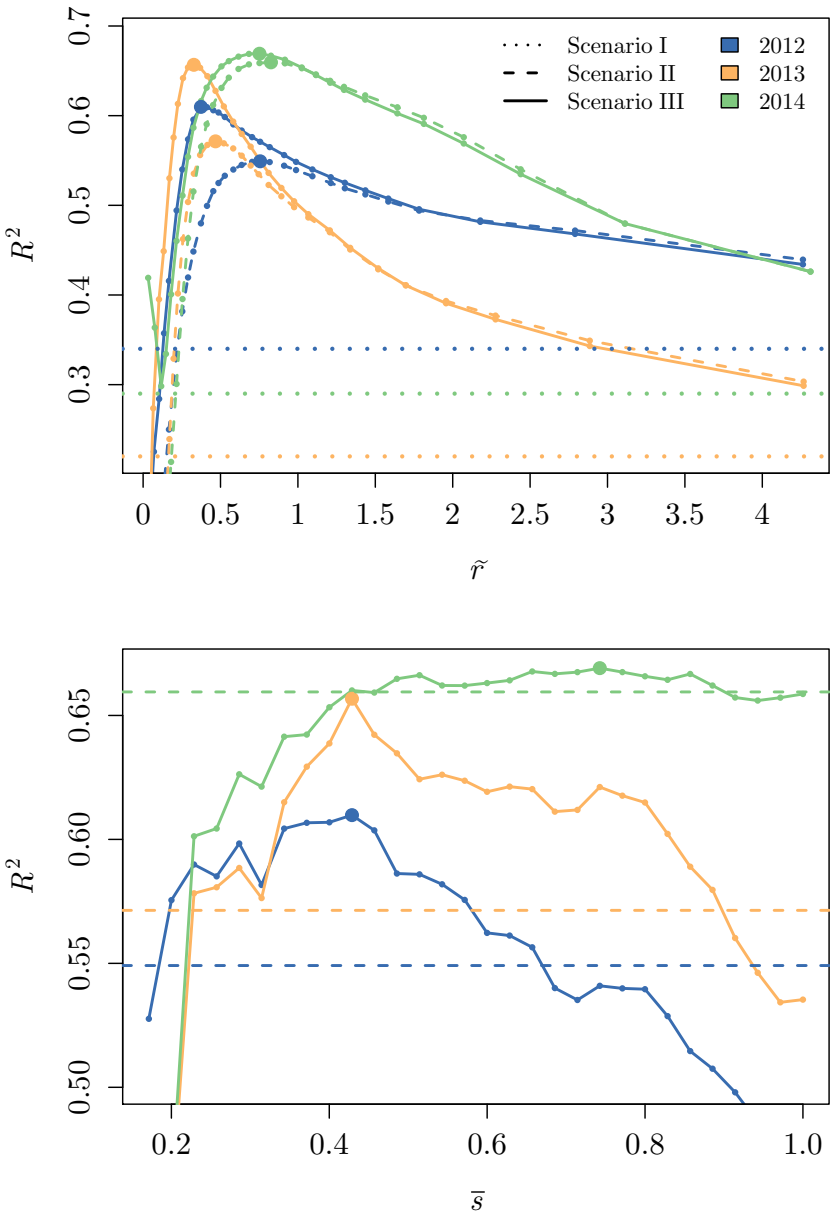


Figure 3.5: Coefficient of determination of the model prediction for several transportation pseudo cost values for Scenario I ($r = \infty$), Scenario II ($0 < \tilde{r} < \infty$ and $\bar{s} = 1$) and Scenario III ($0 < \tilde{r} < \infty$ and $0 < \bar{s} \leq 1$).

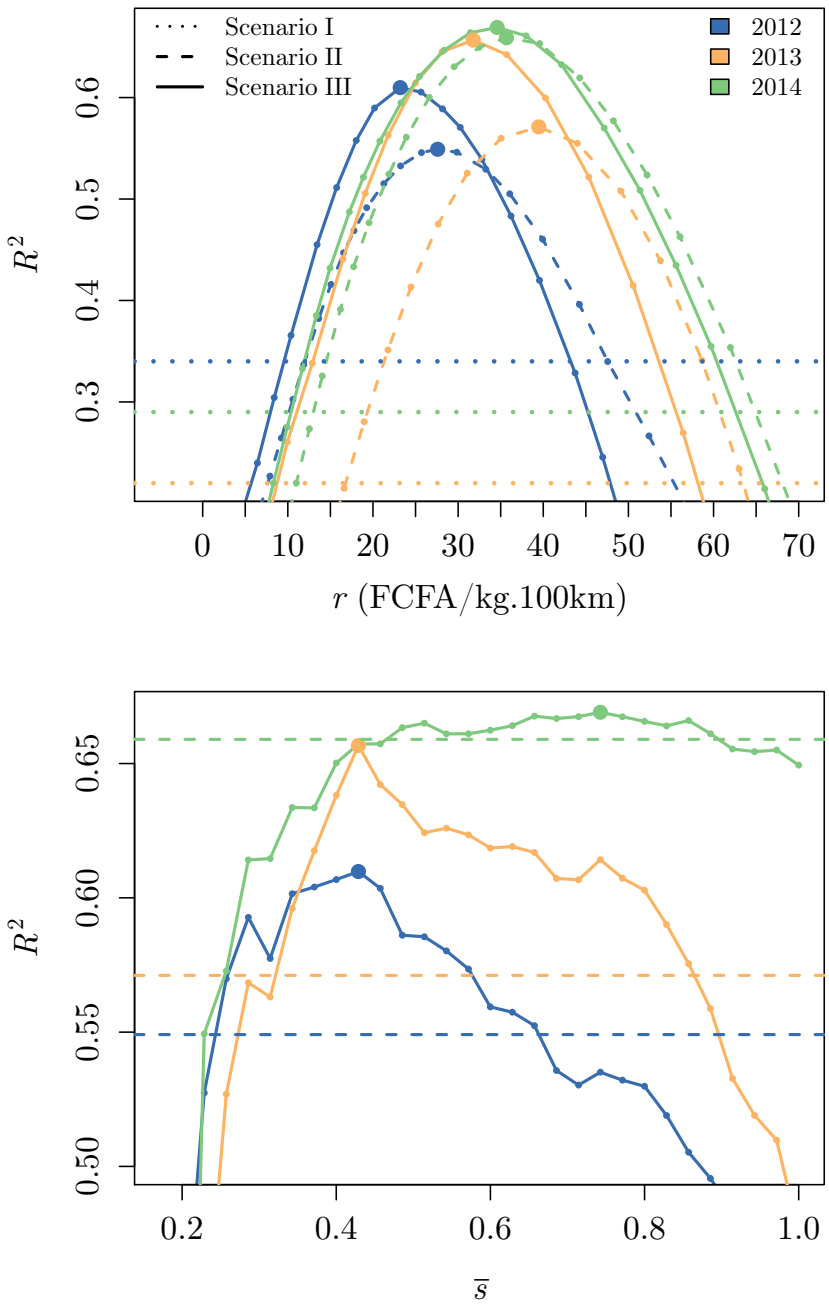


Figure 3.6: Coefficient of determination of the optimal model (using $\alpha_{0,opt}$,$\alpha_{1,opt}$ and $\alpha_{2,opt}$) prediction for several transportation cost values for Scenario I ($r = \infty$), Scenario II ($0 < \tilde{r} < \infty$ and $\bar{s} = 1$) and Scenario III ($0 < \tilde{r} < \infty$ and $0 < \bar{s} \leq 1$).

3.4.3 Scenario III ($0 < r < \infty$ and $0 < \bar{s} \leq 1$)

When the transaction costs s were introduced, the performance of the model improved for each year (R^2=0.61-0.67, see the solid lines on Figure 3.5). The impact was higher for 2013, with 9% of the variance potentially explained by the transaction costs compared with 6% for 2012 and 1% for 2014 (see panel II in Table 3.1). The 10 random allocations of s gave, as expected, no performance increase between Scenario III and II. It is also worth noting that, even though it was not its primary objective, the model was efficient to estimate prices, especially in light of the few parameters involved and the disparate nature of the data sources.

The performance of the model evolved as a function of $\bar{s}$ (Figure 3.5 and 3.6 on the right). The years 2012 and 2013 appeared to follow a similar trend ($\bar{s}_{opt} \sim 0.4$), although with a different amplitude, whereas 2014 exhibited a better performance for lower transaction costs ($\bar{s}_{opt} \sim 0.7$). Yet, the convergence is not perfectly smooth. At each iteration, 40 pairs of market were removed from trading ($s = 1$ to $s = 0$). It could therefore lead to abrupt changes within the optimization process. Furthermore, depending on the configuration of transaction costs, the production flow takes a preferential path that does not necessarily lead to linear change in Pr^{eq} and therefore R^2. Having mentioned that, the trend remained clear but the accuracy of the optimum values must be interpreted with caution.

From these results, the impact of social capital on millet prices (regardless the distances between markets), approximated by the business network between agents of different markets, was clearly demonstrated. Transaction costs, as defined in this study, reflect the business network of people working in a given market. Trading is easier between trustworthy people belonging to the same social network. In the model, a high s prevents a trader from directly taking advantage of a profitable arbitrage opportunity. It forces the trade flow to follow a less risky path. In other words, s determines the preferential flow path of trade. Transaction costs might also reflect heterogeneous r that could be related to differences in road quality for instance. However, the fact that s could be asymmetric (i.e. $s_{ji} \neq s_{ji}$) is a sufficient evidence against this hypothesis as road quality is, in most cases, similar in both direction.

The specificity of the impact of social capital, in particular the difference between the years, is challenging to assess because Pr^{eq} – the only variable that changes in the model – depends on the interaction between Pr^{harv}, r, s and d. The combination of these four parameters determines the production flow and ultimately, Pr^{eq}. Furthermore, the validation data set was incomplete (Figure 3.4) which could result in local over-fitting and flaws in the interpretation of the results.

In light of theses observations, one should not jump to any conclusion on the year-specific impact of the transaction costs. However, some observations can be drawn from the difference of production between the years. The agricultural

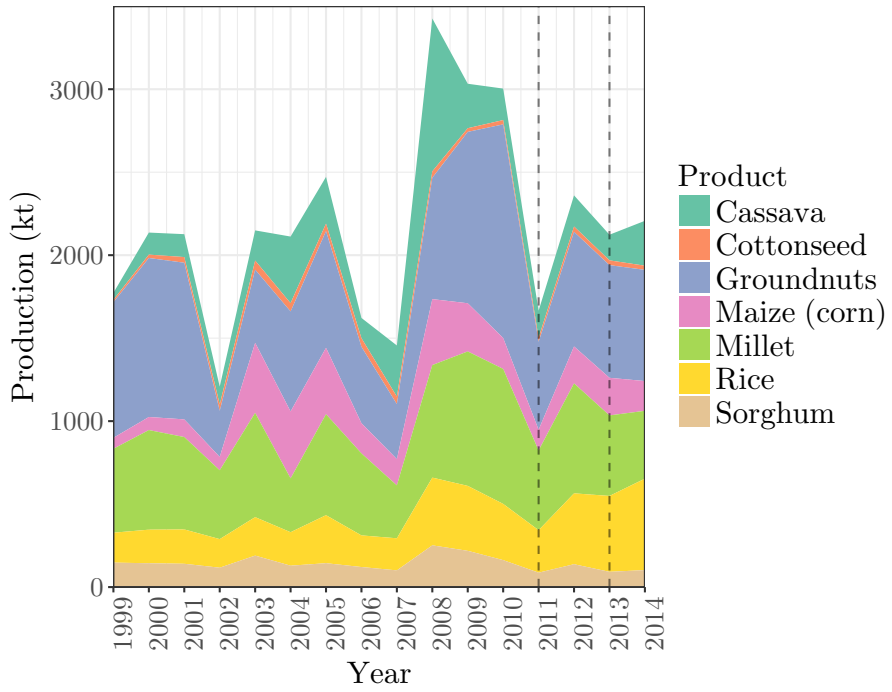


Figure 3.7: Cumulative production of the main crops in Senegal from 1999 to 2014 (data from DAPSA, same as in Chapter 1). Dotted lines show the period covered in this Chapter. The agricultural production in 2011 was in total lower than for 2012 and 2013 which were average years.

production in the 2011/2012 harvest was lower than usual (Figure 3.7), due to late onset of the rainy season, dry spells, early cessation of rains, and late provision of inputs (WFP, 2012a). In comparison 2012/2013 and 2013/2014 were average years, the year with the highest effect of social capital on prices. While it is unlikely that social network radically changes from year to year, small changes can have an great impact if there are focused on important market pairs.

3.4.4 Residuals

Based on the specific values of the residuals, May 2012 was systematically overestimated (+9.8 FCFA/kg in average) due to unusual low actual price values related to food assistance (Figure 3.4) (WFP, 2012b). Discarding this month yielded significant improvement in the model fit (R^2=0.64, n=308 compared to R^2=0.61, n=354) for the same r_{opt} and $\bar{s}_{opt}$ values.

The largest residuals (absolute value > 20%) generally corresponded to unexpected values in the price database, such as the price in Thiodaye for May and August 2012, or in Bignona for February 2012 (Figure 3.4). These outliers

could be explained by either sampling or encoding errors, or market failures. In the latter case, the difference between the actual and predicted price could then be used as an indicator of market failures associated with unexpected events such as religious feasts, storage effects, food aid or product substitution.

Some of the markets associated with the outliers were also poorly estimated, suggesting that their under or overestimation could be explained by systematic inaccuracies for these markets. This could be due to inaccuracies in the delineation of the catchment area, resulting in poor estimation of local production and population.

3.4.5 Trade Flows

The average trade flows can be used to classify the markets into sinks and sources of production (Figure 3.8). Gossas (16 – the numbers between brackets refer to market number-name matching from Figure 3.2), Mbar (14) and Kaffrine (10) were the main sources of millet production, and Tilène (40), Thiaroye (38) and Touba (17) were the main sinks. Unsurprisingly, these corresponded to the main production and populated areas, respectively. Most interesting were intermediary situations involving assembly markets such as Kaolack (15) or Diourbel (19) that had similar inflows and outflows. These markets are believed to be critical for the functioning of national trade. Interestingly, a great number of market pairs did not exchange any goods, but no market was completely isolated, i.e., with null flows. Trade flow was not necessarily direct between origin and destination markets, since production could transit through intermediary markets. Therefore, markets with null flow might temporarily host some production that was subsequently transported to another market until gradually reaching its final destination. Such intermediary transits were not described by the model. Caution should therefore be taken when interpreting flow values. Since they may not correspond to actual estimation of production flows which are challenging to validate, but rather to an indicator of trade intensity between two markets.

As an illustration, Figure 3.9 shows the impact of the transaction costs on the trade flows for three main markets in 2013 and 2014. Flows in 2012 were very similar to 2013 but lower in intensity.

According to our model, Mbar (14) is a major producer market and one of the main source of production. Therefore, the integration *from* Mbar *to* the other markets drove the trade flows. In 2014, the trade possibilities were numerous and around half of them led to transfers of millet. The other half were markets which were either located in distant regions or associated with a low population. In 2013, trade opportunities were very limited and, as a result, the intensity of flows was higher and more concentrated. Interestingly, the market of Ndindy (20) and Birkelane (13) received 7 kt and 6 kt in 2013 from Mbar, but nothing in 2014. This clearly illustrates the non-linearity of the trade flows. A high quantity of trade does not necessarily imply that the receiving markets

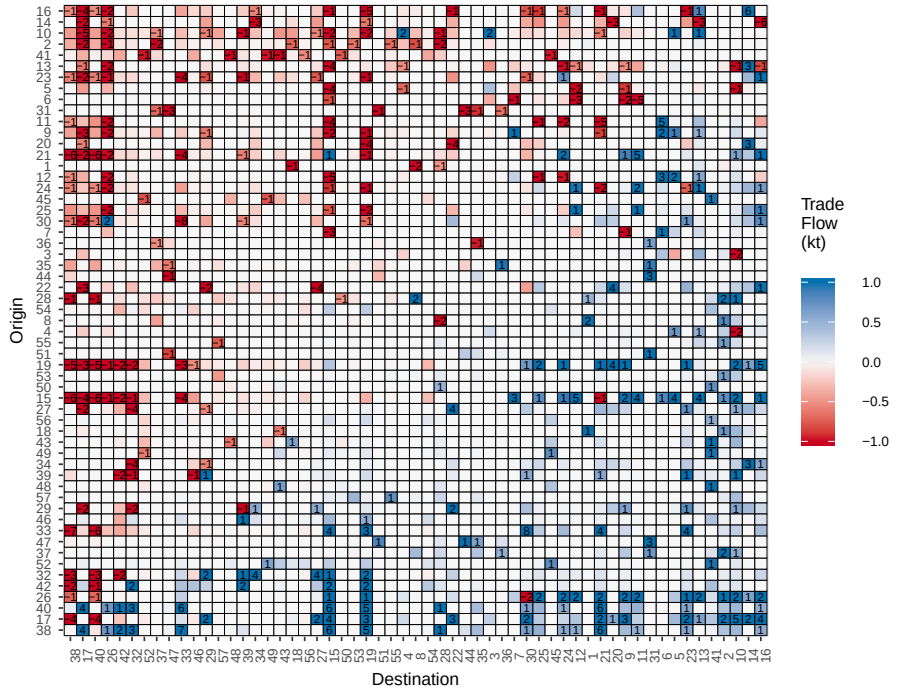


Figure 3.8: Average net trade flows (computed in kt of millet production), for all years, between all market pairs (for number-name matching refer to Figure 3.2). Numbers on trade flows maps indicate flows with higher intensity than 1 kt.

were the final destinations. On the contrary, in 2013, it is likely that a large part of the production was re-transferred to other markets, eventually reaching the same markets than 2014 but using indirect routes. However, taking longer routes leads to market inefficiency in production distribution and therefore, price transmission.

Kaolack (15) is one of the most important assembly markets of the country. It collects the millet production from the cropping areas and distributes it to consumer markets. This role was accurately described by our model: millet was transferred from several rural areas to Kaolack and consumer markets received production from Kaolack. This role seems to be exacerbated when the transaction costs are high such as in 2013 because this market remained well connected to the rest of the country even in case of low $\bar{s}$. It even shipped production to remote markets such as Kedougou (57).

Finally, Tilène (40), the market in Dakar, draws in production from the interior of the country. This market is a good example of the impact of s in the model. In 2013, net flows were distributed to almost all markets trading with Dakar. Since these markets also received production from producer markets, it does not entail that the millet they sent to Dakar was *produced* in their own

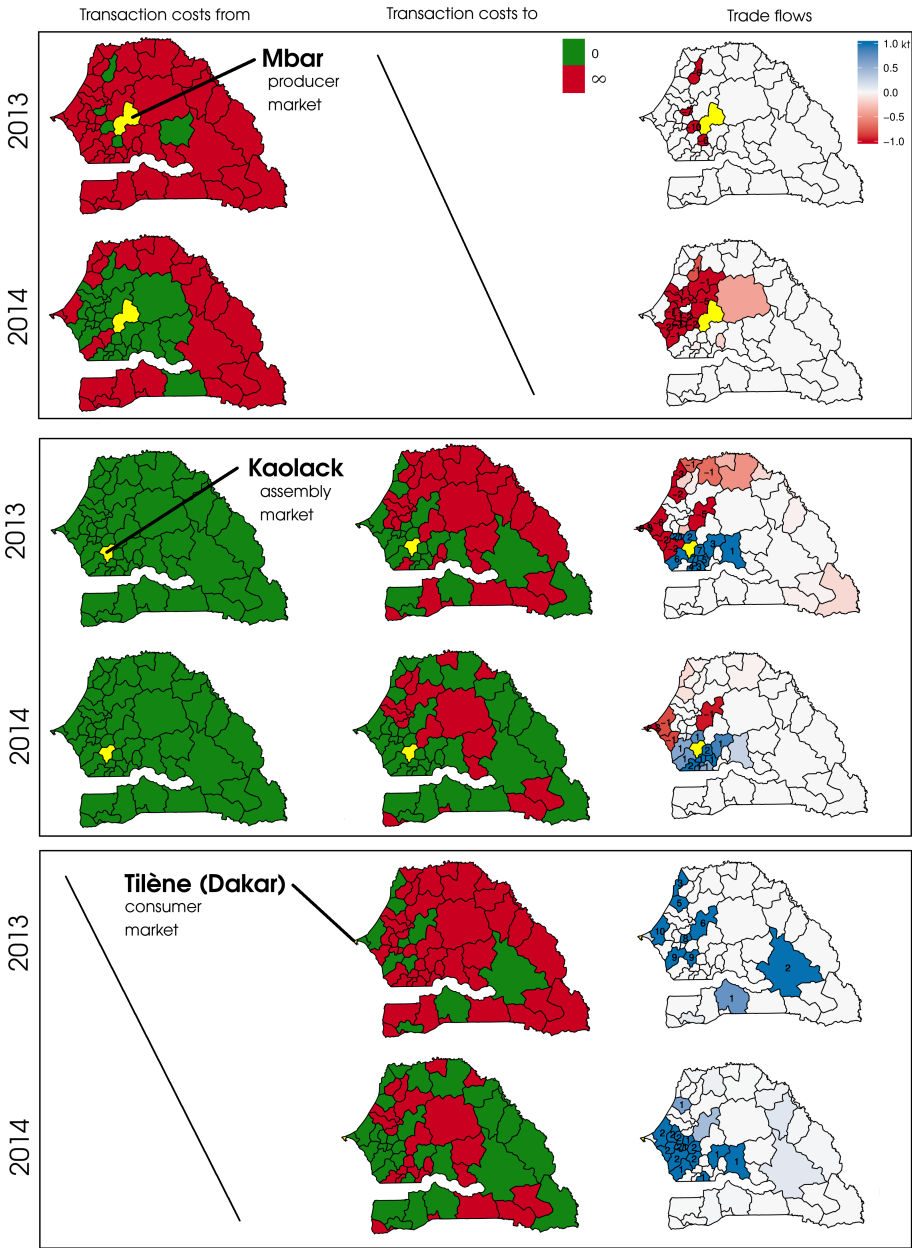


Figure 3.9: Trade flows of, and transaction costs, s, from and to the markets (in yellow) of Mbar (14, Producer), Kaolack (15, Assembly) and Tilène (40, Consumer) in 2013 ($\bar{s}$=0.43) and 2014 ($\bar{s}$=0.74). Numbers on trade flows maps indicate flows with higher intensity than 1 kt. Transaction costs maps from Mbar and to Tilène are not shown as the trade flow uses only one direction for these markets.

catchments areas. Rather, production was successively transported from one market to the other before reaching these areas. Unsurprisingly, Kaolack (15), Diourbel (19), Thiès (33) and Touba (17), the four major assembly markets, were among these markets. In 2014, net flows appeared to come from closer markets, reflecting the smaller trade constraints of that year.

A network visualization of the the trade flow intensity can be found in Figure A3 in Appendix. Trade flows of 2012 and 2013 are much more concentrated towards the main producer and consumer areas than in 2014. It is explained by the more important number of market pairs trading together in 2014 (503) than in 2012 and 2013 (373 and 375 respectively) due to the difference in $\bar{s}_{opt}$.

3.4.6 Limitations

Lacking data on self-consumption, we made the hypothesis that all the millet production was available for trade or that the individual self-consumption was similar everywhere. Millet was also assumed to be planted evenly within the cropland, which is inaccurate. However, the aggregation of production at catchment area level is likely to mitigate the effects of this simplification. The demand function would have been more realistically estimated if it accounted for the household income as well as taste and preferences (see Chapter 4 for an approach to get socioeconomic level estimation at fine granularity). However, these data are generally unavailable at local scale. Furthermore, substitution effects with other crops such as rice were not considered.

The roads data were checked and edited manually so that all major access routes were taken into account. The speed of these roads (due to legal limitations or quality) was unknown and therefore neglected. This could lead to an underestimation of the catchment areas of urban markets (or the opposite for rural markets) by underrating their attractiveness since they are often equipped with better roads compared to rural markets. Besides, historical and cultural preferences were not taken into account in the definition of catchment areas. Additionally, transport by rail or boat was not included since neither are used extensively for millet transportation (Bertholet *et al.*, 2004). Finally, inter-border trade was not considered but it occurs only marginally due to the constraints of international trade for small producers and the weak import demand for millet (Ndiaye and Niang, 2010).

When using CDR data to analyze the whole population, several bias might arise and limit the generalization of the analysis (see Chapter 2). First, we assumed that every agent (at least every trader) owns a mobile phone. In 2013, there were 93 mobile phone subscriptions per 100 inhabitants which implies that most of the Senegalese population owns cell phones (ITU World Telecommunication, 2016a). However, by using only CDR data, some sociodemographic sub-groups, in particular the poorest, are still left out the analysis. This means that the social network of poorer traders could have been underrated. Second, Sonatel's market share reached nearly 62% of the cell phone market in 2013

which entails that a selective bias may arise from the different demographic sub-groups targeted by each operator (Autorite de Régulation des Telecommunications et des Postes, 2013).

We selected phone calls as the proxy of the business network between two market areas. This choice relies on the hypothesis that the social network and business network are similar as business calls are indistinguishable from the others. In doing so, some business ties could have been overestimated when the social network is stronger than the business network between two market areas, and the other way around. For the sake of simplicity, the transaction costs associated with social capital were defined as a dummy variable, i.e., null or infinite depending on a threshold on the number of calls. However, a large range of situations exists between these two extreme cases and could have been considered by rewriting Eq. 3.3 as Eq. 3.11:

$$p_i - p_j \leq d_{ij}r + f(s_{ij}) \tag{3.11}$$

with $f(.)$ some increasing and convex cost function.

This approach would be interesting to gain further insight into the actual relationship between social capital and transaction costs. However, without proper validation data (i.e. the actual transaction costs), one can only speculate on the functional form $f(.)$. In this study, a simple binary model was preferred instead of arbitrary decide on $f(.)$.

3.4.7 Policy implications

Several policies could be put in place to cut down transportation costs, the main source of inefficiency in millet markets. As pointed out previously by Teravaninthorn and Raballand (2009), it is not so much that transportation costs in Africa are higher than those in other developing regions such as China, but transportation prices are much higher. Administrative barriers are at least as important as poor roads in hindering the market functioning, particularly in western and central Africa. Removing restrictions on the entry of new transportation companies into the market should stimulate competition and reduce the high profits of local trucking companies. Rwanda is a well-known example of an African country that deregulated its transport sector and saw a dramatic drop in transport prices almost overnight (Teravaninthorn and Raballand, 2009). The trucking industry in Senegal is dominated by a large number of very small operators who own and operate an obsolete trucking fleet (Hamilton, 2010). Improving infrastructure and the trucking industry are therefore expected to have a major impact on market functioning and prices.

Understanding the role that social capital plays in market exchanges is essential for policy design. Finding approaches to facilitate search and fostering trust will likely improve trade exchange. The only gateway for policymaker is to work on social structures via formal institutions (e.g., legal institution,

public market information system) or interpersonal relationships (e.g., fostering traders' associations or the learning of different languages). Functioning institutions and strong governance make transactions impersonal leading to economic efficiency (Rashid *et al.*, 2010; Durlauf and Fafchamps, 2005). Law and court should, therefore, be strengthened especially in developing countries where many transactions are small and buyers and sellers are too poor for court action to yield reparation (Bigsten *et al.*, 2000; Fafchamps and Minten, 2002). Whether or not social capital simplifies market trade is an indicator of the efficiency and reach of formal institutions.

3.5 Conclusions

In this study, we demonstrated the effect of social capital on millet prices in Senegal. Social capital was approximated by the business network of economic agents using a unique data set on mobile phone communications between 9 million people. Our approach was a spatial equilibrium model that accounts for both transportation (r) and transaction costs (s) and successfully estimated millet prices in 57 markets in Senegal for three contrasted years.

The transportation costs were modeled proportionally to the distance and accounted on average for the majority of price differentials in the country ($\sim$20-40%). Clearly, this is the main source of inefficiency in millet markets in Senegal. Several components of freight cost are probably included in this value such as maintenance, opportunities, etc. Other transaction costs proportional to the intermarket distance could also be involved but they remain challenging to isolate. In particular, the impact of social capital on market functioning could already be accounted for in this value due to, for instance, mistrust in remote traders.

The transaction costs, modeled as null or infinite, accounted for between 1 and 9% of the price variance, demonstrating the effect of social capital on millet prices. Interpreting s and its specific impact is not straightforward and remains challenging to validate. The impact of s is marked for two years following a poor harvest, e.g., in 2012 and 2013. In this situation, the assumption is that events result from traders managing their risk by focusing their commercial transactions on well-known and therefore less risky markets, i.e., the aversion to risk is higher following a poor production. However, additional data from other years are still required before reaching firm conclusions on this point.

Further insights can be expected from expanding the model to other countries in the Sahel as well as exploiting multiple years of mobile phone data. In particular, specifically target traders in the call network would allow to validate the assumption regarding the correlation between business network and social network of the whole population. It would be also interesting to investigate the seasonality (i.e. month by month) of the effect of social capital on transaction costs. Finally, data to validate the trade flows extracted from the spatial equilibrium model would allow to go further in the understanding of the factor driving the production flows at national scale.

This work opens new avenues for (i) research on social capital and market integration, (ii) better integration of the two first pillars of food security, i.e., availability of and access to food, and (iii) more comprehensive implementation of early warning systems for food security in the region. Further insights can be expected from expanding the model to other countries in the Sahel as well as exploiting multiple years of mobile phone data.

Chapter 4

Combining Disparate Data Sources for Improved Poverty Prediction and Mapping[1]

Don't ask me what poverty is because you have met it outside my house. Look at the house and count the number of holes. Look at my utensils and the clothes that I am wearing. Look at everything and write what you see. What you see is poverty.

A poor man, Kenya 1997 (Narayan-Parker and Patel, 2000)

Highlights

- Spatially finest poverty maps are essential for improved diagnosis and policy planning, especially keeping in view the SDGs (**relevance**).
- Big data sources like CDRs and EO data have shown promise to provide inter-censal statistics (**timeliness**).
- We outline a computational framework to efficiently combine disparate data sources, like environmental data, and mobile data, to provide accurate predictions of poverty and its individual dimensions for finest spatial micro-regions in Senegal (**accuracy**).
- We use Gaussian Process regression, a Bayesian learning technique, providing uncertainty associated with predictions. We perform feature selection using elastic net regularization to prevent over-fitting.
- The combination of both data sources provide better results than using each dataset separately or together in a single model, especially for individual dimension of poverty (**accuracy**).
- The different data ecosystems need not share any data between them. The individual data sets remain private within their specific ecosystems, and only the output predictions and the associated variances are shared (**access, protection**).

[1]Adapted from Pokhriyal, N.[†] and **Jacques, D. C.**[†] 2017. Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences, 114(46):E9783 – E9792. ([†]equally contributed to this work)

Abstract

More than 330 million people are still living in extreme poverty in Africa. Timely, accurate and spatially fine-grained baseline data is essential to determine policy in favor of reducing poverty. The potential of *Big Data* to estimate socio-economic factors in Africa has been proved. However, most current studies are limited to using a single data source. We propose a computational framework to accurately predict Global Multi-Dimensional Poverty Index (MPI), at a finest spatial granularity and coverage of 552 communes in Senegal using environmental data (related to food security, economic activity and accessibility to facilities) and call data records (capturing individualistic, spatial and temporal aspects of people). Our framework is based on Gaussian Process regression, a Bayesian learning technique, providing uncertainty associated with predictions. We perform feature selection using elastic net regularization to prevent over-fitting. Our results empirically prove the superior accuracy when combining disparate data than using each data sources separately or together in a single model (Pearson correlation of 0.91). Advantageously, the different data ecosystems need not share any data between them. The individual data sets remain private within their specific ecosystems, and only the output predictions and the associated variances are shared. Our approach is used to accurately predict important dimensions of poverty: health, education and standard of living (Pearson correlation of 0.84-0.86). All predictions are validated using deprivations calculated from census. This method can be used to generate poverty maps frequently, and its diagnostic nature is, likely, to assist policy makers in designing better interventions for poverty eradication.

4.1 Introduction

More than 330 million people are still living in extreme poverty in Africa (Beegle *et al.*, 2016). Consequently, the goal to "eradicate extreme poverty for all people everywhere by 2030" tops the list of the 17 Sustainable Development Goals adopted by world leaders, at the United Nations summit in September 2015. The lack of good-quality and fine-grained data to assess poverty regularly features in discussions of the development agenda for Africa (Devarajan, 2013; Jerven, 2013a). Timely measurement and availability of data are vital in ending poverty. Despite the nature of the strategies used to reduce poverty, governments and development agencies need a baseline depiction. Poverty maps provide such a spatial distribution of the socio-economic deprivations, and, help policy makers assess the impact of interventions. For efficient targeting of policies at micro-regions and specific demographics, poverty maps should be made available at the finest administrative unit of planning. Also these values should be dis-aggregated into individual dimensions of poverty, like deprivations in education, standard of living, health, etc (Alkire *et al.*, 2014).

Currently, the most reliable way to estimate poverty is through intensive socio-economic household surveys. However, this approach is costly and time consuming and can only be realistically carried out for a small sample of households. The extrapolation of the local poverty estimation to a larger scale is traditionally done by exploiting links between census (wide area) and survey (smaller area coverage) data through small area estimation methods (Elbers *et al.*, 2003; Rao and Molina, 2015). These techniques depend on the timely availability of census, which is typically collected every 10 years, and whose analysis is delayed, for poorer economies by years, making timely updates of poverty challenging (see Introduction).

Recently, there has been a growing interest in realizing the potential of *Big Data* to understand societal development in Africa. However, most current studies are limited to using single source data sets, such as mobile phone data (Blumenstock *et al.*, 2015a), or satellite imagery (Jean *et al.*, 2016). Since poverty is a complex phenomenon, understanding it using multiple *lenses* obtained from diverse datasets, will help to chart more accurate maps for poverty.

Several studies highlight that significant spatial variation of poverty may be due to a variety of geographic factors, including agro-meteorological conditions, accessibility and proximity to markets, access to land, etc. (De Sherbinin *et al.*, 2008; Berdegué *et al.*, 2015) (Table 4.2). Earth Observation satellites collect data on metrics such as nighttime lights, vegetation cover or meteorological conditions. The unique features of such data sets are their global coverage, high revisit capability and free availability. A complementary resource lies in Geographic Information Systems (GIS) analysis. In particular, proximity to important services (schools, hospitals) and density of infrastructure (such as roads) are all factors that might contribute to alleviate poverty (Okwi *et al.*,

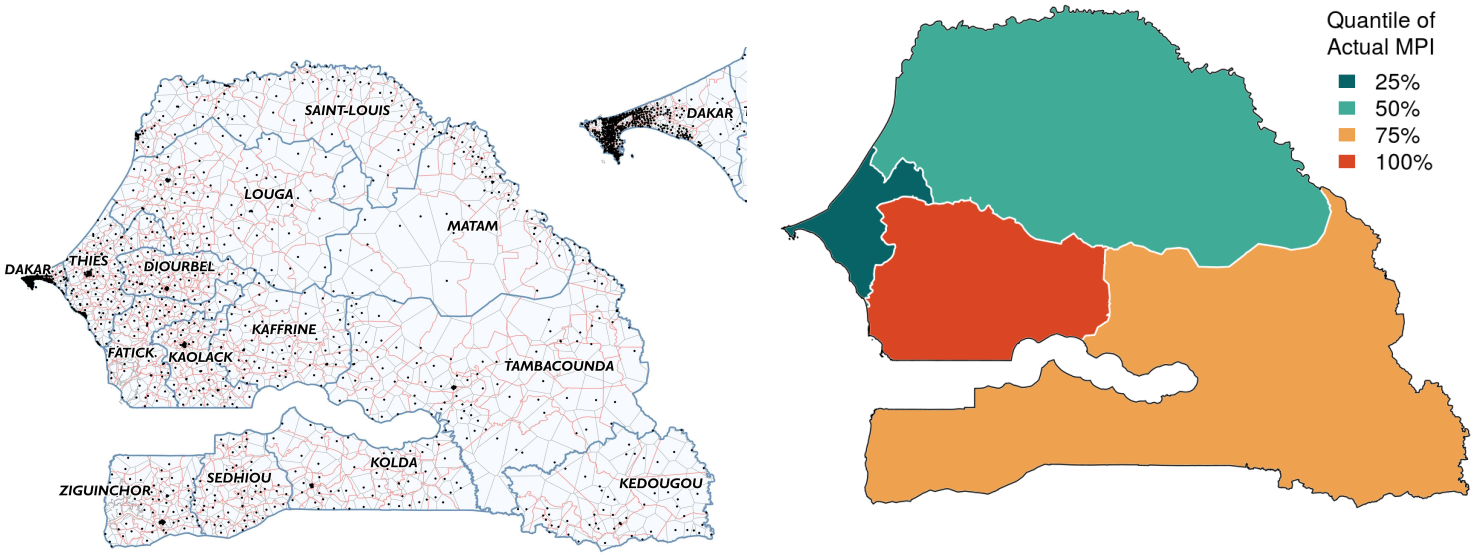


Figure 4.1: On the left, is a composite map of Senegal. Black dots depict the location of the 1666 mobile towers (antennas). The Voronoi tessellation formed by these towers is shown in gray. The commune (which is the finest administrative unit in Senegal) boundaries are shown in red. There are 552 communes with 431 rural communes and 121 urban centers. The navy blue boundaries are those of regions, which are the coarsest administrative unit in Senegal. There are 14 regions, which are named in the map. On the right, is the current (2016) map of Global MPI for 4 divisions of the country (West, North, South and Center).

2007).

While satellite and other geographic data (e.g. roads, market locations...) are apt to observe and understand the availability and access to natural resources and man-made structures, they lack information about population structure, especially the socio-economic ties, cultural interactions and micro and macro-behavior that is essential to understand poverty. One way to study societal interactions is provided by the widespread use of digital technologies (Eagle *et al.*, 2009b). Internet is still finding ground in Sub-Saharan Africa. However, mobile phones is a prevalent technology with adoption rates of more than 70%, even with 43% of population living in abject poverty (Peña-López *et al.*, 2016). Such widespread use of mobile phones generates an unprecedented volume of data called call data records (CDRs). CDRs capture how, when, where and with whom individuals communicate. This data, traditionally used by the telecommunication companies for billing purposes, capture both micro and macro patterns of human interaction, while preserving the individual anonymity via spatial and temporal aggregation. Recently, CDRs have been studied to estimate poverty in different countries (Table 4.3).

Poverty has traditionally been measured in one dimension, usually income (or consumption); also known as income poverty. Another internationally comparable measure is the Global Multidimensional Poverty Index (MPI[2]), which complements income poverty, and is created from nationally representative Demographic and Health Surveys, and Multiple Indicator Cluster Survey (DHS-MICS) (Alkire and Santos, 2010). It was developed by the Oxford Poverty & Human Development Initiative (OPHI) and the United Nations Development Program. It is a composite of 10 indicators across three critical dimensions - education (years of schooling, school enrollment), health (malnutrition, child mortality), and living conditions (Cooking Fuel, Sanitation, Access to Drinking Water, Electricity, Floor, Asset Ownership).

MPI is calculated as a product of the incidence of poverty (H) and the average intensity across the poor (A). The incidence or headcount ratio of poverty is the proportion of the population that is multi-dimensionally poor. The average intensity of their poverty is the average proportion of indicators in which poor people are deprived. MPI is robust to decomposition within relevant sub-groups of populations, like urban vs. rural, geographic regions (districts/provinces/states), and gender; so that targeted policies can be planned for specific demographics.

Global MPI has some limitations. Though it has been defined from available variables in global surveys (DHS-MICS), some of the *potential* dimensions of poverty (like gender, income, employment) are not directly incorporated. However, due to the wide availability of these surveys, Global MPI can easily be estimated in more than 100 countries covering 5.2 billion people (Alkire and Santos, 2010). Consequently, it represents a benchmark index, more interesting

[2]If not specified, MPI refers to global MPI throughout the Chapter.

than the single dimension poverty line, for replication of this study in another country. It is worth mentioning that countries can also adapt the multidimensional poverty approach to select different indicators and/or update weights that align better with their nation's poverty measure. For instance, Mexico, Colombia and Chile, have implemented their own version of national MPI using additional dimensions than global MPI such as employment and social protection, when data are available (www.mppn.org). Throughout the Chapter, 'poverty' refers to the Global MPI, and 'dimensions' refers to education, health and standard of living.

The study focuses on Senegal, a Sub-Saharan country, that suffers from persistently high poverty. The communes towards the interior of the country have more poverty compared to the rest (see Figure 4.5 – bottom, representing MPI at commune level). The west regions containing the capital city Dakar, and communes neighboring the coastal boundary are less poor than the rest of the country. Of special interest is the spatially large division in the south, consisting of the regions of Tambacounda, Kedougou and Kolda, which are depicted as one color on the current map in Figure 4.1, but have communes of varying poverty values spread throughout. Interestingly, the communes in Kedougou region in the extreme southeast corner of Senegal are predicted as wealthier than other communes within the region. The communes in the region of Ziguinchor, in the southwest corner, are wealthier as compared to other communes in the south. This is attributed to the fact that Ziguinchor is the second largest city in Senegal, with economic advantage from being a port and a tourist center.

The 121 urban centers are shown as small circles on the map, and, in general, have less poverty values when compared to rural communes. The population in urban centers is generally richer than the population living in adjacent rural communes. This is true even for very poor communes of Senegal in the regions of Kaffrine and Tambacounda in the center for which the contrast is even higher. The urban centers bordering with the neighboring country, Mauritania, in the north-east, are wealthier; this could be attributed to the economy of the Senegal river basin and to cross-border trade. The predominantly urban areas in Dakar are shown enlarged in the map. All communes in Dakar are well-off than the rest of Senegal, because of the concentration of economic activity over the years.

This study uses mobile phone data in the form of CDRs, and data related to food security (availability and access components), economic activity and access to services are grouped together as environmental data (Table 4.1). The CDR variables capture not only the basic phone usage statistics of a user, but also the regularity, diversity, and spatio-temporal variability in the user's mobile interactions. The poverty maps are produced at the spatially finest level of policy planning, called *communes*, and validated at that level using the concurrent census data. Current poverty maps, based on Global MPI (see Figure 4.1) and consumption-based measures (Le Soleil, 2017), do not exist uniformly for all communes of Senegal. The map produced by our analysis is available for all 552

Summary Statistics	CDRs	Environment Data	Census	OPHI MPI data
Timeline	Jan-Dec 2013	1960-2014	2013	2013
Number of total calls & text	11 Billion	N/A	N/A	N/A
Number of unique individuals	9.54 M	N/A	14 M	N/A
Spatial granularity of available data	Antenna-level (1666)	vector data - 100 m -1 km	Household-level	Region-level (14)
Cost incurred in data collection & preparation	Low/no cost (data exhaust)	Low/no cost (data exhaust)	USD 29 Million	Very high cost, and human expertise
Frequency of update of data	Real-time	∼1 year	10 years	3-5 years

Table 4.1: Summary statistics and characteristics of the data used - CDRs, environment, census, OPHI MPI poverty index.

communes (see Figure 4.5). Such maps can be generated frequently in between cycles of surveys and censuses, since CDR and environment data are available at fine temporal granularity.

Our objective is to present a computational framework that integrates disparate data sources to accurately predict the Global MPI and its individual dimensions, at finest level of spatial granularity. This framework consists of models trained independently on each data source. Each source specific model employs Gaussian Process (GP) regression (Rasmussen and Williams, 2006) to infer poverty values. GP falls under the class of kernel methods, where the choice of different kernel functions, enables one to learn different non-linear relationships between the independent and target variables. Each GP based model provides a probabilistic estimate of poverty for a given commune, including the mean and variance of the estimates. The variance provides a measure of uncertainty, which allows combining the predictions from the multiple data sources.

An important advantage of this methodology is that the different data ecosystems need not share any data between them. The individual data sets remain private within their specific ecosystems, and only the output predictions and the associated variances are shared.

Table 4.2: Brief review of poverty estimation methods based on environmental data.

Reference	Poverty proxy	Model	Important variables	Main conclusions	Region
Dasgupta *et al.* (2005)	daily consumption expenditure	regression, correlation	indoor air pollution (wood/charcoal use), access to clean water, no sanitation, Diarrhea, outdoor air pollution (number of deaths from PM10)	substantial variability across countries	Cambodia, Lao PDR, Vietnam
Vista and Murayama (2011)	per capita income	regression	mean road density, share in internal revenue allotment, agrarian reform accomplishment rate, population growth, distance to major cities, mean elevation, percentage of slope with agri. limitations and mean annual rainfall	spatial variation in poverty is mainly caused by disparities on access to road infrastructure	Philippines
Amarasinghe *et al.* (2005)	food expenditure	regression and clustering	proportion of irrigation land, average landholding sizes	poverty maps show significant spatial clustering of poor and non-poor areas	Sri Lanka
Okwi *et al.* (2007)	per capita expenditure	spatial regression	slope, soil type, distance/travel time to public resources, elevation, type of land use, demographic variables	increasing access to roads and improving soil conditions would result in decline in poverty	Kenya
Minot *et al.* (2006)	per capita expenditure	regression	distance to town, soil quality, slope	poverty in the remote areas is linked to low agricultural potential and lack of market access	Vietnam
Rogers *et al.* (2006)	household expenditure	discriminant analysis	distance to market, agro-climatic variables, diseases risk, livestock density	satellite-derived variables tended to dominate the list of selected variables that determine poverty predictions	Uganda
Jean *et al.* (2016)	household consumption expenditure, asset wealth	transfer learning (deep learning)	roofing material, distance to urban areas	interesting potential of machine learning method using limited training data	Nigeria, Tanzania, Uganda, Malawi, Rwanda

Benson *et al.* (2005)	household expenditure	spatial regression, geographically weighted regression	crop diversity, education, non-agricultural economic activities	spatial non-stationarity of the relationship between poverty and its determinants	Malawi
Kam *et al.* (2005)	household income	geographically weighted regression	education, accessibility and services	high poverty incidence that correspond with ecologically depressed areas. However, other livelihood-influencing factors such as education, accessibility and services are significantly correlated with poverty	Bangladesh
Watmough *et al.* (2016)	relative welfare (female literacy, land ownership, deprived class, and water source)	random forests	travel time to market towns, percentage of a village covered with woodland, and percentage of a village covered with winter crop	satellite sensor data are strongly associated with aspects of rural welfare for an extensive region of a developing country	India

Table 4.3: Brief review of poverty estimation methods based on CDR data.

Ref	Data Source	Model (Number of features)	Sample Size	Time period	Pearson's R	Spatial Resolution of Validation	Poverty Measure	Region
Blumenstock *et al.* (2015a)	CDR & Phone survey	Linear Regression (5088)	1.5 M (CDR) + 856 (Survey)	9 months	0.68	492 DHS Clusters	DHS composite wealth index	Rwanda
Soto *et al.* (2011)	CDR	Support Vector Machine (279)	500K	6 months	0.80	1200 geographical regions	Socio-economic levels (A,B,C)	Urban area in a Latin American city
Smith *et al.* (2013)	CDR	Pearson correlation (12)	5 M	5 months	0.78-0.85	11 sub-prefecture level	MPI (OPHI)	Ivory Coast
Pokhriyal and Dong (2015)	CDR	Linear Regression (33)	9 M & 150K	12 months	0.82	14 regions in Senegal	MPI (OPHI)	Senegal

4.2 Data

4.2.1 Target Country

Senegal is a Sub-Saharan country which ranks 162 with a Human Development Index of 0.49 (UNDP, 2017). As one of the poorest country in the world, it has 75% of population living in multidimensional poverty (OPHI, 2013). Senegal is composed of 14 coarsest administrative units called regions, which are further divided into 45 administrative units called departments. The finest level of administrative units are called communes. There are 552 communes (121 as urban centers, and 431 rural (Figure 4.1).

4.2.2 Data Sources

CDRs A call data record (CDR) consists of an identifier with the caller, and callee, the antenna location of the caller and callee, the time of the call, duration of the class, and a flag indicating if the record is a text or a call (see Chapter 2). A CDR is generated each time a call or text is placed. The data belongs to the subscribers of Sonatel – Orange, which is the dominant telecom provider in Senegal. It is anonymized, and spans a period from January 1 to December 31, 2013. It contains more than 9.54 million unique aliased mobile phone subscribers. By comparison, the population of Senegal in 2013 was 14.13 million. Additionally, the geographical coordinates of the mobile antennas are known, and shown in Figure 4.1.

Environmental Features Based on literature, several environmental features that may have a relationship with poverty have been explored (Table 4.4). They are either based on GIS (e.g. roads, market locations), Earth Observation data or weather stations.

Census The *Agence Nationale de la Statistique et de la Demographie* (ANSD), which is the National Statistics Office of Senegal, provided us with a 10% sample of the 2013 census (called *Recensement General de la Population de l'Habitat de l'Agriculture et de l'Elevage* – RGPHAE). The data is evenly sampled across the entire population of Senegal. It has data from 1.4 million individuals, spread across 150,000 households, characterizing information related to demographic statistics (mortality, fertility, migration, literacy, occupation etc.), along with habitat features, such as type of roof, floor, access to drinking water, sanitation and agriculture practices. The advantage of census is that it represents important national statistics at the level of individuals. Brief statistics of the data sources are given in Table 4.1.

Table 4.4: Source, unit and expected relationship to poverty of each environmental variables used in this study.

Feature (Number of Statistics)	Unit	Type of data	Endogeneity	Data Sources	Expected relationship to poverty
			Food security (Availability)		
Temperature (annual, annual range, diurnal range, warmest month, warmest quarter, coldest month, coldest quarter, wettest quarter, driest quarter, isothermality) (11)	degree Celsius	ground	exogenous	WorldClim database (1960-1990) (Hijmans *et al.*, 2005)	high temperature (+)
Precipitation (annual, wettest month, wettest quarter, driest month, driest quarter, warmest quarter, coldest quarter, coefficient of variation) (8)	mm	ground	exogenous	WorldClim database (1960-1990) (Hijmans *et al.*, 2005)	low precipitation (+)
Elevation (1)	m	remote sensing	exogenous	CGIAR-SRTM data aggregated to 30 seconds (http://www.diva-gis.org/)	high elevation (+)
Slope (1)	degree	remote sensing	exogenous	CGIAR-SRTM data aggregated to 30 seconds	high slope (+)
Soil Type (14)	% of territory	ground	exogenous	Soil and Terrain Database for Senegal and the Gambia (version 1.0), scale 1:1 million (SOTER Senegal Gambia, www.isric.eu/projects/soter-senegal-and-gambia)	poor agronomic soil (+)
NDVI (2)	-	remote sensing	endogenous exogenous	10-day temporal synthesis of 1 km SPOT-VEGETATION satellite images (2000-2013) (www.vgt.vito.be)	low NDVI (in rural areas) (+)
Crop Production (7)	t	ground	endogenous	Direction de l'Analyse, de la Prévision et des Statistiques Agricoles (DAPSA) 2000-2014 database (Direction de l'Analyse, de la Prévision et des Statistiques Agricoles, 2013)	low production (in rural areas) (+)

Food security (Access)					
Millet Price (1)	FCFA/kg	ground	endogenous	Modeling based on local supply and demand Jacques *et al.* (2015)	high millet price (+)
Proximity to Urban centers (Market) (1)	km	GIS	endogenous	ANSD	far from urban centers (+)
Proximity to Main Roads (1)	km	GIS	endogenous	Open Street Map (www.openstreetmap.org)	far from main road (+)
Economic activity					
Nighttime Lights (2)		remote sensing	endogenous	Version 4 of the 2013 nighttime lights time series captured by the Operational Linescan System of the Defense Meteorological Satellite Program (stable lights)	low density of light (+)
Density of Roads (1)	km	GIS	endogenous	Open Street Map	low density of roads (+)
Land Cover					
Land Cover (20)	% of territory	remote sensing	exogenous / endogenous	2005 1:100.000 scale Senegal Land Cover Map produced by the Global Land Cover Network Leonardi (2008) based on GlobCover 2005 map (Defourny *et al.*, 2009)	urban areas (-), cropland (+), forest (+), grassland (+)
Access to facilities					
Proximity to School / University (1)	km	GIS	endogenous	Open Street Map	far from School / University (+)
Proximity to Water tower (1)	km	GIS	endogenous	Open Street Map	far from water tower (+)
Proximity to Hospital (1)	km	GIS	endogenous	Open Street Map	far from hospital (+)

4.2.3 Feature Extraction

CDRs We have access to more than 11 billion mobile phone transactions involving call and text for a year in Senegal. Each time a call or text is placed, it is logged as a transaction. Missed, forwarded and other *undelivered* calls were removed from the logs.

To extract important features that quantify the mobile usage pattern of a subscriber, we focus on well-studied metrics capturing the individualistic, spatial and temporal patterns of the subscriber (Frías-Martínez *et al.*, 2013; Bogomolov *et al.*, 2014; de Montjoye *et al.*, 2013c). The individual aspects quantify the typical usage pattern of an individual. Some of the metrics that belong to this category are the number of active days, the number of contacts, the average call duration, percent nocturnal, etc. Spatial metrics are the ones that quantify the typical movement pattern of an individual. Examples of spatial metrics for a subscriber include radius of gyration, entropy of antennas, etc. There are 43 core features (briefly described in Table 4.5), extracted using the *Bandicoot* toolbox (de Montjoye *et al.*, 2016). All features were calculated at monthly granularity capturing some of the temporal aspect of a subscriber, resulting in 43×12 CDR based features.

Second step is to localize each subscriber, i, to his/her home antenna. A home antenna, h_i, is calculated as one from where the subscriber makes the most nocturnal calls (from 7 pm - 7 am) during each month. We filtered out individuals who made less than 5 calls during each month, and who were not active for at least half of the year within the range of their home antennas. This ensures that individuals are reliably allocated to their home antennas. After the filtering step, the sample contained 6.19 Million individuals (65% of the original subscriber population).

We then compute the average feature value for each antenna site by computing the average of the feature values for all individuals that consider that antenna as their home:

$$m_a^{(f)} = \frac{1}{N_a} \sum_{i:h_i=a} m_i^{(f)} \tag{4.1}$$

where $m_i^{(f)}$ is the f^{th} feature value.

Finally, we compute the feature value for each commune as the weighted average of all antennas whose Voronoi polygon intersects with the commune boundary as:

$$m_c^{(f)} = \frac{1}{\sum w_{c,a}} \sum w_{c,a} m_a^{(f)} \tag{4.2}$$

The weight $w_{c,a}$ is the ratio $\frac{Area(c \cap a)}{Area(a)}$, which is a measure of how much of the Voronoi cell for antenna a falls within the boundary of commune c. To study how well has the *Voronoi-based* approach performed in assigning people to their communes, we study the correlation of the commune population estimated by

this approach and that calculated from census. The Pearson's correlation is reported as 0.85 with a p-value of < 0.00001, thus ensuring the validity of our approach.

Environmental Features In this study, we focus on 3 broad categories of environmental features: food security (divided into the availability and access components), economic activity, and access to services (Table 4.4). These three categories cover most of the *environmental* features that have been shown to be significantly related to poverty in the literature (Table 4.2).

Food security is mainly described by agro-meteorological measurements (temperature, precipitation, slope, elevation, soil type) that drive the agricultural production (crop production), one of the most important input, along with livestock and fishing, of food availability in the country. On the other hand, access to staple food can be approximated by the average millet prices observed in the markets (retail prices in 56 local markets). Millet serves as the main local staple food crop in the country which makes it a potentially good indicator of poverty (see Chapter 3). In addition, proximity to main road and urban centers were also computed to describe the connectivity to major markets.

The economic activity corresponds to the intensity of urbanization. In the literature, the nighttime lights are the most frequently used to describe poverty using remote sensing data (Njuguna and McSharry, 2017). Moreover, a clear link between household wealth and the level of night light emissions has been shown before (Weidmann and Schutte, 2016). The underlying hypothesis is that economic activity and urbanization are strong indicators of living standards.

Finally, the access to services can help to predict some of the individual indicators of poverty. In particular, the proximity to school, water towers and hospitals can be used to determine the deprivation in education, water and health, respectively.

The raw environmental data is available either in raster grid (at different spatial resolutions) or in vector format. As a first step, all vector data were converted into raster grid format. Then, all data layers were resampled (using nearest neighbor approach) at a spatial resolution of 100m. Pixel values falling within each commune's boundary were averaged to give a unique value for that commune.

All environmental data are available at high spatial resolution with the exception of crop production and millet prices (see Table 4.4 for the data sources). Millet prices were available in 56 local markets, potentially missing some of the local heterogeneity. Production estimation features were derived from the Direction de l'Analyse, de la Prévision et des Statistiques Agricoles (DAPSA) database. The granularity of these features is at the department level. Cultivated areas were masked using the 2005 1:100.000 scale Senegal Land Cover Map produced by the Global Land Cover Network based on GlobCover 2005

map (Defourny *et al.*, 2009), which is the most accurate map for Senegal (Waldner *et al.*, 2015b). Since reliable information on the spatial distribution of each crop is unavailable, we make an assumption that they were grown evenly within the cultivated areas of a specific department. Therefore the production of a specific department was distributed evenly among all the 100m pixels that fell within the cropland of this department. This raster was then used to aggregate the production estimations by communes.

The Normalized Difference Vegetation Index (NDVI) is used as a proxy of potential agricultural production within a department. The NDVI, defined as the difference between near-infrared and red reflectances normalized by the sum of the two parameters, is a useful yield proxy in regions where water or soil fertility are the main limiting factors, such as Sahel (Samaké *et al.*, 2005; Rockström and De Rouw, 1997). For each pixel within cultivated areas, NDVI values above 0.2 during the growing season (July to November) were integrated (TNDVI), which limited the contribution of bare soil to the signal.

Table 4.5: List of core features extracted for each individual from CDR data using the Bandicoot toolbox (de Montjoye *et al.*, 2016). Features are grouped into categories based on prior research (Bogomolov *et al.*, 2014). These features are calculated for each month, so in total there are $43 \times 12 = 516$ features.

Features (Number of statistics)	Description
	Regularity
Interevent time (4)	The interevent time between two records of the user.
	Diversity
Number of contacts (2)	The number of contacts the user interacted with (call and text handled separately).
Entropy of contacts (2)	The entropy of the user's contacts, both for call and text.
Balance of contacts (4)	The balance of interactions per contact. This feature is calculated for text and call. For every contact, the balance is the number of outgoing interactions divided by the total number of interactions (in+out)
Interactions per contact (4)	The number of interactions a user had with each of its contacts.
Percent pareto interactions (2)	The percentage of user's contacts that account for 80% of its interactions.
Percent pareto durations (1)	The percentage of user's contacts that account for 80% of its total time spend on the phone.
	Active Behavior
Percent nocturnal (2)	The percentage of interactions the user had at night (call and text).
Percent initiated conversations (1)	The percentage of conversations that have been initiated by the user both for call and text.
Percent initiated interactions (1)	The percentage of calls initiated by the user.
Response delay (2)	The response delay of the user within a conversation (in seconds). This is calculated for text (standard deviation and mean of the response delay).
Response rate (1)	The response rate of the user (between 0 and 1).
	Basic Phone Use
Active days (1)	The number of days during which the user was active.
Call duration (2)	The standard deviation and the mean of the duration of user's calls.
Number of interactions (6)	The number of interactions.
Ratio of text & call interactions (1)	This computes the ratio of the text and call interactions.
	Spatial Behavior
Number of antennae (1)	The number of unique places visited.
Entropy of antennas (1)	The entropy of visited antennas.
Percent at home (1)	The percentage of interactions the user had while he was at home.
Radius of gyration (1)	Returns the radius of gyration, the equivalent distance of the mass from the center of gravity, for all visited places.
Frequent antennas (1)	The number of location that account for 80% of the locations where the user was.
Churn rate (2)	The standard deviation and mean of the frequency spent at every antenna each week.

4.3 Method

4.3.1 Gaussian Process Model for Predicting Poverty from a Single Data Source

To predict poverty for a commune from a single data source (CDR or environment), the following model is assumed:

$$y_i = \beta^\top \mathbf{x}_i + f(\mathbf{x}_i) + \epsilon \tag{4.3}$$

where y_i is the target poverty value and $\mathbf{x}_i$ is a vector of independent variables derived from the particular data source for the i^{th} commune. The first term is a linear combination of the independent variables. The function $f()$ models the non-linear relationship between y_i and $\mathbf{x}_i$. The residual term, ϵ, models the remaining unexplained noise, and is modeled as a zero-mean Gaussian random variable, i.e., $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

Without the non-linear term, $f()$ in (4.3), the model is equivalent to ordinary linear regression. However, a simple linear model is not rich enough to capture the relationships between the target and the independent variables (Figure 4.2 in Appendix), thus motivating the need for a non-linear term.

Instead of assuming a fixed parametric form for $f()$, we adopt a non-parametric approach, by assuming a Gaussian Process (GP) prior on $f()$. A GP is a stochastic process, indexed by $\mathbf{x} \in \mathbb{R}^d$. Any finite sample generated from it is jointly multivariate normal (Rasmussen and Williams, 2006). The generative process is defined by:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{4.4}$$

$$y_i \sim \mathcal{N}(\beta^\top \mathbf{x}_i + f(\mathbf{x}_i), \sigma_n^2), \forall i \tag{4.5}$$

where $m(\mathbf{x})$ is the mean of $f(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ is a kernel function that defines the covariance between any two evaluations of $f(\mathbf{x})$, i.e., $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$, and $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. For model simplicity, we assume that $m(\mathbf{x}) = 0$, which is a standard practice in GP based methods (Rasmussen and Williams, 2006).

Given a training set of examples, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, the GP prior on $f()$, and other terms in (4.3); the posterior distribution of y_* (for an unseen input vector, $\mathbf{x}_*$), is a Gaussian distribution, with the following mean and variance (See Appendix 'Gaussian Process Regression Model' for details):

$$\bar{y}_* := \mathbb{E}[y_*] = \beta^\top \mathbf{x} + \mathbf{k}^\top (K + \sigma_n^2 I)^{-1} \mathbf{y} \tag{4.6}$$

$$\sigma_*^2 := \text{var}[y_*] = k_* - \mathbf{k}^\top (K + \sigma_n^2 I)^{-1} \mathbf{k} + \sigma_n^2 \tag{4.7}$$

Here, $\mathbf{y} = [y_1, y_2, \ldots]^\top$, and K is a matrix which contains the kernel function evaluation on each pair of training inputs, i.e., $K[i, j] = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}$ is a vector of the kernel computation between each training input and the test input, i.e., $\mathbf{k}[i] = k(\mathbf{x}_*, \mathbf{x}_i)$, $k_* = k(\mathbf{x}_*, \mathbf{x}_*)$, and I is an identity matrix.

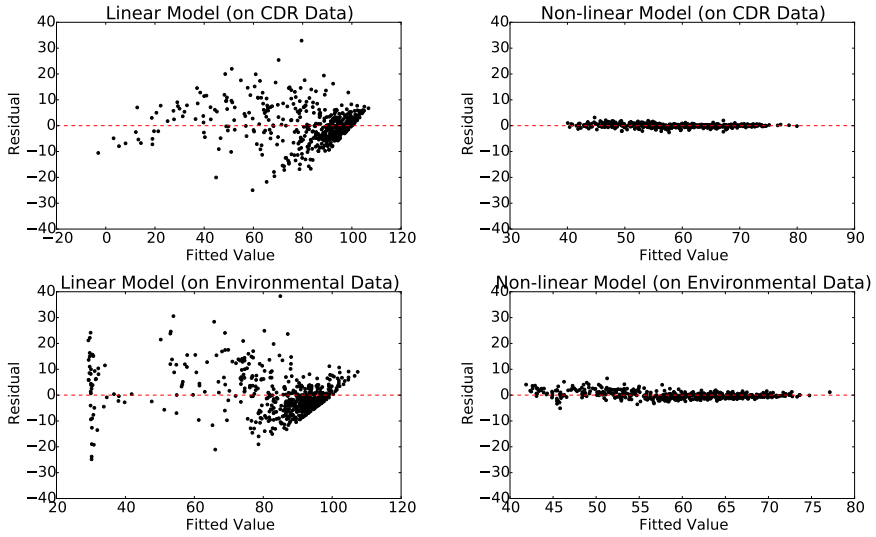


Figure 4.2: Residual vs. fit plots to predict incidence of poverty (H) using CDR (top panel) and environmental (bottom panel) data. *Left*: linear (Elastic Net Regression); *Right*: non-linear (Gaussian Process Regression, GPR). Linear model fits indicate non-linearity in the data. The residuals for GPR are normally distributed. *Shapiro-Wilk* test statistic - CDR: 0.97 (p-value $< 10^{-9}$); Environmental: 0.95 (p-value $< 10^{-9}$).

Understanding Model Uncertainty The predictive variance associated with the GP model, as calculated using Eq. 4.7, indicates the model uncertainty for a test target. The variance does not depend on the observed target values, but only on the inputs. The variance at a given test commune is directly related to how many similar communes (in terms of the CDR, environmental and spatial features) are available in the training data. For instance, if the predictive variance is high for a given test commune, it would mean that the relative density of the training feature vectors in proximity of the feature vector corresponding to the test commune is low, and hence the GP model will yield a higher predictive variance.

Choice of kernel function The role of the kernel function is to specify how the function values, $f(\mathbf{x})$ and $f(\mathbf{x}')$, vary as the function of their corresponding inputs, $\mathbf{x}$ and $\mathbf{x}'$. We use the following kernel function:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right) \exp\left(-\frac{\|\mathbf{u} - \mathbf{u}'\|^2}{2\ell_u^2}\right) \tag{4.8}$$

where $\mathbf{u}$ and $\mathbf{u}'$ are the spatial coordinates (latitude, longitude) of the commune centers corresponding to $\mathbf{x}$ and $\mathbf{x}'$, respectively. The first exponent term captures non-linear dependencies in the feature space. The second exponent term plays the same role, but in the geographic space and models the spatial auto-correlation as a continuous function. The parameter σ_f^2 is the variance of

the stochastic process f, ℓ is the process length-scale for the feature space part, and ℓ_u is the process length-scale for the spatial part.

The quantities $\beta, \ell, \ell_u, \sigma_n^2$, and σ_f^2 are estimated by maximizing the marginalized log-likelihood of the training data. To remove the effect of spurious features (i.e. perform feature selection), we couple the GP model with Elastic-net regularization (Zou and Hastie, 2005) during the model learning phase. This allows for automatic relevant feature selection and learning a parsimonious model that improves interpretability.

Model Training The unknown parameters of each source-specific model in (4.3) are: the parameter β of the linear component, the hyper-parameters of the kernel function, ℓ, ℓ_s, σ_f^2, and the variance of the error term, σ_n^2. These are estimated by maximizing the marginalized likelihood of the target poverty values in the training data, $\mathbf{y}$. The marginalized likelihood is obtained by taking the integral of the likelihood times the prior:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{x})d\mathbf{f} \tag{4.9}$$

where the matrix $\mathbf{X}$ contains the training input vectors as rows, and $\mathbf{f}$ is a vector containing the latent function values for the inputs in $\mathbf{X}$. The GP prior means that $p(\mathbf{f}|\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, K)$ and the likelihood is a Gaussian, i.e., $p(\mathbf{y}|\mathbf{f}, \mathbf{X}) \sim \mathcal{N}(\mathbf{X}\beta + \mathbf{f}, \sigma_n^2 I)$. The integration of (4.9) yields the following marginalized log likelihood (Rasmussen and Williams, 2006) of the training data:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) = &-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top (K + \sigma_n^2 I)^{-1}(\mathbf{y} - \mathbf{X}\beta) \\ &-\frac{1}{2}\log|K + \sigma_n^2 I| - \frac{N}{2}\log 2\pi\end{aligned} \tag{4.10}$$

where N is the number of training examples.

Regularization Regularization techniques, such as those employed in Lasso (Tibshirani, 1996a) or Ridge regression (Hoerl and Kennard, 1988), are often used to improve model performance, especially when the data contains several irrelevant features. The L_2 penalty, imposed by Ridge regression, ensures shrinkage of regression coefficients to avoid over-fitting. On the other hand, the L_1 penalty imposed by Lasso forces the coefficients to be sparse, thereby providing feature selection. However, neither of the two regularization methods have been found to universally dominate the other (Tibshirani, 1996a). For instance, in the presence of groups of correlated features, Lasso tends to select only one feature within each group, which leads to poor interpretability of the estimated coefficients. Elastic net regularization (Zou and Hastie, 2005) is a weighted addition of L_1 and L_2 penalties and combines the strengths of both Lasso and Ridge regression. It is known to select a greater number of influential features than Lasso, and has lower false positive rate than ridge regression.

We use elastic net regularization to penalize complexity of the solution (i.e. perform feature selection) and to avoid over-fitting on the limited training data set. The elastic net penalty is computed as:

$$\alpha\lambda\|\beta\|_2^2 + (1-\alpha)\lambda\|\beta\|_1 \tag{4.11}$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ (L_1 penalty, Lasso) and $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ (L_2 penalty, Ridge). The value for α was varied from 0.1 to 1, and chosen by empirically verifying the results; whichever α gave the best results was chosen. The value for λ was set to 0.5 attributing the same weight for both penalties.

Our empirical results show that elastic net regularization results in better prediction accuracy, when compared to ordinary least squares ($\lambda = 0$), Ridge ($\alpha = 1$), and Lasso regression ($\alpha = 0$).

To regularize the coefficients in β, we apply Elastic net regularization on the marginalized log likelihood to obtain the following objective function:

$$J(\beta, \ell, \ell_s, \sigma_n^2, \sigma_f^2) = \log p(\mathbf{y}|\mathbf{X}) - (\alpha\lambda\|\beta\|_2^2 + (1-\alpha)\lambda\|\beta\|_1) \tag{4.12}$$

The function J is maximized to estimate the hyper-parameters using conjugate gradient descent (Rasmussen and Nickisch, 2010).

4.3.2 Combining Source-specific Models

To predict poverty for a commune, we use two independently trained models specified in (4.3), corresponding to the two data sources, viz., CDRs and environmental data. Each model produces a posterior Gaussian distribution, denoted by $y_{ic} \sim \mathcal{N}(\bar{y}_{ic}, \sigma_{ic}^2)$ and $y_{ie} \sim \mathcal{N}(\bar{y}_{ie}, \sigma_{ie}^2)$ for the CDR and environmental data, respectively. The combined poverty estimate, y_i, is assumed to be a mixture distribution consisting of two Gaussians, defined above, and the mixing weights defined as:

$$w_{ic} = \frac{\frac{1}{\sigma_{ic}^2}}{\frac{1}{\sigma_{ic}^2} + \frac{1}{\sigma_{ie}^2}}, \qquad w_{ie} = \frac{\frac{1}{\sigma_{ie}^2}}{\frac{1}{\sigma_{ic}^2} + \frac{1}{\sigma_{ie}^2}} \tag{4.13}$$

The weights assign greater importance to the source that provides a smaller predictive variance, signifying higher confidence in the prediction for the particular commune. The mean and the variance for the combined poverty estimate are (see Appendix 'Estimating Moments of a Mixture Distribution'):

$$\begin{aligned} \mathbb{E}[y_i] &= w_{ic}\bar{y}_{ic} + w_{ie}\bar{y}_{ie} \\ \text{var}[y_i] &= w_{ic}\sigma_{ic}^2 + w_{ie}\sigma_{ie}^2 + w_{ic}w_{ie}(\bar{y}_{ic} - \bar{y}_{ie})^2 \end{aligned} \tag{4.14}$$

It is worth mentioning that the variance of such mixture is always higher than one of the two Gaussians taken separately. It is higher than *both* Gaussians if the variances of each distribution (σ_{ic}^2 and σ_{ie}^2) are very close. The difference between the mean ($\bar{y}_{ic} - \bar{y}_{ie}$) of the two Gaussians further increase the variance

of the mixture and it can, therefore, be interpreted as an indicator of uncertainty.

This might be counterintuitive as additional information should lead to lower variance. However, one should recall that using our approach, the different data ecosystems need not share any data between them. It allows to preserve privacy of sensitive data (such as CDRs) as only the output predictions and the associated variances are shared. Therefore, the combination of information from the data sources is made at the highest level, i.e. using outputs of the models, losing the potential of using raw data. This is a methodological choice where the priority is put on privacy and security over advanced data fusion techniques.

4.3.3 Model Validation

This section details the steps followed to validate our model, namely creating commune-level poverty statistics from census data, and methodology for spatial cross-validation.

Creating Commune Poverty Statistics from Census The 10% random sample of the 2013 RGPHAE census covering all the country, used here, has survey responses for 150,000 households and 1.4 M individuals, pertaining to their socio-economic indicators (literacy, birth and death in the family etc.), habitat (type of house, access to electricity, drinking water etc.). Some survey responses are individualistic (like literacy, profession), while others are associated to the entire household (like type of roof, sanitation, electricity).

The first step is to assign the individuals to their respective households using information from the fields in the census. Second step is to calculate per-household deprivations in the poverty indicators of interest. Global MPI computation (Alkire *et al.*, 2015) requires deprivations along three dimensions (with 10 indicators, namely, health (child mortality, nutrition), education (child school attendance, years of schooling) and standard of living (electricity, sanitation, drinking water, flooring, cooking fuel, assets).

We follow the procedure similar to the widely used Alkire-Foster methodology for computing MPI (Alkire and Foster, 2011a). First, we create a deprivation vector $depvec_{i,d}$ corresponding to each household i in poverty-indicators $d = 1, \ldots, D$. Each vector entry is either 1 if $y_{i,d} \leq z_d$, where $y_{i,d}$ is the achievement of household i in indicator d, and z_d is the cut-off score in indicator d, or 0 otherwise. A value of 0 for $depvec_{i,d}$ implies non-deprivation of the household in that particular indicator.
For the values of cut-off scores for different indicators, please see Table A1 in Appendix. We aggregate all households that are deprived in a particular indicator, for each commune, and divide by the total number of households in that commune. This score gives the proportion of households deprived in a particular indicator within a commune.

Since MPI is a multiplicative combination of H (headcount of poverty) and A (intensity of poverty), i.e., $MPI = H \times A$, we first calculate H and A. For H, we introduce a weight, w_d, for each indicator d. For each household we compute a weighted deprivation score, $c_i = \sum_{d=1}^{D} w_d depvec_{i,d}$. The weights w_d are assigned as follows. The education and health related indicators are given a weight of $\frac{1}{6}$, while each of the 6 standard of living indicator are given a weight of $\frac{1}{18}$. Thus each dimension has a weight of $\frac{1}{3}$.

H_j, which is the relative headcount of poor households in commune j, is calculated as:

$$H_j = \frac{1}{N_j} \sum_{i=1}^{N_j} I(c_i > \theta) \tag{4.15}$$

where θ is a cutoff, whose higher values mean higher cutoff for household achievement, and $I(c_i > \theta)$ is the indicator function. N_j is equal to the total number of households in the j^{th} commune.

To calculate A, we count only the poor households, and their deprivations, as follows:

$$A_j = \frac{1}{\sum_{i=1}^{N_j} I(c_i > \theta)} \sum_{i=1}^{N_j} I(c_i > \theta) * c_i \tag{4.16}$$

The value of threshold θ is taken as 0.3. We varied θ from 0.2-0.75, and the H, A values obtained in each run was correlated with region level H, A available from University of Oxford's MPI calculation (OPHI MPI data in Table 4.1). The results were stable, and peaked at 0.3, which is also the threshold value taken by OPHI for their calculations. The Spearman's rank correlation coefficient between each indicator (MPI, H, A) and dimensions of poverty is given in Figure 4.3.

Spatial Cross Validation A standard cross-validation (CV) is often performed to ensure that the model generalizes to out-of-sample data. We performed a standard 10-fold CV, where the data is randomly split into 10 folds. Each time, 9 folds are used for training, and 1 fold is used for evaluation, meaning we randomly assign 90% of communes to the training set, and evaluate on the remaining 10% of communes. This procedure is repeated 250 times to provide a robust assessment of the variability of model parameters and prediction statistics.

Though training and evaluation data are selected randomly, the above described method of validation may prove to be insufficient, as the poverty deprivations tend to be spatially correlated. Thus a model may appear to perform well when evaluated this way, even though it may have poor extrapolation power in the spatial sense. The above results are provided for comparison.

To measure the extrapolation capacity of the model to spatial areas that were not represented in the training data, spatial cross-validation techniques, where

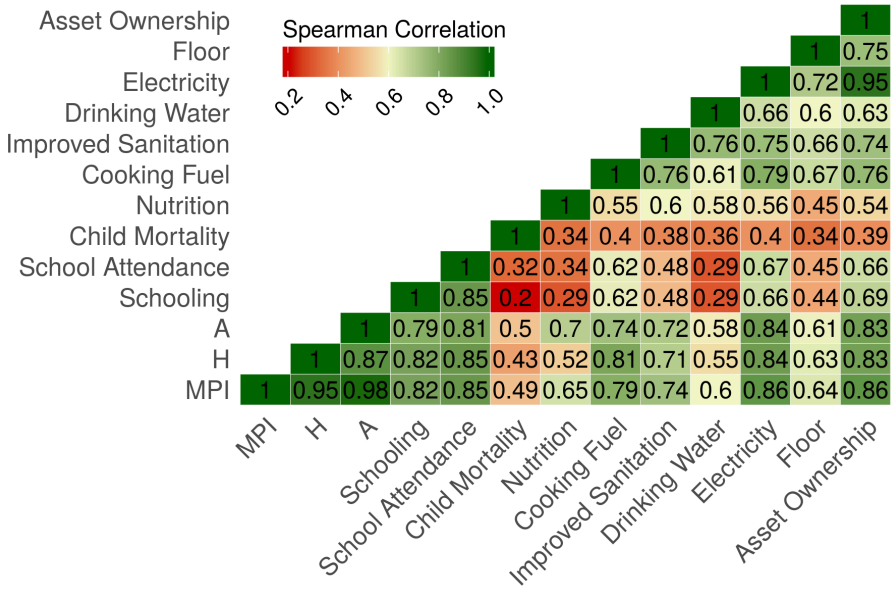


Figure 4.3: Spearman's rank correlation matrix between individual deprivations, H (Headcount of poverty), A (Intensity of poverty) and MPI at commune level

training and evaluation sets are sampled from geographically distinct regions, are more robust (Deville *et al.*, 2014a; Bahn and McGill, 2013). The following spatial cross-validation strategy was adopted: for each cross-validation run, we first randomly sample a region r from the set of 14 regions, and then randomly sample a commune c belonging to r. All communes that lie within distance d of the commune c are included in the training data set. The remaining communes are included in the evaluation data set.

This strategy ensures that communes from all regions of Senegal are represented in the training and evaluation data sets during cross-validation. To ensure that the training data set has enough examples, we force at least 40% of the communes (225) are included in the training data set. To achieve this, d is initially set to 100 km, and is increased by 50 km until the size of the training data set meets the threshold.
Spatial cross-validation is repeated 250 times. We report the mean predictive performance (using Pearson's and Spearman's correlation, and RMSE values) on the evaluation data set, along with the standard deviation across multiple runs.

All code to replicate the results can be obtained online (https://github.com/damienjacques/).

4.4 Results

4.4.1 Predicted MPI Poverty Values

The predicted map of MPI for micro-regions, i.e., 552 communes of Senegal, is depicted in Figure 4.4 (left) and Figure 4.5 (top). Compared to the current poverty map in Figure 4.1, our map highlights heterogeneity in the existence of poverty within each macro-region. A quantitative validation of the predictions is provided against commune level poverty values estimated from census data (see Figure 4.5 - bottom) using cross-validation procedures. Using standard CV, the model gives a Pearson's correlation of 0.94, with a p-value of <0.0001. These results are provided for comparison as they overestimate the performance of the model because the spatial auto-correlation is not taken into account. Using spatial CV, the predictions in Figure 4.5 (top) have a Pearson's correlation of 0.91 and rank correlation of 0.87, with p-values less than 10^{-20} for both tests, indicating strong significance. This emphasizes the efficacy of our model in predicting poverty values accurately at finest spatial granularity, using multi-source data.

As a comparative study of how our model performs using multi-source, and single-source data, we experimented with three datasets – Multi-source, CDR and Environment – to predict headcount of poverty (H), intensity of poverty (A), and poverty index (MPI) at commune-level (see Table 4.6). We report highly accurate results for all three targets (H, A, and MPI). Rank correlations are preserved, as we report Spearman's correlation of 0.85 for both H and A. The values of Pearson's r correlation are much higher than rank correlation, across all prediction tasks, indicating the linear correspondence of the poverty values with the predicted ones. We report significantly low p-values ($< 10^{-34}$) for spatial cross-validation compared to standard cross-validation, signifying more stable performance. For detailed results see Table A2 in Appendix. Table 4.6 shows that combining multiple data sources (CDRs and environmental data) results in a consistent improvement of accuracy over using the individual data sources. The improvement is more pronounced in detailed results for all the indicators of poverty, and given in Table A2 in Appendix. Interestingly, our approach performs also better than using concatenated data (CDR and environmental data) in a single model (Table A2 in Appendix). The implication of this result is that sensitive data (CDRs) can remain stored behind the firewall of the data provider without reducing the prediction accuracy. In our approach, only the privacy-compliant model outputs (aggregated) need to be share to provide the final predictions.

The left panel of Figure 4.4 plots the relationship between MPI values predicted by our model, and those estimated from census (by multiplying by ten the population of each commune provided by the 10% random sample). We observe a linear relationship, in general, for MPI, with lower values for urban areas (shown in red) and higher values for rural areas (shown in blue). Predominantly urban communes of Dakar, and few urban centers are underestimated

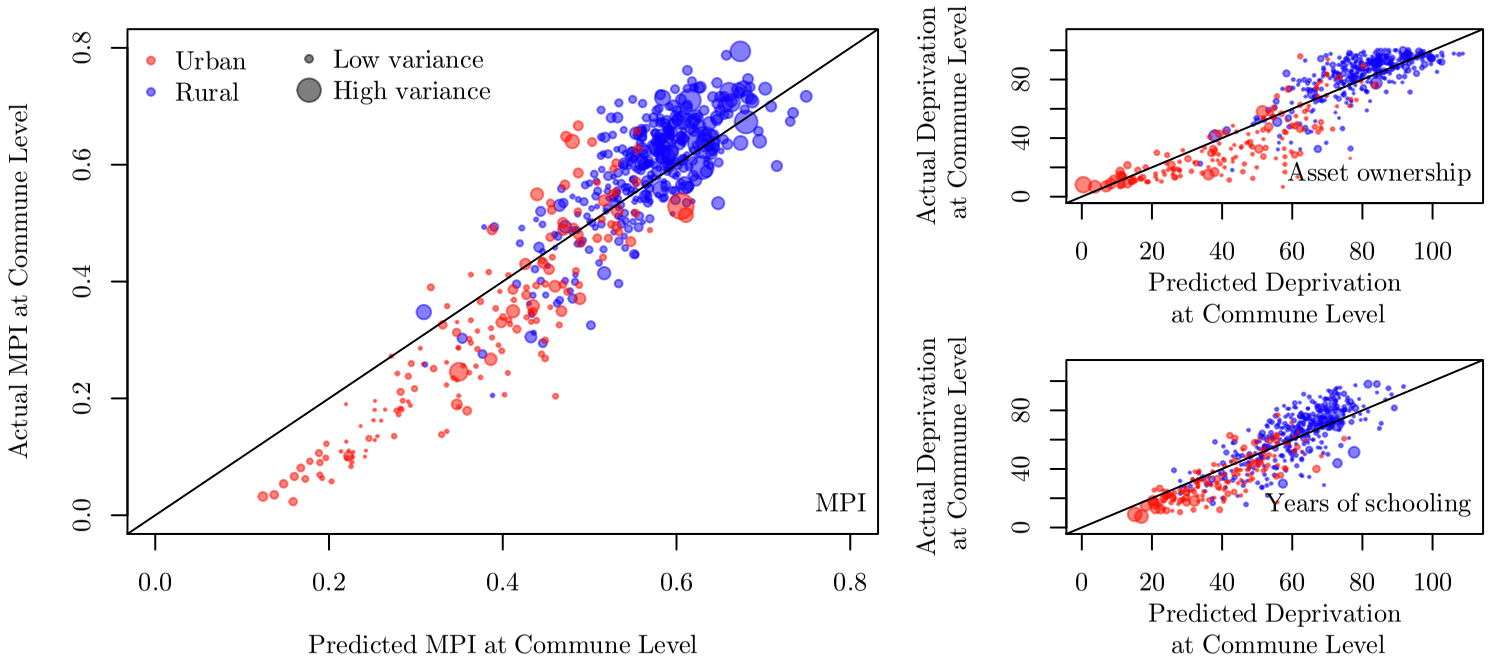


Figure 4.4: The left panel denotes the comparison of actual and predicted MPI values for all communes and urban areas of Senegal. The rural and urban areas are differentiated using blue and red colors respectively. The size of the circle denotes the variance of MPI prediction for that commune. The top right panel shows how the actual and predicted values compare for asset ownership, while the one on the bottom shows the comparison for years of schooling. A bias exists (might be due to omitted variable) that can simply be corrected using a linear regression (OLS) between actual and predicted values (Figure A1 in Appendix). The only impact would be a decrease in RMSE values (not reported here as all conclusions remain unchanged).

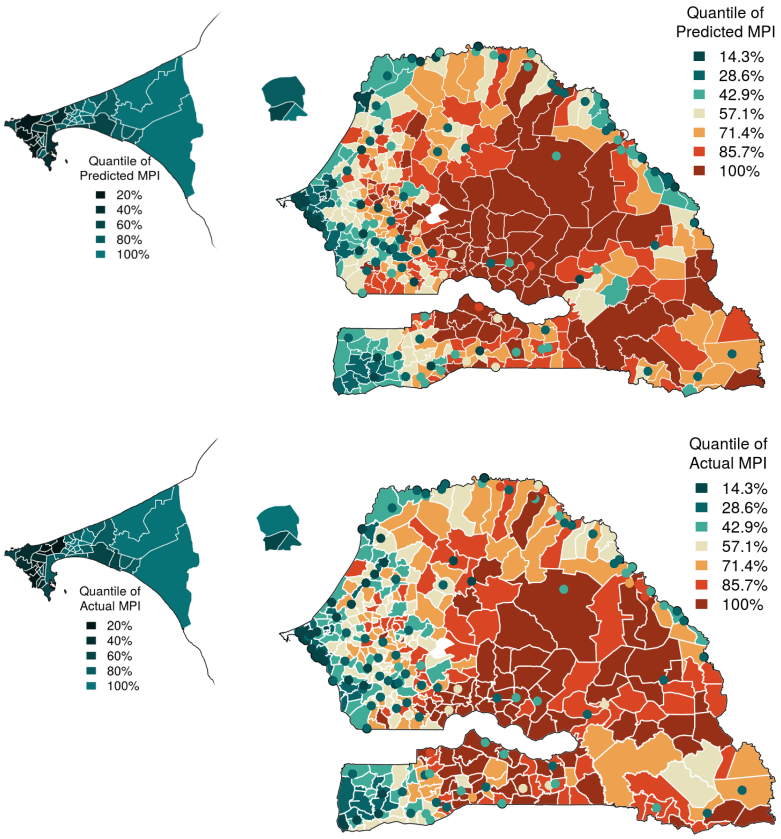


Figure 4.5: Quantiles of predicted (top) and actual (bottom) MPI at commune level. The urban centers are depicted by small circles on the map. The communes in Dakar and Thiès regions are shown enlarged.

for poverty (i.e. they are predicted richer than they are). Likewise there are very few rural communes, where poverty is overestimated. We also observe that for communes with lower population densities the predicted variance is comparatively high, than for communes with higher densities, signifying that lesser number of data-points in the vicinity of a given commune contribute to its higher variance (Figure 4.6).

4.4.2 Predicted Values for the Dimensions of Poverty

Global Multidimensional poverty index (MPI) consists of 10 individual deprivation indicators grouped along three dimensions: (i) education (indicators - Years of Schooling and School Attendance), (ii) health (indicators - Child Mortality, Nutrition), and (iii) standard of living (indicators - Cooking Fuel, Sanitation, Access to Drinking Water, Electricity, Floor and Asset Ownership). Each individual deprivation indicator is taken as the target of our model, and

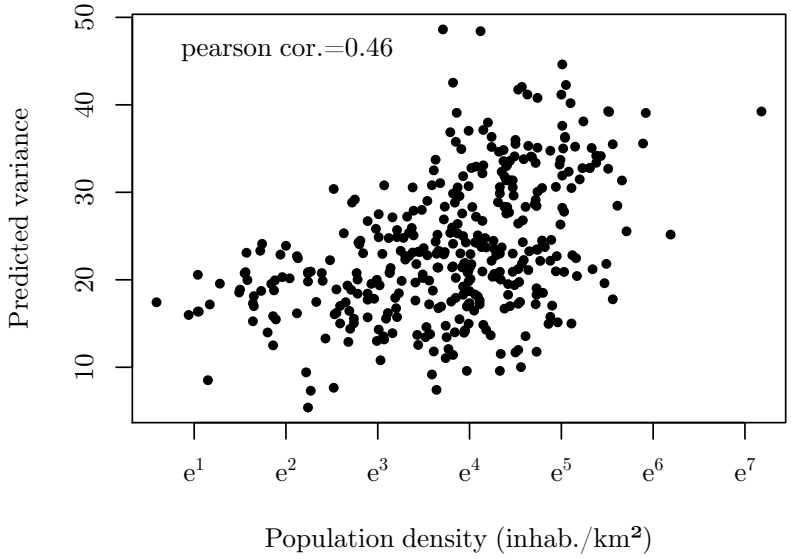


Figure 4.6: Relationship between precision of estimates of poverty and the population density of each commune.

the averaged spatially cross-validated results, along the three dimensions, are reported in Table 4.6. Detailed results for each of the 10 indicators are given in Figure A5 and Table A2 in Appendix.

Referring to Table A2 in Appendix, we note that the accuracy of the model is high for some deprivations, and good for most deprivations. All deprivations are better predicted using CDR data, probably, because it characterizes the individual behavior while environmental data depict conditions that might have an influence on poverty (Table 4.5). Figure 4.4 (top right panel) compares our predictions for asset ownership with those estimated from the census. Rural communes depicted (by blue) are clustered closely towards high deprivation. The urban areas have, generally, lower deprivation than rural areas, though it is spread out.

Indicators related to education - years of schooling, and school attendance, are predicted well, because usage of short message service (SMS) is indicative of literacy (Sundsøy, 2016). The environmental data also performs well, because it captures the distance to schools, main roads and urban centers, which facilitate access to educational attainment. Figure 4.4 (bottom right panel) shows that all areas of Senegal are deprived in education, as the rural (in blue) and urban (in red) points are spread evenly on the plot. Though, rural areas tend to dominate at the very high deprivation index, while very low deprivation areas are urban.

Table 4.6: Spatially-cross validated results of the predictions of MPI, Headcount of poverty (H), and Intensity of poverty (A), along with the individual indicators for poverty given by our model using disparate datasets. The results are compared when single source data is available. corr. – Pearson's r correlation, rank corr. – Spearman's rank correlation, and RMSE – Root Mean Square Error. For both types of correlations, all p-values were less than 10^{-20}. A standard deviation associated with the multiple runs for each measurement is reported within simple brackets.

Poverty Indicators	Multi-source Data		
and dimensions	corr.	rank corr.	RMSE
MPI	0.91 (0.06)	0.88 (0.06)	0.08 (0.01)
H	0.91 (0.07)	0.85 (0.08)	10.79 (3.96)
A	0.86 (0.05)	0.85 (0.07)	4.71 (0.96)
Education	0.86 (0.05)	0.84 (0.05)	11.84 (1.88)
Health	0.49 (0.15)	0.50 (0.16)	12.76 (2.12)
Standard of Living	0.83 (0.11)	0.75 (0.13)	14.82 (3.92)
Poverty Indicators	CDR		
and dimensions	corr.	rank corr.	RMSE
MPI	0.89 (0.07)	0.86 (0.07)	0.08 (0.01)
H	0.90 (0.08)	0.84 (0.08)	10.76 (2.60)
A	0.83 (0.07)	0.82 (0.08)	4.98 (1.14)
Education	0.82 (0.05)	0.81 (0.07)	13.08 (1.68)
Health	0.50 (0.12)	0.52 (0.12)	12.91 (1.92)
Standard of Living	0.81 (0.11)	0.74 (0.11)	15.24 (3.45)
Poverty Indicators	Environment		
and dimensions	corr.	rank corr.	RMSE
MPI	0.84 (0.09)	0.80 (0.10)	0.10 (0.02)
H	0.83 (0.11)	0.75 (0.11)	13.65 (4.86)
A	0.81 (0.07)	0.79 (0.08)	5.36 (0.75)
Education	0.76 (0.07)	0.74 (0.07)	14.98 (3.03)
Health	0.36 (0.23)	0.35 (0.23)	13.91 (2.32)
Standard of Living	0.73 (0.18)	0.64 (0.20)	17.88 (4.50)

The model performs poorly for the indicators within the health dimension, i.e., child mortality and nutrition. This is attributed to the fact that our data are not representative of the children population, and, thus, the features extracted from CDR data do not capture this deprivation. Similar inference can be drawn for poorer correlations for nutrition. Moreover, the validation of deprivation values computed from the census for nutrition indicators are based on two hunger related questions, as detailed nutritional information is not available to us (See Table A1 in Appendix for details).

4.4.3 Dimensions of Poverty - Interpretation of Weights

Figures A6 and A7 in Appendix display the features deemed important by our model for the environment and CDR data respectively. The important features are those for which the corresponding entries in the coefficient vector, β, are high in magnitude.We ignore child mortality and nutrition, as our model does not perform very accurately for these two indicators. The following interpretations are given for information purposes. These are, by no means, indicators of causality.

Referring to Figure A6 in Appendix, nighttime lights appear to be the most important feature regardless of the predicted dimensions, conforming to current research (Jean *et al.*, 2016; Njuguna and McSharry, 2017). It was expected as nighttime lights shows a strong correlation with MPI (Spearman correlation of -0.66). Urban areas and road density, two other important indicators of economic activity are relevant, but to a lesser extent. Even though the coefficient values of each dimension are not directly comparable since each dimension was taken as a separate target, it is interesting to note that the weights of nighttime lights intensity for electricity and asset ownership deprivation are the highest. This result confirms previous findings (Min *et al.*, 2013) that access to electricity is correlated with nighttime lights (Spearman correlation of -0.67).

Several features related to the presence of water in the commune (water bodies, water, mudflat soil, hydro-morphic soil and elevation) are positively correlated with water deprivation although the opposite is observed for the other dimensions. One interpretation is that natural non potable water would be used for drinking in these areas. On the other hand, access to water is interesting for irrigated agriculture, watering livestock or fishing that can increase income and life quality which explains the negative relationship for the other dimensions. Interestingly, the distance to water tower is not quite correlated with this deprivation. Alternatively the proximity to water forage would have probably been a more interesting feature.

The food security (access) features (like distance to main roads and urban centers) are also prominent, stressing their importance for development. Millet price has a mixed behavior. Depending on the dimensions, its coefficient is sometimes positive, sometimes negative without an evident explanation could have been found.

The effect of temperature is clear. The higher the maximum temperature and the range, the higher the poverty. Temperature plays a role in crop growth but it also impacts the environment quality of the people who lives in warm (and cold during the night) areas. The effect of precipitation is less obvious. The amount and the period of rainfall affects the availability of water, which is the main limiting factor in Sahel for crop and forage production. The precipitation seasonality, described by the period of time during which the water is available and the precipitation of the warmest quarter (critical period), are logically negatively correlated with poverty. However, the annual precipitation and the

precipitation of the wettest month and quarter have a positive coefficient (except for education deprivation). In other words, the more it rains in an area, the more poor it is. The intuition would have been that it was the opposite. But looking more closely, it appears that several features related to agriculture (groundnut production, cassava production, rain-fed croplands) show the same patterns. We interpret that these features define a suitable environment for agricultural areas which, itself, is linked to the presence of rural community tending more to poverty than urban population.

Similar analysis for the CDR features reveal several interesting insights regarding the relationship between poverty and the individual characteristics captured in CDR features. While we considered CDR features for each month individually, for the ease of visualization (See Figure A7 in Appendix), we average the monthly values of the weights associated with each feature.

Here we discuss the CDR features that were selected by the model as the strongest predictors for the various targets. These features are listed in Table A3 in Appendix. One of the strongest negative predictors for most of the targets is the number of active days (for call and text), which characterizes that individuals in wealthier communes have monetary resources to recharge phone, and make/receive calls. While the ratio of calls vs. text, shows the preference to calls, and emerges as an important factor to predict education based deprivations. The feature, inter-event time call, measures the irregularity in the responding to calls/text and emerges as a positive predictor for deprivations. Features that indicate diversity in communication, such as entropy of contacts and interactions per contact (call and text), report a negative relationship to poverty. These results confirm previous findings (Eagle *et al.*, 2010a; Soto *et al.*, 2011; Blumenstock *et al.*, 2015a) that diversity of an individuals' relationships is positively correlated with their economic well-being. However, for features such as percent pareto interactions and balance of contacts, which are proportional to an individual's diversity in communication, we report a positive relationship with poverty. This counter-intuitive relationship needs to be further studied in the context of telecommunication patterns in Senegal.

We observe a negative relationship between the *activeness* of an individual in his/her mobile interactions and poverty. For instance, the delay in responding to text has a positive relationship to poverty. Interestingly, the feature percent initiated interactions (calls), has, again, a positive relationship to poverty, signifying that in Senegal, individuals living in more deprived communes are more likely to initiate calls (for request of resources, etc.) than those living in less deprived communes. The mobility patterns of individuals, captured using spatial features such as number of frequent antennas, entropy of antennas, and total number of antennas used by an individual, indicate a negative relationship to poverty. Thus individuals living in more deprived communes tend to move fewer antennas than those living in less deprived communes. This observation should be viewed cautiously because of sparse antenna density in rural communes.

4.5 Discussion

The technological advances over the past decade has led to building of communication devices (like phones) and sensors (like satellite, weather and ground sensors) that produce and store a myriad of data. In this work, we show how these novel sources of data, which are characterized by their volume, variety and associated uncertainty, can be used to generate accurate poverty maps.

We outline several challenges that lie in establishing relationships between alternative data sources (that are not collected to directly measure socio-economic deprivations) and poverty. First challenge occurs due to the varying spatial granularity at which the different data sets are available; this requires for an aggregation mechanism to link them. CDR data is available for each subscriber while environmental data have mixed spatial resolution, from very accurate vector data to low resolution satellite imagery (1 km). On the other hand, census data is available for individuals or households, depending on the response variable. However, given that the individual information is anonymized for both CDRs and census data, there is no obvious way to link the records across these two data sets. In this work, we localize the individuals and/or households to their respective communes, or urban centers, by using their census information (details in Methods). This lets us calculate the commune level deprivations. For CDRs, the individuals are localized to their home antennas based on their most frequent night location. The CDR and environmental data are aggregated to commune levels. Though we have taken a commune as the level of aggregation, the framework allows for the same analysis at even finer spatial resolutions.

A key concern associated with using CDR data for population level analyses is the selection bias arising from mobile phone ownership. In Senegal, however, there were 92.93 mobile phone subscriptions per 100 inhabitants in 2013, which implies that most of the population owns cell phones (ITU World Telecommunication, 2016a). Second is the bias arising when utilizing data from only one provider. However, the provider of the data used here, Sonatel, has nearly 62% of the cell phone market in 2013 (Autorite de Régulation des Telecommunications et des Postes, 2013). Third concern is that some demographic sub-groups like children and ultra poor, are left out by the analysis while only using CDR data. Also results may be biased towards urban regions, than rural, because of factors like lack of electricity in rural areas. However, it is worth mentioning that

Here, we used two distinct types of environment data. The first type includes static natural/physical environment variables (like elevation, soil types, etc.) or long term dynamic phenomenon (like climate). Second type includes human induced aspects, like urban areas, roads, access to facilities, etc. Though natural environment acts as a constraint in designing poverty eradication plans, effective policies and sustainable approaches should be made an integral part for policy planning. Environmental features derived from satellite images (nighttime lights, NDVI, etc.) have the potential to be computed in near real-time to monitor the impact of shocks such as natural hazards, armed conflicts, or crop

Data Source	Model	Results (Pearson's R)
Nighttime lights	Linear Regression	0.39
Njuguna and McSharry (2017)	Linear Regression	0.84
Our Model	Gaussian Process Regression	0.91

Table 4.7: Comparative table showing how our model performs compared to only only nightlights, and a previous work (used as a baseline) using only 4 features, namely call volume and mobile ownership per capita, nightlights and population density.

pests that can rapidly cause serious deprivations. However, for reliability, these variables need to be aggregated for a longer period, typically at an annual level for night-time lights and for the growing season for NDVI. OpenStreetMap (OSM) data, which is used to map facilities and roads, is crowd-sourced and therefore, have the (theoretical) potential to be updated in near real time. Though this capability could be limited in African countries. Due to above constraints, one year is probably the relevant period for consistent monitoring of poverty with our method (compared to 3-5 years for a detailed and costly census).

Another challenge is the ease of availability of data. Environment data sets are available to researchers for free, and typically have no privacy constraints, especially at the resolution at which it is analyzed here. CDR data is collected by commercial telecommunication entities, and might suffer from lack of accessibility to researchers due to sharing constraints between different organizations. However, our methodology requires no raw data to be shared between different data owning entities; only the output predictions from each individual model and the associated uncertainties are combined at the final step.

An important consideration is the number of features extracted from the data. Recent work (Njuguna and McSharry, 2017) has used 4 features, namely call volume and mobile ownership per capita, night-lights and population density to estimate MPI of sectors in Rwanda, using a linear regression model. As a baseline for our model, we used the same features and model to predict MPI values at commune level in Senegal. A spatially cross-validated Pearson's correlation of 0.84 was achieved with significant p-value (<0.0001) (Table 4.7). Although less features provide computational tractability of analysis, they offer no insight into other features that could be useful in understanding poverty. Also, linear models are limited in their ability by the linearity assumption and sensitivity to outliers.

An important advantage of our GPR model is that each predicted poverty value is associated with an uncertainty (generated by the model). This highlights the strength of confidence in the predictions, and can be used as a guidance by the policy makers. Comparing these source-specific uncertainties can reveal which data holds better signal for a specific prediction (see Figure 4.7). We note that for predicting A, the predictions of CDRs and environment data are comparable for most of the communes. For predicting headcount of poverty (H), CDRs perform with lower uncertainties than environment data. These variations

maybe attributed to multiple reasons, including resolution and concurrency of data, demographics and mobile penetration of the cellular provider, and spatial heterogeneity of poverty deprivations.

Though, we have discussed the methodology for predictions at commune level, our predictions of MPI and associated dimensions can be successfully aggregated to coarser administrative units, if needed for policy planning. Since we use global MPI as the poverty index, its limitations, as noted by global MPI researchers (Alkire *et al.*, 2014), are applicable to our study as well. In particular, global MPI does not include characteristics such as parents education, social norms and beliefs, empowerment, etc. Additionally, it will be interesting to see how well can this methodology be used to predict other indicators of deprivation and inequality, like GINI index, at micro-regional level. Apart from being useful in producing interim statistics in between long cycles of census and surveys, such methodology can be also extended to places of conflict or remote areas that are difficult to reach by census takers.

As described in the results, the interpretation of the model coefficients provide some insights on the dimensions of MPI. However, due to the number of variables, this interpretation is still complex and not necessarily straightforward for policy intervention .Conversely, the MPI dimensions predicted by the model, are well known factors for which policy planning is feasible (Alkire and Santos, 2014). As an illustration, Figure A5 in Appendix shows the highest predicted deprivation for each commune within each dimension. If the objective is to decrease the MPI, these areas should be prioritized.

Lastly, though GPR model uncertainty is impacted by the bias and inaccuracy of each data sources (quality of soil type map, interpolation of climatic data, missing facilities, mobile operator's market share), a higher resolution and accuracy of the input data should benefit the modeling relevance and quality.

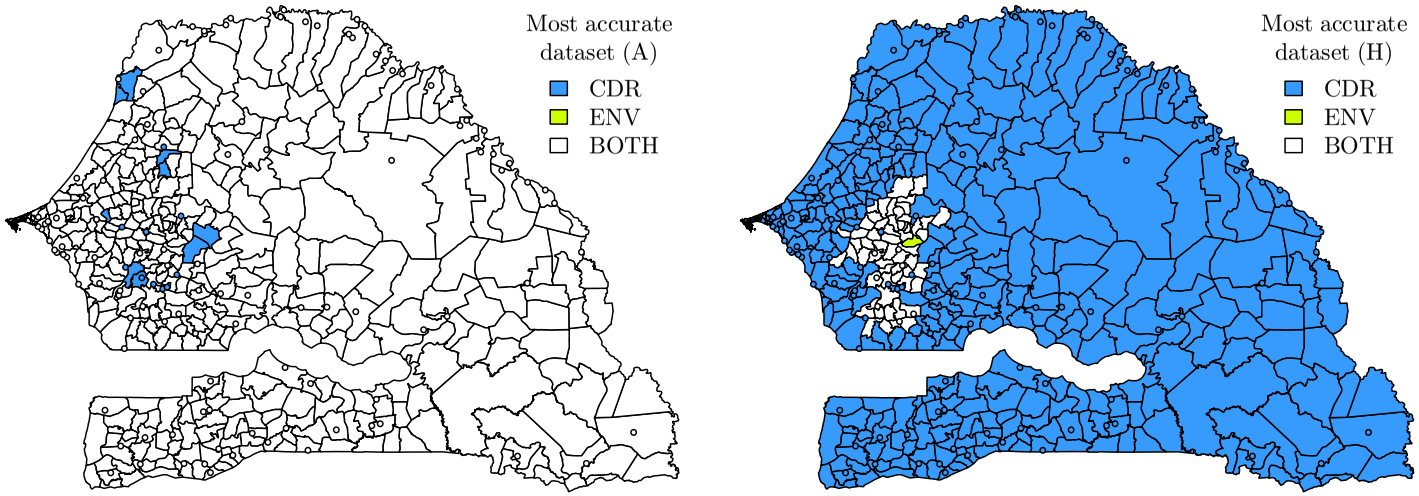


Figure 4.7: The uncertainty associated with each dataset evidenced by the most accurate one (denoted as CDR and ENV) for the prediction of the Headcount of poverty (A) (shown left), and the average Intensity of Poverty (H) (shown right).

4.6 Conclusions

Spatially finest poverty maps are essential for improved diagnosis and policy planning, especially keeping in view the Sustainable Development Goals. *Big Data* sources like CDRs and EO data have shown promise in providing inter-censal statistics at fine granularity. This Chapter outlined a computational framework to efficiently combine disparate data sources, like environmental data and mobile data, to provide accurate predictions of poverty and its individual dimensions for finest spatial microregions in Senegal.

Our approach was to use Gaussian Process Regressions (a Bayesian learning technique) with each data source and then combine the outputs based on the uncertainties of each prediction; giving more prominence to less uncertain prediction. This approach allows to obtain more accurate results than the models taken separately or than running a single model on concatenated data sources. It is worth mentioning that source-specific models already gave accurate prediction, confirming the potential of CDRs and EO data for poverty mapping in Africa.

An important advantage of our methodology is that sensitive data (CDRs) can remain safely stored behind the firewall of the data providers without compromising the prediction accuracy. Only the privacy-compliant model outputs (i.e. aggregated data) need to be share to combine the source-specific models. This offers a practical solution to the data access and protection issue specific to personal data collected and processed by private companies.

Conclusions

Main Findings

Achieving the SDGs by 2030 require relevant, accurate and timely data to track progress and identify the root causes of deprivations. Given the current data ecosystem, it requires to bring about a data revolution in sustainable development. This means making better use of existing traditional data but also exploring innovative approaches based on alternative data sources. In this thesis, we explored how the unique features of two types of Big Data sources – mobile phone data and EO data – can supplement official statistics and contribute to bridge the knowledge gap in Africa. We presented three contrasted applications in Senegal, related to the two first SDGs: no poverty and zero hunger by 2030. First, we defined the accuracy requirements for estimators of crop area and crop yield needed to provide early estimations of production and anticipate potential food crises. Second, we demonstrated the impact of the traders' social capital on market functioning and millet prices, an important indicator of food access. And third, we developed a computational framework to efficiently combine disparate data sources, such as mobile phone and EO data, to provide more accurate predictions of poverty and its determinants for fine spatial micro-regions.

Table 1 summarizes the main findings along with the data used for each Chapter of the thesis. In Chapter 1, a methodological framework was proposed to define the accuracy requirements of early warning estimators of cropland area, crop area and crop yield (based on EO data) using historical time series of official agricultural statistics as reference (**accuracy**, **timeliness**). We showed that the inter-annual variability of crop yield was the main factor limiting the accuracy of pre-harvest production estimates. Chapter 2 introduced CDRs, a promising data source to tackle development challenges with a potential of providing near real time information (**relevance**, **timeliness**). This Chapter also highlighted some of the bias sources (selection, temporal ans spatial) inherent to such data (**accuracy**). Finally, a meaningful discussion on the access and privacy issues was undertaken (**access**, **protection**). In Chapter 3, CDRs were proved an effective way to approximate social capital in food markets and model its impact on millet prices (**relevance**). Results demonstrated that accounting for the social capital in the transaction costs could explained a significant part of the price variance with a year-specific effect. Finally, in Chapter 4, a computational framework to combine disparate data was introduced providing

highly accurate, intercensal fine-grained poverty map (**accuracy**, **timeliness**). A clear advantage of this methodology was that the different data ecosystems need not to share any data between them. This means that sensitive data, such as mobile phone data, could be leveraged without jeopardizing individual privacy (**protection**) enhancing data access for research (**access**). Furthermore, this last work confirmed the high potential of mobile phone metadata to assess the socio-economics status of population in developing countries (**relevance**).

What innovative use of CDRs can deliver relevant information to support the achievement of main development challenges such as poverty and food security?

Estimate social capital in food markets.

In Chapter 3, we demonstrated the impact of social capital, approximated by mobile phone calls between market areas, on millet prices in Senegal. To the best of our knowledge, this is the first time that CDRs were used to approximate social capital in agricultural markets. This work contributed to a better comprehension of the root causes of price dispersion in Senegalese millet markets. Understanding the role that social capital plays in market exchanges is essential for policy design. Our results suggest that finding approaches to facilitate search and fostering trust could potentially improve trade exchange (e.g., legal institution, public market information system, traders' associations, etc.). CDRs are particularly relevant in this case as they are a good proxy of social network of population. More insights are expected from research focusing specifically on traders' network.

Characterize the socio-demographic profiles of an entire population and map poverty at high resolution.

In Chapter 4, we showed how mobile phone activity combined with environmental indicators approximated by EO data, can be used to estimate multidimensional poverty in Senegal at fine level of granularity. While similar results have already been shown before (see Chapter 2), this works has pioneered the use of CDRs *combined* with other data sources such as EO data to map poverty at fine granularity. A similar approach has been developed in parallel by Steele *et al.* (2017) in Bangladesh confirming the promising avenue of multi-source techniques to predict poverty. Furthermore, it confirms the strong value of CDRs to get intercensal and near real-time estimation of poverty indicators. Still, it is worth noticing that this method will always be impacted by the selection bias inherent to CDRs data. In particular, the most deprived are potentially left out from such analysis. However, a multi-source approach as proposed in Chapter 4, should reduce the impact of this effect.

Study social networks, mobility and population dynamic.

In Chapter 2, we introduced the unique features of CDR (geo-spatial, dynamic, directed weighted Network) that make it a tremendous source of information. They are particularly adapted to research projects in the field of social network, mobility and socio-demographic studies. The network dimension allows to identify social community by looking at nodes interacting more with each other. The temporal features can be used to inform on the resilience and evolution of social ties after a shock or change in the community structure. Mobility analysis can be carried out at the individual level by combining both spatial and temporal dimension. Finally, land use and population dynamic can be derived from temporal profile of BTS activity.

With the emergence of initiatives such as the Data For Development (D4D) challenge, the number of projects making use of CDRs to tackle development issues have exploded. In Chapter 2, we gave five contrasted examples in five different countries of such applications. Probably the most promising avenue is in epidemiological studies of human infectious diseases and the assessment of socio-demographic profile to map poverty (see Chapter 4). However, less obvious applications, such as assessing electricity demand or optimizing traffic routes, are also encouraging. Chapter 3 offered a good illustration of an innovative use of CDRs, never shown before, for concrete economic questions (approximate social capital in food markets).

What are the accuracy requirements of alternative data sources (CDRs, EO data) to adequately supplement official statistics?

The accuracy requirements of early estimators of cropland area, crop area and crop yield depend on time, space and crop.

Early warning systems for food security play an essential role in preventing food crisis. The most popular approach to get early estimation of production is to estimate cropland area, crop area and crop yield using EO data. These early estimators should be accurate enough to provide more accurate estimations of production than using the average and the trend of historical agricultural statistics. In Chapter 1, we developed a methodological framework to define the accuracy requirements of such estimators needed to support early warning systems for food security. Results showed that requirements depend on time, crop and space. In particular, the inter-annual variability of crop yield was the main factor limiting the accuracy of pre-harvest production estimates in Senegal. Furthermore, estimators of cropland area were useful to improve early production prediction of the main crops in Senegal stressing the value of the cropland mapping for food security.

CDRs are not statistically representative of the population as a whole.

While massive, CDRs still remain a sample of the whole population as they are not designed to be statistically representative. This introduces a selection bias explained by the market share, penetration rate and the different user behavior of each demographic groups. On the other hand, the spatial resolution of CDRs depends on the antenna density which is lower in rural areas where the most deprived generally live. Lastly, the event-based nature of CDRs limits the temporal granularity of mobility analysis. These limitations should be acknowledged when using such data.

CDRs have an added-value for research and knowledge gap filling.

All the limitations of CDRs challenge the very idea of using them as an alternative data sources to compute official statistics (see Chapter 2). However, for research, there are enormous possibilities (as exemplified in Chapter 3 and 4). Furthermore, sometimes these are the only source of information available (e.g., in remote areas or war zones). They can, for instance, be used to fill the gap in-between more conventional data collection (e.g., intercensal poverty maps as presented in Chapter 4). Their near-real time capacity is also an asset to support post-disaster management (e.g. following an earthquake or flood). These observations were made for mobile phone metadata but in essence, several other similar alternative data sources (social networks data, credit cards data, etc.) held by private companies share the same limitations and potential.

An important asset of EO data comes from the availability of long time series allowing to observe change at the Earth Surface (e.g. deforestation, urbanization, desertification...) over extended period of times ($>$ 40 years[1]). On the other hand, research based on large mobile phones datasets ($>$ 1 million of users) covering more than 1 year are rare. The most impressive datasets from the literature were used by Bagrow *et al.* (2011) (4 years, 1.4 M in an African country) and Eagle *et al.* (2009b) (3 years, 10 M in an European country). In comparison, we used in this thesis a dataset of CDRs covering 1 year of communications between 9 millions users. In 2016, the global number of mobile cellular subscriptions has exceeded for the first time the 100%. It means that, from now, extensive dataset (covering most of the world population) will start accumulate. This will open new avenues of research focusing on historical socio-behavior potentially covering the worldwide population. The smartphones will follow and bring about a revolution in personal data sensors.

[1]The beginning of the Landsat program dates back from 1970s.

Can we secure access to sensitive data such as CDRs while protecting individual and business privacy?

Access to CDRs is limited because of the risks to privacy.

In Chapter 2, we discussed the important concerns arising from the extraction of personal information from CDRs (locations, movements, social network...). On the other hand, data access may be limited by private companies to preserve business privacy. Anonymise the data is generally insufficient to prevent re-identification and the best approach remains to share only aggregated data. However, aggregate the data to preserve privacy goes with a loss of information. There is, therefore, a trade-off between privacy and utility apart from the aggregation of users for which some applications, such as dynamic population or antenna-to-antenna traffic, are still possible.

Sharing model outputs instead of raw data allows to preserve privacy.

An alternative approach to aggregation is to keep the data in a safe environment on companies' servers and only share model outputs. This is the solution that we proposed in Chapter 4. We showed that outputs from Gaussian Process regression applied to different data sources could be combined based on the uncertainty of each model while keeping the sensitive data (CDRs) behind the firewall of the data provider. Furthermore, this approach comes with a consistent improvement of accuracy for multi-source model, especially regarding specific dimension of poverty, over using the individual data sources separately. OPAL, a project undertaken by high profile institutions and companies (briefly described in Chapter 2), aims to realize a similar approach by building a bank of validated algorithm that can be run on the servers of the data providers.

Table 1: Thesis main findings by Chapter along with details on the data used.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4
SDG	Availability component of food security (SDG 2)	All	Access component of food security (SDG 2)	Poverty (SDG 1)
Main Finding	The inter-annual variability of crop yield is the main factor limiting the accuracy of pre-harvest production estimates.	While CDRs have some limitations (bias, access, privacy), their potential for SDG applications is tremendous (mobility, socio-econmics levels and social network).	Accounting for the social capital in the transaction costs can explain a significant part of the millet price variance.	Superior accuracy can be obtained for fine scale poverty mapping using disparate data sources without jeopardizing individual privacy.
MP Data	-	CDRs, VLRs (+ signaling data, top-up amount...)	CDRs (call network)	CDRs (individualistic, spatial and temporal patterns)
EO Data	Accuracy requirements for crop area/yield estimated from EO data	-	NDVI	NDVI, Nighttime lights
Official Statistics	Crop statistics	-	Market prices	Census and Household survey
Data Features	Accuracy - Timeliness	Relevance - Accuracy - Timeliness - Access - Protection	Relevance	Relevance - Accuracy - Timeliness - Access - Protection

General Discussion

Public-private Partnership and Data Philanthropy

Most of the Big Data sources are generated and held by private companies. Access to these data is, therefore, dependent on the goodwill of the enterprises. As we have described in Chapter 1, the dependence on companies' cooperation to access data for scientific research, leads to a power imbalance between the private and public sector.

The provision of datasets to the research community by private companies is regarded by some as a form of corporate social responsibility or charity and referred to as 'data philanthropy' (Stempeck, 2014). It would be naive to believe that companies would open their dataset only to enhance their brand image and reputation. This would mean forgetting that personal data were qualified as a new asset class by the World Economic Forum (Schwab *et al.*, 2011). Porter and Kramer (2002) studied the competitive advantage of corporate philanthropy and introduced the concept of 'strategic philanthropy'. The true worth of philanthropy would not be in the positive public relations such activity generates, but rather in the enhancement of business competitiveness. This was acknowledged by Robert Kirkpatrick, the director of UN Global Pulse, in the context of development cooperation (Kirkpatrick, 2011):

> "The companies that engage with us [...] don't regard this work as an act of charity. They recognize that population well being is key to the growth and continuity of business. For example, what if you were a company that invested in a promising emerging market that is now being threatened by a food crisis that could leave your customers unable to afford your products and services? And what if it turned out that expert analysis of patterns in your own data could have revealed all along that people were in trouble, while there was still time to act? [...] The time has come for companies to recognize the importance of using their data to help the United Nations understand what is happening to the world's citizens — their customers."

There is no question here of blaming companies for trying to make a profit out of their data. But to consider operational applications, further assurance should be given that data will remain accessible independently from business decisions. This implies to foster public-private partnerships which may take the form of commercial agreements. The ultimate goal is to avoid situations such as with the Ebola crisis during which data access was very difficult (see Chapter 1 - Data Access).

Furthermore, while private corporations hold the data, the question still remains if they belong to them or rather to their individual emitters. This is an important debate which cannot be disregarded.

A New Digital Divide

Relying on Big Data analysis to tackle development issues may encounter a major risk in creating what has been called a 'new digital divide'. This effect can be explained by a lack of digital access and literacy in developing countries coming from lower analytical capacities and technical infrastructures. Countries with the most data and capacities would be in better position to exploit data science for economic benefit, even if they claim to use it to serve others. This is referred to as the 'power paradox' of Big Data by Richards and King (2013). Because Big Data sensors and Big Data pools are predominantly in the hands of powerful intermediary institutions, the data revolution may disempower and making dependent, the very communities and countries it promises to serve. This has been a long time concern of the humanitarian field, in which technologies can sometimes be seen as reinforcing existing power dynamics and social inequalities rather than disrupting them (Burns, 2015).

Another perverse impact of the asymmetry of technical capacities is that research on Big Data for development tends to be conducted remotely by non-social scientists. This makes understanding the local context challenging and inevitably impact how the research is conducted (Taylor, 2015). During the first D4D (Ivory Coast) conference, the organizers asked by a show of hands who in the audience, among those who had submitted a paper, had ever been to Ivory Coast. Out of the 150 participating research teams, only a few dozen of researchers had already visited the country, and only one researcher team said they had conducted field interviews in order to gain insights into the data's potential biases and limitations (Letouzé, 2016). On the other hand, it is worth mentioning that only one group was based in Africa (a team from Cameroon). One of the participating researchers from the first challenge was well aware of this problem:

> "Of course, this was the most interesting phenomenon of all – that we were just sitting here in the Netherlands getting this data, taking a network from the internet, getting all the other data from the internet, then, of course, the strange thing – going to a conference where nobody from Ivory Coast is, and we're telling them, "Here's your transport model"."
>
> Peter van der Mede, Director, Goudappel Coffeng, interviewed 2 May 2013 (Taylor, 2015)

During the second edition in Senegal, 11 research teams (over 260) coming from Senegal were granted access to the data, but only one eventually submitted the project on time. Consequently, important efforts should be made to bridge this 'digital divide'. This means improving collaboration between local institutions and computational experts as well as investing in technical training of local staffs to foster self-governance.

From Privacy to Transparency

In today's digital world, preserving our privacy is becoming increasingly challenging. Taylor (2015) made an interesting comparison between current digital ecosystem and the Panopticon, an ideal prison designed by the 19th- century reformer Jeremy Bentham. The Panopticon was conceived to allow a single guard to monitor all the inmates at once by inducing self-control (Figure A8 in Appendix). Although it is physically impossible for a single watchman to observe all the inmates' cells alone, the fact that they do not know when they are being watched means they are incentivized to act as if they were being observed at all times. The current data ecosystem makes obsolete such a mutual awareness between the subject and the observer because all our doing are now encoded in our digital traces. The watchman might still be alone but is has now thousand of eyes.

Our digital traces provide information on where we live, where we work, how much do we earn, our sexual orientation, whom we vote for, how healthy we are, what we buy, etc. Target, the second-largest discount store retailer in the United States, predicted that a young woman was pregnant based solely on her purchasing behavior. They sent her coupon for maternity products and nursery furnitures at her family home. The story of the father becoming furious and complaining to the manager was reported in the New York Times Magazine (Duhigg, 2012):

> "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"
>
> [...] The manager apologized and then called a few days later to apologize again.
>
> "On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

This now classic example illustrate how serious the invasion of privacy can be. In the famous suitcase Boring v. Google, Inc[2], Google's motion quoted the Restatement of Torts[3] (Google, 2009):

> "Complete privacy does not exist in this world except in a desert, and anyone who is not a hermit must expect and endure the ordinary incidents of the community life of which he is a part."

[2]In April, 2008, the Borings brought an action against Google, Inc. asserting claims for invasion of privacy, trespass, injunctive relief, negligence, and conversion based on Google's "Street View" program. The Borings alleged that their residence was only accessible by a road that was clearly marked as "private" and with "no trespassing" signs posted, but that Google still entered their property to photograph their house. The District Court granted Google's motion to dismiss as to all of the Borings' claims.

[3]Prosser, supra note 31, at 391-92; RESTATEMENT (SECOND) OF TORTS § 652D cmt. c (1977)

Google's six-lawyer team then added that:

> "Today's satellite-image technology means that even in today's desert, complete privacy does not exist. In any event, Plaintiffs live far from the desert and are far from hermits."

The District Court granted Google's motion to dismiss as to all of the Borings' claims. This case exemplifies the difficulty for plaintiffs to prevail in a suit against an intrusion on their privacy in the United States. In Europe, things are likely to improve as of 25 May 2018, the General Data Protection Regulation (GDPR) has come into force. This regulation[4] on data protection and privacy for all individuals within the European Union aims primarily to give control to citizens and residents over their personal data (European Union, 2016). The regulation includes (but is not limited to) the following key requirements:

1. **consent** must be explicit for data collected and the purposes data is used for (Article 7; defined in Article 4) ;
2. the **right to access** our personal data and information about how this personal data is being processed (Article 15) ;
3. the **right to request erasure** of our personal data (Article 17).
4. privacy settings must be set at a high level **by default** and the data processing must be **design** to guarantee data protection (Article 25).

This a big step forward for privacy protection in Europe. It is worth mentioning that, under the assumption that appropriate safeguards are implemented, research may provide a legitimate basis for processing data without the approval of the data subject. The GDPR uses a broad definition of research, without distinction between the activities of public and private entities (Recital 159). Consequently, it is unclear exactly how far the GDPR's research exemption will extend. In particular, if Big Data analytics will fall under the umbrella of research activities. (Maldoff, 2016). But this change will only applied for EU citizens while the knowledge gap is mainly a concern for the developing world.

One could argue that with the erosion of privacy, the debate will increasingly shift from privacy protection to transparency fostering. Regarding transparency, the key is not the amount of information publically available, but the symmetry of access to information. Sometimes the volume is mainly used to make things less transparent. Taylor and Kelsey (2016) take the example of courts, which are one of the few places where equal access to information is essential to prevent one party from having an unfair advantage over the other. In theory, any material accessed by one side must be made available to the other party. But the tactics of many lawyers is then to drown the other party under the information to make its ability to find key information more difficult.

[4] A regulation is binding and directly applicable while a directive requires national governments to pass enabling legislation.

They further identify several levels of transparency. At the zero level, the citizen provides data to the 'allocator' (government, corporation, hospital, etc.) without having anything in return. At level 1.0, the citizen provides data, and the allocator offers some others in return that are not necessarily related to the data provided. For example, a hospital collects data from a patient and in return provides you with data on waiting time or cleanliness. At the 2.0 level, allocator collect data, analyze it and return it to users in a more personalized way to help them make decisions. The underlying proposal is to sketch several degrees of transparency, allowing to attribute a form of labeling or rating depending on the level of openness of organizations.

Empowerment of National Statistics Office and Better Use of Traditional Data

The obvious solution to improve the data ecosystem, not discussed in this thesis, is to increase investments in national statistical offices. The NSOs are concerned about the risk that the data revolution would only be limited to new data sources and Big Data analysis. They fear (i) a potential work overload from the development of new approaches, (ii) of being bypassed by the technology, thereby making traditional data collection obsolete, and (iii) that information extracted from alternative data would not be accurate enough (Stuart *et al.*, 2015).

The first apprehension has its roots in the most fundamental challenge of NSOs in developing countries: the lack of funding and resources. This makes them vulnerable to political and interest group influence (including donors). It is the reason why, paradoxically, some developing countries are eager to embrace the Big Data as the source of official statistics. They see it as a shortcut to avoid having to build up expensive statistical capacities. Data quality should be protected and improved by strengthening NSOs, ensuring they are functioning autonomously, independent of sector ministries and political influence. If alternative data have to play a role in official statistics, serious efforts should be directed towards enhancing technical capacities of NSOs, fostering data governance and independence.

The second fear is unrealistic. NSOs will always be the primary providers of (national) official statistics as independent, accurate and trusted systems are needed. Unofficial data sources should be harnessed to fill data gaps, complement official data sources and ensure more regular reporting on important indicators but they will never supplant official statistics based on traditional data. Therefore, investments in alternative approaches to support SDG monitoring should not be to the detriment of the strengthening of existing systems.

While new technologies offer several opportunities, sometimes the most straightforward and the lowest-tech solutions perfectly responds to the data need (Bellagio Big Data Workshop Participants, 2014). In the remote region of Bawomataluo in Indonesia, children weight, height, and date of weighing are monitored thanks to stickers stuck on front doors (Coghlan, 2014). Vol-

unteers and midwives in the community update information on the sticker every month and hold a database with every child's picture, birthdate, and nutritional status. Before this mapping exercise began, nobody knew exactly how many children were in the area and how they were. With this population monitoring programme, every child and pregnant women are accounted for, and this information can be used to target essential health assistance better.

Technologies can also be used to cut down the cost of traditional data collection such as household surveys. For instance, the World Food Program has launched the Mobile Vulnerability Analysis and Mapping (mVAM) program in 2013 to track food security trends in near real-time using mobile phone survey. This approach has a high potential to get fast information on very precise topics. However, it suffers from the same selection bias as mobile phone metadata due to, for instance, mobile phone ownership. Interestingly, respondents are willing to answer more private questions during a phone call than in face-to-face meeting. On the other hand, during a phone call, the respondents can easily deceive the operator about their current situation. Still, this approach might be the most promising technological innovation for data collection in remote and dangerous areas.

The final concern that alternative data tend to be inaccurate is certainly the most grounded. Like illustrated in this thesis, while covering large parts of the population, Big Data might be less accurate than rigorous statistical sampling because of their inherent bias. This issue is true for developed countries as well. The head of French NSO, INSEE, explained why he thinks that Big Data has, for now, low operational potential for official statistics (CNIS, 2013):

> "INSEE is following Big Data attentively. However, all articles on the subject refer to very advanced indicators that, for the time being, present little interest, in that they can only save a few days compared to the production of cyclical statistical indicators and nothing that seems to be operational in that respect"

In view of all these elements, the first priority regarding the data revolution should be to support NSOs through funding and technical training. However, this effort has a cost that should be acknowledged. For instance, assuming that a census has an average cost of USD 2[5] per capita, a worldwide population census would reach a total cost of 15 billion USD. By neglecting to appropriately fund and train statistical offices, we take the risks to see low-cost solutions emerging and becoming the reference.

[5] 2 USD per capita is a reasonable estimate according to Jerven (2017). However, at national scale large disparities exist. For instance, US census is the most expensive census of the world with a cost of ∼USD 50 per capita while the one of Denmark costs as little as USD 0.05 per capita (UN, 2009)

From Data to Decision

There is a risk that analyses based on Big Data will focus too much on correlation and prediction rather than on causes, diagnostics or inference, without which policy is virtually blind. Correlation is not necessarily informative. With a big enough dataset, it is easy to find correlations between almost anything. When looking at sparse data, these artifacts can be manually spotted. However, the task is more challenging for large data sets. The potential consequence is misleading and potentially harmful recommendations for policies. Research, therefore, needs to develop new methods and algorithms that can handle Big Data and address issues regarding how one can compare data sets and draw robust inferences.

To turn the data revolution into actions, data have to be translated into information that is easy to understand and useful for the end-users. The old saying is misleading; data do not speak for themselves (boyd and Crawford, 2012). This requires efforts to improve 'research literacy', in particular, a better communication about the limits and the uncertainty on data and findings. Finally, the best data in the world – timely, relevant, accurate and easy to understand – will not by themselves lead to change. Change is only possible if governments and citizens leverage the data knowledge to make a difference in people's lives.

On the other hand, one should not neglect the political and economic context in which the data revolution is conducted. Historical factors, institutional organizations, as well as incentives influence how and why informative systems operate (Krätke and Byiers, 2014). This might explain some of the obstacles to improving the data ecosystem and/or using good official statistics. Technological solutions are a promising avenue but are not sufficient as official statistics are inherently political.

Appendix

Annexes to Chapter 1

Correlation between Rainfall and Crop Yield

The two interesting rainfall features to predict crop yield variability at the country level are the cumulative annual rainfall variability and temporal characteristics (frequency and intensity) (Berg *et al.*, 2010). Here, average cumulative rainfall over the country was used. Figure A1 shows the correlation between cumulative rainfall over the rainy season and the crop yield. The most correlated period depends on each crop.

Most of the rice is irrigated in Senegal River Valley, and its dependence on rainfall is therefore limited which is illustrated by the very low correlation observed for this crop. Cotton is mainly cultivated in the South of the country (Figure 1.2) where the rainy season starts early compared to the rest of the country. This could explain the higher correlation observed for this crop early in the season (July).

Groundnuts, sorghum, and millet were better estimated later in the season (September) while maize and cassava showed higher correlation for the end of July. The correlations observed for the yield were in line with previous results in Senegal (Ndao and Breuer, 2013). For the area, an unexpected high correlation was found for millet (∼0.77).

Rainfall data were downloaded from the seasonal explorer developed by the Vulnerability Assessment and Mapping of the World Food Programme.

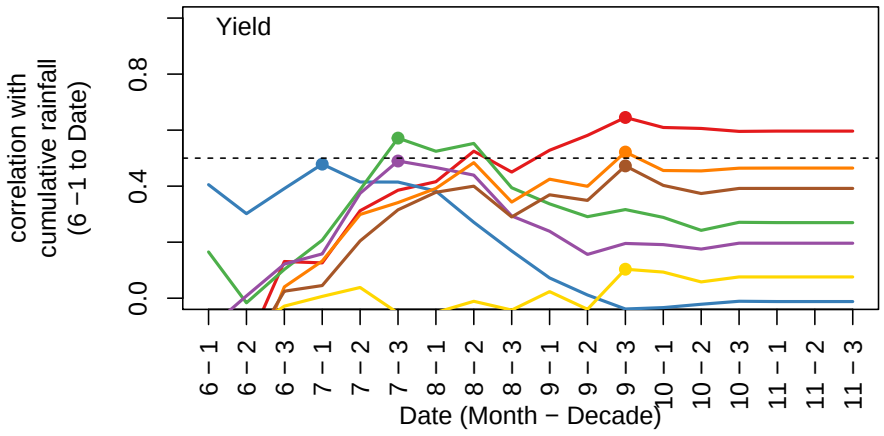


Figure A1: Correlation between crop yield and average accumulated rainfall over the country (from the first decade of June to each decade of the rainy season) for 20 years (1997-2016). All correlation higher than 0.5 (dotted line) were statistically significant (p-value$<$0.01). For color code, refer to Figure 1.4.

The primary data source is the CHIRPS gridded rainfall dataset produced by the Climate Hazards Group at the University of California, Santa Barbara. CHIRPS is a 35+ year quasi-global rainfall dataset incorporating satellite imagery with in-situ station data to create gridded rainfall time series. Full details on the underlying methodology can be found in (Funk *et al.*, 2015). CHIRPS data were shown to have low systematic errors (bias) and low mean absolute errors (Peterson *et al.*, 2013). CHIRPS is free to use and accessible online (http://chg.geog.ucsb.edu/data/chirps/).

Distribution of errors for identical CV(RMSE)

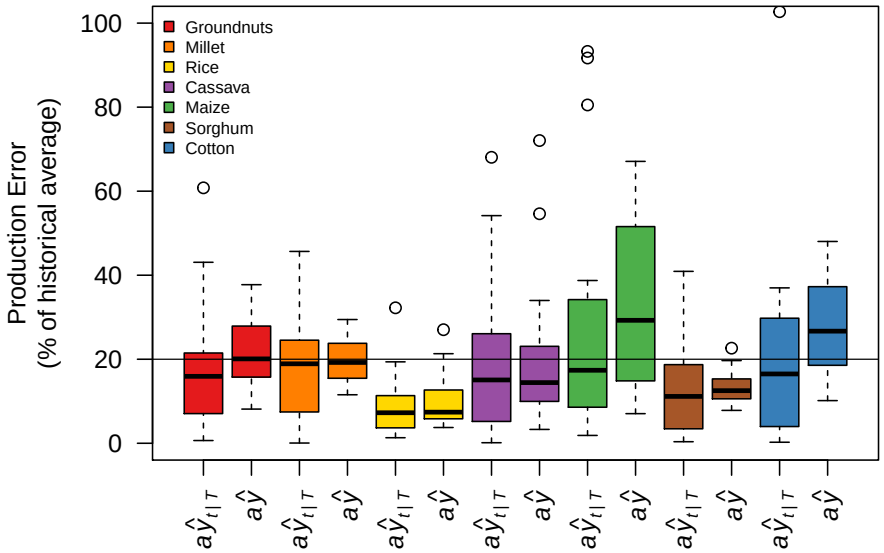


Figure A2: Box-plots of the distribution of production errors (% of historical average) for $\hat{p}_{aug} = a\hat{y}_{t|T}$ and $\hat{p}_{aug} = a\hat{y}$ where the accuracy of $\hat{y}$ corresponds to the accuracy requirement of crop yield defined in Table 1.3. Both distribution give the exact same CV(RMSE). However, $a\hat{y}$ has a higher median value but with lower number of outliers.

Annexes to Chapter 3

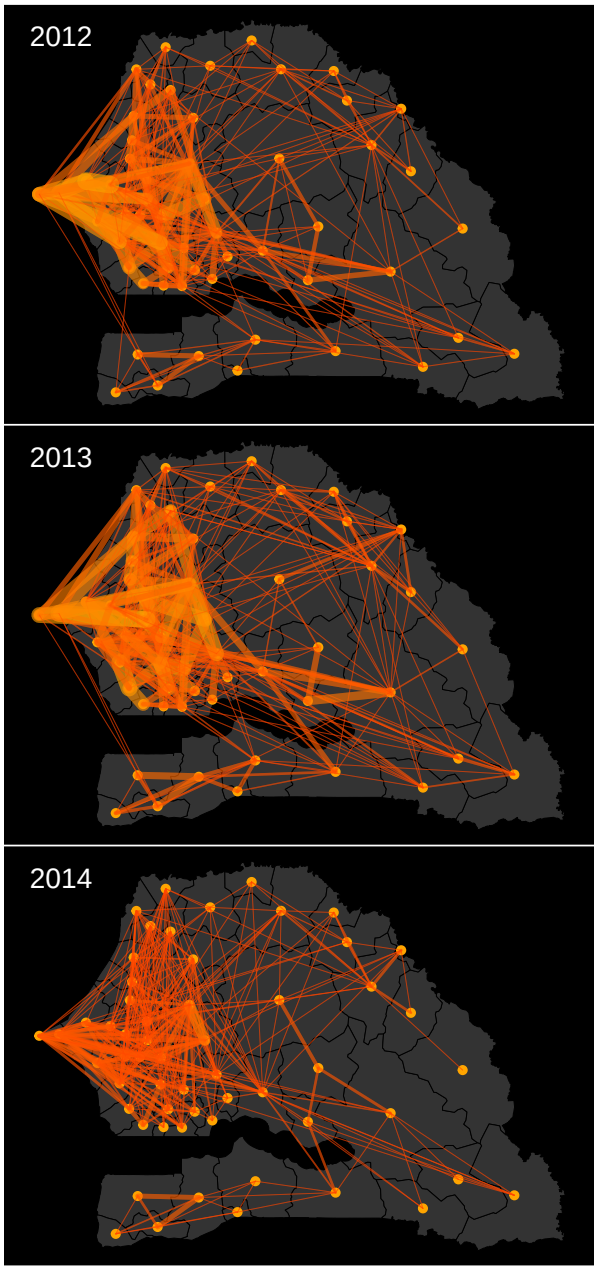


Figure A3: Network visualization of Figure 3.8 for each year separately. The width of each line is proportional to the trade flow intensity. The flow direction is not shown.

Annexes to Chapter 4

Gaussian Process Regression Model

The following model is assumed to predict poverty for a commune from a single data source (CDR or environment):

$$y_i = \beta^\top \mathbf{x}_i + f(\mathbf{x}_i) + \epsilon \tag{17}$$

where y_i is the target poverty value and $\mathbf{x}_i$ is a vector of independent variables derived from the particular view for the i^{th} commune. Instead of assuming a fixed parametric form for $f()$, we adopt a non-parametric approach, by assuming a Gaussian Process (GP) prior on $f()$, with zero mean function, and kernel function, $k()$. The generative process thus becomes:

$$\begin{aligned} f(\mathbf{x}) &\sim GP(0, k(\mathbf{x}, \mathbf{x}')) \\ y_i &\sim \mathcal{N}(\beta^\top \mathbf{x}_i + f(\mathbf{x}_i), \sigma_n^2), \forall i \end{aligned}$$

A GP is a stochastic process, such that any finite sample generated from this stochastic process is jointly multivariate normal (Rasmussen and Williams, 2006).

The posterior distribution of $f(\mathbf{x}_*)$ at a test input, $\mathbf{x}_*$, can be computed given a training set of examples, $\{\mathbf{x}_i, f(\mathbf{x}_i)\}_{i=1}^N$. The joint distribution of the training outputs, $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots$, and the test output, $f(\mathbf{x}_*)$, according to the GP prior is:

$$\begin{bmatrix} f(\mathbf{x_1}) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_N) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x_1}) & \ldots & k(\mathbf{x}_1, \mathbf{x_N}) & k(\mathbf{x}_1, \mathbf{x}_*) \\ k(\mathbf{x}_2, \mathbf{x_1}) & \ldots & k(\mathbf{x}_2, \mathbf{x_N}) & k(\mathbf{x}_2, \mathbf{x}_*) \\ \vdots & \ddots & \vdots & \vdots \\ k(\mathbf{x}_N, \mathbf{x_1}) & \ldots & k(\mathbf{x}_N, \mathbf{x_N}) & k(\mathbf{x}_N, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x_1}) & \ldots & k(\mathbf{x}_*, \mathbf{x_N}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

For notational simplicity, let K denote a $N \times N$ matrix which contains the kernel computation on each pair of training inputs, i.e., $K[i,j] = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}$ be a vector of the kernel computation between each training input and the test input, i.e., $\mathbf{k}[i] = k(\mathbf{x}_i, \mathbf{x}_*)$, and k_* be the self-covariance for $\mathbf{x}$, i.e., $k_* = k(\mathbf{x}_*, \mathbf{x}_*)$. Moreover, let $\mathbf{f}$ be a $N \times 1$ vector, such that $\mathbf{f}[i] = f(\mathbf{x}_i)$. The above equation can be written as:

$$\begin{bmatrix} \mathbf{f} \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K & \mathbf{k} \\ \mathbf{k}^\top & k_* \end{bmatrix}\right)$$

Since $\mathbf{f}$ and $f(\mathbf{x}_*)$ are jointly Gaussian, one can make use of the well-known Gaussian identity (Mises, 1964) for the conditional distribution of $f(\mathbf{x}_*)$, i.e.,:

$$f(\mathbf{x}_*)|\mathbf{f} \sim \mathcal{N}\left(\mathbf{k}^\top K^{-1}\mathbf{f}, k_* - \mathbf{k}^\top K^{-1}\mathbf{k}\right) \tag{18}$$

We assume that the observed poverty for the i^{th} commune, y_i is equal to the sum of the linear term, the latent function value, with zero mean GP prior, and an independent and identically distributed Gaussian noise ($\sim \mathcal{N}(0, \sigma_n^2)$). Thus, the prior on the observed data will be:

$$\begin{aligned} \mathbb{E}[y_i] &= \beta^\top \mathbf{x}_i \\ \text{cov}[y_i, y_j] &= k(\mathbf{x}_i, \mathbf{x}_j) + \delta_{ij}\sigma_n^2 \end{aligned}$$

where δ_{ij} is the Kronecker delta, such that $\delta_{ij} = 1$, if $(i = j)$, and 0, otherwise. For the entire training data set:

$$\begin{aligned} \mathbb{E}[\mathbf{y}] &= \mathbf{b} \\ \text{cov}[\mathbf{y}] &= K + \sigma_n^2 I \end{aligned}$$

where $\mathbf{b}$ is a N length vector, such that $\mathbf{b}[i] = \beta^\top \mathbf{x}_i$ and I is the $N \times N$ identity matrix. The joint distribution of $\mathbf{y}$ and $f(\mathbf{x}_*)$ can be written as:

$$\begin{bmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{b} \\ \beta^\top \mathbf{x}_* \end{bmatrix}, \begin{bmatrix} K + \sigma_N^2 I & k \\ k^\top & k_* \end{bmatrix} \right)$$

Using the conditional Gaussian result, similar to (18), and noting the relation between y_* and $f(\mathbf{x}_*)$ from (4.3), the conditional distribution for the prediction, y_*, becomes:

$$\begin{aligned} \mathbb{E}[y_*] &= \beta^\top \mathbf{x}_* + \mathbf{k}^\top (K + \sigma_n^2 I)^{-1} (\mathbf{y} - \mathbf{b}) \\ \text{var}[y_*] &= k_* - \mathbf{k}^\top (K + \sigma_n^2 I)^{-1} \mathbf{k} + \sigma_n^2 \end{aligned}$$

Estimating Moments of a Mixture Distribution

Let random variable y represent a mixture of two unimodal normal distributions, $y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and mixing probabilities w_1 and w_2, such that $w_1 + w_2 = 1$, i.e.,

$$y = w_1 y_1 + w_2 y_2$$

Any moment of y can be computed as (Kim and White, 2003):

$$\mathbb{E}[y^k] = w_1 \mathbb{E}[y_1^k] + w_2 \mathbb{E}[y_2^k]$$

Which directly gives:

$$\mathbb{E}[y] = w_1 \mu_1 + w_2 \mu_2$$

The expression for the variance of y can be derived as follows:

$$\begin{aligned} \text{var}[y] &= \mathbb{E}[y^2] - (\mathbb{E}[y])^2 \\ &= w_1 \mathbb{E}[y_1^2] + w_2 \mathbb{E}[y_2^2] - (w_1 \mu_1 + w_2 \mu_2)^2 \\ &= w_1 (\text{var}[y_1] + \mu_1^2) + w_2 (\text{var}[y_1] + \mu_1^2) - (w_1 \mu_1 + w_2 \mu_2)^2 \\ &= w_1 \sigma_1^2 + w_2 \sigma_2^2 + w_1 \mu_1^2 + w_2 \mu_2^2 - w_1^2 \mu_1^2 - w_2^2 \mu_2^2 - 2 w_1 w_2 \mu_1 \mu_2 \\ &= w_1 \sigma_1^2 + w_2 \sigma_2^2 + w_1 w_2 \mu_1^2 + w_1 w_2 \mu_1^2 - 2 w_1 w_2 \mu_1 \mu_2 \\ &= w_1 \sigma_1^2 + w_2 \sigma_2^2 + w_1 w_2 (\mu_1 - \mu_2)^2 \end{aligned}$$

The last result makes use of the fact that $w_1 + w_2 = 1$. This allows us to understand the formula as stating the variance of the mixture is the mixture of the variances plus a non-negative term accounting for the (weighted) dispersion of the means.

Table A1: A summary of poverty indicators and associated deprivations, with emphasis how our methodology calculates them using the RGPHAE census data, keeping in view the OPHI guidelines.

Poverty indicators	Deprivation of a household used by OPHI for MPI calculation	RGPHAE Census questionnaire response used by our methodology for MPI calculation
Health		
Child Mortality	At least, one child has died	At least, one child has died
Nutrition	Any member is undernourished	At least, one member has skipped a meal during the last 12 months
Education		
School attendance	Any school-aged child is not attending school up to grade 8	Any school-aged (5-18 years old) is not attending school
Years of schooling	No member that has completed at least 5 years of education	About higher schooling of any member
Standard of Living		
Cooking fuel	Uses solid fuels for cooking	Household does not use electricity or natural gas for cooking
Electricity	No access to electricity	No electricity, or generator
Sanitation	No access to adequate sanitation or if it's shared	Household has no sewer connection or pit
Drinking water	No access to safe drinking water	No water tap in household
Flooring	Has dirt/earth/dung floor	Household has dirt/earth/dung floor
Assets	Has a maximum of one small assets (radio, TV, refrigerator, phone, bicycle, motorbike) AND it has no car	Household has one asset (radio, TV, refrigerator, phone, bicycle, motorbike) AND has no car

Table A2: Spatially-cross validated results of the predictions of MPI, Incidence of poverty (H), and Intensity of poverty (A), along with the individual indicators for poverty given by our model using disparate datasets. The results are compared to models learned on single source and on concatenated feature space. **corr.** – Pearson's r correlation, **rank corr.** – Spearman's rank correlation, and **RMSE** – Root Mean Square Error. For both types of correlations, all p-values were less than 10^{-20}. A standard deviation associated with the multiple runs for each measurement is reported within parenthesis.

	Multi-source Data			CDR			Environment			Concatenated		
	corr.	**rank corr.**	**RMSE**	**corr.**	**rank corr.**	**RMSE**	**corr.**	**rank corr.**	**RMSE**	**corr.**	**rank corr.**	**RMSE**
Poverty Indicator												
MPI	0.91	0.88	0.08	0.89	0.86	0.08	0.84	0.80	0.10	0.90	0.85	0.10
	(0.06)	(0.06)	(0.01)	(0.07)	(0.07)	(0.01)	(0.09)	(0.10)	(0.02)	(0.06)	(0.07)	(0.02)
H	0.91	0.85	10.79	0.90	0.84	10.76	0.83	0.75	13.65	0.90	0.83	11.34
	(0.07)	(0.08)	(3.96)	(0.08)	(0.08)	(2.60)	(0.11)	(0.11)	(4.86)	(0.07)	(0.08)	(3.87)
A	0.86	0.85	4.71	0.83	0.82	4.98	0.81	0.79	5.36	0.84	0.82	5.52
	(0.05)	(0.07)	(0.96)	(0.07)	(0.08)	(1.14)	(0.07)	(0.08)	(0.75)	(0.07)	(0.08)	(1.40)
Individual Indicators of Poverty												
Education												
Years of Schooling	0.85	0.85	12.00	0.81	0.80	13.30	0.76	0.75	15.42	0.85	0.84	12.06
	(0.04)	(0.04)	(1.21)	(0.05)	(0.06)	(1.55)	(0.07)	(0.08)	(2.48)	(0.04)	(0.04)	(1.01)
School Attendance	0.86	0.83	11.68	0.82	0.81	12.85	0.75	0.72	14.54	0.85	0.83	11.60
	(0.05)	(0.06)	(1.83)	(0.07)	(0.07)	(1.73)	(0.09)	(0.09)	(3.06)	(0.05)	(0.06)	(2.05)
Health												
Child Mortality	0.45	0.46	10.91	0.45	0.48	11.32	0.34	0.33	11.54	0.45	0.45	10.85
	(0.15)	(0.16)	(0.58)	(0.13)	(0.13)	(0.73)	(0.19)	(0.21)	(0.65)	(0.14)	(0.16)	(0.49)
Nutrition	0.52	0.53	14.61	0.54	0.55	14.49	0.38	0.37	16.28	0.47	0.46	15.33
	(0.15)	(0.15)	(3.65)	(0.11)	(0.11)	(3.10)	(0.26)	(0.25)	(3.99)	(0.21)	(0.22)	(4.24)
Standard of Living												
Cooking Fuel	0.86	0.70	13.82	0.83	0.68	12.98	0.76	0.58	16.49	0.86	0.70	15.56
	(0.14)	(0.18)	(8.76)	(0.14)	(0.16)	(7.00)	(0.20)	(0.25)	(8.78)	(0.13)	(0.18)	(9.19)

Sanitation	0.79	0.70	16.99	0.74	0.69	18.05	0.72	0.61	18.64	0.77	0.66	18.69
	(0.17)	(0.18)	(3.42)	(0.17)	(0.17)	(3.14)	(0.22)	(0.26)	(4.33)	(0.20)	(0.23)	(3.91)
Water	0.75	0.72	14.60	0.74	0.71	14.70	0.67	0.61	16.97	0.68	0.62	17.15
	(0.14)	(0.14)	(3.22)	(0.13)	(0.12)	(2.98)	(0.20)	(0.21)	(3.25)	(0.21)	(0.22)	(3.20)
Electricity	0.88	0.84	15.09	0.86	0.83	16.67	0.79	0.72	20.27	0.84	0.80	18.61
	(0.04)	(0.07)	(0.98)	(0.04)	(0.06)	(1.25)	(0.10)	(0.13)	(1.72)	(0.05)	(0.09)	(1.65)
Floor	0.78	0.68	15.79	0.79	0.70	15.24	0.64	0.54	17.87	0.74	0.63	16.58
	(0.15)	(0.14)	(5.79)	(0.13)	(0.12)	(4.93)	(0.24)	(0.23)	(6.22)	(0.19)	(0.16)	(5.81)
Asset ownership	0.89	0.86	12.61	0.87	0.85	13.81	0.80	0.75	17.05	0.85	0.82	15.37
	(0.04)	(0.05)	(1.33)	(0.04)	(0.04)	(1.20)	(0.11)	(0.11)	(2.69)	(0.05)	(0.06)	(1.48)

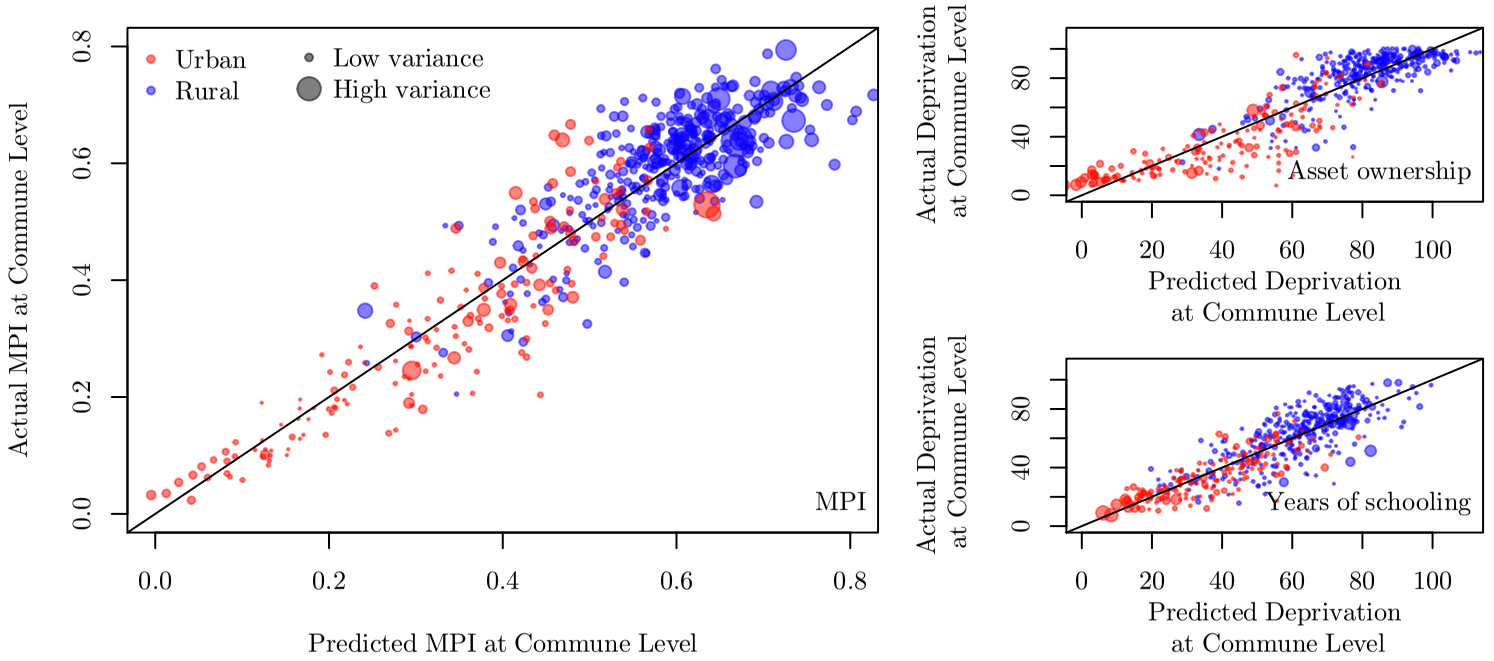


Figure A4: The left panel denotes the comparison of actual and predicted MPI values for all communes and urban areas of Senegal. The rural and urban areas are differentiated using blue and red colors respectively. The size of the circle denotes the variance of MPI prediction for that commune. The top right panel shows how the actual and predicted values compare for asset ownership, while the one on the bottom shows the comparison for years of schooling. This figure is the same as Figure 4.4 but a bias correction has been applied using a simple linear regression between predicted and actual values.

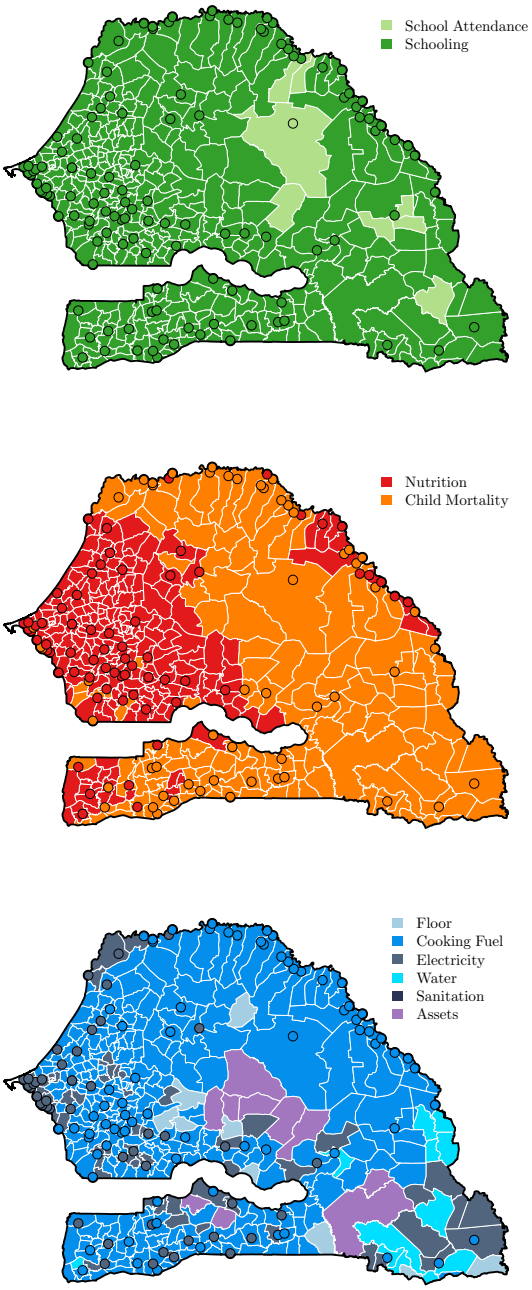


Figure A5: The highest deprivation by commune as predicted by our model for each dimension of global MPI (from top to bottom: education, health and standard of living)

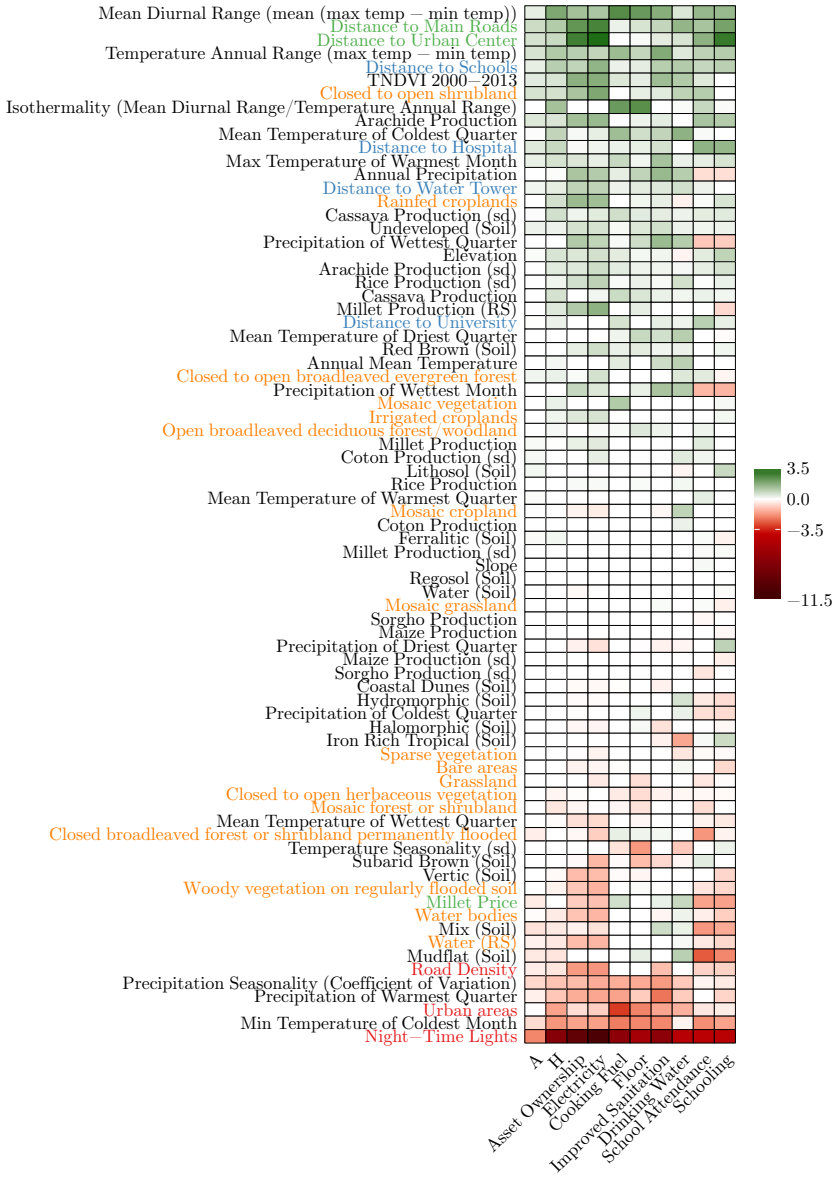


Figure A6: Visualization of selected features using elastic net regularization on environment data for prediction of selected deprivations. The rows represent the features, which are ranked according to their weights from positive (marked green) to negative (marked red). Different features groups are color coded. Features related to food availability are given black color, while those related to food accessibility are colored green. The land cover features are colored yellow, and the features detailing economic activity are given red color. Finally, features depicting access to services are shown in blue. The cells in white were given 0 weights by our model.

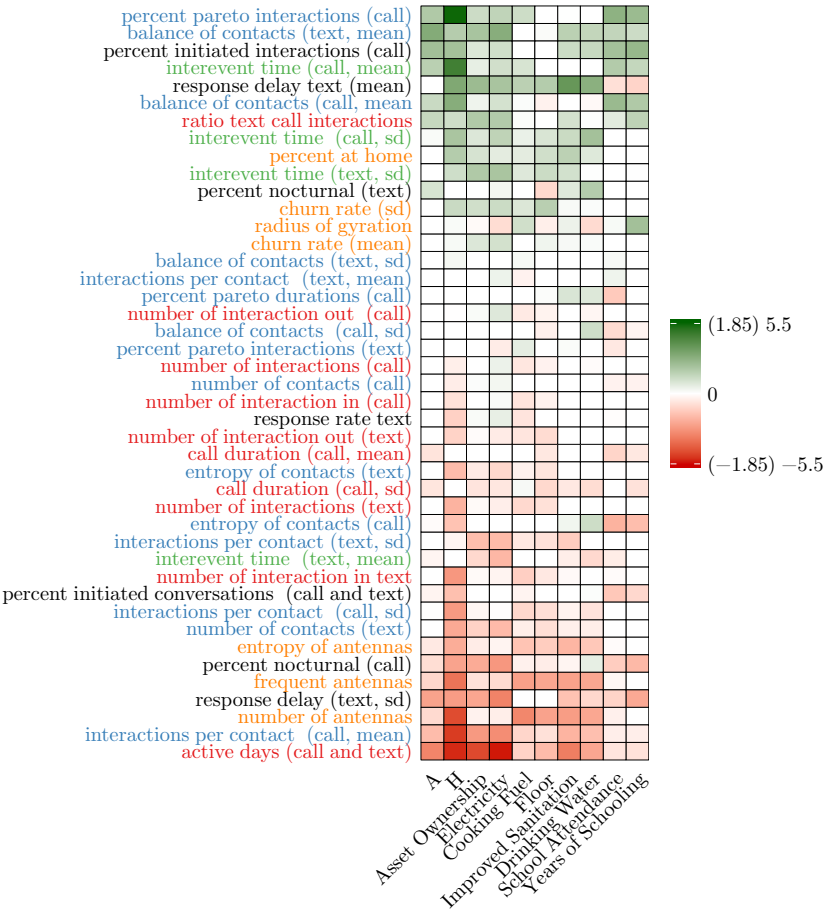


Figure A7: Visualization of selected features using elastic net regularization on CDR data for prediction of selected deprivations. The rows represents features, which are ranked according to their weights from positive (marked green) to negative (marked red). The columns are the various deprivations. The features groups are color coded. Features related to diversity features are colored blue. Those related to spatial aspects are colored yellow. The features related to active behavior are marked in black. The feature related to basic phone usage are given red color, and those related to regularity The cells in white were given 0 weights by our model. Legend in parenthesis correspond to the different variation in weights. H and A weights vary between 1.85 to -1.85, for others the weights vary between 5.5 to -5.5.

Table A3: List of the important features chosen by our model to predict each of H, A, Schooling, School Attendance, Cooking Fuel, Sanitation, Water, Electricity, Floor, Assets. The features having positive relationship with the various deprivations are marked as + in the cell corresponding to the feature name and the deprivation. Otherwise it is marked as -. The various semantic groupings under which the different features fall is also listed.

Feature Type	H	A	Schooling	School Attendance	Cooking Fuel	Sanitation	Water	Electricity	Floor	Assets
Basic										
Active days call & text	-	-			-	-	-	-	-	-
Ratio of call/text interactions			+			+		+		+
Number of interactions in text					-					
Regularity										
Inter-event time call (mean)	+	+	+	+						
Inter-event time call/text (std)						+	+	+	+	+
Diversity										
Balance of contacts text (mean)		+	+	+		+	+	+		+
Percent Pareto Interactions call	+	+	+	+	+			+		+
Interactions per contact call (mean)	-	-			-	-	-	-		-
Entropy of contacts call			-	-				-		
Active										
Response delay text (mean)	+		-		+	+	+	+	+	+
Response delay text (std)		-	-	-		-	-	-		-
Percent initiated interactions (call)		+	+	+		+	+			
Percent initiated conversations (call & text)			-	-						
Spatial										
Frequent antennas	-				-	-	-		-	
Number of antennas	-				-	-	-		-	
Radius of gyration			+		+					
Entropy of antennas					-	-	-		-	

Annexes to Conclusions

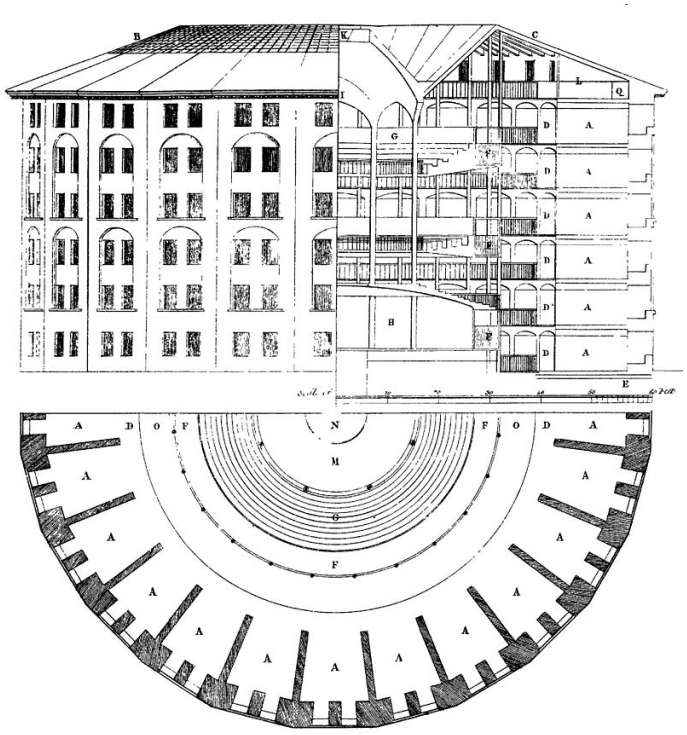


Figure A8: Elevation, section and plan of Jeremy Bentham's Panopticon penitentiary, drawn by Willey Reveley, 1791 (Bentham and Bowring, 1843).

References

Abu-Ismail, K., Imady, O., Kuncic, A., Nojoum, O., and Walker, J. 2016. Syria at War, Five Years On. Technical report, United Nations Economic and Social Commission for Western Asia.

Ahas, R., Silm, S., Järv, O., Saluveer, E., and Tiru, M. 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of urban technology, 17(1):3–27.

Aker, J. C. 2010. Information from markets near and far: Mobile phones and agricultural markets in Niger. American Economic Journal: Applied Economics, 2(3):46–59. Doi: 10.1257/app.2.3.46.

Aker, J. C. and Blumenstock, J. E. 2014. *The economic impacts of new technologies in Africa*. The Oxford Handbook of Africa and Economics: Policies and Practices.

Aker, J. C. and Mbiti, I. M. 2010. Mobile phones and economic development in Africa. The Journal of Economic Perspectives, 24(3):207–232.

Aksoy, M. A. and Isik-Dikmelik, A. 2008. Are low food prices pro-poor? Net food buyers and sellers in low-income countries.

Alexandratos, N., Bruinsma, J., *et al.*. 2012. World agriculture towards 2030/2050: the 2012 revision. Technical report, ESA Working paper FAO, Rome.

Alkire, S. and Foster, J. 2011a. Counting and multidimensional poverty measurement. Journal of public economics, 95(7):476–487.

Alkire, S. and Santos, M. E. 2010. Acute multidimensional poverty: A new index for developing countries. United Nations development programme human development report office background paper, (2010/11).

Alkire, S. and Santos, M. E. 2014. Measuring acute poverty in the developing world: Robustness and scope of the multidimensional poverty index. World Development, 59:251–274.

Alkire, S., Samman, E., *et al.*. 2014. Mobilising the Household Data Required to Progress toward the SDGs. Technical report, University of Oxford.

Alkire, S., Roche, J. M., Ballon, P., Foster, J., Santos, M. E., and Seth, S. 2015. *Multidimensional poverty measurement and analysis*. Oxford University Press, USA.

Amarasinghe, U., Samad, M., and Anputhas, M. 2005. Spatial clustering of rural poverty and food insecurity in Sri Lanka. Food Policy, 30(5):493–509.

Atzberger, C. 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sensing, 5(2):949–981.

Autorite de Régulation des Telecommunications et des Postes. 2013. Observatoire de la Telephonie Mobile: Tableau de bord au 31 decembre 2013. Technical report.

Bagrow, J. P., Wang, D., and Barabasi, A.-L. 2011. Collective response of human populations to large-scale emergencies. PloS one, 6(3):e17680.

Bahn, V. and McGill, B. J. 2013. Testing the predictive performance of distribution models. Oikos, 122(3):321–331.

Ballivian, A. 2014. Using Big Data for the Sustainable Development Goals.

Banda, J. P. 2003. Nonsampling errors in surveys. United Nations Secretariat ESA/STAT/AC, 93(7).

Banerjee, A. and Duflo, E. 2012. *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.

Barabasi, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. Nature, 435(7039):207.

Barabási, A.-L. 2010. *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin.

Barabási, A.-L. 2016. *Network science*. Cambridge university press.

Barr, A. 2000. Social capital and technical information flows in the Ghanaian manufacturing sector. Oxford Economic Papers, 52(3):539–559. Doi: 10.1093/oep/52.3.539.

Barrett, C. B. 2010. Measuring food insecurity. Science, 327(5967):825–828.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. 2013. Human mobility characterization from cellular network data. Communications of the ACM, 56(1):74–82.

Beegle, K., Christiaensen, L., Dabalen, A., and Gaddis, I. 2016. *Poverty in a rising Africa*. World Bank Publications.

Bellagio Big Data Workshop Participants. 2014. Big data and positive social change in the developing world: A white paper for practitioners and researchers. Technical report, Oxford Internet Institute.

Benson, T., Chamberlin, J., and Rhinehart, I. 2005. An investigation of the spatial determinants of the local prevalence of poverty in rural Malawi. Food Policy, 30(5):532–550.

Bentham, J. and Bowring, J. 1843. *The Works of Jeremy Bentham*, volume 7. W. Tait.

Berdegué, J. A., Bebbington, A., and Escobal, J. 2015. Conceptualizing spatial diversity in Latin American rural development: Structures, institutions, and coalitions. World development, 73:1–10.

Berg, A., Sultan, B., and de Noblet-Ducoudré, N. 2010. What are the dominant features of rainfall leading to realistic large-scale crop yield simulations in West Africa? Geophysical Research Letters, 37(5).

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., and Sbodio, M. L. 2013. AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 663–666. Springer.

Bertholet, F., Tracz, G., Dinghem, S., Irwin, T., Diouf, I., Ndiaye, I., Wodon, Q., SIaens, C., and Savard, L. 2004. Le secteur des transports routiers au Sénégal. Technical report, World Bank.

Beugelsdijk, S. and Schaik, T. V. 2003. Social capital and regional economic growth. In *43rd Congress of the European Regional Science Association: "Peripheries, Centres, and Spatial Development in the New Europe"*. Louvain-la-Neuve: European Regional Science Association (ERSA).

Bigsten, A., Collier, P., Dercon, S., Fafchamps, M., Gauthier, B., Gunning, J. W., Oduro, A., Oostendorp, R., Patillo, C., Soderbom, M., *et al.*. 2000. Contract flexibility and dispute resolution in African manufacturing. The Journal of Development Studies, 36(4):1–37. Doi: 10.1080/00220380008422635.

Blanford, J. I., Huang, Z., Savelyev, A., and MacEachren, A. M. 2015. Geo-located tweets. enhancing mobility maps and capturing cross-border movement. PloS one, 10(6):e0129202.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008.

Blondel, V. D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., and Ziemlicki, C. 2012. Data for development: the D4D challenge on mobile phone data. arXiv preprint arXiv:1210.0137.

Blondel, V. D., Decuyper, A., and Krings, G. 2015. A survey of results on mobile phone datasets analysis. EPJ Data Science, 4(1):1. Doi: 10.1140/epjds/s13688-015-0046-0.

Blumenstock, J. and Eagle, N. 2010. Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 6. ACM.

Blumenstock, J., Cadamuro, G., and On, R. 2015a. Predicting poverty and wealth from mobile phone metadata. Science, 350(6264):1073–1076.

Blumenstock, J. E. 2015. Calling for better measurement: Estimating an individual's wealth and well-being from mobile phone transaction records.

Boerma, J. T. and Stansfield, S. K. 2007. Health statistics now: are we making the right investments? The Lancet, 369(9563):779–786.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. 2014. Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 427–434.

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., and Smoreda, Z. 2015. Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations. Transportation Research Procedia, 11:381–398.

boyd, d. and Crawford, K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5):662–679.

Budescu, D. V., Broomell, S., and Por, H.-H. 2009. Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. Psychological science, 20(3):299–308.

Burchi, F. and De Muro, P. 2016. From food availability to nutritional capabilities: Advancing food security analysis. Food Policy, 60:10–19.

Burke, M. and Lobell, D. B. 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proceedings of the National Academy of Sciences, 114(9):2189–2194.

Burns, R. 2015. Rethinking big data in digital humanitarianism: Practices, epistemologies, and social relations. GeoJournal, 80(4):477–490.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., and Ratti, C. 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies, 26:301–313.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. 2008b. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 41(22):224015. Doi: 10.1088/1751-8113/41/22/224015.

Carr, J. A., D'Odorico, P., Suweis, S., and Seekell, D. A. 2016. What commodities and countries impact inequality in the global food system? Environmental Research Letters, 11(9):095013.

Carr-Hill, R. 2013. Missing millions and measuring development progress. World Development, 46:30–44.

Castells, M., Fernandez-Ardevol, M., Qiu, J. L., and Sey, A. 2009. *Mobile communication and society: A global perspective*. MIT Press. Doi: 10.1111/j.1944-8287.2008.tb00398.x.

Centre de Gestion et d'Économie Rurale de la Vallée du fleuve Sénégal. 2014. Analyse Économique - Les Exploitations Agricoles Familiales du Sénégal. Technical report.

Challinor, A., Wheeler, T., Craufurd, P., Slingo, J., and Grimes, D. 2004. Design and optimisation of a large-area process-based model for annual crops. Agricultural and forest meteorology, 124(1):99–120.

Chen, M., Mao, S., and Liu, Y. 2014. Big data: A survey. Mobile networks and applications, 19(2):171–209.

CNIS. 2013. Assemblée plénière - Réunion du 24 janvier 2013. Technical report, Conseil National de l'information statistique.

CNN. 2016. More Africans have access to cell phone service than piped water. https://edition.cnn.com/2016/01/19/africa/africa-afrobarometer-infrastructure-report/index.html. Accessed: 2018-04-25.

Coghlan, R. 2014. More than numbers. Why better data adds up to saving the lives of women and children. Technical report, World Vision International.

Commissariat de la Sécurité Alimentaire. 2000-2014. Food and Commodity Prices Data.

Cox, M. and Ellsworth, D. 1997. Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th conference on Visualization'97*, pages 235–ff. IEEE Computer Society Press.

Dasgupta, P. and Ray, D. 1986. Inequality as a determinant of malnutrition and unemployment: theory. The Economic Journal, 96(384):1011–1034.

Dasgupta, S., Deichmann, U., Meisner, C., and Wheeler, D. 2005. Where is the poverty–environment nexus? Evidence from Cambodia, Lao PDR, and Vietnam. World Development, 33(4):617–638.

Davies, S., Smith, B., and Lambert, R. 1991. *Early warning in the Sahel and Horn of Africa: the state of the art. A review of the literature. Volume 1 of a three-part series.*, volume 20. Institute of Development Studies.

De Hoyos, R. E. and Medvedev, D. 2011. Poverty effects of higher food prices: a global perspective. Review of Development Economics, 15(3):387–402.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. 2013a. Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3.

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

de Montjoye, Y.-A., Quoidbach, J., Robic, F., and Pentland, A. S. 2013c. Predicting personality using novel mobile phone-based metrics. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 48–55. Springer.

de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. 2014. D4D-Senegal: the second mobile phone data for development challenge. arXiv preprint arXiv:1407.4885.

de Montjoye, Y.-A., Rocher, L., and Pentland, A. S. 2016. bandicoot: a Python Toolbox for Mobile Phone Metadata. Journal of Machine Learning Research, 17(175):1–5.

De Sherbinin, A., VanWey, L. K., McSweeney, K., Aggarwal, R., Barbieri, A., Henry, S., Hunter, L. M., Twine, W., and Walker, R. 2008. Rural household demographics, livelihoods and the environment. Global Environmental Change, 18(1):38–53.

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J.-M., Krings, G., Gutierrez, T., Blondel, V. D., and Luengo-Oroz, M. A. 2014. Estimating food consumption and poverty indices with mobile phone data. arXiv preprint arXiv:1412.2595.

Defourny, P., Blaes, X., Bogaert, P., *et al.*. 2007. Respective contribution of yield and area estimates to the error in crop production forecasting. In *ISPRS Archives XXXVI-8/W48 Workshop proceedings: Remote sensing support to crop yield forecast and area estimates*.

Defourny, P., Schouten, L., Bartalev, S., Bontemps, S., Cacetta, P., De Wit, A., Bella, C. d., Gerard, B., Giri, C., Gond, V., *et al.*. 2009. Accuracy assessment of a 300 m global land cover map: The GlobCover experience. In *The 33rd International Symposium on Remote Sensing of Environment*.

Delincé, J., Lemoine, G., Defourny, P., Gallego, J., Davidson, A., Fisette, T., Mcnairn, H., Daneshfar, B., Ray, S., Neetu, Rojas, O., Achard, F., Malheiros de Oliveira, Y., Mollicone, D., and Latham, J. 2017. Handbook on Remote Sensing for Agricultural Statistics. Technical report, Global Strategy to improve Agricultural and Rural Statistics (GSARS), Rome.

Dennett, M., Elston, J., and Speed, C. 1981. Rainfall and crop yields in seasonally arid West Africa. Geoforum, 12(2):203–209.

DeSalvo, K. B., Olson, R., and Casavale, K. O. 2016. Dietary guidelines for Americans. Jama, 315(5):457–458.

Devarajan, S. 2013. Africa's statistical tragedy. Review of Income and Wealth, 59(S1):S9–S15.

Devereux, S. 2009. Why does famine persist in Africa? Food security, 1(1):25–35.

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

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. 2014b. Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45):15888–15893. Doi: 10.1073/pnas.1408439111.

Diallo, Y., Gueye, M. T., Sakho, M., Darboux, P. G., Kane, A., Barthelemy, J.-P., and Lognay, G. 2013. Importance nutritionnelle du manioc et perspectives pour l'alimentation de base au Sénégal (synthèse bibliographique)/Nutritional importance of cassava and perspectives as a staple food in Senegal. A review. Biotechnologie, Agronomie, Société et Environnement, 17(4):634.

Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. Numerische mathematik, 1(1):269–271. Doi: 10.1007/bf01386390.

Direction de l'Analyse, de la Prévision et des Statistiques Agricoles. 2013. Rapport de présentation des résultats définitifs de l'enquête agricole 2012-2013.

Dong. 2011. Millet has many faces. Global Agricultural Information Network.

Duhigg, C. 2012. How companies learn your secrets. The New York Times, 16:2012.

Durlauf, S. and Fafchamps, M. 2005. Social capital. In Elsevier, editor, *Handbook of Economic Growth*, volume 1, page 1639–1699. Elsevier. Doi: 10.1016/S1574-0684(05)01026-9.

Duveiller, G. and Defourny, P. 2010. A conceptual framework to define the spatial resolution requirements for agricultural monitoring using remote sensing. Remote Sensing of Environment, 114(11):2637–2650.

D'Alessandro, S., Fall, A. A., Grey, G., Simpkin, S., and Wane, A. 2015. Senegal. Agricultural sector risk assessment. Technical report, World Bank.

Eagle, N. and Pentland, A. S. 2006. Reality mining: sensing complex social systems. Personal and ubiquitous computing, 10(4):255–268.

Eagle, N., Pentland, A. S., and Lazer, D. 2008. Mobile phone data for inferring social network structure. In *Social computing, behavioral modeling, and prediction*, pages 79–88. Springer. Doi: 10.1007/978-0-387-77672-9_10.

Eagle, N., Pentland, A. S., and Lazer, D. 2009b. Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences, 106(36):15274–15278.

Eagle, N., Macy, M., and Claxton, R. 2010a. Network diversity and economic development. Science, 328(5981):1029–1031.

Easterly, W. 2009. How the millennium development goals are unfair to Africa. World development, 37(1):26–35.

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

Enke, S. 1951. Equilibrium among spatially separated markets: Solution by electric analogue. Econometrica: Journal of the Econometric Society, pages 40–47. Doi: 10.2307/1907907.

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L119:1–88.

Fafchamps, M. 2006. Development and social capital. The Journal of Development Studies, 42(7):1180–1198. Doi: 10.1080/00220380600884126.

Fafchamps, M. and Minten, B. 2001. Social capital and agricultural trade. American Journal of Agricultural Economics, 83(3):680–685. Doi: 10.1111/0002-9092.00190.

Fafchamps, M. and Minten, B. 2002. Returns to social network capital among traders. Oxford economic papers, 54(2):173–206. Doi: 10.1093/oep/54.2.173.

FAO. 1996. Rome Declaration on World Food Security and World Food Summit Plan of Action: World Food Summit 13-17 November 1996, Rome, Italy. Technical report, Food and Agriculture Organization.

FAO. 2008. FAOSTAT: Food balance sheet.

FAO. 2013. *Food Wastage Footprint: Impacts on Natural Resources: Summary Report*. FAO.

FAO, IFAD, UNICEF, WFP, and WHO. 2017. The State of Food Insecurity in the World 2017. Building resilience for peace and food security. Technical report, Food and Agriculture Organization of the United Nations.

Fehling, M., Nelson, B. D., and Venkatapuram, S. 2013. Limitations of the Millennium Development Goals: a literature review. Global Public Health, 8(10):1109–1122.

Fletcher, T., Jütting, J., Eele, G., Greenwel, G., Klein, T., Zbiranski, T., and Keeley, B. 2015. A road map for a country-led data revolution. Technical report, PARIS21.

Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., Mueller, N. D., O'Connell, C., Ray, D. K., West, P. C., *et al.*. 2011. Solutions for a cultivated planet. Nature, 478(7369):337–342.

Frelat, R., Lopez-Ridaura, S., Giller, K. E., Herrero, M., Douxchamps, S., Djurfeldt, A. A., Erenstein, O., Henderson, B., Kassie, M., Paul, B. K., *et al.*. 2016. Drivers of household food availability in sub-Saharan Africa based on big data from small farms. Proceedings of the National Academy of Sciences, 113(2):458–463.

Frías-Martínez, V., Soto, V., Virseda, J., and Frías-Martínez, E. 2013. Can Cell Phone Traces Measure Social Development? In *NetMob*.

Friedman, H. S. 2013. Causal inference and the Millennium Development Goals (MDGs): Assessing whether there was an acceleration in MDG development indicators following the MDG declaration.

Fukuyama, F. 1995. *Trust: The social virtues and the creation of prosperity*. Free press New York.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., *et al.*. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific data, 2:150066.

Gabre-Madhin, E. Z. 2001. Market institutions, transaction costs, and social capital in the Ethiopian grain market. Technical report, International Food Policy Research Institute.

Gantz, J. and Reinsel, D. 2011. Extracting value from chaos. IDC iview, (1142):9–10.

Genesio, L., Bacci, M., Baron, C., Diarra, B., Di Vecchia, A., Alhassane, A., Hassane, I., Ndiaye, M., Philippon, N., Tarchiani, V., *et al.*. 2011. Early warning systems for food security in West Africa: evolution, achievements and challenges. Atmospheric Science Letters, 12(1):142–148.

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., *et al.*. 2014. World population stabilization unlikely this century. Science, 346(6206):234–237.

Glassman, A. and Ezeh, A. 2014. Delivering on the data revolution in Sub-Saharan Africa. Technical report, Center for Global Development.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. 2010. Food security: the challenge of feeding 9 billion people. Science, 327(5967):812–818.

Gómez, C., White, J. C., and Wulder, M. A. 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 116:55–72.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. Nature, 453(7196):779–782.

Google. 2009. Defendant's Motion to Dismiss the Amended Complaint (Boring v. Google, Inc.). Technical report.

Gordon, D. C. 2016. *The Republic of Lebanon: nation in jeopardy*, volume 18. Routledge.

Graef, F. and Haigis, J. 2001. Spatial and temporal rainfall variability in the Sahel and its effects on farmers' management strategies. Journal of Arid Environments, 48(2):221–231.

Grassegger, H. and Krogerus, M. 2017. The data that turned the world upside down.

Green, K., Kempka, D., and Lackey, L. 1994. Using remote sensing to detect and monitor land-cover and land-use change. Photogrammetric engineering and remote sensing, 60(3):331–337.

Grist, J. P. and Nicholson, S. E. 2001. A study of the dynamic factors influencing the rainfall variability in the West African Sahel. Journal of climate, 14(7):1337–1359.

Grootaert, C. 1999. Social capital, household welfare, and poverty in Indonesia. Technical report, World Bank.

Groten, S. 1993. NDVI—crop monitoring and early yield assessment of Burkina Faso. International Journal of Remote Sensing, 14(8):1495–1515.

GSMA Intelligence. 2016. The Mobile Economy 2016.

Guiro, A. T. *et al.*. 2005. Bilan de la recherche agricole et agroalimentaire au Sénégal. Technical report, Institut sénégalais de recherches agricoles (ISRA), Institut de Technologie Alimentaire (ITA), Centre de coopération internationale en recherche agronomique pour le développement (CIRAD).

Guiso, L., Sapienza, P., and Zingales, L. 2004. The role of social capital in financial development. The American Economic Review, 94(3):526–556. Doi: 10.3386/w7563.

Gutierrez, T., Krings, G., and Blondel, V. D. 2013a. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. arXiv preprint arXiv:1309.4496.

Haarsma, R. J., Selten, F. M., Weber, S. L., and Kliphuis, M. 2005. Sahel rainfall variability and response to greenhouse warming. Geophysical Research Letters, 32(17).

Hamilton, B. A. 2010. Dakar-Bamako Corridor Cost of Transport Analysis. Technical report, USAID Senegal.

Hanjra, M. A. and Qureshi, M. E. 2010. Global water crisis and future food security in an era of climate change. Food Policy, 35(5):365–377.

Hardjono, T., Shrier, D., and Pentland, A. 2016. TRUST:: DATA: A New Framework for Identity and Data Sharing. Visionary Future LLC.

Hertel, T. W. 2011. The global supply and demand for agricultural land in 2050: A perfect storm in the making? American Journal of Agricultural Economics, 93(2):259–275.

Hic, C., Pradhan, P., Rybski, D., and Kropp, J. P. 2016. Food surplus and its climate burdens. Environmental science & technology, 50(8):4269–4277.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. International journal of climatology, 25(15):1965–1978.

Hoerl, A. and Kennard, R. 1988. Ridge Regression. Encyclopedia of Statistical Sciences, 8.

Hotelling, H. 1990. Stability in competition. In *The Collected Economics Articles of Harold Hotelling*, pages 50–63. Springer.

Hutchinson, C. 1991. Uses of satellite data for famine early warning in sub-Saharan Africa. International Journal of Remote Sensing, 12(6):1405–1421.

Independent Expert Advisory Group on a Data Revolution for Sustainable Development, U. 2014. A world that counts: mobilizing the data revolution for sustainable development.

Inglada, J., Arias, M., Tardy, B., Morin, D., Valero, S., Hagolle, O., Dedieu, G., Sepulcre, G., Bontemps, S., and Defourny, P. 2015b. Benchmarking of algorithms for crop type land-cover maps using Sentinel-2 image time series. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, pages 3993–3996. IEEE.

International Poverty Centre. 2006. What is poverty? Concepts and measures. Technical report, UNDP International Poverty Centre (IPC).

International Telecommunication Union, I. 2015. Monitoring The Wsis Targets: A Mid-Term Review.

Isham, J. 2002. The effect of social capital on fertiliser adoption: Evidence from rural Tanzania. Journal of African Economies, 11(1):39–60. Doi: 10.1093/jae/11.1.39.

ITU World Telecommunication. 2016a. Key ICT indicators for developed and developing countries and the world. http://www.itu.int/en/ITU-D/Statistics/.

ITU World Telecommunication. 2016b. Key ICT indicators for developed and developing countries and the world (totals and penetration rates). http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2016/ITU_Key_2005-2016_ICT_data.xls. Accessed: 2017-01-25.

Ivanic, M. and Martin, W. 2008. Implications of higher global food prices for poverty in low-income countries. Agricultural economics, 39(s1):405–416.

Jacques, D., Waldner, F., d'Andrimont, R., Radoux, J., *et al.*. 2015. Genesis of millet prices in Senegal: the role of production, markets and their failures. In *Fourth Conference on the Analysis of Mobile Phone Datasets, NetMob*. MIT Media Labs, Boston.

Jacques, D. C., Marinho, E., d'Andrimont, R., Waldner, F., Radoux, J., Gaspart, F., and Defourny, P. 2018. Social capital and transaction costs in millet markets. Heliyon, 4(1).

Jalloh, A., Nelson, G. C., Thomas, T. S., Zougmoré, R. B., and Roy-Macauley, H. 2013. *West African agriculture and climate change: a comprehensive analysis*. International Food Policy Research Institute.

Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F., and Hlavacs, H. 2015. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. IEEE Transactions on Intelligent Transportation Systems, 16(5):2551–2572.

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

Jensen, R. 2007. The digital provide: Information (technology), market performance, and welfare in the South Indian fisheries sector. The quarterly journal of economics, pages 879–924. Doi: 10.1162/qjec.122.3.879.

Jerven, M. 2013a. Comparability of GDP estimates in Sub-Saharan Africa: The effect of Revisions in Sources and Methods Since Structural Adjustment. Review of Income and Wealth, 59(S1):S16–S36.

Jerven, M. 2013b. *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press.

Jerven, M. 2017. How much will a data revolution in development cost? Forum for Development Studies, 44(1):31–50.

Kam, S.-P., Hossain, M., Bose, M. L., and Villano, L. S. 2005. Spatial patterns of rural poverty and their relationship with welfare-influencing factors in Bangladesh. Food Policy, 30(5):551–567.

Keita, N., Frederic, V., Ferraz, C., Gallego, J., and Galmés, M. 2012. Handbook on Master Sampling Frame for Agriculture. Technical report, Global Strategy to improve Agricultural and Rural Statistics (GSARS), Rome.

Kempeneers, P. and Soille, P. 2017. Optimizing Sentinel-2 image selection in a Big Data context. Big Earth Data, 1(1-2):145–158.

Kim, T.-H. and White, H. 2003. On More Robust Estimation of Skewness and Kurtosis: Simulation and Application to the S&P500 Index. University of california at san diego, economics working paper series, Department of Economics, UC San Diego.

Kindornay, S., Bhattacharya, D., and Higgins, K. 2016. Implementing Agenda 2030: Unpacking the Data Revolution at Country Level.

Kirkpatrick, R. 2011. Data Philanthropy is Good for Business. http://www. forbes. com/sites/oreillymedia/2011/09/20/data-philanthropyis-good-for-business.

Knack, S. and Keefer, P. 1997. Does social capital have an economic payoff? A cross-country investigation. The Quarterly journal of economics, 112(4):1251–1288. Doi: 10.1162/003355300555475.

Krätke, F. and Byiers, B. 2014. Implications for the data revolution in Sub-Saharan Africa. Technical report, Paris 21.

Kummu, M. and Varis, O. 2011. The world by latitudes: A global analysis of human population, development level and environment across the north–south axis over the past half century. Applied geography, 31(2):495–507.

Lambert, M.-J., Waldner, F., and Defourny, P. 2016. Cropland mapping over Sahelian and Sudanian agrosystems: A knowledge-based approach using PROBA-V time series at 100-m. Remote Sensing, 8(3):232.

Lambert, M.-J., Blaes, X., Traoré, P. S., and Defourny, P. 2017. Estimate yield at parcel level from S2 time serie in sub-Saharan smallholder farming systems. In *Analysis of Multitemporal Remote Sensing Images (MultiTemp), 2017 9th International Workshop on the*, pages 1–7. IEEE.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. 2010. A survey of mobile phone sensing. IEEE Communications magazine, 48(9).

Laney, D. 2001. 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., *et al.*. 2009. Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915):721.

Le Soleil. 2017. Poverty Map in Senegal: A Disparity in Sharing Wealth.

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., San Miguel, M., and Ramasco, J. J. 2015. Comparing and modelling land use organization in cities. Royal Society open science, 2(12):150449.

Leonardi, U. 2008. Senegal Land Cover Mapping. Technical report.

Letouzé, E. F. 2016. *Applications and Implications of Big Data for Demo-Economic Analysis: The Case of Call-Detail Records*. Ph.D. thesis, University of California, Berkeley.

Levine, S. S., Apfelbaum, E. P., Bernard, M., Bartelt, V. L., Zajac, E. J., and Stark, D. 2014. Ethnic diversity deflates price bubbles. Proceedings of the National Academy of Sciences, 111(52):18524–18529. Doi: 10.1073/pnas.1407301111.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., and Tatem, A. J. 2012. Population distribution, settlement patterns and accessibility across Africa in 2010. PloS one, 7(2):e31743. Doi: 10.1371/journal.pone.0031743.

Lobell, D. B., Thau, D., Seifert, C., Engle, E., and Little, B. 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment, 164:324–333.

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

Lucci, P., Bhatkal, T., and Khan, A. 2018. Are we underestimating urban poverty? World Development, 103:297–310.

MacDonald, G. K., Brauman, K. A., Sun, S., Carlson, K. M., Cassidy, E. S., Gerber, J. S., and West, P. C. 2015. Rethinking agricultural trade relationships in an era of globalization. BioScience, 65(3):275–289.

Maldoff, G. 2016. How GDPR changes the rules for research. https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/. Accessed: 2018-04-20.

Martinez-Cesena, E. A., Mancarella, P., Ndiaye, M., and Schläpfer, M. 2015. Using mobile phone data for electricity infrastructure planning. arXiv preprint arXiv:1504.03899.

Matton, N., Canto, G. S., Waldner, F., Valero, S., Morin, D., Inglada, J., Arias, M., Bontemps, S., Koetz, B., and Defourny, P. 2015. An automated method for annual cropland mapping along the season for various globally-distributed agrosystems using high spatial and temporal resolution time series. Remote Sensing, 7(10):13208–13232.

Mawejje, J. and Terje Holden, S. 2014. Does social network capital buy higher agricultural prices? A case of coffee in Masaka district, Uganda. International Journal of Social Economics, 41(7):573–585. Doi: 10.1108/IJSE-03-2013-0066.

McDonald, S. 2016. Ebola: a big data disaster. Privacy, property, and the law of disaster experimentation. CIS Papers.

McNew, K. and Fackler, P. L. 1997. Testing market equilibrium: is cointegration informative? Journal of Agricultural and Resource Economics, pages 191–207.

Metcalfe, H. 2013. Mobile for Development Impact Products and Services Landscape. Technical report, GSMA.

Min, B., Gaba, K. M., Sarr, O. F., and Agalassou, A. 2013. Detection of rural electrification in Africa using DMSP-OLS night lights imagery. International Journal of Remote Sensing, 34(22):8118–8141.

Minet, J., Curnel, Y., Gobin, A., Goffart, J.-P., Mélard, F., Tychon, B., Wellens, J., and Defourny, P. 2017. Crowdsourcing for agricultural applications: A review of uses and opportunities for a farmsourcing approach. Computers and Electronics in Agriculture, 142:126–138.

Minot, N., Baulch, B., Epperecht, M., *et al.*. 2006. Poverty and inequality in Vietnam: Spatial patterns and geographic determinants. Technical report.

Mises, R. V. 1964. Chapter Ix - Analysis of Statistical Data. In Mises, R. V., editor, *Mathematical Theory of Probability and Statistics*, pages 431 – 493. Academic Press.

Misturelli, F. and Heffernan, C. 2010. The concept of poverty a synchronic perspective. Progress in Development Studies, 10(1):35–58.

Morton, T. A., Rabinovich, A., Marshall, D., and Bretschneider, P. 2011. The future that may (or may not) come: How framing changes responses to uncertainty in climate change communications. Global Environmental Change, 21(1):103–109.

Mule, T. 2012. Census coverage measurement estimation report: Summary of estimates of coverage for persons in the United States. Washington, DC: US Census Bureau.

Murray, C. J. 2007. Towards good practice for health statistics: lessons from the Millennium Development Goal health indicators. The Lancet, 369(9564):862–873.

Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. 2016. Large-scale mobile traffic analysis: a survey. IEEE Communications Surveys & Tutorials, 18(1):124–161.

Narayan-Parker, D. and Patel, R. 2000. *Voices of the poor: can anyone hear us?*, volume 1.

Ndao, M. M. and Breuer, I. M. 2013. Climate risk and food security in Senegalanalysis of climate impacts on food security and liveSenegal:. Technical report, National Agency for Civil Aviation and Meteorology and World Food Programme.

Ndiaye, M. 2007. Senegal Agricultural Situation: Country Report 2007. Global Agricultural Information Network (GAIN) Report.

Ndiaye, M. and Niang, M. 2010. De l'étude sur la transmission des fluctuations et le calcul de prix de parité à l'importation/exportation dans la sous région: cas pratique du Sénégal.

Njuguna, C. and McSharry, P. 2017. Constructing spatiotemporal poverty indices from Big Data. Journal of Business Research, 70:318–327.

Noba, K., Ngom, A., Guèye, M., Bassène, C., Kane, M., Diop, I., Ndoye, F., Mbaye, M. S., Kane, A., and Ba, A. T. 2014. L'arachide au Sénégal: état des lieux, contraintes et perspectives pour la relance de la filière. OCL, 21(2):D205.

Okwi, P. O., Ndeng'e, G., Kristjanson, P., Arunga, M., Notenbaert, A., Omolo, A., Henninger, N., Benson, T., Kariuki, P., and Owuor, J. 2007. Spatial determinants of poverty in rural Kenya. Proceedings of the National Academy of Sciences, 104(43):16769–16774.

OPHI. 2013. Country Briefing: Senegal. http://www.ophi.org.uk/wp-content/uploads/Senegal-2013.pdf?79d835.

Peña-López, I. *et al.*. 2016. World development report 2016: Digital dividends. Technical report, World Bank.

Peterson, P., Funk, C., Husak, G., Pedreros, D., Landsfeld, M., Verdin, J., and Shukla, S. 2013. The Climate Hazards group InfraRed Precipitation (CHIRP) with Stations (CHIRPS): Development and Validation. In *AGU Fall Meeting Abstracts*.

Pickett, J., Anderson, D., Bowles, D., Bridgwater, T., Jarvis, P., Mortimer, N., Poliakoff, M., and Woods, J. 2008. Sustainable biofuels: prospects and challenges. The Royal Society, London, UK.

Pinney, A. 2011. An Afghan Population Estimation. In *Snapshots of an Intervention: The Unlearned Lessons of Afghanistan's Decade of Assistance (2001–2011)*. Afghanistan Analysts Network.

Pokhriyal, N. and Dong, W. 2015. Virtual Network and Poverty Analysis in Senegal. D4D Challenge Senegal Scientific Papers, Netmob.

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

Porter, M. E. and Kramer, M. R. 2002. The competitive advantage of corporate philanthropy. Harvard business review, 80(12):56–68.

Portes, A. 2000. The two meanings of social capital. In *Sociological forum*, volume 15, pages 1–12. Springer. Doi: 10.1023/A:1007537902813.

Portes, A. 2014. Downsides of social capital. Proceedings of the National Academy of Sciences, 111(52):18407–18408. Doi: 10.1073/pnas.1421888112.

Poushter, J. 2016. Smartphone ownership and internet usage continues to climb in emerging economies. Technical report, Pew Research Center.

Putnam, R. D., Leonardi, R., and Nanetti, R. Y. 1994. *Making democracy work: Civic traditions in modern Italy*. Princeton university press.

Rao, J. N. and Molina, I. 2015. *Small area estimation*. John Wiley & Sons.

Rashid, S., Minot, N., Lemma, S., and Behute, B. 2010. Are staple food markets in Africa efficient? Spatial price analyses and beyond. In *COMESA policy seminar" Food price variability: Causes, consequences, and Policy Options*, pages 25–26.

Rasmussen, C. E. and Nickisch, H. 2010. Gaussian Processes for Machine Learning (GPML) Toolbox. J. Mach. Learn. Res., 11:3011–3015.

Rasmussen, C. E. and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. The MIT Press.

Reddy, S. G. and Kvangraven, I. H. 2015. Global Development Goals: If at All, Why, When and How? Why, When and How.

Rembold, F., Atzberger, C., Savin, I., and Rojas, O. 2013. Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sensing, 5(4):1704–1733.

Renier, C., Waldner, F., Jacques, D. C., Babah Ebbe, M. A., Cressman, K., and Defourny, P. 2015. A dynamic vegetation senescence indicator for near-real-time desert locust habitat monitoring with MODIS. Remote Sensing, 7(6):7545–7570.

Reuveny, R. 2007. Climate change-induced migration and violent conflict. Political geography, 26(6):656–673.

Rhee, J., Im, J., and Carbone, G. J. 2010. Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data. Remote Sensing of Environment, 114(12):2875–2887.

Ricciato, F., Widhalm, P., Craglia, M., and Pantisano, F. 2015. Estimating population density distribution from network-based mobile phone data. Technical report, Joint Research Centre.

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

Richards, N. M. and King, J. H. 2013. Three paradoxes of big data. Stan. L. Rev. Online, 66:41.

Rockström, J. and De Rouw, A. 1997. Water, nutrients and slope position in on-farm pearl millet cultivation in the Sahel. Plant and Soil, 195(2):311–327.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin III, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., *et al.*. 2009. A safe operating space for humanity. nature, 461(7263):472.

Rogers, D., Emwanu, T., and Robinson, T. 2006. Poverty mapping in Uganda: An analysis using remotely sensed and other environmental data. Technical report.

Samaké, O., Smaling, E., Kropff, M., Stomph, T., and Kodio, A. 2005. Effects of cultivation practices on spatial variation of soil fertility and millet yields in the Sahel of Mali. Agriculture, ecosystems & environment, 109(3):335–345.

Saramäki, J. and Moro, E. 2015. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. The European Physical Journal B, 88(6):164.

Schwab, K., Marcus, A., Oyola, J., Hoffman, W., and Luzi, M. 2011. Personal data: The emergence of a new asset class. In *An Initiative of the World Economic Forum*.

Seekell, D., D'Odorico, P., and Pace, M. 2011. Virtual water transfers unlikely to redress inequality in global water use. Environmental Research Letters, 6(2):024017.

Sen, A. 1981. *Poverty and famines: an essay on entitlement and deprivation*. Oxford university press.

Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., and Dabalen, A. 2015. Data deprivation: another deprivation to end. Technical report, World Bank.

Sghir, M. *et al.*. 2015. Rapport synthétique de l'analyse des chaines de valeur. Eléments Techniques, Economiques et Financiers pour la mise en place des Agropoles. Technical report, Organisation des Nations Unies pour le Développement Industriel (ONUDI).

Shepon, A., Eshel, G., Noor, E., and Milo, R. 2018. The opportunity cost of animal based diets exceeds all food losses. Proceedings of the National Academy of Sciences, page 201713820.

Smith, C., Mashhadi, A., and Capra, L. 2013. Ubiquitous Sensing for Mapping Poverty in Developing Countries.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. 2010a. Limits of predictability in human mobility. Science, 327(5968):1018–1021.

Soto, V., Frías-Martínez, V., Virseda, J., and Frías-Martínez, E. 2011. Prediction of Socioeconomic Levels Using Cell Phone Records. In *Proceedings of the 19th International Conference on User Modeling, Adaption and Personalization*, pages 377–388. Springer.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., *et al.*. 2017. Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface, 14(127):20160690.

Stehman, S. 2005. Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. Remote Sensing of Environment, 96(3):466–474.

Stempeck, M. 2014. Sharing data is a form of corporate philanthropy. Harvard Business Review. Available from: https://hbr. org/2014/07/sharing-data-is-a-form-of-corporatephilanthropy (accessed 26 January 2016).

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one, 10(2):e0107042. Doi: 10.1371/journal.pone.0107042.

Stuart, E., Samman, E., Avis, W., and Berliner, T. 2015. The data revolution: Finding the missing millions. Technical report, Overseas Development Institute.

Sultan, B., Baron, C., Dingkuhn, M., Sarr, B., and Janicot, S. 2005. Agricultural impacts of large-scale variability of the West African monsoon. Agricultural and forest meteorology, 128(1-2):93–110.

Sultan, B., Roudier, P., Quirion, P., Alhassane, A., Muller, B., Dingkuhn, M., Ciais, P., Guimberteau, M., Traore, S., and Baron, C. 2013. Assessing climate change impacts on sorghum and millet yields in the Sudanian and Sahelian savannas of West Africa. Environmental Research Letters, 8(1):014040.

Sundsøy, P. 2016. Can mobile usage predict illiteracy in a developing country? CoRR, abs/1607.01337.

Swenson, C., Moore, T., and Shenoi, S. 2006. GSM Cell Site Forensics. In *Advances in Digital Forensics II*, pages 259–272. Springer.

Takayama, T. and Judge, G. G. 1971. *Spatial and temporal price and allocation models*.

Tartarelli, S., d'Heureuse, N., and Niccolini, S. 2010. Lessons learned on the usage of call logs for security and management in IP telephony. IEEE Communications Magazine, 48(12):76–82.

Taylor, L. 2015. No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environment and Planning D: Society and Space, 34(2):319–336.

Taylor, R. and Kelsey, T. 2016. *Transparency and the open society: Practical lessons for effective policy*. Policy Press.

Teravaninthorn, S. and Raballand, G. 2009. *Transport prices and costs in Africa: a review of the main international corridors*. World Bank Publications.

The Economist. 2008. How to promote the spread of mobile phones among the world's poorest. http://www.economist.com/node/11465558. Accessed: 2017-01-25.

The Economist. 2014. Ebola and big data: Waiting on hold. https://econ.st/2IydaT2. Accessed: 2018-02-11.

Thornton, P. K. 2010. Livestock production: recent trends, future prospects. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1554):2853–2867.

Tibshirani, R. 1996a. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society (Series B), 58:267–288.

Tiru, M. 2014. Overview of the sources and challenges of mobile positioning data for statistics. In *International Conference on Big Data for Official Statistics, Beijing*.

Tollens, E. F. 2006. Market information systems in sub-Sahara Africa challenges and opportunities. In *the International Association of Agricultural Economists Conference, Gold Coast, Australia*.

Tucker, C. J., Vanpraet, C. L., Sharman, M., and Van Ittersum, G. 1985. Satellite remote sensing of total herbaceous biomass production in the Senegalese Sahel: 1980–1984. Remote sensing of environment, 17(3):233–249.

UN. 2009. Main results of the UNECE-UNSD survey on the 2010 round of population and housing censuses. Technical report, Economic and Social Council.

UN General Assembly. 1948. Universal declaration of human rights. UN General Assembly.

UN General Assembly. 2000. United Nations millennium declaration.

UNDP. 2017. Human Development Report 2016-Human Development for Everyone. Technical report, United Nations Development Programme.

UNFPA, F. 2014. Population Estimation Survey 2014 for the Pre-War Regions of Somalia. Available at somalia.unfpa.org/sites/default/files/pub-pdf/Population-Estimation-Survey-of-Somalia-PESS-2013-2014.pdf. Accessed March 22, 2018.

United Nations. 2015. *The millennium development goals report 2015*. United Nations Publications, Department of Economic and United Nations and Department of Public Information.

Valerio, D., D'Alconzo, A., Ricciato, F., and Wiedermann, W. 2009. Exploiting cellular networks for road traffic estimation: a survey and a research roadmap. In *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pages 1–5. IEEE.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K.-I., Kondor, I., and Barabási, A.-L. 2006. Modeling bursts and heavy tails in human dynamics. Physical Review E, 73(3):036127.

Vista, B. M. and Murayama, Y. 2011. Spatial determinants of poverty using GIS-based mapping. In *Spatial analysis and modeling in geographical transformation process*, pages 275–296. Springer.

Vulnerability Analysis and Mapping unit. 2000-2014. Food and Commodity Prices Data.

Waldner, F., Canto, G. S., and Defourny, P. 2015a. Automated annual cropland mapping using knowledge-based temporal features. ISPRS Journal of Photogrammetry and Remote Sensing, 110:1–13. Doi: 10.1016/j.isprsjprs.2015.09.013.

Waldner, F., Fritz, S., Di Gregorio, A., and Defourny, P. 2015b. Mapping priorities to focus cropland mapping activities: Fitness assessment of existing global, regional and national cropland maps. Remote Sensing, 7(6):7959–7986.

Waldner, F., Hansen, M. C., Potapov, P. V., Löw, F., Newby, T., Ferreira, S., and Defourny, P. 2017a. National-scale cropland mapping based on spectral-temporal features and outdated land cover information. PloS one, 12(8):e0181911.

Waldner, F., Jacques, D. C., and Löw, F. 2017b. The impact of training class proportions on binary cropland classification. Remote Sensing Letters, 8(12):1122–1131.

Wallach, D., Mearns, L. O., Rivington, M., Antle, J. M., and Ruane, A. C. 2015. Uncertainty in agricultural impact assessment. In *Handbook of Climate Change and Agroecosystems: The Agricultural Model Intercomparison and Improvement Project Integrated Crop and Economic Assessments, Part 1*, pages 223–259. World Scientific.

Wang, X., Feng, H., Xia, Q., and Alkire, S. 2016. On the relationship between income poverty and multidimensional poverty in China. OPHI Working Papers, (101):1–21.

Watmough, G. R., Atkinson, P. M., Saikia, A., and Hutton, C. W. 2016. Understanding the Evidence Base for Poverty–Environment Relationships using Remotely Sensed Satellite Data: An Example from Assam, India. World Development, 78:188–203.

Webb, P., Coates, J., Frongillo, E. A., Rogers, B. L., Swindale, A., and Bilinsky, P. 2006. Measuring household food insecurity: why it's so important and yet so difficult to do. The Journal of nutrition, 136(5):1404S–1408S.

Weidmann, N. B. and Schutte, S. 2016. Using night light emissions for the prediction of local wealth. Journal of Peace Research, page 0022343316630359.

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

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

WFP. 2012a. Food Security Assessment in Zones at Risk: Harvest 2011-2012. Technical report.

WFP. 2012b. WFP Launches Targeted Food Distributions in Senegal. http://www.wfp.org/stories/wfp-launches-targeted-food-distributions-senegal. Accessed: 2018-04-20.

White, J. and Wells, I. 2002. Extracting origin destination information from mobile phone data. In *Eleventh International Conference on Road Transport Information and Control*. IET.

Wodon, Q. and Zaman, H. 2009. Higher food prices in Sub-Saharan Africa: Poverty impact and policy responses. The World Bank Research Observer, 25(1):157–176.

Woolcock, M. and Narayan, D. 2000. Social capital: Implications for development theory, research, and policy. The world bank research observer, 15(2):225–249. Doi: 10.1093/wbro/15.2.225.

World Bank. 2012. *Information and Communications for Development 2012: Maximizing Mobile*. World Bank Publications.

World Bank. 2016. Poverty and Shared Prosperity 2016: Taking on Inequality. Technical report, World Bank.

World Bank. 2017. Monitoring Global Poverty: Report of the Commission on Global Poverty. Technical report, World Bank.

Xiong, J., Thenkabail, P. S., Gumma, M. K., Teluguntla, P., Poehnelt, J., Congalton, R. G., Yadav, K., and Thau, D. 2017. Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. ISPRS Journal of Photogrammetry and Remote Sensing, 126:225–244.

Youyou, W., Kosinski, M., and Stillwell, D. 2015. Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4):1036–1040.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B, 67:301–320.

List of Publications

Peer-Reviewed Publications

Jacques, D. C. and Defourny, P. 2018. Accuracy requirements for early warning of crop production in Senegal. Royal Society Open Science. *Under Review.*

Fritz, S., See, L, Laso Bayas, J.C., Waldner, F., **Jacques, D. C.**, Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J.,Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu B., Yan, N., Lianghzi, Y., Gilliams, S., Mucher, S., Moorthy, I., McCallum, I. 2018. A Comparison of Global Agricultural Monitoring Systems and Current Information Gaps. Agricultural Systems, *Under Review.*

Zufiria, P. J., Pastor-Escuredo, D., Ubeda-Medina, L. A., Hernandez-Medina, M. A., Barriales-Valbuena, I., Morales, A. J., **Jacques, D. C.**, Nkwambi, W., Diop, M. B., Quinn, J., Hidalgo-Sanchis, P., Luengo-Oroz, M. 2018. Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. Plos One, 13(4): e0195714.

Jacques, D. C., Marinho, E., d'Andrimont, R., Waldner, F., Radoux, J., Gaspart, F., and Defourny, P. 2017. Social capital and transaction costs in millet markets. Heliyon, 4(1).

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

Tennant, J. P.*, Dugan, J. M.*, Graziotin, D.*, **Jacques, D. C.***, Waldner, F.*, Mietchen, D.*, Elkhatib, Y.*, Collister, L. B.*, Pikas, C. K.*, Crick, T.*, et al.*. 2017. A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Research, 6.

Radoux, J.* , Chomé, G.* , **Jacques, D. C.*** , Waldner, F.* , Bellemans, N.* , Matton, N.* , Lamarche, C.* , d'Andrimont, R.* , and Defourny, P. 2016. Sentinel-2's Potential for Sub-Pixel Landscape Feature Detection. Remote

* equally contributed to this work.

Sensing, 8(6):488.

Tennant, J. P.*, Waldner, F.*, **Jacques, D. C.***, Masuzzo, P.*, Collister, L. B.*, and Hartgerink, C. H.* 2016. The academic, economic and societal impacts of Open Access: an evidence-based review. F1000Research, 5.

Ardekani, M. R. M., **Jacques, D. C.**, and Lambot, S. 2016. A Layered Vegetation Model for GPR Full-Wave Inversion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(1):18–28.

Renier, C., Waldner, F., **Jacques, D. C**., Babah Ebbe, M. A., Cressman, K., and Defourny, P. 2015. A dynamic vegetation senescence indicator for near- real-time desert locust habitat monitoring with MODIS. Remote Sensing, 7(6):7545–7570.

Jacques, D. C., Kergoat, L., Hiernaux, P., Mougin, E., and Defourny, P. 2014. Monitoring dry vegetation masses in semi-arid areas with MODIS SWIR bands. Remote Sensing of Environment, 153:40–49.

Conference Papers

Jacques, D. C., Marinho, E., d'Andrimont, R., Waldner, F. and Radoux, J. 2015. Genesis of millet prices in Senegal: The role of production, markets and their failures. Fourth Conference on the Analysis of Mobile Phone Datasets, NetMob 2015 (MIT Media Labs, Boston).

Ardekani, M. R., Neyt, X., Nottebaere, M., **Jacques, D. C.** and Lambot, S. 2014. GPR data inversion for vegetation layer. 15th International Conference on Ground Penetrating Radar (GPR) (Square Brussels Meeting Centre, Brussels).

[5]contributed equally to this work