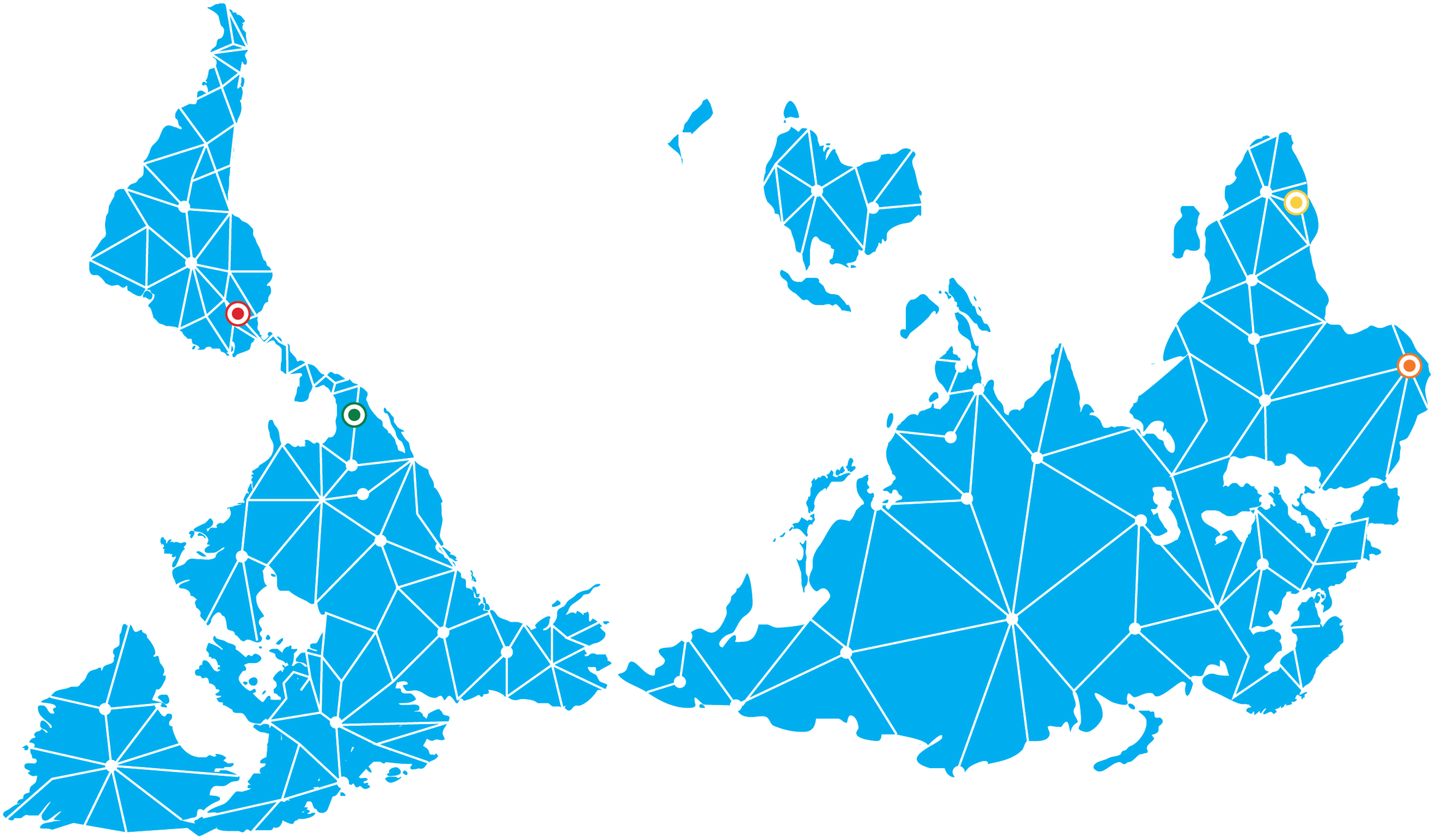
BIG DATA
TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES
2019

DATA-POP ALLIANCE



1. Socio-physical Vulnerability to Flooding in Senegal.
2. Characterizing and analyzing urban dynamics in Bogota.
3. Understanding the Relationship between Short and Long Term Mobility.
4. Large-Scale Mapping of Citizens' Behavioral Disruption in the Face of Urban Crime Shocks.

DATA-POP ALLIANCE

AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

Introduction

These four research papers were developed in collaboration with and funded by the Agence française de développement (AFD) between 2016 and 2019 under a joint program with Data-Pop Alliance and research partners titled "*Strengthening the evidence-base for leveraging Big Data to address global development challenges*".

The starting point was the realization, in 2016, that 'we' still had little hard academic evidence on Big Data's potential to understand and address some of the world's most pressing societal challenges—violence, vulnerability, inequality, mobility, among others. Even today, in 2019, it is difficult to identify one "killer case" that unambiguously and powerfully demonstrates how computation analysis of 'our' data—of these little digital breadcrumbs we leave behind—can shed light on and improve human processes and outcomes, in other words, our lives.

For those of us, development researchers, practitioners and 'experts', who argue that analyzing these data can and should be undertaken with appropriate safeguards and systems, it was and remains critical to present evidence supporting our claims— especially given concerns over privacy, widening power imbalances, human rights violations, etc.—and to provide pointers on how to mitigate those risks. In turn, these kind of projects are opportunities to improve, facilitate and strengthen our research practices and partnerships to both reflect and promote key determinants of sustainable development, including smoother, fairer and safer access to data, and tighter links between analysts, local decision-makers and communities.

We designed this research program and papers with a couple of different objectives and criteria in mind. First, we wanted to focus on various development challenges in different local contexts to ensure relevance. Second, we sought to work with trusted partners, so as to ensure academic quality and data security. After considering many options and combinations, we decided to analyze the following topics: climate resilience and flood vulnerability in Senegal, led by Cloud to Street; urban crime dynamics in Colombia with researchers from Fondazione Bruno Kessler (FBK) in Trento; migration in Senegal and Namibia with Flowminder Foundation, and gender impacts of crime in Mexico with researchers from MIT Media Lab. The development and coordination of these research papers involved core staff members of Data-Pop Alliance in New York City, Cambridge, Bogota and Mexico City.

Individually, these papers show specific cases and examples of how computational analysis of behavioral data, combined with other datasets, can paint a finer-grained, more complex and dynamic picture of human reality than 'traditional' data allows. They also state limitations and call for prudence when attempting to draw system-wide conclusions or design policies on the basis of their findings—including because of potential flaws and biases in traditional data used to train some of their models. Collectively, they sketch the contours of a world where public decisions, in the form of policies and programs, may someday be designed, implemented and evaluated using the best available data and approaches, so as to ground them more firmly in facts than is currently the case.

For DPA, additional projects, papers and partnerships have followed directly and indirectly from this program, building on their results and lessons. In Mexico, we are developing research partnerships and activities with Oxfam Mexico ("DataMex"), UNODC and different private data providers such as Banorte, focusing on inequality and criminality. In Colombia, we have been implementing since 2017 a major project funded by The InterAmerican Development Bank (IDB) called "Ciudata Segura" with 6 city governments, that aims to expand and deepen the analysis conducted in our initial Bogota-study—through qualitative analysis and interviews. In Senegal, DPA has been a co-founder of the Open Algorithms (OPAL) project since 2017, funded primarily by AFD, which seeks to develop technological systems and governance standards to allow sensitive data such as call-detail records to be accessed and analyzed in a safer, more ethical and more sustainable manner than has been the case so far—including, it has to be noted, for these four papers.

In the next few months and years, we will strive to contribute further to the strengthening and 'structuring' around appropriate technological and governance standards of this (still) nascent field of research and practice called "Big Data (and increasingly AI) for social good" through our three pillars: research, data literacy and strategic support, to address real-world development challenges in local contexts. We hope these papers spur interests, questions, ideas, among their readers.

BIG DATA
TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES
2018

DATA-POP ALLIANCE

SOCIO-PHYSICAL VULNERABILITY TO FLOODING IN SENEGAL: AN EXPLORATORY ANALYSIS WITH NEW DATA & GOOGLE EARTH ENGINE

Bessie Schwarz, Cloud to Street
Beth Tellman, Cloud to Street
Jonathan Sullivan, Cloud to Street
Catherine Kuhn, Cloud to Street
Richa Mahtta, Cloud to Street
Bhartendu Pandey, Cloud to Street
Laura Hammett, Cloud to Street
Gabriel Pestre, Data Pop Alliance

Funded by:

DATA-POP ALLIANCE
Cloud to Street
AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

SOCIO-PHYSICAL VULNERABILITY TO FLOODING IN SENEGAL: AN EXPLORATORY ANALYSIS WITH NEW DATA & GOOGLE EARTH ENGINE

Bessie Schwarz, Cloud to Street
Beth Tellman, Cloud to Street
Jonathan Sullivan, Cloud to Street
Catherine Kuhn, Cloud to Street
Richa Mahtta, Cloud to Street
Bhartendu Pandey, Cloud to Street
Laura Hammett, Cloud to Street
Gabriel Pestre, Data Pop Alliance

Prepared for Dr. Thomas Roca, Agence Française de Développement Beth &Bessie Inc.(Doing Business as Cloud to Street)

Abstract: Each year thousands of people and millions of dollars in assets are affected by flooding in Senegal; over the next decade, the frequency of such extreme events is expected to increase. However, no publicly available digital flood maps, except for a few aerial photos or post-disaster assessments from UNOSAT, could be found for the country. This report tested an experimental method for assessing the socio-physical vulnerability of Senegal using high capacity remote sensing, machine learning, new social science, and community engagement. This scientific approach to flood analysis developed in this report is far faster and more responsive than traditional flood mapping, but is only a fraction of the cost. First Cloud to Street's customized water detection algorithms were run for several publicly available satellites (MODIS, Landsat) to map major floods from the last 30 years and second machine learning approach to hydrology in Google Earth Engine was trained on the maps of past floods. Third, a Principal Component Analysis, run on customize designed Census Senegal variables, revealed five underlying dimensions of social vulnerability to flooding. Overall, the research predicts a floodplain in Senegal of 5,596 km 2 , 30% of which is high-risk zone where over 97,000 people live. Approximately 5 million people live in the 30 arrondissements that have very high social vulnerability profiles compared to other arrondissements. In a future version, this risk platform could be set to stream satellite imagery public and other sensors, so that the vulnerability analysis for Senegal can be updated with the mere refresh of a browser page – no downloading is required.

TABLE OF CONTENTS

1. Introduction: New threats, New Innovations, and Bringing New Vulnerability Analysis to Senegal

The risk and impact of natural hazards is increasing faster than at any other time in human history due climatic change and major population migration. Floods account for almost half of all weather-related disasters over the last two decades, affecting 2.3 billion people. Each year, millions of people and billions of dollars' worth of assets around the world are affected by floods, which cause more economic, social, and humanitarian losses worldwide than any other type of hazard (UNISDR, 2015). By 2030, the number of people and GDP exposed to flooding will double as a result of climate change and population migration (World Resources Institute, 2015). This is simply not a level of risk that the world is able to absorb. Today, disaster management is overwhelmingly a reactionary practice, with 87% of disaster funding spent on immediate response (Keating et al., 2014). In developing countries, 80% of people exposed to flood, including countries like Senegal, do not have flood insurance (Keating et al., 2014). The authors estimate that many people living in the floodplain today would likely not appear on an official flood map and therefore cannot fully prepare or be protected by government.

Traditional vulnerability models, used to assess river geomorphology and hydrology, are physically based, time-intensive, costly, do not incorporate social dimensions of the vulnerability to disasters, and are not responsive to or inclusive of communities. First and foremost, the process of building new hydrologic and hydraulic models to reflect geomorphic changes can costs millions of dollars and take years to calibrate and validate. These barriers can prohibit development of hydrologic models in a timely manner, especially considering that new assessments are required each time a new major event comes through the area.

Second, traditional flood risk analysis requires a significant amount of expert data that is rarely publically available or easy to create. Initiatives like CLUVA (Climate Change and Urban Vulnerability in Africa) have invested in building GIS systems in Saint-Louis, Senegal's second largest city, but program managers note that data gaps remain regarding stakeholder knowledge about flood vulnerability between communities, regional, and state governments. While it is recognized by Senegalese academics that that "appropriate information presented in appropriate ways can have a catalytic role in risk prevention" (Diagne, 2007), the authors could find no publicly available digital flood maps, save for a few aerial photos or post-disaster assessments from UNOSAT (UN Satellite Service).

In addition, most hydrologic models do not consider the social dimensions of vulnerability, an equally important element for disaster response and preparedness. Finally, with a few key exceptions like Building Resilience and Adaptation to Climate Extremes and Disasters (BRACED) and a select set of other projects, there are few good models of inclusive and genuine community engagement around resilience and vulnerability assessment in Senegal. Further, there are arguably no approaches that integrate community input and treat local knowledge and expertise with the same value afforded to external scientific assessments of vulnerability in the area.

Despite a relative lack of risk data and vulnerability modeling for Senegal, it is critical to understand what makes the area socially and physically vulnerable to flooding, especially as climatic changes are likely to exacerbate this hazard. Flood risk is constantly changing in Senegal due to changing climate and urban settlement. As in much of the Sahel, Senegal has experienced a history of highly uncertain climatic conditions, varying between cycles of drought to eras of frequent and severe flooding. After several very dry decades between 1968–1997, regional Senegalese climate has shown a 35% increase in average rainfall between 2000 and 2005 (Nicholson 2005). In addition to changing climate, Senegal has undergone significant land use change triggered by extreme drought in the 1970s, 80s, and 90s which forced rural populations into urban areas (Goldsmith, Gunjal, & Ndarishikanye, 2004). The peak urbanization rate of Senegal's capital, Dakar, was estimated around 7-8%, and 44% of Senegalese currently live in urban areas (Mbow, Diop, Diaw, & Niang, 2008).

As the frequency of intense flood events increases, the results of rapidly changing human and natural dynamics in Senegal have increased vulnerability to floods. In 2005, continuous heavy rains from August into early September caused flooding in Dakar, leading to 46 deaths, a cholera epidemic, and the evacuation of 60,000 people (Tschakert, Sagoe, Ofori-Darko, & Codjoe, 2010). Again in 2009, Dakar floodwaters destroyed 30,000 homes, which affected over a half a million people and resulted in $44.5 billion (USD) in damage and loss. In 2012, another catastrophic flood devastated already fragile public infrastructure and contaminated over 7,700 drinking water sources. The United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA) found that between 100,000 and 300,000 Senegalese are affected by floods, including in rural areas and cities such as Saint-Louis and Kaolack (UNOCHA, 2013). These floods were major disasters for the region, not just because of the physical threat, but also because of the social, economic, and political conditions of the people and communities that were affected. The social dimensions of Senegal are rapidly changing with urbanization and climate change, which highlights the need for dynamic flood vulnerability assessments.

Borrowing from the Global Resilience Partnership and Zurich Flood Resilience Program, this report defines flood resilience as enabling sustained development of human, financial, natural, social, and physical capital over time. Flood resilience is not an endpoint, but represents an evolving effort to adapt as flood vulnerability shifts with climate, land use, economic, and demographic changes. A flood resilient society must be able to learn from the consequences of its own adaptation efforts as well as changes in vulnerability due to internal (land use, levee construction) and external (storm frequency) forces. This process of "learning" will only build resilience if knowledge about the changing system translates into more transparent and democratic interventions (Pahl-Wostl, Becker, Knieper, & Sendzimir, 2013). Different types of knowledge about vulnerability can come from both analysis by researchers and experiences of people in flood-affected communities. New technology can combine these disparate but complementary sources.

Lack of information means that practitioners lack clear guidance on how to prioritize spending and other resources, and how and where to design programs to increase preparedness and reduce the degree of impact. Accessibility of this information is seen as vital to enhancing people's capacity to deal with the impacts of climate change (Hellmuth, Mason, Vaughan, Van Aalst, & Choularton, 2011; T. Mitchell et al., 2010). Likewise, the production of climate information for decision-making is increasingly being seen as an entry point for joining up work on climate change adaptation, disaster risk reduction, and development in climate-sensitive places (Ahmed, Kodijat, Luneta, & Krishnamurthy, 2015; Foresight, 2012). Heavy investments in drainage and infrastructure systems have been funded by the World Bank ($90 million USD) through the Storm Water Management and Climate Change Adaptation Project. However, gaps remain in coordination and communication between key stakeholders, residents, and government officials, and between different government agencies (Diagne, 2007; Vedeld, Coly, Ndour, & Hellevik, 2015).

This report focuses on the preparedness stage of the disaster cycle and seeks to address critical information gaps for answering the question of where and how to prepare Senegal for extreme flooding today and in years to come.

Climate information generated through monitoring and analysis activities is becoming an integral part of risk management and resilience programming. Fortunately, the abundance of cheap physical and digital sensors, data-collective satellites, and higher capacity computing power has created a wealth of data at finer resolution, faster speed, and lower cost than previously imaginable. An open, big data revolution enables scientists to understand more than ever about disasters and create insights at the speed and scale needed to make practical decisions for more adequate disaster management. This new resource, however, requires new scientific methods.

This report combines new big data analysis tools with the best available rapid assessment tools in social and physical science to explore the potential for understanding and addressing information gaps about flood risk in Senegal. Streaming satellite imagery available in Google Earth Engine (GEE) is one alternative, which can generate flood vulnerability maps quickly and cheaply for immediate planning and decision-making after a flood event, while more precise hydrologic models are developed. GEE is a geographic data repository coupled with a cloud-computing platform that provides access to the historical library of public satellite imagery and other scientific map products and analytical tools for the development of scientific algorithms. The GEE platform offers unique benefits for vulnerability assessments in flood prone developing countries for three primary reasons: 1) the amount of data it stores and provides access to, 2) its high-volume data processing capability, and 3) the use of a web browser interface. GEE's data catalog is a multi-petabyte archive of georeferenced datasets essential for disaster assessment and prediction, including images from earth observing satellites (e.g. Landsat, MODIS) and airborne sensors, weather and climate datasets, digital elevation models, and others. GEE can process these high-volume datasets extremely quickly by parallelizing the processing among thousands of central processing units (CPU). Finally, this analytical power is accessible from any computer with a good internet connection, allowing regional to global analyses to be run even on low-configuration desktop computers, evading the need for expensive software, processing, or data management systems. Lastly, the use of a web browser interface allows users to share data and analyses immediately by sending out a simple browser link. The license to use GEE is currently free for scientific, governmental, and even some commercial use.

We leveraged the modeling capabilities of GEE and R to assess the current geomorphology and hydrology of the region based on satellite remote sensing data. Furthermore, the authors used state-of-the-art tools and methods to assess social vulnerability of the region based on multi-source socio-economic data. A critical piece of predicting future change of floods and preparing for this threat is understanding where floods have occurred in the past and what kind of mitigation investments have been successful. Flood inundation data are also a necessary input for the new, data-driven hydrology methods developed by Cloud to Street and adapted to Senegal for this report. Therefore the authors first built a historic flood inventory for Senegal based on two multi-decade satellite data repositories: MODIS and Landsat. This is described in Chapter 2a. Next, the authors use these past floods as training data for a machine learning model in five priority watersheds in order to estimate the probably floodplain in those places. This is described in Chapter 2b. Chapter 3 details the social vulnerability of Senegal that the authors created with a sample of raw Senegalese Census data provided exclusively to this research team and its partners.

The results estimate the number and nature of major floods that have occurred in Senegal in the recent past and predict which parts of the country and population are at risk from future extreme flood. It also describes the main social conditions expected to lead to more loss and which areas/communes have the most vulnerable populations.

This risk information can help to more quickly answer questions regarding which areas will be affected hardest and why, and where government should spend its limited resources for disaster mitigation and resilience. In addition to the technical strengths, the tool affords impressive communication capabilities for decision-makers. Not only are the resulting maps highly engaging, easy to understand, and interactive, but the results are presented with the generally recognizable Google Maps base layer.

Just as new algorithms and scientific methods are required to harness big data, new approaches to management – how the authors govern and engage communities in resilience – are required for applying these insights in order to take full advantage of the insights produced by the tools. Providing responsive flood vulnerability maps can play an important role in shifting disaster mitigation efforts to where they are most needed, and engaging local decision makers in future generations of the work to tailor their own tool-building, as outlined by the final section, will have the potential to transform disaster management. Thereby, the authors argue that this combination of big data and community input has the power to turn big data on its head, equipping non-experts with data and capacities rather than just extracting and crunching data from people. The localized science and analysis can help individuals understand the climate crisis and take control when preparing and responding to hazards. The platform streams the most recent satellite data collected, and so analysis can be updated with the mere refresh of a browser page – no downloading is required. In regions undergoing rapid land-use change like Senegal, GEE's constantly up-to-date data catalog can provide critically responsive analysis. This responsive analysis, combined with biophysical and social vulnerability assessments, provides actionable information of where to focus investments in disaster resilience. The alternative – waiting for hydrologic model updates from scientific experts – may lead to slow results and an information gap at times when timely information is sorely needed.

In following chapters, this report aims to: 1) holistically assess the current threat from floods and 2) outline the opportunities and limitations of these new approaches by understanding vulnerability in Senegal.

2. Biophysical risk: history of flood in Senegal (Chapter 2a)
3. Biophysical risk: the hydrology of the landscape (Chapter 2b)
4. Social vulnerability to disaster (Chapter 3)
5. Combined socio-physical vulnerability (Chapter 4)
6. Participatory engagement for flood resilience (Chapter 5)

This report builds the foundations of a tool to assess biophysical and social dimensions of risk that is flexible enough to include adjustments by local experts with knowledge and context of important variables of flood risk in their region. When fully built, this tool can also be used to analyze how vulnerability changes over time by running the model over specific years and months, when land use, geomorphology, and human settlement patterns may have shifted. It could dynamically identify population and infrastructure at risk for flooding by drawing on open global satellite data, the national census, mobile phone call detail records, and the crowd. The tool, designed for governments, residents, communities, aid agencies, and researchers alike, is deeply rooted in three key strategies for transformation: human-centered scientific modeling, community-based learning, and government-level development impact. Our model relocates resilience into the hands of communities and reshapes traditional scientific modeling to be inclusive of those traditionally not engaged in the knowledge creation process. The co-produced vulnerability tool will be rooted in community needs, but our framework is designed to complement and integrate with existing resilience efforts at the national level.

2.a Biophysical Risk: Building a Historical Flood Database in Senegal

There is no consistent data source for flooding for the globe. Flood data is usually collected on an event-by-event or country-by-country basis and the only collection of geospatial flood events that is global and historic in nature is the Dartmouth Flood Observatory (DFO), which maintains an inventory of major historical flood events. While this database offers useful data for each flood event from 1985 to present, such as estimates of flood size, number of people affected, and approximately 200 mapped floods (1999-2011) for various countries, these data sources do not culminate in a dataset robust enough to detect regional trends and drivers of changes in flood behavior. There is currently no existing spatial data for flood events in Senegal in the DFO database. The lack of spatial flood data prevents the hydrology community from being able to scale inundation prediction maps, including application of machine learning techniques that could inform mitigation and adaptation programs.

We created of list of flood events in Senegal that could be gathered from publicly available information sources. These sources include: existing databases, academic articles, institutional reports, and news articles. The principal data source used to identify the occurrence of historical flood events in Senegal was the DFO database where 7 floods were identified (Table 1). Several additional sources of information corroborated the information found within the DFO database, including UNITAR's Operational Satellite Applications Programme (UNOSAT) and Copernicus Emergency Management Service (EMS) . Additional sources that yielded evidence of additional flood occurrence included news briefs from Building Resilience and Adaptation to Climate Extremes and Disasters (BRACED) of heavy rain events in 2015 that caused flooding in Dakar and Saint-Louis (Building Resilience and Adaptation to Climate Extremes and Disasters, 2015).

The results of this preliminary analysis that relied upon publicly available information are summarized in Table 1. Our analysis does not represent an exhaustive list of floods within Senegal and in future collaboration with local partners the authors would seek to add further detail.

2a.1. Historical Flood Events

There is no consistent data source for flooding for the globe. Flood data is usually collected on an event-by-event or country-by-country basis and the only collection of geospatial flood events that is global and historic in nature is the Dartmouth Flood Observatory (DFO), which maintains an inventory of major historical flood events.[1] While this database offers useful data for each flood event from 1985 to present, such as estimates of flood size, number of people affected, and approximately 200 mapped floods (1999-2011) for various countries, these data sources do not culminate in a dataset robust enough to detect regional trends and drivers of changes in flood behavior. There is currently no existing spatial data for flood events in Senegal in the DFO database. The lack of spatial flood data prevents the hydrology community from being able to scale inundation prediction maps, including application of machine learning techniques that could inform mitigation and adaptation programs.

We created of list of flood events in Senegal that could be gathered from publicly available information sources. These sources include: existing databases, academic articles, institutional reports, and news articles. The principal data source used to identify the occurrence of historical flood events in Senegal was the DFO database where 7 floods were identified (Table 1). Several additional sources of information corroborated the information found within the DFO database, including UNITAR's Operational Satellite Applications Programme (UNOSAT)[2] and Copernicus Emergency Management Service (EMS)[3]. Additional sources that yielded evidence of additional flood occurrence included news briefs from Building Resilience and Adaptation to Climate Extremes and Disasters (BRACED) of heavy rain events in 2015 that caused flooding in Dakar and Saint-Louis (Building Resilience and Adaptation to Climate Extremes and Disasters, 2015).

The results of this preliminary analysis that relied upon publicly available information are summarized in Table 1. Our analysis does not represent an exhaustive list of floods within Senegal and in future collaboration with local partners the authors would seek to add further detail.

Register #	Detailed Locations	Date Began	Date Ended	Affected (km^2)	Source
C2S0001	Dakar, Ngor, Saint-Louis	8/6/2015	11/1/2015	N.D.	BRACED
DFO3971	Dakar, major cities in interior	8/24/2012	8/29/2012	79242.7	DFO, UNOSAT
DFO3531	Dakar	8/24/2009	8/26/2009	8510.69	DFO, UNOSAT
DFO3180	Senegal River Valley; Mauritania - Gorgol region - Maghama, Mbout, Kaedi Assaba - Barkeol, Kankossa; Senegal - Thiès, Louga, Matam, Kaolack, Tamba and Dakar	8/31/2007	9/20/2007	167997.63	DFO
DFO2729	Dakar area.	8/20/2005	9/10/2005	333.207	DFO
DFO2315	Senegal - Northern Kanel region, Central Nioro region, Matam region. Kaolack, Kaffrine. Assaba, Gorgol, Brakna and Adrar; Southeastern Mauritania - Affole area, Timbedra Gambia - Upper River Division; Guinea-Bissau - eastern area. Bafata and Gabu provinces. Geba river valley	8/9/2003	11/5/2003	78728.11	DFO
DFO1866	Saint-Louis Region - Districts: Podor, Dagana and Matam. Louga Region - Districts: Kacbacmer, Linguaure and Louga	1/9/2002	1/12/2002	62705.79	DFO
DFO1008	Southwest Region: Nouakchott	9/24/1995	10/8/1995	47808.22	DFO

Table 1: List of historical flood events occurring within Senegal.

[1] The Dartmouth Flood Observatory conducts global remote sensing-based fresh water measurement and mapping in "near real time" and records such information into a permanent archive. http://floodobservatory.colorado.edu/

[2] The UNOSAT Flood Portal provides free access to satellite-derived flood data in GIS vector format. The portal includes data for flood events occurring since 2007 for which UNOSAT did satellite image analysis. http://floods.unosat.org/geoportal/catalog/main/home.page

[3] The Copernicus Emergency Management Service platform allows "users" to provision satellites within hours or days for disaster response. The results of these "activations" are published on Copernicus EMS. http://emergency.copernicus.eu/mapping/list-of-activations-rapid

2a.2. Mapping Historical Flood Events

The identification of historical flood events and dates of their occurrence allows for the utilization of a large library of earth-observing satellite sensors for flood detection. There are a number of satellite missions available for earth observation including: the Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat missions 1-8, and most recently Sentinel-1 (operation late-2014). Each sensor has unique advantages and challenges but in general image collection frequency, spatial resolution, and spectral resolution define the utility of each sensor (Table 2). The two MODIS satellites, Terra and Aqua, have been used extensively to develop a number of flood algorithms (Boschetti, Nutini, Manfron, Brivio, & Nelson, 2014; Feng et al., 2012; Islam, Bala, & Haque, 2009; Xiao et al., 2006) given that the mission produces global coverage every one to two days, making it ideal for rapid response to flood events. However, these sensors have notoriously low spatial resolution (250 meters per pixel, and the flood must cover the entire pixel to be detected). On the other hand, Landsat satellites are higher-resolution (30 meters per pixel) but have a return period of 16-days, making the coincidence of flood events and imagery rare. Still, a number of water detection algorithms have been developed for Landsat sensors with a few applications to flood extent (Chignell, Anderson, Evangelista, Laituri, & Merritt, 2015; Donchyts, Schellekens, Winsemius, Eisemann, & van de Giesen, 2016; Feyisa, Meilby, Fensholt, & Proud, 2014; Yang et al., 2014). Lastly, Sentinel-1, a synthetic aperture radar (SAR) sensor, is able to address the common challenge of clouds that obfuscate areas of analysis and limit the utility of both MODIS and Landsat. The Sentinel-1 technology is relatively new, having been launched in late-2014, which limits the historical reach of this data source; however, development of flood detection algorithms has been generated for SAR technologies in general based on private satellites (Martinis, Twele, Strobl, Kersten, & Stein, 2013; Martinis, Twele, & Voigt, 2009; Mason, Giustarini, Garcia-Pintado, & Cloke, 2014).

For the purposes of flood detection, there is a suite of satellites to choose from that, together, can overcome the respective limitations of each. GEE is an ideal computing platform to build a database of historical floods that requires the fusion of multiple satellite imagery data sources. GEE brings together the full libraries of MODIS, Landsat, and Sentinel-1, and provides the computational power to integrate these products over the historical stack of imagery. The following is a description of the methods used to detect floods across each type of satellite sensor and the benefits and shortcomings of each.

Agency	Sector Name	Operational	Spectral Resolution	Spatial Resolution	Image Extent	Return Period
National Aeronautics and Space Administration (NASA)	Terra Moderate Resolution Imaging Spectroradiometer (Terra – MODIS)	2000 – Present	36 spectral bands	250m (bands 1-2) 500m (bands 3-7) 1000m (bands 8-36)	Global	Daily
National Aeronautics and Space Administration (NASA)	Aqua Moderate Resolution Imaging Spectroradiometer (Aqua – MODIS)	2002 – Present	36 spectral bands	250m (bands 1-2) 500m (bands 3-7) 1000m (bands 8-36)	Global	Daily
National Aeronautics and Space Administration (NASA)	Landsat 8 Operational Land Imager (OLI)	2013 – Present	11 spectral bands	30m (bands 1-7 & 9) 15m (panchromatic band) 100m (TIRS bands 10-11)	170km x 185km	16-days
National Aeronautics and Space Administration (NASA)	Landsat 7 Enhanced Thematic Mapper (ETM+)	1999 – Present	8 spectral bands	30m (bands 1-7) 15m (panchromatic band)	170km x 185km	16-days
National Aeronautics and Space Administration (NASA)	Landsat 5 Thematic Mapper (TM)	1984 – 2013	7 spectral bands	30m (bands 1-5 & 7) 120m (thermal band)	172km x 185km	16-days
National Aeronautics and Space Administration (NASA)	Landsat 5 Thematic Mapper (TM)	1982 – 2093	7 spectral bands	30m (bands 1-5 & 7) 120m (thermal band)	170km x 185km	16-days
European Space Agency (ESA)	Sentinel-1A/B	Sentinel-1A: 2014 – Present Sentinel-1B: 2016 – Present	1 synthetic aperture radar (SAR) band	5m (wave & strip mode) 20m x 40m (wide mode)	20km x 20km to 400km x 400km (depending on mode)	3 to 6-days (from two satellite constellation)
European Space Agency (ESA)	Sentinel-2A/B Multispectral Imager (MSI)	Sentinel-1A: 2014 – Present Sentinel-1B: 2016 – Present	13 spectral bands	10m (visible & near-infrared) 10m (shortwave infrared) 60m (atmospheric correction)	290km x 300km	2 to 5-days (from two satellite constellation)

Table 2: Summary of available satellite sensors for the observation of flood events.

2a.3. Flood Detection Methods

There is a large variety of water detection algorithms that can be applied to the imagery from the flood detection sensors listed above (see Coltin et al. (2016) for a review for the MODIS sensor alone). For flood detection in Senegal, the authors chose four sensors, MODIS and Landsat 5-8, principally due to their availability over the time frame of historical flood events. MODIS also has the advantage of daily scenes that increases the likelihood of getting a cloud-free look at flood events. In this analysis, different water detection techniques were used for MODIS and Landsat images. For MODIS, the authors utilized the method developed by the Dartmouth Flood Observatory and NASA's Near Real Time Flood Mapping[4] platform. In addition, an automatic threshold detection technique known as Otsu thresholding was used to optimize the selection of water versus land pixels. For Landsat, a recently developed Automated Water Extraction Index (AWEI), an improvement over other water indices, was applied to available Landsat scenes. The following section describes the methods in full.

2a.3.1. MODIS Imagery

Daily MODIS satellite images that coincided with the dates of the identified floods in Senegal were collected using GEE's satellite sensor data catalog. Only 7 of the 8 flood events identified above coincided with the availability of MODIS imagery (2000 – present). Across these 7 events 839 images were collected and analyzed. Of the water detection techniques available, the method utilized by the Dartmouth Flood Observatory and NASA's Near Real Time Flood Mapping[5] platform was chosen for use in Senegal (Brackenridge, Anderson, & Caquard, 2009). This method allows for the detection of discrete flood events from daily MODIS imagery. In particular, the DFO algorithm is able to avoid a common misclassification of cloud shadows and hill shade areas as water due to their similar spectral signatures. The DFO algorithm overcomes this issue by applying either 2 or 3-day composites of images that maintain stationary elements (water) and eliminate mobile elements (cloud shadows) between daily images. The 2-day composites are able to capture highly transient flood events though more "noise" may be present due to the coincidence of cloud shadows, leading to more false positives. On the other hand, 3-day composites reduce the coincidence of cloud shadows across scenes but miss rapid or flash flood events that are highly transient, creating more false negatives. The choice of 2- or 3-day composites comes with this inherent trade-off. Overall, however, this method provides a relatively accurate approach for observing flood events. In a quantitative comparison among several flood detection techniques, the DFO algorithm was found to have a relatively high measure of precision and recall when compared to other water detection algorithms (Coltin et al., 2016).

2a.3.2. Landsat Imagery

Water detection techniques for Landsat often use band thresholding, primarily Normalized Difference Water Index (NDWI) and the Modified Normalized Difference Water Index (MNDWI) (Gao, 1996; Xu, 2006). Feyisa et. al. (2014) presented a new water index, the Automated Water Extraction Index (AWEI), that addresses several shortcomings of other water indices such as NDWI or MNDWI. In particular, it has been recognized that water indices face two major problems: 1) results obtained using different indices are inconsistent; 2) threshold values applied to distinguish water from non-water are unstable, varying with scene and locations (Ji, Zhang, & Wylie, 2009). These problems are pronounced in classifications with significant areas of low-albedo surfaces and the presence of shadows. To address these issues, Feyisa et. al. (2014) formulated two equations based on Landsat 5 "blue" and "green" bands termed *non-shadow* (*nsh*) and *shadow* (*sh*). $AWEI_{nsh}$ is primarily formulated to eliminate non-water pixels including build-up urban areas and $AWEI_{sh}$ is designed to further improve accuracy by removing cloud shadows. As a result, these equations can be used in isolation or combination depending on the specific challenges of a scene or location to minimize misclassifications.

The AWEI algorithms developed by Feyisa et. al. (2014) were implemented over Senegal where Landsat images were available. Across the 8 flood events, 5 flood events had available Landsat imagery that totaled 266 images and were analyzed using the methods of Feyisa described above. The AWEIsh and AWEInsh equations were generalized to all Landsat sensors including 4, 5, 7, and 8 allowing for potentially more "looks" during each flood event. To restrict the AWEI thresholds to appropriate pixels a cloud mask was applied to each available Landsat image (Zhu & Woodcock, 2012).

[4] The Land, Atmosphere Near real-time Capability for EOS (LANCE) supports application users interested in monitoring a wide variety of natural and man-made phenomena. Near Real-Time (NRT) data and imagery from the AIRS, AMSR2, MISR, MLS, MODIS, OMI and VIIRS instruments are available much quicker than routine processing allows. Most data products are available within 3 hours from satellite observation. NRT imagery are generally available 3-5 hours after observation. https://earthdata.nasa.gov/earth-observation-data/near-real-time

[5] The Land, Atmosphere Near real-time Capability for EOS (LANCE) supports application users interested in monitoring a wide variety of natural and man-made phenomena. Near Real-Time (NRT) data and imagery from the AIRS, AMSR2, MISR, MLS, MODIS, OMI and VIIRS instruments are available much quicker than routine processing allows. Most data products are available within 3 hours from satellite observation. NRT imagery are generally available 3-5 hours after observation. https://earthdata.nasa.gov/earth-observation-data/near-real-time

2a.3.3. Automatic Thresholding

A common problem with the threshold technique for water detection is determining the appropriate threshold for each satellite image used. The spectral reflectance of land and water (see Figure 1 land-water histograms for example) are known to vary from region to region and even among images of the same locations due to changing environmental variables such as depth, water turbidity, chemical composition and surface appearance. As case in point, it was found for Landsat 8 images over the Murray-Darling Basin in Australia that optimal thresholds for MNDWI across the region ranged from -0.20 to 0.40 (Donchyts, Schellekens, et al., 2016).

The DFO flood detection method described above utilizes three thresholds to identify water pixels within an image including: a ratio of NIR and RED bands (NIR/Red Ratio), a threshold of the Red band, and a threshold of the SWIR band with standard values of 0.70, 2027, and 675, respectively. Using these standard thresholds for this analysis, several misclassifications occurred over the extent of Senegal, highlighting the need for adjustments to these thresholds (Figure 1).

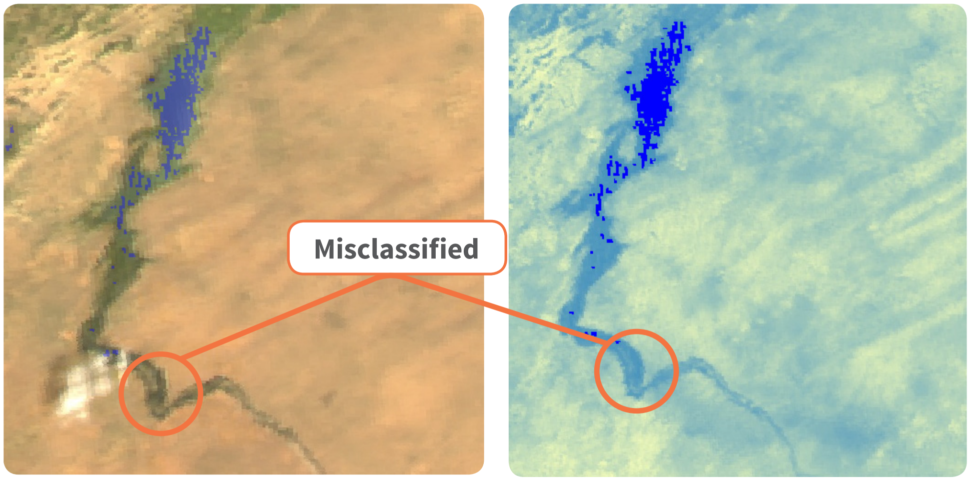


Figure 1: Areas of water detection where the standard threshold values of he DFO algorithm misclassified water bodies as land. (left: Lac de Guiers in visible spectrum; right: Lac de Guiers in Shortwave Infrared - SWIR).

To determine optimal thresholds for distinguishing land from water a technique known as Otsu thresholding was used (Otsu, 1975). In short, the Otsu threshold determines the interclass variance within a land/water histogram to find the threshold of the greatest interclass variance. The maximum interclass variance indicates the optimal threshold for detection of water versus land. To determine the highest interclass variance and thus the optimal threshold, different thresholds were tested in a stepwise fashion (~100 thresholds) to identify where the interclass variance peaks. Otsu thresholding is known to work best when the water class represents a significant portion of pixels within the histogram and are not obfuscated by clouds. This technique has been successful employed in several studies including detection of watercourses in the Murray-Darling Basin in Australia using Landsat-8 (Donchyts, Schellekens, et al., 2016), river delineation of the Brahmaputra River in India using Landsat-5 (Yang et al., 2014), and surface water detection in the Yangtze River Basin in China (Li et al., 2013).

For this assessment of Senegal, the internal MODIS Quality Assurance (QA) bands were used to identify the least cloudy image and remove cloudy pixels, and a range of threshold values were tested. A buffer region around a permanent water mask, provided by a dataset mapping forest cover in the 21st century by Hansen et. al. (Hansen et al., 2013), was used to ensure a high proportion of water pixels in the sample area. The resulting histograms were used to determine the optimal threshold using methods presented by Otsu (1975).

2a.3.4. UNOSAT Data

UNOSAT Spatial flood extend data was used as comparison with other datasets to better understand flooding in Senegal and to shed light on the advantages of SAR data. The UNOSAT Flood Portal listed several flood events for Senegal and, in one case, also hosted spatial data of inundation extent. Specifically, an image was captured by the Canadian Space Agency Radarsat-2 sensor on September 5th, 2012, and was analyzed for flood extent surrounding in the Saloum Delta surrounding Kaolack city and the Kaolack and Fatick provinces. The methodology behind this flood detection was not publicly available and cannot be compared to other methodologies. This dataset does, however, provide insight into the benefits of SAR data in seasons of high cloud coverage and targeted specifically to areas experiencing severe flooding.

2a.4. Flood Detection Results

2a.4.1. DFO Algorithm Implementation

In general, the DFO algorithm identified flooding in deltaic wetlands and low-lying areas surrounding rivers. Within major cities such as Dakar, Saint-Louis, or Kaolack the DFO algorithm was unable to detect significant flooding that had otherwise been reported. The lack of detection in urban areas can be explained by "mixed" pixels where the spectral response is a mix of multiple land use land covers such as impervious surface, vegetation, and water within one 250-m2 pixel. Conversely, in rural areas and wetlands, where flooded areas occupy larger natural features rather than smaller, urban features, this effect is less pronounced. Figure 2 below shows the results of the DFO algorithm flood detection in urban (Saint-Louis) and rural (Ziguinchor) areas by illustrating the number of times an area or pixel was observed as flooded across the recorded flood events. This map highlights areas that regularly flood during heavy rain events.

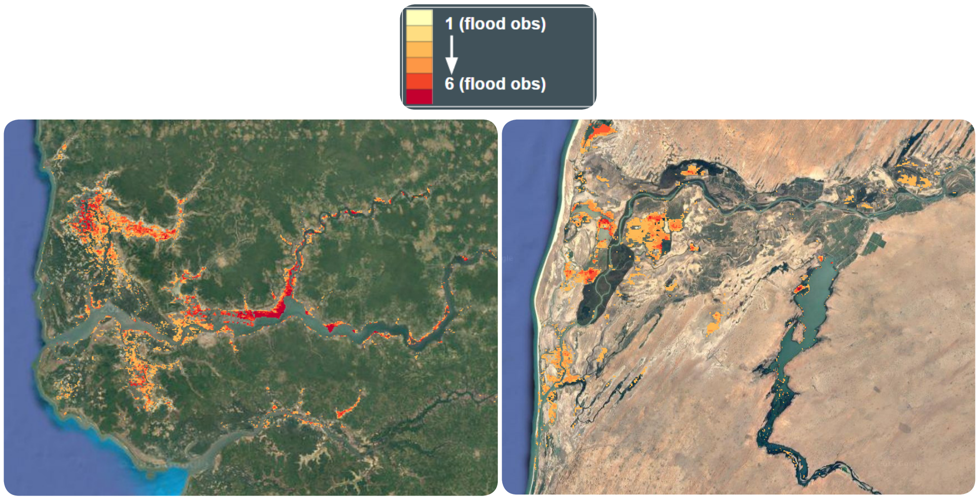


Figure 2: Number of times an area (pixel) flooded from 2003–2015 in Senegal using the DFO algorithm (left: Ziguinchor, Senegal; right: Saint-Louis, Senegal and Senegal River).

Methods for improving the DFO algorithm were also explored due to observable error in classification of water and land when using standard thresholds. Specifically, the authors applied an Otsu thresholding technique that automatically selects an optimal threshold based on the interclass variance between the reflectance of land and water. The results obtained confirm the necessity for implementation of Otsu thresholds as the range for the NIR/Red Ratio and SWIR threshold were 0.49 – 0.85 and 290 – 885, respectively. Although the range of values of the NIR/Red Ratio and the SWIR threshold clustered around the standard DFO default values of 0.70 and 675, respectively, these ranges indicate that improvements in flood detection can be obtained by updating the thresholding per event. This can be explained by variations in image artifacts such as haze, changing phenology of vegetation, or turbidity of water within individual scenes. Figures 3 and 4 show the results of the Otsu thresholding across the flood events tested and for the NIR/ Red Ratio and SWIR bands.

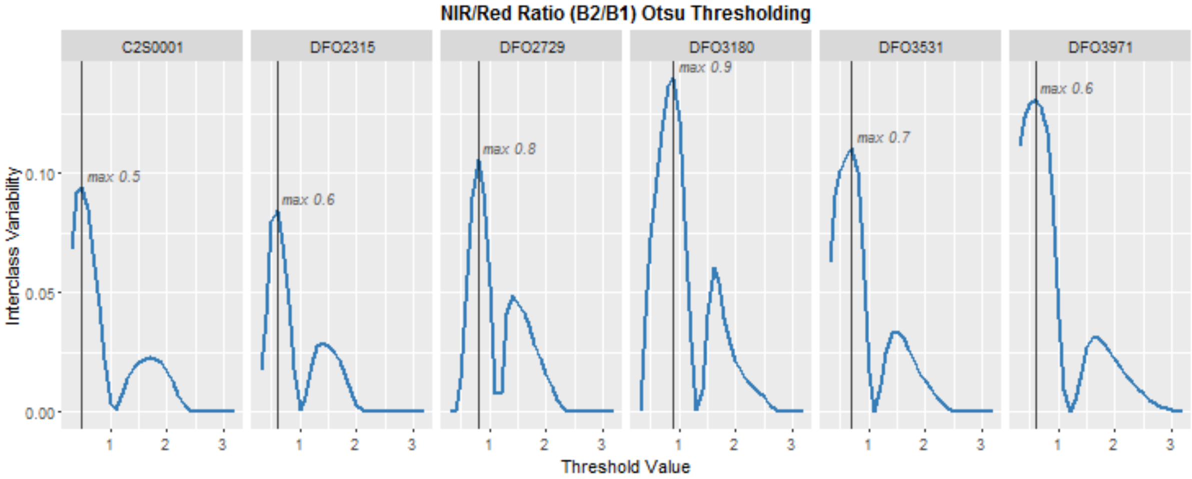


Figure 3: Otsu thresholding results for the NIR/ Red Ratio

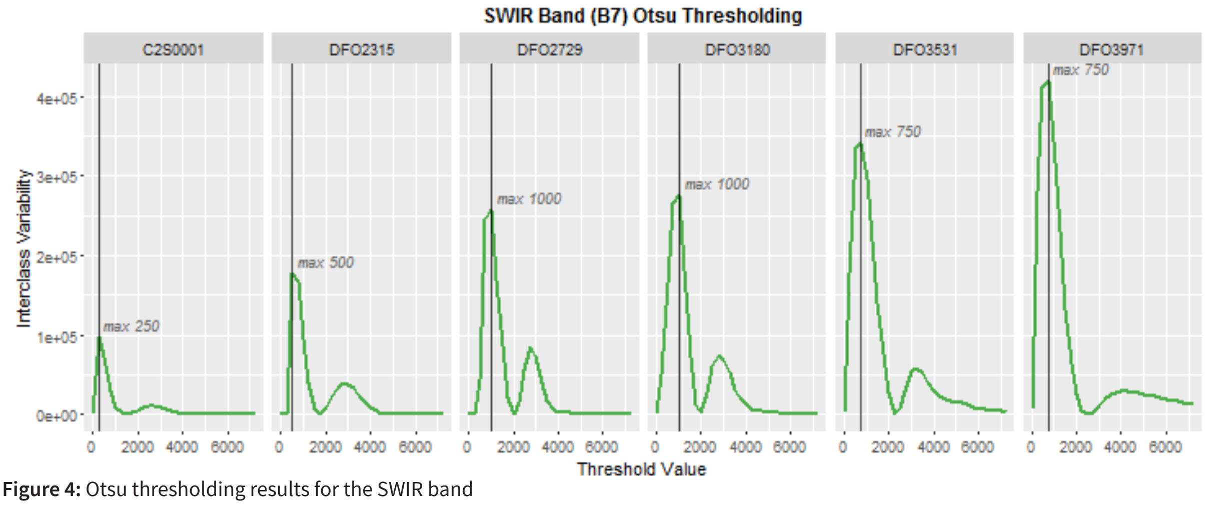


Figure 4: Otsu thresholding results for the SWIR band

2a.4.2. Feyisa Algorithm Implementation

A flood detection algorithm for Landsat was also prepared with the aim of improving the detection of flooding in urban areas. An Automated Water Extraction Index (AWEI) presented by Feyisa et al (2014) was implemented across available Landsat 4 – 8 imagery that coincided with flood events. The results of this approach can be seen in Figure 5, which compared to the DFO algorithm has improved detection capabilities in Saint-Louis, Senegal. From these images, flooding was observed in the southeastern portions of Saint-Louis (Figure 5 – right) as well as in surrounding deltaic wetlands located in rural areas (Figure 5 – left).

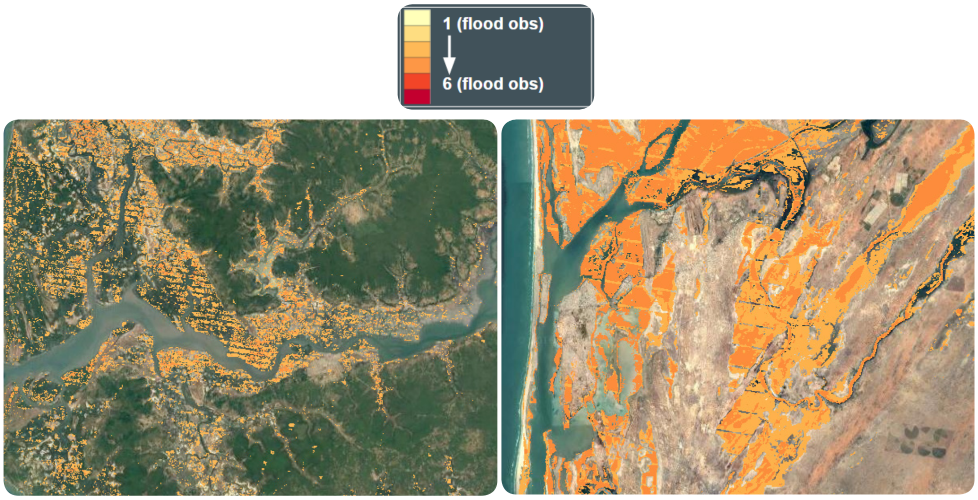


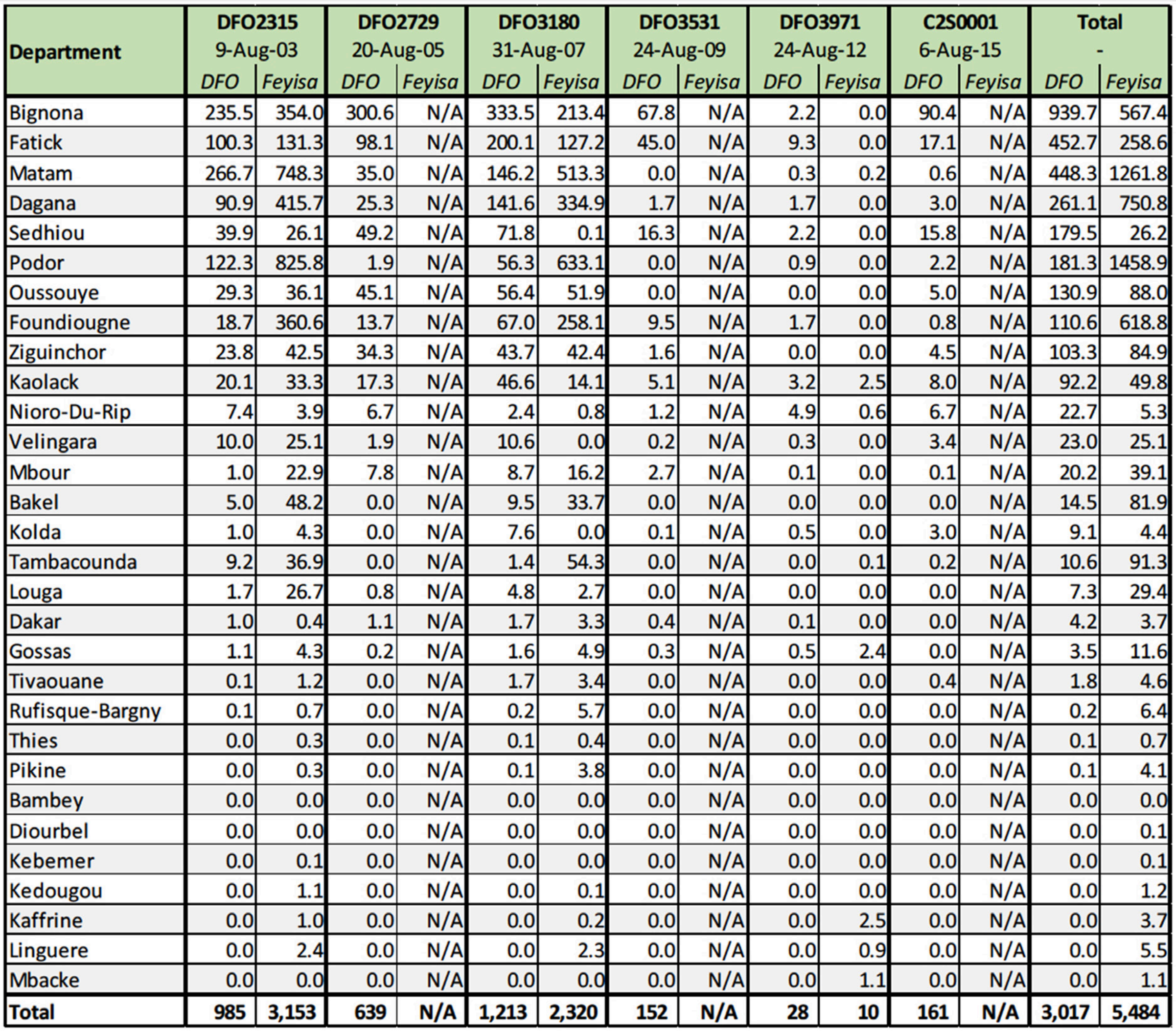
Figure 5: Number of times an area (pixel) flooded from 2003–2015 in Senegal using the Feyisa algorithm (left: Ziguinchor, Senegal; right: Saint-Louis and Senegal River).

It’s important to note that, unlike the DFO algorithm, smaller portions of Senegal are observable during flood events. Compared to the DFO algorithm, only 3 of the 6 catalogued flood events had available images to apply the Feyisa flood detection method, minimizing the temporal utility of this method. Additionally, Landsat 7, which covers a significant time period from 1999 to present, is of marginal utility because a sensor error in 2003 created “striping” within images (see Figure 5). As a result, the coverage of Landsat is compromised but, in return, higher detail in urban areas is possible.

2a.5. Flood History in Senegal

The area of each flood event was calculated over each department in Senegal to demonstrate which areas have the most persistent flooding (Table 3). On an area basis, the results demonstrate that the most extensive flooding occurs in departments consisting of predominantly rural areas. This is consistent with our observations above where rural areas had significant flooding in deltaic wetlands. According to our results, the most extensive flood in terms of area was from rain events in August – September 2007 (DFO 3180) that covered 1,213 to 2,320 km^2, depending on the flood detection method used. These results are consistent with the DFO catalogue in identifying the most extensive flood events, though the magnitude differs as the DFO catalogue reported 167,997 km^2 for the August – September 2007 floods. The DFO estimate is nearly the total area of Senegal and likely is calculated beyond the boundaries of Senegal, using impacted boundaries (departments, watersheds) as a basis for calculation, or both. Alternatively, our results provide a more realistic estimate of actual area inundated within departments in Senegal providing a more detailed understanding of affected areas.

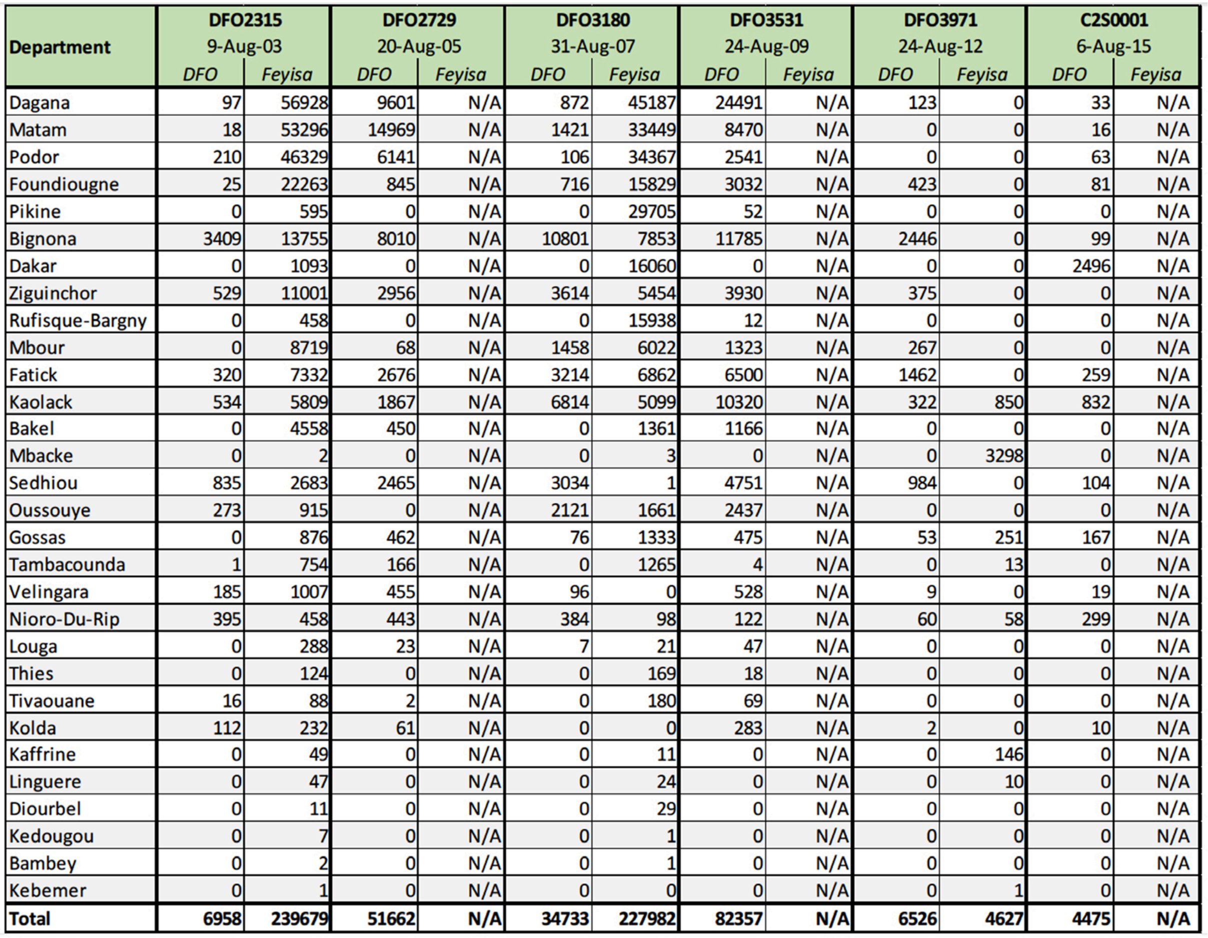
Depending on the flood detection method used, different estimates of affected area are found, highlighting the need to use multiple sensors and observation when possible. In general, the Feyisa method consistently predicts greater flooded areas than the DFO flood detection method, even in rural areas where the DFO algorithm is considered to perform best. The greater resolution of Landsat provides greater sensitivity to flooded pixels, though the number of observations through time is limited by the fact that only 3 of the 6 catalogued floods had available Landsat imagery.



Department	DFO2315 9-Aug-03		DFO2729 20-Aug-05		DFO3180 31-Aug-07		DFO3531 24-Aug-09		DFO3971 24-Aug-12		C2S0001 6-Aug-15		Total -	
	DFO	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*
Bignona	235.5	354.0	300.6	N/A	333.5	213.4	67.8	N/A	2.2	0.0	90.4	N/A	939.7	567.4
Fatick	100.3	131.3	98.1	N/A	200.1	127.2	45.0	N/A	9.3	0.0	17.1	N/A	452.7	258.6
Matam	266.7	748.3	35.0	N/A	146.2	513.3	0.0	N/A	0.3	0.2	0.6	N/A	448.3	1261.8
Dagana	90.9	415.7	25.3	N/A	141.6	334.9	1.7	N/A	1.7	0.0	3.0	N/A	261.1	750.8
Sedhiou	39.9	26.1	49.2	N/A	71.8	0.1	16.3	N/A	2.2	0.0	15.8	N/A	179.5	26.2
Podor	122.3	825.8	1.9	N/A	56.3	633.1	0.0	N/A	0.9	0.0	2.2	N/A	181.3	1458.9
Oussouye	29.3	36.1	45.1	N/A	56.4	51.9	0.0	N/A	0.0	0.0	5.0	N/A	130.9	88.0
Foundiougne	18.7	360.6	13.7	N/A	67.0	258.1	9.5	N/A	1.7	0.0	0.8	N/A	110.6	618.8
Ziguinchor	23.8	42.5	34.3	N/A	43.7	42.4	1.6	N/A	0.0	0.0	4.5	N/A	103.3	84.9
Kaolack	20.1	33.3	17.3	N/A	46.6	14.1	5.1	N/A	3.2	2.5	8.0	N/A	92.2	49.8
Nioro-Du-Rip	7.4	3.9	6.7	N/A	2.4	0.8	1.2	N/A	4.9	0.6	6.7	N/A	22.7	5.3
Velingara	10.0	25.1	1.9	N/A	10.6	0.0	0.2	N/A	0.3	0.0	3.4	N/A	23.0	25.1
Mbour	1.0	22.9	7.8	N/A	8.7	16.2	2.7	N/A	0.1	0.0	0.1	N/A	20.2	39.1
Bakel	5.0	48.2	0.0	N/A	9.5	33.7	0.0	N/A	0.0	0.0	0.0	N/A	14.5	81.9
Kolda	1.0	4.3	0.0	N/A	7.6	0.0	0.1	N/A	0.5	0.0	3.0	N/A	9.1	4.4
Tambacounda	9.2	36.9	0.0	N/A	1.4	54.3	0.0	N/A	0.0	0.1	0.2	N/A	10.6	91.3
Louga	1.7	26.7	0.8	N/A	4.8	2.7	0.0	N/A	0.0	0.0	0.0	N/A	7.3	29.4
Dakar	1.0	0.4	1.1	N/A	1.7	3.3	0.4	N/A	0.1	0.0	0.0	N/A	4.2	3.7
Gossas	1.1	4.3	0.2	N/A	1.6	4.9	0.3	N/A	0.5	2.4	0.0	N/A	3.5	11.6
Tivaouane	0.1	1.2	0.0	N/A	1.7	3.4	0.0	N/A	0.0	0.0	0.4	N/A	1.8	4.6
Rufisque-Bargny	0.1	0.7	0.0	N/A	0.2	5.7	0.0	N/A	0.0	0.0	0.0	N/A	0.2	6.4
Thies	0.0	0.3	0.0	N/A	0.1	0.4	0.0	N/A	0.0	0.0	0.0	N/A	0.1	0.7
Pikine	0.0	0.3	0.0	N/A	0.1	3.8	0.0	N/A	0.0	0.0	0.0	N/A	0.1	4.1
Bambey	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A	0.0	0.0
Diourbel	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A	0.0	0.1
Kebemer	0.0	0.1	0.0	N/A	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A	0.0	0.1
Kedougou	0.0	1.1	0.0	N/A	0.0	0.1	0.0	N/A	0.0	0.0	0.0	N/A	0.0	1.2
Kaffrine	0.0	1.0	0.0	N/A	0.0	0.2	0.0	N/A	0.0	2.5	0.0	N/A	0.0	3.7
Linguere	0.0	2.4	0.0	N/A	0.0	2.3	0.0	N/A	0.0	0.9	0.0	N/A	0.0	5.5
Mbacke	0.0	0.0	0.0	N/A	0.0	0.0	0.0	N/A	0.0	1.1	0.0	N/A	0.0	1.1
Total	**985**	**3,153**	**639**	**N/A**	**1,213**	**2,320**	**152**	**N/A**	**28**	**10**	**161**	**N/A**	**3,017**	**5,484**

Table 3: Summary of inundated area (km^2) detected by the *DFO* and *Feyisa* methods within each department.

Inundated area is one measure of flood severity, though more important is an understanding of the human impact of flood events. With new advances in estimating population and poverty using machine learning techniques and remote sensing data (Jean et al., 2016; Stevens, Gaughan, Linard, & Tatem, 2015), it is possible to quantify affected areas in terms of population. This was done using an available dataset for Senegal provided by World Pop on population density. Table 4 summarizes the results.



Department	DFO2315		DFO2729		DFO3180		DFO3531		DFO3971		C2S0001	
	9-Aug-03		20-Aug-05		31-Aug-07		24-Aug-09		24-Aug-12		6-Aug-15	
	DFO	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*	*DFO*	*Feyisa*
Dagana	97	56928	9601	N/A	872	45187	24491	N/A	123	0	33	N/A
Matam	18	53296	14969	N/A	1421	33449	8470	N/A	0	0	16	N/A
Podor	210	46329	6141	N/A	106	34367	2541	N/A	0	0	63	N/A
Foundiougne	25	22263	845	N/A	716	15829	3032	N/A	423	0	81	N/A
Pikine	0	595	0	N/A	0	29705	52	N/A	0	0	0	N/A
Bignona	3409	13755	8010	N/A	10801	7853	11785	N/A	2446	0	99	N/A
Dakar	0	1093	0	N/A	0	16060	0	N/A	0	0	2496	N/A
Ziguinchor	529	11001	2956	N/A	3614	5454	3930	N/A	375	0	0	N/A
Rufisque-Bargny	0	458	0	N/A	0	15938	12	N/A	0	0	0	N/A
Mbour	0	8719	68	N/A	1458	6022	1323	N/A	267	0	0	N/A
Fatick	320	7332	2676	N/A	3214	6862	6500	N/A	1462	0	259	N/A
Kaolack	534	5809	1867	N/A	6814	5099	10320	N/A	322	850	832	N/A
Bakel	0	4558	450	N/A	0	1361	1166	N/A	0	0	0	N/A
Mbacke	0	2	0	N/A	0	3	0	N/A	0	3298	0	N/A
Sedhiou	835	2683	2465	N/A	3034	1	4751	N/A	984	0	104	N/A
Oussouye	273	915	0	N/A	2121	1661	2437	N/A	0	0	0	N/A
Gossas	0	876	462	N/A	76	1333	475	N/A	53	251	167	N/A
Tambacounda	1	754	166	N/A	0	1265	4	N/A	0	13	0	N/A
Velingara	185	1007	455	N/A	96	0	528	N/A	9	0	19	N/A
Nioro-Du-Rip	395	458	443	N/A	384	98	122	N/A	60	58	299	N/A
Louga	0	288	23	N/A	7	21	47	N/A	0	0	0	N/A
Thies	0	124	0	N/A	0	169	18	N/A	0	0	0	N/A
Tivaouane	16	88	2	N/A	0	180	69	N/A	0	0	0	N/A
Kolda	112	232	61	N/A	0	0	283	N/A	2	0	10	N/A
Kaffrine	0	49	0	N/A	0	11	0	N/A	0	146	0	N/A
Linguere	0	47	0	N/A	0	24	0	N/A	0	10	0	N/A
Diourbel	0	11	0	N/A	0	29	0	N/A	0	0	0	N/A
Kedougou	0	7	0	N/A	0	1	0	N/A	0	0	0	N/A
Bambey	0	2	0	N/A	0	1	0	N/A	0	0	0	N/A
Kebemer	0	1	0	N/A	0	0	0	N/A	0	1	0	N/A
Total	**6958**	**239679**	**51662**	**N/A**	**34733**	**227982**	**82357**	**N/A**	**6526**	**4627**	**4475**	**N/A**

Table 4: Summary of population in inundated areas determined two different flood detection methods (DFO and Feyisa) and World Pop data.

The estimates of population in inundated zones in Table 4 differs substantially by flood detection method, more so than the estimates of area of flooded zones. The Feyisa method, for 2 out the 3 flood detections available, greatly overpredicts the population affected as compared to the DFO algorithm. In the case of the 2003 and 2007 floods, the Feyisa method estimates are an order of magnitude greater. These differences can be explained, again, by the fact that the Feyisa method is able to detect flood occurring in urban areas, where population is much greater, while the DFO method performs best in rural or deltaic wetland areas with low populations. Since the population is uneven across Senegal and rural/urban contexts, the Feyisa method greatly over predicts compared to the DFO method, emphasizing its utility when it comes to flood impacting urban areas. Of course, where no Landsat imagery is available, the DFO method still provides estimates and relative areas that are impacted, though likely largely underestimated.

2a.6. Limitations

Several limitations of the data analysis provided above should be noted to highlight the bounds of its utility. First and foremost, our estimates of flood extent and thus affected populations do not provide the potential full coverage of flood events in Senegal. Our exercise in cataloguing flood events was based on publically available information; additional flood events beyond those highlighted are therefore likely. Additionally, our observations of floods are limited by available and usable (i.e. cloudless) imagery, which is rare during major rain events. Lastly, the estimates of flooded areas are not certain and are lacking validation data, which is consistently a challenge for ephemeral flood events. Together, these limitations restrict the utility of our flood maps where potentially high severity flood events were not observed that would deserve equal attention. The use of these flood maps is best applied to further modeling where validation and uncertainty in flood plains can be communicated.

2b. Biophysical Risk: Data-Drive Hydrology with Machine Learning

New data-driven methods hold the promise of overcoming limitations in traditional flood modeling and allowing us to predict floods faster and more dynamically for new places in which data are currently lacking. This chapter details the exploratory methods used to develop a machine learning flood prediction model for Senegal and provides a preliminary estimation of areas most vulnerable to flooding in the watersheds mostly likely to be at risk from these extreme events. Building on the flood database described in the previous chapter, the research described here has two corresponding goals: 1) understanding the methods, the benefits and the current limitations of applying new data tools to Senegal, and 2) developing the science to optimize machine learning algorithm parameters across climatological and ecological gradients.

These results show that machine learning (ml) algorithms have the potential to be able to reproduce benchmark historic floods as detected by remote sensing, especially using Random Forest on the flood detected with MODIS satellites (accuracy of 97%). The Saint-Louis region was the primary testing ground for customizing the algorithm, where the authors designed and assessed four machine learning approaches on 11 flood conditioning factors. Scaling the tool to an additional five watershed regions, which were selected in consultation with at Agence Française de Développement and which cover 34% of the country, the authors predicted a floodplain of 5,596 km^2 and 30% of that floodplain (1,641 km^2) is considered high risk, meaning the model predicted flooding in 100% of the trials. Over 97,000 people could be at high risk of exposure to flooding according to analysis conducted using the WorldPop gridded population dataset.

To describe this research and its promise for physical vulnerability assessment in Senegal, this chapter first provides contextual background on traditional and new strategies for flood modeling (Section 2b.1). The authors describe the methods used to develop a Senegal machine learning model and its outputs (Sections 2b.2 and 2b.3). In the final two sections, the authors discuss the limitations and further research, and provide guidance on using the science developed for the report as a tool for future analysis of vulnerability in Senegal.

2b.1. Introduction

Machine learning (ml), defined as advanced programming strategies that provide computers with the ability to learn without being explicitly programmed, is a cutting-edge new tool increasingly used to analyze flooding. Applying these tools to International River Basins (IRBs) allows hydrologists and programmers to overcome the current limits of understanding the river dynamics and to better predict floods and vulnerability. However, because the data-driven flooding approach developed for this report is so new, its results are not yet fully tested.

Traditionally, large-scale flood modeling of IRBs relied on physically-based flood models that are rooted in equations describing the physical movement of water. These models are usually expensive to build, require significant expertise to calibrate, and can take days of computation time to generate a single set of results. Model outputs also only represent a snapshot of flood risk because the parameters used in the model are time-specific (rainfall, land use, population). Being static as such can quickly renders the outputs from traditional models irrelevant, especially in areas of rapid development.

As a response, many experts are arguing for simpler, satellite-based approaches, even at the cost of significant decreases in accuracy, in order to address the urgent need for hydrologic data in IRBs (Hossain, Katiyar, Hong, & Wolf, 2007). Machine learning algorithms and remote sensing are increasingly used in lieu of process-based methods to advance the field of hazard forecasting by producing flood maps at higher speed and lower cost (Naghibi & Pourghasemi, 2015; Rasouli, Hsieh, & Cannon, 2012; Solomatine & Xue, 2004). Initial applications of machine learning in the field of hydrology included the use of neural networks and support vector machines to predicting flood extent (Han, L, & N, 2007; Liong & Sivapragasam, 2002) and rainfall-runoff flow rates (Campolo, Andreussi, & Soldati, 1999; Lin, Cheng, & Chau, 2006). These and other studies (Hong, 2008; Pradhan, 2010; Tehrany, Pradhan, & Jebur, 2013; D. Wang et al., 2013) showed the algorithms could prove useful in modeling extreme events. This study, which applies machine learning to generate flood predictions in Senegal, tests this ground-breaking methodology and has high potential to transform the way global inundation modeling is done.

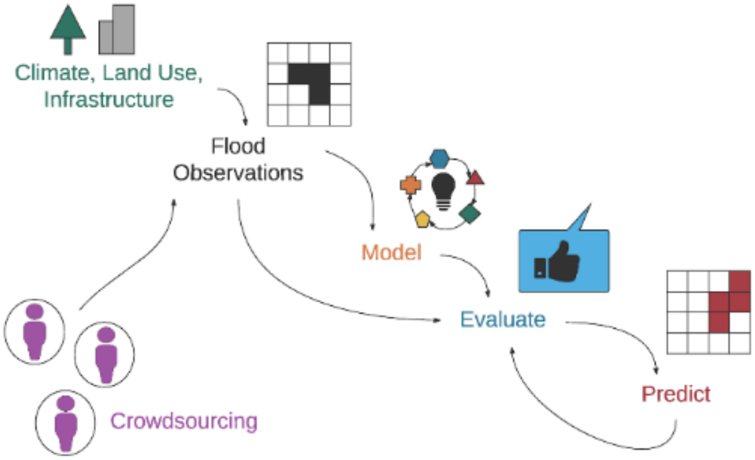


Figure 6: Conceptual model workflow for machine learning based flood plain prediction.

2b.2. Methods

2b.2.1 Study Areas

Three river valleys (Senegal River, Saloum River, Casamance River) and the Dakar area were identified as candidates for modeling because of the availability of training data and based on feedback from personnel at Agence Française de Développement. In-depth model training, and validation was conducted in along the Senegal River Valley (SRV) in the Saint-Louis region, in northwestern Senegal. The region was selected as the primary prototype because of its high population, its history of flooding, and the robust library of training data. Flooding is considered the greatest hazard on the risk continuum in this area (Pelling & Wisner, 2012). In addition to this in-depth study, less in-depth modeling was done for another part of the SRV; for the Saloum and Casamance river valleys; and for the Dakar area.

These study areas span 32 different arrondissements in the following administrative regions: Saint-Louis, Matam, Fatick, Kaolack, Dakar, Sédhiou, and Ziguinchor, as shown in Figure 7. The full list of arrondissements is given in Table A2.

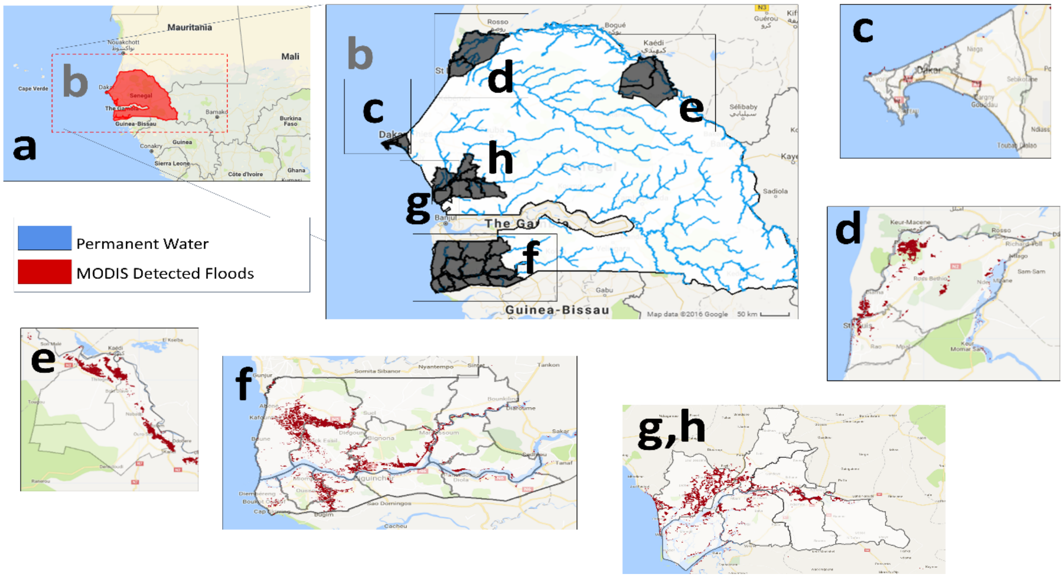


Figure 7: Maps of Senegal (a), showing the study areas (b) chosen for the machine learning model: floods were predicted in the area around Dakar (c); in the Senegal River Valley for selected arrondissements of the Saint-Louis (d) and Matam (e) regions; in the Casamance River for selected arrondissements of the Ziguinchor and Sedhiou regions (f); and in the Saloum River Valley for selected arrondissements of the Kaolack (g) and Fatick (h) regions. Red pixels indicate areas of MODIS-detected floods that are used as training data for the machine learning model.

2b.2.2. Flood Conditioning Factors

Flood conditioning factors describe the environmental conditions contributing to physical flood risk in a given watershed. While local scale data may be scarce, cutting-edge new efforts to generate global layers for many of these variables can be of use. For this study, flood conditioning factors (Table A1) were selected based on a literature review of both traditional and statistical flood models (Tehrany et al., 2013; Z. Wang et al., 2015). The significance of any given conditioning factor is expected to vary across landscape types. Factors were constrained to open source datasets in order to ensure the model could be replicated easily by anyone with access to the Internet.

Eleven total flood conditioning factors were chosen including: slope, digital elevation model (DEM), curvature, stream power index (SPI), topographic wetness index (TWI), impervious surface, normalized difference vegetation index (NDVI), slope, event precipitation, height above nearest drainage (HAND), and Euclidean distance from river. All variables were subsampled at a 30-m resolution and metadata details for each dataset can be found in Table A1.

2b.2.3. Model Development

The machine learning algorithms rely on training or reference data to determine landscape patterns of flooding. See Chapter 1 for a detailed explanation of how training data was created. Training data from the two different sources described in Chapter 2a, including the Dartmouth Flood Observatory algorithm (MODIS, 250 km resolution) and the Feyisa algorithm (Landsat, 30 m resolution), were tested for this analysis. This process, allowed us to explore how imagery resolution and detection approach impacted model predictions. Training data from both approaches in the form of a gridded binary raster (0 = flood, 1 = no flood) was stratified into the two classes and then randomly subsampled at 30-m resolution.

Using this training data, four different types of supervised machine learning algorithms were tested: (1) Random Forest (RF), (2) Support Vector Machine (SVM), (3) Fast Naive Bayes (NB) and (4) Classification & Regression Trees (CART). These four algorithms range from very simple (NB) to very complex ensemble classifiers (RF). While each algorithm relies on different statistical decision rules, they all use a similar framework, where flood conditioning factors and training data are inputs that generate a floodplain as an output (Figure 8).

2b.2.4. Performance Metrics

In a process called k-fold validation, model training and testing is repeated 10 times, withholding a separate 10% of training pixels each time. At the conclusion of the modeling exercise, each algorithm has a training and validation score indicating how well it identifies flooded pixels on familiar pixels (training data) or unfamiliar data (validation data) relative to the benchmark data. The average score is used to assess overall performance (Mannel, Price, & Hua, 2011) and is recorded in a table called a confusion matrix. Model results are then evaluated on a suite of metrics (Table 5) based on how many pixels in each class (flooded or not flooded) were correctly labeled in the modeled floodplain (Am) when compared to the benchmark or training data (Ab). These metrics, derived from the confusion matrix, have been used to evaluate other flood models (Alfieri et al., 2013; Bates, 2004; Werner, Hunter, & Bates, 2005) and measure accuracy with and without penalties for overprediction and underprediction.

Metric	Explanation	Equation
Hit Rate (H)	Pixels labeled as flooded in the training data (A_b) intersected with (∩) those predicted to be flooded by the model (A_m) (Sampson et al., 2015a)	$H = \frac{A_m \cap A_b}{A_b}$
False Alarm Rate (F)	Measure overestimating of the floodplain (between 0-1, 1 means all pixels are "false alarms" falsely labeled as flooded) (Wu et al., 2012)	$F = \frac{A_m \backslash A_b}{A_m \cap A_b + A_m \backslash A_b}$
Critical Success Rate (C)	Penalties for under and other prediction ratio of the total intersection of predicted and benchmark flood pixels divided by the total number or union (∪) of flooded pixels in both sets. Ranges from 0 -1 (1= perfect match) (Sampson et al., 2015b)	$C = \frac{A_m \cap A_b}{A_m \cup A_b}$
Mean Error (E_a)	Mean absolute error (E_a) where B is benchmark flooded fraction, M is modeled flooded fraction and N is number of grid cells formed by aggregating test case raster results to ~1 km scale. (Sampson et al., 2015a)	$E_a = \frac{\sum_{i=1}^{N} \lvert M-B \rvert}{N}$
Error Bias (B)	Score of 1 and greater indications a tendency to overpredict and scores between 0-1 indicate underprediction. (Sampson et al., 2015a)	$B = \frac{A_m/A_b}{A_b/A_m}$

Table 5: Metrics for evaluating model performance, error and bias

2b.3. Results

This model generates a 1 arc-second (30 meters) flood prediction extent across the six specified study regions (Figure 8) and can be applied anywhere in the country for any storm occurring within the Landsat and MODIS histories. The parallelized GEE computing framework divides the study region into tiles for simultaneous processing which allows the model to be run for any watershed in Senegal on a laptop from the internet within a matter of minutes.

2b.3.1. Saint-Louis Prototype

Four machine learning classifiers using two different types of training data were tested for the Saint-Louis region (Figure 8) with a range of model parameters. Accuracies for both types of training data are reported in Table A3. Overall model accuracies ranged from 47–92% with Hit Rates between 52-98%, comparable to accuracies reported in other applications of these algorithms to classification problems in GEE (Dong et al., 2016; Goldblatt, You, Hanson, & Khandelwal, 2016; Johansen, Phinn, & Taylor, 2015). Algorithms trained using the DFO flood detection algorithm outperformed those trained using the Feyisa method by 11% (overall accuracy) and 12% (Hit Rate). Model calibration improved RF Hit Rates by 5%. These results suggest that models trained with MODIS, rather than Landsat, training imagery will have higher accuracy. Among the four algorithms tested, the Random Forest algorithm has the greatest potential for accurately predicting floodplain extent based on training data, with an average Hit Rate of 97% and the lowest False Alarm rate (14%) relative to the other algorithms. Based on these results, RF should be prioritized when using machine learning for hydrologic assessment in Senegal. However, performance metrics for other algorithms are still in the range of that reported in flood literature, suggesting that these approaches should not be discarded entirely.

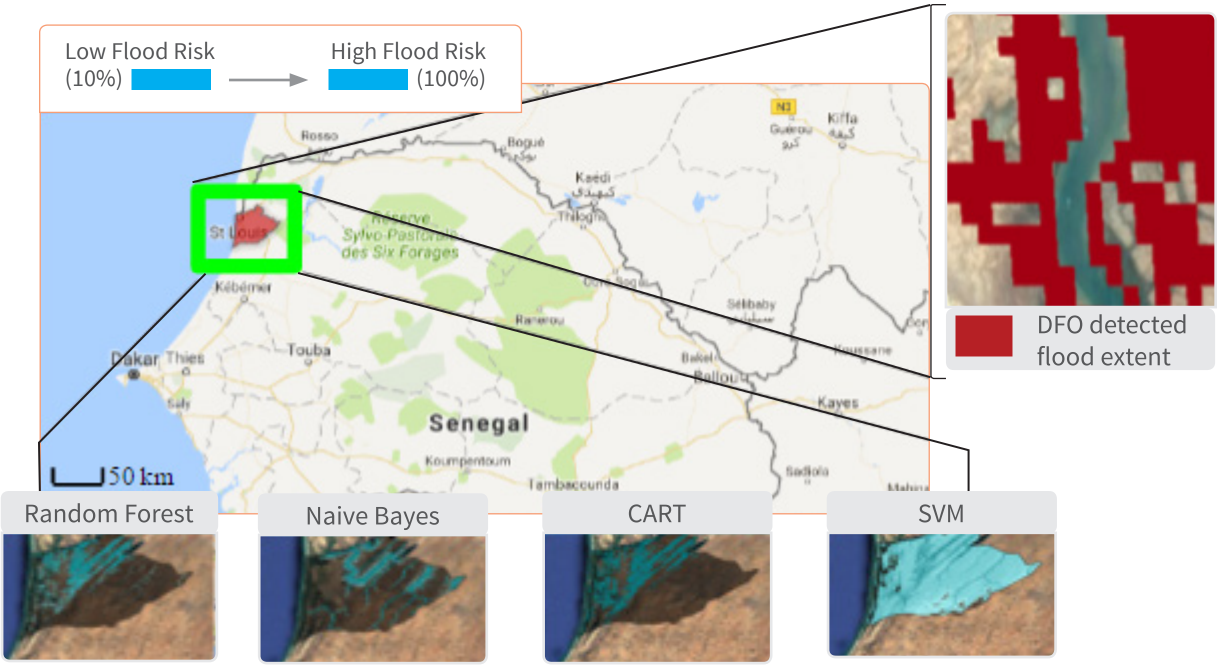


Figure 8: Example flood predictions based on the DFO historic flood #3180 detected from the Global Flood Database detection algorithm (upper right corner) in the Saint-Louis region along the Senegal River. The color gradient (lower images) indicates the number of times (1-10) a pixel was marked as flooded across the ten k-fold validation trials.

2b.3.2. Predictions for Flood-Prone Regions across Senegal

The Saint-Louis prototype was used to inform a Random Forest flood prediction model for five other regions in Senegal (Figure 9) covering an area of 24,992 km2, or 34% of the total area of the country. The model estimates a total predicted floodplain of 5,596 km2 and 30% of that floodplain (1,641 km2) is considered high risk, meaning the model predicted flooding in 100% of the trials. Over 14,000 people in each study region on average are at high risk of exposure to flooding according to analysis conducted using the WorldPop gridded population dataset. For all the regions analyzed in this study, over 97,000 people are estimated to be at very high risk of flood exposure.

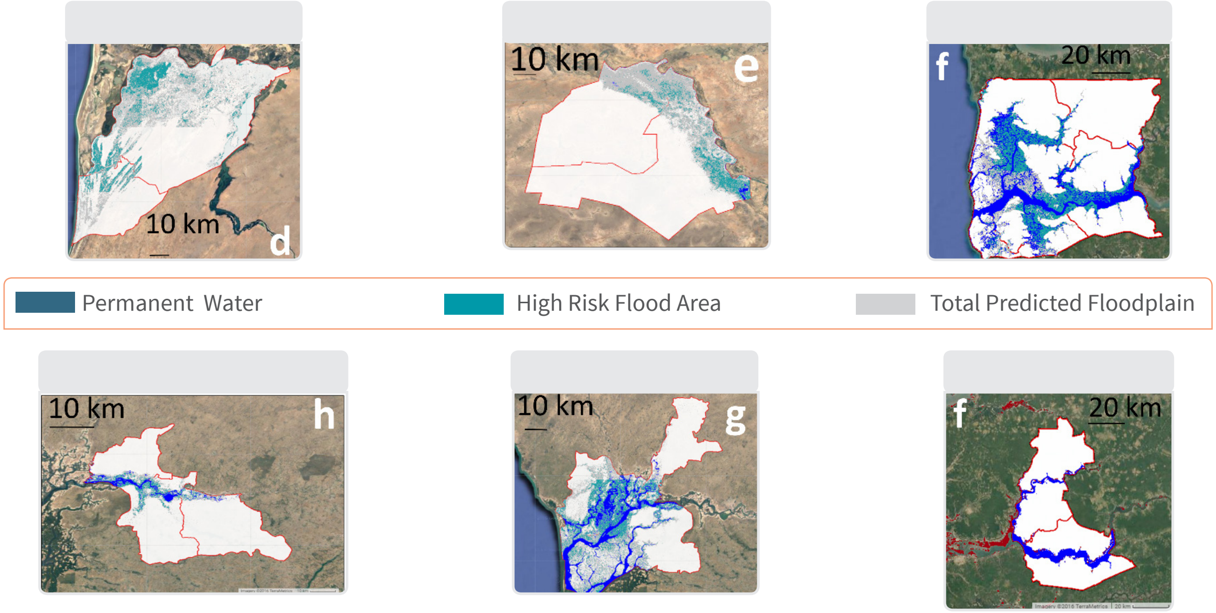


Figure 9: Floodplain extents for each of the five focus regions. Flood extents within the focus regions (white outlined in red) are separated into the total predicted floodplain (any area marked as flooded during validation, marked as grey) and the high-risk floodplain (any area marked as flooded during 10 out of the 10 validation trials, marked as dark turquoise). The Casamance River Valley was divided into two separate models for flood-prone arrondissements in the Ziguinchor and Sédhiou regions because of memory restrictions in GEE.

	Area Analyzed (km2)	Total Risk Area (km2)	% in Predicted Zone	High Risk Area (km2)	People at Risk
Matam	5,135	1,051	20%	114	38,400
Fatick	3,162	1,085	34%	528	17,038
Kaolack	1,906	204	11%	89	2,109
Saint-Louis	3,990	1,399	35%	523	8,208
Dakar	559	0	0%	0	0
Ziguinchor	7383	1,616	22%	349	31754
Sédhiou	2,855.81	241	8%	39	4,426

Table 5: Results showing the region area, the total risk area (any pixels classified as flooded in any trial, the high-risk area and the total population in the high-risk floodplain for each region. These results were generated using the Random Forest machine learning classifier trained with the MODIS September 2007 historic detected flood raster.

2b.4. Error, Model Limitations, and Further Research

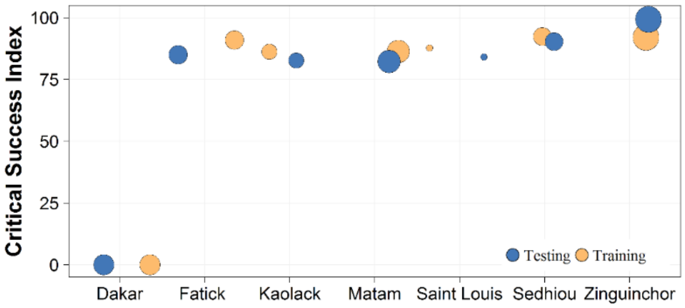


Figure 10: Critical Success Indices for training and validation (testing) rounds across the six River regions tested and Dakar. Bubble size is the relative areas of training data for each region.

Model bias shows the Random Forest algorithm has a bias to overprediction. RF has an average Error Bias of 1.16; an Error Bias between 0-1 show underprediction and > 1 shows overprediction. This overprediction is reflected in the false alarm rate, which ranges from 5-15%. Therefore total floodplain extent estimates are expected to be ~90% of the predicted total as a result of this model bias. Despite this bias toward overprediction, mean error still falls between 0.04-0.13, an exciting finding that shows the Random Forest model can be used to make reasonably accurate predictions with only a modest amount of parameterization. However, currently the model can only be run on areas smaller than 10,000 km2 because of memory constraints in GEE. Therefore the authors recommend the development of country-wide model based on a nested mosaic of smaller models tailored to the specific conditions of local regions instead of using a generic national approach. Also, the model currently is most successful in areas with robust training data, achieving high Critical Success rates in areas with training data. Figure 10 shows the impact of the lack of training data on the Critical Success Index of Dakar, whereas regions with training data are able to achieve high Critical Success Rates. As the results from Figure 10 and Table 6 show, in areas where the flood detection algorithm cannot identify floods to be used for training, the model is completely unable to generate predicted floodplains. The inclusion of higher resolution training data with broader spatial coverage could significantly improve model capabilities.

2b.5. Conclusion

Accurate flood mapping is crucial for protecting vulnerable populations and mitigating the catastrophic economic losses that can result from flood events. This dynamic socio-political and environmental context requires rapid, on-demand analysis executable within the constraints of sparse field data. Results from this research demonstrate the potential of using machine learning to revolutionize flood modeling for this region.

This project tested four machine learning algorithms trained with two different resolutions of training data for the September 2012 floods in the Saint-Louis region.

These results show that machine learning algorithms have potential to be able to reproduce benchmark historic floods as detected by remote sensing with Hit Rates between 51-98%. DFO-trained models had higher performance metrics than the Landsat-trained models by 5-19%. Model calibration improved Hit Rates by up to 11%. The RF model, trained with DFO-MODIS data, had the highest success rate with an overall accuracy of 91.5% and an average hit rate of 97%. A third of the focus regions lie within the modeled floodplain in the Senegal, Souma, and Casamance River Valleys and over 100,000 people are at very high risk of flood exposure.

	Hit Rate	False Alarm Rate	Mean Error	Error Bias
Dakar	0%	0%	0.01	0.00
Fatick	90%	15%	0.13	1.09
Kaolack	89%	5%	0.06	1.23
Matam	90%	11%	0.11	1.29
Saint-Louis	87%	11%	0.11	1.07
Sédhiou	95%	5%	0.04	1.18
Ziguinchor	98%	11%	0.05	1.11

Table 6: Accuracy rates across several metrics indicate how the machine learning model succeeded in each of the test watershed in Senegal.

3. Social Vulnerability to Disaster in Senegal

Social conditions that make one community more likely to experience loss from a disaster – loss of life, loss of livelihood, lack of recovery – is critical to understanding the threat of and resilience to flooding in Senegal. The field of social vulnerability investigates the ways the non-physical systems of an area contribute to the population's capacity to absorb and recover from a disaster. Although social vulnerability and resilience sciences have advanced immensely in the last two decades, the social science – particularly for developing countries – lags considerably behind the geophysical study of disasters. Yet, it is possibly even more important to understand what makes developing communities vulnerable where the climatic changes are likely to hit hardest and where existing inequality is often the greatest. This chapter examines two questions: 1) what social characteristics drive vulnerability in Senegal?; and 2) which arrondissements are more likely to experience loss during extreme flooding and other fast onset disasters?

To answer these questions, the authors conducted a literature review and a factor analysis to assess social vulnerability for Senegal. This analysis was built on a sample of anonymized individual level census data from 2013, provided by Agence Nationale de la Statistique et de la Démographie du Sénégal (ANSD). Using the IPCC definition of vulnerability, the Cutter conceptualization of disaster, and our literature review of vulnerability for the region, the authors selected 19 variables that are expected to contribute to social vulnerability to flooding in Senegal and which were not correlated with each other as shown in Table 7.

Important socio economic characteristics for Senegal in general include i) a population of 3.031 million (21.7% of the population) internet users by 2015 estimates (Central Intelligence Agency, 2016), ranking 14th in Africa; ii) a low median age in highly rural arrondissements (13 years in the Naming arrondissement, in the southern part of the country); iii) a large youth population, even in the arrondissements with the highest median age (26 years in Grand Dakar and Dakar Plateau). According to a 2012 estimate by the UN Economic Commission for Africa, this places Senegal's median age below that of Africa as a whole.

We found five underlying dimensions to drive vulnerability in Senegal: 1) a lack of basic informational resources, 2) age (elderly populations), 3) disabilities, 4) dense hubs, and 5) population increase from internal migration. The resulting social risk index reveals 30 arrondissements to be the most socially vulnerable. In total, approximately 5 million people live in arrondissements that have very high social vulnerability profiles compared to other arrondissements.

There are many ways to improve this work and further develop this science in order to refine our resultant profile of who is vulnerable in Senegal and make statistical predictions of which groups are going to be more vulnerable. To mention a few: i) incorporating big data options like cell phone data to provide information that the census cannot, and ii) eliciting stakeholder input. In terms of adding cellphone data, these resulting population estimates may add temporal scales so the authors can tell how vulnerability changes seasonally or even hourly. Cell phone data may also add finer spatial resolution and additional dimensions like social cohesion. Third and most importantly, the next phase of the social vulnerability analysis would have a strong focus on getting feedback from leaders in government, NGOs, and communities in vulnerable areas. This feedback would be used for obtaining and selecting social vulnerability variables. Chapter 5 discusses options and recommendations for conducting this local engagement.

3.1. Introduction

Disasters including floods are not just physical phenomena. They are deeply influenced by the social, demographic, economic, and political conditions of the human populations they affect. As a result, two communities hit by the same hazard will likely experience different amounts of loss in the short and long term. In the Chicago heat wave of 1995, black communities that had the same rates of violence, poverty, and were located in the same area experienced significantly different death rates (33 vs. 3 deaths for every 100,000 residents in one example); depending on how frequently their community members interacted with each other. Knowing one another – from church, talking on the street, or meeting in local stores – proved to be lifesaving (Klinenberg, 2003). When Katrina hit land, New Orleans had a relatively low elderly population. Sixteen percent of the city's residents were over 60 according to the 2005 US Census. Yet 75% of deaths from the hurricane were people in this age bracket (Brunkard, Namulanda, & Ratard, 2008). In Senegal's July 2016 floods, around 12000 people were affected; out of which more than 75% were poor farmers[6] whose crops were destroyed, putting their livelihoods at stake (ACAPS, 2016). This chapter explores several of these trends in Senegal.

The field of social vulnerability investigates the ways in which the non-physical systems of an area contribute to its population's capacity to absorb and recover from a disaster. Quantitative social vulnerability distills the social dimensions of that put people at risk into measurable numeric proxies and holistic indexes of overall risk. These social dimensions can range from economic or social conditions of a household – such as poverty status, dependency ratio, and other factors – to physical characteristics – such as disabilities, age, and gender, of an individual. Given influence factors in determining the outcome of a disaster, integrating these dimensions is critical in order to understand the threats posed to a region holistically.

The Intergovernmental Panel on Climate Change's defines vulnerability as:

> *"the propensity or predisposition to be adversely affected. Vulnerability encompasses a variety of concepts and elements including sensitivity or susceptibility to harm and lack of capacity to cope and adapt" (IPCC, 2014)*[7]

Social vulnerability is defined as the potential of a community or individual to experience loss from a hazard due to risk dimensions that are social in nature, rather than physical or ecological (Cutter, Boruff, & Shirley, 2003)[8]. After decades of research, there is some consensus in the social science research community around the demographic, behavioral, and psychological characteristics that make people and communities vulnerable at least in a general sense (Cutter et al., 2003). These dimensions of vulnerability fall along a spectrum of universality or generalizability; some dimensions are fairly well documented and consistent across geographies (Cutter et al., 2003), while others vary significantly across time, place, and context.

People who have more financial resources, who are not especially young or old, and have strong community support are less vulnerable. The elderly are vulnerable because of their health, disability, lack of transport, and lack of access to information and other resources (Ngo, 2001). Communities with a majority of their population above age 65 are likely to be more vulnerable than with a majority population between ages 30 and 45. Conversely, children, particularly infants and young children, are vulnerable because of their dependence on adults and their psychological impressionability (Peek, 2008). Crime can indicate reduced community cohesion, and prevent evacuation in rapid onset events like fires and floods. Governance may be weak in violent areas, leading to corruption of disaster aid, and preventing help from getting to those most in need (Tellman, Alaniz, Rivera, & Contreras, 2014). In many scenarios women are more vulnerable than men because of their lack of resources – both material and informational (A. Fothergill, 1996a; Neumayer & Plümper, 2007a). Psychological factors are increasingly recognized as significant at every stage of disaster response (Werg, Grothmann, & Schmidt, 2013a). Furthermore, culture has a strong influence on risk perception and requires a very local and nuanced analysis to understand, which often demands qualitative study (Adger, Barnett, Brown, Marshall, & O'Brien, 2013).

[6] http://reliefweb.int/disaster/fl-2016-000089-sen

[7] Physical hazards combine with existing vulnerabilities to create a disaster. Resilience, borrowed from ecology, broadly refers to the ability of a system to recover after shock (Holling, 1973; Pimm 1993), as opposed to vulnerability, which is typically employed to identify specific social conditions pre-disaster and define post-disaster impacts. The IPCC defines resilience as "The capacity of social, economic, and environmental systems to cope with a hazardous event or trend or disturbance, responding or reorganizing in ways that maintain their essential function, identity, and structure, while also maintaining the capacity for adaptation, learning, and transformation" (IPCC AR5 WGII). As the IPCC vulnerability definition includes "capacity to cope and adapt", it accounts for what many mean by resilience.

[8] However, these different vulnerabilities are intertwined, and their distinctions and relationships are not clearly determined in the literature.

While the dimensions of social vulnerability have mostly been explored through qualitative methods, there has been an academic effort over the last two decades to quantify these dimensions in order to estimate or even predict social vulnerability. These assessments have primarily come in the form of geospatial indices (de Sherbinin, 2014). Over the last two decades, social vulnerability researchers have begun to distill the dimensions of social vulnerability into empirically based indicators. When combined in summary indices, typically using demographic information, these tools describe who is most vulnerable and where the most vulnerable are located before, during, and after a crisis (Tate, 2012). If measured using benchmarks and monitored over time, these indicators may serve as diagnostic tools.[9]

Our methodology is based on the Social Vulnerability Index (SoVI) developed by Dr. Susan Cutter at the University of South Carolina. SoVI uses Principal Components Analysis (PCA) based factor analysis on a large set of county- or tract-level US Census variables in order to determine a set of underlying dimensions of vulnerability *e.g*. Hispanic ethnicity, special needs individuals, Native American ethnicity, and service industry employment (Cutter et al., 2003). An updated model in 2010 added new dimensions such as family structure, language barriers, vehicle availability, medical disabilities, and healthcare access in the preparation for and response to disasters. Other approaches to social vulnerability or similar themes that use other methods and datasets exist, but those are not as widely relied upon within the scientific or practitioner community. The factor analysis-based approach and other methods have been used in a small set of countries and regions around the world.

Although social vulnerability and resilience sciences have advanced immensely in the last two decades, the social science – particularly for developing countries – lags considerably behind the geophysical study of disasters. Yet, it is possibly even more important to understand what makes developing communities vulnerable where the climatic changes are likely to hit hardest and where existing inequality is often the greatest.

Qualitative and statistically descriptive assessments of Senegal consistently describe several attributes that make certain groups more vulnerable in general.

Poverty and marginalized communities, which in Senegal are chiefly concentrated in rural communities, are consistently estimated to be at higher risk and subject to a variety of other threats like violence that compound existing vulnerability. Poverty and concentration of marginalized groups is a determinant of higher social vulnerability as these groups are considered more sensitive than others and have less adaptive capacity (Alice Fothergill & Peek, 2004; Holmes, Sadana, & Rath, 2010; O'Hare, 2001).

Generally, communities with more women are more sensitive to hazards due to gendered risks and vulnerabilities (A. Fothergill, 1996a; Holmes et al., 2010; Ray-Bennett, 2009). Also, in Senegal, the level of literacy amongst women is low due to major drop-outs from school (UNESCO, 2012). Possible reasons are early marriage, teenage pregnancy, and socio-cultural norms regarding the role of women in the society. Educational progress does play a crucial role in increasing human adaptive capacity (Reid & Vogel, 2006; Tschakert, 2007). However, a high female population with low literacy can be related to high social vulnerability as an outcome.

The primary sector, viz. Agriculture, which is concentrated in rural areas (with more than half the Senegalese population), contributes 20% of Senegal's GDP. On the other hand, the secondary and tertiary sectors, viz. industries and services located in the cities, contribute 80% of Senegal's GDP. This huge disparity in income to secondary and tertiary sectors and their concentration in urban centers act as a major driver of rural to urban migration (Urban Habitat, 2014).

As noted earlier in this chapter, rural communities are at much higher risk of loss from disasters; however, in Senegal, places with the most vulnerable populations tend to be peri-urban areas, as they consist primarily of informal settlements (World Bank, 2012). People from rural areas migrate to these areas and develop neighborhoods without drainage canals and sewage systems. In Dakar, within one decade (from 1998 to 2008), around 40% of new inhabitants moved into zones of high flood potential (Geoville Group, 2009; World Bank, 2010). Besides urban/rural setting, several other characteristics make some communities or groups more vulnerable than others, specifically in Senegal. Table 7 outlines the indicators used in the report's analysis, with references to literature that supports their use in identifying vulnerable people. This report explores a quantitative vulnerability assessment of some of the generalizable dimensions of vulnerability through an exploratory model. Overall, some papers argue that inequities build into governance structures and the cultural history of the country creates a cycle of vulnerability driven by underlying systemic conditions. Vulnerability therefore goes beyond the directly measurable characteristics of communities (Sané, Gaye, Diakhaté, & Aziadekey, 2015).

Social vulnerability is critical to understanding the threat of and resilience to flooding in Senegal. Some research on vulnerability in the country argues that social dimensions of vulnerability were more critical to the overall stability of Senegal than pressures from environmental and climatic changes. "Compared to this pervasive manifestation of social vulnerability, climate extremes appear to be a minor hazard, although the recent heavy rainfalls did significantly disrupt rural livelihoods" (Tschakert, 2007).

[9] The more contextual dimensions of vulnerability, the third category presented above, do not lend themselves to generalizable proxies like census data and are therefore cannot be full captured by a quantitative index.

To examine the social nature of flooding risk in the country in this report, the authors ask:

1. *What social characteristics drive vulnerability in Senegal?*
2. *Which arrondissements are more likely to experience loss during extreme flooding and other fast onset disasters?*

3.2. Methods

Social vulnerability cannot be measured directly, at least in full, so scientists use variable proxies that can been directly measured and monitored in order to model underlying relationships, both positive and negative (Cutter et al., 2003). This report uses a common factor analysis-based approach to assess social vulnerability for Senegal. The goal of this model is to reduce the measurable (and available) characteristics of Senegal to the latent dimensions that may determine social vulnerability to disaster.

3.2.1. Data for Social Indicators

Through a partnership with Data-Pop Alliance and the Agence Nationale de la Statistique et de la Démographie du Sénégal (ANSD), the authors were given access to Senegal's official census, the Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage (RGPHAE) (Agence Nationale de la Statistique et de la Démographie du Sénégal, 2013). The ANSD supplied documentation about the census collection process, including the original questionnaires and census workers' handbook, and provided support on accessing and understanding the data.

The 2013 edition of the RGPHAE was conducted over the 21-day period from November 19 to December 14 of that year, and collected information at the household level (on a variety of topics including family structure, asset ownership, agricultural practices, and living situation) as well as detailed information about each individual living in the household (such as demographic information and education/work history). The ANSD provided access to a 10% sample of all responses, resulting in a dataset of 145,952 household records comprising 1,245,551 individual inhabitants (roughly 1/10th of Senegal's population of 14 million people).

At the third administrative level (generally referred to as "CAV"), Senegal is mainly divided into arrondissements, but also into areas called communes and villes (generally large towns and cities, respectively), which are administered separately from arrondissements. However, each commune and ville generally shares boundaries with an arrondissement that it has historically been associated with, and it is not uncommon to refer to all areas of 3rd administrative level simply as "arrondissements". For the purposes of this paper, the authors take "arrondissement" to mean a 3rd administrative level area that includes an official arrondissement and its contiguous towns and cities, to ensure that every census record can be geolocated to a single arrondissement. Using a combination of spatial merges in GIS, shapefiles from the GADM database of Global Administrative Areas (GADM, 2015), and Senegal's laws on changes in the administrative division of the country (République du Sénégal, 1996, 2013), the authors were able to associate each census response with one of the 122 arrondissements that existed at the time the census was undertaken.

The household and individual results from the census were used to build 25 indicators relevant to social vulnerability. Using the IPCC definition of vulnerability introduced in Section 3.1, the Cutter conceptualization of disaster and the authors' literature review of vulnerability for the region, the authors selected available variables from the Senegalese census that are expected to contribute to social vulnerability to flooding in Senegal. Each variable has been known to contribute to vulnerability in the region or more generally (see Table 7) based on an extensive literature review by the authors. Certain indicators were drawn directly from responses to a specific question in the census (ex: Are you male or female?); others were built from combinations of responses to multiple questions (ex: Are you male or female? + Are you the head of the household?). All indicators (with the exception of population density) were built at the individual/household level; and later aggregated at the arrondissement level by taking the arithmetic mean (for numeric indicators) or the percentage of "true" values (for binary indicators). Population density was calculated by dividing the surface area in the GADM shapefiles by the number of individual census responses. Table 7 summarizes the selected indicators, their origin, and supporting literature.

3.2.2. Demographic Profile of Senegal

Senegal has a population of 13.1 million people, of which about 54% live in rural areas. The country is characterized by growing youth population (Janneh, 2012), with a median age of 18 country-wide as shown in Figure 11.

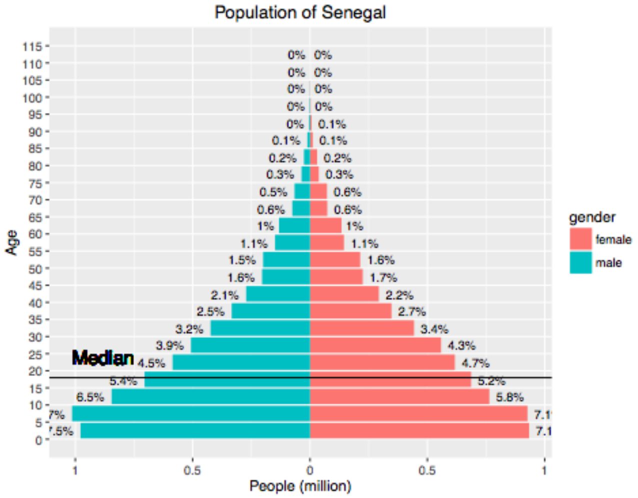


Figure 11: Population attributes of Senegal

Median age is especially low in highly rural arrondissements (13 years in the Naming arrondissement, in the southern part of the country), but even the arrondissements with the highest median age (26 years in Grand Dakar and Dakar Plateau) have a large youth population.

There are a half dozen major language groups, the largest of which are Wolof (38.7%), Pular (26.5%), and Serer (15%). Roughly one third of the population can read and write at least one language, but rates range a lot from one arrondissement to another: for instance, in Almadies (a wealthy part of Dakar inhabited by diplomats and expats) roughly 60% of inhabitants can read and write, whereas the rate is as low as 7% in Koulor (in the Tambacounda region). Higher literacy rates are observed in urban areas and their peripheries, especially in the west of the country near the cities of Dakar, Thiès, and Ziguinchor, and to a lesser degree around the eastern city of Tambacounda. Unemployment also ranges widely, from 35% in Fafacourou (a highly rural department in the southern region of Kolda) to 8% in Cabrousse (a coastal village in the south west of Senegal), with a national average of 16%. Generally speaking, unemployment is concentrated in the northeastern part Senegal and, to a lesser degree, in the central southern portion of the county (Central Intelligence Agency, 2016).

The country has 3.031 million (21.7% of the population) by 2015 estimates ranking 14th in Africa for number of internet users (Central Intelligence Agency, 2016). The percentage of people who have internet and a computer in their household is below 10% in all but 5 arrondissements; of those 5 arrondissements, all are located in the Dakar region. However, mobile phone ownership is widespread, with 81% of households reporting that they own a mobile phone. At the arrondissement level, household mobile phone ownership ranges from 45% in the Dar Salam department (in the Kédougou region in southeastern Senegal) to 92% in Almadies (a wealthy part of Dakar).

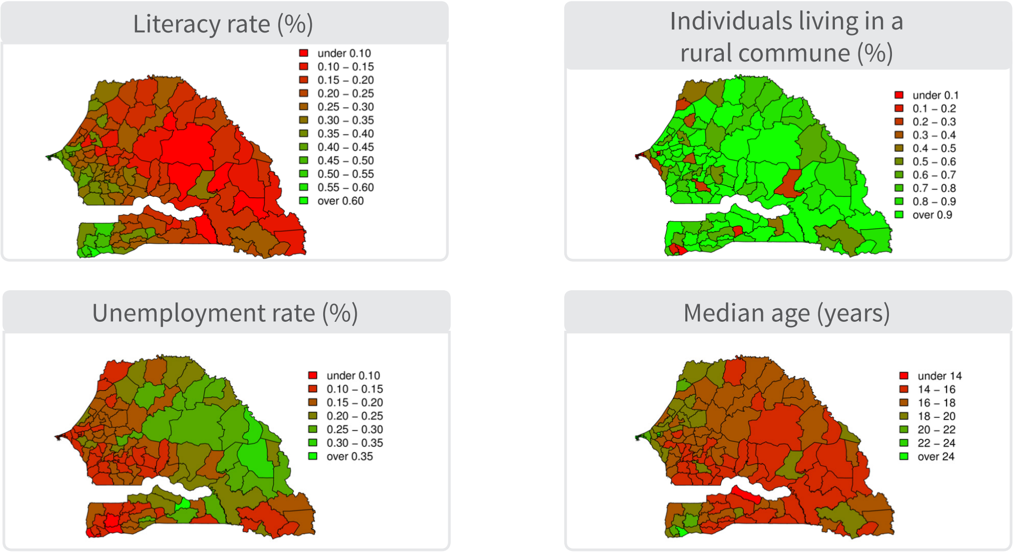


Figure 12: Spatial demographic profile of Senegal (by arrondissement): (a) Literacy rate; (b) Urban/Rural distribution (c) Unemployment rate, (d) Median age.

3.2.3. Spatial hotspots of Key Dimensions that Affect Social Vulnerability

In order to understand spatial vulnerability outcome profile, it is important to gain an understanding of how key social and economic dimensions vary across the geographic space. Here, the authors carried out a spatial hotspot analysis using the ESRI ArcMap Gettis-ord Gi* statistic on the following variables: pop_size, pop_density, pct_female, pct_skipped_meal_7days, pct_unemployment, pct_information_internet. Descriptions of the variables referred to in this section can be found in Table 7. The authors find that population size and density (given by the pop_size and pop_density variables, respectively) and access to resources (proxied through the pct_information_internet variable) are concentrated in the western part of the country where major cities such as Dakar and Thiès are located (Figure 13). In addition, the western region has higher food security, as the pct_skipped_meal_7days variable indicates a cold spot in the region. Similarly, the western region shows high employment (based on the pct_unemployment variable). On the contrary, the eastern region has low employment levels and the entire region is identified as a hotspot of unemployment. Interestingly, the pct_female variable does not show a major hotspot in the country, suggesting that women are not concentrated in any one region but rather the levels are randomly distributed across geographic space. In summary, these hotspots suggest that there are regions in the country where a particular social or economic dimension has significant spatial relationship. However, this does not suggest that the social vulnerability outcome profile will display emergent spatial patterns, but rather that emergent patterns are likely.

Hotspots of key variables that affect vulnerability

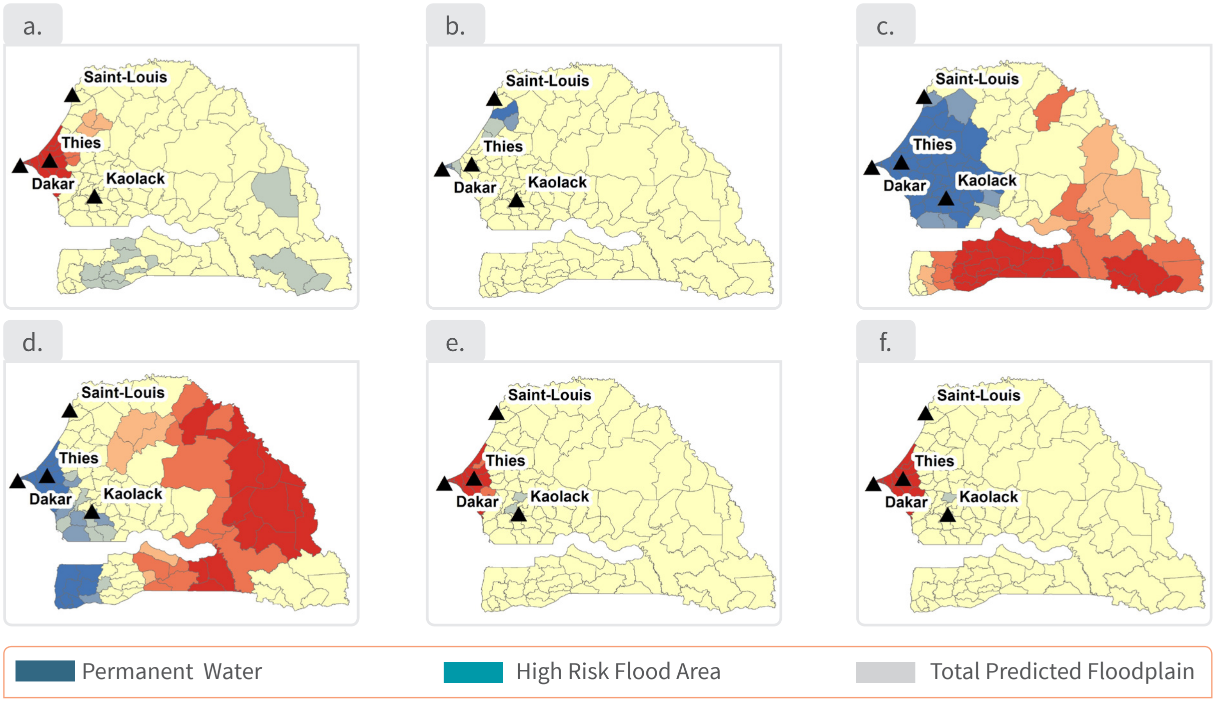


Figure 13: Spatial hotspots of key variables that affect vulnerability, measured at the arrondissement level: a) Population size; b) Percentage of females; c) Percentage of households where a member missed a meal in the last 7 days due to lack of resources; d) Percentage of unemployed individuals; e) Percent of households with access to internet and a computer; f) Population density

3.2.4. Variable Selection

Using the criteria descried in Section 3.2 of this chapter, the authors initially selected 25 variables from the aggregated ground census data to assess social vulnerability in Senegal (as shown in Table 7). In order to reduce redundancies in the data, the authors performed pair-wise comparison of variables to identify multiple collinearity in the data and dropped some variables on the basis these pair-wise correlations. The authors removed the variables iteratively until the selected variables were sufficiently non-redundant. Specifically, this was achieved with two iterations. In the first iteration, the authors identified variables that were correlated with at least three other variables with a correlation coefficient greater than |0.7|. Here, the authors dropped the pct_child, pct_female_hoh, and pct_top_quantile_children variables. During the next iteration, the authors used the criteria to identify variable pairs with correlation coefficient greater than |0.8|. On this basis, the pct_literacy and pct_difficulty_bathing variables were also dropped. After our iterative pair-wise comparison, the authors selected 19 variables with to carry out Principal Component Analysis (PCA) based factor analysis.

Dimension	Name of Variable in model	Description	Data Source	Literature Reference
Crowded	avg_num_residents	High number of people per household	Senegal Census	(Brenkert & Malone, 2005; Maheu, 2012; Mbow et al., 2008)
Population density	pop_size OR pop_density*	people per arrondissement people per km2	Senegal Census & GADM	(Brenkert & Malone, 2005; Gencer, 2013; Mbow, Diop, Diaw, & Niang, 2008; Rufat, Tate, Burton, & Maroof, 2015)
Female	pct_female*	High proportion of female persons	Senegal Census	(Chatterjee & Sheoran, 2007; A. Fothergill, 1996b; Holmes et al., 2010; Neumayer & Plümper, 2007b; Reid & Vogel, 2006; UNESCO, 2012)
Age (youth)	pct_youth	% of children below 4	Senegal Census	Sané, 2015*; Newport and Godfrey, 2003** ; Chatterjee and Sheoran, 2007**
Age (elderly)	pct_elderly*	% of people over 45	Senegal Census	(Maharaj, 2012; Mbaye, Ridde, & Kâ, 2012; Ngo, 2001; Parmar et al., 2014)
Female-headed households	pct_female_hoh		Senegal Census	(UNESCO, 2012)
Youht-headed households	pct_child_hoh	Households headed by people 14 or younger	Senegal Census	(International Monetary Fund, 2010; Vanderbeck & Worth, 2015)
Mother with many dependents	avg_n_infants* OR pct_top_quantile_children	Woman with infants OR a high % of people with an extreme number of childeren based on the country average	Senegal Census	
Disability (vision)	pct_difficulty_vision*	High difficulty seeing	Senegal Census	(Chatterjee & Sheoran, 2007; Drame & Kamphoff, 2014; Jonkman & Kelman, 2005)
Disability (hearing)	pct_difficulty_hearing*	High difficulty hearing	Senegal Census	(Chatterjee & Sheoran, 2007; Drame & Kamphoff, 2014; Jonkman & Kelman, 2005)
Disability (mobility)	pct_difficulty_walking*	High difficulty walking or going up stairs	Senegal Census	(Chatterjee & Sheoran, 2007; Drame & Kamphoff, 2014; Jonkman & Kelman, 2005)
Disability (memory)	pct_difficulty_memory*	High difficulty with memory or concentration	Senegal Census	(Chatterjee & Sheoran, 2007; Jonkman & Kelman, 2005)
Disability (personal care)	pct_difficulty_bathing	High difficulty when caring for his/herself e.g. bathing	Senegal Census	(Chatterjee & Sheoran, 2007; Jonkman & Kelman, 2005)
Education	pct_undereducation*	People 15 or older with less than a 5th grade level education	Senegal Census	(Brenkert & Malone, 2005; Drame & Kamphoff, 2014; Tschakert, 2007)

Dimension	Name of Variable in model	Description	Data Source	Literature Reference
Social cohesion/ community stability	pct_migration_internal* pct_migration_external*	Communities in which many members moved in the past year Communities in which many members moved out the past year	Senegal Census	(Kahn et al, 2003; Kane et al., 1993; Newport & Jawahar, 2003; Werg, Grothmann, & Schmidt, 2013b)
Literacy	pct_literacy	% of people who can read and write	Senegal Census	(Brenkert & Malone, 2005; Rufat et al., 2015)
Unemployment	pct_unemployment*	% who want to work but do not have a job	Senegal Census	
Access to Information (stationary source)	pct_information_stationary*	% of Households with a radio, landline or Television	Senegal Census	(Ngo, 2001; Tschakert, 2007)
Access to Information (mobile phone)	pct_information_mobile*	% of Households with a Mobile Phone	Senegal Census	
Access to Information (internet)	pct_information_internet*	% of Households with a computer/laptop and Internet/Wifi	Senegal Census	(Jonkman & Kelman, 2005)
Wealth income/ resources	pct_skipped_meal_7days* OR pct_skipped_care_12months*	% of Households where a member had to skip a meal in the last 7 days OR failed to receive care in the last 12 months due to lack of resources	Senegal Census	(Brenkert & Malone, 2005; Alice Fothergill & Peek, 2004; Sané et al., 2015)
Rural (population)	pct_rural*	% of community that is rural	Senegal Census	(Brenkert & Malone, 2005; Tschakert, 2007)
Note: Variables that are marked with "*" are finally included in Factor Analysis.				

Table 7: Input variable matrix and literature review

3.2.5. PCA Based Factor Analysis

Principal Component Analysis (PCA) is a statistical dimension reduction algorithm which uses an orthogonal transformation technique to convert a set of correlated variables into a new reduced set of uncorrelated variables. These new sets of uncorrelated principal components can be used to summarize the original data based on relatedness between different variables (Cutter et al., 2003). Each variable, both in the original data with pre-selected variables and the newly-derived uncorrelated set of factors, should in some way affect the social vulnerability outcome. While variables in the original data can be labeled in how they affect social vulnerability based on existing literature and ground knowledge, the factors obtained from PCA-based factor analysis needs to be reinterpreted in order to determine their relationship with social vulnerability. Once these relationships are identified, the factors are rescaled by applying a directional (multiplying by a value of -1 or +1 depending upon how the given factor is related to social vulnerability) and a summation of factors will then reflect the final social vulnerability scores.

In this study, the authors performed PCA-based factor analysis using *varimax* rotation with the selected 19 variables, after reducing multi-collinearity in the data, in the R programming platform (Revelle, 2016). Of the several principal components obtained, the authors selected components or factors that explained maximum variability in the data. Here the authors used a scree plot, which shows the relation between Eigen values and the number of factors considered. the authors found that for five factors Eigen values remained greater than one. Factors with Eigen values less than one are unstable and have much less variability, owing to the fact that in PCA the first few components account for a significant majority of the variation in the original data. Thus, the authors selected these five factors considering the cut-off value of 1 (Figure 14). These five factors explain ~69 % of the variation in the original dataset.

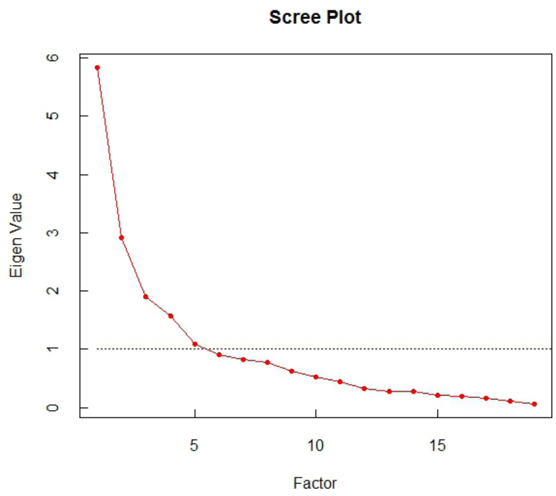


Figure 14: Scree plot showing Eigen values and factors obtained from PCA. Blue dotted line shows the threshold considered for factors selection.

3.3. Results

Factor	Interpreted name	% Variation explained	Governing Variables	Correlation coefficients	Expected relation with Social Vulnerability (Directional)
1	Lack of basic and informational resources	17%	pct_skipped_meal_7days pct_information_stationary pct_skipped_care_12months pct_information_mobile	(-)0.91 0.79 (-)0.76 0.71	-
2	Elderly population	15.5%	pct_elderly	(-)0.85	-
3	Disabilities	15%	pct_difficulty_walking pct_difficulty_hearing pct_difficulty_vision pct_difficulty_memory	0.9 0.89 0.84 0.71	+
4	Dense hubs	15%	pct_information_internet pop_density pct_rural	0.78 0.77 (-)0.71	+
5	Population increase from internal migration	6%	pct_migration_internal	0.82	+

Table 8: Key dimensions of Social Vulnerability in Senegal

A significant variation among the 19 input variables, which the authors identified to be closely related to the social vulnerability outcome, is captured by five factors. These factors will differentiate administrative units considered in this study based on their relative social vulnerabilities as determined from the underlying dimensions of the data. Table 8 lists all five factors with the percentage of variation in the original data they, their governing variables, and their significant correlates in the original data. After examining the nature and direction of correlation of each factor with individual variables in the original data, the authors determined their expected relationship with the final social vulnerability.

Lack of basic and informational resources

The first factor, which describes most of the variation among the variables, captures social vulnerability due to food insecurity, access to resources in terms of healthcare, and information from stationary sources like television and radio. This factor is strongly correlated with the pct_skipped_meal_7days variable (percentage of households where someone had to skip a meal in the previous week due to lack of resources). The greater the percentage of people skipping meals due to resource unavailability, the more vulnerable a region may be to environmental hazards like flooding, which will likely further exacerbate and disrupt the ability to grow or purchase food. Two other dominant variables explaining this factor are pct_skipped_care_12months (percentage of households where a member failed to receive health care in the previous year due to lack of resources) and pct_information_stationary (percentage of households with access to a radio, television, or landline phone).

Access to regular health care (Sané et al., 2015) reflects higher resilience of the population towards natural hazards. In the absence of accessibility to food and healthcare, the population is more socially vulnerable, especially since disaster may further disrupt access to medicine, which may be in even higher demand during flood periods due to new diseases that can occur. Our first factor variable is negatively correlated with the pct_skipped_meal_7days and pct_skipped_care_12months, meaning that a higher value of factor variable is associated with fewer skipped meal and healthcare needs; and therefore this factor is negatively associated with the social vulnerability outcome, meaning that a higher value of factor variable is associated with lower social vulnerability. In addition to the aforementioned two variables, the factor variable is also positively loaded with access to information (Tschakert, 2007), which reflects higher resilience of the population towards environmental hazards including flood events, for example if they lack access to flood warning. This suggests that, in general, the population that is otherwise deprived of access to food and healthcare services still has access to sources of gaining information. The same populations that lack access to information, which can be important in early warning, may also be unable to store resources to draw upon in a disaster if they have trouble meeting basic food and health needs. Overall, this factor variable explains 17% of the variation in the data.

Elderly Population

Our second factor explains 15.5% of the variation in the data. Note that this factor is negatively correlated ($\rho = -0.85$) with the pct_elderly variable (percentage of people age 45 or older). Overall, the elderly are more sensitive to environmental risks and have less overall adaptive capacity, which makes them more vulnerable to environmental disasters (Filiberto et al., 2009). In Africa specifically, aging is closely related to increased vulnerability (Parmar et al., 2014) due to high illiteracy rate, informal employment, and early retirement, particularly in rural areas. Informal employment, which is predominant in the case of women, usually does not provide any pension arrangements. According to (International Monetary Fund, 2010) only 17% of elderly people in Sub Saharan Africa receive pension. Therefore, the authors expect that a higher value of this factor is associated with lower social vulnerability.

In general there is more demand for medical facilities from older populations (Maharaj, 2012) and low availability of such facilities in poor countries like Senegal (Leye et al., 2013) increases risk to this group. In Senegal, the “Plan Sesame” was introduced in 2006 to reduce social vulnerability (Mbaye et al., 2012) and provides free access to public healthcare services to elderly people of 60 years and over (Maharaj, 2012; Parmar et al., 2014). These trends are further indicative of the high social vulnerability of elderly populations during flood periods.

Disabilities

One in every ten children in Africa has some type of disability (Drame & Kamphoff, 2014). Disabilities make people vulnerable as they may pose barriers to their participation in mainstream education and employment. Furthermore, disabilities increase the risk of loss during floods in poor countries where government do not have enough capacity and resources to take special steps to evacuate and prepare this population. Nontheless, several steps have been taken in Senegal to close the gap in education between disabled populations and others. For example, the Senegalese parliament passed a law of “Social Orientation” in 2010, giving children and youth with disabilities a right to free education in their nearest neighborhood school with mainstream school settings (ACPF, 2011; Plessis & Reenen, 2011); however, considering the poverty level in Senegal, this “Education for all” vision will likely take time to enact and implement.

Our factor variable takes into account disabilities related to hearing, walking and vision, explaining 15% of the variation in the data. The pct_difficulty_walking and pct_difficulty_hearing variables (percentage of people who experience high levels of difficulty with walking/stairs and hearing, respectively) are the dominant variables loaded in the factor variable, followed by pct_difficulty_vision and pct_difficulty_memory (percentage of people who experience high levels of difficulty with vision and memory/concentration, respectively). All four variables are positively correlated with the factor variable, suggesting that a higher value of the factor variable is associated with a higher percentage of population with the aforementioned disabilities. Keeping in consideration the present socio-economic and political conditions in the country, people with disabilities are generally more sensitive to environmental hazards, because it can be difficult to receive evacuation instructions, rebuild livelihoods quickly, or take advantage of relief programs and compete with others for resources. For these reasons, disabled populations may to a certain extent have less adaptive capacity, eventually leading this group to be more socially vulnerable than others (Drame & Kamphoff, 2014). This also suggests that a higher value of the factor variable represents higher social vulnerability.

Dense Hubs

This factor describes “Dense Hubs” , viz. highly connected areas, and explains 15% of the variation in the input data. The dominant variables explaining this dimension of vulnerability are pct_information_internet (percetage of households with access to a computer and/or internet), pct_rural (percentage of individuals living in a commune classified as rural) and pop_density (population density). The pct_information_internet and pop_denisty variables are positively loaded, indicating that the factor represents high population density areas with good internet connectivity. The pct_rural variable is negatively loaded, which reinforces the fact that this factor captures information on dense urban regions. With regions of high population density, exposure to flood risks would be higher. Urban areas with lack of urban planning increases the likelihood of a disaster event associated with floods (Gencer, 2013; Mbow et al., 2008; Sané et al., 2015). This is due to the fact that these areas are also poorly managed and governed when it comes to policies for integrating rural and urban areas (Urban Habitat, 2014). For example, in Dakar, an increased rate of urbanization has resulted in an increase in informal settlements, which covers almost 30% of the urban areas. Furthermore, unplanned urban areas, with lower rents or housing costs, remain attractive to low-income households or poor rural migrants to settle in hazard-prone flood zones (Maheu, 2012; Simon, 2010). Lastly, low quality building material, poor transportation networks, and lack of healthcare facilities make these areas more vulnerable. For example, housing can easily collapse, evacuation can become difficult, and often post-flood epidemics are not dealt with adequately. Therefore, this factor is positively associated to social vulnerability.

Besides the explanation given above, there is also an overlap between vulnerability and urban poverty, as asserted in the urban vulnerability science literature (Gencer, 2013). Nevertheless, not all poor people are vulnerable to disasters and often times people who are relatively rich are vulnerable as well. Social demographics do not completely determine vulnerability outcomes, as human choice can drive community development and interact with general sociodemographic variables.

Population Increase from Internal Migration

This fifth and last factor describes the social vulnerability of a community as affected by migration of a significant number of people in a year. These migrants may be returning family members, distant relative visiting, or in general a new family migrating into the community. This factor explains 6% of the variation in the data with a correlation coefficient of 0.82. With a community experiences an increase in incoming migrants, it may become less stable and more exposed to various threats like occupational hazards from specific jobs and communicable diseases, which migrants often bring with themselves (Kahn et al, 2003). A study carried by Kane et al. (1993) found that 27% of male Senegalese migrants were HIV positive against 1% of the non-migrants Senegalese males in the same area. Thus, this factor positively affects social vulnerability. On the other hand, migration can represent increase in off-farm income, and this cash flow and increased adaptive capacity help diversify income, and can enable investments to reduce risk.

3.4. Social Vulnerability Profile of Senegal

Social Vulnerability index is generated for Senegal using principal component analysis based factor analysis with nineteen indicators. It is classified into four categories: very low, low, high, and very high. Figure 15 shows the spatial patterns of social vulnerability for Senegal that the authors identified. Out of 122 arrondissements in Senegal, the resulting social risk index reveals thirty arrondissements to be the most socially vulnerable, i.e., with very high social vulnerability profile (Table 9). In total, the authors found that roughly 5 million people live in arrondissements that have a very high social vulnerability profile.

Our analysis showed that very high social vulnerability profiles in Senegal were mostly concentrated in arrondissements with major cities such as Dakar, Thiès, Kaolack, Ziguinchor, and others, as well as arrondissements located near these cities. In Senegal, urban population living in peri-urban areas has often been identified as the most vulnerable group to natural disasters. Due to uneven income distribution and the fact that major industries are located chiefly in cities, rural to urban migration is quite common in Senegal. Rapid and large-scale rural-to-urban population migration leading to unplanned urban expansion has been identified as a major driver for changes in regional hydrology leading to flooding in Senegalese cities such as Saint-Louis, Kaolack, Tambacounda, and Dakar (World Bank, 2012). In fact, the rural-to-urban migrants living in outlying areas of major cities are deprived of urban infrastructure and amenities and are often counted under rural population; Dakar is a good example with more than 30% of rural population. Furthermore, in these rapidly urbanizing regions, often the changes in policies lag behind the rate of rural-urban migration. This lagging political response to high rates of migration forces the migrant population from rural areas to end up residing in the outskirts of the cities, which are mostly low lying flood-prone areas and prohibited construction zones. Gradually, the migrant population developed urban neighborhoods which are generally deprived of proper drainage and sewage systems. These complex migration and urban expansion dynamics have led to increased social vulnerability of the Senegalese population. This is best exemplified by the heavy rains of 2012 that resulted in a major flood disaster due to the combined effect of climate change and this unplanned development. Essentially, unplanned and unorganized construction in the outlying areas of the city changed the regional hydrology, resulting in the obstruction of water flow towards the ocean.

Besides the high social vulnerability of regions in or near major cities or towns, the authors identified a few arrondissements in central, northern, and south-eastern Senegal that have very high social vulnerability. Our in-depth analysis of the data suggests that the northern arrondissements are hotspots of population with different kinds of physical disabilities. Our analysis, however, does not identify the reason for which these arrondissements have a relatively higher proportion of people with physical disabilities compared to other arrondissements. Nonetheless, in addition to a relatively greater proportion of population with disabilities, the authors identified several arrondissements in the central, northern, and south-eastern Senegal with lack of access to food, healthcare, and information, leading to very high social vulnerability profiles. The central and northern arrondissements, however, especially, have higher elderly populations that are generally more vulnerable to natural disasters. A key unifying feature of the central, northern, eastern arrondissements that the authors have identified to have very high social vulnerability is that these have predominantly rural characteristics. In several of these arrondissements farming is the principal occupation. As previous literature suggests, flooding in Senegal has severely affected agriculture and thus the livelihood of the population (ACAPS, 2016).

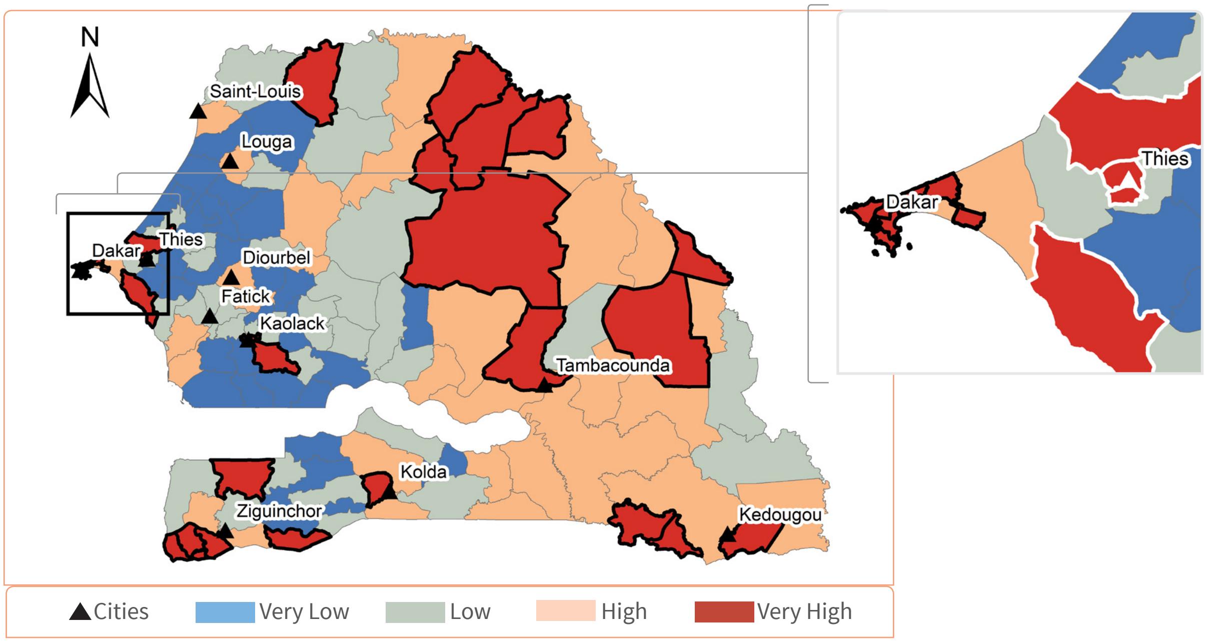


Figure 15: Social vulnerability profile for Senegal generated using factor analysis of select social vulnerability indicators. Locations of few cities have been shown in the map for reference.

S.No.	Region	Arrondissement	Estimated Population (2015)
1	Dakar	Almadies	290100
2	Dakar	Dakar Plateau	358000
3	Dakar	Grand Dakar	342400
4	Dakar	Parcelles Assainies	447600
5	Dakar	Guediawaye	496900
6	Dakar	Niayes	238900
7	Dakar	Pikine Dagoudane	421800
8	Dakar	Rufisque	217600
9	Kédougou	Fongolimbi	17600
10	Kédougou	Dakately	7900
11	Kédougou	Dar Salam	14600
12	Kaolack	Koumbal	337200
13	Kolda	Sare Bidji	88700
14	Matam	Agnam Civol	79300
15	Matam	Velingara	68900
16	Sédhiou	Djibanar	107600
17	Saint-Louis	Mbane	88500
18	Saint-Louis	Cas Cas	128200
19	Saint-Louis	Salde	93100
20	Tambacounda	Moudery	98600
21	Tambacounda	Boynguel	45000
22	Tambacounda	Koussanar	48800
23	Thiès	Sindia	412800
24	Thiès	Thiès Nord	135800
25	Thiès	Thiès Sud	149900
26	Thiès	Pambal	126200
27	Ziguinchor	Sindian	58100
28	Ziguinchor	Cabrousse	18000
29	Ziguinchor	Loudia Ouolof	24000
30	Ziguinchor	Nyassia	16900

Table 9: The 30 most social vulnerable arrondissements.

3.5. Data Limitations

Though the dimension reduction conducted in this report was very successful, the conclusions that the authors can make from the results are limited for several reasons. Variable selection is arguably the most important part of a PCA-based approach. The data used in a social vulnerability index needs to be current, robust, and qualitatively verified by experts. Likewise, the outputs must be interpreted and used with intimate knowledge of the region and with the statistical limitations of the data in mind.

There could be important dimensions of vulnerability that were not represented as variables in the available data. The authors suspect these omitted variables could be important to social vulnerability in Senegal and other contexts. There are most likely more aspects that the authors are not aware of but that are critical for social vulnerability in the region. However, with the comprehensive nature of the available data, all the major aspects and dimensions of social vulnerability have been incorporated.

Beyond these data considerations, it is critical to note that PCA is not a predictive statistical model. The results presented here describe the natural groupings of the variables input into the model and not validated externally with disaster outcome data in a statistical sense. However, this essential point reinforces the importance of variable selection for this approach to social vulnerability. Again, the model is only as good as the variables that go into it and only describes the information it is given.

3.6. Recommendations for Application and Further Research

Of the several ways to improve the scientific understanding of social vulnerability, the most important is incorporating more, locally contextualized input data. As the authors explain in the introduction of this section, many of the factors that make communities vulnerable differ widely across different cultural, political and other contexts, and at different stages of the disaster cycle. Many of those differences and the variables therefore necessary to include in an social vulnerability assessment can only be detected through local knowledge. For instance, what qualifies as relatively low-income varies between communities and across time. Also, culture has a strong influence on risk perception and requires a very local and nuanced analysis to understand, which often demands qualitative study (Adger et al., 2013).

The variables for a social vulnerability index must be constructed and reviewed in collaboration or consultation with local scientists, practitioners, and/or community members. The authors recommend a deep literature review on the region, or Senegal in general, conducted in collaboration with organizations or groups like local development staff, governments, and NGOs.

To cover the full set of indicators, the authors would almost certainly have to include ground information from conventional national surveys and incorporate big data options like cell phone data to provide information that the census cannot. This may add temporal scales so the authors can tell how vulnerability changes seasonally or even hourly. It may also add finer spatial resolution and additional dimensions like social cohesion.

We further recommend the exploration of other assessment models. Primarily, other statistical techniques, such as other data reduction methods and predictive models, could be useful. Regression-based modeling is the next frontier in social vulnerability analysis (Fekete, 2009) because it ensures that the models are describing an external reality of disasters, rather than just interpreting characteristics of input data. The authors are building such a model for the U.S. over the course of 2015. This model was not created for the sake of this report for two reasons: 1) the approach has not been widely developed and verified for social vulnerability in the academic literature and 2) the geospatial damage data necessary to build such a model were not available to the authors at the time of this analysis.

Finally, two other critical potentials for moving forward with vulnerability analysis are: 1) determining the appropriate scale of analysis, and 2) exploring the responsive or even real time applications of vulnerability analysis. Though conducting quantitative social vulnerability analysis at higher spatial resolution offers new insight into the social conditions that lead to vulnerability, the geographic scale at which those new results are most meaningful remains a largely unexplored research area. The authors recommend using variograms or another scale sensitivity-analysis techniques to determine the areas in which the data is most different.

Despite its conceptual and scientific limitations, integrating the social dimensions of hazards into the disaster cycle is necessary for fully successful emergency planning and response (National Academy of Sciences, 2012). Indices can help reduce people's social risk before a disaster hits if appropriately integrated into planning and response. When fully developed, they can identify areas most in need of assistance when a disaster strikes. The appropriate index can also suggest areas most in need of recovery assistance post-disaster by knowing which communities had a low coping capacity prior to the disaster, and where they were located. A validated and fully functioning version of the model presented here may serve these purposes.

Lastly, there are a few ways to customize the social vulnerability assessment through the model based on the user's needs and interests. Some important decisions concern the variables to include, accurate interpretation of the identified factors, and assigning appropriate weights to the factors to derive final social vulnerability scores. These decisions could be made through expert consultation and incorporating decision making methodologies (Saaty, 2008).

4. Combined Socio-physical Vulnerability of Senegal

As previously explained, the authors tested our machine learning based approach to elucidate the spatial profiles of the biophysical risk in pre-selected river valleys. In fact, 33 arrondissement overlap with the Senegal, Saloum, and Casamance river valleys where the authors assessed the biophysical risk profiles. Our results showed that several of these river valleys have high biophysical risk. Specifically, there is higher population exposed to flood risk in Matam, Ziguinchor, Fatick, and Saint-Louis regions. Our social vulnerability assessment results suggest that specific arrondissements of these regions have very high social vulnerability (Table 10). For example, Sindian, Cabrousse, Loudia Ouolof, and Nyassia arrondissements of the Ziguinchor region has very high social vulnerability. Similarly, Agnam Civol in the Matam region has very high social vulnerability. There are also arrondissements that have high biophysical risk but has high social vulnerability (as opposed to very high social vulnerability). For example, Ogo, Rao, and Tendouck (and Niaguis) arrondissements in the Matam, Saint-Louis, and Ziguinchor regions, respectively, have high social vulnerability.s

While these preliminary results are encouraging in identifying regions that have high biophysical risk and high social vulnerability, a complete nation-wide assessment of the biophysical risk profiles of the entire country is necessary to yield insights into the combined social vulnerability and biophysical flood risk profiles of the country. Therefore, a key next step is to strengthen the machine learning-based approach to delineate biophysical risk to flooding for the entire country. This will allow for a combined a combined national-scale assessment of food risk and social vulnerability. The most social vulnerable arrondissements within the flood risk zones of the Senegal watersheds analyzed in this report are shown in Table 10. The map of Senegal in Figure 16 shows the combined socio-physical vulnerability to flooding in Senegal for the watersheds modeled.

Region	Department	Arrondissement	Social Vulnerability
Sédhiou	Sédhiou	Djibabouya	Very Low
Sédhiou	Sédhiou	Djiredji	Very Low
Fatick	Fatick	Ndiob	Low
Kaolack	Kaolack	Ngothie	Low
Saint-Louis	Dagana	Ndiaye Mberess	Low
Sédhiou	Bounkiling	Bona	Low
Ziguinchor	Bignona	Tenghory	Low
Matam	Matam	Ogo	High
Saint-Louis	Saint-Louis	Rao	High
Ziguinchor	Bignona	Tendouck	High
Ziguinchor	Ziguinchor	Niaguis	High
Kaolack	Kaolack	Koumbal	Very High
Matam	Matam	Agnam Civol	Very High
Sédhiou	Goudomp	Djibanar	Very High
Ziguinchor	Bignona	Sindian	Very High
Ziguinchor	Oussouye	Cabrousse	Very High
Ziguinchor	Oussouye	Loudia Ouolof	Very High
Ziguinchor	Ziguinchor	Nyassia	Very High

Table 10: Social vulnerability profiles of the select arrondissements where machine learning-based approach to estimate flood risk profiles was implemented.

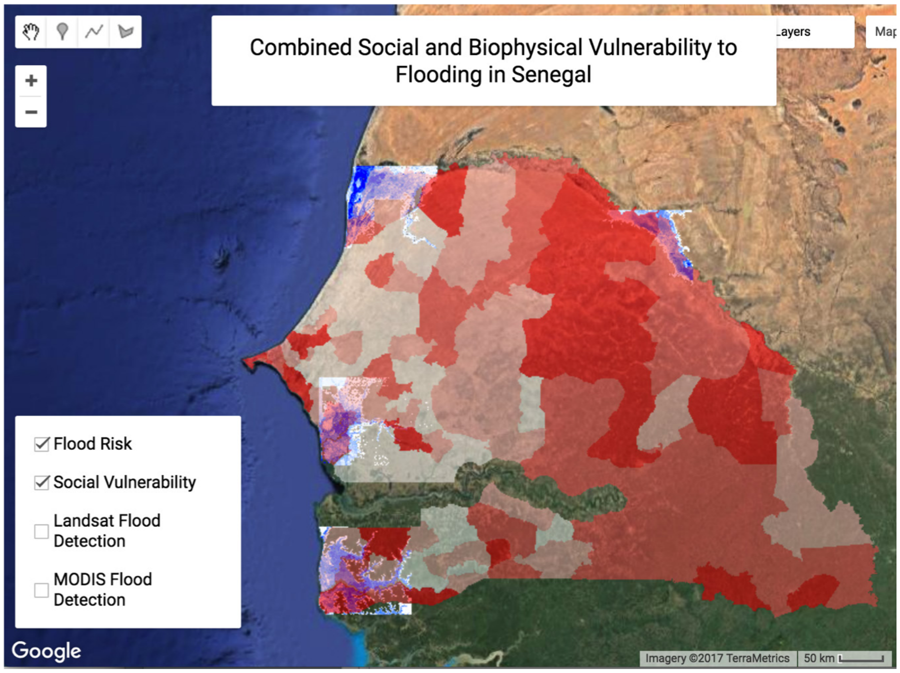

Combined Social and Biophysical Vulnerability to Flooding in Senegal
Layers
Map
+
−
☑ Flood Risk
☑ Social Vulnerability
□ Landsat Flood Detection
□ MODIS Flood Detection
Google
Imagery ©2017 TerraMetrics
50 km

Figure 16: Combined socio-physical vulnerability map for the five test watersheds of Senegal.

5. Participatory Engagement for Flood Resilience: A Blueprint for Engaging Local Senegalese in this Assessment

This chapter lays out a comprehensive methodology for engaging with stakeholders to further develop the flood vulnerability assessment described in the preceding chapters of this report. The flood vulnerability assessment was conducted remotely and derived from global or national remotely collected data sets. However, past international development and disaster risk reduction programs have found that involving local people in vulnerability and scientific assessments can be essential for improving the accuracy of the assessment as well as create significant resilience co-benefits for the communities engaged. Involving local people in risk assessment can lead to more durable response intervention programs that meet the needs of the community in question.

When implemented, stakeholder engagement enhances the accuracy, perception, and robustness of the scientific inputs of a vulnerability assessment. Stakeholders at the local level can provide researchers with detailed understanding of local flood extent and which communities are most at risk from flooding. Through this process researchers can also uncover additional information useful in estimating and predicting the flood vulnerability of the country. Secondly, beyond enhancing the quality of scientific data, stakeholder engagement can increase the preparedness of the communities involved by raising awareness of risk factors and building the durability of disaster policies and infrastructure investments.

The engagement proposed here is designed to involve community-level stakeholders in participatory ways throughout the vulnerability assessment process in order to verify the location of historic flooding, “ground-truth” the social vulnerability assessment, and add fidelity to the machine learning based physical flooding analysis. In doing so, the interactions proposed here will build the resilience of communities involved and lay the ground work for more inclusive decision-making processes between relevant national and international stakeholder groups, such as development banks, governments, and other vulnerable communities.

5.1. Introduction

We define stakeholders as those individuals or groups who can affect or be affected by the operations of an organization or project (R. K. Mitchell, Agle, & Wood, 1997). For this flood vulnerability research in Senegal, stakeholders include those parties who will contribute to, or be beneficiaries of, the flood mapping described in the previous chapters.

There are many case examples of how the quality and results of environmental planning, disaster risk reduction, and international development/aid projects are improved by including stakeholders in decision-making (Beierle, 2002; CARRI, 2013; Chambers, 1994; IIED, 2016; Pandey & Okazaki, n.d.; Pretty, Guijt, Thompson, & Scoones, 1995). The methodology outlined in this chapters builds on the success and challenges in the sector in order to provide background and a blueprint for a future participatory science strategy within the vulnerability assessment described in this report.

Stakeholder theory has its roots in many disciplines, including urban planning, international development, communications, and the business world (R. K. Mitchell et al., 1997). Thus, stakeholder engagement has many definitions. The World Bank defines *stakeholder engagement* as the process of "building and maintaining an open and constructive relationship with stakeholders [to] thereby facilitate and enhance a company's or a project's management of its operations, including its environmental and social effects and risks" (World Bank, n.d.). Regardless of the definition used, the term stakeholder engagement has become an all-encompassing reference to many processes that include (but are not limited to) stakeholder identification and analysis, outreach, communications, consultation, and partnership development (International Finance Corporation, 2007).

Engaging with stakeholders necessitates first understanding who those stakeholders are and how best to involve them in the project process. Through *stakeholder identification and analysis*, a business leader or project manager seeks to understand which individuals or groups have a stake, or vested interest, in the project at hand, and what motivates those stakeholders or entities towards action. Understanding this landscape can help a project team proactively engage with a stakeholder group and provide information or reach out to better understand the project context and risks of failure.

In international development and disaster risk reduction, stakeholder analysis is often a helpful first step in the project proposal process, and it can be a critical part of any subsequent project interactions at the local level (ODI, n.d.). The engagement methodology outlined in this chapter primarily pulls from international development examples, though there are valuable lessons and principles to learn from other disciplines.

In international development, stakeholder engagement processes are often referred to as *participatory engagement*. *Participatory stakeholder engagement* processes seek to expand beyond external stakeholder analysis that determines potential risks to the project to involve the community in the project itself. Such practices champion local knowledge and imply a two-way street of information flow: rather than simply delivering information to a known audience about a project, the development or implementing agency is listening as well (GTZ, 2007, p. 9). Such engagement helps the project keep its goals relevant to the beneficiaries and assists with capacity building or training; both elements can be critical to long-term project adoption and success. Furthermore, involving people in a project process is a way of increasing ownership and democratizing the outcomes (Tschakert, 2007).

Processes that encourage stakeholder participation have long been a core of international development. However, the quality of decisions made through stakeholder participation is strongly dependent on the nature of the process or method leading to them (Reed, 2008). In international development, approaches to stakeholder participation have evolved over time: from largely awareness raising about a given project in the community in the late 1960s, to incorporating local perspectives in data collection and planning in the 1970s, to encouraging local stakeholders to participate in the project process from start to finish (Reed, 2008, p. 2418). An important development in the evolution of participatory engagement are "rapid rural appraisal" (RRA) and "participatory rural appraisal" (PRA). These approaches aim to incorporate the knowledge and opinions of rural people in the planning and management of development projects and programs. The Institute of International Environment and Development (IIED) and the Institute of Development Studies (IDS) were early collaborating authors on guiding frameworks and methodologies for RRA and PRA methods in the 1980s and 90s (Chambers, 1994; IIED, 2016; Pretty et al., 1995). These methodologies built on the traditions of activism and anthropology and evolved into participatory learning and action (PLA) systems that have increasingly been mainstreamed in the development world and has adapted over time to take on specific thematic challenges such as community-level adaptation to climate change (IIED, 2016).

Many development agencies have developed toolkits that set forth best practice methodologies and activities for participatory stakeholder engagement. These toolkits provide facilitation tips and activities that strive to keep development activities inclusive and "human-centered," or focused on the needs of the beneficiary communities rather than an external agenda (GTZ, 2007; ideo.org, 2015; Pretty et al., 1995; Tekman, Hassapi, Chrysostomou, Konnaris, & Neophytou, 2012). Participatory stakeholder engagement activities include participatory planning (to include community mapping and participatory budgeting), survey and interview techniques, and educational games (ideo.org, 2015; Pretty et al., 1995).

For example, the ideo.org toolkit provides specific methodologies for gathering information in collaborative, inclusive ways within the context of a community in the developing world. They outline ways to invite and include people from across a community, as well as specific interview techniques for group settings (ideo.org, 2015). As technology changes and mobile technologies are increasingly available in rural areas, methods of participation are changing rapidly to incorporate digital survey tools and GPS devices as well as make use of cellular networks (Gordon, Schirra, & Hollander, 2011). For example, the World Food Programme's mobile vulnerability mapping project utilizes SMS and call centers to gather data on food security in remote areas via the existing phone network. Using this method WFP was able to gather 100,000 questionnaires in 2015 (Bauer, Attia, & Clough, 2016).

In the scientific community, one method of participatory stakeholder engagement is citizen science. Citizen science, also referred to as civic, participatory, or community science, is the act of involving citizens in science as researchers (Conrad & Hilchey, 2010). This inclusion of people who are not necessarily traditionally trained as scientists in data collection and research can democratize scientific processes, hold governments and companies accountable, and also increase the reach of a research team, or their ability to gather data (Conrad & Hilchey, 2010). Citizen science can either be citizen contributions to scientific studies that are ongoing, or science that is developed by and completed by citizens (Kruger & Shannon, 2000).

In both social science and physical science communities, crowdsourcing, or the act of gathering data from a large group of non-experts via survey, online, or other means, is often an umbrella term that includes citizen science practices (Lauriault & Mooney, 2014). Because of its capacity to scale data collection and/or feedback, crowdsourcing can be a useful technique for quickly gathering large amounts of data. For example, the US Geological Survey's Tweet Earthquake Dispatch utilizes crowdsourced data in the form of Twitter updates to track aftershocks following major earthquakes in real time ("Federal Crowdsourcing and Citizen Science Toolkit," n.d.; USGS, n.d.).

Disaster risk reduction is "the concept and practice of reducing disaster risks through systematic efforts to analyze and reduce the causal factors of disasters" (UNISDR, n.d.). This report focuses on the disaster risk reduction process of vulnerability analyses for the purposes of preparing communities and economies for disaster. Public stakeholder participation in disaster risk reduction planning specifically is critical for long term preparation and planning for crisis.

Integrating stakeholders into disaster risk reduction processes either via crowdsourcing or in-person workshops has several benefits. First, this gives a voice to the communities at risk from disasters (Tschakert, 2007, p. 382). Secondly, in vulnerability analyses, engaging local people early and often can help include local knowledge, contextualizing the analysis and giving fidelity to the contributing variables. In addition, such involvement also can educate members of a community community and increase their resilience by making them more aware of risks. Finally, involving stakeholders can contribute to the lasting success of any ensuing interventions as it offers opportunities for capacity building and training of stakeholders during the project to ensure that follow-up can be accomplished locally (CARRI, 2013, p. 3; GTZ, 2007, p. 5).

Beyond involving local stakeholders, it is important to think about which stakeholders are involved. "Putting the vulnerable first" and including those demographic groups such as minorities, women, youth, or others who may not be a part of conventional city or disaster risk planning process can help grow the adaptive capacity of a community and increase resilience over time (Paavola & Adger, 2006). Indeed, some believe that the primary goal of any participatory stakeholder engagement intervention should be to be inclusive of as many different stakeholders/stakeholder groups as possible in order to reach and hear from those communities that often don't have a voice in existing decision-making or governance processes (Reed, 2008). This inclusion can help ensure that these populations are accounted for in the actions that are planned to reduce disaster risk (such as evacuation plans, infrastructure improvements, etc.). In addition, such inclusion and resulting capacity building can give vulnerable people and communities the skills to respond to disasters themselves, reducing death toll when events do happen (Pandey & Okazaki, n.d.).

However, being inclusive in participatory stakeholder engagement is difficult, especially when projects are conducted by outside parties. These challenges are shown in the case of Mercy Corps' work in Nepal, where routinely men were the only ones showing up to community disaster planning meetings. This meant the "least needy," the men, were the most aware of what to do in the event of a disaster, while dependents such as women and children were not well educated on evacuation or disaster response protocols and their needs were not voiced during the drafting of disaster plans. The organization had to become more creative in its outreach, turning to using street drama performances to make sure that the women and children in the community were included in risk awareness activities (Nepali Red Cross, 2009). As this case demonstrates, in order to realistically operationalize inclusive participatory stakeholder engagement with vulnerable communities, one must consider ethical lines, cultural and gender norms, and the capacity of the implementing organizations.

5.2. Participatory Engagement for Flood Resilience in the Senegalese Context

As outlined in earlier sections of this report, flooding events in Senegal are frequent and destructive. Yet, in Senegal and across the Sahel region, drought is a much more pressing concern for many local leaders than flooding. This situation has led to information and policy response gaps around flooding (Tschakert et al., 2010). Additionally, many communities are not able to respond in order to keep themselves safe when flooding events happen. The remainder of this chapter describes a participatory stakeholder engagement process designed specifically for the flood vulnerability assessment for Senegal.

Following extreme 2009 floods, the Government of Senegal created its first recovery plan after a post-disaster needs assessment, which was conducted with the support of the international community and funded by the World Bank (The World Bank, 2014). Priority actions outlined in the report include: creating infrastructure to respond to urban flooding in Dakar and preparing a master plan for storm water management and preventing and mitigating disasters by a) developing an urban development plan containing the mapping of flood risks, b) strengthening the management of flood risks, and c) educating affected communities. This plan for education and outreach was part of Senegal's effort to create a culture of "proactive preparedness" (World Bank, 2012). It shows interest in reaching out to the community about flooding, but not explicitly in learning what the community's needs are regarding flooding events.

In 2012, further flooding inspired additional flood risk management approaches and the government launched a revised ten-year flood management program. This new program aimed to involve local officials in the flood planning process, but does not explicitly speak to local citizen involvement (The World Bank, 2014). In the Fall of 2016, extreme floods have again devastated Senegal. Outside of major urban centers, in the central part of the country, floods have highlighted a lack of communication and warning systems available as well as inadequate local awareness or capacity to react to the flooding (Trust.org, 2016). Clearly, and despite planning efforts, there remains today a missing link between local capacity to respond and government-level policy action.

Key NGOs operating in the area are helping to bridge this gap, build local capacity, and involve citizens in flood management and response. Building Resilience and Adaptation to Climate Extremes and Disasters (BRACED) has a program titled “Live With Water” that helps urban Senegalese to adapt to flood conditions safely and even use water to expand their livelihoods by launching small agricultural enterprises (BRACED, 2016). The participatory engagement the authors propose to support our flood vulnerability assessment would also strive to close this policy gap by working to connect people with data to help them lend their voice to the decisions being made around flood resilience.

In order to ensure Cloud to Street is using the best physical and social risk data, it is necessary to engage with stakeholders across the area being mapped in Senegal. Determining the physical extent and frequency of flooding events in Senegal is not the only critical input for decision-making by government, NGOs, and other planning entities; it is also critical for these decision-makers to understand the ways that people react to floods and are vulnerable because of their social standing. In addition, as described in Chapter 3, flooding in Senegal does not affect every community in the same way. To fully understand flood risk, it is critical to analyze both the social and biophysical risk in areas affected by flooding and be able to quantify and qualify social vulnerability indicators appropriately.

5.3. Methodology for Participatory Stakeholder Engagement

The participatory component of our vulnerability assessment would engage flood vulnerable populations in Senegal and include two kinds of interactions. The goal of these interactions is first to help verify or “ground truth” the spatial extent of the country’s past floods through digital and in-person engagement. The ultimate ambition of this goal is an online crowdsourcing tool that enables local communities to effectively refine the training inputs for machine learning based flood vulnerability assessments in their community, and improves the underlying algorithm overall. Secondly, multi-stakeholder engagement events will be aimed at building capacity in the digital platform, garnering feedback on social drivers of risk, and giving local partner organizations the ownership to carry flooding risk lessons into the future. These two components to the engagement strategy – digital tool and in-person workshop – work together to both improve the accuracy and precision of flooding vulnerability maps themselves, spread awareness of risk amidst affected populations, and facilitate increased adaptive capacity at the local level.

Through these two kinds of interactions, the authors aim to garner ground-level feedback and make adjustments to their data validation interface and flooding maps. In addition, these interactions would allow local partners to learn the flood vulnerability science and risk specific to their community. As a stand-alone tool, the power of a technical vulnerability assessment is small compared to the potential of empowering local decision-makers with access to big data and computing power in order to tailor their own tool-building in ways that the authors believe have the potential to transform disaster management. Once trained through the interventions described in this chapter, local citizens/ stakeholders can continue to contribute to the assessment after the official project is complete.

Recommendations and options for the twin components listed above are described below. The first steps in stakeholder engagement are site selection and an outreach plan to engage the right people. After these are complete, multi-stakeholder in-person workshops will be conducted in each site and a digital feedback method/user interface for flood extent updates will be launched and monitored over time. In addition, Cloud to Street will develop a user interface for digital data collection to integrate the new data collected into the existing assessment and deploy this via the stakeholder workshop and other external methods. For more detail on this digital tool, see Section 5.3.1.

5.3.1. On-line Participatory Digital Flood Map Verification Tool

The flood vulnerability assessment for Senegal, described in Chapters 2b, 3 and 4, relies on the fidelity of large data sets spanning remote sensing and census information. In addition to utilizing available data from the cloud, Cloud to Street intends to draw on crowd-sourced flood observations from mobile devices and/or computers to automatically update the assessment over time. In addition to the in-person workshops, a digital interface tool will be utilized to gather widespread data on the areas that flood in Senegal. This tool would allow citizens to answer the question: “does/did this area flood or not?” in order to validate the mapped predictions and keep training the data to be more accurate, thus leading to more accurate understanding of past floods and prediction of the floodplain. As mentioned above, this tool would be utilized in the workshop setting as much as possible. The platform will also be designed with an eye towards scalability so that the authors can replicate the pilots proposed in this chapter as easily as possible.

To support this data collection tool, the authors will be working with digital visualization and user experience design firms (such as Development Seed) to explore user interfaces that are appropriate for such data validation in the context of Senegal. The data validation tool could fall at one or multiple points along a spectrum of complexity – from a simple text yes/no to a more complex mobile application to a website. The aim will be to keep the technology used as well as the application itself as simple as possible to accommodate high and low tech users. It will be critical to test this tool in the field during its development (before rollout) as well as garner feedback during its use and make adjustments as necessary.

The goal will be for utilization of this tool across all areas of Senegal prone to flooding. The data from the tool will be most valuable if the tool is understood by the users, utilized in a timely and accurate manner during and after flooding events, and if the users are located throughout the flood-prone areas of the country. The workshops' training-of-trainers approach can help facilitate these processes, but the workshops themselves will not be able to recruit users from all flood vulnerable areas. Cloud to Street could work with AFD and other partners to utilize existing networks as well as Facebook and other digital outreach tools to spread the use of the tool. However, no matter how a user is introduced to the tool it is important for users to understand the importance of both reporting the information and reporting it correctly. Therefore, there would need to be easily understandable guides that explain the tool's use as well as a training-of-trainers with other potential recruiters.

As they are collected both at the workshops and over time, these data would be combined into a final web map showing both physical and social vulnerability to flooding. The user interface would enable automatic updates to not only the physical map of flooding extent (i.e. are the flood contours correct?), but also eventually to the social vulnerability layers (i.e. are the demographics of this area represented correctly in the analysis?) through addition of separate questions in the future.

5.3.2. Workshop Site Selection

A sample of communities (towns within specific arrondissements), representing the spectrum of flood risk in Senegal and/or the demographics and socio-economic profile of its population, will be selected for workshop and digital engagement. Using the socio-physical flood vulnerability risk index from this report and the demographic data provided to us by the Agence Nationale de la Statistique et de la Démographie du Sénégal, the sites can be selected to represent one or more of the categories listed below:

Most vulnerable arrondissements: Where are the areas that are most at risk from flooding (physical, social, and both) – both historic and current/future?

Highest vulnerability and lowest vulnerability locations: choosing a mix of high- and low- vulnerability locations based on past flooding events will allow for some comparison of responses.

Representative of Senegal's demographic and physical makeup: choosing a variety of terrain, location, poverty level, urban/rural makeup, etc. to represent Senegal as well as possible across all participatory data collection sites.

We propose piloting the effort in six communities in order test the method before scaling up. The initial pilot is described in the section below. Further consultation with local partners in Senegal and with AFD will be necessary to determine which sites will be included in the eventual participatory engagement effort.

Success in implementation will depend on many factors such as rate of participation. To recruit the right participants and ensure the interventions are culturally relevant and appropriate, the implementers should plan each interaction in collaboration with local organizations that are familiar with the communities and have existing projects and trust relationships with community members (e.g. BRACED, Red Cross, the UN's World Food Programme).

5.3.3. In-Person Workshops

5.3.3.1. Audience and Stakeholder Analysis

To determine the audience for the workshops and craft a participant invitation list, a comprehensive stakeholder analysis should be completed for the community. Cloud to Street would work with AFD and local partners to complete this analysis, asking for key local community groups, community leaders, local and national NGO representatives, and community members representative of a broad demographic. It is critical when analyzing stakeholders participation that local norms and power dynamics are taken into account so that all members of a community are actually represented. For example, it may be the case that one gender or social group has a more dominant place in the community, and reaching other social groups may take some creativity. This was the case in reaching women in the Nepal example cited above in Section 5.1 (Nepali Red Cross, 2009).

The result of this stakeholder analysis will be a stakeholder list or tracking document. This could include information for each stakeholder, such as the individual or group's name, the contact person and contact information, how the stakeholder could contribute to a workshop, whether there are any key considerations to keep in mind, how influential the stakeholder is over other potential participants, and the strategy for inviting the stakeholder to participate in the workshop(s). Such a tracking document is sometimes called a stakeholder map because it can show connections between the various stakeholders and stakeholder groups. Keeping this tracking list as simple as possible and utilizing a list format that is prevalent in many office scenarios (such as Microsoft Excel or Word) will make it easier to collaborate with local community groups in gathering this information and to maintain this information in the future.

After initial stakeholder audience research is complete, an invitation to participate in the workshop can be sent far in advance, through local community organizations and leaders. This can include the proposed agenda and outcome of the workshop itself, demonstrating the value-added to the participants.

5.3.3.2. Goals and Activities

One Participatory, multi-stakeholder, workshop will be conducted at each site during the course of the project process. Each workshop will be a first engagement and will lay the ground work for future digital or in-person interventions conducted by Cloud to Street or their partners. As such, the workshop will set up the data validation process and serve as an opportunity for training and strengthening of local partnerships. The format of a workshop can be adjusted to the context as needed based on stakeholder availability, anticipated attendance numbers, and participants' level of prior experience with mapping and/or participatory planning. These adjustments should be made in consultations with local community leaders.

The workshops would draw upon elements of focus group interviews and participatory mapping. Some questions that are general to the region or country can be asked of all participants to gain insight into demographic trends and risk to floods. In addition, activities that utilize maps and other visual tools in a hands-on manner can allow participants from across different education and language capacities to contribute meaningful information on the flood risk in their communities. And, finally, there can be activities that train the participants on how to use the interface of the digital tool that will collect information on the physical extent of flooding. Specific activities and the best communications tools to be utilized during the workshops can be determined in cooperation with AFD and in-country partners.

The goals for the in-person participatory stakeholder engagement workshops and a few design options for practically achieving that goal are detailed below:

Primary Goal - Data collection: *collect data on social and physical flood vulnerability in the community*. This information could be gathered through on or more of the following: map validation, participatory GIS, or focus group interview activities during an engagement workshop:

Charrette-style map validation: Present a printed copy of the machine learning flooding information map from Cloud to Street to local partners and community members. Encourage community members to mark the map with areas that are correctly shown vs. areas that lack flooding vulnerability data. Record feedback via note-taking and saving and georeferencing the map copies.

> ***Potential participation:*** 10-40 community members and local partner organization representatives.
>
> ***Data Output:*** Confirmed pixel location of historical flooding areas, identification of floods or non-flooded areas that Cloud to Street was previously unaware of.
>
> ***Additional Output:*** Increased community ownership over spatial information, relationship-building between community members and partners and with Cloud to Street, and increased risk awareness for community participants.
>
> ***Example:*** In Indonesia, participatory mapping and map validation has been used to help communities communicate spatial information to the government, including the boundaries of conservation areas. The process of participatory community mapping has helped resolve land ownership disputes and bring community members together (IFAD, 2009).

Participatory GIS: This builds on the previous activity. Participatory GIS can take place through in-person map drawing by community members in a workshop/group scenario and/or through a tour of the community with several local residents and partners with a printed map, digital map or GPS unit to mark areas that were flooded. If the latter option is chosen, a small group is preferable.

> ***Potential participation:*** 2-10 community members and local partner organization representatives.
>
> ***Data Output:*** Confirmed pixel location of historical flooding areas, identification of floods that Cloud to Street was previously unaware of.
>
> ***Example:*** In 2011 in Gorakhpur, India, a facilitation team working with the Asian Cities Climate Change Resilience Network helped community members map areas in town prone to flood risk by using GPS handheld devices and satellite imagery printouts of the area. The points were then aggregated by the team and input into an overall hazard map for the city (Singh, 2014).

Focus group-level participatory risk assessment and/or interviews: This exercise is intended to include not only physical flooding risk, but also capture information on social vulnerability. Focus groups of 10+ people can participate in interviews where facilitators ask specific questions, or they could be group participatory risk identification exercises where people identify and rank their risks from flooding. Questions could be tailored to flooding in particular and could help identify and validate social vulnerability indicators.

> ***Potential participation:*** 10-40 community members and partners.
>
> ***Data Output:*** Identification and validation of social vulnerability indicators, understanding of the thresholds for measurement for these indicators.
>
> ***Additional Output:*** Increased relationship-building and risk awareness/sharing within community participants.
>
> ***Example:*** The Climate Change Collective Learning and Observatory Network in Ghana completed participatory risk

assessment where participants, grouped by age and gender, were asked to elicit the various problems they face at the community level (free listing), write or draw them on index cards, then rank them, by order of importance, and score their severity or harm to wellbeing and livelihoods (CCLONG, 2009).

Flooding map validation interface training: As detailed in Section 5.3.3 below, Cloud to Street will work with other partner organizations to develop a simple user interface that can be used on a mobile device to validate the extent of physical flooding. This tool can be incorporated into the implementation of activities 1 and 3 above and the workshops can be used as a training-of-trainers for those community members who are willing/able to use such an interface to validate flooding in real time in the future.

Potential participation: Unlimited community members and partners.

Data Output: Confirmed pixel location of historical flooding areas, potential real-time updates during flood events.

Example (of the interface): In the US, the National Oceanic and Atmospheric Administration (NOAA) administers the iPhone and Android app "mPing," that allows citizens to submit local precipitation reports in real-time storm events. This validates and supplements NOAA's weather report data ("NSSL Projects," n.d.).

Secondary Goal: develop ongoing relationship with workshop participants and local partners. The success of community engagement work and participatory data collection depends on the strength of Cloud to Street's relationships with local partners. A primary goal for the workshops, especially at the outset of this project, will be to establish and maintain relationships with organizations and individuals identified as key stakeholders – not just for Cloud to Street, but for local government entities as well.

In addition to these primary and secondary goals, Cloud to Street aims to further the following when planning in-person community engagement and participatory activities, as possible:

Raise community risk awareness and build resilience. This can be accomplished through scenario-planning activities or games (CARE, 2011; Tompkins, Few, & Brown, 2008) that engage workshop attendees and potentially the wider community in flood-risk awareness and safety training.

Facilitate local solutions and community preparedness, utilizing methods of community-based disaster risk reduction. Community-based disaster risk reduction activities focus on capacity building, elevating community awareness of risk, and being inclusive of vulnerable people (Pandey & Okazaki, n.d.). In this context, activities can include alerting local partner organizations of small grant application opportunities, helping train local partner organizations to carry out similar workshops, and working with larger NGO and government partners to facilitate complementary events on disaster risk management during/around the time of these workshops.

Support community empowerment and inclusion in the regional and national disaster risk management planning process. Utilizing stakeholder relationships fostered through this participatory process, Cloud to Street and their partners can connect local communities with national-level decision-makers planning for disaster risk reduction so that their specific vulnerabilities are known. In addition, training local and national partners on big data technology and the design and use of the data validation tool will help further this goal.

5.4. Next Steps and Follow Up

5.4.1. Tracking Success

Effective stakeholder engagement and participatory science efforts rests on a sound foundation of research, communication, and partnership. The success of the workshops and the data validation tool hinge on the strength of local partnerships, the ability to understand community dynamics and encourage key stakeholders to participate, and the clear communication of the goals of Cloud to Street and AFD's work. In short, the success of the workshops and the utility of the tool will depend on the level of stakeholder buy-in and participation.

If conducted, the authors would assess the components of this project based on the data collection/validation goals set forth in Section 5.3. The participatory workshops will be successful if Cloud to Street is able to comprehensively validate and collect physical flood extent data and discuss necessary adjustments to social vulnerability indicators for the six workshop sites. During the events themselves, success can be measured by not only the number of people that participate, but also the quality of interaction. Particularly for the workshops, Cloud to Street can be adaptive on the ground and consistently solicit and capture feedback about what went well and what could be changed from local partners and participants. Cloud to Street plans to be adaptive and change the workshop invitees and format between workshop sessions (and during, as necessary) based on this feedback. For example, if there were too few participants at one workshop, additional outreach may be necessary preceding the next event.

With the data validation tool, success can be measured by rate of adoption as well as the quality and relevance of the data collected, and if it increases accuracy of the machine learning model. This success will depend on whether the goal of the tool is understood by the users and utilized in a timely and accurate manner during and after flooding events. Understanding the margins of error for this action will be helpful for determining the success of the data validation tool.

5.4.2. Overcoming Barriers to Success

The success of the in-person workshops and online tool depend on the stakeholder engagement effort's ability to reach the appropriate audiences and build trust. Language barriers are a potential hurdle and translation assistance will be necessary for both workshops and the data validation tool. In addition, potential demographic differences between sites could hinder success and should be considered at all stages of planning and implementation for both the workshop and the data validation tool. For example, in urban areas of Senegal, large portions of the population live in unplanned informal settlements (Diagne, 2007, p. 553). If urban sites are chosen for workshops, different outreach strategies or workshop activities may be necessary to understand the social vulnerability landscape via workshop feedback. In addition to demographic and geographic differences, Cloud to Street and partners will need to maintain awareness about cultural definitions of vulnerability and risk that may vary for each community.

For the data validation tool, understanding the online/offline divide will be critical to success. How many people have access to what technology? Providing clear training on the tool's use as well as testing to make sure it is designed with the capacities of the user(s) in mind will help avoid potential challenges.

Working with local partners and key stakeholders will help clarify these and other potential challenges and keep such considerations at the forefront of project planning. The project team will need to be adaptable over the course of the project to be able to document and respond to these differences and adjust products and outcome indicators accordingly.

5.5. Conclusions: Next Steps

Cloud to Street is committed to coupling innovations in participatory stakeholder engagement on the ground with innovations in digital mapping in the cloud and finding ways to scale the two together in a way that leaves no one behind. In addition to improving vulnerability information this project seeks to increase the overall resilience and adaptive capacity of vulnerable communities identified here, as well as enable improved equity and ownership in disaster risk management in Senegal moving forward. While this project is centered around flooding risk, many other disaster events and risks to human health and safety can be considered when providing basic risk assessment or emergency response training during a workshop, or have the opportunity to be included in the future as variables measured using the data validation tool interface. Cloud to Street seeks to actively think about the potential co-benefits of its activities in stakeholder engagement and participatory flood risk mapping, and work with communities to help them be able to continue this kind of monitoring in the future, after Cloud to Street's project work has come to a close.

Cloud to Street is at the forefront of real-time vulnerability mapping, harnessing the significant amount of spatial data now available for risk management on the ground. One of the great advantages of big data like satellite imagery is that it is easier than ever to scale such analyses for anywhere in the world. However, looking at the earth from above shows only half of the picture. Connecting satellite data to the people that are most vulnerable is a challenging step, but a necessary one in order to ensure equitable risk representation and paint the true picture of what is happening on the ground.

6. Appendix

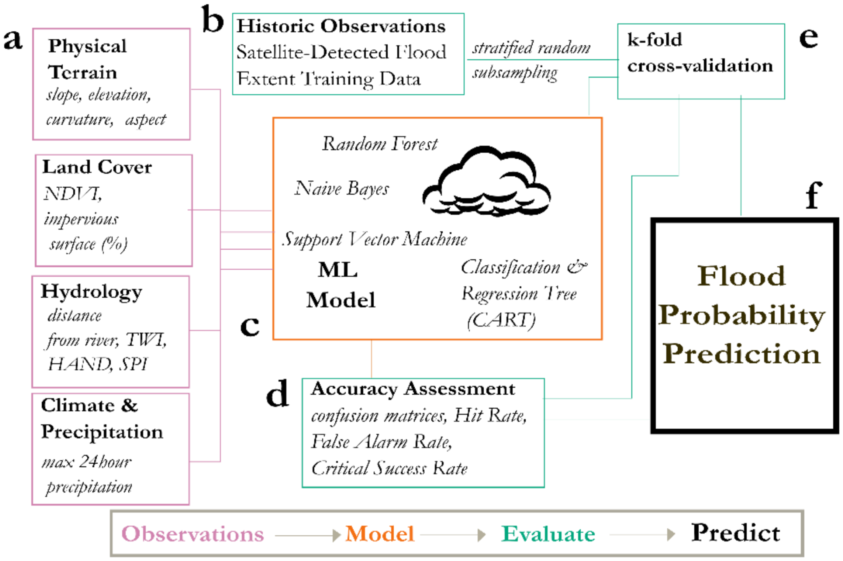

a
Physical Terrain
slope, elevation, curvature, aspect
Land Cover
NDVI, impervious surface (%)
Hydrology
distance from river, TWI, HAND, SPI
Climate & Precipitation
max 24hour precipitation
b
Historic Observations
Satellite-Detected Flood Extent Training Data
stratified random subsampling
k-fold cross-validation
e
Random Forest
Naive Bayes
Support Vector Machine
ML Model
Classification & Regression Tree (CART)
c
d
Accuracy Assessment
confusion matrices, Hit Rate, False Alarm Rate, Critical Success Rate
f
Flood Probability Prediction
Observations
Model
Evaluate
Predict

Figure A1 Model structure and overview showing the key steps involved in creating floodplain maps.

Factor	Global Dataset	Method & References
Precipitation (mm)	NOAA PERSIANN-CDR 0.25 degrees (Ashouri et al., 2015)	Precipitation summed for the duration of the flood
Impervious Surface (%)	2009 ESA GlobCover 300 m (Bontemps et al., 2011)	
Distance from River (m)	Senegal Water Lines (NGA, 2015)	where x and y are vectors of coordinates and d is Euclidean distance between two points.
Topographic Wetness Index (TWI)	WWF Hydrosheds 15 arc-second Flow Accumulation (Lehner, Verdin, & Jarvis, 2006)	TWI= Ln(a/tanβ) where β is local slope in radians, a is the upslope catchment area (Beven & Kirkby, 1979).
Stream power index (SPI)		Erosive power of stream, energy dissipation SPI= As tan β As is the catchment area (m2) β is the local slope gradient (degrees) (Florinsky, 2012)
Local Slope (degree)	SRTM (30 m) (Farr et al., 2007)	Local slope informs overland and lateral flow velocities (Moore et al. 1991)
Elevation (m)		Local elevation informs climate patterns, vegetation communities (Moore et al. 1991)
Curvature		The second derivative of the slope (Farr et al., 2007)
Height Above Nearest Drainage (HAND)	Global 30 m Height Above Nearest Drainage (Donchyts, Winsemius, Schellekens, Erickson, Gao, Savenije, & Giesen, 2016)	Digital elevation model normalized to nearest streamline.
Normalized Difference Vegetation Index (NDVI)	Landsat 7 ETM+ 30 m	(Near Infrared (NIR) - Red) / (Near Infrared (NIR) + Red) bands from Level L1T orthorectified scenes radiometrically corrected to TOA reflectance (Chander, Markham, & Helder, 2009)

Table A1: Flood Conditioning Factors Matrix: The following matrix describes major variables used in our predictive flood model. the authors chose well-established variables as inputs in order to be able to compare our performance with process-based models. Each dataset described below is natively available in the Earth Engine platform or calculated from datasets available in Earth Engine, with the exception of (3) Senegal Water Lines (NGA, 2015) and (13) Global HAND (Donchyts, Winsemius, Schellekens, Erickson, Gao, Savenije, & van de Giesen, 2016).

	Hit Rate*		Specificity**		False Alarm Rate		Critical Success Index		Overall Accuracy	
	MODIS	Landsat	MODIS	Landsat	MODIS	Landsat	MODIS	Landsat	MODIS	Landsat
CART	0.98	0.98	0.76	0.65	0.24	0.35	88.48	84.43	0.87	0.82
NB	0.52	0.15	0.81	0.82	0.20	0.19	60.35	22.72	0.67	0.48
RF	0.97	0.89	0.86	0.77	0.14	0.24	92.14	83.59	0.92	0.83
SVM	0.60	0.60	0.68	0.41	0.33	0.59	62.14	54.72	0.63	0.51

** also called Sensitivity* *** measure of ability to detect non-flooded pixels accurately*

Table A3: Performance metrics for each of the four algorithms for the September 2012 flood in the Saint-Louis region comparing the MODIS versus Landsat-based training imagery. Total modelled floodplains (risk areas) were calculated by summing pixel area of all pixels identified as flooded in any of the ten trials. High risk floodplains were identified by selecting all regions marked as flooded across all ten folds of cross-validation.

	Area Analyzed (km^2)	Total Risk Area (km^2)	Percent in Predicted Zone (%)	High Risk Area (km^2)	People at Risk
Matam 3180	5,135	1,051	20%	114	38,400
Matam 2315	5,135	802	16%	101	41,330
Fatick 3180	3,162	1,085	34%	528	17,038
Fatick 2315	3,162	781	25%	307	7,801
Kaolack 3180	1,906	204	11%	89	2,109
Kaolack 2315	1,906	221	12%	45	774
Saint-Louis 3180	3,990	1,399	35%	523	8,208
Saint-Louis 2315	3,990	1,267	32%	285	10,616
Dakar 3180	559	0	0%	0	0
Dakar 2315	559	0	0%	0	0
Ziguinchor 3180	7383	1,616	22%	349	31,754
Ziguinchor 2315	7383	1,494	20%	282	30,224
Sédhiou 3180	2,855.81	241	8%	39	4,426
Sédhiou 2315	2,855.81	122	4%	5	1,630

Table A4: Comparing modeled floodplains between the September 2007 (DFO #3180) and the August 2003 (DFO #2315) seasonal floods. While model accuracies (not included) do not vary drastically between events, the high risk flood areas and total exposed population predictions are clearly is sensitive to interannual variability.

7. References

ACAPS. (2016). *Briefing Note Senegal Floods*.

ACPF. (2011). *African Child Policy Forum* (Children with disabilities in Senegal: the hidden reality). Addis Ababa, Ethiopia: Africa.

Adger, W. N., Barnett, J., Brown, K., Marshall, N., & O'Brien, K. (2013). Cultural dimensions of climate change impacts and adaptation. *Nature Climate Change, 3*(2), 112–117. https://doi.org/10.1038/nclimate1666

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

Ahmed, A. K., Kodijat, A., Luneta, M., & Krishnamurthy, K. (2015). Typhoon Haiyan, an Extraordinary Event? A Commentary on the Complexities of Early Warning, Disaster Risk Management and Societal Responses to the Typhoon. *Asian Disaster Management News, 21*, 20–25.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). GloFAS–global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci, 17*(3), 1161–1175.

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., … Prat, O. P. (2015). PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies. *Bulletin of the American Meteorological Society, 96*(1), 69–83.

Bates, P. D. (2004). Remote sensing and flood inundation modelling. *Hydrological Processes, 18*(13), 2593–2597.

Bauer, J.-M., Attia, W., & Clough, A. (2016, April 5). When open is not enough: bringing food security data to affected communities. *Open Data Institute*. Retrieved from http://theodi.org/blog/when-open-is-not-enough-bringing-food-security-data-to-affected-communities?utm_content=buffera14af&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Beierle, T. C. (2002). The Quality of Stakeholder-Based Decisions. *Risk Analysis, 22*(4), 739–749. https://doi.org/10.1111/0272-4332.00065

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Bulletin, 24*(1), 43–69.

Bontemps, S., Defourny, P., Bogaert, E. V, Arino, O., Kalogirou, V., & Perez, J. R. (2011). GLOBCOVER 2009-Products description and validation report.

Boschetti, M., Nutini, F., Manfron, G., Brivio, P. A., & Nelson, A. (2014). Comparative Analysis of Normalised Difference Spectral Indices Derived from MODIS for Detecting Surface Water in Flooded Rice Cropping Systems. *PLoS ONE, 9*(2), e88741. https://doi.org/10.1371/journal.pone.0088741

BRACED. (2016). Dakar women grow herb business from floodwater. Retrieved October 7, 2016, from http://www.braced.org/news/i/?id=afd069c6-2d03-4787-a1bd-05b1ffd074ea

Brackenridge, G., Anderson, E., & Caquard, S. (2009). Global active archive of large floods 1985–2007, Dartmouth Flood Observatory (DFO), Hanover, USA. Retrieved from https://scholar.google.com/scholar?cluster=6329912732663934836&hl=en&as_sdt=8005&sciodt=0,7#0

Brenkert, A. L., & Malone, E. L. (2005). Modeling vulnerability and resilience to climate change: a case study of India and Indian states. *Climatic Change, 72*(1–2), 57–102.

Brunkard, J., Namulanda, G., & Ratard, R. (2008). Hurricane Katrina Deaths, Louisiana, 2005. Disaster Medicine and Public Health Preparedness, 2(4), 215–223. https://doi.org/10.1097/DMP.0b013e31818aaf55

Building Resilience and Adaptation to Climate Extremes and Disasters. (2015, November 2). Senegal Floods 2015. Retrieved from http://www.braced.org/reality-of-resilience/i/?id=3dc7ca27-bc7d-466f-a6af-4dae3bc12abd

Campolo, M., Andreussi, P., & Soldati, A. (1999). River flood forecasting with a neural network model. *Water Resources Research, 35*(4), 1191–1197.

CARE. (2011). Decision-making for climate resilient livelihoods and risk reduction: A Participatory Scenario Planning approach. Retrieved from http://www.care.org/sites/default/files/documents/CC-2011-ALP_PSP_Brief.pdf

CARRI. (2013). Success Stories: The Importance of Effective Community Engagement. Retrieved from http://www.resilientus.org/wp-content/uploads/2013/10/Oct-Success-Stories-Compilation-Community-Engagement.pdf

CCLONG. (2009). Participatory Assessment and Learning Tools. Retrieved October 6, 2016, from http://cclong.epa.gov.gh/index.php?option=com_content&view=category&layout=blog&id=39&Itemid=59

Central Intelligence Agency. (2016, October 19). Senegal. Retrieved October 24, 2016, from https://www.cia.gov/library/publications/the-world-factbook/geos/sg.html

Chambers, R. (1994). The origins and practice of participatory rural appraisal. World Development, 22(7), 953–969. https://doi.org/10.1016/0305-750X(94)90141-4

Chander, G., Markham, B. L., & Helder, D. L. (2009). Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote Sensing of Environment, 113*(5), 893–903.

Chatterjee, C. B., & Sheoran, G. (2007). *Vulnerable groups in India*. Centre for Enquiry into Health and Allied Themes Mumbai, India. Retrieved from http://fi.gb.pgstatic.net/attachments/33376_8c7cb59047bd4d6896adaa-2729fe8bd8.pdf

Chignell, S., Anderson, R., Evangelista, P., Laituri, M., & Merritt, D. (2015). Multi-Temporal Independent Component Analysis and Landsat 8 for Delineating Maximum Extent of the 2013 Colorado Front Range Flood. *Remote Sensing, 7*(8), 9822–9843. https://doi.org/10.3390/rs70809822

Coltin, B., McMichael, S., Smith, T., & Fong, T. (2016). Automatic boosted flood mapping from satellite data. *International Journal of Remote Sensing, 37*(5), 993–1015. https://doi.org/10.1080/01431161.2016.1145366

Conrad, C. C., & Hilchey, K. G. (2010). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental Monitoring and Assessment, 176*(1–4), 273–291. https://doi.org/10.1007/s10661-010-1582-5

Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. *Social Science Quarterly, 84*(2), 242–261.

de Sherbinin, A. (2014). Climate change hotspots mapping: what have we learned? *Climatic Change, 123*(1), 23–37.

Diagne, K. (2007). Governance and natural disasters: addressing flooding in Saint Louis, Senegal. *Environment and Urbanization, 19*(2), 552–562. https://doi.org/10.1177/0956247807082836

Donchyts, G., Schellekens, J., Winsemius, H., Eisemann, E., & van de Giesen, N. (2016). A 30 m Resolution Surface Water Mask Including Estimation of Positional and Thematic Differences Using Landsat 8, SRTM and OpenStreetMap: A Case Study in the Murray-Darling Basin, Australia. *Remote Sensing, 8*(5), 386. https://doi.org/10.3390/rs8050386

Donchyts, G., Winsemius, H., Schellekens, J., Erickson, T., Gao, H., Savenije, H., & Giesen, N. van de. (2016). Global 30m Height Above the Nearest Drainage.

Donchyts, G., Winsemius, H., Schellekens, J., Erickson, T., Gao, H., Savenije, H., & van de Giesen, N. (2016). Global 30m Height Above the Nearest Drainage. *HAND, 1000*, 0.

Dong, J., Xiao, X., Menarguez, M. A., Zhang, G., Qin, Y., Thau, D., … Moore, B. (2016). Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2016.02.016

Drame, E. R., & Kamphoff, K. (2014). Perceptions of Disability and Access to Inclusive Education in West Africa: A Comparative Case Study in Dakar, Senegal. *International Journal of Special Education, 29*(3), 69–81.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., … Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics, 45*(2), RG2004. https://doi.org/10.1029/2005RG000183

Federal Crowdsourcing and Citizen Science Toolkit. (n.d.). Retrieved September 5, 2016, from https://crowdsourcing-toolkit.sites.usa.gov/

Fekete, A. (2009). Validation of a social vulnerability index in context to river-floods in Germany. *Natural Hazards and Earth System Sciences, 9*(2), 393–403.

Feng, L., Hu, C., Chen, X., Cai, X., Tian, L., & Gan, W. (2012). Assessment of inundation changes of Poyang Lake using MODIS observations between 2000 and 2010. *Remote Sensing of Environment, 121*, 80–92. https://doi.org/10.1016/j.rse.2012.01.014

Feyisa, G. L., Meilby, H., Fensholt, R., & Proud, S. R. (2014). Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment, 140*, 23–35. https://doi.org/10.1016/j.rse.2013.08.029

Filiberto, D., Wethington, E., Pillemer, K., Wells, N., Wysocki, M., & Parise, J. T. (2009). Older People and Climate Change: Vulnerability and Health Effects. *Generations, 33*(4), 19–25.

Florinsky, I. V. (2012). *Digital terrain analysis in soil science and geology*. Academic Press.

Foresight. (2012). *Reducing Risks of Future Disasters: Priorities for Decision Makers* (Final Project Report). The Government Office for Science, London. Retrieved from https://www.gov.uk/government/publications/reducing-risk-of-future-disasters-priorities-for-decision-makers

Fothergill, A. (1996a). Gender, risk, and disaster. *International Journal of Mass Emergencies and Disasters, 14*(1), 33–56.

Fothergill, A. (1996b). Gender, Risk, and Disaster. *International Journal of Mass Emergencies and Disasters, 14*(1), 33–56.

Fothergill, A., & Peek, L. A. (2004). Poverty and disasters in the United States: A review of recent sociological findings. *Natural Hazards, 32*(1), 89–110.

GADM. (2015). Version 2.8. Retrieved from http://www.gadm.org/

Gao, B. C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment, 58*(3), 257–266.

Gencer, E. A. (2013). Natural Disasters, Urban Vulnerability, and Risk Management: A Theoretical Overview. In *The Interplay between Urban Development, Vulnerability, and Risk Management* (pp. 7–43). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29470-9_2

Geoville Group. (2009). Spatial Analysis of Natural Hazard and Climate Change Risks in Peri Urban Expansion Areas of Dakar, Senegal. Presented at the World Bank Urban Week, Washington D.C.

Goldblatt, R., You, W., Hanson, G., & Khandelwal, A. K. (2016). Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine. *Remote Sensing, 8*(8), 634.

Goldsmith, P. D., Gunjal, K., & Ndarishikanye, B. (2004). Rural-urban migration and agricultural productivity: the case of Senegal. *Agricultural Economics, 31*(1), 33–45. https://doi.org/10.1111/j.1574-0862.2004.tb00220.x

Gordon, E., Schirra, S., & Hollander, J. (2011). Immersive Planning: A Conceptual Model for Designing Public Participation with New Technologies. *Environment and Planning B: Planning and Design, 38*(3), 505–519. https://doi.org/10.1068/b37013

GTZ. (2007). Mainstreaming Participation. Retrieved from http://www.fsnnetwork.org/sites/default/files/en-svmp-instrumente-akteuersanalyse.pdf

Han, D., L, C., & N, Z. (2007). Flood forecasting using support vector machines.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., … others. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342(6160), 850–853.

Hellmuth, M. E., Mason, S. J., Vaughan, C., Van Aalst, M. K., & Choularton, R. (2011). *A better climate for disaster risk management*. Palisades: International Research Institute for Climate and Society.

Holmes, R., Sadana, N., & Rath, S. (2010). Gendered Risks, Poverty and Vulnerability in India: Case Study of the Indian Mahatma Gandhi National Rural Employment Guarantee Act (Madhya Pradesh). *The Overseas Development Institute Research Report, October.* Retrieved from https://www.odi.org/resources/docs/6254.pdf

Hong, W.-C. (2008). Rainfall forecasting by technological machine learning models. *Applied Mathematics and Computation, 200*(1), 41–57.

Hossain, F., Katiyar, N., Hong, Y., & Wolf, A. (2007). The emerging role of satellite rainfall data in improving the hydro-political situation of flood monitoring in the under-developed regions of the world. *Natural Hazards, 43*(2), 199–210. https://doi.org/10.1007/s11069-006-9094-x

ideo.org. (2015). Field Guide to Human-Centered Design. Retrieved from http://d1r3w4d5z5a88i.cloudfront.net/assets/guide/Field%20Guide%20to%20Human-Centered%20Design_IDEOorg_English-ee47a1ed4b91f3252115b83152828d7e.pdf

IFAD. (2009). Good Practice in Participatory Mapping. nternational Fund for Agricultural Development (IFAD). Retrieved from https://www.ifad.org/documents/10180/d1383979-4976-4c8e-ba5d-53419e37cbcc

IIED. (2016). Participatory Learning and Action (PLA). Retrieved September 14, 2016, from http://www.iied.org/participatory-learning-action

International Finance Corporation. (2007). Stakeholder Engagement: A Good Practice Handbook for Companies Doing Business in Emerging Markets. World Bank Group. Retrieved from http://www.ifc.org/wps/wcm/connect/938f1a0048855805beacfe6a6515bb18/Ifc_StakeholderEngagement.pdf?MOD=AJPERES

International Monetary Fund. (2010). *Senegal: Poverty Reduction Strategy Paper Annual Progress Report.* International Monetary Fund.

IPCC. (2014). *Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects*. Cambridge University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=aJ-TBQAAQBAJ&oi=fnd&pg=PA1142&dq=Summary+for+Policymakers.+In:+Climate+Change+2014:+Impacts,+Adaptation,+and+Vulnerability.+Part+A:+Global+and+Sectoral+Aspects.&ots=v0RzOM54HF&sig=MAp9dTJL-dnthKOBuQuZDbOrITk

Islam, A. S., Bala, S. K., & Haque, A. (2009). Flood inundation map of Bangladesh using Modis surface reflectance data. In *International conference on water and flood management (ICWFM) Dhaka, Bangladesh* (Vol. 2, pp. 739–748). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.469.6366&rep=rep1&type=pdf

Janneh, A. (2012). Statement submitted by Abdoulie Janneh, UN Under-Secretary-General and Executive Secretary of ECA. Presented at the Forty-fifth Session of the Commission on Population and Development, New York: United Nations Economic Commission for Africa. Retrieved from http://www.un.org/en/development/desa/population/pdf/commission/2012/country/Agenda%20item%204/UN%20system%20statements/ECA_Item4.pdf

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science, 353*(6301), 790–794.

Ji, L., Zhang, L., & Wylie, B. (2009). Analysis of Dynamic Thresholds for the Normalized Difference Water Index. *Photogrammetric Engineering & Remote Sensing*, (11), 1307–1317.

Johansen, K., Phinn, S., & Taylor, M. (2015). Mapping woody vegetation clearing in Queensland, Australia from Landsat imagery using the Google Earth Engine. *Remote Sensing Applications: Society and Environment, 1*, 36–49. https://doi.org/10.1016/j.rsase.2015.06.002

Jonkman, S. N., & Kelman, I. (2005). An analysis of the causes and circumstances of flood disaster deaths. *Disasters, 29*(1), 75–97.

Kahn et al. (2003). Health Consequences of Migration: Evidence from South Africa's Rural Northeast (Agincourt). Presented at the African Migration in Comparative Perspective, Johannesburg, Sounth Africa.

Kane, F., Alary, M., Ndoye, I., Coll, A. M., M'boup, S., Guèye, A., … Joly, J. R. (1993). Temporary expatriation is related to HIV-1 infection in rural Senegal. *AIDS (London, England), 7*(9), 1261–1265.

Keating, A., Campbell, K., Mechler, R., Michel-Kerjan, E., Mochizuki, J., Kunreuther, H., … others. (2014). Operationalizing resilience against natural disaster risk: Opportunities, barriers and a way forward. *Zurich Flood Resilience Alliance*. Retrieved from https://riskcenter.wharton.upenn.edu/wp-content/uploads/2014/07/zurichfloodresiliencealliance_ResilienceWhitePaper_2014.pdf

Klinenberg, E. (2003). *Heat Wave: A Social Autopsy of Disaster in Chicago*. University of Chicago Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=r22xueipNegC&oi=fnd&pg=PR9&ots=NmQWOB-our&sig=kmkaKhHk8TUKFCBqTLIJwAl98C8

Kruger, L., & Shannon, M. A. (2000). Getting to Know Ourselves and Our Places Through Participation in Civic Social Assessment (PDF Download Available). *Society and Natural Resources, 13*(5). https://doi.org/10.1080/089419200403866

Lauriault, T. P., & Mooney, P. (2014). *Crowdsourcing: A Geographic Approach to Public Engagement* (SSRN Scholarly Paper No. ID 2518233). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2518233

Lehner, B., Verdin, K., & Jarvis, A. (2006). HydroSHEDS technical documentation, version 1.0. *World Wildlife Fund US, Washington, DC*, 1–27.

Leye, M. M. M., Diongue, M., Faye, A., Coumé, M., Faye, A., Tall, A. B., … Tal-Dia, A. (2013). [Analysis of free health care for the elderly in the context of the "Plan Sésame" in Senegal]. *Santé publique* (Vandoeuvre-les-Nancy, France), 25(1), 101–106.

Li, W., Du, Z., Ling, F., Zhou, D., Wang, H., Gui, Y., … Zhang, X. (2013). A Comparison of Land Surface Water Mapping Using the Normalized Difference Water Index from TM, ETM+ and ALI. *Remote Sensing, 5*(11), 5530–5549. https://doi.org/10.3390/rs5115530

Lin, J.-Y., Cheng, C.-T., & Chau, K.-W. (2006). Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal, 51*(4), 599–612.

Liong, S.-Y., & Sivapragasam, C. (2002). FLOOD STAGE FORECASTING WITH SUPPORT VECTOR MACHINES. *Journal of the American Water Resources Association, 38*(1), 173–186.

Maharaj, P. (2012). *Aging and Health in Africa*. Springer Science & Business Media.

Maheu, A. (2012). Urbanization and Flood Vulnerability in a Peri-Urban Neighbourhood of Dakar, Senegal: How can Participatory GIS Contribute to Flood Management? In W. L. Filho (Ed.), *Climate Change and the Sustainable Use of Water Resources* (pp. 185–207). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22266-5_12

Mannel, S., Price, M., & Hua, D. (2011). Impact of reference datasets and autocorrelation on classification accuracy. *International Journal of Remote Sensing, 32*(19), 5321–5330.

Martinis, S., Twele, A., Strobl, C., Kersten, J., & Stein, E. (2013). A Multi-Scale Flood Monitoring System Based on Fully Automatic MODIS and TerraSAR-X Processing Chains. *Remote Sensing, 5*(11), 5598–5619. https://doi.org/10.3390/rs5115598

Martinis, S., Twele, A., & Voigt, S. (2009). Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution TerraSAR-X data. *Natural Hazards and Earth System Sciences, 9*(2), 303–314.

Mason, D. C., Giustarini, L., Garcia-Pintado, J., & Cloke, H. L. (2014). Detection of flooded urban areas in high resolution Synthetic Aperture Radar images using double scattering. *International Journal of Applied Earth Observation and Geoinformation, 28*, 150–159. https://doi.org/10.1016/j.jag.2013.12.002

Mbaye, E. M., Ridde, V., & Kâ, O. (2012). ["Good intentions are not enough": analysis of a health policy for the elderly in Senegal]. *Sante publique (Vandoeuvre-les-Nancy, France), 25*(1), 107–112.

Mbow, C., Diop, A., Diaw, A. T., & Niang, C. I. (2008). Urban sprawl development and flooding at Yeumbeul suburb (Dakar-Senegal). *African Journal of Environmental Science and Technology, 2*(4), 75–88.

Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of who and What Really Counts. *Academy of Management Review, 22*(4), 853–886. https://doi.org/10.5465/AMR.1997.9711022105

Mitchell, T., brahim, H., K., Polack, E., Hall, N., Hawrylyshyn, K., Hedger, M., … Sajjad, M., S. (2010). Climate Smart Disaster Risk Management, Strengthening Climate Resilience. Strengthening Climate Resilience,.

Naghibi, S. A., & Pourghasemi, H. R. (2015). A Comparative Assessment Between Three Machine Learning Models and Their Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping. *Water Resources Management, 29*(14), 5217–5236.

National Academy of Sciences. (2012). *Disaster Resilience: A National Imperative*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/13457

Nepali Red Cross. (2009, April). Community-Based Disaster Risk Reduction Good Practice. Retrieved from http://www.preventionweb.net/files/10479_10479CommunityBasedDRRGoodPracticeR.pdf

Neumayer, E., & Plümper, T. (2007a). The Gendered Nature of Natural Disasters: The Impact of Catastrophic Events on the Gender Gap in Life Expectancy, 1981–2002. *Annals of the Association of American Geographers, 97*(3), 551–566. https://doi.org/10.1111/j.1467-8306.2007.00563.x

Neumayer, E., & Plümper, T. (2007b). The Gendered Nature of Natural Disasters: The Impact of Catastrophic Events on the Gender Gap in Life Expectancy, 1981–2002. *Annals of the Association of American Geographers, 97*(3), 551–566. https://doi.org/10.1111/j.1467-8306.2007.00563.x

Newport, J. K., & Jawahar, G. G. (2003). Community participation and public awareness in disaster mitigation. *Disaster Prevention and Management: An International Journal, 12*(1), 33–36.

NGA, N. G.-I. A. (2015). Senegal Water Lines. Retrieved January 3, 2016, from https://ngamaps.geointapps.org/arcgis/rest/services/Senegal/Senegal_Water_Lines/MapServer

Ngo, E. B. (2001). When disasters and age collide: Reviewing vulnerability of the elderly. *Natural Hazards Review, 2*(2), 80–89.

NSSL Projects: mPING. (n.d.). [text]. Retrieved October 6, 2016, from http://mping.nssl.noaa.gov/

ODI. (n.d.). Planning Tools: Stakeholder Analysis. Retrieved October 3, 2016, from https://www.odi.org/publications/5257-stakeholder-analysis

O'Hare, G. (2001). Hurricane 07B in the Godivari Delta, Andhra Pradesh, India: Vulnerability, mitigation and the spatial impact. *Geographical Journal, 167*(1), 23–38.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica, 11*(285–296), 23–27.

Paavola, J., & Adger, W. N. (2006). Fair adaptation to climate change. *Ecological Economics, 56*(4), 594–609. https://doi.org/10.1016/j.ecolecon.2005.03.015

Pahl-Wostl, C., Becker, G., Knieper, C., & Sendzimir, J. (2013). How Multilevel Societal Learning Processes Facilitate Transformative Change: A Comparative Case Study Analysis on Flood Management. *Ecology and Society, 18*(4). https://doi.org/10.5751/ES-05779-180458

Pandey, B., & Okazaki, K. (n.d.). Community Based Disaster Management: Empowering Communities to Cope with Disaster Risks. United Nations Centre for Regional Development, Japan. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.467.1932&rep=rep1&type=pdf

Parmar, D., Williams, G., Dkhimi, F., Ndiaye, A., Asante, F. A., Arhinful, D. K., & Mladovsky, P. (2014). Enrolment of older people in social health protection programs in West Africa – Does social exclusion play a part? *Social Science & Medicine, 119*, 36–44. https://doi.org/10.1016/j.socscimed.2014.08.011

Peek, L. (2008). Children and disasters: Understanding vulnerability, developing capacities, and promoting resilience—An introduction. *Children Youth and Environments, 18*(1), 1–29.

Pelling, M., & Wisner, B. (2012). *Disaster risk reduction: Cases from urban Africa*. Routledge.

Plessis, I. G., & Reenen, T. H. V. (2011). *Aspects of Disalibility Law in Africa*. PULP.

Pradhan, B. (2010, January 27). Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. J*ournal of Spatial Hydrology.*

Pretty, J. N., Guijt, I., Thompson, J., & Scoones, I. (1995). Participatory Learning and Action: A trainer's guide. IIED. Retrieved from http://pubs.iied.org/6021IIED/

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, 414–415, 284–293.

Ray-Bennett, N. S. (2009). The influence of caste, class and gender in surviving multiple disasters: A case study from Orissa, India. *Environmental Hazards, 8*(1), 5–22. https://doi.org/10.3763/ehaz.2009.0001

Reed, M. S. (2008). Stakeholder participation for environmental management: A literature review. Biological Conservation, 141(10), 2417–2431. https://doi.org/10.1016/j.biocon.2008.07.014

Reid, P., & Vogel, C. (2006). Living and responding to multiple stressors in South Africa—Glimpses from KwaZulu-Natal. *Global Environmental Change, 16*(2), 195–206. https://doi.org/10.1016/j.gloenvcha.2006.01.003

République du Sénégal. (1996). Loi n° 96-06 du 22 mars 1996 portant Code des Collectivités locales. Retrieved from http://www.gouv.sn/Code-des-Collectivites-locales.html

République du Sénégal. (2013). Loi n° 2013-10 du 28 décembre 2013 portant Code général des Collectivités locales. Retrieved from http://www.gouv.sn/Code-general-des-Collectivites.html

Revelle, W. (2016). An overview of the psych package. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.7429&rep=rep1&type=pdf

Rufat, S., Tate, E., Burton, C. G., & Maroof, A. S. (2015). Social vulnerability to floods: Review of case studies and implications for measurement. *International Journal of Disaster Risk Reduction, 14*, 470–486. https://doi.org/10.1016/j.ijdrr.2015.09.013

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences, 1*(1), 83–98.

Sampson, C. C., Smith, A. M., Bates, P. D., Neal, J. C., Alfieri, L., & Freer, J. E. (2015a). A high-resolution global flood hazard model. *Water Resources Research, 51*(9), 7358–7381. https://doi.org/10.1002/2015WR016954

Sampson, C. C., Smith, A. M., Bates, P. D., Neal, J. C., Alfieri, L., & Freer, J. E. (2015b). A high-resolution global flood hazard model. *Water Resources Research, 51*(9), 7358–7381. https://doi.org/10.1002/2015WR016954

Sané, O. D., Gaye, A. T., Diakhaté, M., & Aziadekey, M. (2015). Social Vulnerability Assessment to Flood in Medina Gounass Dakar. *Journal of Geographic Information System, 7*(4), 415–429. https://doi.org/10.4236/jgis.2015.74033

Simon, D. (2010). The Challenges of Global Environmental Change for Urban Africa. *Urban Forum, 21*(3), 235–248. https://doi.org/10.1007/s12132-010-9093-6

Singh, B. K. (2014). Flood Hazard Mapping with Participatory GIS The Case of Gorakhpur. *Environment and Urbanization Asia, 5*(1), 161–173. https://doi.org/10.1177/0975425314521546

Solomatine, D. P., & Xue, Y. (2004). M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *Journal of Hydrologic Engineering, 9*(6), 491–501.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One, 10*(2), e0107042.

Tate, E. (2012). Social vulnerability indices: a comparative assessment using uncertainty and sensitivity analysis. *Natural Hazards, 63*(2), 325–347. https://doi.org/10.1007/s11069-012-0152-2

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2013). Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology, 504,* 69–79.

Tekman, E. D., Hassapi, A., Chrysostomou, G., Konnaris, Y., & Neophytou, H. (2012). Participatory Development Training Manual. UNDP. Retrieved from http://archive.undp-act.org/data/articles/FT%20TRAINING%20MANUAL%20WEB.pdf

Tellman, B., Alaniz, R., Rivera, A., & Contreras, D. (2014). Violence as an obstacle to livelihood resilience in the context of climate change. *United Nations University-Institute for Environment and Human Security, 3*. Retrieved from https://works.bepress.com/ralaniz/10/

The World Bank. (2014, August). Senegal: Urban Floods, Recovery and Reconstruction since 2009. Retrieved from https://www.gfdrr.org/sites/gfdrr/files/Senegal_English_August%202014.pdf

Tompkins, E. L., Few, R., & Brown, K. (2008). Scenario-based stakeholder engagement: Incorporating stakeholders preferences into coastal planning for climate change. *Journal of Environmental Management, 88*(4), 1580–1592. https://doi.org/10.1016/j.jenvman.2007.07.025

Trust.org. (2016, August 16). Senegal Floods Expose Need for Community Warning, Preparation. Retrieved October 6, 2016, from http://floodlist.com/africa/senegal-floods-expose-need-community-warning-preparation

Tschakert, P. (2007). Views from the vulnerable: Understanding climatic and other stressors in the Sahel. *Global Environmental Change, 17*(3–4), 381–396. https://doi.org/10.1016/j.gloenvcha.2006.11.008

Tschakert, P., Sagoe, R., Ofori-Darko, G., & Codjoe, S. N. (2010). Floods in the Sahel: an analysis of anomalies, memory, and anticipatory learning. *Climatic Change, 103*(3–4), 471–502. https://doi.org/10.1007/s10584-009-9776-y

UNESCO. (2012). UNESCO Global Partnership for Girls' and Women's Education - One Year On. Retrieved from http://www.unesco.org/eri/cp/factsheets_ed/SN_EDFactSheet.pdf

UNISDR. (2015). The Human Cost of Weather Related Disasters 1995-2015. United Nations Office for Disaster Risk Reduction.

UNISDR. (n.d.). What is Disaster Risk Reduction? Retrieved November 4, 2016, from https://www.unisdr.org/who-we-are/what-is-drr

UNOCHA. (2013, October). Senegal: Breaking the cycle of annual floods. Retrieved from http://www.unocha.org/top-stories/all-stories/senegal-breaking-cycle-annual-floods

Urban Habitat. (2014). Senegal. Retrieved from https://www.wm-urban-habitat.org/eng/senegal/

USGS. (n.d.). Tweet Earthquake Dispatch. Retrieved November 4, 2016, from http://earthquake.usgs.gov/earthquakes/ted/

Vanderbeck, R., & Worth, N. (2015). *Intergenerational Space*. Routledge.

Vedeld, T., Coly, A., Ndour, N. M., & Hellevik, S. (2015). Climate adaptation at what scale? Multi-level governance, resilience, and coproduction in Saint Louis, Senegal. *Natural Hazards, 82*(S2), 173–199. https://doi.org/10.1007/s11069-015-1875-7

Wang, D., Ding, W., Yu, K., Wu, X., Chen, P., Small, D. L., & Islam, S. (2013). Towards long-lead forecasting of extreme flood events. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13* (p. 1285). New York, New York, USA: ACM Press.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology, 527,* 1130–1141.

Werg, J., Grothmann, T., & Schmidt, P. (2013a). Assessing social capacity and vulnerability of private households to natural hazards – integrating psychological and governance factors. *Nat. Hazards Earth Syst. Sci., 13*(6), 1613–1628. https://doi.org/10.5194/nhess-13-1613-2013

Werg, J., Grothmann, T., & Schmidt, P. (2013b). Assessing social capacity and vulnerability of private households to natural hazards–integrating psychological and governance factors. *Natural Hazards and Earth System Sciences, 13*(6), 1613–1628.

Werner, M. G. F., Hunter, N. M., & Bates, P. D. (2005). Identifiability of distributed floodplain roughness values in flood extent estimation. *Journal of Hydrology, 314*(1–4), 139–157. https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2005.03.012

World Bank. (2010). *World Development Report 2010: Development and Climate Change*. World Bank Publications.

World Bank,. (2012, September 12). Senegal Makes Disaster Preparedness a Priority. *The World Bank.* Retrieved from http://www.worldbank.org/en/news/feature/2012/09/12/senegal-makes-disaster-preparedness-a-priority

World Bank. (n.d.). Stakeholder Engagement and Information Dissemination. Retrieved October 4, 2016, from http://siteresources.worldbank.org/INTRANETENVIRONMENT/Resources/244351-1279901011064/StakeholderEngagement-andGrievanceMechanisms_111031.pdf

World Resources Institute. (2015, March). World’s 15 Countries with the Most Peop...ver Floods World Resources Institute.pdf.

Wu, H., Adler, R. F., Hong, Y., Tian, Y., Policelli, F., Wu, H., … Policelli, F. (2012). Evaluation of Global Flood Detection Using Satellite-Based Rainfall and a Hydrologic Model. Http://Dx.doi.org/10.1175/JHM-D-11-087.1. https://doi.org/10.1175/JHM-D-11-087.1

Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J. Y., Salas, W., & Moore, B. (2006). Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. *Remote Sensing of Environment, 100*(1), 95–113. https://doi.org/10.1016/j.rse.2005.10.004

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing, 27*(14), 3025–3033. https://doi.org/10.1080/01431160600589179

Yang, K., Li, M., Liu, Y., Cheng, L., Duan, Y., & Zhou, M. (2014). River Delineation from Remotely Sensed Imagery Using a Multi-Scale Classification Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(12), 4726–4737. https://doi.org/10.1109/JSTARS.2014.2309707

Zhu, Z., & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment, 118*, 83–94.

Papiers de Recherche de l'AFD

Les *Papiers de Recherche de l'AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ilss'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l'AFD n'en exclut aucune autre.

L'Agence Française de Développement (AFD), institution financière publique qui met en oeuvre la politique définie par le gouvernement français, agit pour combattre la pauvreté et favoriser le développement durable. Présente sur quatre continents à travers un réseau de 72 bureaux, l'AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète. En 2014, l'AFD a consacré 8,1 milliards d'euros au financement de projets dans les pays en développement et en faveur des Outre-mer.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s).

Les *Papiers de Recherche* sont téléchargeables sur : http://librairie.afd.fr/

AFD Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

Agence Française de Développement (AFD), a public financial institution that implements the policy

defined by the French Government, works to combat poverty and promote sustainable development. AFD operates on four continents via a network of 72 offices and finances and supports projects that improve living conditions for populations, boost economic growth and protect the planet. In 2014, AFD earmarked EUR 8.1bn to finance projects in developing countries and for overseas France.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

AFD Research Papers can be downloaded from: http://librairie.afd.fr/en/

AFD, 5 rue Roland Barthes
75598 Paris Cedex 12, France
ResearchPapers@afd.fr
ISSN 2492 - 2846

BIG DATA

TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018

BIG DATA
TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES
2018

DATA-POP ALLIANCE

CHARACTERIZING AND ANALYZING URBAN DYNAMICS IN BOGOTA

Marco De Nadai - University of Trento, denadai@fbk.eu
Emmanuel Letouzé - Data-Pop Alliance, eletouze@datapopalliance.org
Marta C. González - MIT, martag@mit.edu
Bruno Lepri - FBK, lepri@fbk.eu

DATA-POP ALLIANCE FONDAZIONE BRUNO KESSLER mit media lab

Funded by:
AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

CHARACTERIZING AND ANALYZING URBAN DYNAMICS IN BOGOTA

Marco De Nadai,
University of Trento,
denadai@fbk.eu

Emmanuel Letouzé,
Data-Pop Alliance,
eletouze@datapopalliance.org

Marta C. González, MIT,
martag@mit.edu

Bruno Lepri, FBK, lepri@fbk.eu

Please cite this paper as: DE NADAI, M., E. LETOUZÉ, M.C. GONZÁLEZ and B. LEPRI (2017), “Characterizing and analyzing urban dynamics in Bogota”, *AFD Research Paper Series*, No. 2018-70, June.

Abstract

Containing crime without affecting the livability of the urban environment is a major challenge in our society. Traditionally, researchers relate crime to socio-economic disorganization and people's routine activity, as it influences effective control and suitable targets. An important open question is what the role the urban fabric plays. Although empirical research has shown that the physical urban environment is an essential factor for urban vitality and health, we lack evidence of any clear relationship between the structural characteristics (e.g. roads and land use mix) of neighborhoods and crime. Here, by using open data and mobile phone records, we explore this link with a spatial regression model that analyzes the environmental and the social conditions to which each part of the city is exposed. We found that physical characteristics of the city connected to higher urban diversity better explain the emergence of crime than traditional socio-economic conditions and, together, physical characteristics and socioeconomic conditions improve the performance of previous approaches. This result suggests that urban diversity and natural surveillance theories play an important role in the proliferation of crime, and the knowledge of this role can be exploited in urban planning to reduce crime.

Key words: Welfare programs, Well-Being, Poverty, Family Structure and Planning, Economics of Minorities, Races, Indigenous Peoples and Immigrants, Non-labor Discrimination, Demand and Supply of Labor, Urban and Transportation Analysis, Housing, Land Use in Urban Economies

JEL Classification: I38, J12, J13, J15, J18, J21, J23, J61, J64, J68, J71, N36, O18, R14, R21, R23, R28, R41, R52

Original version: English

Accepted: June 2017

TABLE OF CONTENTS

Characterizing and analyzing urban dynamics in Bogota

Marco De Nadai[1,2], Andrés Clavijo[4], Marta C. González[3], Emmanuel Letouzé[4] and Bruno Lepri[2]

Containing crime without affecting the livability of the urban environment is a major challenge in our society. Traditionally, researchers relate crime to socioeconomic disorganization and people's routine activity, as it influences effective control and suitable targets. An important open question is what the role the urban fabric plays. Although empirical research has shown that the physical urban environment is an essential factor for urban vitality and health, we lack evidence of any clear relationship between the structural characteristics (e.g. roads and land use mix) of neighborhoods and crime. Here, by using open data and mobile phone records, we explore this link with a spatial regression model that analyzes the environmental and the social conditions to which each part of the city is exposed. We found that physical characteristics of the city connected to higher urban diversity better explain the emergence of crime than traditional socioeconomic conditions and, together, physical characteristics and socioeconomic conditions improve the performance of previous approaches. This result suggests that urban diversity and natural surveillance theories play an important role in the proliferation of crime, and the knowledge of this role can be exploited in urban planning to reduce crime.

[1] University of Trento.

[2] FBK.

[3] MIT.

[4] Data–Pop Alliance

1. Introduction

The rapid growth of cities and the increase of population mobility have challenged our ability to understand crime. The primary focus of criminology research has been on people that commit crimes, and the reason they are involved. For crime to happen three conditions are supposed to be present and interact: the presence of a motivated offender who is willing to commit a crime, a suitable target, and the absence of guardians that would otherwise prevent the crime from taking place [6]. In this equation crime offenders, victims and guardians are all affected by socioeconomic conditions, the social disorganization (e.g. unemployment) of community [9] and the place where they intersect. Thus, place matters.

Environmental criminology suggests that place not only is logically required, but also influences the likelihood of becoming a crime hotspot through its physical characteristics. Accordingly, place is one of the five necessary and sufficient components that constitute a criminal incident, namely place, time, law, offender and victim [3]. Thus, environmental criminologists are interested in land use, street design, traffic patterns and daily activities of people. However, scholars have virtually ignored other theories (e.g. social disorganization), and bounded their discussion to macro-areas of few cities.

Urban planners and sociologists argue that cities are not a mere artificial construction that group people. A city is a vital process of the people who compose it; and its neighborhoods are the elementary form of cohesion in urban life [15]. One of the seminal books in city planning is Jane Jacobs' The Death and Life of Great American Cities [13]. In this book she introduced the concept of eyeson-the-street, which suggests that safety can be maintained by citizens through urban surveillance. For this to work, some physical qualities need to be present in the neighborhoods (i.e. a mix of residential, commercial and recreational land uses) to guarantee the diversity and continuous presence of people throughout the day. It is thus clear the tight coupling of environmental criminology and urban planning theories.

Traditional approaches on describing crime have failed to provide a clear and broad description of the desirable characteristics the different parts of the city should possess to keep crime events low.

In the present study we seek to fill this gap, formalizing the hypothesis that physical characteristics of the city not only are related to better life conditions [17] and vitality [7], but also greatly influence crime. Thus, we create two types of models. One is focused on *describing* how physical characteristics influence crime in each part of the city. The other *predicts* crime events in a city from the structural features, and answers to the question "can we predict crime from the physical characteristics of the city?". Thanks to several new sources of data and a Negative Binomial model, we model crime through physical and socioeconomic characteristics, but also spatial and routine-activity information.

We find that structural characteristics of the city, namely Jacobs' diversity conditions, are better predictors than socioeconomic status, and that these results are robust across different spatial aggregations. Also, we find that the number of inhabitants' routine movements between the neighborhoods is closely related to crime, with highly-connected points of the city experiencing a higher number of crimes, as suggested by the routine-activity theory [6]. Finally, we observe that the combination of structural, routine and socioeconomic information provides better estimates of crime than each on its own. Together, these observational results suggest that the city structure has a strong connection with crime, and that improving its qualities can discourage criminality.

Thus, our main contributions are: i) we focus on place in a new fashion and show how physical characteristics greatly influence crime events; ii) we built a now-casting model, which is portable from one city to many, able to predict crime counts in a city; iii) we employ new sources of data and combine multiple criminology theories in Bogota.

This paper is organized as follows: in Section 2 we review the literature in this field. In Section 4 we outline the proposed approach and the evaluation process. Finally, we show our results in Section 5, before discussing the implications and limitations in Section 6.

2. Related work

One of the most prolific place that established the hallmark of environmental criminology is Chicago. In the University of this city, sociologists and criminologists started to consider neighborhoods as unit of analysis, both from the social and political (or administrative) perspective.

Ernest Burgess developed a concentric-zone model, based price changes in housing values, to study crime patterns in Chicago [4]. He observed that the distribution of social problems and crime vary in respect to the distances to the center. On the basis of this model, Clifford Shaw extensively researched how young people, juvenile delinquency and adult offenders were distributed in space [8]. He introduced the *spot maps*, *delinquency rate maps*, *radial maps and zone maps* that established a landmark on crime mapping.

In the recent years, many empirical and predictive studies flocked thanks to new methods mainly coming from computer science, and the increasing availability of new sources of data. Graif and Sampson [11] examined the connection of immigration and socioeconomic diversity to homicide. Thanks to a spatial model they found that immigrant concentration is either unrelated or inversely related to homicides, whereas language diversity is negatively correlated to homicides. Verifying how diversity of people influences crime in neighborhoods was the goal of the study of Traunmueller *et al*. [19]. To observe *people dynamics* they used mobile phone records broken by age and gender. They observed that age-diversity and presence of non-residents are linked to lower criminality. Bogomolov *et al*. [2] used mobile phone data in a similar fashion, and predicted crime *hotspots* thanks to a Random Forest regression.

By only using the ambient population characteristics extracted from mobile phone records they were able to predict with almost 70% of accuracy whether an area would have high or low crime levels in time. The assumption of observing ambient population ignoring the movements of people in the city was tackled by Graif *et al.* [9]. They argued that people are continuously exposed to different neighborhoods, and they proposed a *network of neighborhoods* to describe this. This approach is considered important for understanding how changes of activity spaces can influence crime.

Strikingly, very few scholars considered multiple sources of data and theories in their discussions. Moreover, the importance of place is limited to spatial-autocorrelation or the topology of street patterns (e.g. [16]). Contrairily to this, Wang *et al*. [20] examined crime in Chicago by leveraging census counts as well as new sources of data available on the web, such as crowd-generated Points Of Interest (POI) and taxi flows. POIs were expected to be linked to higher crime incident as they represent suitable targets; taxi flows were considered as proxy for trips made by humans. The incremental now-casting model built by Wang *et al*. showedthat crime rate can be estimated with a relative error of 30% by using demographic information, but it could be also reduced by 5% with the POI data. Finally, he showed that the complexity of the crime in Chicago could be further explained by the taxi flow network, which further improved the results by 5%.

3. Description of the data

3.1 Spatial data

Bogotá, the capital city of Colombia, is divided into 20 localities, each of which contains between 1 to 12 zonal planning units, known as UPZ in Spanish. In total, there are 113 UPZ. Each UPZ, in turn, is divided into neighborhoods, which are themselves composed of a set of blocks.

In addition to the UPZ scheme, Bogotá also has spatial divisions defined by the national census. The designations used by the census are (from largest to smallest): urban sector, urban section, and block. A sector is a cartographic census division, roughly equivalent to a neighborhood (especially for large cities), and comprising between 1 and 9 sections. Each section is composed of approximately 20 contiguous blocks, all falling within the same sector. Finally, a block is a lot of land, built or unbuilt, bounded by public paths, roads, crosswalks, etc. Blocks may also be bounded by a natural feature such as a river, stream or channel, as long as it is permanent and easy to locate in the field.

Our main source of spatial data for this analysis is the Capital District's Spatial Data Infrastructure dataset, known as Infraestructura de Datos Espaciales del Distrito Capital (IDECA) in Spanish. Its function is to facilitate the access to geographic information about Bogot´a and support its social, economic, and environmental development.

The IDECA dataset used for this study is a compilation of 10 spatial datasets which cover the following topics: buildings, lots, blocks, localities, land use, points of interest, strata, transport nodes, cycling trails, and road network.

3.2 Socioeconomic and crime data

The main sources for socioeconomic and crime data are the National Statistics Office of Colombia, known as Departamento Administrativo Nacional de Estadística (DANE) in Spanish, and the National Police of Colombia. We use the population census and multipurpose survey from DANE and crime data from the National Police.

3.2.1 Census data

We used census data from the last census held in Colombia in 2005, and the population projections made by DANE for 2015. The Secretary of Planning of Bogotá, using DANE's projections, issued projections of the 2015 population of Bogotá by UPZ. From the census, we have access to socioeconomic data at the block level. In Bogotá, this census was conducted in 1,931,372 households, distributed across 37,473 blocks, and includes 6,778,691 people. The projections estimate a rise in Bogotá's population to 7,878,783 people in 2015.

3.2.2 Multipurpose Survey

The Multipurpose Survey, known as Encuesta Multipropósito (EM) in Spanish, was performed in 2014 by DANE, and financed by the Secretary of Planning of Bogotá. Its objective was to obtain statistical information on social, economic, and environmental aspects of urban households and residents of Bogotá. DANE uses the EM to derive income data, multidimensional poverty indixes, and subjective poverty indexes, among others. The information from this survey is statistically representative at a locality level (which is less granular than the UPZ).

3.2.3 SISBEN Survey

The Identification System of Potential Welfare Recipients, known as Sistema de Identificación de Potenciales

Beneficiarios de Programas Sociales (SISBEN) in Spanish, was created with the purpose of reducing the cost of targeting social benefits receivers and keeping track of these in the whole country.

The SISBEN survey is the tool through which individuals are categorized as recipients of social aid. It contains a set of variables related to durable goods consumption, human capital endowment and current income. The 2012 survey contains 3,545,789 observations distributed through different Colombian departments; and 123 variables, most of them related to socioeconomic conditions.

It will be possible to localize the amount of beneficiaries for each block in Bogotá using the 2012 survey, and thus generate a more granular characterization of socioeconomic status in different zones of Bogotá. The Non-Disclosure Agreement that allows using this data will be signed soon.

3.2.4 OD network

The Origin-Destination network is extracted from mobile phone data, automatically collected for billing purposes. We use data from the largest telecommunications operator in the area. For each area (UPZ) a and b we aggregate the number of people that move from a to b in a typical day. This results in a matrix were for each UPZ we have the information about the routine activity of people, which is related to both the presence of people and the risk of victimization.

3.2.5 Crime data

The criminal cases dataset includes geo-located and timestamped records of reported crime in Bogotá. It consists of 27,863 criminal cases for homicide and theft (burglaries of commercial property, burglaries of houses, and robberies) for 2014. Specifically, the dataset includes the category and subcategory of the crime, the longitude, latitude, and address of where the crime was reported to have occurred, and the responsible police department.

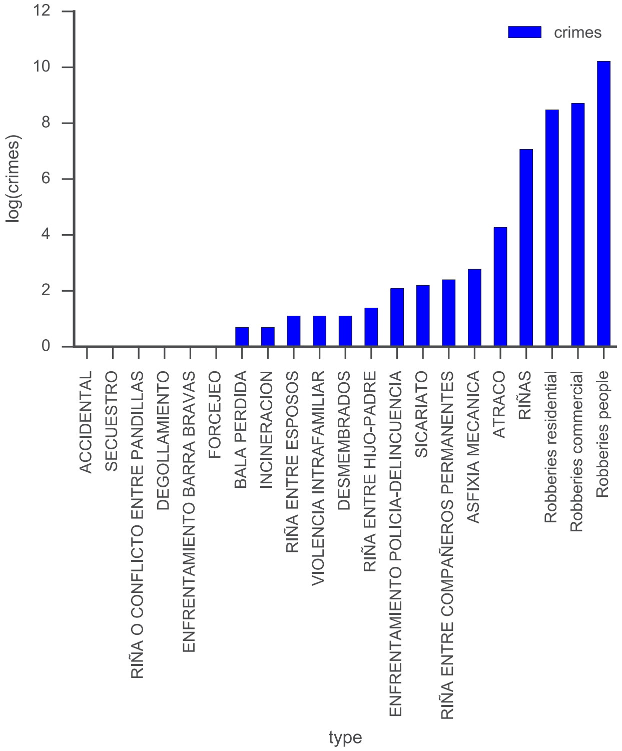
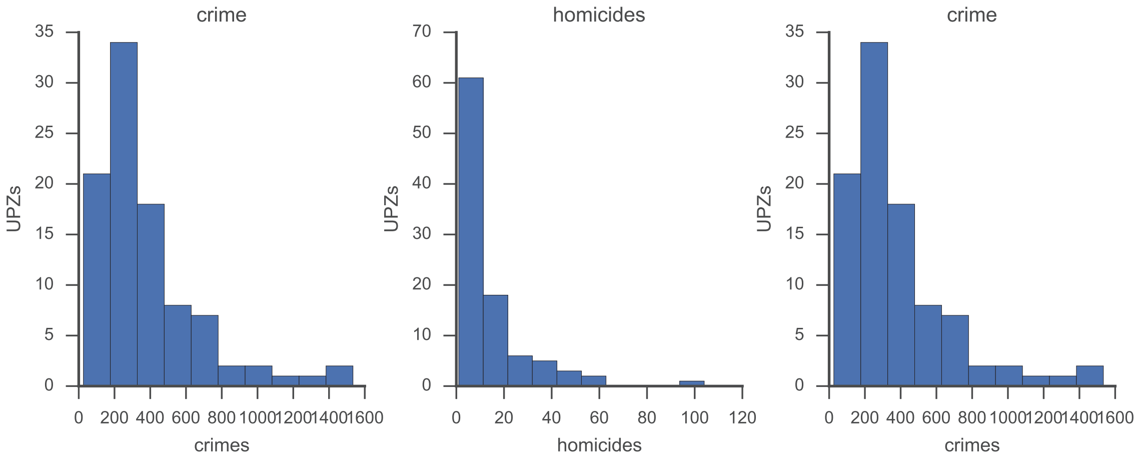


Figure 1: Crimes per type in Bogotá.



(a) (b) (c)

Figure 2: a) Crime distribution in the UPZ; b) Violent crime distribution in the UPZ; c) Robberies distribution in the UPZ.

4. Methods

Crime is considered a rare occurring phenomenon: a small proportion of people are victimized, with also few unreported and undiscovered crime. Crime offenses are not distributed evenly in space; they tend to cluster in parts of the city that exhibits similar characteristics, favorable to the spread of crime. Rare events, expressed through count variables, have been widely modeled on Ordinary Least Squares (OLS) through a logarithmic and square root transformation of the response variable. However, this introduces extra estimation bias, homoschedability assumptions and impossible predictions [14]. Moreover, spatial auto-correlation is rarely accounted for. For this reason, we use a Negative Binomial regression (NB) for count variables, and an eigenvector spatial filter (ESF) to account for spatial auto-correlation.

In this paper we want to *describe* a city but also to be able to *predict* it. Differently from the common meaning, here we use the word “predict” to mean the estimation of a property not directly observed (nowcasting).

Thus, we first employ a descriptive model for Bogota where it is possible to understand the interactions of each component of the model to describe crime events. Then, we create a predictive model validated with a 5-fold Cross-validation with 1000 repetitions (to avoid overfit). This allows to answer to the question “can I predict crime events from the characteristics of the city?”.

4.1 A regression model for rare events

The Poisson regression model, a class of Generalized linear models (GLM), is particularly attractive to model count response variable. However, its restrictive assumption to have identical mean and variance is violated with many real-world situations. The NB introduces an extra parameter k to the Poisson model with parameter λ_i , where the variability of λ_i has a gamma distribution with mean μi and index v. It follows that $Y_i \sim NB(k, \mu_i/k)$ and:

$$E(Y_i) = \kappa \frac{\mu_i}{\kappa} = \mu_i \qquad (1)$$

$$Var(Y_i) =| \mu_i \frac{\mu_i^2}{\kappa} \qquad (2)$$

with k accounting for the extra variability with a quadratic function on μ_i. and as As $k \rightarrow \infty$ the distribution of Y_i converges to a Poisson random variable. With a log-link function on μ_i, the NB model can be written as:

$$Ln[E(Y_i)] = Ln[E(i)] + \beta_0 + \sum_{k=1}^{n} X_k B_k \qquad (3)$$

where $X_0, X_1, \ldots, X_n$ are the covariates, $\beta_0, \beta_1, \ldots, \beta_n$ the regression parameters, and $LN[E(i)]$ is the offset variable and its role is to control for size differences across the units.

4.1.1 Spatial auto-correlation

Models dealing with spatial data analysis require to test for spatial auto-correlation. Positive (negative) spatial auto-correlation refers to the attitude of nearby attributes to have similar (dissimilar) values. There are numerous quantitative methods to measure and deal with spatial auto-correlation. The eigenvector spatial filter (ESF) [12] introduces a set of independent variables that account for the spatial relationship of the variables. These variables are a subset of the eigenvectors extracted from the numerator of the Moran's I coefficient [5]:

$$(I - \frac{11^T}{n})W(I - \frac{11^T}{n}) \qquad (4)$$

where I is a $(n \times n)$ identity matrix, 1 is a $n \times 1$ vector of ones, and W is a generic $(n \times n)$ distance matrix. This matrix can we either be defined with geographical distance or flow of people and freight.

To model the proximity of each region, we define the W distance matrix as the inverse squared distance separating the observations:

$$w_{i,j} = \begin{cases} 0 & ifi = j \\ 1/d^{\gamma}_{i,j} & \text{otherwise} \end{cases} \qquad (5)$$

where $d(i, j)$ is the Euclidean distance between i and j, and i is γ a penalization parameter. We set γ=2 to place a greater weight on close observations and leave a marginal role to distant observations. The matrix is variance stabilized.

The n eigenvectors describe the full-range of uncorrelated spatial patterns; but employing all n eigenvectors in a regression framework is not desirable for reasons of model parsimony. Thus, we first filter the eigenvectors with low (and opposite) Moran's I ($MI_E/MI_{max} \geq 0.25$). Then, from this subset we select the smallest subset of eigenvectors $\{E_1, E_2, . . . , E_p\}$ each time adding, in a step-wise fashion, the eigenvector that reduces the most the Akaike Information Criterion (AIC) of the model. The process stops when the AIC does not decrease anymore. The step-wise process does not guarantee to select the *best* eigenvectors, but it is a very simple and fast method to select the orthogonal eigenvectors to add. For further details on eigenvector selection and implementation strategies see Tiefelsdorf *et al.* [18]. The final subset of candidate eigenvectors represents the spatial filter for the variable analyzed. Thus, the aspatial NB model defined in Equation (3) takes the form:

$$Ln[E(Y_i)] = Ln[E(i)] + \beta_0 + \sum_{k=1}^{n} X_k B_k + \sum_{j=1}^{p} E_j B_{n+j} \qquad (6)$$

4.2 Spatial aggregation

Bogota is regionally recognized for the important strides it has made in reducing violence related to organized crime in recent years. To understand the multitude of aspects that influence crime in a neighborhood, we first have to define it. A neighborhood is a geographical unit composed by people who usually interact with each-other, and sharing common goals. This spatial community has a loose definition [13] and it has to be "big enough (in population) to swing weight in the city as a whole, but small enough so that street neighbourhoods were not lost or ignored". From this description we selected the Unidades de Planeamiento Zonal (UPZ) as a valid spatial aggregation for neighborhoods. The function of a UPZ is to help in the planning in the development of urban norms in the city. There are 113 UPZ and their average population is 63,720, with average area of 3.7 in2009.

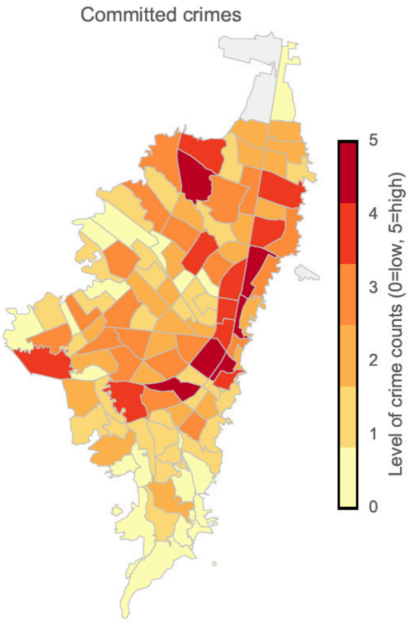


Figure 3: Crime counts in Bogota for each UPZ unit.

4.3 Measure of crime

Crime counts represent a measure to map and understand, in absolute terms, where the crime takes place. Nonetheless, crime rates over residential population are usually preferred as they assess the risk of people to be victimized in a particular location [1]. In a NB model, a rate variable is defined with an offset, which is a variable that is forced to have a coefficient of 1 in the model. This is particular useful to create risk maps, but we think it is a too restrictive constraint. To achieve a greater flexibility we prefer to add population as covariate and estimate its coefficient in the model. A population coefficient (B_p) greater than one means that spatial units with more population have higher rate of criminality. On the contrary, units with less population have fewer crimes per inhabitants when $B_p < 1$.

4.4 Daily routines

As aforementioned, daily routines are supposed to influence the presence of offenders, victims and guardians in a place. Moreover, frequent trips between two places are supposed to influence each other's crime.

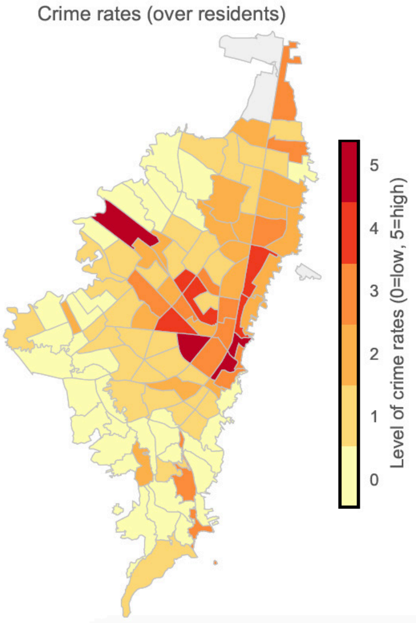

Crime rates (over residents)
Level of crime rates (0=low, 5=high)
5
4
3
2
1
0

Figure 4: Crime rate ($|crimes|/(\beta_p|population|)$) Bogota for each UPZ unit.

Wang *et al.* [20] used the taxi flow network with a spatial lag regression. We instead use the same ESF method used to take into account the spatialautocorrelation. Firstly we compute the total number of trips made in a typical day between spatial units. Then, we define the symmetric weight matrix W_t from the Mobility network dataset, by applying this transformation to the original mobility network W_m:

$$W_t = W_m^T W_m \tag{7}$$

The eigenvectors inserted in the NB model are supposed to be a proxy for the mobility network dependencies between spatial units.

4.5 Jane Jacobs' diversity

As aforementioned, our hypothesis is that Jane Jacobs' diversity conditions have an impact on crime measures. Thus, in this section we describe the metrics, priorly validated [7, 17], that operationalize the Jacobs' theory.

Land use mix. A common way of quantifying the mixture of land uses is through average Shannon entropy. The average entropy, which we here call Land Use Mix (LUM), is defined as:

$$\text{LUM}_{3L,i} = -\sum_{j\in 3L} \frac{P_{i,j}\log(P_{i,j})}{\log(|3L|)} \tag{8}$$

where $P_{i,j}$ is the percentage of square meters having land use j in unit i, and $3L$ = { residential, commercial and istitutional, park and recreational} represents the land uses considered. The LUM ranges between 0, wherein the unit is composed by only one land use (e.g. residential), and 1, where in developed area is equally shared among the n land-uses. The problem with this index is that it depends on the way land uses are grouped together, and no distinction is made on the order of land uses. Thus, an entropy of 0.75 could either mean high land use mix with a major role of residential land use, or high land use mix with a major role of parks. To better represent the different outcome we also employ a second entropy measure, based on the distinction between residential and non-residential land uses:

$$\text{LUM}_{rnr,i} = -\sum_{j\in rnr} \frac{P_{i,j}\log(P_{i,j})}{\log(|rnr|)} \tag{9}$$

where rnr = {residential, non-residential, }.

Jacobs argued for mixing primary uses so that people are on the street at different times of the day. To characterize spatial use in terms of activities, we determine whether each place is used daily (e.g., convenience stores, restaurants, sport facilities) or not. Based on that, we define the average accessibility of the buildings in a spatial unit i as:

$$\text{A}_i = \frac{1}{|B_i|}\sum_{j\in B_i} \text{dist}(j, \text{closest}(j, D))^{-1} \tag{10}$$

where D is the set of places that are used on a daily bases (e.g. convenience and grocery stores), dist(a, b) is the Euclidian distance between a and b, closest(a, C) is a function that finds the closest item in set C from point a, and B_i is the set of buildings in unit i.

Consistent with the methodology used by the website developers to calculate Walk Score[1], we define the weighted *walkability* score as:

$$\text{walk}_i = \frac{1}{|B_i|}\sum_{c\in C} w_c \sum_{b\in B_i} \text{wdist}(b, \text{closest}(b, \text{POI}_c))^{-1} \tag{11}$$

where C = {Grocery, Food, NightLife, Shops, Cultural}, POI_c is the set of POIs of category c, and w_c is the weight of importance to POIs, which is 3 for Food, Nightlife and Grocery POIs, 2 for Shops and 1 for others. wdist(a, b) is a linear function from 1, when the dist(a, b) <= 500 m, to 0 when the dist(a, b) >= 2500 m.

Small blocks. Small blocks are believed to support stationary activities and provide opportunities for short-term and low-intensity contacts, easing into interactions with other people in a relaxed and relatively undemanding way. We compute the average block area among the set B_i of blocks in unit i as:

$$\text{Blocks area}_i = \frac{1}{|B_i|} \sum_{b \in B_i} \text{area}(b) \qquad (12)$$

In addition, we compute the average distance between each building and the nearest street, a proxy for the concept *eyes on the street*, which suggest that the safety of neighborhoods can be maintained through continued surveillance of their inhabitants For each unit i and the set S of streets, it is defined as:

$$Eyes\ on\ the\ street_i = \frac{1}{|B_i|} \sum_{b \in B_i} \text{dist}(b, \text{closest}(b, S))^{-1} \qquad (13)$$

Buildings. Jacobs stressed the importance of having diverse buildings in a district to create vitality. Diverse buildings allow the mix of different socioeconomic groups, as well as the tendency of an easier accommodation of creative people and small enterprises.

Colombia has a fiscal policy, called stratum, that classifies buildings in different regimes of tax payments for utilities and rents. Stratum is based on the external physical characteristics of the building, and it reflects the quality of life of residents with a six-level classification from 1 (lower low) to 6 (high). For this reason, we computed the heterogeneity of a unit i as:

$$\text{Strata}_{\sigma_i} = \sqrt{\frac{1}{|H_i|} \sum_{b \in H_i} (\text{strata}_b - \overline{\text{Strata}_i})^2} \qquad (14)$$

where H_i is the set of houses belonging in district i.

Concentration. Jacobs' fourth and final condition is about having concentration of both buildings and people. First, we determine population density measures by dividing the number of people by the unit's net area.

$$\text{Population density}_i = \frac{|\text{Population}_i|}{\text{area}_i} \qquad (15)$$

Then we compute the floor-area density per each unit i as:

$$\text{Buildings density}_i = \frac{|\text{Buildings}_i|}{\text{area}_i} \qquad (16)$$

[1] http://www.walkscore.com

Border Vacuums. Border vacuums are places that act as physical obstacles to pedestrian activity. For instance, parks can be a hub of pedestrian activity, if efficiently managed [13], but they could also be deplorable places in which criminality flourishes (especially at night). Thus, we compute the average closeness of each building from the nearest park as:

$$\text{Closeness to LP}_i = (\frac{1}{|B_i|} \sum_{j \in B_i} \text{dist}(j, \text{closest}(j, LP)))^{-1} \tag{17}$$

where dist(j, closest(j, LP) is the distance between block j and its closest large park.

4.6 Covariates

The number of committed crimes is mainly influenced by the number of residents, their social disorganization and routine activity. Social disorganization is the inability of the neighborhood to maintain effective social control. Social disorganization is higher in deprived areas, social heterogeneous units and in places with high unemployment rate, which is also a proxy for motivated offenders. We define the social heterogeneity through unemployment rate:

$$\text{unemployment}_i = \frac{|\text{unemployed residents}_i|}{|\text{residents}_i|} \tag{18}$$

and a proxy of income heterogeneity through the weighted standard deviation of property values:

$$\text{social heterogeneity}_i = \sqrt{\frac{\sum_{b=1}^{n} w_b^2}{\left(\sum_{b=1}^{n} w_b\right)^2} \sigma^2} \tag{19}$$

where x_b is the property value of a block in spatial unit i and x_b is the residential population count of block b.

The number of residents is calculated as:

$$\text{population}_i = |\text{residents}_i| \tag{20}$$

4.7 Evaluation

The R^2 statistic is an intuitive interpretation of the proportion total variation in outcome that is accounted for by a OLS model. Concerning GLMs there is no directly analogous R^2 measure. For this reason, GLMs models are usually evaluated through their AIC, Pseudo-R^2 and Root Mean Squared Error (RMSE). One of the most interesting Pseudo-R^2 measures is the McFadder Pseudo-R^2, which:

$$\text{McFadder Pseudo-}R^2 = 1 - \frac{\log \widehat{L}(M_{full})}{\log \widehat{L}(M_{intercept})} \tag{21}$$

where $\hat{L}(M_{full})$ is the log likelihood of the full model and $\hat{L}(M_{intercept})$ is the log likelihood of the null model. It is worth to remember that this is not a true measure of fit, because it only compares the log likelihood of the full model with the one of the null model.

In the predictive model, evaluated through the K-fold Cross-validation (K=5), we create multiple models that use a subset of the features. Thus, the subsets are:

- Socioeconomic: demographic and social disorganization variables;
- City: Jane Jacobs' diversity variables;
- Dynamics: daily routines variables;

and the combinations of them. This allows to understand the contribution of each subset to the description of the crime events in a city.

5. Results

Our contribution in this paper is twofold. At first we focus on describing the relation of the various factors with crime. Then, we built a predictive model to understand how the results can be generalize in different cities.

5.1 Descriptive model

From the results (in Table 2) we can observe the β coefficients of features to understand the importance of each variable, holding the others as constant. The most important variables to describe crime are building density, population density and closeness to daily-use buildings. Particularly, the concentration variables are very important in describing crime, with building density that is positively correlated with crime (0.498). By contrast, the higher the population density is, the less crime events there are.

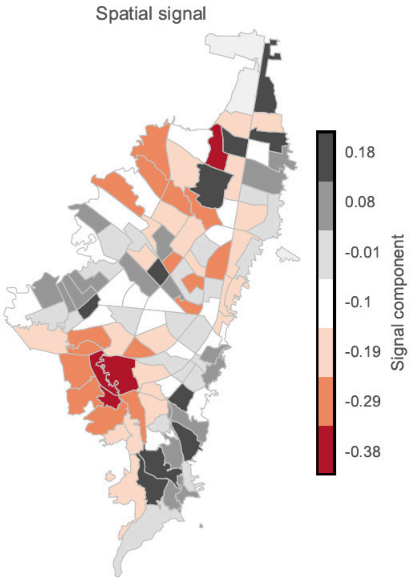


Figure 5: Stochastic signal component $B_e E$ representing the spatial auto-correlation for each UPZ in Bogota.

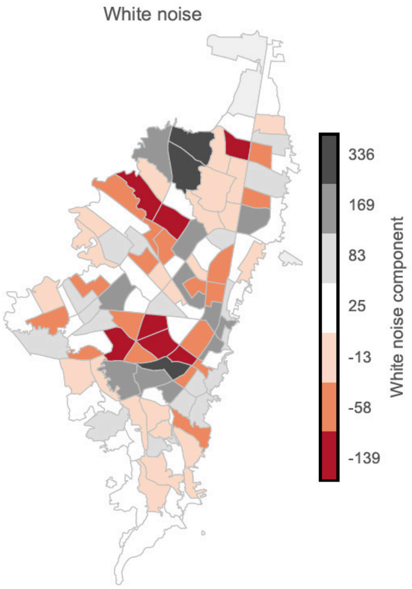


Figure 6: White noise component $y - (XB + B_e E)$ representing the undetected number of crimes for each UPZ in Bogota.

	β coefficient	std. error	p-value	95% CI
Land use				
Land use mix (LUM_3L)(8)	-0.078	0.029	**	[-0.135, -0.021]
Land use mix (LUM_{rnr})(9)	-0.098	0.034	**	[-0.166, 0.032]
Closeness daily buildingsl(10)	0.212	0.035	***	[0.144, 0.281]
Walkability score (11)	0.146	0.060	*	[0.029, 0.263]
Small blocks				
Block area (12)	-0.033	0.054		[-0.140, 0.073]
Eyes on the street (13)	0.270	0.055	***	[0.163, 0.378]
Buildings				
Strata diversity^s(14)	0.070	0.030	*	[0.011, 0.129]
Concentration			***	
Population density (15)	-0.363	0.045	***	[-0.451, -0.276]
Building density (16)	0.498	0.054	***	[0.393, 0.603]
Vacuums				
Closeness parks^l(17)	0.031	0.039		[-0.045, 0.108]
Covariates				
Unemployment (18)	0.110	0.036	**	[0.039, 0.181]
Population (20)	0.6647	0.058	***	[0.550, 0.779]
Social heterogeneity (19)	0.159	0.041	***	[0.078, 0.240]
OD eigenvectors (Sec. 4.4)	-0.184	0.042	***	[-0.267, -0.101]
Spatial eigenvectors	2			
McFadder Pseudo-R^2 †	0.145			
RMSE	92.81			
Moran’s I (p-value)	0.02 (0.42)			

† This is not a true measure of fit, and not bounded to 1. It indicates the degree to which the model parameters improve upon the prediction of the null model.

Table 1: Negative Binomial regression model that describes the number of crimes in each spatial unit.

	S	C	D	S+D	C+D	S+C	FULL
McFadder Pseudo-R^2 †	0.077	0.113	0.085	0.106	0.120	0.141	**0.143**
RMSE	231.93	145.04	312.70	181.76	133.36	143.35	**127.76**

† This is not a true measure of fit, and not bounded to 1. It indicates the degree to which the model parameters improve upon the prediction of the null model.

Table 2: Negative Binomial regression models that predict the number of crime in each spatial unit. The results are average across 1000 iterations of a 5-fold Cross-validation. S: demographic and social disorganization variables only; C: Jane Jacobs’ diversity variables only; D: daily routine variables only; S+D: demographic, social disorganization and daily routine variables; C+D: Jane Jacobs’ diversity and routine variables; S+C: demographic,

Small blocks are also very related to crime, as the distance of buildings from the nearest street has a positive correlation with crime events (0.270). This is in accordance to the *eyes on the street* theory of Jane Jacobs, that generates a virtuous loop which, in turn, increases public safety.

Covariates coming from criminology literature are also significative. We found that Social heterogeneity and unemployment are positively correlated with crime. Thus, higher deprivation and disorganization might stop the mechanisms by which residents themselves achieve guardianship and public order.

Finally, we find that social isolation increase crime, as it supposedly limits neighborhood possibilities and social capital [10]. Thus, we see that highly-connected points of the city experiencing lower number of crimes.

5.2 Predictive model

Our preliminary findings (see Table 2) indicate that structural characteristics of the city, namely Jacobs' diversity conditions, are a better predictor of the target variables (the number of homicides and robberies) than socioeconomic conditions such as unemployment and deprivation. Mobility networks, and thus the routine activity theory, improve the prediction of the model by 15%. The combination of structural and socioeconomic information provides better predictions than each on its own.

6. Discussion

In this paper we modeled, for the first time, multiple aspects of urban life to describe and predict crime. We have done so by operationalizing the Jane Jacobs theory to describe the urban fabric, the social disorganization and the routine activity variables from criminology. We can now discuss some implications of our work.

Descriptive maps. Maps are invaluable tools in criminology to understand where the problems are, and to evaluate initiatives for crime prevention. Thus, our framework allows policy makers to visually analyze crime rates (Figure 2), crime counts (Figure 1) and, most importantly, spatial auto-correlations (Figure 3) and heterogeneous effects (Figure 4).

Factors for crime. With our descriptive model we have shown that it is possible to have a static description of what happens in the city, and how the multitude of complex features of city life come together. Therefore, now more than ever, it is important to control for the many relations that come together, especially as a consequence of urban fabric and mobility. Urban diversity and natural surveillance theories play an important role in the proliferation of crime, and the knowledge of this role can be exploited by policy makers to reduce crime.

Generalization. Our predictive model that the combination of structural and socioeconomic information, and mobility, provides better predictions than each on its own. However, it is striking that the urban fabric has such an important role in the prediction. Thus, we think it is critical to consider it in new models and analysis in different cities in the world.

This work is not without any limitation. First, we don't analyze crime over time. Then, cities are not to be considered island unto themselves, as they are embedded in a country-wide complex system of social interactions. Routine of residents exposes them to different cities, conditions and possibilities on a daily basis. Thus, we think that in the next future it is important to consider also these factors.

Our findings are apt to crime control and prevention action plans for Colombian cities. This study provides valuable insights for local governments so that they can base urban management decisions on empirical evidence on the deterrents of crime.

7. References

[1] Martin A Andresen. Crime measures and the spatial analysis of criminal activity. *British Journal of criminology*, 46(2):258–285, 2006.

[2] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Moves on the street: *Classifying* crime hotspots using aggregated anonymized data on people dynamics. Big Data, 3(3):148–158, 2015.

[3] Paul J. Brantingham. *Environmental Criminology*. Waveland Press, 1991.

[4] Ernest Watson Burgess. *The growth of the city: an introduction to a research project*. Ardent Media, 1967.

[5] Andrew Cliff and Keith Ord. Spatial autocorrelation. Technical report, 1973.

[6] Lawrence E Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.

[7] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: A mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web*, pages 413–423. International World Wide Web Conferences Steering Committee, 2016.

[8] Neva R Deardorff and Clifford R Shaw. Delinquency areas: A study of the geographic distribution of school truants, juvenile delinquents, and adult offenders in chicago, 1930.

[9] Corina Graif, Andrew S Gladfelter, and Stephen A Matthews. Urban poverty and neighborhood effects on crime: Incorporating spatial and network perspectives. *Sociology Compass*, 8(9):1140–1155, 2014.

[10] Corina Graif, Alina Lungeanu, and Alyssa M Yetter. Neighborhood isolation in chicago: Violent crime effects on structural isolation and homophily in inter-neighborhood commuting networks. *Social Networks*, 2017.

[11] Corina Graif and Robert J Sampson. Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates. *Homicide studies*, 2009.

[12] Daniel A Griffith. *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*. Springer Science & Business Media, 2013.

[13] Jane Jacobs. *The death and life of great American cities*. Vintage, 1961.

[14] Robert B O'hara and D Johan Kotze. Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122, 2010.

[15] Robert E Park. The city: Suggestions for the investigation of human behavior in the city environment. *The American Journal of Sociology*, 20(5):577–612, 1915.

[16] Gabriel Rosser, Toby Davies, Kate J Bowers, Shane D Johnson, and Tao Cheng. Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*, pages 1–26, 2016.

[17] Hyungun Sung and Sugie Lee. Residential built environment and walking activity: Empirical evidence of jane jacobs' urban vitality. *Transportation Research Part D: Transport and Environment*, 41:318–329, 2015.

[18] Michael Tiefelsdorf and Daniel A Griffith. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A*, 39(5):1193–1221, 2007.

[19] Martin Traunmueller, Giovanni Quattrone, and Licia Capra. Mining mobile phone data to investigate urban crime theories at scale. In *International Conference on Social Informatics*, pages 396–411. Springer, 2014.

[20] Hongjian Wang, Zhenhui Li, Daniel Kifer, and Corina Graif. Crime rate inference with big data. In KDD, 2016.

Papiers de Recherche de l’AFD

Les *Papiers de Recherche de l’AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ilss’adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l’ensemble des sujets de travail de l’AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l’ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l’AFD n’en exclut aucune autre.

L’Agence Française de Développement (AFD), institution financière publique qui met en oeuvre la politique définie par le gouvernement français, agit pour combattre la pauvreté et favoriser le développement durable. Présente sur quatre continents à travers un réseau de 72 bureaux, l’AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète. En 2014, l’AFD a consacré 8,1 milliards d’euros au financement de projets dans les pays en développement et en faveur des Outre-mer.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l’AFD. Ce document est publié sous l’entière responsabilité de son (ses) auteur(s).

Les *Papiers de Recherche* sont téléchargeables sur : http://librairie.afd.fr/

AFD Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

Agence Française de Développement (AFD), a public financial institution that implements the policy

defined by the French Government, works to combat poverty and promote sustainable development. AFD operates on four continents via a network of 72 offices and finances and supports projects that improve living conditions for populations, boost economic growth and protect the planet. In 2014, AFD earmarked EUR 8.1bn to finance projects in developing countries and for overseas France.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

AFD Research Papers can be downloaded from: http://librairie.afd.fr/en/

AFD, 5 rue Roland Barthes
75598 Paris Cedex 12, France
ResearchPapers@afd.fr
ISSN 2492 - 2846

BIG DATA

TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018

BIG DATA
TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018

DATA-POP ALLIANCE

UNDERSTANDING THE RELATIONSHIP BETWEEN SHORT AND LONG TERM MOBILITY

Sveta Milusheva - sveta.milusheva@gmail.com
Elisabeth zu Erbach-Schoenberg - elisabeth.zu.es@gmail.com
Linus Bengtsson - University of Southampton, Flowminer
Erik Wetter - University of Southampton, Flowminer
Andy Tatem - Department of Geography and Environment, University of Southampton , Flowminder

DATA-POP ALLIANCE FLOWMINDER.ORG

Funded by:

AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

UNDERSTANDING THE RELATIONSHIP BETWEEN SHORT AND LONG TERM MOBILITY

Sveta Milusheva,
sveta.milusheva@gmail.com

Elisabeth zu Erbach-Schoenberg,
elisabeth.zu.es@gmail.com

Linus Bengtsson,
University of Southampton, Flowminer

Erik Wetter,
University of Southampton, Flowminer

Andy Tatem, Department of Geography and Environment, University of Southampton, Flowminder

Please cite this paper as:

MILUSHEVA, S., E. zu ERBACH-SCHOENBERG, L. BENGTSSON, E. WETTER and

A. TATEM (2017), "Understanding the Relationship Between Short and Long Term Mobility", AFD Research Paper Series, No. 2017-69, June.

Abstract

Populations are highly mobile, both in terms of long term movements of individuals relocating their place of residence as well as shorter term mobility such as commuting, seasonal travel and recreational trips. Working with call detail record data from Namibia and Senegal, we study population migration and its link to short term movement. We compare the short term mobility estimates extracted from call detail records to census data in the two countries and find a strong annual relationship, as well as distinct daily patterns in the relationship between long and short term movement. The relationship is strongest for holidays, and we find it to be consistent both across countries as well as across multiple years. In particular, we observe periods of increased travel on migration routes around holidays, with net short term travel in the opposite direction of the direction of migration before the holiday and net travel in the same direction after. Using the Namibia data set, which spans several years, we investigate the link between short term mobility and long term relocation on an individual level, allowing us to gain insights into the mechanisms of interaction of short and long term mobility. We find that it is common for individuals to both visit the place they will migrate to prior to migration and also visit their place of origin after migrating. Additionally, distance between the origin and destination of a migration has a significant influence on the probability of a short term trip associated with a long term move.

The Senegal dataset provides information on the full network of users, which we use to study the relationship between the location of network contacts and probability of traveling to those locations, investigating the importance of social contacts for mobility. We find that while the majority of social contacts in different regions can be explained by long term migration patterns between regions, which in turn are linked to short term movement patterns, social contacts can explain some of the additional short term movement not captured by the long term migration. We also find non-linear relationships between the probability of visiting a region and the number and strength of contacts, as well as between the duration of a visit and social contacts. These results can help inform evidence-based policies that target some of the negative externalities of short term population movement such as spread of infectious disease, increased congestion, and inadequate infrastructure.

*We would herewith like to give credit to MTC, whose assistance and support made the realization of this Study/Project possible. Anonymous mobile phone data have been made available by Orange and Sonatel within the framework of the D4D-Senegal challenge, and we want to thank Sonatel and Orange for their continued support.

Correspondence: sveta.milusheva@gmail.com, elisabeth.zu.es@gmail.com

Key words: Migration, Big data, Call detail Records

JEL Classification: O15, R23, C8, C88, C89

Original version: English

Accepted: July 2017

TABLE OF CONTENTS

1. Introduction

Migration is observed all over the world, both in terms of international migration between countries as well as internal migration within a country. Urbanization is continuing across Africa, although almost two thirds of people still live in rural areas (Henderson, Storeygard, and Deichmann 2014). Linked to migration and urbanization are short-term movements, which in contrast to permanent relocation on the scale of years are temporary relocations on shorter time scales such as days, weeks or months. These short term movements can be the result of economic activity (seasonal labour) (Atamanov and Berg 2012, Etzold et al. 2014, Rain 1999), disaster or conflict related displacement (Bengtsson et al. 2011, Wilson et al. 2016, Poncelet et al. 2010, Doocy et al. 2015) or recreational travel or pilgrimages during holidays (Ahmed, Arabi, and Memish 2006, Rinschede 1992, Mokashi 1987). Permanent relocation can generate short term recreational travel to visit friends and family. Conversely, having visited friends or relatives that have moved away might inform individuals about economic opportunities away from their current place of residence, in which case short term movements could increase the likelihood of long term migration. Additionally, economic opportunities could lead to seasonal, temporary migration of some members of a household, which later causes permanent change of residence for the whole household. Intuitively, it seems migration and short-term mobility as well as the social networks of migrating individuals might influence each other.

Short term internal population movements have implications for a range of settings. They have been shown to have an impact on spread of diseases (Oster 2012, Prothero 1977, Balcan et al. 2009, Huang and Tatem 2013, Tatem and Smith 2010, Wesolowski et al. 2015b, Wesolowski et al. 2015a, Stuckler et al. 2011, Pison et al. 1993), one example being the relationship between mobility and malaria (Lynch et al. 2015, Osorio, Todd, and Bradley 2004, Siri et al. 2010, Yukich et al. 2013, Alkhalife 2003, Littrell et al. 2013). Increases in short term mobility will likely increase pollution, congestion in urban areas and have potential implications for economic activity. In order for policy makers to address some of these potentially negative externalities of population movement, it is necessary to understand short term mobility patterns as well as some of the factors that influence them in order to be able to predict what short term movement fluctuations will look like and to incorporate this knowledge into existing policies. Yet, the drivers of short term mobility have not been studied due to the lack of available data on fine spatial and temporal scales.

Migration is often measured in the context of a census as change of permanent residence on a given spatial scale, and thus migration data is available for many countries. Short-term movements on the other hand are harder to capture and the extent of seasonal mobility is often unknown or only studied in the context of travel surveys for particular populations. Mobile phone data provide a powerful source to measure these shorter term mobility patterns for whole countries (Wesolowski et al. 2012, Tatem et al. 2014, Wesolowski et al. 2015b). Call detail records (CDRs), collected by cell phone companies for billing purposes, provide geo-location and timing for calls and text messages, allowing researchers to observe changes in location for individuals within a country. Where CDR data are made available to researchers in an anonymised form, it is possible to determine country-wide mobility patterns with fine temporal and spatial resolution. These data are frequently used to estimate population mobility over short periods, however, it has been shown that they can also be used to estimate migration, where census based migration estimates are not available (Wesolowski et al. 2013). Here, we study the flip side of this question, using CDR data to study the relationship between short term population movement and long term migration and investigating the motivating factors for this relationship in order to predict short term movement when estimates are not available. Understanding the relationship between the two will allow policy makers to better plan for potential increases in short term movement in order to mitigate negative impacts of these increases in movement. In addition, for countries where cell phone data are not available, understanding the link between long and short term movement can inform policy makers in which cases it is valid to use available census or survey data on long term migration to predict short term movements.

In the context of this study, we focus on migration patterns in two countries Namibia and Senegal. Internal migration in Namibia is common, but varies by demographics, migration being slightly more likely for men and most common for the age group 20-34 (Namibia Statistics Agency 2015). In 2004, 27 percent of the Senegalese population was recorded as an internal migrant (Fall, Carretero, and Sarr 2010). Much of the migration is rural to urban, due to a reduction in the agricultural production index, though there is also circular migration, pastoral movements, and temporary rural to rural or urban to urban migration (Fall, Carretero, and Sarr 2010, Goldsmith, Gunjal, and Ndarishikanye 2004, Linares 2003, Adriansen 2008, Herrera and Sahn 2013). A study focused specifically on migrants to Dakar finds that 87% of male migrants and 81% of female migrants visit their home locations, primarily for holidays, family ceremonies and religious festivals (Fall 1998). Similarly, for Namibia, rural to urban migration makes up the largest proportion of internal migration (Namibia Statistics Agency 2015) and the percentage of the population living in urban areas increased from 33% in 2001 to 43% in 2011. Of the people living in regions that contain urban centers, 40% are migrants from other regions. During major holidays, large outflows out of the capital Windhoek can be observed, with most individuals traveling to the north of the country, most likely to visit relatives (Erbach-Schoenberg et al. 2016).

Working with call detail record data from Namibia and Senegal, we study the link between migration and short term movement through three different analyses. We start out by exploring the relationship between short term mobility measured using CDR data and long term migration measured using census data in the two countries. Using the Namibia data set, which spans several years, we then investigate the link between short term mobility and long term relocation on an individual level, allowing us to gain insights into the mechanisms of interaction of short and long term mobility. The Senegal dataset provides information on the full network of users, which allows us to study the relationship between the location of network contacts and probability of traveling to those locations, investigating the importance of social contacts for mobility. By using data from two different countries, we have the opportunity to test how well models relating short and long term mobility translate from one country to another, which will be important in extrapolating our research to other country contexts. Finally, in the last section we discuss the implications and potential uses of the results presented in this paper.

2. Data

2.1 Census data

Data on migration were provided by the Agence Nationale de la Statistique et de la Demographie (ANSD) and the Namibia Statistics Agency for Senegal and Namibia respectively. For Namibia, the migration data used were collected in the context of the 2011 census and the data for Senegal come from a 10 pecent sample of the 2013 census data. For both countries we use a variable that defines migration as a change in usual residence from the previous year compared to the time of the census.[1]

The specific question asked in the census for Namibia is "where did NAME usually live since September 2010?" (Namibia Statistics Agency 2015). The answer from this question is compared with the answer to the question "where does NAME usually live?" (referring to at the time of the census done in 2011) and migration is inferred if the location has changed between 2010 and 2011.[2] The data for this question is only available at the region level for the 13 regions of Namibia. For Senegal, the specific question asked in the census was "where did you reside one year ago." If the location specified is different from the location the respondent states to be the current residence, then a long term migration is inferred between the two locations. While the Senegal census data is available at the commune level, which is a higher granularity, we aggregate the data up to the region level for the 14 regions of Senegal in order to work at a spatial level comparable to Namibia. This helps facilitate the comparison of results between the two countries.

Out of the 12.9 million individuals in the census that responded to the question of where they were located one year ago with a valid response that could be aggregated at the region level, there were 344,345 individuals that had a long term migration in 2013.[3] In Namibia, a migration from one region to another was recorded for 40,867 individuals (Namibia Statistics Agency 2015). The two countries are very similar in that 45% of all long term migrations occur between neighboring regions in both countries.

2.2 Mobile phone data

Call detail records (CDRs) are collected routinely by mobile phone providers for billing purposes. When a subscriber makes or receives a call or text, a record is created by the operator, containing a unique ID, a timestamp for the time of the communication as well as the location of the closest mobile phone mast through which the communication is routed. In some cases, network data is also available. Network data provides the IDs of both the caller and the receiver, thus allowing for the linking of individuals into a communication network. The data are anonymised before they are shared with researchers in order to preserve the privacy of users.

For Senegal, we use CDR data provided by Sonatel and Orange in the context of the Data for Development Challenge. The data consist of call and text records for Senegal between January 1, 2013 and December 31, 2013 for all of Sonatel's user base. In 2013, Sonatel had slightly over 9.5 million unique phone numbers in its network, representing a large portion of the 13.8 million Senegalese population. Out of the three telecom providers in Senegal, Orange Telecom had between 56 and 62 percent of the cell phone market in 2013 (ARTP 2013), though according to the *Listening to Senegal* survey done in 2014, 83% of those surveyed with a cell phone cite Sonatel as their primary provider and 89% cite having a Sonatel SIM card. The data contains information on all calls and texts made or received by an individual, namely the time and date of the communication and the location of the closest cell phone tower, which makes it possible to measure movement as people change locations based on making calls/texts from different towers.

[1] The Namibia data also includes information on lifelong migration, but this information is not used in the current study.

[2] The Namibia census agency notes that the wording of the question concerning the usual location lived since September 2010 could have prompted some respondents to provide their current residence rather than the former residence even if they lived somewhere else in 2010, because "since" could potentially include the current time. If that were to systematically happen, then long term migration would be underestimated with the census data. Implications of this for our analyses are considered in the discussion analysis.

[3] The 10% sample in the data was weighted in order to be representative of the full population. Individuals that responded that they were abroad one year ago are not included, only internal migrants who provided alocation that could be assigned a region within Senegal are included.

Each SIM card has an anonymised identification number associated with it, which allows linking together all communications and locations for a particular card over time. In addition, the data contains not only the anonymous ID of the SIM that makes or receives a call/text, but also the anonymous ID of the SIM that is receiving or making that call/text. With this information, the mobile phone data can be used to create both a contact location network of the locations where an individual has contacts that he or she has been in touch with over phone, as well as a locational network of the places the individual has physically visited.

For Namibia, CDR data were provided by MTC with the purpose of estimating internal mobility and its impact on malaria transmission. The data span a period of 3.5 years, from October 2010 to May 2014 and include all subscribers to MTC, corresponding to 72 Billion entries for a user base of 4.5 Million unique users. MTC's market share is high at 76% in 2012, and 95% of the population are covered by the network (http://www.mtc.com.na/coverage). Each data entry corresponds to a communication, either a call or text, made or received. Each entry contains a unique identifier, time and date for the communication as well as the tower the communication was routed through. There is no network data available for Namibia. While both CDR data sets have a finer spatial resolution than region level, as the locations in the data are on the level of mobile phone towers, we here aggregate the data to the region level in order to match the census data and to further protect the privacy of users.

2.3 Measuring short and long term moves from CDR data

To define short term movement, we first calculate a daily location for each user at the region level. This daily location is defined based on the last call or text of the day, with missing days assigned the location of the closest day with data. In Namibia there are 13 regions, while Senegal has 14 regions, though two of the regions are excluded in the analyses as they experience large pilgrimages during the year. Pilgrimage regions are excluded since the large population movements observed are representative of a different driver of movement that is unrelated to the relationship between short and long term migration patterns. Results including the pilgrimage regions are included and described in detail in Appendix B. On the individual level, we define a short term move as a change in a user's location from one day to the next. Days without calls and therefore without a defined location are interpolated using the user's location from the closest day with data.[4] We aggregated these individual level data to create daily region level mobility networks. In Senegal, there are a total of around 63.5 million short term trips that take place in 2013. On average that is around 174,469 trips per day.[5] In Namibia, the CDR data capture 64 million short trips over the whole period and 20 million short trips in 2013, which is an average of 55,245 trips per day. While the two countries were very similar in the percentage of long term migrations between neighboring regions, they differ some when it comes to short term movement. For Senegal, 63% of short term trips occur between neighboring regions, while for Namibia around 80% of trips are between neighboring regions.

There are some limitations to the data in calculating movement. Since we define a short term move as a change in location from one day to the next, if someone lives on the border between two regions and commutes for work or other reasons, we might pick up this movement, which represents a different mechanism from the one studied in this paper. To minimize this issue we use the last call or text of the day to define the daily location in order to increase the probability of capturing a user at their place of residence. There is still the possibility that an individual might not have any evening calls, meaning that the last call of the day will be early in the day but as we can see in Figure 1, over 80% of calls/texts take place between 6pm and midnight in Namibia. Therefore, we partially mitigate the issue of capturing commuting trips by using the last call of the day to define the daily location. Additionally, while using a very low granularity and defining movement at the region level does not take advantage of the high granularity available with mobile phone data, it helps in decreasing the level of commuting that is picked up with our definition of movement.

An additional limitation of the data is that we can only observe a user's location for days when he or she has a phone communication. For days with no data, we fill in the missing days based on the location of the closest days with data. For individuals that are missing many days of data, this potentially means interpolating a large number of days, and it is possible the person is located somewhere else, which we cannot capture. Figure 2 shows a histogram of the percentage of SIMs in the data with a given number of days of activity in 2013 in Senegal and in Namibia. We can see that for both countries, there are a number of individuals that have less than 30 days of data in 2013, but there are also many that have data for every single day, or for most days in the year. The data in Namibia are a bit more polarized with more than 50% of individuals falling in the lowest or highest category of usage and fewer in the middle, while for Senegal the distribution is more spread out[6]. Nevertheless, the countries are very similar on average, with individuals having 155 days of data in Senegal and 151 days of data in Namibia.

[4] Filling in missing days in this way means that if there is a period of time where the person does not use the SIM card and right before and right after this period the person is seen in different locations, then it is assumed the change in location occurs halfway between the two days when data is available.

[5] Note that we do not have movement for January 1, 2013 for Senegal, so the number of short term movesis based on 364 days. This is due to the fact that to calculate movement on a given date we need to have a location for the day before.

[6] Note that we only fill in undefined locations between the first day of activity and last day of activity for a user. This way users who are only active for a very short period will not be included in the data outside of that period.

Given the length of the Namibia data set, we can also observe long term migrations in the CDR data. While there are many ways of defining long term migration, we use the definition of an individual living in one region for at least 6 consecutive months followed by at least six consecutive months in a different region. This definition was chosen to match the definition of place of residence in the census (Namibia Statistics Agency 2015) and is similar to the methodology used by Blumenstock (2012), though we use 6 months rather than 12 months as the cutoff.[7]

2.4 Example of Movement Patterns in Senegal and Namibia

Figure 3 looks at movement into the regions where the capitals are located in both countries in order to exemplify what the movement looks like. For Senegal, the capital Dakar is located in the Dakar region, while for Namibia the capital Windhoek is located in the region Khomas. The two graphs on the left show movement based on the percent of long term migrants from the census data entering from each region into the capital region. In both countries, we see a relatively even distribution of individuals migrating into the capital, with around the same number of people entering from most regions. In Namibia, there are two regions from which there are very few migrants into the capital and one from which there are more migrants than the rest, but generally, it's a relatively even distribution. In Senegal, we see similarly a relatively even distribution in the western regions of the country that are closer to Dakar, with magnitudes similar to those in Namibia. There are more regions with fewer migrants though in Senegal, with very low numbers coming from the five regions to the west. This is in line with the geography of the countries since having the capital in the middle of the country, like in Namibia, means that it is relatively easier to get there from anywhere in the country. In Senegal, the capital is the western most tip of the country, and therefore it means those living in the eastern part of the country have to travel much further, which limits the number of migrants.

The two graphs on the right based on the CDR data show the percent of people entering from each of the other regions. When we look at the short term moves into the capitals, we see that distance plays an even bigger role in where the majority of short term moves come from. We see that for Namibia, over 60% of short term visits come from regions that border the capital. In Senegal, this is even starker since there is only one region that borders Dakar (Thies), so we see almost half of all short term movement coming from that region. This could represent several different factors. First, it is easiest to travel from a neighboring region, so we might see the most short term movement coming from the one neighboring region. Additionally, it is possible that there might be more commuting from the neighboring regions, and while this is spread out in Namibia across multiple neighboring regions, in Senegal it is concentrated since there is only one.[8] As people travel into Dakar from other regions, they all have to pass through Thies in order to get to Dakar, so it is possible some people spend a night in this region if they are coming from very far away and make a phone call while there; therefore, due to the definition of movement, we are picking up their movement from Thies to Dakar and not from the farther region that the person is originally coming from.

2.5 Short term movements and census migration

We first look at the relationship between short term movement and migration on an annual basis, summing all short term moves going from region x to region y in a year and summing all long term migration trips from region x to region y in the census data. We use 2013 for both Senegal and Namibia in order to be consistent with the timing that we are using to compare the two countries. The observations are directional, with separate observations for movement from x to y and for movement from y to x. We run regressions of total short term movement on total long term migration, clustering the data at both the origin and destination levels using two-way clustering.

To see how the relationship between short and long term movement shifts from day to day, we can look at the relationship between census data and short term mobility on a daily level. We run regressions of daily short term movement on long term migration in the census data for each day of 2013 for both Namibia and Senegal. To investigate travel behavior around holidays further, we look at the net flows on each route to take into account the direction of flows. Around holidays there are two possibilities, with day to day net short term movement being either in the same direction as the net long term movement or in the opposite direction. Movement in the opposite direction might correspond to individuals

that have migrated traveling back to their place of origin or having family returning after a visit. Movement in the same direction could correspond to family from their place of origin traveling to see them or the migrant returning after a visit to the place of origin. We explore this by regressing the net daily movement between locations on the net long term migration.

[7] A region is assigned to each month based on the location where the most days were spent in that month, where a day is assigned a location according to the last call or text of the day and missing days are assigned to the location of the closest day with a call or text.

[8] Though we have already discussed strategies that we have used in order to minimize the amount of commuting that we pick up in the data.

2.6 Drivers of the relationship between short term movements and census migration

In addition to looking at how short and long term movement are related, we also provide evidence for some of the motivating factors behind this relationship. While a number of factors may exist, we focus on investigating three processes:

1. Individuals might travel to a new location and afterwards migrate to it. This could be because they might assess the new location before moving by visiting someone that has already migrated or investigating economic opportunities in person.[9]
2. Individuals travel to their location of origin after migration, most likely to visit contacts, such as friends and family.
3. Individuals that are not necessarily migrants travel along routes with high migration because they want to visit social contacts that have chosen to migrate

In order to explore processes one and two, we can use the data from Namibia which allows us to follow SIMs over a long period of time. Using this data, we can observe long term migrations for individuals in the data and also measure any short term moves that these individuals might make prior and post migration. If an individual has migrated from location x to location y, we count a short term move prior to migration as the individual having traveled from x to y as well as from y to x prior to the long term migration. Similarly, a short term move after the migration is counted as an individual traveling from y to x and from x to y after the long term migration.[10]

We study all of the long term moves that take place in the Namibia dataset and identify all short term moves taken by the same individual between the origin and destinations prior to the long term move and after the long term move. We break down visits before and visits after by the origin and destinations of individuals, since the likelihood of a visit might depend on the specific locations. To more explicitly explore the influence of location, and specifically the influence of distance on the likelihood of a visit before or a visit after, we estimate two logistic models. In the first, we have a dummy for whether or not the migrant made a visit before migrating on the left hand side, and we have dummies for the region of origin and dummies for the region of destination, along with controls for month and year fixed effects. The same model is also run including distance. These two are analogously estimated for visits after.

We then turn to explore process three. First, we compare the annual relationships between short term movement, long term movement, and social contacts. For these comparisons we use total flows instead of directional flows, summing the number of people going from x to y and from y to x. For social contacts, we sum the total number of social contacts between regions x and y.[11] We then use regression analysis to look at how short term movement, long term movement and social networks are related to each other.

On the individual level, we can study how having a social contact in a region influences individuals' likelihood of visiting the region. We first estimate a conditional logit model with a dummy on the left hand side for having visited a particular region. On the right hand side we have the number of people in that region that the individual interacts with as well as the total number of calls/texts that the person has with people in that region, the latter measuring the strength of the relationship[12]. Additionally, we include a dummy for whether the person has contact with at least one person in the region. The reason for this is that the relationship between contacts and visits might not be linear, with the first contact being more important in making the decision to visit or not and each additional contact influencing the decision to a lesser degree, leading to diminishing returns for each additional contact. The difference in the likelihood of visiting when a person has one contact versus two in a location is potentially much larger than the difference in the likelihood of visiting when a person has 10 versus 11 contacts in a location. We test this by also running a specification where we include a squared term for number of contacts as well as a squared term for number of calls/texts.

3. Results

3.1 Relationship between Migration and Short Term Movement

For Senegal we observe around 127 short term moves between regions x and y in 2013 for each long term migrant going from region x to region y recorded in the census. In Namibia, this number is higher, with around 506 short term moves per migrant in the census. In both cases, the R^2 is between 0.3 and 0.4, with long term migration patterns being able to explain around 30-40% of variation in short term movements happening on an annual basis. We find a strong positive

[9] Based on this analysis, it will not be possible to disentangle whether the motivation for a prior visit iseconomic opportunity or visiting a contact, though in many cases, it is probably a combination of the two.

[10] This definition is relatively restrictive as indirect travel, for example if an individual does not travel between x and y directly but instead stops at an intermediary location for a day or more, will not be counted.

[11] A social contact between x and y is counted as a person whose home location is x making or receiving calls from a person whose home location is y. The home location is based on the location where the most number of days are spent in 2013.

[12] Using the assumption commonly made that individuals are more likely to call and text individuals with whom they have tronger ties.

relationship between census migration numbers and short term movement measured through CDRs for both countries. Table 1 shows the results from the annual regression.

On a daily level, the relationship between movement and census migration remains relatively consistent for both countries, but with significant jumps occurring at distinct points (Figure 4). On average, there are around 0.35 daily moves per long term migration in Senegal and around 1.38 daily moves per long term move in Namibia. The spikes in the relationship between short and long term movement in both countries are related to the major holidays, marked as vertical red lines. This implies that people are more likely to make trips between migrant places of origin and destination around holidays.

Next, we look at net movement to investigate the directionality of these increases in travel on migration routes during holiday periods. We find that before a holiday the coefficient between net short term movement and net long term movement is negative while after a holiday we observe a positive relationship (Figure 5). This signifies net short term movement in the opposite direction of the long term migration patterns before a major holiday, whereas after the holidays the increased travel is in the same direction as the long term move.[13]

Using the daily coefficients of the regression on the net movement data to describe the relationship between short and long term movement not only tells us about the direction of the short term movement, but additionally highlights the timing of the movements. We observe differences in the length of the time period during which the holiday associated movements occur. For example, looking at Easter in Namibia (the first red vertical line), there is one distinct spike, meaning most people travel on the same day, whereas for Christmas in Namibia, travel is more spread out. Visually, we can see this spread of the travel volume over a longer period by the coefficient decreasing more steadily, from the beginning of December for the travel before Christmas and similarly gradually increasing after Christmas. In contrast, for Easter we observe a sudden spike in the coefficient just before and after the holiday. We also see a difference in travel synchronization when comparing travel before and after Christmas, with the return travel period being shorter. This highlights that individuals tend to travel back at a similar time, whereas travel before Christmas is less synchronized. Finally, in addition to the holiday patterns, we see very distinct weekly patterns of movement. These are especially pronounced in Namibia, where it seems there is a steady flow of individuals back and forth every week between places of origin and destination.

Several years of data are available for Namibia making it possible to study whether the observed patterns of holiday related increases in travel as well as differences in travel timing and synchronization for different holidays are consistent across different years. Figure 6 shows daily regression results of net short term movement on net long term movement for 2011, 2012, and 2013. We find that patterns of short term movement around holidays remain consistent from year to year when taking into account the slight change of timing for Easter.

3.2 Relationship Between Short and Long Term Movement for Migrants

In the previous subsection, we saw a strong relationship between long term movement in the census data and short term movement in the CDR data. We now show results for some processes that are likely to influence this relationship.

Looking at the short term movements made by individuals for which we observe a long term migration from the 3 years of mobile phone data available for Namibia, we find that about 26% of individuals migrate having both never visited the place of destination prior to moving and never visit their place of origin after migration.[14] The most common of the four possible cases is for individuals to make short term trips on the migration route both before and after migration, with 38% of individuals falling into this category.[15] Of individuals that only visited either before or after, visiting only after the move is more common with 21% of migrants choosing to do this, versus 15% of migrants visit only before the move. Therefore, we find that while short term movement after a long term migration slightly dominates, much of the relationship between short and long term movement can also be explained by short term movements prior to migration.

Figure 7 shows the percentage of migrants that do not make any short term visits, neither before nor after the move, by region of origin as well as destination. This highlights that most remote regions, such as 1 (in the most eastern part of the country) or 4 (southern most part of Namibia) are the places most likely to never see a visit (and least likely to see a visit both before and after as shown in Appendix Table A2). In contrast, central regions with several nearby regions, such as regions 3 and 11, have the lowest probability for no short term visits associated with a long term move.

We explicitly explore distance as an underlying factor for this finding, as intuitively it is much easier to visit between two regions that neighbor each other, compared to visiting between regions on opposite sides of the country. Remote regions such as 1 and 4 have fewer neighbors thus making it necessary to travel longer distances to any other regions

[13] We should note that at an annual level, there is no significant relationship between net short term movement and net long term movement, see appendix table A1. This is intuitive since the daily correlations go back and forth between negative and positive and average out to zero.

[14] An example of this type of behavior would be a wife whose husband has migrated prior, who then migrates to follow her husband with their children and no longer goes back to the place of origin since the family is now all together in the new destination.

[15] In the previous scenario, the husband who first migrated might have visited first to ensure that he can find work, migrates and then visits home to see his wife and children until they are able to migrate as well to the new destination.

in the country. Upon ranking pairs of regions based on number of visits between them, the top five with the lowest number of visits before and after each have region 4 as one of the regions in the pair. In contrast, the six pairs with the highest percentage of visits both before and after are 11 to 12, 6 to 3, and 11 to 10 (and their reciprocals), which are all neighboring regions.

Table 2 shows results in the form of odds ratios for the probability of visiting a particular region based on a logit model, controlling for region of origin, region of destination, month and year fixed effects. In the first two columns, we have the results for visiting a particularfregion that is the destination prior to migration. In columns 3 and 4, we have the results for visiting a particular region that is the origin after migration. In all cases, the region of comparison is region 13. The table shows that controlling for distance between pairs of regions has very significant effects on the likelihood of visiting. We see that when distance is not controlled for, it looks like there is a significant difference between visiting region 13 and visiting regions 2 and 4. Once distance is controlled for, the probability of visiting region 13 is actually not significantly different from visiting those two regions. The results also show that for some regions such as 1 and 5, the probability of visiting is much lower than region 13 until distance is controlled for, and then we see that once we take account of distance, people are more likely to visit those regions as compared to region 13. Many of these results are also present when looking at visits to the origin after a migration. Finally, the log likelihood on distance is less than one, signifying that the larger the distance between origin and destination, the lower the likelihood of a visit before and similarly for a visit after.

3.3 Relationship Between Social Networks and Visiting Patterns

Table 3 shows regressions exploring the relationship between short term movement, long term migration and social contacts. The first column shows the results from a regression of total short term movements in 2013 on total long term migration. These results are in line with the results presented in Column 1 of Table 1, though a larger R^2 indicates that it is possible to explain more of the variation in short term movement with the long term movements if we do not take into account directionality of movements.[16]

The second column of Table 3 shows total short term movements regressed on number of social contacts between regions x and y. The relationship between contacts and short term movement is positive and highly significant, with a slightly higher R^2 compared to regressing short term movement on long term migration, implying that social contacts potentially provide slightly more information in helping to explain short term movements.

The hypothesis that social contacts might provide additional information as compared to long term migration patterns is explored in Column 4, where we show the results from a regression of total contacts on long term movements. The coefficient on long term movement is large and highly significant, and more importantly, there is a very high R^2 of 0.68. This means that a very large portion of the variation in number of social contacts can be explained by long term migration. Nevertheless, there is still around 30% of the variation that cannot be explained by long term migration and instead is based on other factors that influence the formation of social networks. These non-migrant social networks can also influence the likelihood of making a short term trip.

Finally, in Column 3, short term movement is regressed on both long term migration and number of social contacts. Including both factors in the regression, they each become smaller in terms of the size of the coefficient, but both remain significant at the 0.01 level. Additionally, we see that the R^2 increases to be above 0.5, implying that including both long term migration and social network information provides additional information for explaining the variation in short term movement. This analysis shows that while a large portion of the social network between different regions can be explained by long term migration patterns, the social network provides additional information that can help in determining short term movement patterns because people have some social contacts in locations not determined by long term migration patterns.

The first part of this analysis focuses on social networks and movement aggregated at the region level. We now turn to results from the individual level analyses. Columns 1 and 2 of Table 4 show the results from the conditional logit models as odds ratios. In Column 1, we see that indeed both the number of contacts and the strength of contacts measured by number of calls and texts are significant in explaining the likelihood of visiting a region. They are also greater than 1, indicating that each additional call or text and contact lead to a higher likelihood of visiting. The results also show that there is an extra effect of having at least one contact, with a large coefficient of 5.655 on the dummy for having any contacts.

This shows that most important for a visit is to have at least one contact, with each additional contact increasing the likelihood a little bit. Including the squared terms in Column 2 shows that, as hypothesized, there are diminishing returns to contacts and calls since the coefficients are less than 1. This implies that as individuals have more and more contacts and make more calls or texts, the impact of the increase on the likelihood of visiting decreases.

The results of the duration analysis are shown in Columns 3 and 4 of Table 4. Each additional call or text to a place and each individual contact contribute to the person remaining for longer (0.03 of a day for each call and 0.36 for each social

[16] This result is reasonable since in the directional regression we regress short term movement from x to y on long term movement from x to y, but realistically, the long term move of an individual from x to y might cause both short term movement from x to y and from y to x, which does not get captured in the directional regression, but does get captured in the regression looking at total movements.

contact). Interestingly, when it comes to duration, there does not seem to be the large discrete jump observed for the probability of visiting. Instead, the coefficient on the dummy for at least one contact is not significantly different from the coefficient on total social contacts, so the additional effect of having at least one contact is equivalent to the effect of having one extra contact. However, when including the square terms as shown in Column 4, the coefficient on the dummy becomes negative, so once the diminishing returns of extra contacts are taken into account, there is no additional positive effect on the length of a stay of having at least one contact. Together, all of the results shown in this table point to a very strong non-linear relationship between the location of social contacts and both likelihood of visiting and the length of the visit.

4. Discussion

In this paper, we study the relationship between short term movement and long term migration patterns and dive into the factors that influence this relationship. In the first part of the analysis, we find a strong positive relationship between short term movement and long term migration which confirms previous descriptive research showing that people tend to travel back and forth between the place they originate from and their destination (e.g. Fall 1998). A factor not previously investigated is the consistency of the results across multiple countries and we find the travel patterns to be remarkably similar between the two countries studied, even though Senegal and Namibia differ on a number of factors, such as population size, density and GDP.[17] Results are not only consistent at the annual level, but also for daily regressions. The pattern of increased movement on migration routes around holidays is very consistent, despite the two countries having different holiday patterns due to different majority religions in the two countries (Senegal is majority Muslim while Namibia is majority Christian).

The first regressions we show in Table 1 have an R^2 of between 0.3 and 0.4 for the two countries. With long term migration in the census data as the only variable, we are able to explain around a third of the variation in short term movement. While this demonstrates the importance of long term migration patterns in determining short term movement, as discussed in the introduction, there are many other factors that affect short term movement in addition to long term migration. People might travel for tourism, for business or seasonal work, or for pilgrimages. These all represent different mechanisms that will impact the amount of short term movement. Appendix B provides a first exploration of the impact of pilgrimages on short term movement. While the aim of the appendix is to demonstrate why the two regions with large pilgrimages in Senegal are excluded from the analyses in this paper, it also provides some insight into the impact of large events like pilgrimages on short term movement patterns. Additional mechanisms of short term movement can be explored in future work.

Studying net movements allows us to investigate which pattern of movement is more common: individuals that have migrated traveling back to their place of origin, or contacts (such as family and friends) from their place of origin traveling to visit them. The net movement analysis does not refute the possibility that both types of travel are happening, but instead shows which type of movement is more common. The results show travel in the opposite direction of migration before the holiday and in the same direction after. This is in line with the anecdotal evidence that individuals that have migrated are likely to visit their home location for holidays and return to their new permanent place of residence after the holiday (Fall 1998).

Some additional patterns we observe are that travel is less synchronized for longer holidays when individuals have more time off (for Senegal the two Eids and for Namibia, Christmas), with individuals not necessarily traveling on the same date. This is especially the case for Namibia around Christmas where we see an extended period of movement in the opposite direction of long term movement before the holiday. In addition to increased travel on migration routes around Christmas we also see a weekly pattern, with people moving back and forth between the place of origin and destination on a weekly basis. This signifies that some individuals that have migrated long term choose to travel to their place of origin every weekend. This pattern is present in both countries, but is more marked in Namibia.

We consistently see Namibia having higher levels of short term movement per long term migrant, and there are several things that could explain this. First, as mentioned in the data section, due to the wording of the question in Namibia, it is possible that respondents misinterpreted the question leading to an underestimate of long term migration. This underestimate would mean that when running regressions, the coefficient on long term migration will be larger.

Additionally, in both Senegal and Namibia we are only able to look at short term movement for those individuals represented in the data. If the proportion of individuals represented is smaller in Senegal, then there will be a smaller number of short term moves captured, which would again lead to the difference in magnitude of the coefficient on long term migration we see between the two countries. In Senegal, there are around 9.5 million unique SIMs in 2013 and the population was around 13.8, leading to a ratio of 0.69. In Namibia, on the other hand, there are around 2.9 million SIMs in 2013 and the population was around 2.1 million, leading to a much larger ratio of 1.39 SIMs per person. This likely means that a larger proportion of the population is covered by the data in Namibia, however measuring this effect from

[17] The Namibia population was around 2.5 million, population density was around 3 persons per sq.km., and GDP per capita was $4,673.60 in 2015. In Senegal, the population was around 15.1 million, population density was 79 people per sq.km, and GDP per capita was $899.60 in 2015 according to The World Bank (2017).

the count of unique SIMs alone is not possible due to factors influencing the SIM to population ratio, such as multiple SIM ownership. Individuals frequently own more than one SIM card to make use of promotions or to transfer credit to other users (Stork 2011).[18]

Nevertheless, despite some differences in the magnitude of short term moves per long term migrant, we generally see very similar patterns between Senegal and Namibia. We also finnd that the pattern is very consistent from year to year, as highlighted by comparing results from different years for Namibia. These two findings are important from a policy maker's point of view. While CDR data are an excellent data source for measuring the timing and magnitude of short term movement, mobile phone data are not available to researchers and policy makers in every country context. Additionally, where this data is available, it generally only spans a limited time period. The analyses here show the possibility for policy makers to use information on short term movement from periods for which mobile phone data is available and to combine them with knowledge of main holidays in order to foresee the expected movement patterns that will be critical for many settings, ranging from infrastructure planning to controlling infectious diseases. Additionally, the analyses demonstrate the possibility of using the patterns seen in countries where CDR data is available to extrapolate to other country contexts where the CDR data is not available. In those contexts, if census data is available, it can be combined with information on public holidays to predict fluctuations in short term movement, and this information can be similarly applied to different policy questions.

For example, our first analysis looking at the relationship between short term movement in the CDR data on a daily basis and long term migration from the census data shows that there are very consistent patterns in the relationship, with short weekly spikes representing the weekly movement patterns and large spikes around holidays. Specifically, we are able to identify the periods of movement, which vary with each holiday, but are consistent across years. These periods are times at which large numbers of individuals travel on particular routes, linked to the long term migration routes. In studying issues such as spread of malaria due to travel in areas where the disease is at an elimination stage, this type of information can inform interventions such as mobile surveillance points on high traffic routes or increased malaria screening after individuals return home. To exemplify how this information can be used, we more formally identify these periods of large fluxes of movement between places of origin and destination by highlighting days for which the absolute value of the coefficient is larger than two standard deviations of the absolute daily coefficient for Senegal and larger than three standard deviations for Namibia.[19]

Figure 8 shows the graphs of the coefficients from the daily regressions of net short term movement on net migration (as in Figure 5), but we now highlight the regions of time when the coefficient is larger than two or three standard deviations. This marks the timing of travel, which can provide additional information to guide evidence-based strategies to fight malaria. While it is known that there is a lot of travel around holidays, this analysis shows the duration of the period of increased travel, which might be more difficult to determine from alternative data sources. We can see that for some holidays, the travel happens extremely quickly over the span of a weekend, while for other holidays, it is much longer. This can help policy makers in better targeting resources in order to facilitate decreasing the malariaburden at a lower cost or plan for potential congestion along highly traveled routes.

Looking at more specific aspects of the relationship between short and long term movement patterns provides additional insights that can be exploited by policy makers in crafting policies to address issues related to short term population movement. As might be expected, we saw that the farther origin and destination are, the less likely people are to visit both before and after a long term migration. This is intuitive since the farther the distance, the more difficult it becomes to visit in terms of monetary and time cost, and we are able to quantitatively confirm this. We also saw that social contacts play an important role in determining the location of short term visits and that contacts in other regions are not solely determined by long term migration patterns.

By modeling the relationship between short and long term movement and understanding some of the factors underlying this relationship, we can then consider how changes in long term migration might lead to new patterns of short term movement. This has implications for policies related to population movement, since while previous studies have used mobile phone data to classify sources and sinks of malaria, they do not necessarily provide information on how those sources and sinks might change if the movement patterns shift. Therefore, again, with only limited access to mobile phone data, it is possible to use this data more effectively in order to craft targeted policies that can account for changes in the short term movement even when real-time access to the mobile phone data to measure short term movement is not available.

Finally, in thinking about population movement and applications such as the spread of infectious disease or access to adequate infrastructure, not only does the movement matter, but the length of time that an individual spends in a location can matter as well. For the application of spread of malaria, the longer the person remains in an area, the larger the probability that the person might become infected with malaria (if they are in a high malaria area) and the higher the probability that they might infect mosquitoes that spread the disease if an infected individual enters a low malaria area.

[18] Transfer of credit is only possible between prepaid SIM cards.

[19] We use a larger cutoff for Namibia due to the larger weekly patterns of movement

The combination of the relationship between long term migration and social contacts and the relationship between social contacts and duration of a visit help to provide insights into how long individuals spend in different locations, which could be used to inform risk calculations or assessment of necessity of additional resources.

While CDRs provide extremely detailed information from calls and texts that can help us measure short term mobility, there are nevertheless important limitations. The first is that we can only measure movement for those individuals that use the mobile phone providers for which we have data access, and we have to make the simplifying assumption that those using other providers are not different in some fundamental ways and will have similar movement patterns. This might not necessarily be the case, especially if factors like socioeconomic status or ethnicity influence the choice of phone provider. Additionally, while in the paper we have interchangeably talked about individuals and SIMs, we only have data on SIMs and we do not know how many SIMs might correspond to a given individual. This is especially problematic for Namibia, where we see a ratio of SIMs to the population of 1.39. Nevertheless, multiple SIM ownership might not be a problem for measuring short term movement because of the fact that some users are only active for a short period of time[20], which suggests that there might be sequential multi-SIM ownership, with each individual using only one SIM card at a time but switching SIM cards frequently. For this analysis where we only need two subsequent days with data to observe a short term trip taken, this sequential multiple SIM ownership will not be a problem. Another limitation of the data is that we can only infer movement from people's location based on calls and texts and we have no information on locations at times when they are not using their phone. It is possible that phone usage could be correlated with movement, for example individuals always use their phone when they travel somewhere to call home and let family know they are safe, but they might not call once they are home since they see their family, and so we might miss some movements. If this type of correlation is systematic, it could lead to bias in the short term movements that is measured so that we miss some moves that happen between certain location and not others.

The long term migration data is also mainly helpful in considering short term movements linked to visiting social contacts, but does not make it possible to study additional processes influencing short term movement such as pilgrimages, which are excluded from the current paper. A limitation of the social network analysis is that given that for Senegal only one year of data is available, it is not possible to determine whether the contacts living in a different region migrated there previously or are originally from this area, or if the person is the one that migrated and the contacts are living in the place of origin. These are all important caveats that could influence our results and impact the interpretation of the results.

Using unique data for two different countries, we have exploited the long temporal span of the Namibia data and the detailed network structure of the Senegal data to study the processes behind the relationship between short and long term movement. In addition, this is one of the first papers to combine mobile phone analyses from two different country contexts within sub-Saharan Africa. While it is necessary to study the factors of movement separately due to the different strengths of the two datasets, we are able to conduct similar analyses using census data for both country settings. As more and more work is conducted using mobile phone data in particular contexts, it will become important to examine how much of the analysis and results can be extrapolated to other country settings and to future years.

[20] Table A.1 for Namibia shows that there are many SIMs only active for a very short period of time

5. References

Adriansen, Hanne Kirstine (2008). "Understanding pastoral mobility: the case of Senegalese Fulani". *The Geographical Journal* 174.3, pp. 207-222.

Ahmed, Qanta A, Yaseen M Arabi, and Ziad A Memish (2006). "Health risks at the Hajj". *The Lancet* 367.9515, pp. 1008-1015.

Alkhalife, Ibrahim S (2003). "Imported malaria infections diagnosed at the Malaria Referral Laboratory in Riyadh, Saudi Arabia". *Saudi medical journal* 24.10, pp. 1068-1072.

ARTP, Autorité e Régulation des Télécommunications et des Postes (2013). *Observatoire de la telephonie mobile Tableau de bord au 31 décembre 2013*. url: http : / / www . artpsenegal.net/images/TB_mobile_dec-13.pdf.

Atamanov, Aziz and Marrit van den Berg (2012). "International labour migration and local rural activities in the Kyrgyz Republic: determinants and trade-os". *Central Asian Survey* 31.2, pp. 119-136.

Balcan, Duygu et al. (2009). "Multiscale mobility networks and the spatial spreading of infectious diseases". *Proceedings of the National Academy of Sciences* 106.51, pp. 21484-21489.

Bengtsson, Linus et al. (2011). "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti". PLoS Med 8.8, e1001083.

Blumenstock, Joshua E (2012). "Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda". *Information Technology for Development* 18.2, pp. 107-125.

Doocy, Shannon et al. (2015). "Internal displacement and the Syrian crisis: an analysis of trends from 2011-2014". *Conflict and health* 9.1, p. 33.

Erbach-Schoenberg, Elisabeth zu et al. (2016). "Dynamic denominators: the impact of seasonally varying population numbers on disease incidence estimates". *Population health metrics* 14.1, p. 35.

Etzold, Benjamin et al. (2014). "Clouds gather in the sky, but no rain falls. Vulnerability to rainfall variability and food insecurity in Northern Bangladesh and its eects on migration". *Climate and Development* 6.1, pp. 18-27.

Fall, Abdou Salam (1998). "Migrants' long-distance relationships and social networks in Dakar". *Environment and Urbanization* 10.1, pp. 135-146.

Fall, Papa Demba, María Hernández Carretero, and Mame Yassine Sarr (2010). *Country and Research Areas Report*.

Goldsmith, Peter D, Kisan Gunjal, and Barnabe Ndarishikanye (2004). "Ruralurban migration and agricultural productivity: the case of Senegal". *Agricultural economics* 31.1,pp. 33-45.

Henderson, J Vernon, Adam Storeygard, and Uwe Deichmann (2014). "50 years of urbanization in Africa: Examining the role of climate change". *World Bank Policy Research Working Paper 6925*.

Herrera, Catalina and David E Sahn (2013). "Determinants of internal migration among Senegalese youth". *Cornell Food and Nutrition Policy Program Working Paper 245*.

Huang, Zhuojie, Andrew J Tatem, et al. (2013). "Global malaria connectivity through air travel". *Malar* J 12.1, p. 269.

Linares, Olga F (2003). "Going to the city...and coming back? Turnaround migration among the Jola of Senegal". *Africa* 73.01, pp. 113-132.

Littrell, Megan et al. (2013). "Case investigation and reactive case detection for malaria elimination in northern Senegal". *Malar* J 12, p. 331.

Lynch, Caroline A et al. (2015). "Association between recent internal travel and malaria in Ugandan highland and highland fringe areas". *Tropical Medicine & International Health* 20.6, pp. 773-780.

Mokashi, Digambar Balkrishna (1987). *Palkhi: an Indian pilgrimage*. SUNY Press.

Namibia Statistics Agency (2015). *Namibia 2011 Census Migration Report*.

Osorio, Lyda, Jim Todd, and David J Bradley (2004). "Travel histories as risk factors in the analysis of urban malaria in Colombia". *The American journal of tropical medicine and hygiene* 71.4, pp. 380-386.

Oster, Emily (2012). "Routes of Infection: Exports and HIV Incidence in Sub-Saharan Africa". *Journal of the European Economic Association* 10.5, pp. 1025-1058.

Pison, Gilles et al. (1993). "Seasonal migration: a risk factor for HIV infection in rural Senegal". *JAIDS Journal of Acquired Immune Deciency Syndromes* 6.2, pp. 196200.

Poncelet, Alice et al. (2010). "A country made for disasters: environmental vulnerability and forced migration in Bangladesh". *Environment, forced migration and social vulnerability*. Springer, pp. 211-222.

Prothero, R Mansell (1977). "Disease and mobility: a neglected factor in epidemiology". *International Journal of Epidemiology* 6.3, pp. 259-267.

Rain, David (1999). "Eaters of the dry season: Circular labor migration in the West African Sahel".

Rinschede, Gisbert (1992). "Forms of religious tourism". *Annals of tourism Research* 19.1, pp. 51-67.

Siri, Jose G et al. (2010). "Signicance of travel to rural areas as a risk factor for malarial anemia in an urban setting". *The American journal of tropical medicine and hygiene* 82.3, pp. 391397.

Stork, Christoph (2011). *Namibian Sector Performance 2011*. url: http://www.researchictafrica. net/publications/Evidence_for_ICT_Policy_Action/Stork_C_-_2011_Namibian_ Sector_Performance_Review.pdf (visited on 05/15/2017).

Stuckler, David et al. (2011). "Mining and risk of tuberculosis in sub-Saharan Africa". *American journal of public health* 101.3, pp. 524-530.

Tatem, Andrew J and David L Smith (2010). "International population movements and regional Plasmodium falciparum malaria elimination strategies". *Proceedings of the National Academy of Sciences* 107.27, pp. 12222-12227.

Tatem, Andrew J et al. (2014). "Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning". *Malaria journal* 13.1, p. 52.

The World Bank (2017). *World Development Indicators*. http://data.worldbank.org/datacatalog/world-development-indicators. [Accessed 31 May 2017].

Wesolowski, Amy et al. (2012). "Quantifying the impact of human mobility on malaria". *Science* 338.6104, pp. 267-270.

Wesolowski, Amy et al. (2013). "The use of census migration data to approximate human movement patterns across temporal scales". *PloS one* 8.1, e52971.

Wesolowski, Amy et al. (2015a). "Impact of human mobility on the emergence of dengue epidemics in Pakistan". *Proceedings of the National Academy of Sciences* 112.38, pp. 11887-11892.

Wesolowski, Amy et al. (2015b). "Quantifying seasonal population uxes driving rubella transmission dynamics using mobile phone data". *Proceedings of the National Academy of Sciences* 112.35, pp. 11114-11119.

WHO (2010). "Water woes in Senegal's holy city". *Bulletin of the World Health Organization* 88.1. http://www.who.int/bulletin/volumes/88/1/10-020110/en/.

Wilson, Robin et al. (2016). "Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal Earthquake". *PLOS Currents Disasters*.

Yukich, Joshua O et al. (2013). "Travel history and malaria infection risk in a low-transmission setting in Ethiopia: a case control study". *Malar* J 12, p. 33.

Figure 1: Timing of the Last Call of the Day in Namibia.

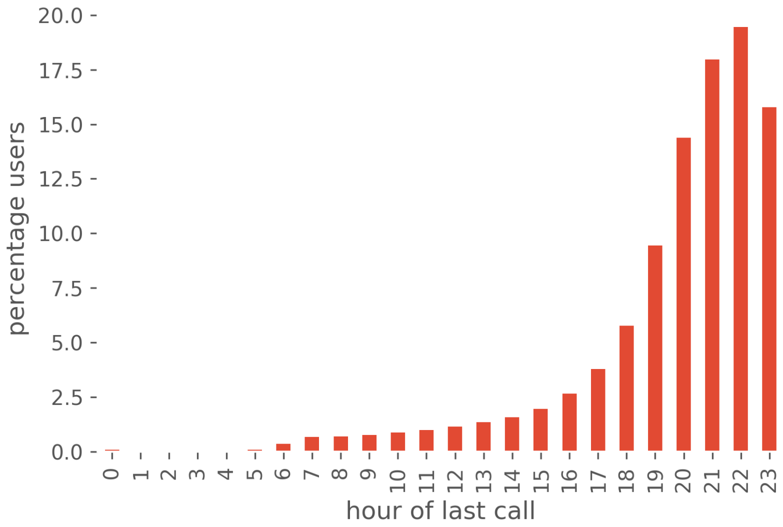
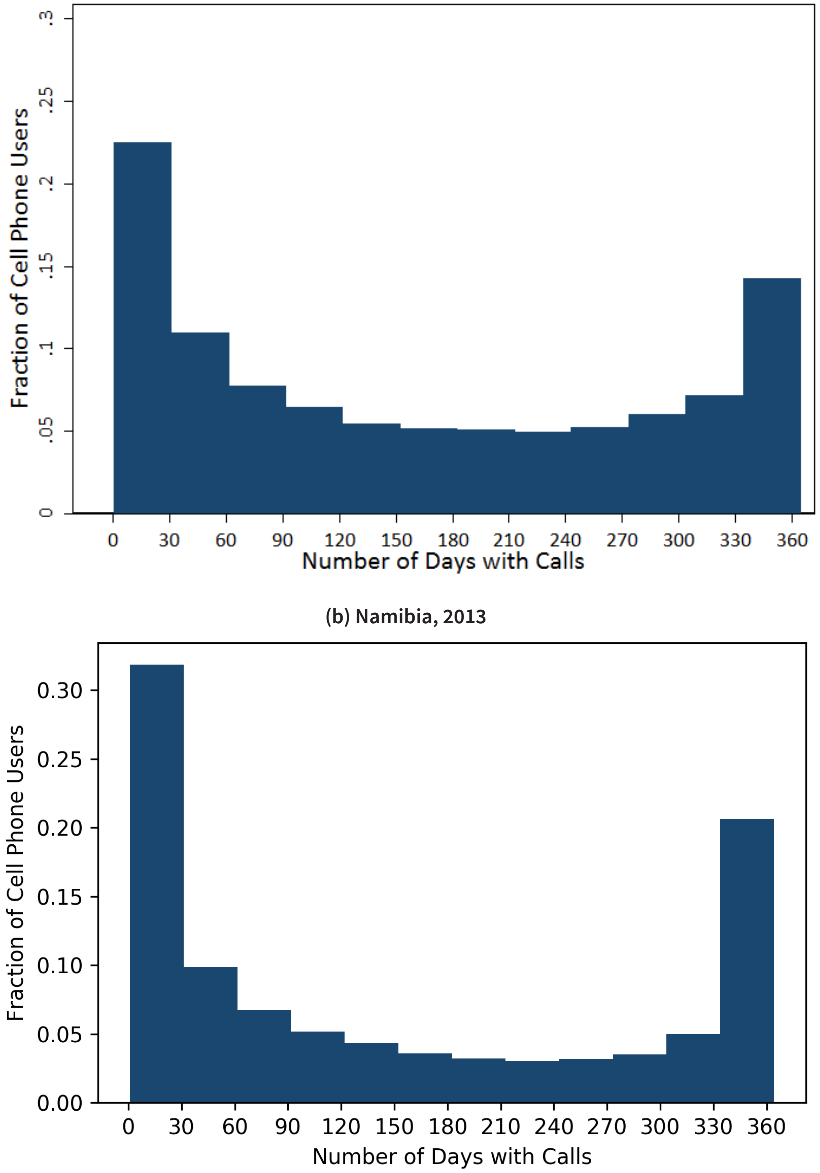
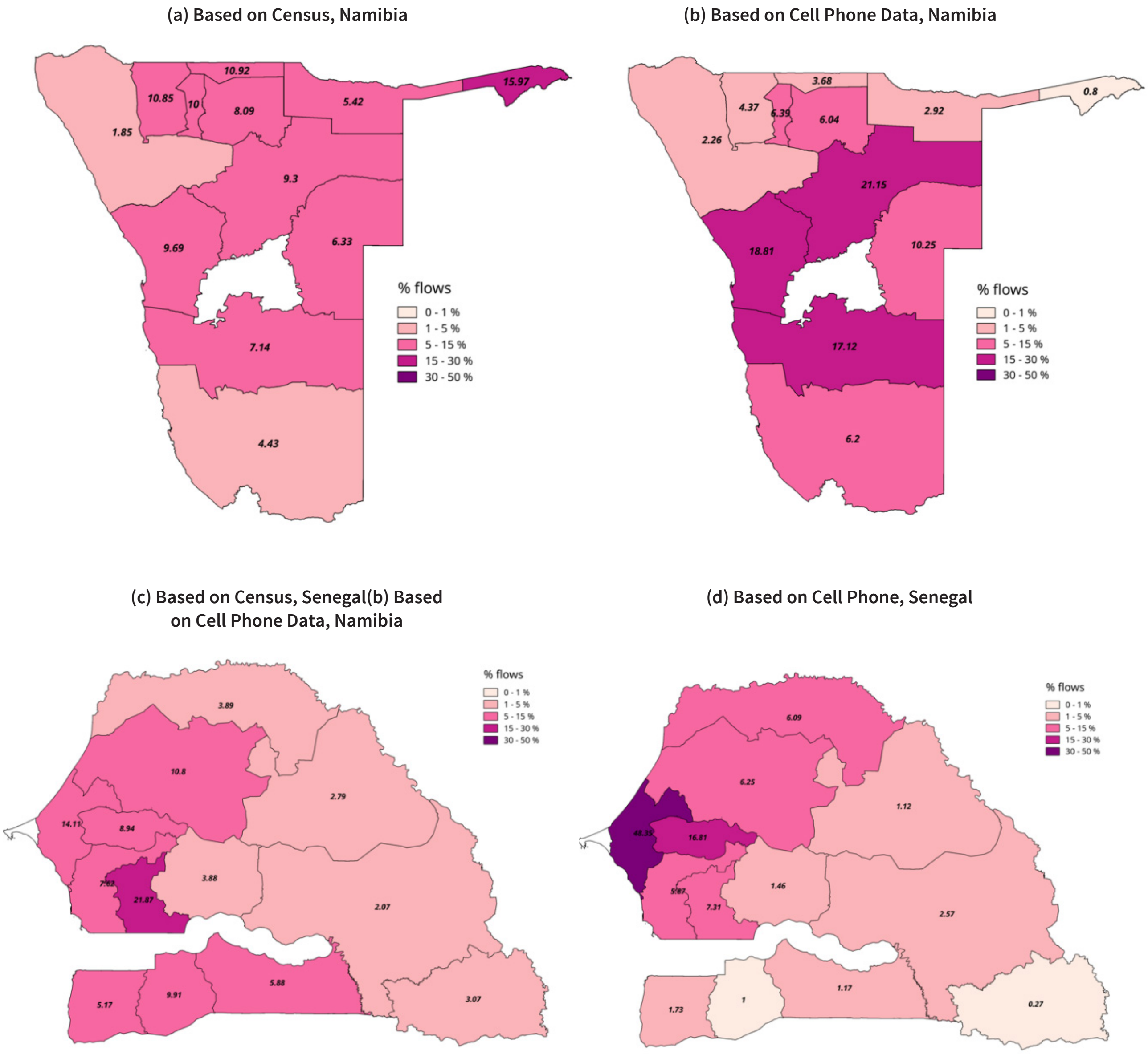


Figure 2: Histogram of Number of Days with Cell Phone Usage



(b) **Namibia, 2013**

Figure 3: Movement into the Two Capital Regions



(a) Based on Census, Namibia

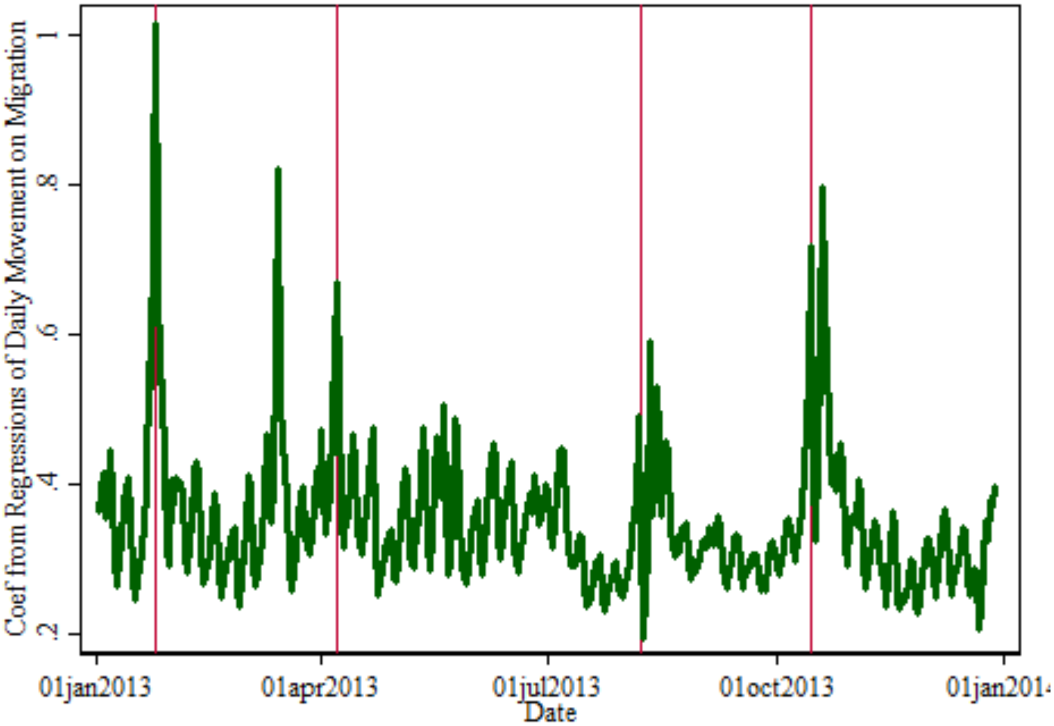
(b) Based on Cell Phone Data, Namibia

(c) Based on Census, Senegal(b) Based on Cell Phone Data, Namibia

(d) Based on Cell Phone, Senegal

Figure 4: Short Term vs Long Term Movement

(a) Senegal, 2013



(b) Namibia, 2013

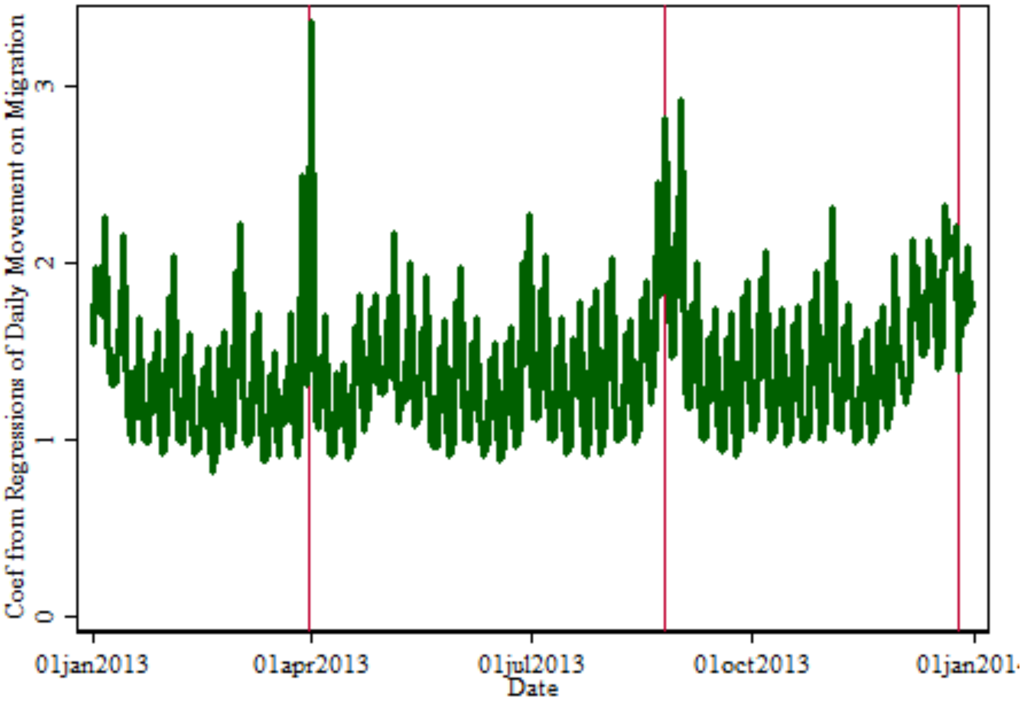
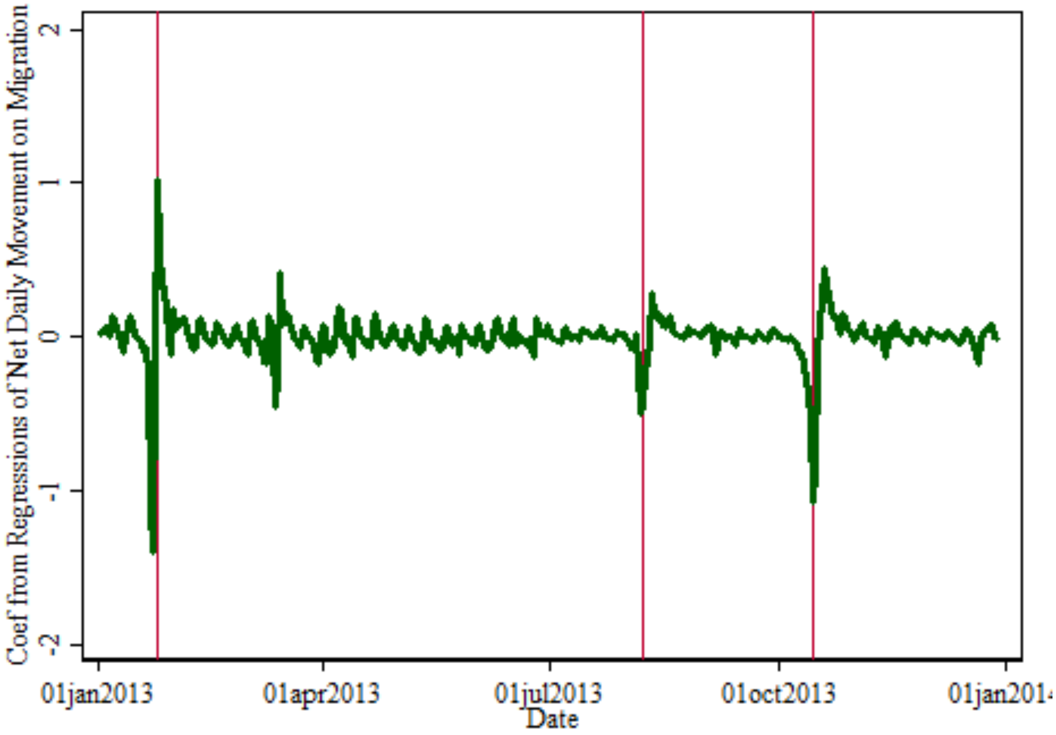


Figure 5: Net Short Term vs Long Term Movement

(a) Senegal, 2013



(b) Namibia, 2013

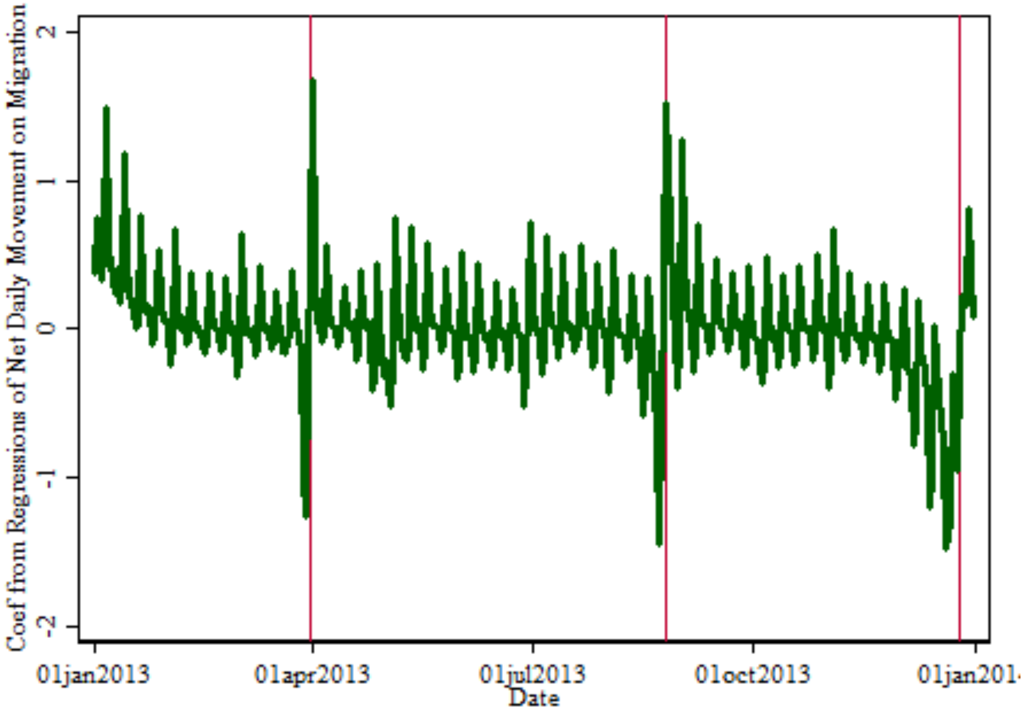
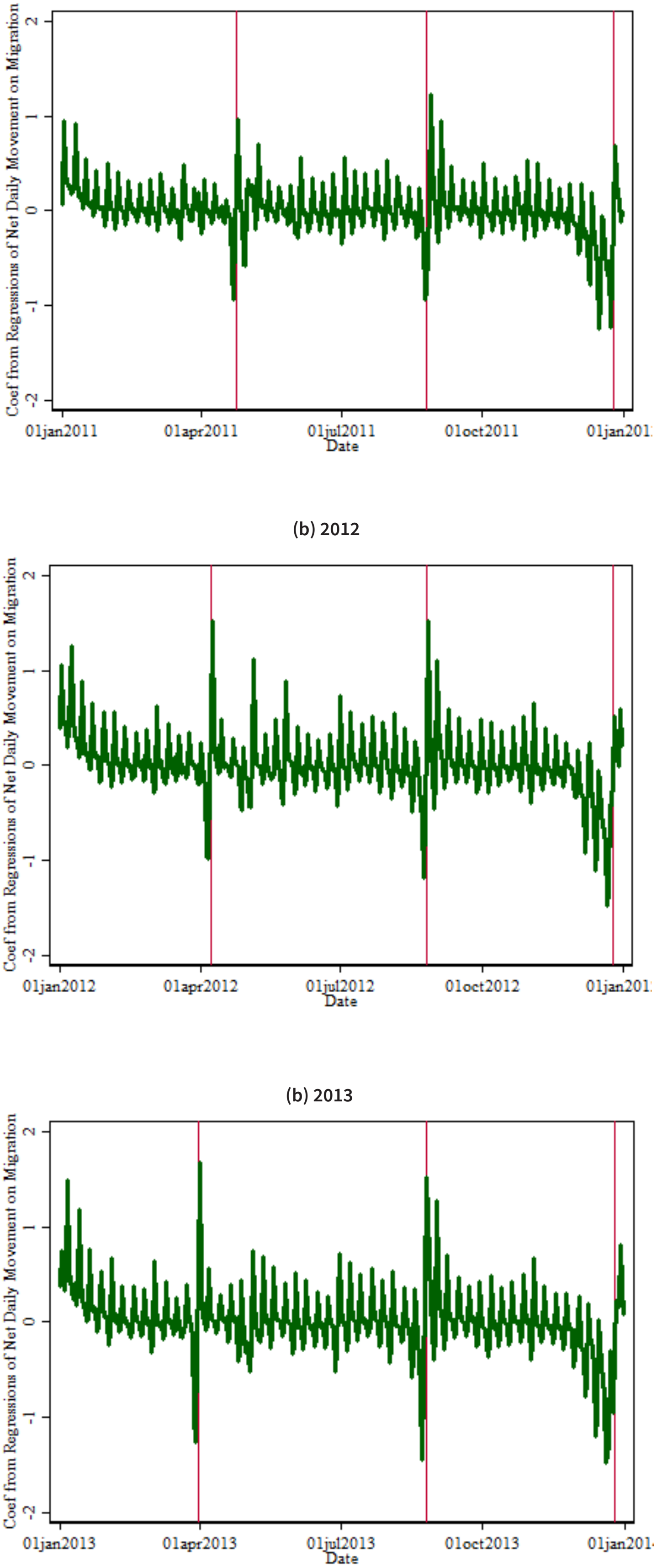


Figure 6: Net Short Term vs Net Long Term Movement for Namibia

(a) 2011



(b) 2012

(b) 2013

Figure 7: Percentage never visiting by admin unit

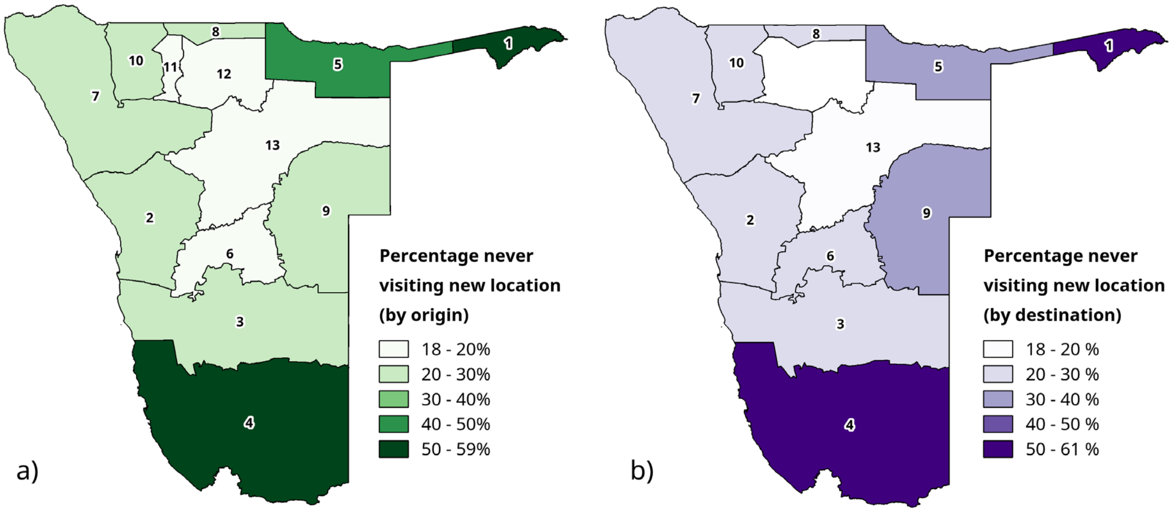
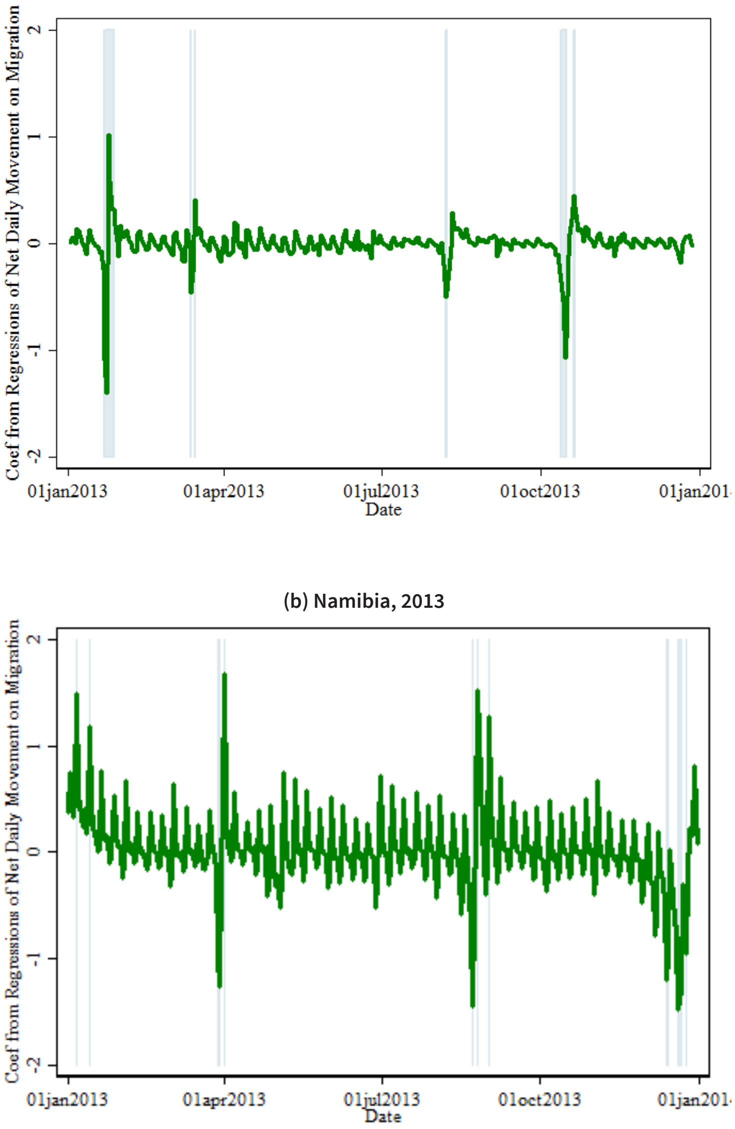


Figure 8: Timing for Targeting Travelers

(a) Senegal, 2013



(b) Namibia, 2013

Table 1: Annual Regressions of Short term on Long Term, Senegal and Namibia

	(1) Senegal	(2) Namibia
Migrants entering	127.3*** (17.22)	505.5*** (164.8)
Constant	57,336*** (17,715)	-8,044 (25,175)
Observations	132	156
R-squared	0.328	0.394

Robust standard errors in parentheses, two-way clustered
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Logit Model of Probability of Visiting a Given Region Before and After Migration to It

	(1) Visit Before	(2) Visit Before with Distance	(3) Visit After	(4) Visit After with Distance
Distance		0.996*** (0.000242)		0.996*** (0.000319)
Region 1	0.285*** (0.136)	1.701** (0.453)	0.227*** (0.0144)	1.527*** (0.229)
Region 2	0.620*** (0.0676)	0.804 (0.170)	0.606*** (0.0142)	0.792*** (0.0284)
Region 3	0.861 (0.275)	1.019 (0.136)	0.794** (0.0926)	0.872*** (0.0255)
Region 4	0.223*** (0.0515)	1.059 (0.316)	0.226*** (0.00666)	1.068 (0.126)
Region 5	0.658* (0.166)	1.342** (0.191)	0.576*** (0.0665)	1.268*** (0.0694)
Region 6	1.006 (0.200)	1.570** (0.282)	1.146** (0.0731)	1.740*** (0.0860)
Region 7	0.644 (0.180)	1.296 (0.283)	0.709*** (0.0274)	1.664*** (0.142)
Region 8	0.932 (0.454)	1.186 (0.259)	1.037 (0.0441)	1.580*** (0.0699)
Region 9	0.526* (0.174)	0.589** (0.135)	0.566*** (0.0563)	0.608*** (0.0181)
Region 10	0.787 (0.368)	1.045 (0.271)	0.984 (0.0547)	1.460*** (0.0630)
Region 11	1.674 (0.863)	1.372 (0.363)	1.690*** (0.148)	1.617*** (0.0200)
Region 12	1.223 (0.448)	1.144 (0.216)	1.248*** (0.0700)	1.319*** (0.0400)
Month FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	139,103	139,103	139,103	139,103

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Annual Regressions of Short term on Long Term Movement and Social Network

	(1) Short Term	(2) Short Term	(3) Short Term	(4) Contacts
Long Term Migration	173.7*** (26.51)		86.46*** (27.25)	1,018*** (133.5)
Person Contacts		0.144*** (0.0239)	0.0857*** (0.0251)	
Constant	11,617 (48,058)	141,985** (52,676)	52,983 (39,450)	-482,603*** (147,952)
Observations	66	66	66	66
R-squared	0.449	0.466	0.502	0.681
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1				

Table 4: Individual Level Relationship Between Short Term Movement and the Social Network

	(1) Probability of Visiting	(2) Probability of Visiting	(3) Interpolated Time in Region	(4) Interpolated Time in Region
Total Calls to Region	1.002*** (0.000261)	1.004*** (0.000278)	0.0267*** (0.00273)	0.0256*** (0.00249)
Total Calls to Region2		0.999998*** (1.63e-07)		-4.25e-06*** (1.04e-06)
Total Social Contacts in Region	1.144*** (0.00379)	1.144*** (0.00374)	0.355*** (0.0366)	0.536*** (0.0273)
Total Social Contacts in Region2		0.99986*** (7.24e-06)		-0.000709*** (0.000103)
Dummy at Least 1 Contact	5.655*** (0.150)	5.533*** (0.147)	0.380*** (0.0841)	-0.218*** (0.0554)
Constant			-0.323 (0.355)	-2.009*** (0.249)
Observations	212,520	212,520	451,902	451,902
R-squared			0.246	0.265
Region FE	YES	YES	YES	YES
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1				

A. Additional Tables and Figures

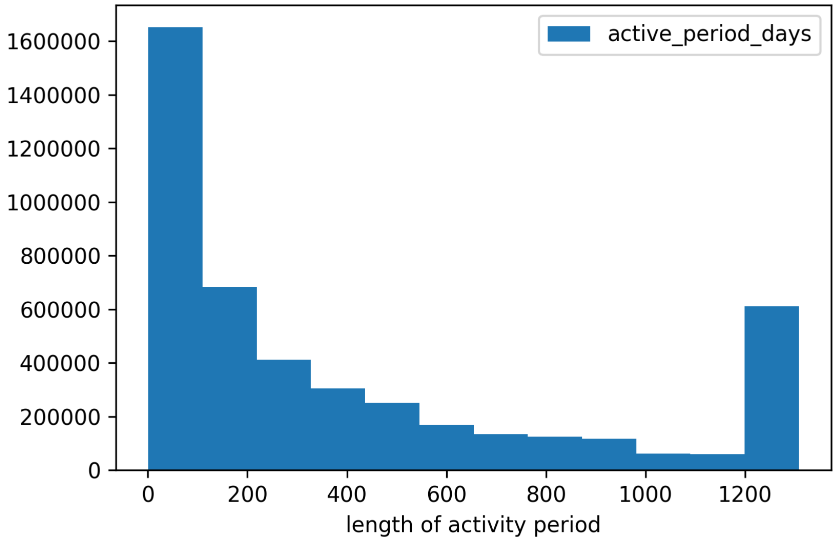
Table A1: Annual Regressions of Net Short term on Net Long Term, Senegal and Namibia

	(1) Senegal	(2) Namibia
Net Migrants	-1.079 (1.635)	-2.097 (2.289)
Constant	-1,194 (1,160)	616.9** (244.8)
Observations	66	78
R-squared	0.010	0.022
Robust standard errors in parentheses, two-way clustered *** p<0.01, ** p<0.05, * p<0.1		

Table A2: Percent of Migrants Making a Trip both Before and After or Never Based onOrigin and Based on Destination

Region	Origin		Destination	
	(1) Visit Before and After	(2) Never Visit	(3) Visit Before and After	(4) Never Visit
1	12.6	58.9	14.0	55.9
2	36.3	26.1	31.5	27.5
3	49.8	24.2	48.9	25.4
4	17.5	56.6	14.4	60.7
5	19.8	40.4	23.9	35.3
6	44.9	18.4	38.1	22.9
7	38.5	25.5	37.5	23.1
8	34.3	26.9	40.8	24.0
9	39.5	29.0	37.8	33.0
10	31.8	29.7	36.7	27.3
11	49.1	18.2	54.2	14.9
12	41.7	18.9	46.4	16.9
13	44.2	19.0	43.4	18.1

Figure A.1: Length of Active Period over 3 Years for Namibia



B. Relationship Between Short Term Movement and Migration when Pilgrimage Regions Included in Senegal

As mentioned in the text, there are two important pilgrimage sites in Senegal. The first is Tivaouane in the Thies region and the second is Touba in the Diourbel region. The pilgrimage site that draws the most people is Touba, where for the Grand Magal over a million people come to the city. In 2013, this Magal occurred twice, on January 1 and on December 21. Additionally, there is a smaller pilgrimage that occurred in June, 2013. In Tivaouane, there is a pilgrimage for the Prophet's birthday, Maouloud, on January 24.

These pilgrimages represent a mechanism driving short term movement that is different from the one discussed throughout this paper. In the paper, we focus on long term migration patterns as a driver of short term migration, whereas in the case of pilgrimages the underlying driver is an important site that draws large numbers of people for specific events. Therefore, if we were to include the regions of these pilgrimages, especially since they draw such large numbers of people, it causes a distortion in the relationship between short term movement and long term migration patterns measured in the census

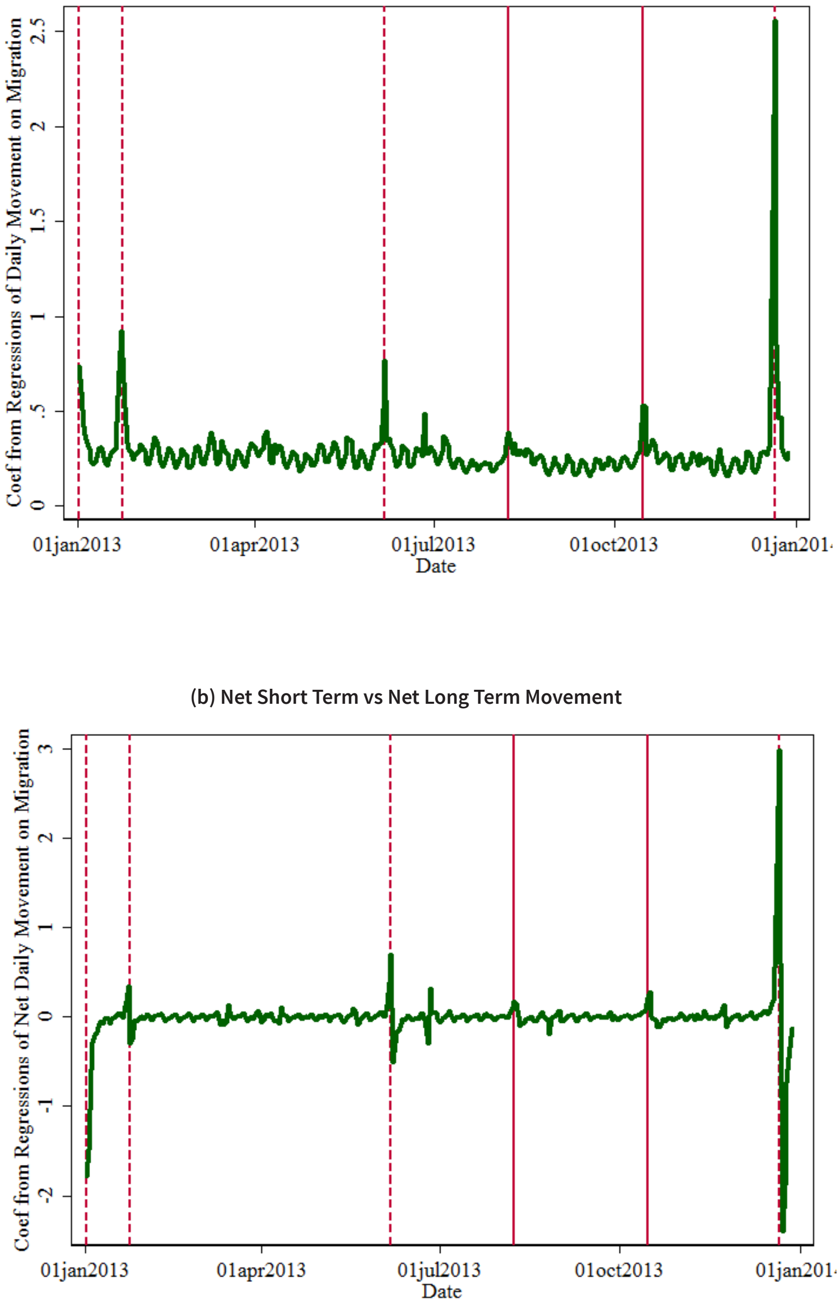
Figure B.1 shows results of daily regressions when all regions of Senegal are included in the analysis, including the two regions that receive large pilgrimages. In these graphs, in addition to solid red lines marking the most important Muslim holidays, dashed red lines mark the various large pilgrimages that occur. In the top panel we see that rather than some of the biggest holidays jumping out as large peaks in the correlation, we are seeing the largest peak as the Grand Magal in December. This is because there is some migration in the census data to the region where Touba is located; therefore, when running the regression for the day of the pilgrimage we see a very strong positive relationship as a large number of people enter the region. Yet, this does not represent a pattern of migrants that are visiting their place of origin or returning after a visit, but instead represents lots of pilgrims entering the region from places where there is also some migration.

The difference in the mechanism driving the short term movement becomes even clearer in the bottom panel. Rather than the expected relationship we saw between short and long term net movement in the paper, with individuals going in the opposite direction from where they migrated before a holiday (large negative coefficient) and then going in the same direction as the migration after the holiday (large positive coefficient), we see that for the pilgrimages there is a large positive coefficient on the day of the pilgrimage and a large negative coefficient on the day after the pilgrimage. This is because the typical migration pattern is for individuals to migrate into these areas.[21] Therefore, when regressing net short term movement on long term migration, we get a large positive coefficient when a pilgrimage occurs and people enter the area and a large negative coefficient as they exit the following day. When including the two pilgrimage sites, it masks the relationship we are trying to capture, which is long term migration and the expansion of social networks through long term migration leading to short term moves between two regions. This is especially true for the Prophet's birthday, which is not only a day when an important pilgrimage occurs to Tivaouane, but is also a public holiday when there is generally no work or school. This time off leads to lots of short term movement along migration routes, which we are able to capture in the main figures of the paper when we exclude the two regions with large pilgrimages.

The large movements that occur when there are pilgrimages are still important and can have important consequences for spread of disease, congestion, and economic activity, but they represent a different mechanism from the one studied in this paper. Additional work can explore this mechanism and potentially how to predict the size of the influx of pilgrims as well as the locations where pilgrims are likely to come from

Figure B.1: Senegal Results with Pilgrimage Regions

(a) Short Term vs Long Term Movement



(b) Net Short Term vs Net Long Term Movement

[21] Based on conversations with census officials in Senegal, this is especially true for Touba in recent years since the city offers many social services, which can lead individuals to migrate there during economic downturns in the country. Additionally, the city has a policy of free land and free water, which also has led some to migrate there (WHO 2010).

Papiers de Recherche de l'AFD

Les *Papiers de Recherche de l'AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ilss'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l'AFD n'en exclut aucune autre.

L'Agence Française de Développement (AFD), institution financière publique qui met en oeuvre la politique définie par le gouvernement français, agit pour combattre la pauvreté et favoriser le développement durable. Présente sur quatre continents à travers un réseau de 72 bureaux, l'AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète. En 2014, l'AFD a consacré 8,1 milliards d'euros au financement de projets dans les pays en développement et en faveur des Outre-mer.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s).

Les *Papiers de Recherche* sont téléchargeables sur : http://librairie.afd.fr/

AFD Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

Agence Française de Développement (AFD), a public financial institution that implements the policy

defined by the French Government, works to combat poverty and promote sustainable development. AFD operates on four continents via a network of 72 offices and finances and supports projects that improve living conditions for populations, boost economic growth and protect the planet. In 2014, AFD earmarked EUR 8.1bn to finance projects in developing countries and for overseas France.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

AFD Research Papers can be downloaded from: http://librairie.afd.fr/en/

AFD, 5 rue Roland Barthes
75598 Paris Cedex 12, France
ResearchPapers@afd.fr
ISSN 2492 - 2846

BIG DATA

TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018

BIG DATA
TO ADDRESS GLOBAL
DEVELOPMENT CHALLENGES
2018

DATA-POP
ALLIANCE

THE IMPACT OF CRIME SHOCKS ACROSS GENDER AND SOCIOECONOMIC GROUPS: A LARGE-SCALE MAPPING OF BEHAVIORAL DISRUPTION

Rodrigo Lara Molina* - Data-Pop Alliance, Media Lab-HHI-ODI.
Alejandro Noriega* - MIT Media Lab.
Eaman Jahani* - Institute for Data, Systems, and Society; MIT.
Julie Ricard - Data-Pop Alliance, Media Lab-HHI-ODI.
Alex Pentland - MIT Media Lab, Data-Pop Alliance, Media Lab-HHI-ODI, Institute for Data, Systems, and Society; MIT.
*Authors contributed equally to this work.

DATA-POP ALLIANCE

mit media lab

Funded by:
AFD AGENCE FRANÇAISE DE DÉVELOPPEMENT

Additional support from:
MIT MANAGEMENT LATIN AMERICA OFFICE

THE IMPACT OF CRIME SHOCKS ACROSS GENDER AND SOCIOECONOMIC GROUPS: A LARGE-SCALE MAPPING OF BEHAVIORAL DISRUPTION

Rodrigo Lara Molina*
Data-Pop Alliance, Media Lab-HHI-ODI.

Alejandro Noriega*
MIT Media Lab.

Eaman Jahani*
Institute for Data, Systems, and Society; MIT.

Julie Ricard
Data-Pop Alliance, Media Lab-HHI-ODI.

Alex Pentland
MIT Media Lab, Data-Pop Alliance, Media Lab-HHI-ODI, Institute for Data, Systems, and Society; MIT.

*Authors contributed equally to this work.

In recent decades the world has seen a simultaneous trend towards becoming more peaceful overall, but also towards higher homicide rates surging in focal regions in the developing world. Although abundant research exists on the nature and sociology of crime, few studies look into the damaging impact of crime and violence on the daily lives of affected communities. The present study proposes the use of societal-scale behavioral data—card transactions' metadata—to elicit such impact. On the crime side, we use detailed homicide records for an entire middle-income country to identify salient crime shocks at the local level. On the behavioral side, we use debit card transaction volumes throughout the country to extract behavioral indices. We show that crime shocks have a substantial effect on communities' consumption patterns. Moreover, we show that the effects of crime shocks distribute differently across population subgroups defined by gender and socioeconomic status— e.g., with reductions of up to 7% in females' average volume of transactions—potentially exacerbating social inequalities. We conclude this work with policy recommendations on the use of 'big data' sources to monitor and help.

TABLE OF CONTENTS

1. Introduction

1.1 Crime and Violence

In recent decades the world has seen a trend towards becoming more peaceful overall (*1*). Simultaneously, higher civilian murder rates are still surging in focal regions in the developing world (while they are decreasing in rich countries): in 2016, 68 percent of violent deaths worldwide were murders, which is more than 3 times the number of deaths caused by war (*2*). Homicide rates are often used as a proxy for the overall regional levels of violence, among other reasons because they are the most systematically recorded crime.

Increasing violence in developing countries can be attributed to a multiplicity of factors. In some cases, it is strongly associated to factors such as organized crime, trafficking, and gang activity (*3*). Moreover, crime is explained to varying degrees by a broad range of structural and societal circumstances, such as demographic and socioeconomic factors (*4*, *5*), access to education (*6*) and weakness of institutions (*5*, *7*). In addition, more recent place-centric approaches have emphasized the impact of physical characteristics of cities on crime, such as the proportion of uninhabited homes (*8*).

Excluding war zones, Latin America and the Caribbean (LAC), is considered the most violent region in the world, where homicide rates registered in 2015 were four times higher than the world average (*3*). In fact, levels of violence in LAC are so disproportionately high, that they are considered at epidemic levels by the World Health Organization. Mexico ranks among the most violent countries in the region, where homicide rates have surged to unprecedented levels since 2007 (*9*), when president Felipe Calderon began a frontal assault on drug-trafficking organizations (*10*). Recently, in 2017, violence levels broke historic records, registering more than 29,000 homicides (INEGI, 2017).

The present work focuses on the effects that prevalent violence may have on the daily lives of local communities. Although the majority of violence is credited to conflict among criminal organizations, or among them and government forces, a recent study published by Open Society Foundations reports numerous atrocity crimes perpetrated against the civilian population since 2006, including killings, torture, and disappearances (*11*). Consequently, violence and insecurity have become the number one concern of Mexicans, above unemployment, corruption, and poverty—with 70% of Mexicans declaring it their top concern, according to the 2015 National Survey on the Quality and Impact of Government (ENCIG 2015).

1.2 Impact of Crime on Citizens' Lives

Whether originated by war or waves of crime, violence has pervasive impacts on communities and the development of cities and countries. Recent studies in Mexico have shown the impact of violence at a societal level, such as deteriorating effect on democratic institutions (*12*), economic costs (*9*, *13–15*) and urban transformation (*16*).

Moreover, violence creates fear and uncertainty, affecting not only the primary victim but also his/her family and community, also known as secondary victims (*17*). Only a few studies worldwide have explored the effects of crime and violence on daily activities of both primary and secondary victims, even the less focusing on LAC countries. A few have revealed the impact of victimization or fear of victimization on risky routine activities (*18*), on nighttime activities (*19*) and on the emotional and psychological health of individuals (*20*).

In Mexico, the National Victimization and Perception of Public Security Survey (ENVIPE), attempts to measure behavioral changes in correlation with perception of security. According to ENVIPE, in 2017 the three daily activities most affected by fear of crime are 'Letting children go out alone', 'Use jewelry' and 'Go out at night'. Further studies based on ENVIPE results, underlined gender differences in both perceptions of security and changes in behaviors in a high-crime environment (*21*, *22*).

However, to our knowledge, the disruption of daily activities such as mobility or consumption, and its differentiated effects on population subgroups, as defined by gender and socioeconomic status, has never been quantified or systematically evaluated.

1.3 Private Sector's Big Data for Public Good

Big data are defined as the digital traces (or digital breadcrumbs) passively emitted through the use of digital devices, such as call records, credit card transactions, GPS locations, etc. (*23*). Given its ubiquitous nature, crumbs can yield accurate information about human behavior (e.g., purchasing behavior or physical mobility and communications) at unprecedented levels of spatiotemporal granularity. Derived insights have been used to diagnose a diverse range of situations, opening to applications such as poverty mapping (*24*), infrastructure (*25*) and transportation planning (*26*). In the field of citizen security and crime prevention, researchers were able to predict crime ”hotspots” in cities such as London, Boston, and Bogota using aggregated human behavioral data captured from the mobile network infrastructure, together with basic demographic information (*27*). To date, these approaches have not looked into the impact of crime on the daily behavior of individuals.

This paper. In this paper, we map individuals' behavioral disruption in the face of urban crime shocks, by using societal-scale debit card transaction records. Section 2 describes the crime and behavioral datasets and how indices were computed. Section 3 summarizes the main results and Section 4 discusses results and future work.

2. Data and Methods

The data and methods used in this paper consist of two types: on the one hand crime data and the elicitation of crime shocks, and on the other hand consumption data and the construction of behavioral indices.

2.1 Crime Shocks

Crime data. We collected detailed crime records from the National Public Safety System of Mexico. More specifically, we computed monthly homicide rates—homicides for each 10,000 individuals—for each of the +2.4 thousand municipalities in the country. The rationale for focusing on homicides is threefold: their high prevalence in the country of focus, their high impact on citizen's perception of safety, and the fact that homicide statistics are among the most reliable among several crime types (e.g., as opposed to rape, which is often under-reported).

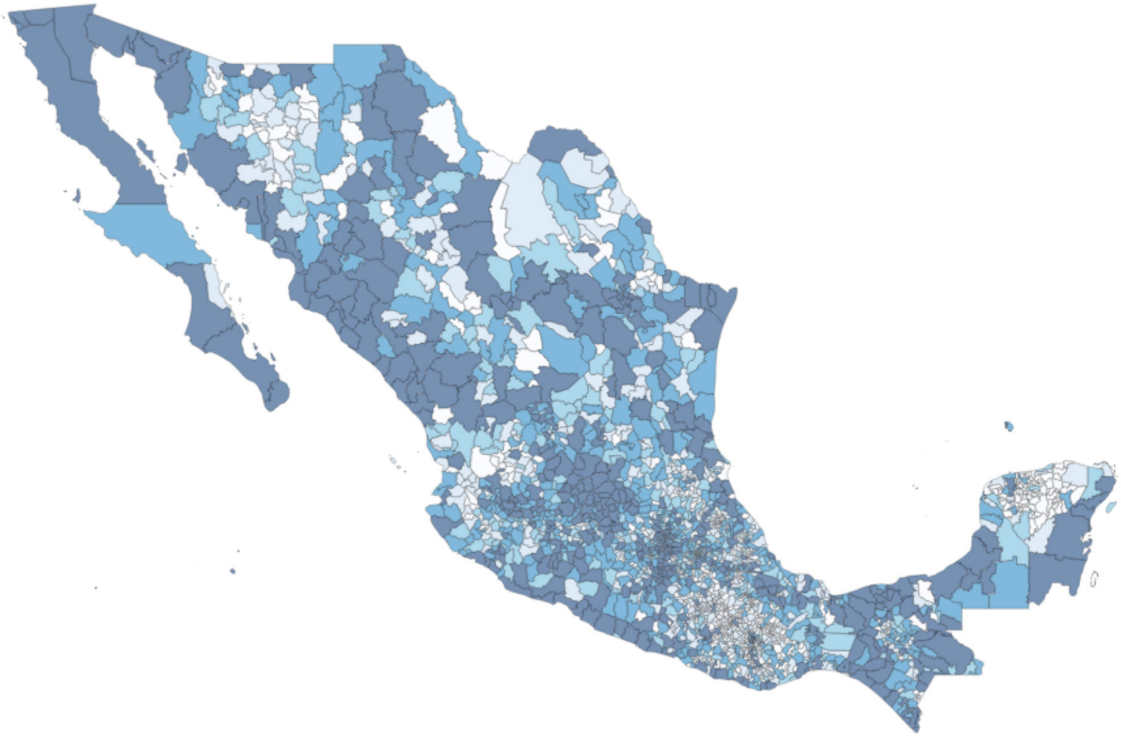


Figure 1: **Spatial partitions and homicide rates**. The map shows the 2.4k municipalities, color-coded by their respective murder rate during the entire period studied. Shades of blue correspond to quartiles of homicide rates, where darker blue indicates higher rates.

Crime shocks. We define a crime shock as a period of sustained and relatively low homicide rates, followed by a period of sustained and relatively high crime rates. In particular—and given the available temporal longitudinality at the intersection of the crime and behavioral datasets (one year from October 2014 to September 2015)—we use six-month windows, such that a shock consists of six months of relative peace, followed by six months of high crime rates. Figure 2 shows such shocks elicited for the municipalities of Mugica, Jalacingo, and Venustiano Carranza. More generally, Figure 3 shows the distribution of percentage change in homicide rate for all municipalities. In what follows, we consider that a municipality endures a crime shock when its six-month homicide rate increases by 75% or more, which corresponds to the .85 percentile (as shown in Figure 3).

As mentioned above, the current ubiquitous use of digital services and infrastructure generates "digital footprints" that can provide statistical information about societal dynamics. For example, mobile phone metadata can provide an approximation of individuals' mobility traces, and credit and debit card transaction records can provide an approximation of consumption and the activities that citizens engage with throughout the day.

2.2 Behavioral Indices.

Consumption data[1]. Here we focus on anonymized and aggregated card transaction records, from one of the leading banks operating in the Mexican market. From these we compute municipality- and monthly-level indices across the period of study (from October 2014 to September 2015) for all urban areas in the country. In particular, we measure the empirical distribution of average expenditure per person and compare them before and after the crime shocks. We

[1] We refer to consumption and expenditure interchangeably.

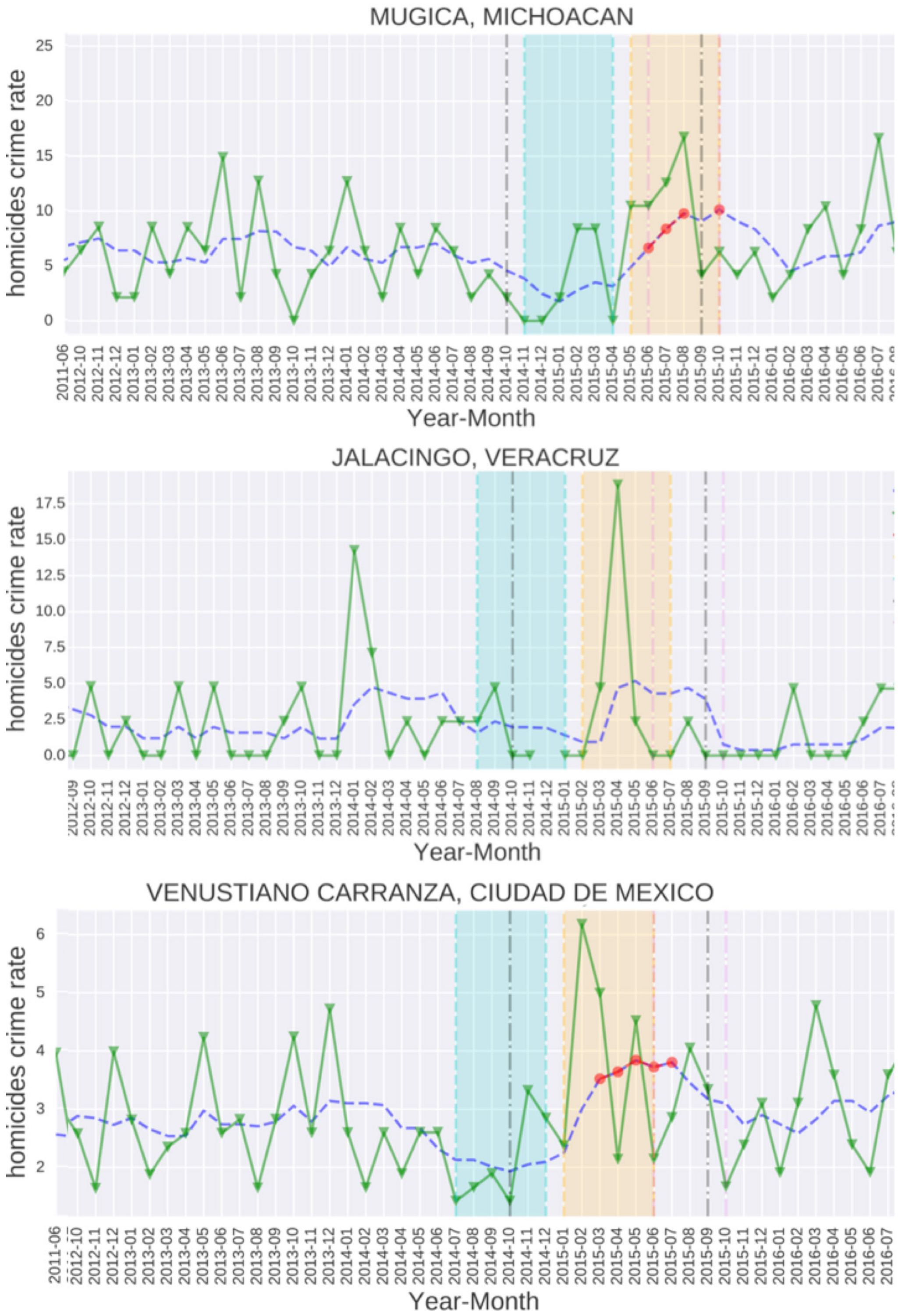


Figure 2: Crime shock examples. Examples of crime shocks in three different municipalities. The six-month windows of the relative peace are shown in blue, followed by six-month windows of high violence, shown in orange. Green lines denote monthly crime rates and dashed blue lines denote six-month moving averages. Red dots indicate six-month percentage changes which are 1.95 standard deviations greater than the average percentage change within the municipality, considering all available homicide data.

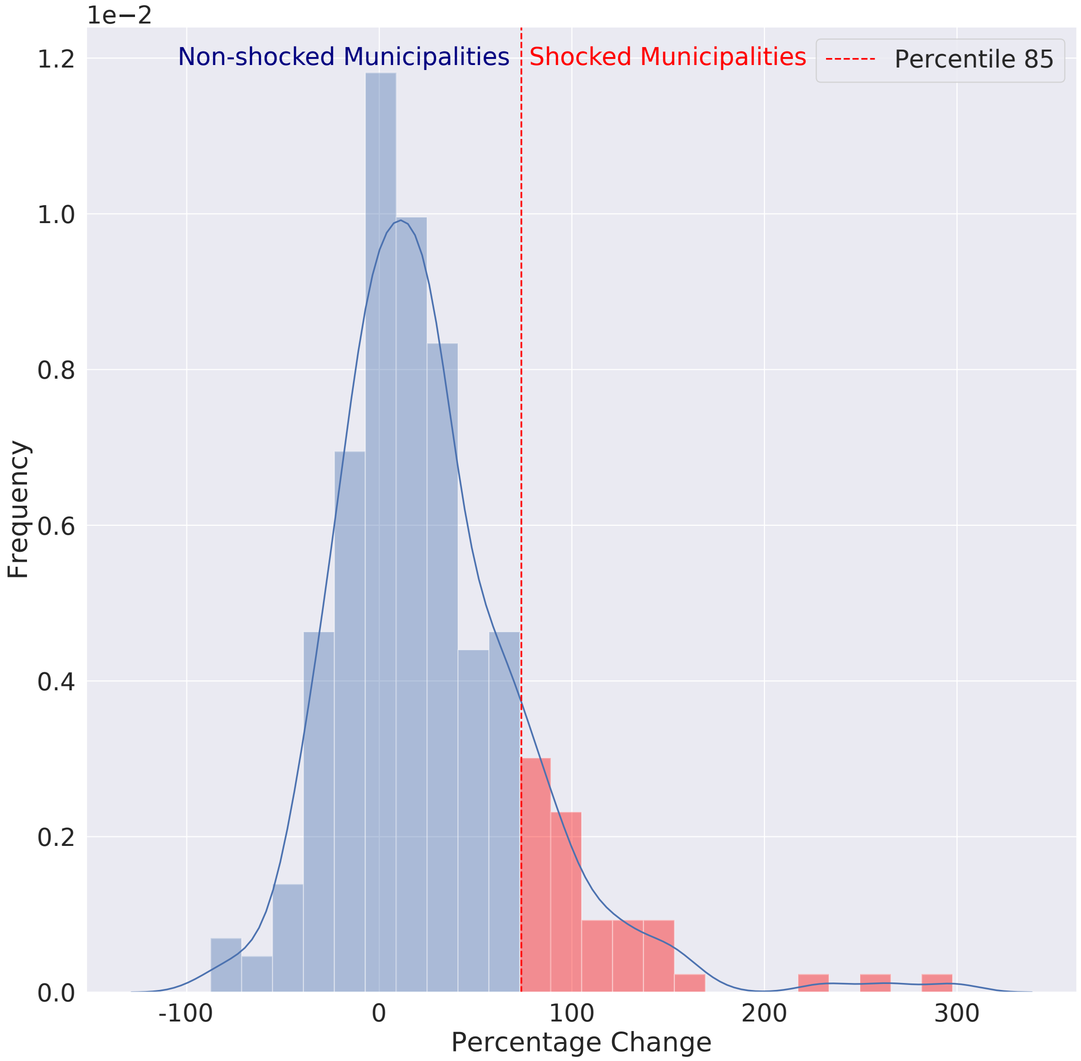


Figure 3: Distribution of changes in homicide rates of municipalities. Observations beyond the 75% change threshold are considered crime shocks. X axis is the percentage change of crime rate in the second six month period from the first six month period in October 2014 to September 2015.

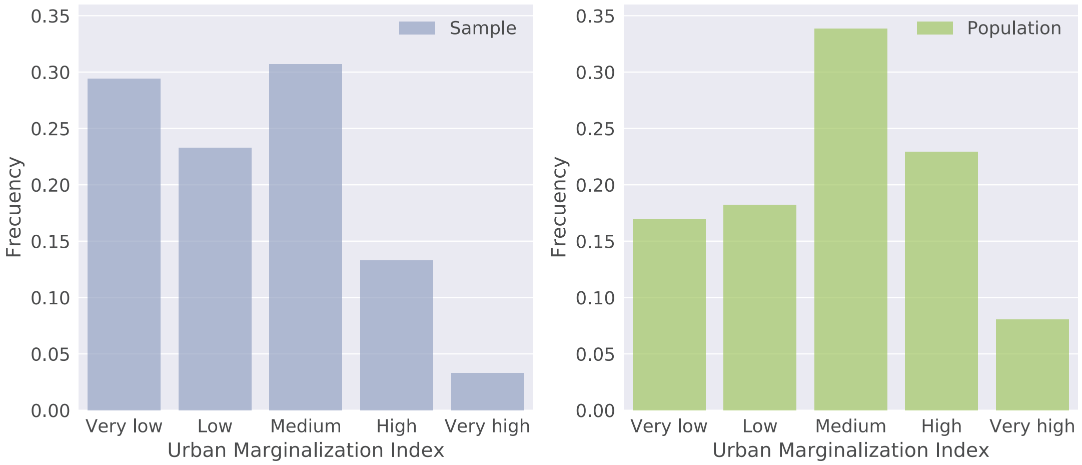


Figure 4: Socioeconomic distribution of consumption data sample vs. the country's population. The *Urban Marginalization* Index is the national statistic used to measure multidimensional deprivation and poverty, in particular associated with schooling, housing, income and residence in urban localities.

further subdivide the distribution per gender (male and female) and socioeconomic groups (defined by very low, low, medium, high and very high marginalization index) to investigate the differential effects across groups.

Socioeconomic distribution and marginalization index. It is relevant to note that our data sample is likely to differ from that of the overall population of the country, as we are restricted to individuals with access to basic financial services. To assess this socioeconomic representativity, we study the distribution of marginalization according to the urban marginalization index as defined by the national statistical office (INEGI)—a multidimensional measure of deprivation and poverty, in particular associated with schooling, housing, income and residence in urban localities. Figure 4 compares the sample and population distributions in terms of urban marginalization index. We note that, as expected, the sample is biased towards individuals in low marginalization segments. However, the sample does include a substantial amount of individuals across very-low, low, medium, and high neighborhood marginalization segments.

Section 3.2.2 studies heterogeneous effects of crime across these segments.

3. Results

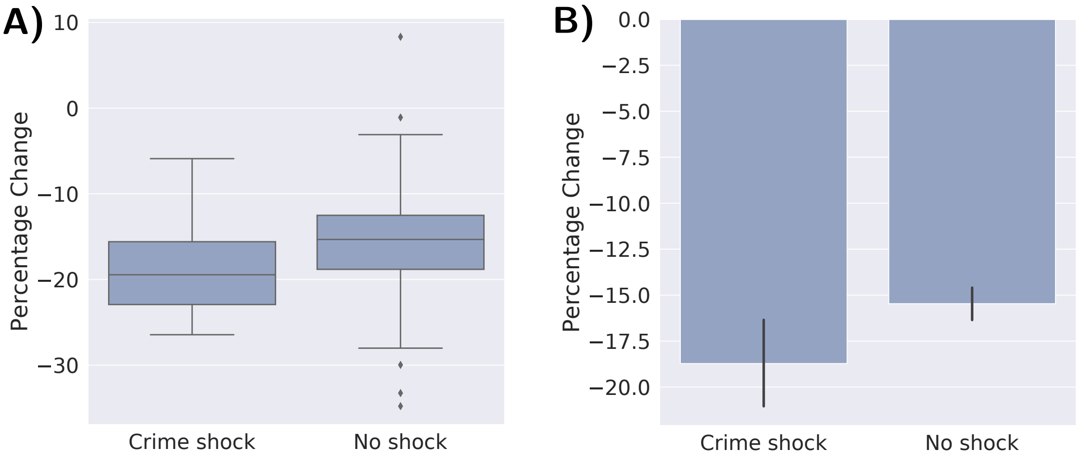


Figure 5: Effect of crime shocks on citizen's consumption patterns. A) Box-plot of average change of individual expenditure in each municipality. B) Bar chart representation with the 95% confidence interval. The shock condition shows that expenditure decreases significantly more in municipalities under crime stress. Expenditure is measured as the average expenditure per client within each municipality. The no shock condition corresponds to municipalities that did not experience a crime shock, and acts as a baseline comparison. Bars correspond to 95% confidence interval around the mean percentage change in expenditure.

3.1 Behavioral Disruptions in the Face of Crime Shocks

Figure 5 summarizes findings about the economic impact of crime shocks on the communities' consumption patterns and the overall economy. The y-axis for the crime shock condition is negative as it corresponds to the percentage decrease in average expenditure activity per person. The expenditure change of no-crime shock condition is also negative, mainly due to seasonality effects of the December period, and it serves as a baseline comparison against the shock condition. As expected, it is observed that the expenditure drop in the shock condition is larger than in the no shock condition, since a persistent wave of criminal activity dissuades households to consume within their municipality and encourages preservation in the face of uncertainty. The difference between the two conditions is above 3% (*p-value* < .05).

Results in Figure 5 are generated from average expenditure of residents from municipalities where the debit card transaction data covers more than 2% of the overall population, according to population extrapolations based on the last available census. Figure 8 in Supplementary Materials shows the distribution of sample coverage across municipalities, as well as the 2% threshold applied. In addition, we restrict the analysis to urban populations with more than 75k residents (in accordance with official population-based urban/rural classification). Finally, the no shock condition serves as counterfactual baseline for comparison against expenditure in municipalities that undergo a crime shock, hence controlling for seasonality effects, such as increased expenditure in December.

3.2 Differential Impact to Population Subgroups

As noted in (*28*), different subgroups in the population are affected differently by waves of violence. In this section, we explore these differential impacts across two subgroup dimensions, gender and socioeconomic status. If the effect of violence across gender groups (male and female) are different, one would expect it can exacerbate already existing inequalities especially if violence interferes more with female lives than males'. The same mechanism that lead to differential impacts across gender can lead to deepening inequalities across other dimensions of social or socio-economic status, as suggested by (*28*). Hence, we investigate whether waves of violence disrupt economic activity at different rates across neighborhoods with low and high urban marginalization index (as described in Section 2.2).

3.2.1 Across Gender

Women's concerns and behaviors in the public space differ from men's in ways that can be broadly associated with gender roles and a set of beliefs about safety (*29*). As evidenced by researchers, the experience of violence, including victimization and perception of security, is different among men and women (*18–20*), leading to differential behavioral changes. These changes in behavior can further exacerbate existing inequalities, such as unequal endowment with social capital (*30*) and access to new economic opportunities, ultimately slowing prospects of personal growth. Given these facts, any difference in how crime dissuades economic activity and social mobility across gender groups would mean that a crime shock would potentially exacerbate the existing level of inequality between men and women. In particular, women could be more constrained for a long period after a crime shock, hence limiting their access to opportunities in comparison to men.

This study quantifies the differential effect of crime across gender groups. Figure 6 shows our main results comparing the effect of crime shocks between men and women, based on changes in their respective expenditures aggregated at the municipality level. As hypothesized, the behavioral disruption effected by crime shocks is stronger for females than males. Crime suppresses female activity (6.1% decline on average) at a rate more than three times larger than males (1.6% decline on average), leading to a gender gap of 4.5%. The diff-in-diff activity across genders is statistically significant at a 0.05 level (p=0.024) obtained after bootstrapping the distribution of mean difference.

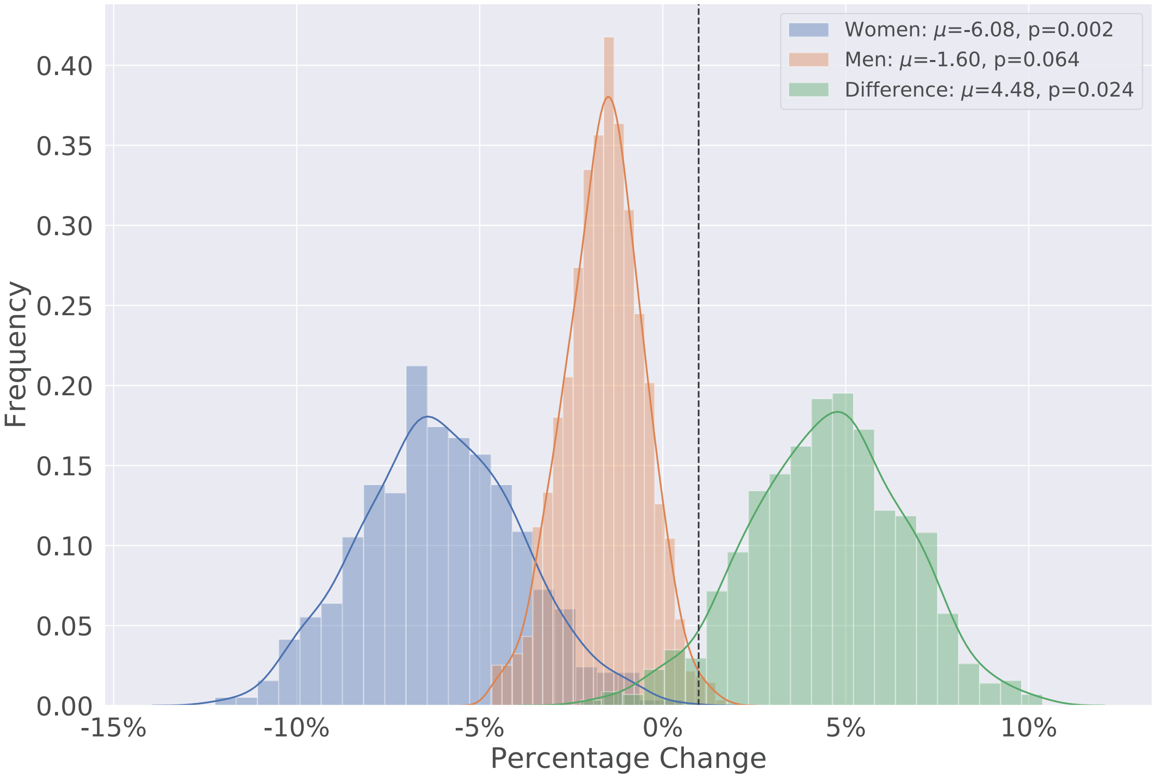


Figure 6: Effects of crime shocks on expenditure across gender. X-axis is the percentage change of expenditure after the crime shock when the trend from no crime shock condition is subtracted. The mean observed effect for males is −1.6%, and −6.1% for females, yielding a gender disparity of 4.5%. The p-values for male and female refer to the hypothesis that the activity change after the crime shock is zero. The p-value for the difference refers to the hypothesis that the difference between the two groups is zero. The distribution of difference is obtained through bootstrapping of difference between males and females.

3.2.2 Across Socio-economic Status

We repeat the same analysis for groups with different socioeconomic status. We use the marginalization index released as part of annual Mexico Census for different districts (there are multiple districts within one municipality). Districts with high marginalization tend to have lower access to education and other public services, and districts with low marginalization tend to be from high socio-economic status. We expect behavioral impacts of crime to hit areas with higher marginalization harder, due to higher mobility and economic constraints. For the purpose of comparison, we merged groups with very low and low marginalization into one group denoted as low and the rest into the second group denoted as high marginalization.

Figure 7 shows our main results on differential effects of crime shocks on districts with different marginalization status. Each district is assigned with a percentage change in average per-person expenditure after subtracting the baseline value from the no crime shock condition. The figure illustrates the distribution of these extra changes in expenditure after a crime shock across different districts with high and low marginalization. The results indeed confirm our expectation that within a municipality affected by a crime shock, districts of higher socioeconomic status (low marginalization) are less impacted by crime shocks, than districts with lower socioeconomic status (high marginalization), as compared to districts in a municipality not affected by a crime shock. Crime suppresses economic activity in high and low marginalized districts by −6.7% and −3.7%, on average respectively, leading to a 3.0% socioeconomic gap in the behavioral impact of crime. The diff-in-diff activity across low and high marginalization is statistically significant at a 0.1 level obtained after bootstrapping the distribution of differences across the groups. This mean difference is only marginally significant (*p-value* = .063) mainly due to small sample size. As mentioned in Section 4, in near-future work we will increase the size of crime shock samples by accessing larger longitudinal periods of overlapping card and crime data, in order to increase the precision of differential effects' estimates, as well as move into a cross-sectional approach across gender and socio-economics.

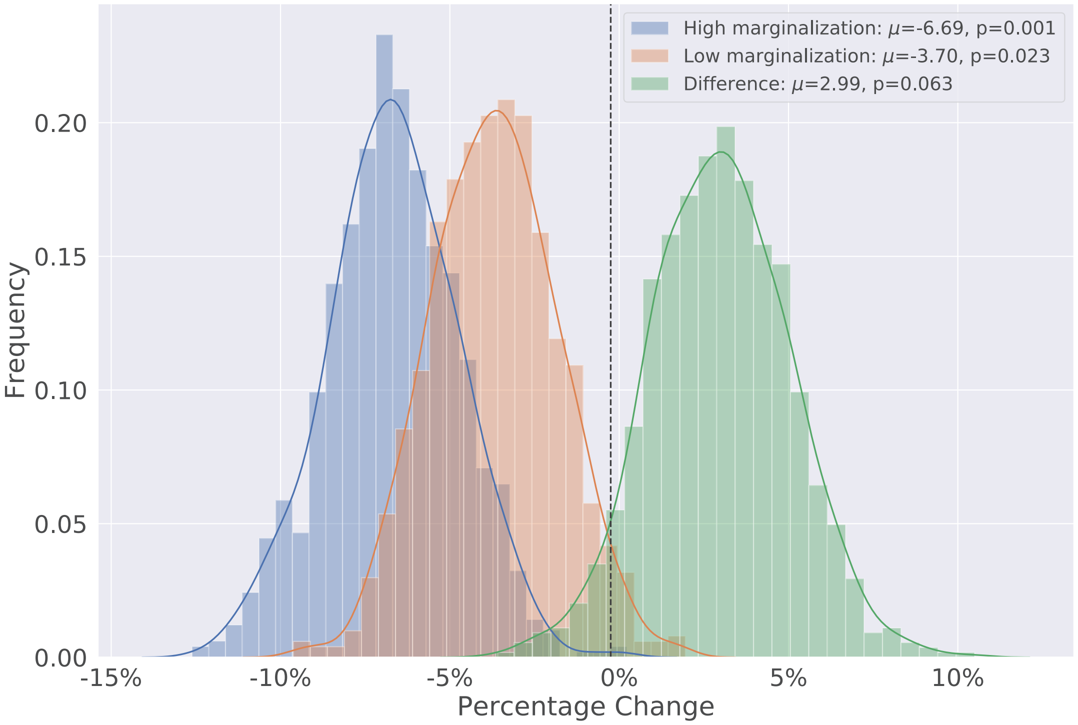


Figure 7: Effects of crime shocks on expenditure across different socioeconomic status. X-axis is the percentage change of expenditure after the crime shock when the trend from no crime shock condition is subtracted. The p-values for high and low marginalization refer to the hypothesis that the activity change after the crime shock is zero. The p-value for the difference refers to the hypothesis that the mean difference between the two groups is zero. The distribution of difference is obtained through bootstrapping of difference between the two groups.

4. Discussion and Future Work

In this work we have shown the effect that crime and violence have on citizens' daily lives. We undertook a novel approach that uses societal-scale information from anonymized card transaction metadata, to compute indices related to the consumption patterns of the population. We show that crime shocks have a negative effect on citizens' consumption patterns and that this effect can be unequally distributed across population subgroups.

A couple of data-related challenges constrained the extent of the analysis here presented. Most importantly, the present study relied on available data spanning one calendar year, for which both card transaction and crime data overlapped. This longitudinality constraint fixed the pre- and post-shock six-month window periods, hence reducing substantially the amount of crime shocks studied. Consequently, finer grain analysis, such as cross-sectional socioeconomic and gender analysis was not attainable due to power constraints; and similarly the study of heterogeneities across municipality types. In contrast, we have recently started a close collaboration with Banorte, the second largest bank in the Mexican market. Through a direct, on-site collaboration with their analytics team, we look forward to build and extend substantially on the results and framework here presented. In particular, by leveraging the project's relevant longitudinality from one to five years.

The results here presented are first steps towards mapping, understanding, and helping manage the effects of crime shocks on affected communities. Several avenues of work lie immediately ahead. First, regarding categories of consumption, it exists a rich categorization of merchants, which can be used to elicit category-specific effects. Future work will extract hierarchical meta-categories, to provide insight into how different domains of people's routine are differentially affected (e.g., entertainment and transportation).

Second, future work will explore at a finer grain how crime shocks unequally affect different community subgroups. In particular, by conducting a cross-sectional analysis at the intersection of gender and socioeconomic status, which is of great relevance towards translating observational insights into policy interventions. Third, future work will explore complementarities of other sources of data—mobile phone records in particular—and the additional insight they can provide in terms of effects on the mobility patterns of individuals. This additional data can ultimately be used to map more complex effects of high relevance, such as the relationship between crime shocks and community segregation. Finally, we expect future work to pursue causal insights from these wealth of observational data, by building detailed studies that leverage natural experiments in the data, or more sophisticated econometric estimators.

Overall, this paper proposes the use of digital footprints (metadata) to build indices of communities' behavioral patterns, at a large-scale and fine grain, in order to monitor the effects of crime and other types of shocks to local communities. The guiding vision of this work is a prospective ecosystem where the public sector, the private sector, and civil society participate in monitoring, communicating and managing the dynamics of disruption and recovery of communities in the face of crime, violence, and other adverse events. Today, several elements of such ecosystem already exist—the societal-scale behavioral data, technologies for safe data sharing (*31*), and policy-support platforms that aim at aiding local policy decisionmaking (*32*)—thus we expect these elements to connect and form such capabilities in the near future.

5. References and Notes

1. S. Pinker, *The better angels of our nature*: A *history of violence and humanity*. Penguin, 2011.

2. T. Economist, “How to cut the murder rate.” https://www.economist.com/leaders/2018/04/05/how-to-cut-the-murder-rate, 2018 (accessed October 7, 2018).

3. C. J. Vilalta Perdomo, J. G. Castillo, and J. A. Torres, “Violent crime in latin american cities,” tech. rep., Inter-American Development Bank, 2016.

4. C. Vilalta and R. Muggah, “Violent disorder in ciudad juarez: A spatial analysis of homicide,” *Trends in organized crime*, vol. 17, no. 3, pp. 161–180, 2014.

5. C. Vilalta and R. Muggah, “What explains criminal violence in mexico city? a test of two theories of crime,” *Stability: International Journal of Security and Development*, vol. 5, no. 1, 2016.

6. S. Machin, O. Marie, and S. Vujic, “The crime reducing effect of education,” ´The *Economic Journal*, vol. 121, no. 552, pp. 463–484, 2011.

7. I. Chatterjee and R. Ray, “Crime, corruption and the role of institutions,” *Indian Growth and Development Review*, vol. 7, no. 1, pp. 73–95, 2014.

8. C. M. Fuentes Flores, “El impacto de las viviendas deshabitadas en el incremento de delitos (robo a casa habitacion y homicidios) en ciudad juárez, chihuahua, 2010,” *Frontera norte*, vol. 27, no. 54, pp. 171–196, 2015.

9. G. Robles, G. Calderon, and B. Magaloni, “Las consecuencias econ ´omicas de la violencia del narcotrafico en m éxico,” tech. rep., IDB Working Paper Series, 2013.

10. J. Chabat, “Combatting drugs in mexico under calderon: the inevitable war,” 2010. ´

11. O. S. J. Initiative et al., “Undeniable atrocities: Confronting crimes against humanity in mexico,” *Open Society Foundations*, *New York*, 2016.

12. G. Trejo and S. Ley, “Municipios bajo fuego,” *Nexos*. Febrero. En: http://www. nexos. com. mx, 2015.

13. L. Jaitman, D. Caprirolo, R. Granguillhome Ochoa, P. Keefer, T. Leggett, J. A. Lewis, J. A. Mejía-Guerra, M. Mello, H. Sutton, and I. Torre, “The costs of crime and violence: New evidence and insights in latin america and the caribbean,” 2017.

14. N. Ajzenman, S. Galiani, and E. Seira, “On the distributive costs of drug-related homicides,” *The Journal of Law and Economics*, vol. 58, no. 4, pp. 779–803, 2015.

15. D. Rebolledo Sanchez, “La violencia como limitante para el desarrollo y el crecimiento económico en el estado de guerrero,” *Revista Mexicana de Ciencias Agr ícolas*, no. 12, 2015.

16. A. T. V. Montemayor, *Drug Violence, Fear of Crime and the Transformation of Everyday Life in the Mexican Metropolis*. University of California, Berkeley, 2016.

17. U. N. O. on *Drugs and Crime, Global study on homicide 2013: trends, contexts, data*. UNODC, 2013.

18. M. Averdijk, “Reciprocal effects of victimization and routine activities,” *Journal of Quantitative Criminology*, vol. 27, no. 2, pp. 125–149, 2011.

19. G. S. Mesch, “*Perceptions of risk, lifestyle activities, and fear of crime,” Deviant Behavior*, vol. 21, no. 1, pp. 47–62, 2000.

20. J.-A. Gale and T. Coupe, “The behavioural, emotional and psychological effects of street robbery on victims,” *International Review of Victimology*, vol. 12, no. 1, pp. 1–22, 2005.

21. R. B. Gutierrez and E. F. Molina, “El miedo al delito en las mujeres en méxico. expresión y determinantes,”

22. M. E. Avila, B. Martínez-Ferrer, A. Vera, A. Bahena, and G. Musitu, “Victimization, perception of insecurity, and changes in daily routines in mexico,” *Revista de saude publica*, vol. 50, p. 60, 2016.

23. A. . Pentland, “The data-driven society,” *Scientific American*, vol. 309, no. 4, pp. 78–83, 2013.

24. J. Blumenstock, G. Cadamuro, and R. On, “Predicting poverty and wealth from mobile phone meta-data,” *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

25. E. A. Martinez-Cesena, P. Mancarella, M. Ndiaye, and M. Schlapfer, “Using mobile phone data forelectricity infrastructure planning,” *arXiv preprint arXiv*:1504.03899, 2015.

26. M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M. L. Sbodio, “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*, pp. 663–666, Springer, 2013.

27. A. Bogomolov, B. Lepri, J. Staiano, E. Letouze, N. Oliver, F. Pianesi, and A. Pentland, “Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics,” *Big data*, vol. 3, no. 3, pp. 148–158, 2015.

28. I. Carvalho and D. A. Lewis, “Beyond community: Reactions to crime and disorder among inner-city residents,” *Criminology*, vol. 41, no. 3, pp. 779–812, 2003.

29. C. B. Gardner, “Safe conduct: Women, crime, and self in public places,” *Social problems*, vol. 37, no. 3, pp. 311–328, 1990.

30. N. Lin, “Inequality in social capital,” *Contemporary Sociology*, vol. 29, no. 6, pp. 785–795, 2000.

31. T. Hardjono, D. Shrier, and A. Pentland, “Opal/enigma,” in TRUST:: DATA: *A New Framework for Identity and Data Sharing*, ch. 3, pp. 79–99, ,: Visionary Future LLC, 2016.

32. “Plataforma preventiva.” Secretaria de Desarrollo Social (SEDESOL), Mexico, http://plataformapreventiva.gob.mx/. Accessed: 2018-06-15.

6. Supplementary Material

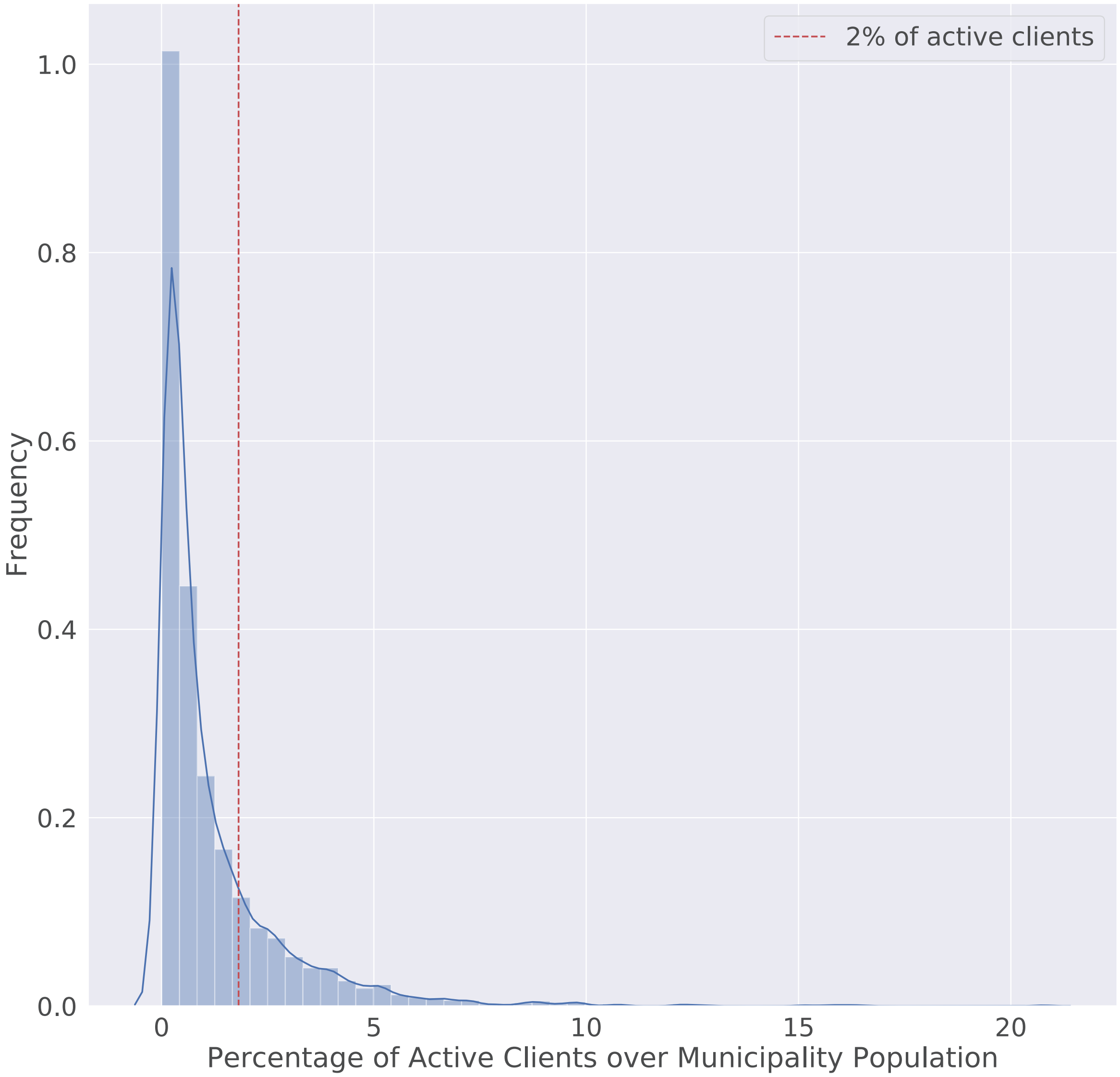


Figure 8: **The distribution of percentage of municipality population present in the debit transaction data. The municipalities below the 2% threshold in dashed red are discarded.** Only municipalities where active bank clients—clients which made at least one transaction during the period of analysis—represent more than 2% of the official estimated population were included in the study, in order to ensure a minimum degree of representativity in levels of expenditure in each municipality.

Papiers de Recherche de l’AFD

Les *Papiers de Recherche de l’AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ilss’adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l’ensemble des sujets de travail de l’AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l’ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l’AFD n’en exclut aucune autre.

L’Agence Française de Développement (AFD), institution financière publique qui met en oeuvre la politique définie par le gouvernement français, agit pour combattre la pauvreté et favoriser le développement durable. Présente sur quatre continents à travers un réseau de 72 bureaux, l’AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète. En 2014, l’AFD a consacré 8,1 milliards d’euros au financement de projets dans les pays en développement et en faveur des Outre-mer.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l’AFD. Ce document est publié sous l’entière responsabilité de son (ses) auteur(s).

Les *Papiers de Recherche* sont téléchargeables sur : http://librairie.afd.fr/

AFD Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

Agence Française de Développement (AFD), a public financial institution that implements the policy

defined by the French Government, works to combat poverty and promote sustainable development. AFD operates on four continents via a network of 72 offices and finances and supports projects that improve living conditions for populations, boost economic growth and protect the planet. In 2014, AFD earmarked EUR 8.1bn to finance projects in developing countries and for overseas France.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

AFD Research Papers can be downloaded from: http://librairie.afd.fr/en/

AFD, 5 rue Roland Barthes
75598 Paris Cedex 12, France
ResearchPapers@afd.fr
ISSN 2492 - 2846

BIG DATA

TO ADDRESS GLOBAL DEVELOPMENT CHALLENGES

2018