Prediction and Inference on Big Data in Development

Gabriel Cadamuro

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Sham Kakade, Chair

Joshua Blumenstock

Kurtis Heimerl

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Prediction and Inference on Big Data in Development

Gabriel Cadamuro

Chair of the Supervisory Committee:
Associate Professor Sham Kakade
Allen School of Computer Science, University of Washington

Recent years have seen an explosion of large-scale data sets pertinent to developing regions. The interest now being paid to country-wide satellite imagery and mobile network data has strong parallels to the proliferation of earlier work being performed on datasets such as ImageNet and the Facebook social network. The hope is that the techniques developed to process and analyze the data in this first iteration of Big Data can be now be turned to datasets from developing regions. Applications in data science for development include increasing business efficiency and competitiveness in these regions, as well as directly improving human development and well-being. This thesis seeks to make Big Data work for applications in the developing world through a comparison of several different projects, including predicting regional wealth and inferring the impact of violence from call data, and determining the quality of a road network from satellite imagery. With this breadth of applications and data types, an integrated approach comprising statistics, economics, and machine learning is vital in data science for development.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The past decade has seen an explosion of large-scale datasets in a huge array of different formats. Social media networks such as Facebook and Twitter have billions of users and record millions of interactions each minute. The ubiquity of high-quality cameras on cellphones have produced billions of images and videos which are annotated and shared by humans: YouTube alone sees 300 hours of videos updated every minute. Overhead, satellites collect imagery at sub-meter resolution at a scale that covers the globe with ever-increasing regularity while on the other end of the scale over 200 million wearable computing devices generate a constant stream of data[2].

However, this deluge of data is not simply a data warehousing challenge: it has also spurred the development of new machine learning algorithms and analytic techniques. The increasing power and versatility of Deep Neural Nets has much to do with the development of both well-curated training datasets such as ImageNet[65] and huge unsupervised corpuses like Twitter's firehouse[165]. A field that was previously exploring single-layer neural nets can now be said to have literal zoos [4] of different model formulations, handling everything from classifying videos[186] to translating different languages[146]. Having billions of records for how people interact with everything from news media to movies[83] has proven a rich source for the development and improvement of dimensionality reduction techniques [130].

These technical innovations have a number of applications; however, the focus has primarily been on servicing the large tech companies that generate them. The gigantic advertisement-serving industry that fuels companies such as Google and Facebook have constantly innovated on learning how to serve the best ads for customers[97]. The ability to classify and detect patterns in images appears in a plethora of different business ideas from the serious mat-

ter of generating captions for the visually impaired [56] to entertaining diversions such as swapping the faces of two people in photo. However, as people in developed societies have become more familiar with big data and more cognizant of its analytic power, interest has percolated widely. Analysis of large data sets can now be seen in the medical industry, from neural net analysis of medical images[150, 80] all the way to outlier analysis on millions of medical records[51].

The story of these trends' impact on developing regions is at once both similar and different. On one hand, we see familiar technologies such as mobile phones reaching maturity and peak penetration[7] in these regions, and enormous social networks datasets rivalling in size those that been generated in developed regions a few short years earlier. Indeed, we can already see these datasets making positive impact on business intelligence in new regions much as in the old. Large CDRs provide ways for telecommunications companies to understand how to incentivize customers[127], or for micro-lending companies to estimate credit- worthiness[31]. However, a variety of social, economic and political factors lead to interesting differences. Sometimes these hamper the growth of certain types of datasets and mitigate the potential benefit of analysis. For example, the considerable profusion of languages[11] and substantial gap in literacy rates make the analysis of huge online text datasets such as Twitter a less attractive option than it would be in a region like North America.

However, these differences often reveal great opportunities for new applications. In some cases, economic conditions have spurred innovative new technologies such as mobile money[112], which have enabled entirely new datasets and analysis[196]. Equally importantly, the more resource-constrained nature of governments in developing regions means that the ability to approximate economic or infrastructure surveys at a fraction of the cost[36, 49] becomes far more valuable than it would be in the developed world. In these ways, we perceive how we might leverage techniques honed in the past decade to improve human development and quality of a life in a substantial way.

This thesis concerns itself with this exciting new field of applications and the skills re-

quired to properly execute it. We examine four different projects that, while all falling under the analysis of Big Data for development purposes, vary enough in all other aspects as to provide a broad tableau for comparison. We will start in chapter 2 where we will dive into detail on call detail records and satellite imagery, two of the most useful and exciting sources of data in this area. In chapter 2.3 we provide an overview of the work done with Big Data in developing regions that concern themselves specifically with predicting quantities of socio-economic interest. We trace the development of machine learning algorithms to make sense of social networks and imagery from their conception to their application in several key papers in this area. In chapter 3, we highlight our own work in predicting the wealth of individuals in Rwanda from their calling history, and show how this method might be applied to create intermediate survey data. We then contrast this to another prediction problem of an entirely different nature. Chapter 4 shows how we developed a methodology to effectively predict the quality of roads and discusses how it might be put to use for goals such as econometric inference.

However, predicting values of socio-economic importance is only one of the uses that Big Data, as chapter 5 makes clear. Using the problem of emergency event management as a motivator, we discuss the importance of being able to attribute and infer the underlying cause of phenomena. In chapter 6 we explain provide a statistical analysis of a bias problem that underlies the study of emergencies from passively collected data. Then, in chapter 7 we perform an analysis that seeks to determine the extent to which violence impacts social cohesion in Afghanistan. This analysis leverages a country-wide call detail record that spans several years, as well as ethnolinguisitic survey data and detailed records of violent events. Finally, in chapter 8 we recapitulate all that we have learned over the projects and discuss what skills and mentalities are needed to pursue a truly integrated approach to this problem.

Chapter 2

BACKGROUND: PASSIVELY COLLECTED DATA IN DEVELOPING REGIONS

The past decade has seen an huge increase in projects that attempt to directly predict some important value using machine learning. However the single critical component of any machine learning task is the underlying data and so we introduce these here. We start by introducing both the Call Detail Record as well as the fundamentals of satellite imagery. We follow on by discussing why these datasets in particular are pertinent to the emerging world and why not several other sources of data commonly investigated in developed regions. Finally we give a board overview of machine learning that had already been done on these datasets to understand how these datasets can be mined for useful insights.

2.1 Call Detail Records (CDRs) in the Emerging World

2.1.1 Mobile Phones and the Developing World

Mobile phones have been among the most widely adopted new technologies in the developing world. In sub-Saharan Africa for example, it was estimated that mobile technology had was used by 45% of possible consumers in 2018 and that this represented a startling 20% increase in the number of customers from just a year ago[7]. In addition to usage patterns familiar in a developed context, users in the developing world have found exciting new uses for phones that highlight their versatility. They have been used as a way to find information vital to their work, from prices of fish [119] all the way to getting regular advice on agricultural practices. Moreover, technology for the transfer of money via cell phone (and associated mobile banking) has been primarily driven by demand in developing regions.

This impact of this technology's adoption had been profound. The mobile phone offers

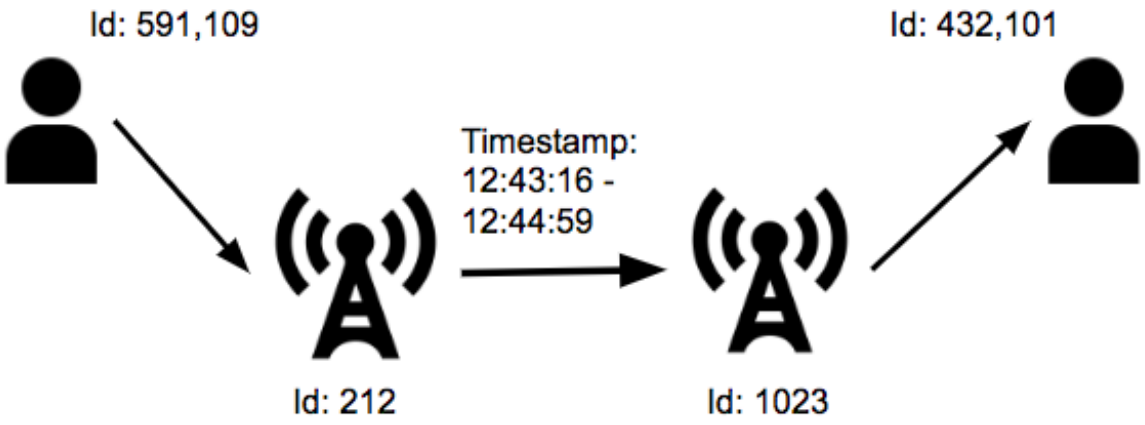
great potential for economic improvement through better profusion of information and social networking opportunities[21]. Moreover, economic studies have shown that development such as the lower barriers to banking enabled by mobile money have a inequality-reducing effect [20]. As the feature phones transition into internet capable smart phones this may further boost internet access by allowing a more suitable model of tower based Internet to thrive over the wired connections designed with a developed economy in mind. Indeed 23% of sub-Saharan phone owners report accessing the internet through their phone[7].

A call detail record (CDR) consist simply of the metadata, not the actual content, of all the calls and texts made through a mobile phone network provider. This metadata will always include who is calling whom (through anonymized hashed identifiers) as well as when these calls or texts were made and from/to which cell-phone towers the communication was transmitted. CDRs are *transactional* datasets, by which we mean they act like a database with one one row per transaction (in this case a call or text). An example of how a transaction is recorded in a CDR can be found in figure 2.1.

As this figure illustrates, there are several different ways in which a CDR could be visualized[34]. One one level, the CDR represents a weighted and directed social network where calls (or texts) from one person to another represent an edge in the network. In this way, we can see that algorithms developed to study large social networks, such as Pagerank or graph decompositions (CITE), could be applied. However, CDRs are strictly richer, since they contain information such as the time at which a call was made for each interaction. This suggests that less thoroughly-explored methods like tensor decomposition[128] or hypergraph embeddings [233] are the more natural approach.

2.1.2 Processing CDRs

In order to properly inform a discussion of the prior work in the area and our proposed extensions, it is necessary to introduce some fundamental concepts in both CDR processing and statistics. The first one is the geo-location abilities of CDRs and how we can use these to define a *home tower* for each individual in the CDR.



Timestamp	Id_out	Id_in	Tower_out	Tower_in	Dur	Hour
170123-124314	591,109	432,101	212	1023	103	12

Figure 2.1: Different columns represent different types of metadata. For example the caller id metadata (Id_out and Id_in) and tower id (Tower_in and Tower_out) are both drawn from discrete distributions though the former has a much higher dimensional than the later. In contrast the duration ('Dur') column is integer valued.

CDRs will almost always have the id of either the tower that received or sent a call/text (and often both). Since the telecommunications company knows where each tower was erected, it is possible to assign a latitude and longitude to each one. These in turn allow us to determine where an individual was when this call was made/received within some boundary of uncertainty, as explained in figure 2.2. This information can be used to provide mobility information, but it also allows one to estimate where an individual lives. We do this by determining the *home tower*: namely, the closest tower to the individual's home. This done by looking at all the towers that send or receive calls for the individual during the period of time in which they are assumed to be at home: this is often assumed to be between 8pm and 6am. The tower with the highest number of calls is then assumed to be the closest one to the individual's home (via the logic explained in figure 2.2) and assigned

as the home tower. Since we know the latitude and longitude of the tower we can then assign this individual to a city/region/province etc: often an essential part of a CDR based study. To account for the possibility of internal migration, the home tower computation is repeated for each month or week of CDR data.

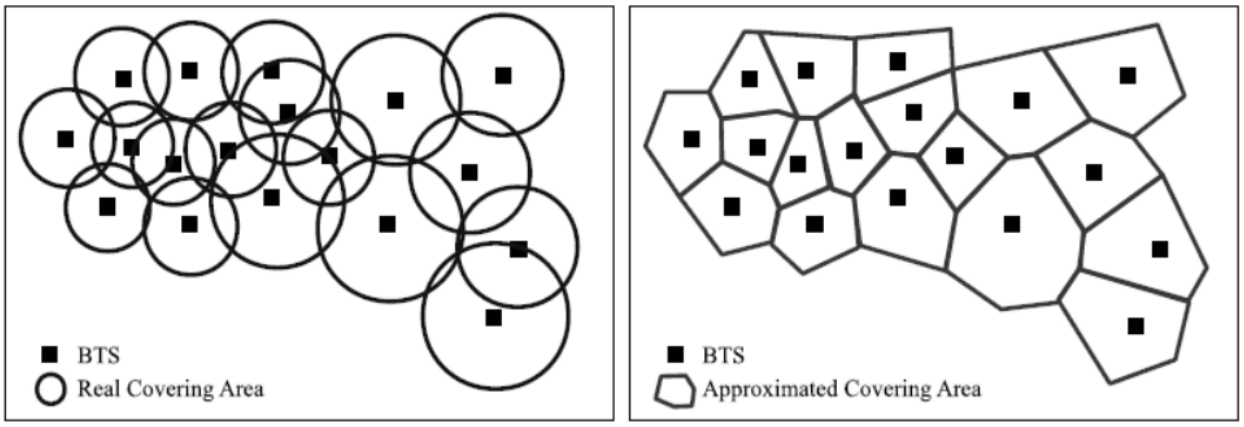


Figure 2.2: Consider a tower network where the squares represent the positions of different cell towers. Reasonably a phone call at some given point might be served by any tower which contains that point within its service radius (left image). For analysis purposes we simplify this by assuming that every point is uniquely served by the tower closest to it: inducing a Voronoi tessellation (right). Figure reproduced from [89].

The second of these is the idea of processing a tabular CDR database into individual-level *metrics*. A *metric* is hence simply a function that takes a CDR and returns a real-numbered value for each individual identifier (id) present in the database. A simple example might just be the function which counts the number of calls made by each id in the CDR: we would call this metric 'number of outgoing calls'.

A few metrics bear specific introduction: a class of these are distribution-based metrics. If we take all the calls made by an individual and look at a specific metadata type, say the id of the recipient individual, then we can aggregate the occurrences of this metadata in each call into a discrete probability distribution. We are then interested on functions that can be done on discrete distributions: mainly, the size of the support set (unique number of call recipients) and the Shannon entropy (entropy of call recipients). The former gives us

a classical social network property known as the *degree* of an individual and the latter has been linked to important social and economic properties of individuals[77].

A second metric that bears introduction is the *radius of gyration* (R_G). This is a metric that estimates the spatial mobility of an individual based on which towers they receive or make calls from. In the most general case the R_G is computed by having L measurements of an individuals position $\{R_i\}$ at several different times. From this, a center of mass (R_{cm}) can be estimated, and the standard deviation from that gives the radius of gyration, as seen in equation 2.1. In the case of a CDR, the R_i will be the latitude-longitude vector of the cellphone tower from which the ith call was sent or received.

$$R_{cm} = \frac{1}{L}\sum_{i=1}^{L} R_i, \quad R_G = \sqrt{\frac{1}{L}(R_i - R_{cm})^2} \tag{2.1}$$

2.1.3 Analytic Uses and Limitations

CDRs are propriatary datasets owned by cellphone network providers, and while there is serious research on their industrial applications[127], they have also proven to be an invaluable source for economic, sociological and policy [34, 154] research. In fact, even careful analysis of uncomplicated CDR properties can yield interesting conclusions. In [41] it was shown that the raw number of international calls closely followed international trade patterns, while following the movement of individuals via CDRs closely correlates with independently gather migration statistics[42]. Economic quantities such a region's wealth have been noted to relate to large scale calling habits[147] such as the number of calls and proportion of sent and received calls. In the realm of sociological study, correlating call behavior with linked demographic statistics of callers also surfaces notable trends in gender and age [37] as well as wealth and education [88] using simple metrics. Even without linked demographic data, researchers can understand the contrast between social networks in cities and the countryside[76] through analysis of their calling habits.

However, recent studies of CDRs have made use of more advanced metrics. One of the

more influential papers looking at the meaning of these metrics examined how the network entropy and location entropy (based on which regions are called) of regions in a country related to their wealth[77]. This found strong evidence to suggest that increased wealth is positively related to more diverse social networks. Advanced mobility metrics such as radius of gyration and location entropy have been likewise found to generally correlate with different measures of wealth [167, 88]. These metrics also show promise for untangling more tricky social properties of interest, such as discrimination[45], which cannot be seen through more simple measures such as call volume.

2.2 Satellite imagery and remote sensing

The field of remote sensing is one that is seeing growth not only in the quality and number of existing platforms but also in the breadth of new technologies and approaches. Remote sensing was originally the purview of militaries in the 1970s and 80s, with explicitly military objectives such as nuclear weapon monitoring or espionage. However, experience with both the platforms (such as satellites) and the sensors apparatus (high resolution cameras) led to this technology spreading first to more public government roles and then to the free market. Prominent examples of the former include the NASA Landsat family of satellites which have underpinned research in applications ranging from meteorology[216] to agriculture tracking[126] and monitoring deforestation[75]. The latter development has led to several companies investing in fleets of satellites (known as constellations) and upwards of 2000 satellites currently in orbit[8]. While some of these constellations, such as Telesat, were launched for communication purposes there are now several companies with dedicated observational constellations. DigitalGlobe, for example, provided the imagery that underlies most of Google's mapping service. Another competitor, Planet Labs, has a constellation of over 150 active "cubesat" satellites.

These two companies exemplify two exciting trends in the quality of remote sensing imagery. DigitalGlobe imagery includes some at resolutions of 50cm per pixel, a resolution that until partway through this decade had been exclusively for military use. This high

resolution has enabled projects that require being able to distinguish individual houses[193] or roads[49] to function on a world-wide scale. Planet Labs' approach of large constellation of smaller cubesats have paid dividends in another direction: that of image cadence. A large constellation enables imagery of a given location to be refreshed daily, which is crucial for applications that need to track events in real time such as monitoring volcanic eruptions[16]. This is in stark contrast to the situation even a decade ago when the most commonly used source of daily updating imagery, NASA's VIIRS, had a 750 meter per pixel resolution.

However, the development in novel sensory equipment and platforms is an equally important development. Imagery taken from an aerial platform, especially an unmanned aerial vehicle (UAV) has become more important with the corresponding growth of that industry. While UAVs sacrifice the global reach of satellites they enable much greater resolution at a fraction of the cost due to the much smaller altitude these vehicles fly at. These have been of particular use when monitoring urban areas[93] due to smaller area of analysis and relative ease of acquiring imagery. The platform need not even be as sophisticated as a drone: an innovative work[114] showed how a mobile camera attached to a small balloon could be useful for monitoring small farms in India! While visual imagery is perhaps the type of remote sensing that first jumps to mind it is important not to ignore the other types of sensory information available. Spectral imagery outside of the visual range has been long used as an important proxy for crop growth and light-based ranging technology (LIDAR, the visual spectrum equivalent of radar) has been operated from airborne platforms[148].

Remote sensing, and in particular that based on satellite platforms have a number of advantages in the context of developing world applications. The first is that satellites are truly global and do not incur any transportation costs to obtain imagery of any part of the globe. This is significant advantage over measurements that have to transport specialized equipment or professionals since developing regions, and especially rural or remote districts, will be harder to access. Limited transportation infrastructure and rugged terrain would be comparatively bigger problems in these cases. The second advantage is that satellites can persistently measure an area without having to deal with local factors such as security

or supply. This in invaluable for studies involving areas in the midst of conflict[52] or natural disasters. The third advantage, a natural by-product of the other two, is that satellite measurement can be done in a truly unbiased manner if desired. Other passively collected data sources such as social media or CDR rely upon a set of users that might have a bias towards being younger or more urban. Manual measurements or surveys may focus their attention on more densely populated or easy to access areas to the detriment of rural communities. There is no reason for either of these biases to occur when using satellite sensing data.

2.3 Machine Learning on developing region data sets

Though data analysis has been performed on CDRs for a while, more focused approaches to studying them from a machine learning lens did not start to arrive until the last half-decade. One notable exception was the application of a graph partitioning algorithm to uncover hidden linguistic structure in a call network of Belgium [35]. More recent work has tended to focus on the idea of using CDR-derived metrics as the input into a machine learnt model to attempt to predict important sociological and economic metrics. The actual learning algorithms cover the spectrum from regression[36] to Bayesian inference[73] but one commonality is an evolution away from smaller set of easily interpreted features to more powerful large feature sets: we discuss this further in chapter 3.3. Regardless of the underlying machinery: these projects have shown promise in accurately predicting a wide variety of attributes including economic (wealth[36]), demographic[73] and sociological (crime rates in city districts[46]).

Machine learning on satellite imagery has a long history, but an understanding of its possible role in development came a little later than it did for CDRs. Initial attempts to detect objects from satellite imagery would apply different machine learning algorithm (AdaBoost, SVMs etc) by featurizing the image into pixel histograms or other such features[57]. This led to projects to extract roads[151], detect buildings[99] as well as machine learning approaches to crop yield estimation based on special spectral data[123]. A seminal paper

matching spatially tagged economic survey paper with a convolution neural net approach resulted in accurate wealth prediction over several countries[117]. While deep nets had been used in satellite imagery before this showed that CNNs were sufficiently powerful to infer economic variables that were previously considered too complex to estimate from just remote sensing data. Several works have followed in this vein by applying deep learning to novel difficult tasks like mapping slums[14] or understanding the quality of a road[49]. Not all recent work involves application of deep nets: in particular predictions of agricultural yield has a long history with several different approaches with continued success in that are[223]. An interesting question is whether machine learning can be done on CDRs and satellite imagery in a synergistic way, since they together cover both a social and physical concept of wealth. Some work has been done in this area[194] but this remains a rich opportunity to explore.

2.3.1 Learning on Internet and Social Media Datasets

Of course, passively collected databases from the internet such as social media have been the focus on a great deal of attention in the machine learning community. Key to this are two key attributes. The first is the considerably higher interaction rate of individual with these technologies: Twitter alone has 500 million new tweets a day. The second is the rich text data that is freely and legally accessible (unlike the payload of a call or text) for analysis. This has enabled some excellent work with applications to issues of socioeconomic interest. One project predicted regional unemployment rates in Spain based on the tweets being made in that region[142]. This utilizes indicators present in CDR datasets (such as what times of the day tweets were made during) but also features based on sentiment analysis of Tweets not possible on a CDR. Likewise, social media analysis has shown the ability to accurately track international migration[225] among OECD countries. This would be significantly harder to do with CDRs due to limited range of a given network provider.

The main issues with transferring these methodologies and learnings to the developing world is that, for the moment, there is significantly less penetration of both the internet and social media than there is for for phone activity. Moreover, the usage rates vary significantly

across the developing world in a way that usage of mobile phones does not. For example, 75% of internet connected Jordanians use social media while only 26% of internet connected Indonesians say the same[169]. This would lead to concerns that any findings on these datasets are heavily skewed towards a younger/wealthier subset of the population. The second is that alot of the NLP learnings on large text datasets simply cannot be replicated in the linguistic conditions of the developing world. Many countries are both multi-lingual (see chapter 7.2.2 for example) and have languages with only minimal NLP study. Even in countries where the official language is one with a well-developed literature researchers must be aware that dialectal variation can cause significant performance issues[33].

Chapter 3

CASE STUDY: PREDICTING INDIVIDUAL WEALTH IN RWANDA VIA CDRS

3.1 Motivation

One of the key problems in addressing economic and social challenges in the developing world is the lack of available data. The geographic distribution of poverty and wealth is used to make decisions about resource allocation and provides a foundation for the study of inequality and the determinants of economic growth. However, such statistics can be difficult to come by in parts of sub Saharan Africa. They may have serious data issues [120] or simply not have had a survey for many years. Angola, for example, did not have a national survey for over 40 years[44] . Collecting information about household income or asset ownership with a survey is the best course of action in such cases, but such an endeavour is expensive and time-consuming. Resource-constrained governments may simply opt to skip data collection entirely in such a case.

This work wondered if it might be possible to approximate the impacts of a full survey with a much more lightweight approach involving a smaller target phone survey and inference on a CDR. The idea would be to use the survey answer to form a numerical estimate of the wealth of the respondents. This would then be the objective function for a machine learning tasks that uses the calling records of these individuals to learn a model mapping calling behaviour to wealth. Applying this model to an entire CDR would enable a broad estimation of wealth levels across the country. While this would be no means replace surveys, which remain the gold standard for collecting such information, they would act as a helpful substitute when such a survey is unavailable.

One important note about Rwanda is that it has a strong record of consistently monitoring

internal demographic and economic data. The National Institute of Statistics of Rwanda runs Demographic and Health Surveys (DHS) at regular intervals to which asks questions about asset ownership and health outcomes. These are asked at the household level in a random sampling of villages. This is helpful for the study since it will enable our own estimates of regional wealth (as derived from CDR data) to be compared to the DHS outcomes aggregated at the regional level.

3.2 Surveys and Creating the Wealth Composite

In Summer 2009, we coordinated a phone survey of a geographically stratified group of Rwandan mobile phone users. Using a trained group of enumerators from the Kigali Institute of Science and Technology (KIST), a short, structured interview was administered to roughly 900 active mobile phone subscribers[37]. The survey instrument contained approximately 80 questions that focused on basic socioeconomic and demographic information. This included questions about asset ownership, such as whether the respondent owned a motorcycle, whether the respondent's house was electrified etc. In the interest of protecting privacy we did not solicit any personally identifying information such as first name, last name, or address (apart from the phone number used to contact them).

Care was taken to try and ensure that this survey was as accurate and representative of the Rwandan mobile phone owning population as possible. Phone numbers were assigned to hometowers (as described in 2.1) and then districts based on their calling patterns. Callers were then selected using a geographically stratified sampling technique based on the total number of subscribers per district. The contact rate was roughly 61%; non-contacts were largely the result of phones that were turned off or disconnected. The cooperation rate was 97%; almost everyone who picked up the phone was enthusiastic to participate in a study with university researchers, with whom they generally had little prior contact. We thus interpret the survey sample as representative of the population of the aforementioned mobile phone using population.

In Rwanda, as in most developing countries, it is difficult to estimate the socioeconomic

status of a survey respondent with a single survey question. Instead, household surveys typically rely on a large number of questions which can be used to infer the consumption or permanent income. The first principal component of these responses is commonly treated as a proxy indicator of the respondents unobserved wealth. In the phone survey for this project the final set of asset ownership questions (and the baseline positive response rate) used as the input for this computation were as follows:

- Do you own a refrigerator? (11%)
- Is your household electrified? (60%)
- Do you own a television? (49%)
- Do you own a bicycle? (30%)
- Do you own a motorcycle/scooter? (11%)
- Do you own a radio? (96%)

3.3 Learning methodology

The survey in 3.2 has provided several possible objective functions to train on. This included the asset ownership questions (boolean objectives for a logistic regression task) as well as the wealth composite (a continuous objective for a regression task). This gives us 867 training points, but the next challenge was how to design the feature set that would be used for training as well as choosing a learning methodology that would effectively run on the large CDR dataset.

3.3.1 Feature Engineering on the CDR

At the time, the CDR operator had roughly 90 percent market share, and 1.5 million registered Subscriber Identification Modules (SIM cards). The CDR included domestic calls,

international calls and SMS records. We decided to focus our analysis on the year of calling history that precedes the telephone survey run in June of 2009. In the end, this amounted to over a billion call and SMS records which needed to be processed into a feature set appropriate for machine learning.

Converting a transactional dataset like a CDR to training features is not a problem with an intuitive or canonical solution. A training set requires us to have one row per unit w we wish to predict on, in this case an individual subscriber, with a variety of information pertaining to these units making up the columns. Unlike the structured, limited range numerical values of a image (which is structurally amenable to a neural net treatment) the data relating to a single individual might be spread all over a CDR. Moreover, the data making up the CDR takes a variety of different formats from high dimensional categorical features (caller/tower ids), real valued features (call duration) and temporal features (date, time of day). The general approach to the featurization problem had been to manually select a small set of features know to be relevant to the prediction problem at hand and use those as the features [87]. While this did incorporate expert knowledge usefully and gave a parsimonious model, it left open the possibility of bias or missing sources of data based on the individual researcher.

We instead desired to develop measures of poverty and wealth that maximized predictive accuracy, possibly at the expense of the interpretability of the model. Thus, instead of devising a parsimonious set of metrics based on intuition, we take a brute force to feature engineering that is designed to capture as much variation as possible from the raw call detail records. Specifically, we developed a method based on a deterministic finite automaton (DFA) to generate a large number of potentially correlated metrics[171]. A DFA is defined by a graph with a set of states (nodes) and legal transitions that can be taken depending on the current state (edges). The DFA also has a specified state to start on and another one to end on.

In our featurization methodology, the DFA represented all the ways to generate a legal feature. The start state represented all the initial data and each transition represented a

database style operations legal at that state. For example, filtering all calls to only include calls made in the morning might be one transition: returning the average duration when grouping by caller-ID might be another. In this case there is still an element of expert knowledge (in knowing what types of transactions are likely to lead to interesting features) but the feature space is much more thoroughly explored than it would be if each individual feature had to be hand picked. The idea was that even though this might result in a number of redundant or useless features, the step of removing such features could be handled by a data driven regularization. In the end, the DFA we designed produced over 3000 different feature definitions.

3.3.2 Supervised Learning

From the several thousand behavioral metrics constructed by the DFA, we used supervised learning techniques to identify a smaller subset of features that are the best joint predictors of the response variable, using the sample of 856 survey respondents to train the model. Specifically, we use elastic net regularization [234] to penalize model complexity and reduce the likelihood that the model is overfit on the small number of training instances. For each possible model parameter β_j, the elastic net imposes a penalty equal to:

$$\lambda(\alpha\beta_j^2 + (1-\alpha)|\beta_j|) \tag{3.1}$$

This formulation enables both L_1 and L_2 regularization to act on the feature set. L_1 regularization is particularly important since this will shrink the feature set, which is desirable given that we expect the feature set will contain many variables. However, L_2 prevents any one feature from becoming overly dominant, which might contribute to stability when applying this model to the larger CDR outside of the training set. Note that while L_1 adds a significant computational load, the actual training set only has 850 individuals and so this does not pose a computational problem.

We experimented with several learning methods including decision forests and SVM but

did not see substantial differences over the regression. As such, we chose to go with regression, due to the ease of computation of applying this model to over a million subscribers and the relative ease with which one can explain the model. Models were then trained on each of the six ownership variables as well as the wealth aggregate using 5-fold cross validation. Models attempted to predict the wealth composite were evaluated by the R^2 metric which shows what percentage of the variance in wealth is correctly predicted by the model. Models attempting to predict asset measurement were evaluated by the Area Under the Curve (AUC) metric. This metric was chosen instead of accuracy since the different asset ownership metrics we were attempting to predict had very different background ownership rates, and comparing AUC rates across these was more of an apples-to-apples comparison since it accounts for the inherent difficulty of predicting on unbalanced classes.

One of the advantages of the feature selection method we employed becomes clear when we look at the underlying structure of the different models trained. Models trained to evaluate different objective functions all end up selecting roughly the same number of features (around 100 out of the 3000 available) but the actual choice varies greatly. Figure 3.1 for example shows how features summarizing the mobility of an individual is relatively more important for estimating motorcycle ownership while international call activity is more helpful for determining aggregate wealth. This shows that generating a large set of base features enables a wide wider variety of models predicting different objectives to built on top of this set than would be possible on a smaller set of hand-crafted features targeted for a single objective.

3.4 Evaluation and Application

We note that there are two aspects we want to carefully evaluate. The first is whether the machine learning is succesfull in a narrow sense: whether it can accurately predict wealth and asset ownership on this set of 865 people from their calling patterns. We are curious to see if some assets are easier than others to predict and what we can infer about the amount of data required for accurate predictions. The second is whether the output of such a model is useful more broadly to help inform policy. In this light, we will attempt to apply

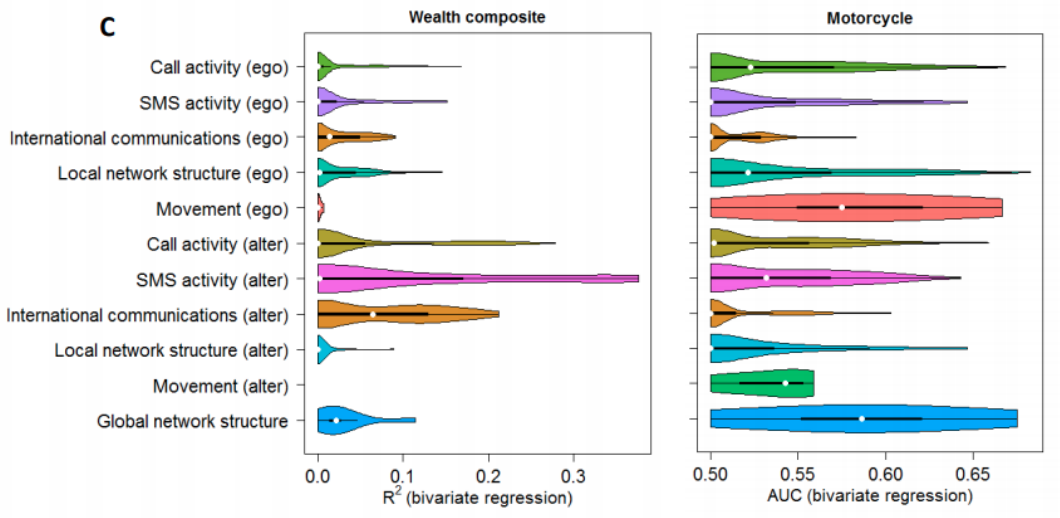


Figure 3.1: The graph compares the relative importance of different groups of features for predicting Wealth and Motorcyle ownership. The bars shows the distribution of all features within the group in terms of their R^2 (for regression) or AUC (for logistic regression) if a model was trained with that feature alone.

our predictive model to the entirety of Rwanda and attempt to aggregate estimates of asset ownership and wealth at a regional level. This can then be compared to DHS surveys to understand how applicable our methodology would be in practice.

3.4.1 Evaluating the Predictions

We first evaluated the performance of the model on predicting the wealth composite. Our final model obtained an R^2 score of 0.46, obtained through five-fold cross validation. As figure 3.2 (A), the prediction is accurate on aggregate though not necessarily for any given individual. Indeed, the later restriction would be unrealistic for (unlike an image in an image recognition) an individual's phone record will never suffice to give a complete economic and social reconstruction of a person. We also investigate the role of our training sample size on the prediction quality. By randomly subsampling our full training set at different rates and then evaluating it on the test set we can see how our predictive score would decrease with less data. This relationship is explored in figure 3.3 and highlights the importance of having

at least 400 or so survey respondents to train model on. An ideal future work would be to expand the respondent set to several thousand to explore this relationship more fully.

We realize that in some cases the policy interest may be in identifying people below a certain wealth threshold, such as a poverty line, instead of a granular wealth estimate. As such, we performed logistic regression experiments to see how well this method could identify the bottom 5 and 25% by wealth. The ROC curves of this task are show in figure 3.2 (C) and the results are summarized in table 3.2. It should be noted that in developing regions, less affluent subscribers will make fewer calls and hence have a smaller call history. This makes distinguishing the bottom 5% actually harder than the bottom 25% since there is so little data to work with and highlights an important limitation of this work.

In addition to wealth, we were interested in seeing if we could predict the asset ownership variables that made up the wealth composite. One could imagine scenarios where an NGO or government might be interested in a particular aspect (such as study on mobility understanding where bike or scooter ownership is lagging) instead of a more generalized wealth index. As such we preformed logistic regression tasks on the six asset ownership questions introduced in chapter 3.2. The ROC curves are shown in figure 3.2 (B) and the scores are summarized in table 3.1. Though overall the predictive scores are strong it is noticeable that some assets are easier to predict than other. In particular, assets revolving around access to electricity have stronger predictive performances: possibly due to the changes in call activity during the late night or early morning.

3.4.2 Application to District Estimates

While we have seen that our machine learning approach has performed well on a variety of tasks we have not addressed the central goal of this work: can models trained on CDR approximate national survey results? While the ability to predict well within a set is a necessary pre-requisite for the goal it still remains to apply the model to the entire country and then aggregate it at some level. Applying the model to all subscribers in the Rwandan dataset took about a day: this was primarily spent computing features for all individuals

Asset	Baseline	Accuracy	AUC
Owns fridge	0.11	0.75	0.88
House electrified	0.60	0.72	0.85
Owns television	0.49	0.73	0.84
Owns bicycle	0.30	0.64	0.68
Owns motorbike/scooter	0.11	0.72	0.67
Owns radio	0.96	0.92	0.50

Table 3.1: Classification results for asset ownership

Wealth strata	Baseline	Accuracy	AUC
Bottom Quartile	0.25	0.79	0.81
Bottom 5%	0.05	0.90	0.72

Table 3.2: Classification results for wealth strata.

since the regression model was very quick and easy to apply. Once estimates for wealth and each asset ownership variable were made we used the predicted home tower for each individual to assign them to a district and then aggregated.

Now it would be possible to compare the district level aggregate estimates to averages for the same district as reported in the 2010 DHS survey. Specifically, we take the average of all mobile-phone owning households (since our own estimates are necessarily predicated on phone ownership) recorded in a district as the target variables. Regressing our predictions versus these survey average finds a very strong relation $r = 0.917$. Our district level predictions can be visually compared to the district survey results in figure 3.4 (A) and (B) respectively. In figure 3.4 (C) we can see the scatterplot comparing the predicted and survey wealth. It is important to note that, in contract to figure 3.2, that even the worst deviation from the regression line is relatively small. This is important as it signifies that the errors that might be made during an individual prediction are largely random and uncorrelated (rather than say a systematic error in how we predict individuals from a certain region).

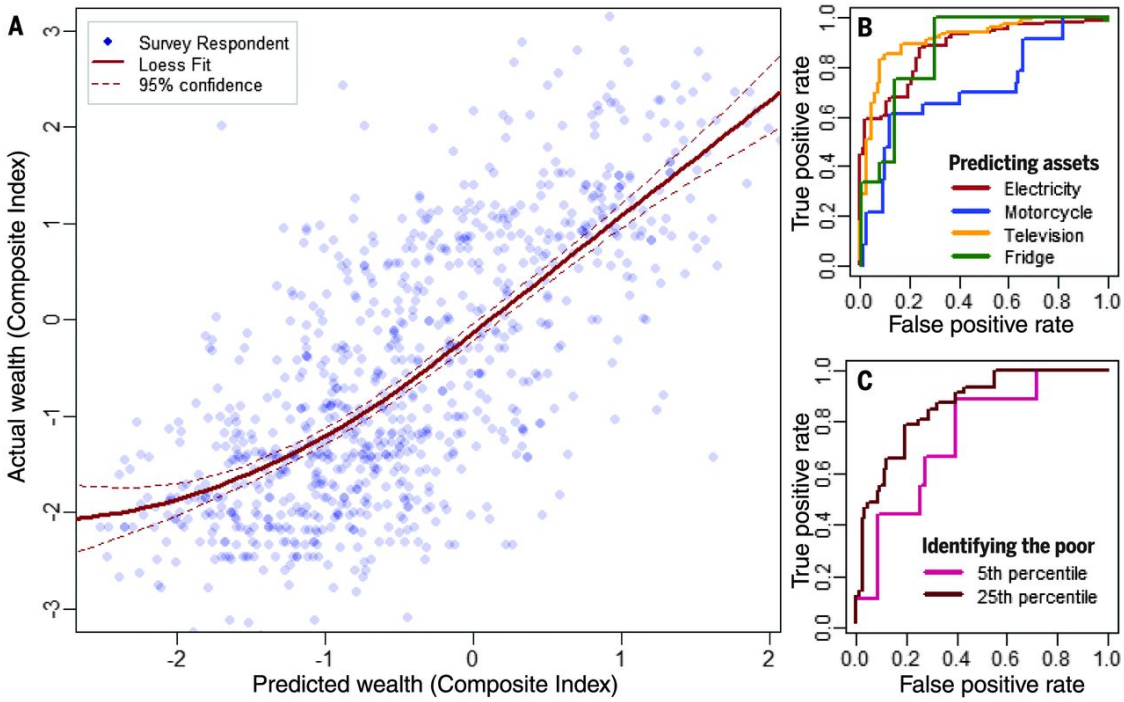


Figure 3.2: (A) Scatterplot comparing the wealth predicted from features versus the actual wealth composite predicted from asset ownership. (B) The ROC curves for the asset ownership logistic regression tasks. (C) The ROC curves for predicting the wealth strata of an individual.

Similarly strong results occur for asset ownership questions: for example our predicted electrification rates also have a strong relation $r = 0.93$ to survey results as seen in 3.4 (D). Note that since we at no time used any DHS survey during the creation and testing of our machine learnt model, this test runs no risk of overfitting.

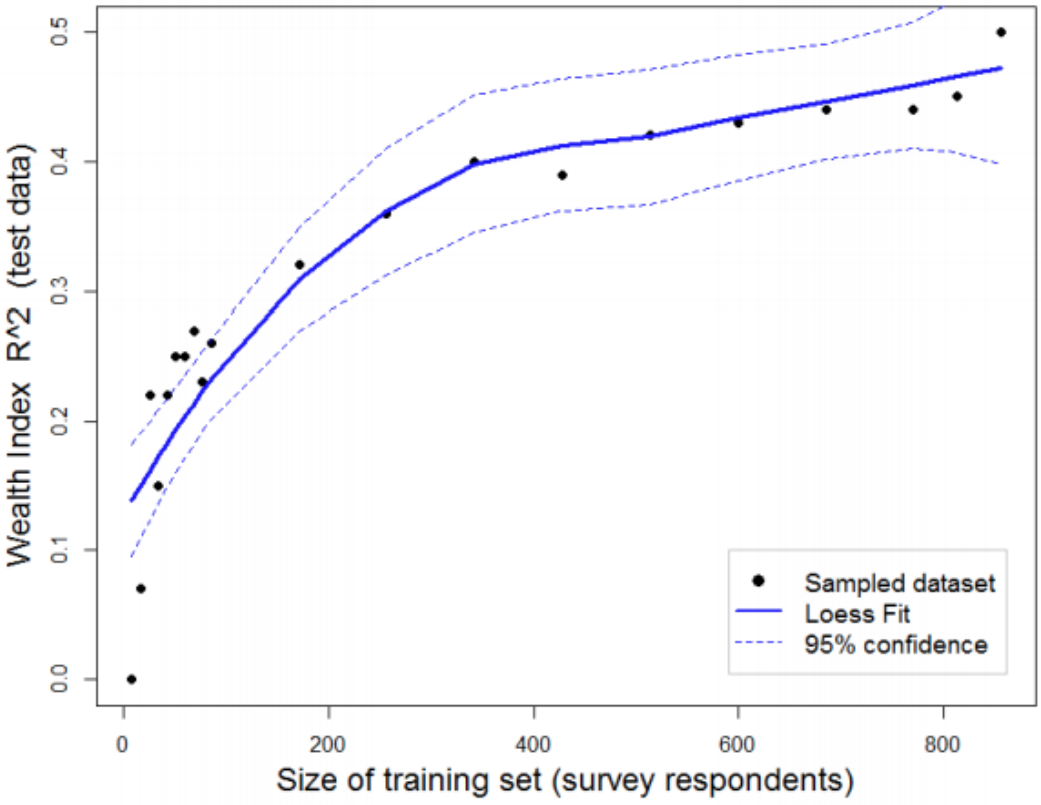


Figure 3.3: Prediction quality as a function of survey size

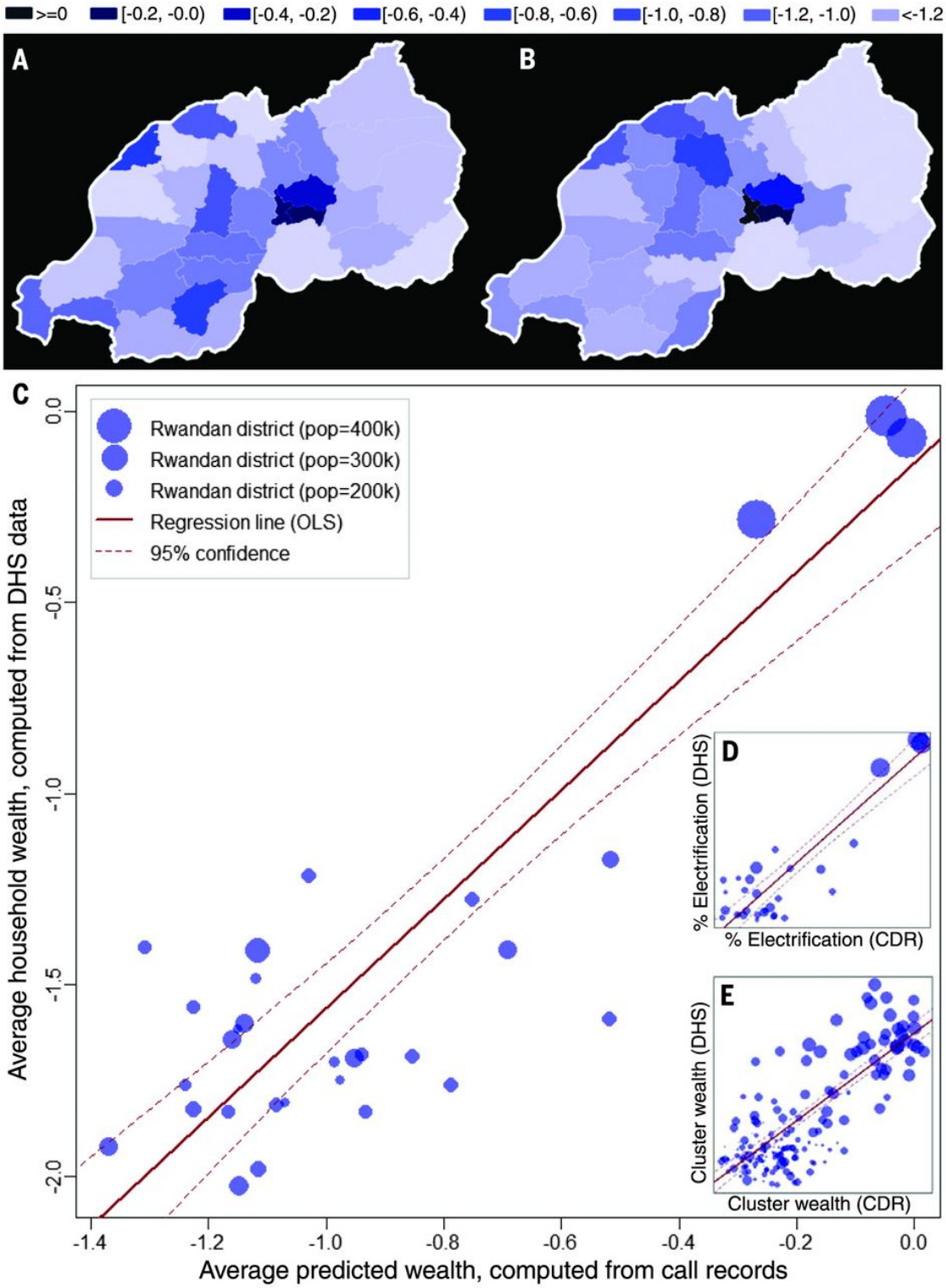


Figure 3.4: Scatterplot comparing predicted district wealth aggregate versus survey derived wealth aggregate.

Chapter 4

CASE STUDY: ESTIMATING ROAD QUALITY THROUGH SATELLITE IMAGERY

While roads are a critical part of economic development for any nation, monitoring the quality and health of such a network is particularly difficult in developing regions. In this chapter we describe a method we developed a model for monitoring the quality of road infrastructure using satellite imagery. Applying a variety of deep learning techniques to this problem enabled us to create a system that provided accurate estimates over a large and varied network of 50 Kenyan roads comprising over 7000 kilometers. In addition to the obvious utility of tracking infrastructure quality, there are many possibilities for econometric analysis enabled by such a system. We explore one such possibility in our work but describe several other exciting ideas in our conclusions.

4.1 Motivation

High-quality roads are among the foremost infrastructure for hastening societal development. Roads enable goods, people, and ideas to travel easily, leading to better equity in service provision, faster economic development, and ultimately, better human outcomes. Though enormous sums are spent on roads – for example, in sub-Saharan Africa, 1.5% of total GDP is spent on roads [199] – funds for road maintenance consistently fall short, a problem arising from an inability to prioritize investments [30]. Understanding the current state of a road network is a fundamental prerequisite for optimal investment of transportation funding. Knowing which roads have fallen into a state of disrepair will inform maintenance crews and seeing which roads are not adequate for certain types of traffic would be important to prioritizing road construction

Being able to monitor road quality, especially in larger countries with diverse and rugged terrain, is a challenging task. Governments in developing countries with oversubscribed budgets for infrastructure can seldom afford to pay for the expensive specialized equipment and carry out this procedure on a regular basis [102]. In urban developing settings, where road usage is heavier and increasingly more people travel with sensor-laden smartphones, crowd sourcing data on road quality is possible [213, 85]. However, but the cheap sensors in phones are incapable of handling continuous acceleration and vibration intensity for more than a few minutes without losing calibration. Moreover, relying exclusively on this would largely exclude roads in rural settings since they are not as conducive to smartphone-based solutions and possibly worsen the gap in urban and rural investment rates.

In this context, using satellite imagery for measuring road quality in rural, resource-constrained settings seems a tempting option. Unlike smartphone-based solutions, remote sensing options can capture imagery in the most remote region as easily as urban centers. Moreover, satellites are persistent and can capture imagery of the same region with a regular cadence: avoiding the cost of constant excursions to all parts of the infrastructure network. Opportunely, a proliferation of satellite companies has resulted in increasingly higher resolution images collected more frequently; in developing regions, this imagery is as high as 30-50cm resolution and some urban areas are imaged on a near-daily basis.

The applicability of Deep Neural Nets to satellite imagery has blossomed in recent years and generated a varied literature as discussed in chapter 2.3. We also note that there has been work in road detection using satellite imagery, which has spawned substantial research [151, 190], competitions [140, 64], and even companies [1]. Though the road detection problem is a related, and possibly complimentary, problem it is distinct from the idea of predicting the quality of an already known road. Interesting new projects about inferring the safety of roads [229] or traffic [124] via deep nets are likewise associated but distinct.

4.2 Processing the Datasets

4.2.1 Road quality though IRI

In order to properly infer the quality of a road from remote sensing data, it is vital to first quantatively define the property of "quality". The road quality measure used throughout our study is called the International Roughness Index (IRI), developed by the World Bank in 1986 [181]. IRI measures cumulative vertical displacement of a vehicle along a stretch of road due to the roughness of the road surface and is typically provided in units of m/km, and is commonly collected using a specialized vehicle with a mounted laser. IRI values can be any positive real number, where higher IRI values imply worse road quality, and typical values fall between 0 and 30. While not explicitly a measurement of road quality, in practice IRI has been found to have very high correlation with user perception of road smoothness.

The dataset of road quality measurements used to train our models consists of IRI measurements conducted at a resolution of 10m along a diverse set of 57 roads throughout Kenya, resulting in samples over a total length of 7000km [115]. A map of the dataset is available in Figure 4.2. This dataset was collected as the result of a partnership between the Kenya National Highway Authority (KenHA) and the Japanese International Cooperation Agency (JICA). Each measurement is tagged with a latitude and longitude ("lat-lon") and a date of survey (during 2013-2015). The roads can vary from tens to several hundred kilometers in length and, as we show in Figure 4.1, span a wide variety of road sizes, terrain types, and land usage. Additionally, IRI measurements are often bucketed into 5 *road quality classes*: great (0-7), good (7-12), fair (12-15), poor (15-20) and bad (20+). Figure 4.1 also shows examples of roads falling into these categories. Roads in our data can also be split into three administrative classes: Class A, linking centers of international importance; Class B, linking national centers within the country; and Class C, linking provincially-important centers. These roads comprise the fabric of Kenya's road transport system, serving as the primary interlinkage between major towns throughout the country.

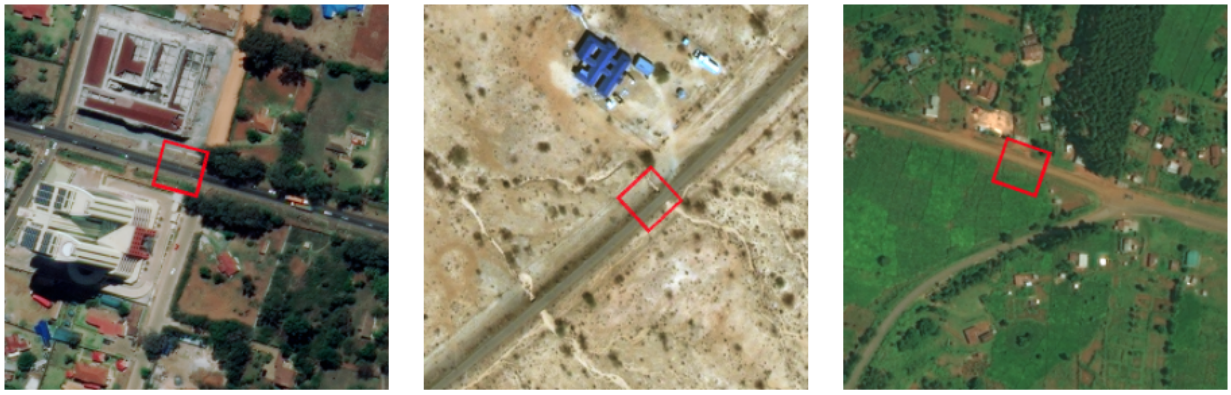


Figure 4.1: These three different roads highlighting the challenging diversity of our dataset. Left: an urban environment along the A104 highway. A104 is a major highway in Kenya and the selected road tile is 'great' quality. Center: the C47 minor road. It passes through an arid environment and the road segment has 'poor' quality. Right: the C67 minor road. It passes through large forests and cropland and the road segment in the image has 'good' quality.

4.2.2 Satellite imagery

The satellite imagery we are using is the DigitalGlobe Basemap+Vivid product [67] and the coverage is the entirety of Kenya. We employ two iterations of this imagery product, each of which is a mosaic of roughly 6300 tiles that forms the illusion of a continuous map by stitching together several images collected at different points in time by multiple different satellites. The +Vivid product is post-processed to account for orthorectification, color correction, and cloud cover, though the latter is still a problem in some remote areas. The first mosaic, compiled in November 2014, consists of imagery from the QuickBird-02 and WorldView-02 satellites, and is composed of tiles with collection dates ranging from 2002 to 2014. The second mosaic, compiled in September, 2017, consists of imagery from the QuickBird-02, GeoEye-1, WorldView-02, and WorldView-03 satellites, and is composed of tiles from 2002 to 2017. The typical resolution of the tiles is roughly 50 cm per pixel, and each of the two image mosaic datasets is $7 - 8TB$.

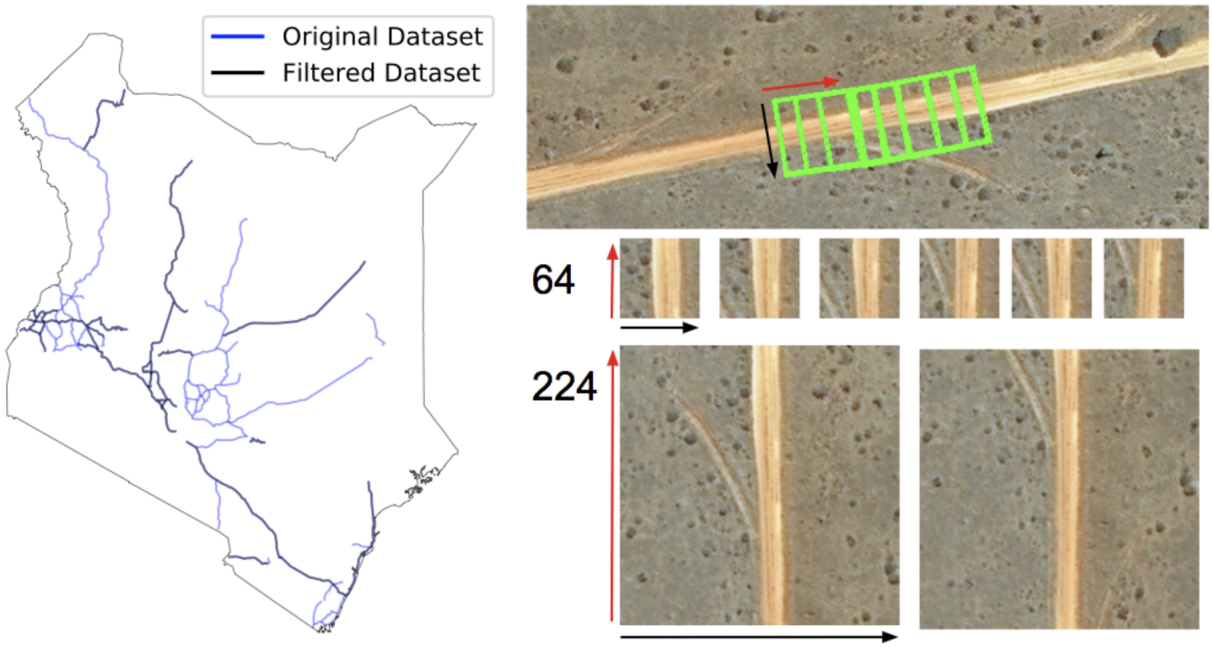


Figure 4.2: Left: Roads with labeled quality data as collected by the Kenya National Highway Authority (KenHA) and our filtered dataset. Right: Example of how a road is segmented. The top segment shows a road divided into overlapping 64x64 squares; this generates the tiles shown in the middle segment. The tiles are always aligned in the direction of the road (red arrow). The bottom shows the same segment if it were divided into 224x224 tiles instead.

Road	Length (km)	Bad (%)	Poor (%)	Fair (%)	Good (%)	Great (%)
A104	269.2	2.2	1.1	1.5	6.3	**88.9**
A109	86.54	0	0	1.9	10.5	**86.6**
A23	10.31	**44.9**	17.9	12.6	24.2	0.3
B8	47.56	3.4	9.8	18.3	**50.5**	17.9
B9	16.64	13.6	**31.9**	21.3	31.1	2
C31	39.33	0	0	0	1.5	**98.5**
C32	30.03	**61.7**	20.5	8.8	7.9	1.2
C33	44.79	0.1	1.0	2.4	19.3	**76.2**
C36	23.05	11.2	9.5	8.3	21.5	**49.6**
C42	40.88	31.9	13.5	6.7	9.9	**38.0**
C47	104.92	**40**.3	35.8	14.8	8.9	2.4
C51	40.89	2.1	3.8	4.9	16.3	**72.9**
C54	27.62	4.4	1	2.8	13.9	**77.8**
C67	37.86	**90.3**	4.4	3.1	2.3	0
C68	16.94	**56.3**	20	15.1	8.6	0
C69	95.07	0	0.1	1.7	26.1	**72.0**
C76	41.89	**79.5**	16	4.4	0	0
C77	110.40	22.2	16.6	21.8	**36.1**	3.2
C78	28.71	16	20	17	**39.3**	8
C83	21.38	**100**	0	0	0	0
C96	19.07	5	23	29	**39**	4
All	1153	19.2	9.5	7.6	16	**47.7**

Table 4.1: A summary of the diverse set of roads in our labeled and filtered data set recording both the length and distribution of road quality labels. For each road, the modal road quality class is in bold. The set ranges from first-class highways (*e.g.*, A104) to rough dirt roads (*e.g.*, C67) and includes roads with significant internal variation (*e.g.*, C77).

4.2.3 Preparing the Training Data

While the satellite imagery and road quality datasets are both impressively large, the wide range of dates covered by the tiles of each mosaic coupled with the range of dates of the IRI measurements creates a mismatch. This issue appears often when learning on satellite imagery [117], but is particularly acute in our scenario since road quality can experience sudden and potentially substantial changes (*i.e.*, due to weather or construction) in a way that only a serious emergency may impact other attributes commonly predicted via satellite imagery (like wealth). In an ideal data-collection scenario, the maximum time period requirement would be a month or even a week. However, given the reduced frequency of data collection in developing regions, this is untenable. Ultimately, to ensure that imagery reasonably matches the condition on the ground when the IRI sample was collected, we decided to restrict our label dataset to only those samples where the difference between the two dates was 12 months or less. Selecting any period of time shorter than 1 year for the maximal time discrepancy would have significantly decreased the amount of labeled data. This left us with a tradeoff between not having enough data to properly train a deep net and possibly having some incorrect labels. Our ideal scenario of using data no more than one month out of date would have left us with only 340 kilometers of road and 40% of the unique roads in our 1-year set. Given that anything more than three months already entails a possible shift between seasons it was decided that the extra data provided by a maximal discrepancy of 1 year was tolerable. This design decision results in a subset of the samples from the larger IRI dataset used for training; this subset consists of samples covering 1153km over 21 roads, as detailed in Table 4.1, which also includes a breakdown of the classes of labels for each road. Additionally, a map of the filtered dataset is available in Figure 4.2. This set of roads includes paved and dirt roads, consistently high-quality and consistently low-quality roads, and roads with high variability in quality.

Having chosen which roads will make it into our training set we then needed to decide upon the training set specifics: defining the fundamental data point and objective function.

Regarding the former, we call this fundamental unit a *patch*, and define it as a quadrilateral such that the length of the patch is parallel to the course of the road and the width is perpendicular as seen in Figure 4.2. Our IRI measurements are at intervals of approximately 10 meters (20 pixels in our imagery data) so possibilities for length are bounded below by 20 pixels. Arguments for a smaller patch size include greater granularity and the road forming a greater proportion of the patch's area. However patches of lesser dimensions may sometimes not include some or all of the road due to random noise in the latitude-longitude pairs associated with IRI values. We settled on a compromise of 64x64 pixels which was robust enough to account for this noise and also neatly covers 3 IRI measurements per patch.

Regarding the latter choice, we decided the objective function would be the real-valued IRI itself. As such, the project becomes a regression problem and will record results in terms of the mean square error (MSE) and the R^2 coefficient. MSE gives us an absolute averaged error while R^2 explains how much of the total variance in the IRI is explained by our prediction. Instead of directly using the IRI values y_i^*, we establish a maximum threshold $T = 30$ and train on the labels $y_i = \frac{min(y_i^*, T)}{T}$. We feel this is justified since anything above an IRI of 20 is already bad and the visual/practical difference between a tile of IRI 30 and another with IRI 40 is minimal. Note that since any predicted IRI value can easily be mapped to a prediction of one of the five *road quality classes* we can also measure the accuracy of our predictions if they were used for a classification task as opposed to a regression. We report this accuracy on the inferred five classes throughout our results for comparative purposes even though we do not at any time train for classification accuracy.

4.3 Machine Learning Methodology

4.3.1 Training and Test Splits

One aspect particular to our application is that the question of how to train and evaluate a model is not as intuitive as making a random train-test split. Since the training data has a sequential nature, randomly splitting the data into train and test sets would result

in cases where patches that appear next to each other in the satellite imagery might be in both the training and test set. In addition to the problem of test data contamination, this testing scenario would be very different from the use case that we are targeting where one would seek to predict on an entirely unseen road. As such, we devise two more appropriate methods of generating training and test sets. The first is done by splitting the entire set into 1-kilometer long "runs", which are then randomly assigned to the train or test set with proportion 70%-30% – we call this the *standard* method since it more closely resembles the random train-test split. The second method is to assign an entire road to the test set and the remaining 20 roads to the train set and average the result over the 21 possible splits (one with each road held out): we call this the *held-out* split procedure. Though this very closely approximates a real application we note that this method breaks an often central assumption of machine learning methods: that the train and test sets are drawn from the same distribution. As the held-out problem is much harder to predict, results reported using the *held-out* methodology are significantly worse than those reported using the *standard* methodology. However, *held-out* predictions are potentially more impactful, as results can generalize to unseen contexts. We recorded results for both methods throughout the work for completeness.

4.3.2 Learning Algorithms

The intuitive starting point for this problem was to train a Convolutional Neural Net that would predict the quality of each road *patch*. Thus, we began with Resnet, AlexNet and VGG-11 [103, 131, 187] as initial network structures and then simply replaced the last layer of fully connected layer nodes with a single sigmoid function instead. We then trained using our 64x64 tiles scaled to 224x224 pixels. All the CNNs were trained over 10 epochs of the data, augmented by random horizontal and vertical flips, and completed within a few hours when trained on a GPU cluster. We found that after around 10 epochs, the training loss was roughly flat, and continuing to train would likely only result in overfitting. This was accomplished on a GPU cluster using pyTorch, taking about two hours to train each model.

These led to our first set of results as shown in table 4.2.

While our labeled dataset is already fairly large, it should be noted that this only represents 15% of the total of the roads in our data-set; the remainder had to be discarded since the labels might be out of date. However auto-encoders provide an alternative to supervised CNNs that allow us to leverage that large set without relying on the labels. Convolutional auto-encoders consist of two parts: an encoder which compresses the images down to k features and a decoder which attempts to reverse the encoding back to the original image. Training this network to attempt to recreate the original image as closely as possible should ideally lead to a k-dimensional representation of the image that preserves as much information as possible. We can leverage this by training the auto-encoder over the larger, complete set of roads to learn a very efficient representation of any given tile and then doing an L2-regularized regression of these features on our training set. We perform this with a 2-convolution auto-encoder with $k = 1000$ alongside retraining the aforementioned CNNs. The auto-encoder was trained overnight on the unlabelled data set first for 20 epochs and then simply regressed with an L2 penalty. Though training the auto-encoder itself is time-consuming, this is only a one-time task and it can be later used to quickly featurize any road. The performance of the autoencoder model is comapred to the CNNs in table 4.2.

Another avenue for exploration is how much the sequential structure of roads can be leveraged to more accurately predict road quality. In the simplest sense we can keep the same fundamental aim of predicting y_i but instead of using only the tile image $\mathbf{x_i}$, one can use the last s $\{\mathbf{x_j} \| i - s < j \leq i\}$. Slightly more complex would be the case where we use the same segments of satellite imagery but attempt to instead predict the average IRI of the entire segment $\bar{y}_i = \sum_{j=i-s}^{i} y_i$.

We handle both of these cases by first using the auto-encoder to featurize all the roads and grouping them into contiguous sequences of length s. We will use the same simple 1-layer LSTM with 500 internal nodes, changing only the objective to optimize between the two. The LSTM is then trained with L2 regularization to prevent overfitting and we record the *held-out* results in table 4.3.

	5-class accuracy		Regression R^2	
Base net	Standard	Held-out	1-KM Split	Held-out
Resnet	0.69	0.44	0.79	0.24
VGG-11	0.71	0.47	0.78	0.26
AlexNet	0.73	0.49	0.66	0.21
Auto-encoder	0.65	0.41	0.78	0.31

Table 4.2: 5-class accuracy and regression R-squared results under *standard* train-test and *held-out* conditions for the single-tile regression problem.

4.4 Evaluation and Application

4.4.1 Evaluating the Models

One immediate observation that can be made from scanning table 4.2 is that the more realistic *held-out* test case is significantly harder than the *standard* scenario. However the results are encouraging given that achieving even the baseline result (accuracy of 0.20 and an R^2 of 0.0) is not guaranteed when the train and test sets are entirely different. This provides evidence that the problem is indeed approachable using standard machine learning techniques. The second observation we wish to highlight is that accuracy is fairly good for a 5-class problem even though we at no point directly optimize for accuracy. This is a consequence that estimating the IRI value fairly accurately will translate to a correct estimate of the road quality class and seems to bolster the idea that directly regressing the IRI and then transforming it to less granular measures as the application calls for it is a viable idea. We note that measuring accuracy does not distinguish between one error mistaking a "fair" tile for a "good" tile and another error mistaking a "poor" tile for an "excellent" tile, thus potentially understating the predictive quality of the CNNs.

We find that there is not much to separate between the different CNN classes in terms of performance. Similar initial learning rates and dropout rates were used in all, though VGG-11 had more problems with overfitting compared to the other two and did not move

	5-class accuracy		Regression R^2	
Sequence length	Last	Mean	Last	Held-out
1	0.41	0.41	0.31	0.31
10	0.42	0.43	0.34	0.35
25	0.43	0.43	0.35	0.32

Table 4.3: Results for LSTMs in the *held-out* test scenario as a function of the length of sequence trained on. Regressing on the final tile value (last) are compared to regressing on the average tile value (mean).

beyond transfer learning. In terms of the key regression metric, we found that the auto-encoder regression outperformed the other CNN methods. This is likely due to its superior generalization performance on unseen roads. In this we find a notable advantage of being able to leverage the entire set of roads; the feature representation from the auto-encoder was much more stable in the *held-out* scenario than it was on the other CNNs. This was likely due to the fact that the auto-encoder could look at a much more diverse set of roads to determine how to featurize a patch, as opposed to the only 21 roads that a CNN could use.

An important caveat is that the overall final R-squared is not the only evaluative measure that should be considered. The train and test sets are highly hetereogenous, with some roads being very different from others while some roads may just be inherently harder to predict on. One of the advantages of the autoencoder is that it has less spread in its predictive quality compared to a CNN. Figure 4.3 illustrates this by comparing the results on individual roads for Alexnet to those of the auto-encoder regression. In addition to better overall performance, the auto-encoder has fewer roads with very high error: this is important for real-world application as we would like to be confident of some guarantee of predictive power.

The results for the LSTMs are summarized in Table 4.3. These show a modest improvement over the non-sequential LSTM, although these initial experiments do not seem to suggest much improvement when increasing the sequence length over 10. However, this is likely influenced by the well-known difficulties of training on longer sequence LSTMs and

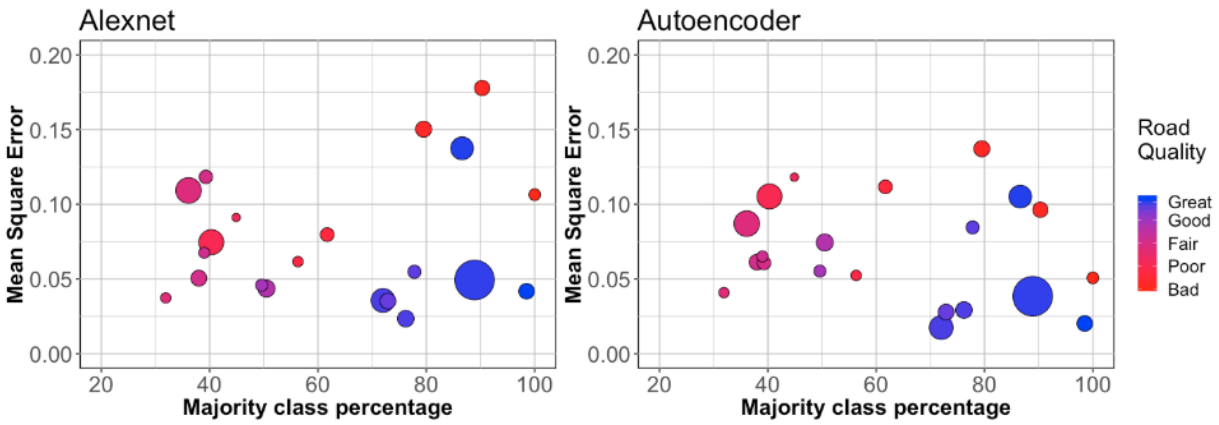


Figure 4.3: Figures showing the distribution of Mean square errors (y-axis, lower indicates better predictive power) of different roads using a Resnet CNN (left) and auto-encoder regression (right). The x-axis is a measure of the heterogeneity of the road, the color provides the average road quality, and the circle size indicates the relative sizes of the roads. Comparisons to VGG-11 and Resnet yield similar results.

may not reflect the actual limit of this technique. Further investigation in both the structure of the LSTM as well as its training will be important

4.4.2 Case Study: Application to an Econometric Analysis

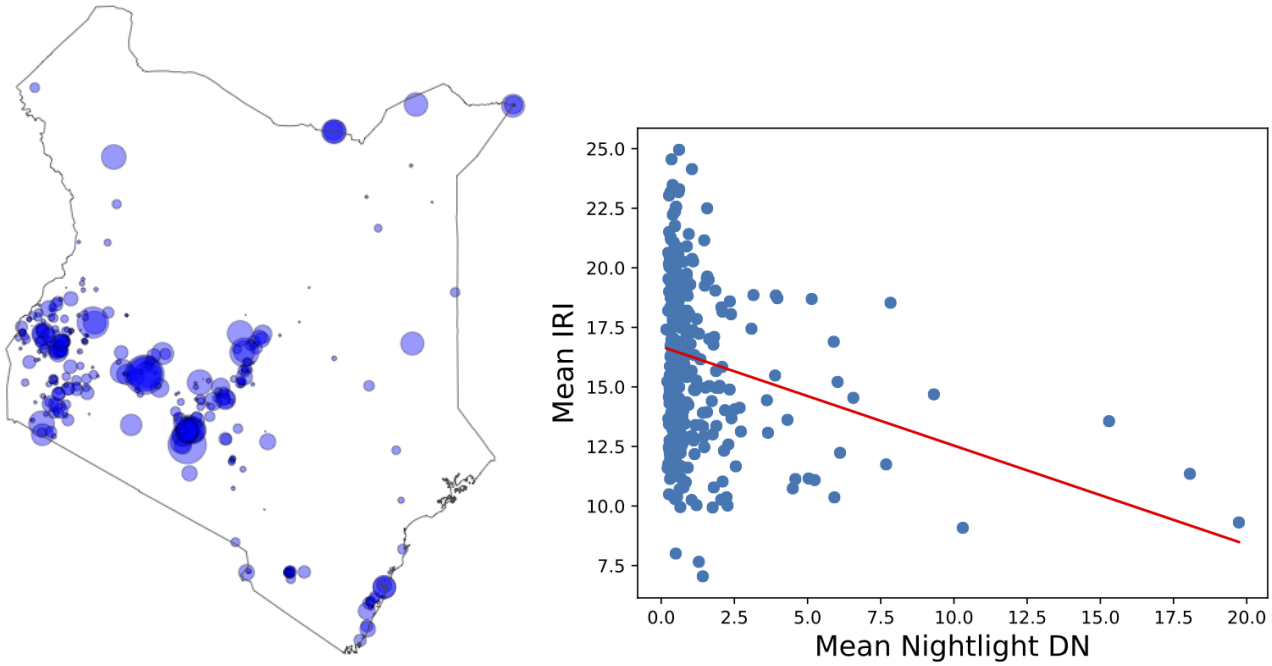
We felt confident that these results characterized a new measurement tool with strong accuracy, high resolution, and applicability in a wide array of terrain types, but we wanted to explore possibilities beyond prediction. There are without a doubt many applications for being able to simply return accurate road predictions. For example, it could enable a country with limited resources to scan its entire network all at once and locate weakness. In another scenario, if combined with recent satellite imagery, it could be used to quantify the damage to infrastructure caused by a natural disaster. However, we felt there was an entire avenue of investigation relating to the econometric impact of road quality that should be highlighted. So, as a test case, we attempted to test the relationship between high-quality roads and the prosperity of the towns that they connect, performed at the scale of an entire country.

For measuring economic conditions, we employ a dataset on nighttime light illumination collected in an imaging product called the Visible Infrared Imaging Radiometer Suite (VIIRS) Day-Night Band [157]. The data used to create this imaging product are collected via a daily fly-by of the NASA/NOAA Suomi National Polar-Orbiting Partnership; results are post-processed to create monthly data on illumination level for every pixel of the earth's surface, where each pixel is roughly 450m x 450m and data are available for each month since April 2012 [158]. The pixel data range in value from -1.4011 to 32641.72, with higher values representing more illumination.

A wide range of studies have employed these data for granular measurements in developing regions, including measuring total economic activity [104], the incidence of poverty [117], electrification [13], electricity infrastructure corridors [157], and many others. In our case, we use the nightlights data as a generic proxy for economic activity; through this simplified lens, we can estimate relative differences in economic activity both spatially and temporally.

For our analysis, we choose to compare quality of intercity roads and the economic activity of major towns (as measured via nightlights) throughout Kenya. Using data from Open Street Maps (OSM), we select all OSM-named "places" in Kenya with a population above 5000 within 2km ($n = 924$). Then, for each of these places, we select all roads incident upon each place with length greater than 5km; this targets intercity roads. For this analysis, we exclude any places in Nairobi County, which is nearly continuously urban. To only consider more highly-connected places and to expedite computation, we limit our road quality analysis to exclude places that have fewer than 4 segments entering town and we restrict those segments to only the final 10km entering each town. Finally, we randomly selected half of the remaining places, leaving us with a dataset of 322 places and 11,376km of associated roadway. Figure 4.4(a) shows a map of the town locations in Kenya, with each location marker sized according to its population.

Figure 4.4(b) shows a comparison of road quality and nighttime illumination for each of these towns. For each roads, we estimate road quality by selecting 10 meter patches of imagery and applying an autoencoder and a fully-connected layer for making the predictions,



(a) Towns in Kenya examined in our case study (b) Comparison of Road Quality and Nighttime Illumination for each town

Figure 4.4: Comparison of mean intercity road quality (measured by International Roughness Index) and mean recent 6-month nighttime illumination (measured as a "Digital Number") for a sample of towns in Kenya ($n = 322$). For (a), place marker size represents relative population size, and for (b), a line of best fit is plotted ($R^2 = -0.25$). Note that lower values of IRI correspond to better road quality.

as discussed in Section 4.3. This provides estimates of IRI for each patch, which we aggregate to produce the mean and standard deviation of IRI values for each road segment. We combine these mean IRI values to quantify the mean road quality for each place weighted by the length of each road (capped at 10 km). For nighttime illumination, we use the mean of the most recent six months of data (June to December 2018). From Figure 4.4(b), we can see a weak relationship between road quality and nighttime illumination (recall that lower values of IRI correspond to better road quality). While the correlation is weak ($R^2 = -0.25$), the relationship is robust to removal of places with low illumination (*i.e.*, Nightlight $DN < 1.0$) or places with high illumination ($DN > 10.0$). While it is not at all expected that road

quality would be the sole determinant of local economic activity, we believe that these results do show a level of correlation that bears further examination.

We acknowledge a number of other shortcomings of this approach. For one, nightlight illumination is only an approximate proxy for local economic activity, especially as many rural areas have very low levels of nightlights, satellite flyover times are often in the middle of the night, and an increasing number of locations are being electrified by more efficient and decentralized electricity sources that may produce less ambient and wasted light. Additionally, not considering the entirety of the entering roads may poorly reflect the actual ability for goods, services, and people to enter each town. Last, the question of mismatched timing between when images were taken and when nightlight values were recorded contributes some disparity. Last, we acknowledge the accuracy challenges that our models have for predicting quality for individual patches. While this is of concern, we believe this is somewhat mitigated by the consideration that most use cases we envision are focused on the quality of segments of roads rather than that of individual patches; for segments, our approach should provide suitable accuracy. On the other hand, for applications that require high patch-by-patch accuracy like pothole detection, our methods may be insufficient.

Despite these challenges, we believe that this case study shows that the measurement technique that we have developed enables examination of road quality throughout an entire country, an unprecedented scale. Beyond the valuable purposes of improving investments in construction and maintenance of roads, we believe that our work presents the opportunity to correlate road quality to many other societal development indicators. This includes access to other infrastructure or services (*e.g.*, water resources, cellular towers, financial services, education and health facilities, markets, cropland, *etc.*), household survey responses (*e.g.*, education and health indicators, wealth measurements, perspectives on governance, *etc.*), and other remote sensing data (*e.g.*, weather and climate information). While our work represents only a beginning, it can enable a wide range of studies on the impact of road quality on different aspects of societal development.

4.5 Conclusions

In this work we describe a methodology to infer the quality of intercity roads in developing regions, with the primary goal of enabling useful and practical applications. To do this, we trained models using satellite data and road roughness data from Kenya and demonstrated that the models performed well in some cases in locations previously unseen, while remaining cognizant of remaining challenges. We saw that while the normal train-test paradigm can be approached readily, achieving reliable results on the held-out case is significantly harder. We also demonstrated a novel use case of our road quality measurement at a larger scale than traditional methods would feasibly allow.

From a technical standpoint there are several exciting improvement we could make to improve predictive accuracy. Athough auto-encoders are one way to leverage a large amount of unsupervised data: a recent work suggests a new idea based on spatial co-location that might be even better suited to satellite based tasks [118]. Even more parimonious might be to directly merge the unsupervised and sequential improvements made into one system by using recurrent autoencoder[222].

We feel that the potential applications are far-ranging and important. We have already talked the possibility of econometric analysis but it should be highlighted that satellite imagery is particularly useful for this since it might be one of the only ways to estimate quality of roads in the past that have since been improved or changed. Such an ability might be vital in the case of an analysis when the quality of the road was not recorded for all periods in the analysis. In addition to econometric analysis, simply being able to infer, even roughly, the quality of an entire road network in a matter hours could serve as a great starting point for various infrastructure projects.

Chapter 5

MEASURING AND ATTRIBUTING THE IMPACTS OF EMERGENCIES

During any emergency situation, be it an outbreak of violence or a natural disaster, reliable and timely reporting is critical to mitigate damage and aid recovery. Understanding the conditions during an emergency event will help allocate immediate relief to those most affected while post-event monitoring is important to understand exactly what the long-term impacts are and how to target effective treatment. Such monitoring is often a complicated and multi-modal operation in which several different types of data sources have to be considered. For example, a flood may cause significant physical damage that can be estimated via weather reports or satellite imagery but the long-term consequences on political behavior and unrest would have to be monitored via surveys or social data[81]. In this chapter we draw together two lines of work which we seek to combine in chapters 6 and 7. The first is measuring a response to emergencies using passively collected data. The second is the literature on causal inference (the ability to attribute an outcome to some given cause) with a particular focus to seeing how it can be applied to natural disasters and violence.

5.1 Tracking the Impacts of Emergencies

Many developed countries will have specialized departments for various types of emergency events (such as FEMA in the United States) with integrated detection and evaluation infrastructures for emergency events. However, in developing countries such infrastructure is often not present. Moreover, situations in these countries such as poor transport infrastructure or endemic political instability may limit the applicability of direct human monitoring and intervention. For these reasons, significant utility has been found in the use of large-scale

passively collected data for the monitoring of emergency situations in these regions.

CDRs are well suited to this tracking task since they are able to record events at very high temporal granularity. Several initial papers, making use of the outlier detection literature, developed methods to detect the statistical aberration in the number of calls made through a tower[224, 125, 69, 101]. In these methods, the amount of aberration would relate to the scale of the disasters and could be validated against post-hoc damage assessment[125] or via detecting other unrelated events[69, 101]. For some of the papers, significant differences from an expected "normal" could be seen for several days after the event[125]. Though most of these methods concern themselves with call volumes, one[69] was also able to use detect elevated movement rates during the days following an earthquake.

Interesting developments occur when focusing on smaller and more localized events with correspondingly weaker signal. One attempt to detect and classify smaller events such as concerts and bombings noted that some parts of the social network appeared to be more sensitive to emergency events[18]. Two other projects on the same dataset noted significant differences in the reaction between different types of emergencies (for example, comparing a plane crash to a bombing to a small earthquake)[24, 139]. This included both simple statistics like how many extra calls were made as well as more complex phenomena such as how information about an emergency propagated through the system. In all of these cases the analysis was constrained to only a few hours after the events: a concession to the fact that the signal for these smaller events can fade quickly.

Remote sensing also offers a strong ability to track emergencies though on a significantly different range of events than CDRs. Forest fires are an ideal case for remote sensing, real-time detection and tracking can be provided by UAVs[53] or geostationary satellites[221] (which are constantly over the same area). Tracking disruptive large natural disasters are an intuitive use for satellite imagery. In addition to forest fire, flooding severity can also be assessed and localized through simple pixel analysis[81]. Harder, from a remote sensing perspectives, are periods of violence. Events such as bombings or armed conflict show up clearly in CDRs, but satellite analysis of such events are focused on very specific material

outcomes[52].

Though we have already noted the problems with coverage that using internet-based sources encounters in developing countries, it is still informative to report on their progress. Using social media to inform the immediate response to disasters has in fact been extensively studied[110] and in this we spot several contrasts with CDR methods. In emergencies the reaction on a rough patchwork of different platforms (e.g Twitter, GoFundMe, Facebook) can be analyzed holistically[61] as opposed to relying on a single network with fixed metadata. Moreover, the ability to actually read message unlocks a whole new avenue of advanced analysis unavailable on simple metadata. This can range from looking at keywords[133] or automated sentiment analysis as a useful signal during emergencies[27] to more involved study of phenomena such as a rumor propagation[227]. Another interesting line of though is the idea that credit card transaction data can likewise detects emergencies and disruptions. This is accomplished both in ways similar to CDR data (by acting as a proxy for mobility) but also by the type of items purchased[72].

5.1.1 Persistent and Meaningful Changes

While the aforementioned works might work well as to track emergencies and measure their severity in different regions they lack relevance to socio-economic study. Learning something important about how emergency events impact society requires two conditions to hold. The first is that the quantity being measured should have some insight into social or economic behaviour. For example, tracking abnormally high call volume during an earthquake is great for understanding where people feel endangered but does not reveal anything new: it is well known people will call more when faced with danger. The second condition is that the impact should persist across time: thus providing an example of long-term impact that may permanently damage social or economic outcomes.

One example of the former is an innovative study of the Haiti earthquake was able to predict the migration patterns of individuals impacted by the earthquake for three weeks after the earthquake: leveraging the fact that the CDR including spatial information that

overlapped with a holiday preceding the earthquake[145]. The idea of tracking migration in the aftermath of a disaster has been very influential and a real-time analysis of CDR data after the 2015 Nepali earthquake of significant use to humanitarian agencies[218]. In this case, the inference about how people reacted in an emergency had immediate socially beneficial consequences.

A work that follows a long-term change in meaningful metrics is a study of the impact of a factory closure (though this is stretching the definition of an emergency)[201]. This work used a set of heuristics that managed to separate the residents of a town into those who suddenly became unemployed and a control set that didn't. It was then noted that many metrics, including those related to mobility and social network entropy, could be seen to suddenly drop and stay noticeably depressed over a period of several months after the closure. Another work analyzed the long-term impacts of violence in Afghanistan on commerce by specifically looking at enterprise calling records[39]. This work showed that during a major escalation in violence, such as the 2015 Battle of Kunduz, firms and their employees left affected areas and that this absence can persist for up to six months after the violence ends.

5.2 Causal Inference and Emergencies

Measuring and tracking changes in metrics occurring in the aftermath of an event is certainly valuable, but being able to attribute them with statistical certainty to the emergency represents an even higher contribution. This requires not only seeing a signal but being able to distinguish it from artifacts related to locality, other co-occurring trends, periodicity effects etc. The methods developed to study this are known as *causal inference* and is a well studied problem with many different ideas to establish a *counterfactual* to compare to measurements from a *treatment set*[210].

We note that causal inference has been an issue of paramount importance when studying socio-economic questions in developing regions more broadly. Most often we see this because econometrists are interested in understanding the true driver of a change in some important trend. Sometimes, this can be a positive trend and helps governments understand how to

plan policy; an example is Duflo's work quantifying how Indonesia's investment in school building led to better education levels and wage growth[74]. More pertinently to the study of emergencies, sometimes the study will focus on negative outcomes such as the study quantifying (also based in Indonesia) how much lack of government auditing increased project costs via corruption[160].

Most of these seminal papers in developmental economics have a few central similarities. These works generally draw both the dependent variable and the independent variable from traditional data sources such as government conducted surveys, census data or measurements conducted by the economists themselves. This leads to trustworthy data with low error rates but limits the data in terms of granularity (e.g: when relying on a census) or scope (e.g: if relying on specifically collected measurements). In this sense, using large passively-collected data (or inferences made on such a dataset) to generate the independent and/or dependent variables represents a possible solution to these problems. In fact, satellite imagery has long been used by economists to augment traditional datasets[71]. One study looking at the impact of air pollution caused by forest fires in Indonesia on early-life mortality was able to use satellite imagery (along with other sensors) to quantify the independent variable of air quality[116]. Also using satellite imagery to determine the independent variable was an examination of how the severity of flooding in Pakistan contributed to political mobilization[81]. Less common is using passively collected data to measure the independent variable, which is normally of some economic or sociological importance. The aforementioned study of violence's impact on business in Afghanistan did this by tracking the spatio-temporal behaviour of corporate cell accounts and quantifying this into internal migration rates[39].

Once an inference problem has been decided on and the independent, dependent and confounding variables defined the methodology for determining the causal effect must be selected. The gold standard is of course a controlled experiment in which a treatment and control set are picked with all confounding variables controlled for except for the independent variable on interest. While still a relevant paradigm for marketing studies or small scale

sociological experiments, such an approach is ill-suited and impractical for large studies in developing regions. More common in this literature is to look for a "natural experiment", build a regression with the help of an instrumental variable or by looking for regression discontinuities[23]. The latter looks for situations when a treatment is applied based on an exact cut-off: the intuition being that observations on either side look very similar in terms of confounders. Classic examples are educational outcomes defined by a minimum score on a test[113] or an age threshold[50].

In cases where this is not possible, another possibility is to find or generate a counterfactual to compare against a treatment directly. This can be done using machine learning, carefully calibrated to ensure that the resultant model will indeed correctly capture counterfactual behaviour for some period of time[210]. Synthetic controls such as these are a recent innovation and have been used for analysis of questions such as how government intervention impacted the sale of Tobacco in California[9]. Expanding this same approach to longitudinal/panel using a matrix completion approach has also recently been suggested[22]. Moreover, if there is a large selection of untreated sets to examine a researcher could check to see if these sets have some key properties that would make them ideal counterfactual sets [3].

Chapter 6

CASE STUDY: QUANTIFYING AND MITIGATING STATISTICAL BIAS DURING EMERGENCIES

Large passively collected data sets have been used to great effect for the detection and in-depth analysis of emergencies and violence in developing countries. However, researchers must always be mindful of the fact that transactional data may have statistical issues that would not be found in more traditional data sets such as a well designed survey. Even a small statistical bias incurred while collecting the underlying data may make any subsequent analysis, no matter how well-designed, report errant results. In the course of working on such an analysis of violence in Afghanistan, we discovered that the danger of statistical bias is considerably amplified by the burst of activity that inevitably follows emergency events. We determined that studying this phenomena and understanding how to mitigate its impact would be pre-requisite to a confident analysis of the violence data. As such, we not only quantified the danger of statistical bias through a set of natural and synthetic experiments, but also created a simple yet highly efficient heuristic to counteract the bias.

6.1 Motivation

This project developed out of the groundwork done for countrywide analysis of the impact of violence on social cohesion in Afghanistan (which is discussed in its entirety in chapter 7. More specifically, we were looking at a CDR that covered most of Afghanistan and were investigating the responses of individuals who we perceived to be impacted by a series of bombings during the spring and summer of 2015. The idea here would be to look at some set of individuals, pick a socially important *metric* and analyze the change in this *metric* for the week immediately after the event when compares to its value for the week immediately

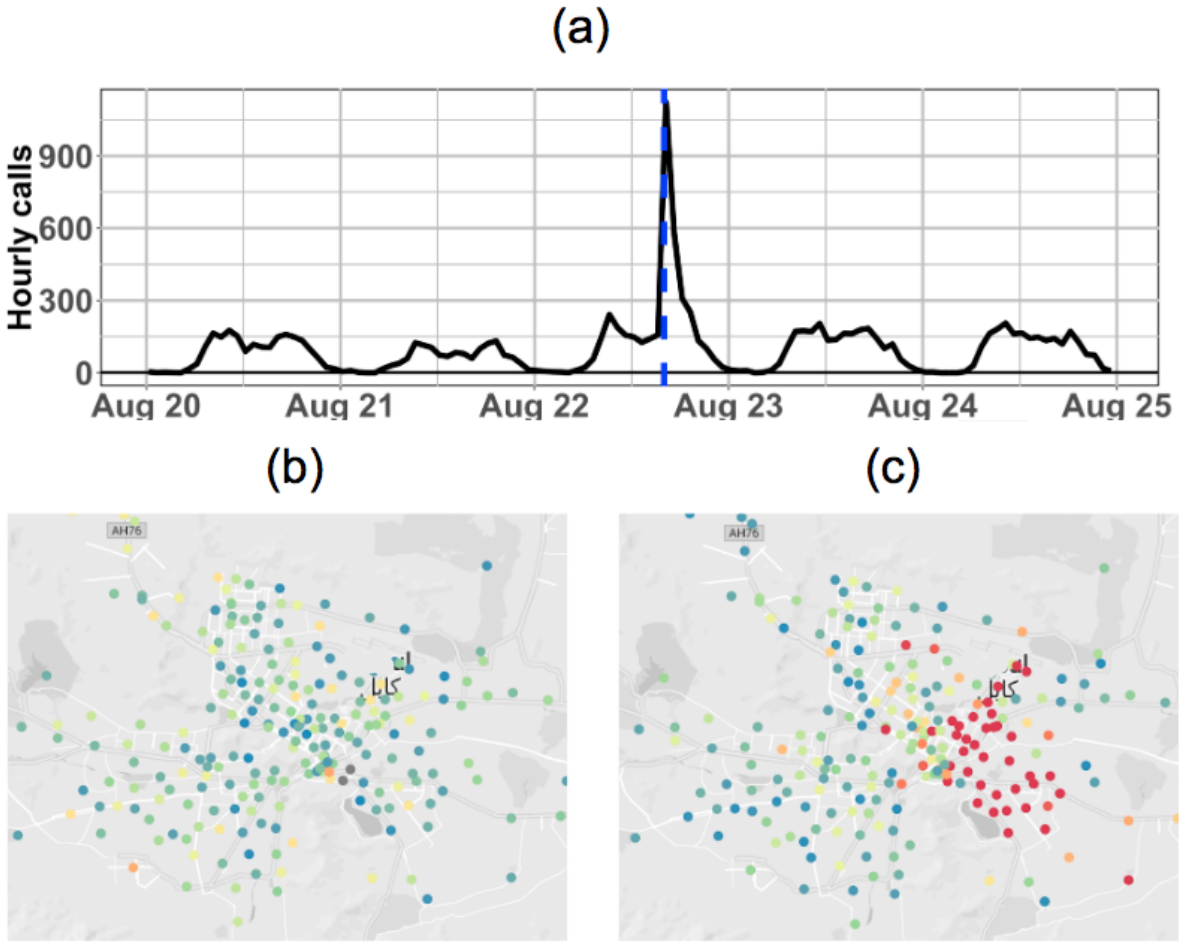


Figure 6.1: Graph (a) shows how the hourly call volume of one of the impacted cell towers experiences a very noticeable surge during the emergency. This is also apparent when mapping the tower-level call volume (b) one hour before and (c) one hour after the bombing. The color of each tower represents how abnormally high the call volume is: with red representing call volumes over 5 standard deviations from the mean.

before the event. Being able to show, with a high level of statistical confidence, that socially important attributes (such as the breadth of someone's social network) changed drastically after being exposed to violence would help researchers understand and quantify the corrosive effect of violence on civil society.

We have briefly discussed the idea of *metrics* in 2.1, but it is important to focus on those which are of interest to sociologists and economists. A common metric to study for example is call volume. While the bombing events we study dramatically raise the average call volume (as seen in figure 6.1) one could draw little insight from this. We instead decided to focus on two broad classes of metrics.

The first set of metrics treat the CDR as a weighted graph and summarize network structure. Specifically, we consider *network degree* (which captures the number of unique connections of each node in the network, also called degree centrality) and *network entropy* (a measure of the dispersion of each individual's network). For any graph, let the number of interactions between node i and node j during a given time period t be $c_{ij}(t)$, and the total volume of i's interactions $c_i(t) = \sum_j c_{ij}(t)$. Degree $D_i(t)$ and network entropy $H_i(t)$ of node i during period t are defined as,

$$D_i(t) = |\{j \mid c_{ij}(t) > 0\}| \text{ and } \quad H_i(t) = -\sum_j \frac{c_{ij(t)}}{c_i(t)} \log \frac{c_{ij}(t)}{c_i(t)} \tag{6.1}$$

A second set of metrics, most relevant in networks with geomarkers, capture the characteristic travel distance or diversity of locations visited. Common examples of metrics here include *location entropy* [231] (defined similarly to the network entropy, but over the distribution of locations visited rather than individuals called) for diversity and **radius of gyration** [95] for travel distance.

Though we were motivated to explore this in the context of CDR it should be noted that the metrics (and subsequent bias issues) are applicable across many types of datasets. For instance, entropy and degree have proven informative in inference tasks ranging from estimating regional unemployment from Twitter usage[142] to predicting wealth from cell-

phone records [76, 43]. Related papers show similar results for mobility metrics [167, 88, 58]. In addition to proving useful on this range of societal-scale social networks, several forms of entropy have shown usefulness in aiding visualization of the DBLP citation network[183].

6.2 Statistical issues

6.2.1 Change detection

Detecting and quantifying the impact of an exogenous geopolitical event on these social metrics of interest (either over time or across locations) can provide important insight into how such events impact the behavior of larger populations. Examples in the literature include using anomalies detected in social media [110, 134] and mobile phone data [224, 125] to infer the severity and location of damage from natural disasters, or the impact of employment shocks [201]. Many of these difference detection techniques transfer smoothly across data-sets: techniques first applied to social media and communications data can be adapted to a data set as dissimilar as credit card transactions[72].

Non-parametric paired tests are used to detect if there is a systematic change in the mean of a metric of interest, say X, over the same population before and after an event or a treatment. For example, the Wilcoxon signed-rank test takes the paired difference of X before and after an event, ranks them in the order of magnitude and then uses the rank and the sign of the difference (discarding the actual magnitude of difference to avoid the effect of heavy-tail noise) to determine whether a change occurred. However, such tests (implicitly) assume that the bias in measuring in X is the same before and after the event. The proportion of times a paired test detects a change when there is no actual change (null hypothesis) is called the *type I error rate* (α) and when the underlying value did indeed change this proportion is called the *power* (β). In our work, we will use the signed-rank test to compare the metric values for a set of impacted individuals before and after an event.

6.2.2 Statistical Bias

However there are statistical nuances in the metric evaluation itself. Metrics that require large number of samples from the distribution may be confounded when the number of samples (for instance, the volume of communication) is much smaller than the support size of the distribution (e.g., the number of individuals in the true distribution). This is the issue of *sparsity* and necessitates the use of *estimator* functions that approximate the true underlying metric. Since we are interested in the predictive accuracy of the estimator, we focus mainly on its bias and variance (Equation (6.2)), the former of which underlies the problem discussed in this paper.

Definition 6.2.1. Let $\hat{\theta}(Y)$ be an estimator of true parameter θ^* using the data Y. The *bias* and *variance* of $\hat{\theta}$ is defined as,

$$bias(\hat{\theta}) = E[\hat{\theta}] - \theta^*, \quad var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]. \tag{6.2}$$

Note that the expectation E[.] is over the randomness of data.

It is important to note that many key social metrics, including all of those defined in chapter 6.1, have serious issues with bias when being estimated. Network degree and any entropy based metric have no unbiased estimator[166]. Obtaining an unbiased estimator for the radius of gyration, since it is related to the standard deviation of locations visited, is known to be a hard problem[217].

Estimating the support size, entropy and general symmetric functions[1] of discrete distributions when the number of observations is much smaller than the support size of the distribution is a fundamental problem that has been very well-studied [84, 96, 79, 17]. It is still an active area of research in statistics, information theory as well as computer science [161, 162, 174, 207, 206, 212, 208, 122, 219, 163, 220, 211, 12, 168]. While this research has improved state-of-the-art estimators, the primary focus has been on estimating a function

[1] A function over a discrete distribution is said to by a *symmetric function* if it remains invariant to relabeling of domain symbols.

on a single distribution, rather than the issue of varying sparsity across the network and over time — which are crucial in the applications of interest.

To give a concrete example, the optimal number of samples needed to estimate the entropy of a discrete distribution within ϵ-error is $\Theta\left(\frac{S}{\log S}\frac{1}{\epsilon}\right)$, where S is the support size of the distribution[2] [122]. In practice, we do not have the luxury of soliciting more samples to meet this bound and consequently the estimation of entropy per individual in a network will incur some non-negligible bias. As we will see in chapter 6.4, this can lead to systematic inference errors in a way that falls outside of this body of statistical literature.

An interesting caveat is that despite the strong literature in the theoretical community, the issue of bias is rarely addressed in the Big Data and Development space. This has been an issue in particular when looking at mobility since key metrics like radius of gyration and location entropy have issues with estimator bias. Using a more densely sampled signal that is normally not available, one work[231] showed a systematic underestimation of mobility metrics using call networks that was greater for individuals making few calls. This has also been previously seen in [173] which found that while key locations were generally well inferred, functions like location entropy or radius of gyration likewise had similarly unbalanced issues with bias. However neither of these works offered general solutions to the problem. Heuristics such as dividing a biased metric by the number of communications[63] have no guarantees in improving accuracy: whether they mitigate or aggravate the problem is entirely dependent on the underlying distributions, functions and sample sparsities. A recent work[209] has analyzed the specific bias induced on location metrics by the location-varying tower density and provided correction specially designed for their specific application setting.

6.2.3 Dynamic Sampling Sparsity

In the case where sparsity is stationary, the number of samples observed before and after an event is the same on average. The generative model for the observed data is as shown

[2]The notation $h(n) = \Theta(g(n))$ means that h is bounded both above and below by g asymptotically.

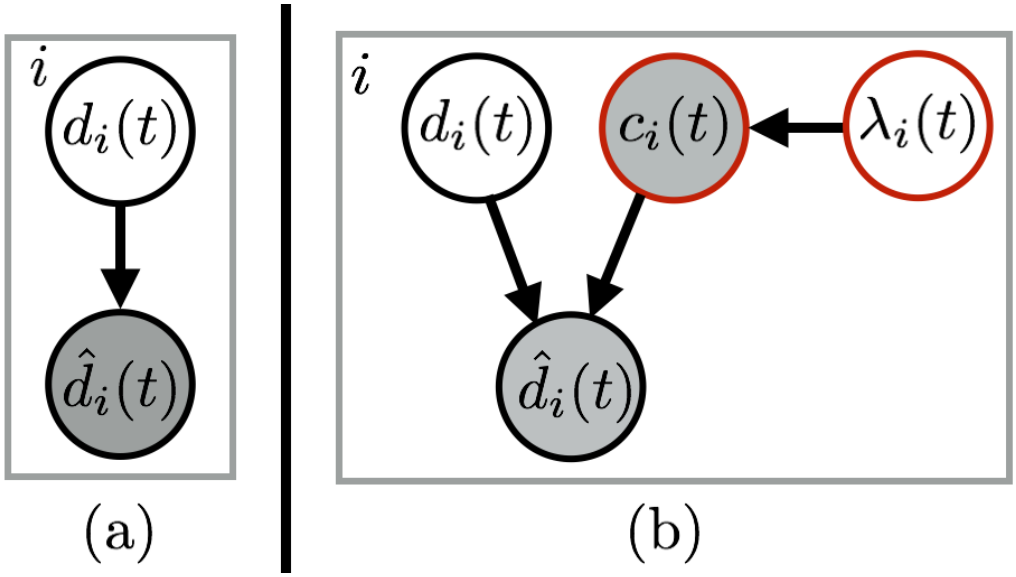


Figure 6.2: Generative model for the data for a single period of time (a) when sparsity is stationary, and (b) when sparsity is non-stationary.

in Figure 6.2(a), where $d_i(t)$ denotes the true distribution and $\hat{d}_i(t)$ denotes the *observed distribution* for an individual i. However, during emergency events the sampling sparsity is emphatically note the same as figure 6.1(a) illustrates.

Let $c \sim \text{Poi}(\lambda)$ denote a random variable drawn according to Poisson distribution with rate parameter λ. At time t, let $\lambda_i(t)$ be the rate of sampling for individual i and $c_i(t) \sim \text{Poi}(\lambda_i(t))$ denote the number of samples observed for an individual i. So, we get to observe the empirical distribution $\hat{d}_i(t)$, which is obtained by drawing $c_i(t)$ samples from the true distribution $d_i(t)$. This generative model is illustrated in Figure 6.2(b).

Let f be the functional (e.g: entropy) we are interested in computing on the distribution d. Let $\hat{f}_i := g\left(\hat{d}_i(t), c_i(t)\right)$ denote the estimator of f for individual i. We note that this estimator is not only dependent on the true underlying distribution $d_i(t)$, but also the sampling rate $\lambda_i(t)$. Therefore, the bias in estimating f_i at time t is also affected by the sampling sparsity, that is,

$$\text{bias}(\hat{f}_i(t)) = \text{E}(\hat{f}_i(t)) - f_i(t) =: B(d_i(t), \lambda_i(t)). \tag{6.3}$$

Since the sparsity is not stationary, that is, $\lambda_i(\text{after}) \neq \lambda_i(\text{before})$, the *bias itself is not stationary*. Even when $d_i(a) = d_i(b)$, the change in sparsity leads to a systematically increased type-I error rate in classical tests like Wilcoxon signed-rank test for detecting change in f.

We are interested in detecting and quantifying the difference in f of the distribution $d(t)$ before and after an event. For each individual i, let the difference be denoted by, $\delta_i := f\left(d_i(a)\right) - f\left(d_i(b)\right)$, where a and b stand for *after* and *before* respectively. However, we do not get see the distribution d_i itself, and instead we get to observe $\hat{d}_i$ which has c_i samples from the true distribution d_i. Given an estimator $\hat{f}_i$, the intuitive way to use it to find the difference δ_i is to estimate on the before and after distributions seperately and then take the difference. This gives us the estimator for the change $\hat{\delta}_i := \hat{f}_i(\hat{d}_i(a)) - \hat{f}_i(\hat{d}_i(b))$, where $\hat{f}_i(\hat{d}_i(t))$ denotes the estimate of f on the observed distribution $\hat{d}_i(t)$. Using Equation (6.3), the expected difference can then be written as,

$$\mathrm{E}(\hat{\delta}_i) = \delta_i + B(d_i(a), \lambda_i(a)) - B(d_i(b), \lambda_i(b)). \tag{6.4}$$

Under the null hypothesis, the underlying distributions remain the same before and after, i.e, $d_i(a) = d_i(b)$. Therefore, under the null, $\delta_i = 0$. When we test for change, we want to control α (the chance of declaring a change when the null is true). If sparsity was stationary, i.e $\lambda_i(b) = \lambda_i(a)$, the mean of the difference would be zero since bias would cancel when we take paired differences (Equation (6.4)). However, since the observed distribution also depends on the non-stationary rate parameter $\lambda_i(t)$, the mean of paired difference is not zero under the null. For $\mathrm{E}(\hat{\delta}_i)$ to be zero under the null hypothesis, we need the following to hold, for all $d_i(a), \lambda_i(a)$ and $\lambda_i(b)$,

$$B(d_i(a), \lambda_i(a)) = B(d_i(a), \lambda_i(b)). \tag{6.5}$$

For functions like entropy, which do not have unbiased estimators [166], such a condition would never hold for any non-trivial distribution d_i and estimator $\hat{f}_i$. This leads to a systematically increased type-I error rate under classic tests like Wilcoxon signed-rank test as

illustrated in Figure 6.3.

6.3 Methodology

6.3.1 The Downsampler correction

We propose downsampling the observed distributions to same number of samples before estimating f as a plug-in solution. Let $c_i^{\min} := \min\{c_i(a), c_i(b)\}$, and $\widetilde{d}_i(a,l)$ and $\widetilde{d}_i(b,l)$ be obtained by drawing $l \leq c_i^{\min}$ samples from $\hat{d}_i(a)$ and $\hat{d}_i(b)$ respectively. The downsampling-corrected version of estimator $\hat{f}_i$ for difference is then defined as follows:

$$\tilde{\delta}_i := \mathrm{E}\left(\hat{f}(\widetilde{d}_i(a,l))\right) - \mathrm{E}\left(\hat{f}(\widetilde{d}_i(b,l))\right), \tag{6.6}$$

where the expectation is over the randomness of drawing $\widetilde{d}_i \sim \hat{d}_i$. In practice this can be approximated by averaging over a few random re-samplings. This solution has a number of advantages. It will on top of whatever estimator already being used no matter how complex. This is important since there is active progress in the field of reducing (single distribution) bias and in this way we do not make any requirements of future estimators. Secondly, downsampling ensures that under the null hypothesis, the bias in estimating f is same for before and after and hence it cancels out when we take paired difference. Thirdly, while the situation where the null hypothesis is false is significantly harder to analyze, the performance of the proposed correction in this case is empirically very strong. We see this in chapter 6.4 where it outperforms all competitors over a wide range of plausible scenarios.

6.3.2 Empirical and Synthetic tests

To verify the pertinence of this problem in real-life analysis we perform a number of empirical studies. In all of these we focus on inferring the change in two metrics: social entropy and network degree. We pick these two since they are both socially informative as well as ubiquitously available over many different types of social graphs. We are interested in how estimates in the change of these metrics are impacted by the variation in sparsity, which we

quantify as the *elevation rate* r.

$$\text{Elevation Rate, } r := \frac{\lambda(\text{after})}{\lambda(\text{before})}.$$

We compare the performance of the following four estimators:

1. **Naive-Estimator**: This simply computes the metric by treating the empirical distribution as the true distribution.

2. **Jackknifed naive**[78]: This is the naive estimator with a jackknife heuristic that averages the naive estimate over all distribution generated by removing one sample from the empirical distribution.

3. **JVHW**[122]: This estimator combines an unbiased estimator for the best polynomial approximation of the function being estimated in the non-smooth region with a bias-corrected estimate on the smooth region.

4. **APML**[168]: Approximate Profile Maximum Likelihood Estimator is a computationally efficient approximation of the profile maximum likelihood [12] which maximizes the probability of the observed profile (multiplicities of the symbols observed ignoring the label).

Note that JVHW is only applicable to entropy and not network degree. We compare these estimators to their *corrected* variants, where we apply the downsampling correction before running these estimators. We found the results broadly similar across the corrected version of these four methods. In the interest of clarity we only show one corrected estimator per graph: the JVHW-correction for entropy and the jack-knifed correction for network degree.

In all of these experiments we ask two questions. Firstly, what is the bias in the estimated difference for each estimator under different values of elevation rate r? Secondly, how does this translate into type-I and type II errors? The first question is simply done by computing the average predicted change and comparing it to the actual average change. The second

question is studied by applying a Wilcoxon signed-rank test to the estimated differences with a desired α of 0.01.

In the first set of experiment we use a country-wide CDR dataset collected over 6 months in Afghanistan. This data comprises data for millions of callers and since our interest concerns changes in specific groups of individuals we restricted this to calls from a set of $N = 1000$ individuals determined to be living near a specific tower in a major city. We take the empirical distribution generated by 6 months of data (with a median of 500 calls per individual) as being sufficiently well sampled to approximate the true social distributions $\{d_i\}$'s and call rates $\{\lambda'_i\}$'s.We take the empirical call rates for six months and scale them down to the equivalent rate for a week $\lambda_i = \frac{7}{180}\lambda'_i$. We then assign before and after distributions to be identically $d_i(a) = d_i(b) = d_i$ and $\lambda_i(a) = \lambda_i$ but we multiply the second calling rate by the elevation rate: $\lambda_i(b) = r\lambda_i$. We repeat 100 trials where we sample using these distributions and λ's as in Figure 6.2(b) and compare the estimated difference between the metric average of sets a and b. We run the Wilcoxon signed-rank test and check if it detects a change. Since the distributions are the same, ideally we would like to estimate that there is no difference.

While experiments on real data are essential to proving the practical concerns around the sampling problem they only provide a fixed set of conditions to experiment with. For this reason we created a synthetic test suite that would allow us to compare our methods against baselines on a variety of distributions and at a significantly more granular level. This allows us to directly set $d_i(a)$ and $d_i(b)$ to both explore different distributions and also be significantly different. As such we can compute the bias of estimators when $E[\delta_i] \neq 0$ as well as for the null case where $E[\delta_i] = 0$.

The experiment then proceeds similarly to the empirical one: with the exception that λ_i and $d_i(a)$ are drawn randomly from a prior distribution. λ_i is consistently distributed by a log-normal with mean of 50 while we perform separate experiments where the $d_i(a)$ are drawn from a distribution of either Dirichlet (with Dirichlet parameter $\alpha_D = 1.0$), geometric (with average probability of success $p = 0.9$) or uniform distributions. For the case where

we wish $E[\delta_i] \neq 0$, we additionally alter the parameter of $d_i(a)$ by some fixed amount to generate $d_i(b)$.

6.4 Evaluation

Before we examine the sensitivity of different methods to the elevation rate we verified first the role of sparsity and set size on accurate predictions. Using our Synthetic test we found that the existing methods will have substantial bias in the null case no matter how large the population N is, as shown in Figure 6.4(a). The variance decreases as a function N, but not the bias. We set $N = 1000$ for the remainder of our synthetic experiments. Figure 6.4(b) shows that decreasing the sparsity, or equivalently increasing the observation rate λ, of course helps all methods: though as previously noted this is rarely possible in practice.

With this established we then moved onto the real-data tests. Recalling that we can only test the scenario were the null case is true we wanted to investigate the bias and type-I error. Figure 6.3 shows the results for social entropy and network degree: though the same trend is present in both. We clearly see that all the methods that do not correct for varying sparsity, including cutting-edge estimation techniques like JVHW and APML, reveal substantial issues with bias at even modest elevation rates which get progressively worse as the rate increases. In contrast, our corrected method consistently returns the correct result no matter the level of imbalance in sparsity. This confirms the theoretical results we had discussed in chapter 6.3.1.

We then proceeded onto the full synthetic tests which handles both the null and non-null scenarios but also tests over a variety of different underlying distributions. The results for the null case (panels a – d of 6.5) further strengthen the conclusions drawn from the earlier real-world test cases: no matter the underlying distribution the downsampler performs perfectly. Note also the variance in performance that each uncorrected method has depending on the metric and distribution for the null case. This illustrates the difficulty of the problem when not accounting for variable sparsity: a non-corrected method that seems to work on one distribution may entirely fail on another.

We also record how often the Wilcoxon singed-rank test records a true positive as a function of the actual average difference(panels e – h of 6.5. We see that the elevation rate has induced an asymmetric change in non-corrected methods and hence worse discovery rates when the true change in entropy is negative. On the other hand, the corrected method is reliable through-out (though no longer perfect). Even in a situation where a given uncorrected method perform well (notably, the APML method is fairly robust in the uniform scenario for both network degree and entropy), the corrected method has comparable or better sensitivity while outperforming it in all other situations. This provides strong evidence that the plug-in correction is an improvement also in the case where there is a difference.

6.4.1 Impact on Socio-economic Analysis

Recalling our initial motivating problem of quantifying the impact of violence on people introduced in chapter 6.1: we give here an example of how dynamic sampling sparsity could impact a real analysis of this problem. In this analysis we had cross-referenced calls made in our CDR with the time and location of a serious bomb attack in Afghanistan and generated a set of 220 individuals who appear to within a kilometer of where the attack took place. Given this set of people, we wanted to analyze how the average network entropy changes in the immediate aftermath of this emergency. For each 24-hour period in our date range we take the difference with respect to the same period one week before. For example, the 24-hour period starting on August 22nd at 10am is paired with the 24-hour period starting on August 15th at 10am, the one starting at August 22nd at 11am is paired with that starting on August 15th at 11am, and so on. We then compare how different methods infer changes based on these differences: our results are shown in Figure 6.6.

While both the basic methods and our correction to JVHW method detect an increase during the emergency period, the uncorrected methods detect anywhere from twice to three times as much of a change. Moreover, the corrected method finds only one 24-hour period to be statistically significantly different: while the other methods declare almost the entire period to show a significant increase in network-entropy. Such a difference between the

uncorrected versions (statistical certainty that average calling entropy has increased by 0.2 the end of the day) versus the corrected version (non-statistically significant change of 0.1) is considerable. It is easy to see how repeating such an error over several events could lead a researcher to draw conclusions not supported by the underlying data.

6.5 Conclusions

This project identified the phenomena *dynamic sampling sparsity*, and highlighted why it is such an important problem for estimation and change detection. Our statistical framework shows that failing to account for varying sparsity in the data frequently leads to systematic errors in the downstream statistical analysis. We demonstrate the severity of this issue through experiments on both real social graph datasets and comprehensive synthetic tests. We created correction that was provably correct in the case where there is no change and was shown to work well in a wide variety of non-null cases by outperforming state of the art estimators.

While we motivated this problem by considering the real-world problem of understanding the impact of emergency events, we note that this problem of varying sparsity is significantly broader. Indeed the issue would likely arise when comparing average values of social metrics (whose bias gets influenced by sampling sparsity) between two different populations with different sampling sparsity rates. Examples in the literature include comparing the structure of social networks in urban locations with that of provincial villages [76], or wealthy provinces to a poorer ones [77, 142]. Moreover, the problem is not strictly restricted to a development focus. Similar behaviour would occur in social media datasets for both emergency and non-emergency (e.g election/sporting event) events that sociologists might be interested in studying.

Though we have given a good solution for the the case where the two distributions being considered are in fact the same the non-null case much harder to analyze. Though the very simple downsampling solution we have supposed beats state of the art single-distribution estimators it is entirely possible that creating a new estimator explicitely designed to deter-

mine the difference in a function (rather than run it directly) might work better. In this case previous works that attempt to estimate functions on joint distributions might be one approach to consider [208]. Another intriguing possibility might to modify functional estimators [122] to strongly favor bias reduction properties since we know that in our scenario that will cause most of the problems.

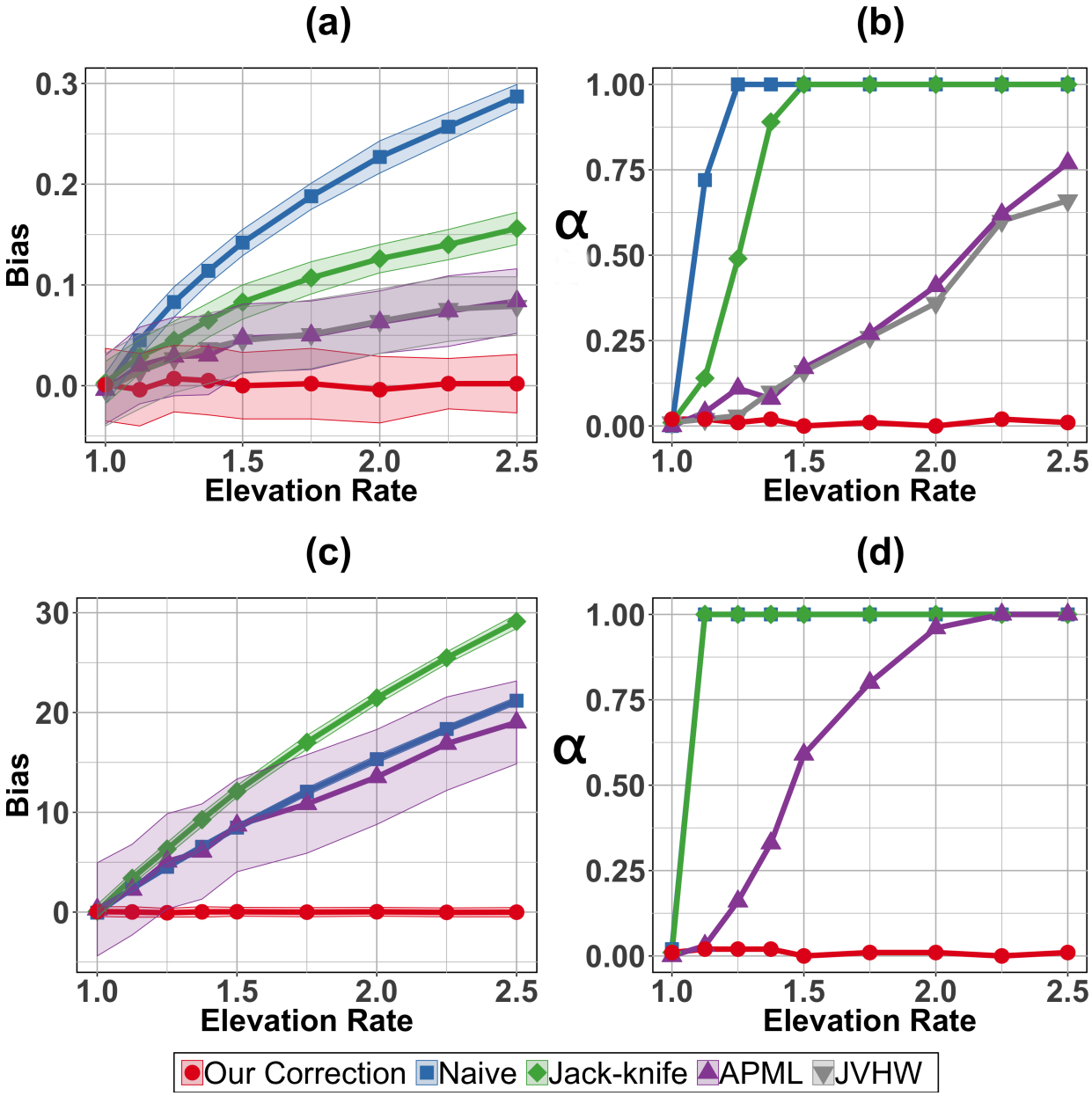


Figure 6.3: Comparison of how the (a) bias and (b) type-I error rate (α) for estimating difference in entropy increases with more variation in sparsity. (c) and (d) show the same for network degree. The bands in (a) and (c) show the variance in estimates of the average difference.

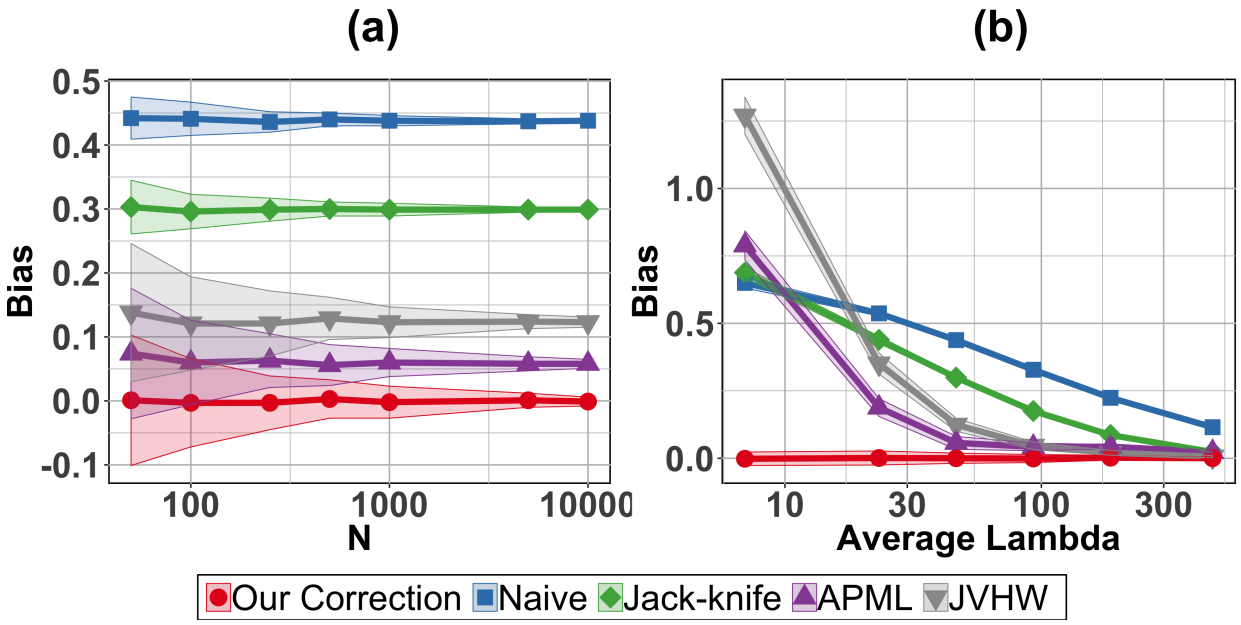


Figure 6.4: Analysis of how (a) varying the number of individuals N and (b) average sparsity impacts entropy bias. Both experiments are run on synthetic data generated from the Dirichlet distribution with elevation rate $r = 2$. The former only improves the variance while the latter decreases bias as the average sparsity drops (as calling rate increases).

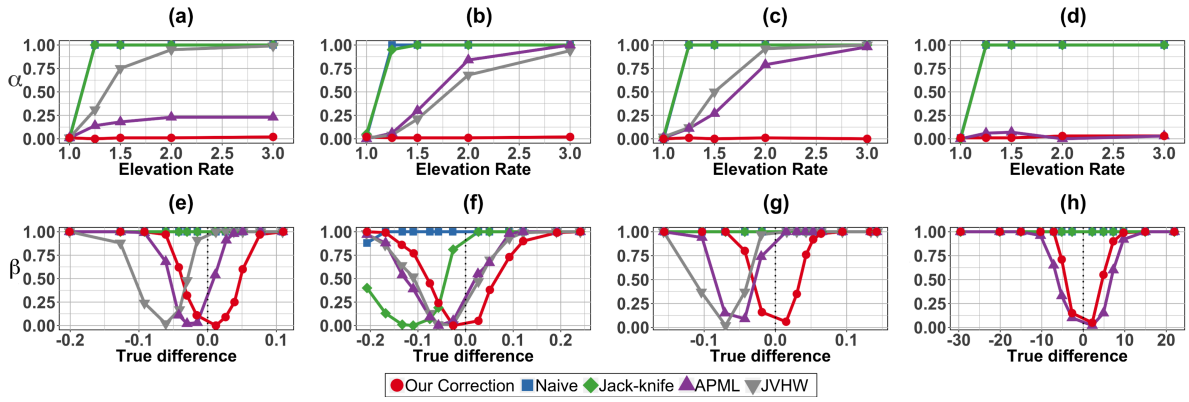


Figure 6.5: Panels (a)-(c): Experiments showing type-I error rates (α) for entropy change for the uniform, geometric and dirichlet scenarios respectively. Panel (d): Type-I error rate for network degree under the uniform scenario. Panels (e)-(g): Power (β) for entropy change detection at an elevation rate of 3 for the uniform, geometric and dirichlet scenarios respectively (h): Power for network degree under the uniform scenario and elevation rate of 3.

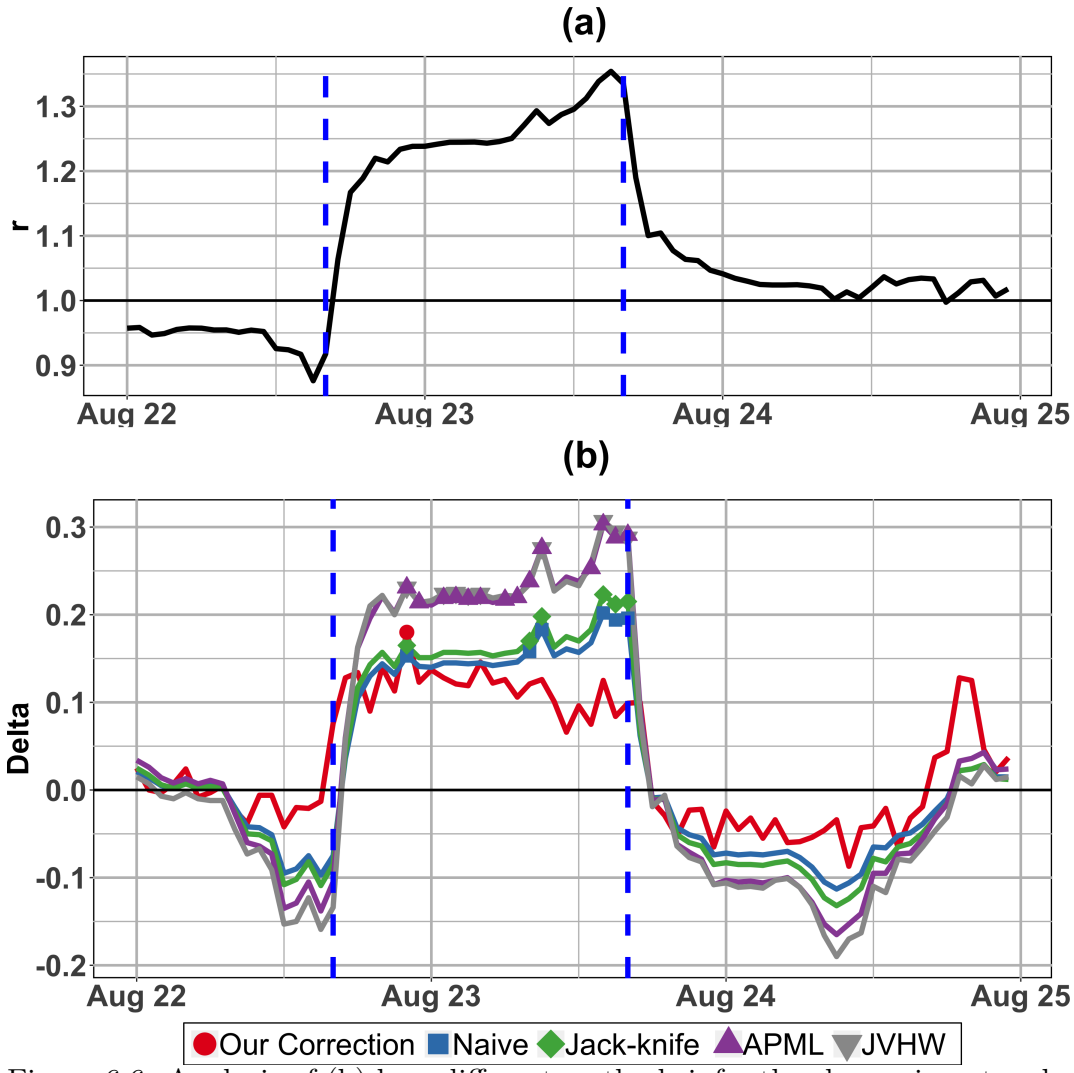


Figure 6.6: Analysis of (b) how different methods infer the change in network entropy in (a) the presence of varying sampling sparsity caused by a violent event. The period between the dotted blue lines indicate when the sliding window contains the bomb blast period. Marked points in (b) indicate a statistically significant difference between this 24-hour period and the same 24-hour period one week prior.

Chapter 7

CASE STUDY: INFERRING THE CAUSAL EFFECT OF VIOLENCE ON SOCIAL COHESIVENESS

7.1 Motivation

Much has been written concerning the effects that violence and warfare has on society. While it is clear that prolonged exposure to intense violence is corrosive to a countries, both in a social and materially economic sense, understanding the particular mechanisms and measuring these effects exactly is an area of great importance[32]. Enduring patterns of violence can impact various economic indicators from GDP growth[10] to house prices[29]. Violence, especially in the form of insurgency or civil war, can also measurably fragment society and trust in institutions[62]. Studying these impacts and understanding them would be especially important in the case of Afghanistan: a country that has been wracked by violence for over 40 years.

In this project we have access to a very comprehensive CDR covering Afghanistan. The telecommunications company owning the CDR has a commanding market share and is available from late 2013 to early 2017. The CDR contains information about foreign calls as well as the tower of the caller though not of the receiver. The period of several years covered by this CDR enables us to perform longitudinal studies of over a year. Additionally, the breadth of coverage over the entire country will enable us to analyze reaction to a variety of different bombings, insurgent attacks and natural disasters. We will seek to specifically answer how does violence impact the social cohesiveness of Afghanistan as measured through its call data.

7.2 Background: Ethno-linguistics and Violence in Afghanistan

7.2.1 Tracking Violence in Afghanistan

The recent history of Afghanistan has been regretfully replete with instances of political violence and upheaval. A communist coup in 1978 that overthrew the 230 year old monarchy [26] was shortly followed by the Soviet invasion and a 10 year insurgency. However, the withdrawal of Soviet forces led to a power vacuum and another decade of fighting before the Taliban was able to solidify its positions in Afghanistan outside of the north. The Taliban rapidly lost their grip on the country during the American invasion of 2002 and then returned to their former role as insurgents. The Taliban have targeted the multinational military coalition forces, Afghan National Security Forces (ANSF), government targets and civilians at large in an insurgency going into its 17th year. In 2017 the situation was further complicated by the increasing presence of the Islamic State Khorasan ("IS-K"), an affiliate of the so-called "Islamic State", which attacked both traditional insurgent targets as well as the Taliban itself.

There have been several initiatives to track and record details of these violent events. The multi-national forces in Afghanistan maintained a detailed SIGACTS (significant acts) database contained over 600,000 events spread over 6 years when it was leaked. Additionally, we had access to the government curated PiX violence dataset which draws from various media and military reports over the period 2015-2016. While it covered a shorter period of time, the reliability and spatial accuracy of the information was judged to be significantly higher quality. However, other datasets with a more global scope also include detailed information on violence in Afghanistan. The Global Terrorism Database[136] is a premier example that has been collected terrorist events since the 1970s all over the world. For Afghanistan in the period covered by our CDR, GTD records about 5800 violent events.

These datasets vary considerably in their scope and completeness but there are a few components which are critical for our analysis. All of them record the time and approximate location of a violent event: though it should be noted the confidence of these estimates vary.

This is vital since it allows us to determine which individuals might have been impacted by looking who reside in hometowers close to the event during the time of violence. Additionally, the datasets need to provide estimates of the total deaths and casualties caused by each event: allowing us to quantitatively scale the violence in each event. The SIGACTS dataset does not provide this but the other two do, grouping the casualties into civilian, military and insurgent.

Past that point the pros and cons of each data source must be carefully weighed. PIX data provides significantly more events of smaller impact than GTD (which generally tracks events large enough to be verified in press reports) while the SIGACTS is perhaps the most comprehensive. An additional concern is that this lack of comprehensiveness may come about from a reporting bias[214]. Moreover, GTD only records events which can be considered 'terrorism'. This means that natural disasters like earthquakes and casualties caused by ANSF operations/airstrikes will not be included. Counter to its relative lack of breadth, GTD has the largest temporal range with span extended all the way back to 1970s. This is contrast with the PIX data (which only covers 2015-2016) and the SIGACTS data (which ends in early 2015). We note that both of these cover only a subset of our CDR data. Finally, there is also the level of detail to consider. In addition to the necessary spatio-temporal details and casualties rates it would be desirable to understand the type of violence (e.g a bombing versus a large scale attack), who committed it and who was the target. While all datasets have detailed (though mutually incompatible) categorizations of the violence type, the GTD dataset is the only one to clearly code the victim and perpetrator column. However, in the end the superior granularity and trustworthiness of the PIX data was considered the most important attribute and we decided to go with that dataset.

There is a significant body of scholarship that seeks to use these event datasets to better understand violence in Afghanistan. SIGACTS has been used since its release in projects such as a verification of counterinsurgency theory [28] to a statistical analysis of the spatiotemporal progression of violence[226]. One recent work[59] used SIGACTS data to track violence during elections and verified not only that greater violence depresses turnout but also that

terrorists tried to avoid civilian casualties while doing so.

7.2.2 Mapping the Ethno-linguistic groups in Afghanistan

Understanding the rich profusion of tribal and linguistic identities that make up Afghani society is as informative as it is difficult. One on hand, Afghanistan is a often place where "tribal and ethnic groups that primacy over the individual." [26], and as such we could hope to learn a great deal from how these groups interact with one another. On the other hand, there is a great deal of overlap on the ground and even understanding what criteria individuals use to group or categorize themselves can be a challenge.

However there are a few broad groups that we may attempt to differentiate between, while acknowledging this is a crude division that misses many internal nuances and smaller communities. A rough map of these different populations can be seen in figure 7.1 and descriptions follow below:

- *Pashtuns* probably comprise the plurality of Afghanistan: making up perhaps 40% of the population. Found mainly in the south and east, with a large population in Pakistan, they primarily define themselves by tribal lineages, the Pashto language and adherence to the Pashtunwali code of conduct[26].

- *Tajiks* are the next largest group, with perhaps 30% of the population. This group can be found in major urban centers as well as the north-east mountains. They are defined by their language, a form of Persian, and their locality: meaning that this definition may be seen as less internally coherent than some others.

- The *Hazara* community resides mainly in the center of Afghanistan and have about 15% of the population. They are noticeable for being Shia minority in an otherwise Sunni Muslim country.

- *Uzbeks* and *Turkmens* speak Turkic langauges, in contrast to the other major groups,

and are found in the north in areas bordering Uzbekistan and Turkmensistan. Together, they comprise about 10% of the population.

- The *Baluch* population found in the south-west of the country are a part of a larger community in Iran and Pakistan.

- The *Aimaq* population is a Persian speaking tribal society with some Turkic cultural mores. Aimaq communities focus more on pastoralism than their regional neighbors and are found in the central and north west of the country.

- The *Nuristanis* of the northeastern mountains are a collection of tribal societies speaking a language significantly different from any other in the region.[26].

Asking about ethnicity and cultural group membership is a politically sensitivity topic due to past instances of discrimination and violence. Questions about linguistics are more acceptable even though there is a significant amount of overlap between linguistics and cultural identity. The two official languages of Afghanistan and *Pashto* and *Dari Persian*. The latter term encompasses the dialects spoken by Tajik (though this inclusion is a politically contentious issue), Aimaq and Hazara communities and is close enough to Iranian Persian to be mutually intelligible. While Pashto is also a part of the Iranian language family it is sufficiently different from Dari and Iranian Persian to not be mutually intelligible; a similar situation applies to the *Balochi* language. As noted, Uzbek and Turkmen are both from the broad family of Turkic languages but are also distinct enough to be considered different languages. The languages spoken by the Nuristani people should more accurately be considered a group of different languages. The current scholarly consensus is these languages are distantly related to Iranian and north Indian languages.

The Central Statistics Office (CSO) of Afghanistan is the government ministry charged with collecting statistics regrading the Afghani people. In particular, this minisitry administered a survey of villages that recorded not only an approximately correct latitude and longitude for each village (with some noise added for anonymity reasons) but also a record of

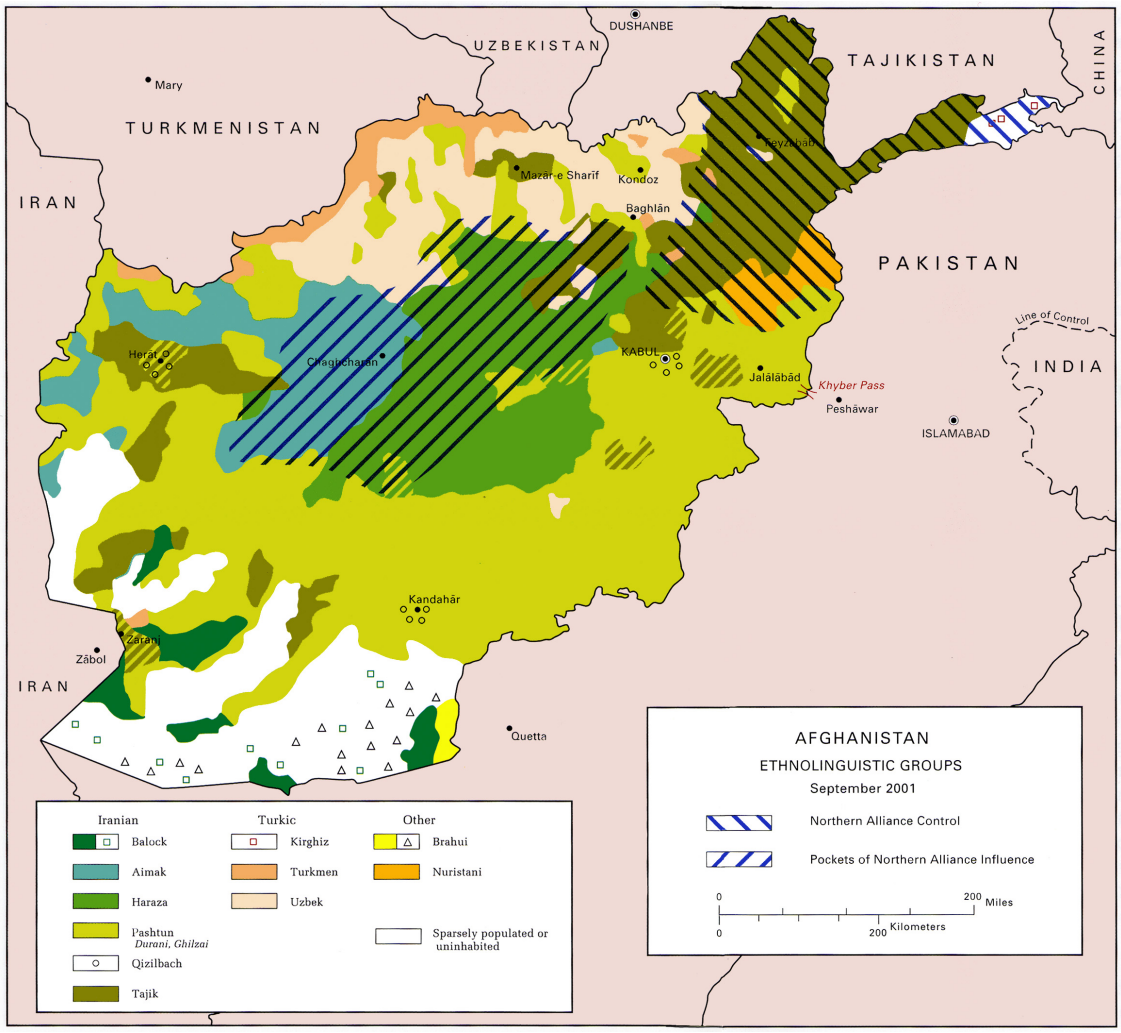


Figure 7.1: A map released by the US military showing the different ethno-linguistic groups in Afghanistan prior to the US invasion of 2001.

the primary language spoken there. There are also non-governmental organizations operating in the same space, such as The Asia Foundation and their annual Survey of the Afghan People[6]. The foundations sends survey members to villages throughout Afghanistan to track economic outcomes such as access to clean drinking water as well as attitudes to social issues such as the role of women. This survey also asks each respondent the languages which we can speak.

Ethnicity and linguistics in can have very strong effects on the political and social development of countries and there is a strong literature to that effect[164, 60]. While traditional studies have tended to study this by looking at the heterogeneous outcomes in different ethnic regions for some variable of interest, large passively collected dataset show great promise in extending the granularity of such analysis. Studies in Estonia[202] and south Asia [38] were able to track segregation at the individual level by looking at the language preferences of an individual's phones. However, these studies focused mainly on quantifying measures of segregation and trying to measure them using CDRs. While sociologists have long known about the relationship between ethnic divisions and violence [82, 159], this relationship has not been analyzed through the lens of large scale passive datasets.

7.3 Methodology

We detail below the steps required to generate the observations that will be used in the regression analysis. These steps are:

1. Use the CSO survey data to give an ethno-linguistic distribution to each cell tower.

2. Assign violent events from our dataset to towers based on spatial proximity.

3. Determine which towers were substantially impacted by violence and infer the change in their ethno-linguistic calling patterns.

At the end of this processing we will be left with a set of observations of the change in ethno-linguistic calling patterns (dependent variable) and the total magnitude of violence by

casualties (independent variable). Moreover, we will add possible confounder variables like ethno-linguistic breakdown of tower, regional fixed effects: this is discussed in 7.4.

7.3.1 Relating Linguistic Distributions to Towers

In order to understand how communication across cultural and linguistic groups are affected by violence, one must first be able to determine where these groups are. In the ideal case, this could be done at the individual level since this would be the most granular and helpful for analysis. However, previous works that have been able to do this have relied on a phone-level variables like language setting[40], which we do not have access to. While it might be possible to assign individuals to groups using network algorithms[35], there is the fear that it would be hard to verify the truth of such an assignment and that concern over its accuracy might impact our conclusions.

As such, a decision was made to focus on instead assigning ethno-linguistic distributions to cell towers based on the CSO linguistic data we had. We recall that each village observation in the CSO survey had a spatial component as well as a primary language and approximate population. The idea was that we could use this information to assign each village's primary language (weighted by its a population) to a given tower and then sum that over all villages. The resulting weights could then be normalized for each tower to provide a distribution over ethno-linguisitic groups.

Recalling the Voronoi nature of tower catchment areas described in chapter 2.1, it might seem that the ideal assignment strategy is to simply assign all the weight of a village to the closest tower. However this ignores certain real-world considerations. Firstly, the survey itself may have added some noise to the latitude and longitude for privacy reasons. Secondly, the village is not a point mass and dwellings may extend over a wide area: especially if its inhabitants practice agriculture or pastoralism. Thirdly, people will move around in the course of daily work and may make or receive calls in other nearby towers. As such we designed a heuristic that instead took all towers within a 5 kilometer radius of a village and shared the population among these towers proportional to the inverse of the towers distance

from the village. Any towers closer to the tower than 1 kilometer were weighed as if they were at 1 kilometer to avoid excessive influence. For example: assume we have a village v with 1450 Pashto speakers and three towers t_1, t_2 and t_3 at distances of 0.5, 4 and 5 kilometers respectively. Then tower t_1 would have weight $\frac{1}{1}$ (since it's closer than 1 kilometer we round up to 1), t_2 has weight $\frac{1}{4}$ and t_3 gets a weight of $\frac{1}{5}$. Then village v would add $(\frac{100}{145})1450 = 1000$ Pashto speakers to t_1, $(\frac{25}{145})1450 = 250$ to t_2 and $(\frac{20}{145})1450 = 200$ speakers to t_3. The result of applying this algorithm to all the towers in our CDR gives us the map seen in figure 7.2 and we can see that at the broad level it matches the map in figure 7.1.

Such a heuristic works well in rural areas, where the relative density of villages to towers provides an informative estimate. However, since the survey mainly inspects villages instead of urban areas we get two problems around towns and villages. The first is that there are many towers in the middle of urban areas with no villages nearby for which we cannot estimate an ethno-linguistic distribution. The second is that towers which are both close to villages and urban areas receive linguistic information from the former and not the latter: potentially biasing our estimates if there are urban-rural differences in linguistics. In our mind the second problem is far more serious, since it might lead to unrepresentative tower estimates while the first problem simply means there are towers which we know to be unknown.

In order to mitigate the first problem we leveraged the fact that in almost all large towns there would always be at least one (and often more) towers without an assignment. All towers without any assigned weight will be considered as 'unknown' and assigned to set U_1. We then run a post-processing that looks at all towers not in U and assigns each tower t that has some tower $t' \in U_1$ within a 5 kilometer radius to U_2. After this is done we declare all towers within $U_1 \cup U_2$ to be 'unknown' and reset their ethno-linguistic assignment to 100% 'unknown'. This cautious approach leaves us more confident in our linguistic assignments but leaves gaps in the major urban centers of Kabul, Kandahar and Mazar-i-Sharif as figure 7.3 illustrates.

Figure 7.2: Inferred linguistic map for Afghanistan

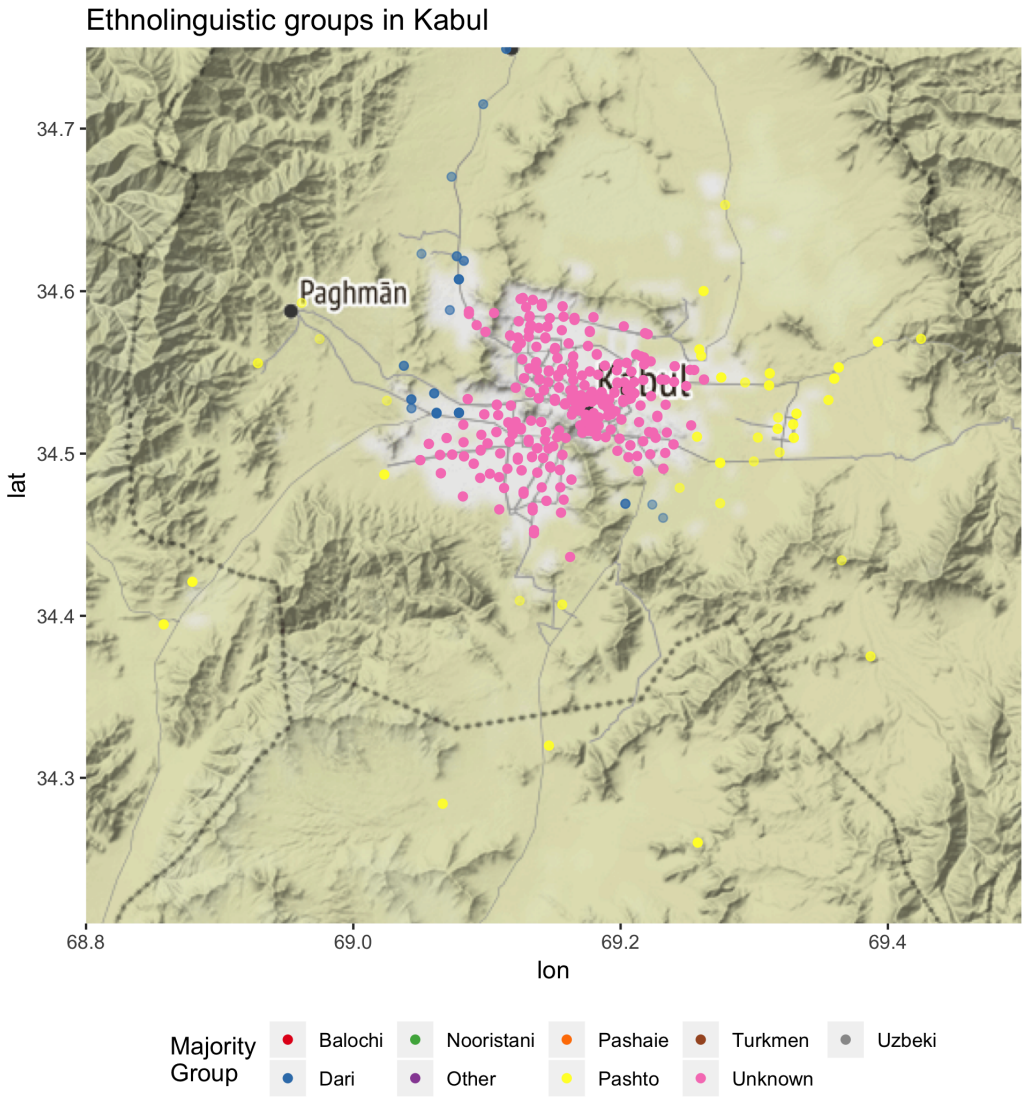


Figure 7.3: Inferred linguistic map for Kabul

7.3.2 Defining Linguistic Call Metrics

Having developed a tower-level mapping that we felt comfortable working with, the next step was understanding how we can use these to determine the calling behavior of a tower. Unfortunately, many of the relevant metrics defined in the literature such as communication segregation[38] or inbreeding homophily[60] require having individual-level linguistic assignments or hard assignments to a single group. As such we defined new metrics that used tower-level assignments.

Let us assume that a given tower i has a probability distribution T_i over the linguistic groups. We also assume that the tower makes a total of c_i calls with c_{ij} calls going to tower j with corresponding linguistic distribution T_j. Then we can define the *outgoing EL-distribution* of tower i as

$$L_i^{out} = \frac{1}{c_i} \sum c_{ij} T_j \tag{7.1}$$

The equivalent *incoming EL-distribution* (L_i^{in}) is defined similarly, except the summation is taken over all c_{ji} instead. Note that c_{ij} does not necessarily have to refer to the total number of calls, but could simply be any generic weight metric. For example c_{ij} could be the total duration of calls from tower i to j, or the number of calls from tower i to j excluding all calls below 1 minutes in duration. Any such weight metric will be explicitly defined when introduced into the analysis.

We note that these in/outgoing linguistic distributions are themselves somewhat simplified. It assumes that the distribution of ethno-linugistic group membership of of individuals calling tower j from tower i is the same as the overall distribution T_i. In actuality, this might be skewed towards certain linguistic groups. For example if T_i was 50% Pashto and 50% Dari while T_j was 100% Pashto it might be reasonable to assume that the group of individuals calling T_j from T_i are disproportionately Pashto speakers.

Having defined in/outgoing linguistic distributions on an individual level we consider now how to convert them into a metric. The most intuitive metric is simply taking the value

of the linguistic distribution for a given language. For example, L_i^{in}(Pashto) would be the proportion of incoming calls estimated to be coming from Pashto speakers. Though simple, these proportional metrics have a number of advantages. Firstly, we do not need to know any individual's own linguistic mapping ahead of time, the change will be manifest in any case. The final advantage is that, unlike the more advanced functions studied in chapter 6 these function are simply count ratios. While there may be random error associated with few measurements there will not be any bias in any measurements: which eliminates the potential issues of dynamic sampling sparsity. As such, we decided that these features would become our dependent variable in our regression analysis.

7.3.3 Relating Violence Datasets to the CDR

With the ability to now compute individual-level metrics that can be aggregated to tower-level dependent variable, we turned our attention to the independent variable: the level of violence. As discussed in chapter 7.2.1, our final dataset of violent events had approximate longitude and latitudes, which are essentially in linking these events to individuals in our CDR. We did this by first assigning an event to a tower if it occurred within ten kilometres of the tower's position. Note that this is one-to-many mapping where one event could be mapped to several different towers.

Recalling that the events our event datasets provide rough event classes as well as temporal information, we can construct an approximate "violence history" of the tower in question. Visualizing this allows one to understand the huge amount of variation in not only the degree but type of violence histories depending on which region and even specific town is analyzed. Figure 7.4 shows the violence history for an area on the outskirts of Jalalabad. We note that after the start of the fighting season in May 2015 there is an endemic sequence of violent events revolving around insurgent attacks, ANSF counter-attacks and airstikes and insurgent infighting (between the Taliban and ISIS). Figure 7.5 shows the situation in Kabul: we note that despite a similar degree of continual high-casualty violence the events in Kabul are predominantly bombings as opposed to gun attacks. In other regions note in the heart

of the fighting, violence can be much more sparse and concentrated into a single month or week. This is the case for a small town in north-east Afghanistan, as figure 7.6 shows.

Analysis of these different violent scenarios is crucial to understanding the regression problem. We see here that the independent variable (violence) is heterogeneous in several ways other than just the gross casualty count. Some areas have a history of little violence and then suffer a sudden burst of violence while others have been living through continuous violence for years. It is reasonable to assume that any reaction to violence will be colored by how common violence is for those living in the area. Moreover, we see that there are different perpetrators of violence: we might see different reactions if casualties are caused by the local Taliban, foreign ISIS fighters or collateral from ANSF operations. Finally the type of violence matters too: the actual presence of insurgent or ANSF forces during a period of violence may provoke a very different reaction than a bomb or mortar attack that causes an equivalent number of casualties. Addressing how to account for these will be a critical component of our regression analysis.

7.4 Causal Analysis

7.4.1 Setting up the problem

In this analysis we are seeking to understand how inter-ethno-linguistic communication (the dependent variable) in Afghanistan is impacted by political violence. An important first step will be to define the spatial and temporal observational granularity. In the ideal case the spatial granularity would be at the individual level: where each subscriber forms a set of observations. However this raises issues of both computation complexity as well as measurement accuracy issues related to sparsity as discussed in chapter 6. As such, we have decided that the fundamental unit of spatial aggregation will be the cell tower. Along the temporal axis the answer is a little less clear-cut. Aggregation at the daily level would reveal more fine-grained patterns but at the risk of making long term-trends hard to detect. Since our primary interest is to look for persistent changes in calling behaviours, we have decided

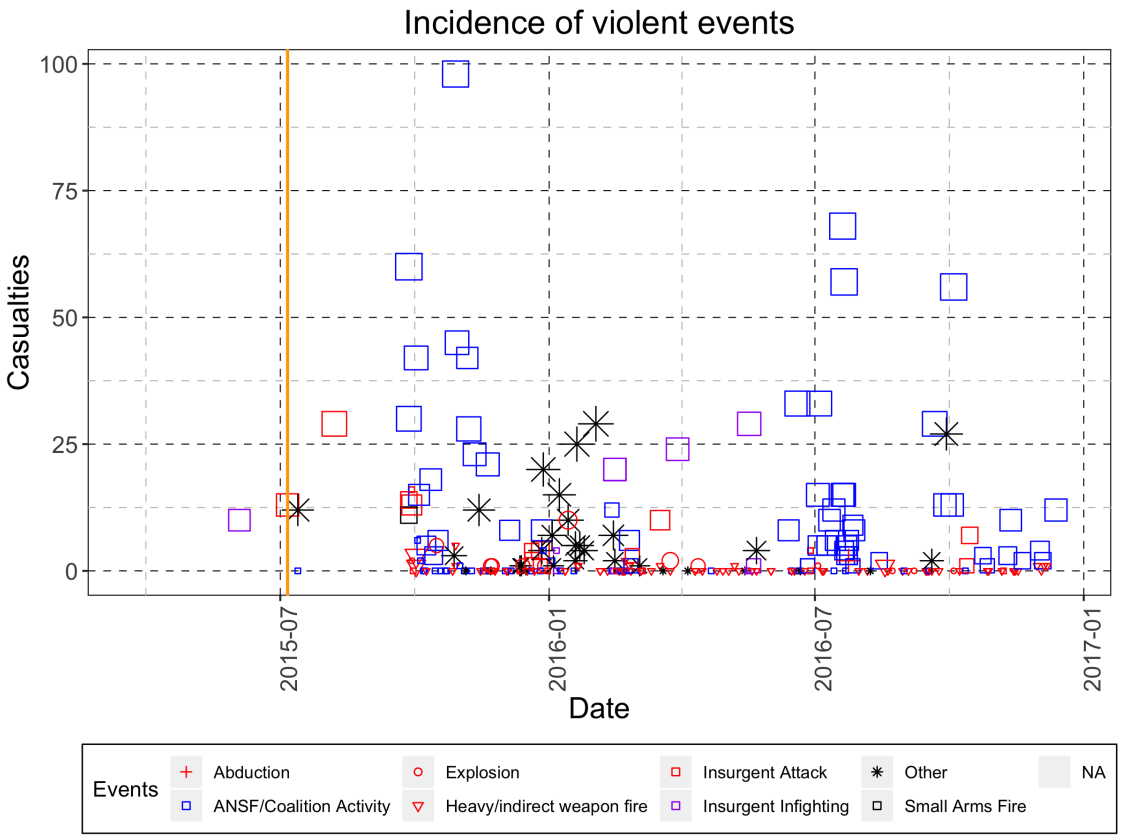


Figure 7.4: Temporal graph of violent events in a semi-rural area on the outskirts of Jalalabad. Each point in the graph is an event where the shape and color of the point give information about the event.

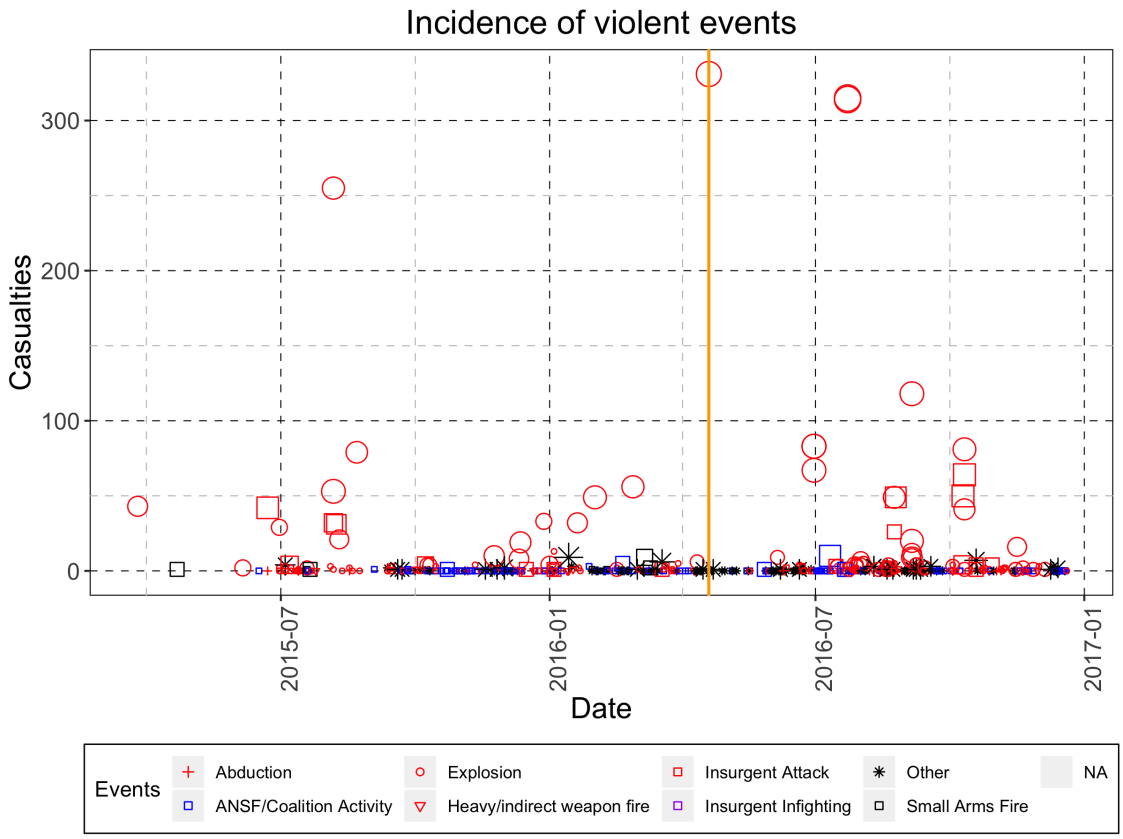


Figure 7.5: Temporal graph of violent events for an area in the center of Kabul. Each point in the graph is an event where the shape and color of the point give information about the event.

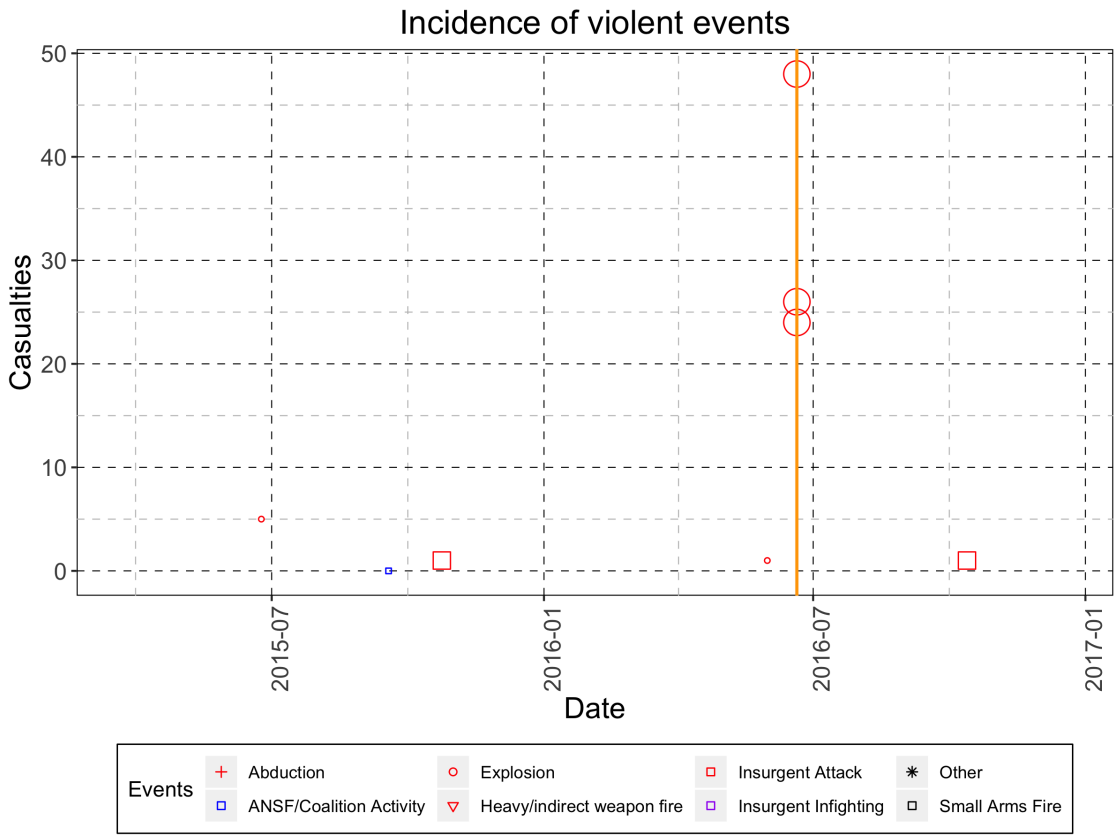


Figure 7.6: Temporal graph of violent events for a small town in Badakhshan (North East Afghanistan). Each point in the graph is an event where the shape and color of the point give information about the event.

to aggregate at the monthly level. However we have also performed analysis at the daily level to understand the robustness of our results and understand what, if any, differences exist in short and long-term changes.

Our independent variable will be whether the tower in question experienced violence during the time period in question. Our violence dataset includes casualties and deaths for a given event broken down by insurgents, ANSF and civilians. For our purposes we have decided to use the total number of deaths as the measure of violence: this would give us an integer valued independent variable. However, in the interest of starting with the simplest possible model we also decided that we should start by modelling violence as a boolean variable before moving to an integer valued variable. Unfortunately, due to high base level of violence in Afghanistan, understanding how to convert these integer-valued number of deaths into a boolean variable is far from simple. Simply deciding any time unit in which a tower has one or more deaths would cover too much of the dataset and might not capture the level of violence that is disruptive to inter-ethnic communications. Indeed by analyzing the distribution of deaths in all the available tower-months, as shown in figure 7.7, we see that this would include over 40% of tower months. We decided instead that any tower with 30 many or more deaths in a month (1 or more deaths for daily aggregation) was to be considered seriously impacted by violence. This was picked by using figure 7.7 to find a threshold that would separate the top 5 percentile by number of deaths.

We have defined the dependent variable as being the ethno-linguistic calling patterns of towers and we use the metric for both incoming and outgoing EL calling patterns defined in equation 7.1. In our analysis we will hence compute the distributions L_i^{out} and L_i^{in} as previously defined: using the total number of calls between two towers as the directed weight. Each distribution is a function over the 8 EL-groups recorded by the CSO survey (Dari, Pashto, Uzbek, Turkmen, Baloch, Pashaie, Nooristani, Other) and a final "unknown" group for towers we cannot infer over: this would give us a total of 18 possible dependant variables. However, we condensed this to four groups instead:

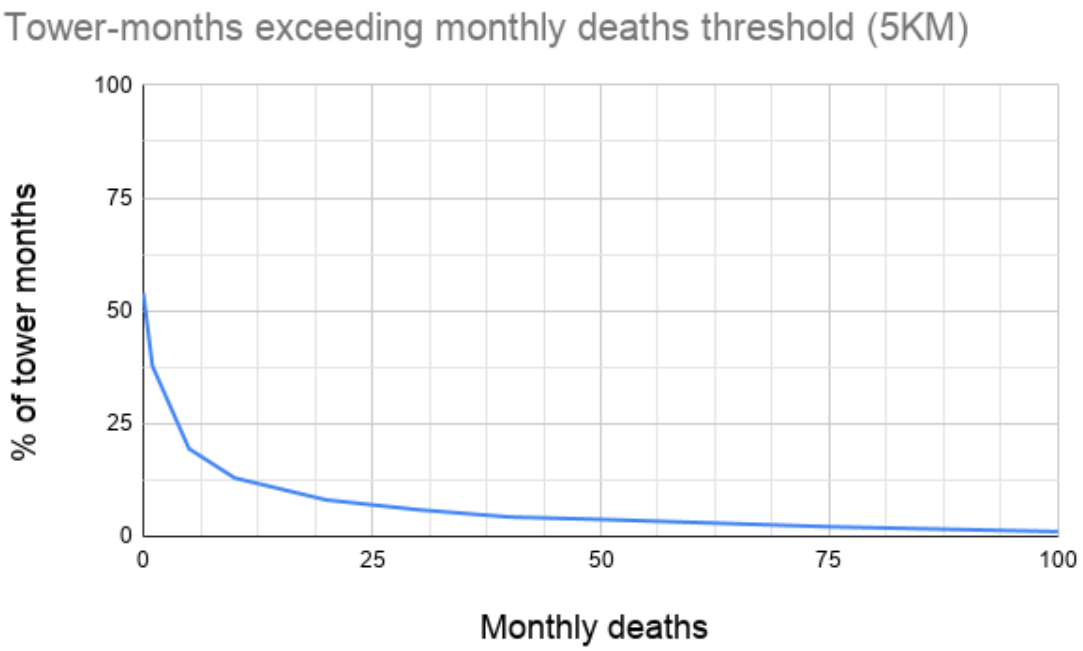


Figure 7.7: Graph showing the distribution of number of deaths per tower-month. Note that 46% of tower-months do not experience a single violent event. Meanwhile, about 6% of tower-months see 30 or more recorded deaths. Note that events occurring close to several towers will be assigned to all towers in range.

- Dari
- Pashto
- Other
- Mixed

where the "Other" group consists of the union of the other 6 EL groups recorded in the CSO survey and "Mixed" corresponds to towers with no clear majorty or an "unknown" assignment.

One additional caveat to consider is that not all calls are equally important in understanding the fragmentation of inter-linguistic bonds. Short calls made during the working day do not imply the same degree of closeness between the call participants as a long call made during the weekend for example. As such, we were curious to understand what would happen if we computed the ethno-linguistic call distributions as in equation 7.1 but using only a subset of the calls. Referring to the original formulations of the EL distribution that uses all the calls as $L_i^{out}(all)$ and $L_i^{in}(all)$, we introduce the distributions $L_i^{out}(long)$, $L_i^{in}(long)$ (which only considers calls of over a minute) and $L_i^{out}(nonwork)$, $L_i^{in}(nonwork)$ (which considers calls made during the weekend and outside of the working day). We expect broadly similar results on these dependant variables though we hypothesize the result will be stronger for the "nonwork" and "long" formulations.

Finally, it is important to discuss the confounder variables we will need to account for. We must start with a fixed effect for each tower, since the geographic and social context of a location will have a very strong influence on the ethno-linguistics of the towers it communicates with. Secondly, we also add a time fixed effect since we know that calling patterns change distinctly seasonally, especially during the month-long religious observance of Ramadan. Noting that religious and cultural observances can be very spatially sensitive, we add a third confounder variable: the cross product of each time period and district. This will enable us to detect and account for culturally specific events which are not as ubiquitous

as Ramadan but still influence calling behaviour. Finally, we consider the cross product of ethno-linguistic group and time-period for much the same reason: this being based on the same logic as discussed for our third confounder variable. Thus we will not use these two confounders together in the same regression, but examine the stability of the result when we use one or the other.

7.4.2 Regression equations

In this analysis we are working with panel data. This is as we have repeated measurements for a given tower i over several time periods: 24 months or 731 days, depending on the granularity we decide upon. As such, our first regression equation will be a simple panel fixed effects model where there are individual fixed effects for the tower and the period of time.

We use the standard notation that y_{it} represents the dependent variable for tower i at time t. There will be a dependent variable for each of the major ethno-lingusitic groups we identified (Dari, Pashto, Other, Unknown) and for different calling conditions (all calls, call outside of working hours, calls exceeding a minute of duration): this will be specified during the analysis of result but does not change the regression equations. X_{it} will represent the independent (binary) variable of whether there was a significant level of violence at tower i during period t and β the parameter we are interested in measuring. Finally, α_i and ϕ_t will be the fixed effect for tower and time period respectively. As such, we can can express our first regression equation for how violence impacts ethno-linguistic coherence as:

$$y_{it} = X_{it}\beta + \alpha_i + \phi_t + \mu_{it} \tag{7.2}$$

where μ_{it} represents the error term.

Regression equation 7.2 then takes 1,387 different clusters (towers). This gives us a total of 28,495 observations when aggregating at the monthly level and 867,110 observations when aggregating at the daily levels. We solved this using the lfe package[90] in R on a computing

cluster.

An different approach to the problem might be to make use of *lagged independent variables*. In this setup, the dependant variable of the previous time periods ($X_{i(t-1)}$) might occur in the equation explaining X_{it} for example. This makes intuitive sense for our regression performed on daily aggregations, where the effect may take time to propagate across the social network and possibly persist over several days.

Our first regression equation is just a straightforward application of lagged variables to the original equation 7.2. Let us define $X_{i,t-k:t}$ as the vector of size k containing $X_{i,t-k}, X_{i,t-k+1}...$ $X_{i,t-1}$. We can similarly define the vector $B_{i,t-k:t}$ as the corresponding parameters for each of the lagged variables. We can then write our new regression equation as:

$$y_{it} = X_{it}\beta + X_{i,t-k:t}B_{i,t-k:t} + \alpha_i + \phi_t + \mu_{it}, \tag{7.3}$$

with all other variables being defined as in equation 7.2. For monthly aggregations we will set k=4, meaning we will be using the dependent variables for the previous 4 months as lag terms. The daily regression will use k=45 and also included 5 lead terms (that use $X_{i(t+k)}$ instead of $X_{i(t-k)}$).

In addition to the lagged dependant variables, we will introduce a few different changes to how we attempt to handle temporal fixed effects. As discussed in section 7.4.1, there is significant hetereogenity in how different areas of Afghanistan experience a given time slice, both in the matter of cultural norms and local geo-political conditions. To model the latter, we introduce a district-time fixed effect: $\phi_{d,t}$ where d is the district to which the tower corresponds. Replacing the previous time fixed effect with this gives us the equation:

$$y_{it} = X_{it}\beta + X_{i,t-k:t}B_{i,t-k:t} + \alpha_i + \phi_{d,t} + \mu_{it}. \tag{7.4}$$

To address the problem of culture-time variation we introduce a similar fixed effect for ethno-linguistic group and time $\pi_{g,t}$ where g is the main ethno-linguistic group if the tower has more

CDR dataset	Dari	Pashto	Other	Mixed
All Outgoing calls	-0.09	-0.161^{*}	0.018	0.234^{***}
All Incoming calls	0.050	-0.124	0.092^{*}	-0.067
Non-work Outgoing calls	0.210^{***}	-0.124	-0.022	0.362^{***}
Non-work Incoming calls	-0.040	-0.153	0.063	0.129

Table 7.1: Summary of % weighted calls to/from given ethnic groups from monthly panel fixed effects model. The number denotes the measured change (in percentage points) and the asterisks define the p-value associated with the change (*p<0.1; **p<0.05; ***p<0.01).

than 70% membership in this group and "unknown" otherwise. This gives us the final variant of our lagged regression equation:

$$y_{it} = X_{it}\beta + X_{i,t-k:t}B_{i,t-k:t} + \alpha_i + \pi_{g,t} + \mu_{it}. \tag{7.5}$$

As such, we have three equations for our lagged model, each of which have almost the same number of observations as the un-lagged equation 7.2 (a few towers come online only partway through our time period).

7.5 Regression results

We begin by analyzing the basic panel fixed effects model as described in equation 7.2 without lag or lead terms. For each possible formulation of the equation (which ethno-linguistic group to focus on, what calls to consider) we performed a full analysis: an illustrative example for outgoing calls to Dari towers at the monthly level is given in table 7.5. However, for brevity, we condensed the results to focus on the parameter of interest β along with the statistical significance in table 7.1 (monthly results) and in table 7.2 (for daily results).

Observing these two tables, we note that the results for the daily aggregations are much stronger. This makes sense intuitively, since this would capture the impact of violence at its most timely, whereas the effect may attenuate over a month if violence occurs towards the start of the month or only be present for a fraction of the time if the violence occurs after the start of the month. Table 7.2 certainly shows strong evidence that the proportion of calls

CDR dataset	Dari	Pashto	Other	Mixed
All Outgoing Calls	-0.184^{***}	-0.256^{***}	-0.004	0.428^{***}
All Incoming Calls	-0.050	-0.208^{***}	0.062^{***}	0.211^{***}
Non-work Outgoing calls	-0.15	-0.076	0.033^{*}	0.436^{***}
Non-work Incoming calls	0.032	-0.160^{***}	0.075^{***}	0.033

Table 7.2: Summary of % weighted calls to/from given ethnic groups from daily panel fixed effects. The number denotes the measured change (in percentage points) and the asterisks define the p-value associated with the change (*p<0.1; **p<0.05; ***p<0.01).

going to/from Pashto towers is significantly decreased (though the magnitude of the effect is modest) by violence: seemingly shifting towards urban areas and more heterogeneous areas ("Mixed" towers) as well as areas inhabited predominately by less populous ethno-linguistic groups ("Other" towers).

Now we will consider the monthly results for panel fixed effects with lag variables (equations 7.3 to 7.5). Since the lag term for k is only k=4, it is possible to summarize the results for both β and B in the same table, as seen in table 7.3. Table 7.3 summarizes the results for equation 7.5, as we felt that the district-time fixed effect did the best job of capturing the spatially hetereogenous effects of time as different districts were impacted by fighting at different times. However, the results for equations 7.3 and 7.4 were broadly similar.

With this more sophisticated regression equation we see that the impact of calls to and from Pashto communities is much more readily apparent. We see a statistically significant long-lasting reduction of up to 4 months (for outgoing calls) and 2 months (for incoming calls) that is not apparent for either Daris nor for the other two categories. In particular, the effect seems to be more pronounced for the non-work calls (which might include a greater proportion of personal calls that signify social connection) than for the overall calls. for the outgoing calls we can see that calls that would previously have been going to Pashto areas seem instead to be going to either Dari areas (although the trend here is ambiguous) and "other" ethno-linguistic communities. A similar trend might be occurring in terms of received incoming calls but the evidence is less clear. One concern that must be addressed is

CDR dataset	Variable	Dari	Pashto	Other	Mixed
All Outgoing Calls					
	β	0.014	-0.251^{***}	0.092^{***}	0.145
	B_{-1}	0.193^{***}	-0.198^{***}	0.071^{**}	-0.005
	B_{-2}	0.206^{***}	-0.164^{**}	0.074^{***}	-0.106
	B_{-3}	0.014	-0.257^{***}	0.064^{**}	0.179
	B_{-4}	0.022	-0.222^{***}	0.004	0.160
All Incoming Calls					
	β	-0.05	-0.251^{***}	0.074	0.227
	B_{-1}	0.052	-0.196^{***}	-0.007	0.151
	B_{-2}	0.219^{**}	-0.240^{**}	0.085	-0.023
	B_{-3}	-0.126	-0.149	0.111	-0.003
	B_{-4}	0.054	-0.116	0.102^{*}	-0.041
Non-work Outgoing Calls					
	β	0.030	-0.269^{***}	0.102^{***}	0.137
	B_{-1}	0.219^{*}	-0.207^{***}	0.073^{**}	-0.085
	B_{-2}	0.142	-0.229^{**}	0.080^{***}	0.007
	B_{-3}	-0.027	-0.272^{***}	0.071^{**}	0.228^{*}
	B_{-4}	0.081	-0.164^{***}	0.041	0.042
Non-work Incoming Calls					
	β	0.013	-0.296^{***}	0.114^{**}	0.169
	B_{-1}	0.167^{*}	-0.211^{**}	0.085^{**}	-0.051
	B_{-2}	0.214^{**}	-0.213^{**}	0.072	-0.072
	B_{-3}	-0.136	-0.188^{*}	0.076^{*}	0.248
	B_{-4}	0.057	-0.220^{**}	0.057	0.112

Table 7.3: Summary of % weighted calls to/from given ethnic groups from monthly lagged model. The number denotes the measured change (in percentage points) and the asterisks define the p-value associated with the change ($^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$). Note that B_{-i} represents the coefficient of the ith lagged independent variable.

CDR dataset	Dari	Pashto	Other	Mixed
All Outgoing Calls	-0.153^{***}	-0.166^{***}	0.012	-0.036
All Incoming Calls	-0.015	-0.133^{***}	0.023	0.138^{***}
Non-work Outgoing calls	-0.112^{**}	-0.147^{**}	-0.013	-0.043
Non-work Incoming calls	0.053	-0.100^{**}	0.021	0.004

Table 7.4: Summary of % weighted calls to/from given ethnic groups from daily lagged model. The number denotes the measured change (in percentage points) and the asterisks define the p-value associated with the change (*p<0.1; **p<0.05; ***p<0.01). Results only shown for the β coefficient, not the B_{-i} lag coefficients.

that the impact of violence is not monotonically decreasing as a function of the lag distance. We would expect that $B_{-i} > B_{-(i+1)}$ generally but in fact, we see that for outgoing calls this is not always the case. We discuss possible causes of this in section 7.6.

For the daily equations we will again look at results from the formulation of equation 7.5 for the same reasons as in the monthly case. As the daily versions of equations 7.3 to 7.5 have k=45 lag terms, it would be impossible to analyze all the coefficients as in table 7.3. Instead we analyze the values of the main coefficient β in table 7.4. To analyze the k lag variables we instead plot them, along with the uncertainty in their values, as shown in figure 7.8.

The results for β in the lagged equation shown in table 7.4 broadly follow the trends seen in table 7.2. The calls to and from Pashto areas during days of heavy violence has a consistent and statistically significant change, though the magnitude of the change is somewhat smaller than for the panel fixed effects model. This is understandable, as we expect that this model is able to account for periods of time with multiple consecutive days of violence (with a likely cumulative effect on calling patterns) in a more comprehensive manner than the previous model.

Figure 7.8 also shows a much clearer and long-lasting impact on outgoing calls to Pashto areas than to others. We note that all the most recent lag coefficients up to B_{-16} have a significant negative values, implying that calls to Pahsto areas are depressed for over two weeks after a day of significant violence. The effect of calls to Dari areas (also shown in

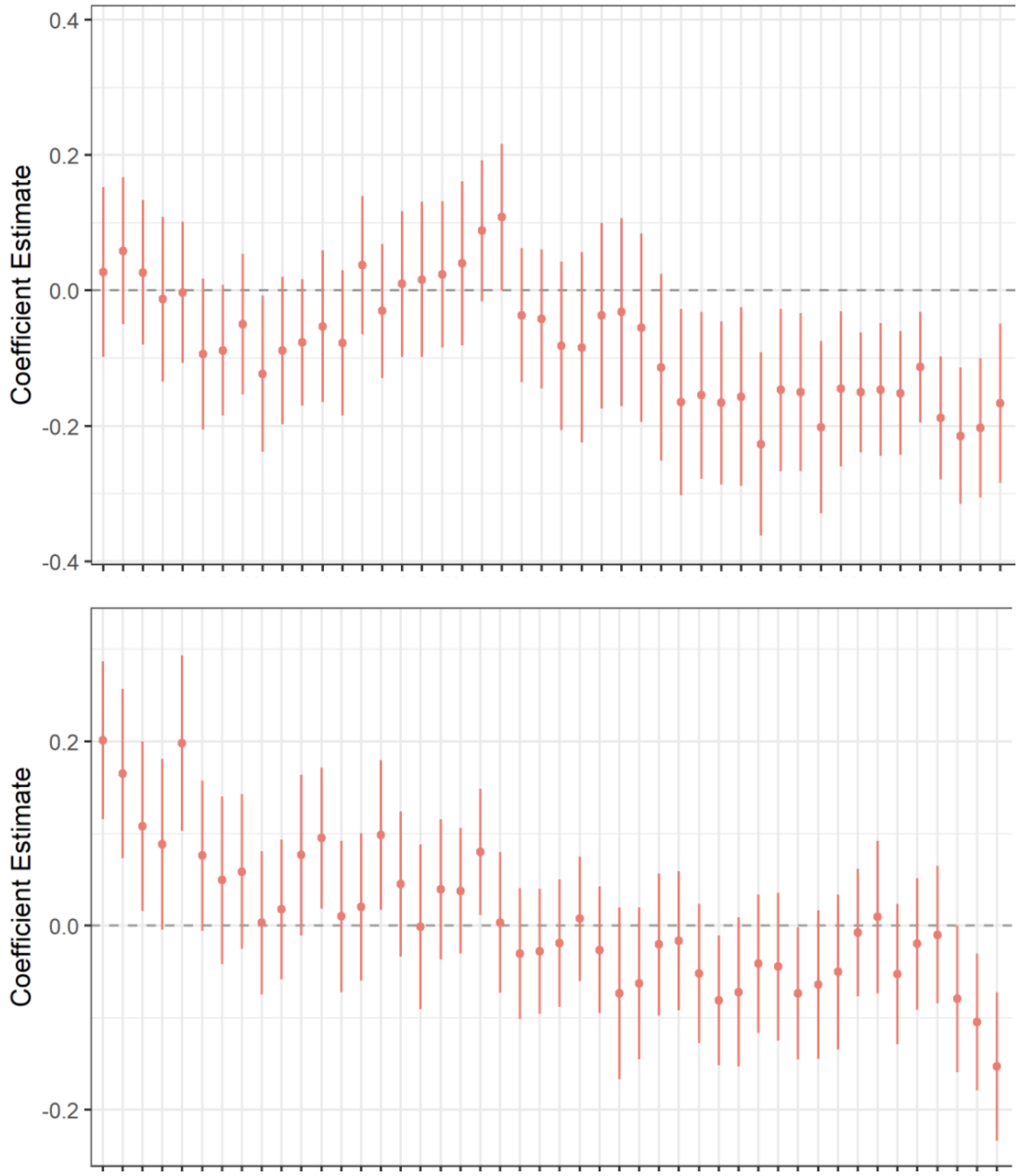


Figure 7.8: Graph comparing the two lag coefficient plots (B_{-i}) of Pashto (top) and Dari (bottom). The plots show the estimated value of each coefficient from B_{-45} on the far left to β on the far right along with an error bar corresponding to the uncertainty of the estimate. These coefficient plots are drawn from applying regression equation 7.5 to the set of all outgoing calls.

figure 7.8) is only statistically significant for the last two lag terms and hence are hard to distinguish from the immediate short-term disruption known to occur after a violent event. The calls to mixed-group and other areas (not included in figure 7.8 for clarity purposes) do not show consistent significant coefficient values.

7.6 Interpreting results and future work

Though the results in section 7.5 have provided an excellent base to build upon additional work to be done in order to substantiate our original thesis. We wished to show that heavy levels of violence can lead to the fragmentation of links between ethno-linguistics groups. We hypothesized that this would be particularly pronounced when considering links between the Pashto community, from whom the Taliban predominantly draws its recruits, and other groups. We have indeed found substantial evidence that the volume of communication to/from Pashto areas is substantially reduced after violence and that the reduction is much more substantial than for other groups. We noted that there appears to be both a more immediate component to this, noticeable on the daily level for two weeks, as well as a more long lasting effects covering multiple months. A plausible explanation for these results is indeed that it is a by-product of inter-linguistic ties being severed by animosity engendered by additional violence. However, there are different plausible causes that must be examined and ruled out as well as a few caveats that do not seem to support these conditions. We will proceed to list these points, explain their provenance and propose future work that could resolve or clarify them.

7.6.1 Confirming inter-linguistic fragmentation via dyadic regression

Though we have seen that calls to and from Pashto areas has dropped in the aftermath of violence, this does not necessarily show that this drop originated from other ethno-linugisitic groups. It is plausible that this drop could have come from Pashto towers themselves, which seek instead to contact friends or relations residing in areas with a primarily different ethno-linguisitic group. Essentially, it is important to confirm that the drop in calls is due

to fewer Dari/Uzbek/Turkmen areas calling Pashto areas rather than Pashto areas making proportionally fewer calls amongst themselves.

The ideal approach to this problem will be to consider the dyads in question as opposed to just the two towers themselves. In this way, instead of labelling communication simply as a function of the incoming (or outgoing tower) one could instead label communication as "co-linugistic" or "cross-lingusitic" based on a similarity metric applied to the ethno-linguistic breakdown of both towers. This could then be aggregated back to the tower, allowing us to see the percent change in the proportion of cross-linguistic calls and how it is impacted by violence. Alternatively, one could directly perform a *dyadic regression*[98] instead of using towers as the cluster unit.

7.6.2 Outlier detection for independent variable assignment

As discussed in section 7.2.1, violence in Afghanistan shows a high degree of spatial variability. Picking two example districts, we note that what might be a violent time period in the Achin district in Nangarhar (average recorded deaths per week: 10.8) might be very different from the northern district of Darzab (average recorded deaths per week: 0.6). This observation interacts problematically with our current method of setting a hard threshold for violence. If we are measuring a reaction to perception of violence it is important to account for the background violence. A constant threshold may hence over-represent very violent districts (where there may be a degree of habituation to political violence) vis-a-vis less violent districts where a smaller number of deaths may provoke a larger response. More problematically, this problem is remains even if the binary independent variable is replaced with a continuous one such as the total number of deaths.

One approach to solve these issues may be to use a district-level probabilistic model that accounts for the background rate of violence. Such a model would take the total number of deaths for each district, look for outlying periods of atypically high violence and then check which towers were responsible for this outburst of violence and so give us a more sensitive binary independent variable. One such approach might be to model the temporal weekly

sequences of deaths using a Poisson point process [109] which allows us to determine the likelihood of each individual week. Any week which has a probability of less than $\alpha = 0.001$ is determined to be an outlier and this positive result is propagated back to the towers that contributed to the violence during this period.

Table 7.5: An example of the output of a regression analysis. This table is for the regression of all outgoing calls aggregated at the monthly level.

	% Out Calls: Pashto All Hours			
	(1)	(2)	(3)	(4)
Violence (month t)	-0.161^{*}	-0.156^{**}	-0.375^{***}	-0.251^{***}
	(0.089)	(0.074)	(0.086)	(0.077)
t − 1		−0.093	-0.342^{***}	-0.198^{***}
		(0.060)	(0.072)	(0.072)
t − 2		0.127^{**}	−0.044	-0.164^{**}
		(0.063)	(0.076)	(0.076)
t − 3		-0.522^{***}	-0.605^{***}	-0.257^{***}
		(0.084)	(0.089)	(0.081)
t − 4		-0.295^{***}	-0.336^{***}	-0.222^{***}
		(0.055)	(0.059)	(0.072)
Tower FE	Y	Y	Y	Y
Time FE	Y	Y	-	-
Ethnicity × Time FE	-	-	Y	-
District × Time FE	-	-	-	Y
Mean Dep Var	26.269	26.501	26.501	26.501
SD Dep Var	23.863	23.972	23.972	23.972
Full R^2	0.992	0.992	0.993	0.997
Full Adj R^2	0.991	0.992	0.992	0.996
Projected R^2	0.068	0.068	0.151	0.594
Projected Adj R^2	0.02	0.018	0.102	0.481
F Stat	140.104	54.354	20.879	473.07
N Clusters	1387	1385	1385	1385
Observations	28,495	27,183	27,183	27,183
Residual Std. Error	2.241	2.174	2.080	1.581

Notes: Standard errors clustered at the tower level. Columns 2-4 include 4 lags. Month-year time FEs, and ethnic group × month-year FEs are included. Towers are assigned a majority ethnic group (Dari, Pashto, Other, or Mixed). Dependent variable: % of total out calls to the relevant ethno-linguistic group. Independent variable: 1 if a tower experiences >30 deaths in a given month, 0 otherwise. Deaths are counted within a 5KM radius of a given tower. *p<0.1; **p<0.05; ***p<0.01

Chapter 8

CONCLUSION

This thesis has described a series of projects which have used Big Data in the context of the developing world. In chapter 2 we introduced the two main types of passively collected data we worked with: transactional databases like Call Detail Records and remote sending data like satellite imagery. We have shown how the former can be used to study phenomena ranging from material wealth (chapter 3) to the impact of violence in society (chapter 7 7). We also explored the statistical nuances of making inferences of such datasets and the potential danger of drawing incorrect conclusion from them in chapter 6. In addition to social and economic inferences we also showed how infrastructure can be analyzed at scale using satellite imagery in chapter 4. The applications of these projects have ranged from standard machine learning predictions to causal inferences for policy (as introduced in chapter 5) but have always been driven by an end-goal of providing useful and actionable information to developing regions.

One point that can be drawn from looking at these projects as a whole is the importance of breadth in this field. This is true in a technical sense: the approaches to the problems described in this work have varied from statistical tests and probabilistic proofs all the way to temporal modelling and deep neural nets. But it is also true from a sense of perspective: in some cases a narrow ability to predict a certain quantity is desired, while in others a causal inference and understanding is essential. In addition to the standard toolkits of machine learning and statistics, several of the works here have relied on sociological and economic knowledge than is more common in the space of information technology for development. For example, knowing how to sample real-life populations in an unbiased way or understanding the social and linguistics components of a country form a key component of several projects

in this work. Though the data scientist need not necessarily be an expert in these fields, enough familiarity and competence to understand how these considerations will impact the machine learning components of a project is essential.

A final point to consider is that trends in the area of data science in developing countries are broadly optimistic. Communication infrastructure, both mobile phones and internet connectivity, in these regions has grown significantly in the last decade and will penetration rates will approach that of the developed worlds. This will mean that analysis performed on CDRs, social media sites etc will become increasingly large and representative of the population as a whole. While issues such as less-well represented languages may continue to be a problem when it comes to NLP applications it will not impact analysis of metadata (which underlies the type of projects discussed in chapter 3 and 7). This positive trend in data quality and quantity is even more noticeable in the realm of remote sensing data. In addition to the well-known sources of public data such as LANDSAT, governments have added new sources such as SENTINEL (launched 2015-2017) which are open to researchers across the world. The development in the private sector has been even more dramatic: for example Planet has launched upwards of 150 imaging satellites in the past decade. While using this data for academic purposes is not as straightforward as it is in using publicly available data sources, there have been multiple research papers making use of these data sets to great effect. We hope that this inexorable trend towards better data, coupled with a continuing interest in using data science to understand developing communities, will make a positive and lasting impact on the lives of the billions who live in these regions.

Acknowledgements

Some of the research in this thesis was supported by the National Science Foundation Grant under award CCF - 1637360 (Algorithms in the Field) and by the Office of Naval Research (Minerva Initiative) under award N00014-17-1-2313.

BIBLIOGRAPHY

[1] CrowdAI. https://crowdai.com.

[2] How much data is created on the internet each day? https://https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/.

[3] Practical Guide to Event Studies. https://github.com/setzler/eventStudy/blob/master/guide/event_study_guide.pdf.

[4] PyTorch model zoo. https://pytorch.org/docs/stable/torchvision/models.html.

[5] Street Bump. www.streetbump.org.

[6] Survey of the Afgan People 2018 . http://asiafoundation.org/where-we-work/afghanistan/survey/.

[7] The Mobile Economy Sub-Saharan Africa 2019. https://www.gsma.com/r/mobileeconomy/sub-saharan-africa/.

[8] University of Minnesota Introduction to Satellite Imagery. https://www.pgc.umn.edu/guides/commercial-imagery/intro-satellite-imagery/.

[9] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[10] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

[11] Jade Abbott and Laura Martinus. Benchmarking neural machine translation for southern african languages. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 98–101, 2019.

[12] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pages 11–21, 2017.

[13] Adedamola Adepetu and Jay Taneja. Filling spatial and temporal gaps in development surveys using night lights. In *UNESCO Chair Conference on Technologies for Development (Tech4Dev 2016)*, 2016.

[14] Adrian Albert, Jasleen Kaur, and Marta C Gonzalez. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1357–1366. ACM, 2017.

[15] Adrian Albert, Jasleen Kaur, and Marta C. Gonzalez. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Knowledge Discovery and Data Mining*, 2017.

[16] Anna Aldeghi, Simon Carn, Rudiger Escobar-Wolf, and Gianluca Groppelli. Volcano monitoring from space using high-cadence planet cubesat images applied to fuego volcano, guatemala. *Remote Sensing*, 11(18):2151, 2019.

[17] John Aldrich et al. Ra fisher and the making of maximum likelihood 1912-1922. *Statistical science*, 12(3):162–176, 1997.

[18] Yaniv Altshuler, Michael Fire, Erez Shmueli, Yuval Elovici, Alfred Bruckstein, Alex Sandy Pentland, and David Lazer. The social amplifier—reaction of human communities to emergencies. *Journal of Statistical Physics*, 152(3):399–418, 2013.

[19] Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S Spiro. How information snowballs: Exploring the role of exposure in online rumor propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 466–477. ACM, 2016.

[20] Simplice Asongu. The impact of mobile phone penetration on african inequality. *International Journal of Social Economics*, 42(8):706–716, 2015.

[21] Simplice A Asongu and Jacinta C Nwachukwu. The role of governance in mobile phones for inclusive human development in sub-saharan africa. *Technovation*, 55:1–13, 2016.

[22] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research, 2018.

[23] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

[24] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[25] Jining Bao, Yunzhou Zhang, Xiaolin Su, and Rui Zheng. Unpaved road detection based on spatial fuzzy clustering algorithm. *Journal on Image and Video Processing*, 26, 2018.

[26] Thomas Barfield. *Afghanistan:A cultural and Political History*. Princeton University Press, 2010.

[27] Ghazaleh Beigi, Xia Hu, Ross Maciejewski, and Huan Liu. An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment analysis and ontology engineering*, pages 313–340. Springer, 2016.

[28] Eli Berman, Jacob N Shapiro, and Joseph H Felter. Can hearts and minds be bought? the economics of counterinsurgency in iraq. *Journal of Political Economy*, 119(4):766–819, 2011.

[29] Timothy Besley and Hannes Mueller. Estimating the peace dividend: The impact of violence on house prices in northern ireland. *American Economic Review*, 102(2):810–33, 2012.

[30] Monica Beuran, Marie Castaing Gachassin, and Gael Raballand. Are there myths on road impact and transport in sub-saharan africa? 2013.

[31] Daniel Björkegren and Darrell Grissen. Behavior revealed in mobile phone usage predicts loan repayment. *Available at SSRN 2611775*, 2018.

[32] Christopher Blattman and Edward Miguel. Civil war. *Journal of Economic literature*, 48(1):3–57, 2010.

[33] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

[34] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.

[35] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[36] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[37] Joshua Blumenstock and Nathan Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 6. ACM, 2010.

[38] Joshua Blumenstock and Lauren Fratamico. Social and spatial ethnic segregation. In *DEV-4 ACM Conference, Cape Town, South Africa*, 2013.

[39] Joshua Blumenstock, Tarek Ghani, Sylvan Herskowitz, Ethan B Kapstein, Thomas Scherer, and Ott Toomet. *Insecurity and industrial organization: Evidence from Afghanistan*. The World Bank, 2018.

[40] Joshua Blumenstock, Ott Toomet, Rein Ahas, and Erki Saluveer. Neighborhood and network segregation: Ethnic homophily in a silently separate society. *Proc. NetMob*, 2015.

[41] Joshua E Blumenstock. Using mobile phone data to measure the ties between nations. In *Proceedings of the 2011 iConference*, pages 195–202. ACM, 2011.

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

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

[44] Joshua Evan Blumenstock. Fighting poverty with data. *Science*, 353(6301):753–754, 2016.

[45] Joshua Evan Blumenstock and Nathan Eagle. Divided we call: disparities in access and use of mobile phones in rwanda. *Information Technologies & International Development*, 8(2):pp–1, 2012.

[46] 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.

[47] Gabriel Cadamuro, Ramya Korlakai Vinayak, Joshua Blumenstock, Sham Kakade, and Jacob Shapiro. The illusion of change: Correcting for biases in change inference for sparse, societal-scale data. In *The World Wide Web Conference*, pages 2608–2615. ACM, 2019.

[48] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. Assigning a grade: Accurate measurement of road quality using satellite imagery. *arXiv preprint arXiv:1812.01699*, 2018.

[49] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. Street smarts: measuring intercity road quality using deep learning on satellite imagery. In *Proceedings of the Conference on Computing & Sustainable Societies*, pages 145–154. ACM, 2019.

[50] Christopher Carpenter and Carlos Dobkin. The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, 1(1):164–82, 2009.

[51] Rich Caruana and Alexandru Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–78. ACM, 2004.

[52] Jesse Casana and Elise Jakoby Laugier. Satellite imagery-based monitoring of archaeological site damage in the syrian civil war. *PloS one*, 12(11):e0188589, 2017.

[53] David W Casbeer, Randal W Beard, Timothy W McLain, Sai-Ming Li, and Raman K Mehra. Forest fire monitoring with multiple small uavs. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 3530–3535. IEEE, 2005.

[54] Ray M Chang, Robert J Kauffman, and YoungOk Kwon. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67–80, 2014.

[55] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[56] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.

[57] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28, 2016.

[58] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[59] Luke N Condra, James D Long, Andrew C Shaver, and Austin L Wright. The logic of insurgent electoral violence. *American Economic Review*, 108(11):3199–3231, 2018.

[60] Sergio Currarini, Matthew O Jackson, and Paolo Pin. Identifying sources of racial homophily in high school friendship networks. In *Proceedings of the National Academy of Science of the USA*, volume 107, pages 4857–4861, 2010.

[61] Dharma Dailey and Kate Starbird. Social media seamsters: Stitching platforms & audiences into local crisis infrastructure. In *CSCW*, pages 1277–1289, 2017.

[62] Giacomo De Luca and Marijke Verpoorten. Civil war, social capital and resilience in uganda. *Oxford Economic Papers*, 67(3):661–686, 2015.

[63] Yves-Alexandre de Montjoye, Luc Rocher, Alex Sandy Pentland, et al. bandicoot: A python toolbox for mobile phone metadata. *J Machine Learn Res*, 17:1–5, 2016.

[64] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *DeepGlobe Workshop at CVPR (DeepGlobe 2018)*, 2018.

[65] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[66] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[67] DigitalGlobe. Basemap +Vivid Product. `dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/2/DG_Basemap_Vivid_DS_1.pdf`.

[68] Matthew Dixon, Spencer P Aiello, Funmi Fapohunda, and William Goldstein. Detecting mobility patterns in mobile phone data from the ivory coast. 2013.

[69] Adrian Dobra, Nathalie E Williams, and Nathan Eagle. Spatiotemporal detection of unusual human population behavior using mobile phone data. *PloS one*, 10(3):e0120449, 2015.

[70] Christopher NH Doll, Jan-Peter Muller, and Jeremy G Morley. Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, 57(1):75–92, 2006.

[71] Dave Donaldson and Adam Storeygard. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98, 2016.

[72] Xiaowen Dong, Joachim Meyer, Erez Shmueli, Burçin Bozkaya, and Alex Pentland. Methods for quantifying effects of social unrest using credit card transaction data. *EPJ Data Science*, 7(1):8, 2018.

[73] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 15–24. ACM, 2014.

[74] Esther Duflo. Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American economic review*, 91(4):795–813, 2001.

[75] Gregory Duveiller, Pierre Defourny, Baudouin Desclée, and P Mayaux. Deforestation in central africa: Estimates at regional, national and landscape levels by advanced processing of systematically-distributed landsat extracts. *Remote sensing of environment*, 112(5):1969–1981, 2008.

[76] Nathan Eagle, Yves-Alexandre de Montjoye, and Luís MA Bettencourt. Community computing: Comparisons between rural and urban societies using mobile phone data. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 144–150. IEEE, 2009.

[77] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[78] Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

[79] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[80] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[81] C Christine Fair, Patrick Kuhn, Neil A Malhotra, and Jacob Shapiro. Natural disasters and political engagement: evidence from the 2010–11 pakistani floods. 2017.

[82] James D Fearon. Ethnic mobilization and ethnic violence. *The Oxford handbook of political economy*, pages 852–868, 2006.

[83] Andrey Feuerverger, Yu He, and Shashi Khatri. Statistical significance of the netflix challenge. *Statistical Science*, pages 202–231, 2012.

[84] Ronald A Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

[85] Lars Forslöf and Hans Jones. Roadroid: Continuous road condition monitoring with smart phones. *Journal of Civil Engineering and Architecture*, 9(4):485–496, 2015.

[86] Lars Forslof and Hans Jones. Roadroid: Continuous road condition monitoring with smart phones. *Journal of Civil Engineering and Architecture*, 9:485–496, 2015.

[87] Vanessa Frias-Martinez, Enrique Frias-Martinez, and Nuria Oliver. A gender-centric analysis of calling behavior in a developing economy using call detail records. In *2010 AAAI Spring Symposium Series*, 2010.

[88] Vanessa Frias-Martinez and Jesus Virseda. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development*, pages 76–84. ACM, 2012.

[89] Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Socio-economic levels and human mobility. In *Qual meets quant workshop-QMQ*, 2010.

[90] Simen Gaure. lfe: Linear group fixed effects. *The R Journal*, 5(2):104–117, 2013.

[91] Felix A. Gers, Jiirgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. In *Artificial Neural Networks*, 1999.

[92] Dimitry Gershenson, Brandon Rohrer, and Anna Lerner. Medium-Voltage Distribution (Predictive). `https://energydata.info/dataset/medium-voltage-distribution-predictive`.

[93] Caroline Gevaert, Richard Sliuzas, Claudio Persello, and George Vosselman. Evaluating the societal impact of using drones to support urban upgrading projects. *ISPRS international journal of geo-information*, 7(3):91, 2018.

[94] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[95] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

[96] IJ Good and GH Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[97] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. Omnipress, 2010.

[98] Bryan S Graham. Dyadic regression. *arXiv preprint arXiv:1908.09029*, 2019.

[99] Ilias Grinias, Costas Panagiotakis, and Georgios Tziritas. Mrf-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS journal of photogrammetry and remote sensing*, 122:145–166, 2016.

[100] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[101] Didem Gundogdu, Ozlem D Incel, Albert A Salah, and Bruno Lepri. Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science*, 5(1):25, 2016.

[102] Ken Gwilliam and Zmarak Shalizi. Road Funds, User Charges, and Taxes. *The World Bank Research Observer*, 14(2):159–186, 1999.

[103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[104] J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, 2012.

[105] Sahar Hoteit, Guangshuo Chen, Aline Viana, and Marco Fiore. Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pages 45–50. ACM, 2016.

[106] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[107] ICPAC. Kenya Land Use Data. `geoportal.icpac.net/layers/geonode\%3Akenyalandcover2015`, 2015.

[108] Idaho Transportation Department. Pavement Performance Report. `apps.itd.idaho.gov/apps/pm/ITD\%202015\%20Performance\%20Report.pdf`, 2015.

[109] Alexander Ihler, Jon Hutchins, and Padhraic Smyth. Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 207–216, 2006.

[110] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):67, 2015.

[111] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015.

[112] William Jack and Tavneet Suri. Mobile money: The economics of m-pesa. Technical report, National Bureau of Economic Research, 2011.

[113] Brian A Jacob and Lars Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *Review of economics and statistics*, 86(1):226–244, 2004.

[114] Aditya Jain, Zerina Kapetanovic, Akshit Kumar, Vasuki Narasimha Swamy, Rohit Patil, Deepak Vasisht, Rahul Sharma, Manohar Swaminathan, Ranveer Chandra, Anirudh Badam, et al. Low-cost aerial imaging for small holder farmers. In *Proceedings of the Conference on Computing & Sustainable Societies*, pages 41–51. ACM, 2019.

[115] Japanese International Cooperation Agency. Summary of Terminal Evaluation. www.jica.go.jp/english/our_work/evaluation/tech_and_grant/project/term/africa/c8h0vm000001rp75-att/kenya_2015_01.pdf.

[116] Seema Jayachandran. Air quality and early-life mortality evidence from indonesias wildfires. *Journal of Human resources*, 44(4):916–954, 2009.

[117] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[118] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[119] Robert Jensen. The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *The quarterly journal of economics*, 122(3):879–924, 2007.

[120] Morton Jerven. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It*. Cornell Univ. Press, 2013.

[121] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l1 distance. *IEEE Transactions on Information Theory*, 2018.

[122] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[123] Michael D Johnson, William W Hsieh, Alex J Cannon, Andrew Davidson, and Frédéric Bédard. Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and forest meteorology*, 218:74–84, 2016.

[124] Lynn H Kaack, George H Chen, and M Granger Morgan. Truck traffic monitoring with satellite images. In *Proceedings of the Conference on Computing & Sustainable Societies*, pages 155–164. ACM, 2019.

[125] Ashish Kapoor, Nathan Eagle, and Eric Horvitz. People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In *AAAI spring symposium: artificial intelligence for development*, 2010.

[126] Richard J Kauth and GS Thomas. The tasselled cap–a graphic description of the spectral-temporal development of agricultural crops as seen by landsat. In *LARS symposia*, page 159, 1976.

[127] Muhammad Raza Khan, Joshua Manoj, Anikate Singh, and Joshua Blumenstock. Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty. In *Big Data (BigData Congress), 2015 IEEE International Congress on*, pages 677–680. IEEE, 2015.

[128] Yong-Deok Kim and Seungjin Choi. Nonnegative tucker decomposition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[129] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[130] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[131] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[132] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[133] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3):e1500779, 2016.

[134] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. Rumor detection over varying time windows. *PloS one*, 12(1):e0168344, 2017.

[135] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[136] Gary LaFree and Laura Dugan. Introducing the global terrorism database. *Terrorism and Political Violence*, 19(2):181–204, 2007.

[137] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[138] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[139] Yu-Ru Lin and David Lazer. The effect of social contexts on network response to emergencies. Technical report, HARVARD UNIV CAMBRIDGE MA, 2011.

[140] David Lindenbaum. Introducing the SpaceNet Road Detection and Routing Challenge and Dataset. `https://medium.com/the-downlinq/introducing-the-spacenet-road-detection-and-routing-challenge-and-dataset-7604de39`

[141] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, 2016.

[142] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

[143] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[144] Shiwei Lu, Zhixiang Fang, Xirui Zhang, Shih-Lung Shaw, Ling Yin, Zhiyuan Zhao, and Xiping Yang. Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. *ISPRS International Journal of Geo-Information*, 6(1):7, 2017.

[145] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[146] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[147] Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan Bollen. Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to côte d'ivoire. *EPJ Data Science*, 4(1):15, 2015.

[148] Hank A Margolis, Ross F Nelson, Paul M Montesano, André Beaudoin, Guoqing Sun, Hans-Erik Andersen, and Michael A Wulder. Combining satellite lidar, airborne lidar, and ground plots to estimate the amount and distribution of aboveground biomass in the boreal forest of north america. *Canadian Journal of Forest Research*, 45(7):838–855, 2015.

[149] George Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 1955.

[150] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[151] Volodymyr Mnih and Geoffrey E. Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision (ECCV 2010)*, 2010.

[152] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pages 210–223. Springer, 2010.

[153] A Morales, W Creixell, J Borondo, J Losada, and R Benito. Understanding ethnical interactions on ivory coast. In *Proceedings of the Third Conference on the Analysis of Mobile Phone Datasets*, pages 116–122, 2013.

[154] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161, 2016.

[155] Diala Naboulsi, Razvan Stanica, and Marco Fiore. Classifying call profiles in large-scale mobile traffic datasets. In *INFOCOM, 2014 Proceedings IEEE*, pages 1806–1814. IEEE, 2014.

[156] Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga. Combining satellite imagery and open data to map road safety. In *AAAI*, pages 4524–4530, 2017.

[157] NASA. Suomi NPP VIIRS Land. `https://viirsland.gsfc.nasa.gov/index.html`.

[158] NOAA EOG. Version 1 VIIRS Day/Night Band Nighttime Lights. `https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html`.

[159] Anthony Oberschall. *Conflict and peace building in divided societies: Responses to ethnic violence*. Routledge, 2007.

[160] Benjamin A Olken. Monitoring corruption: evidence from a field experiment in indonesia. *Journal of political Economy*, 115(2):200–249, 2007.

[161] Alon Orlitsky, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 426–435. AUAI Press, 2004.

[162] Alon Orlitsky, NP Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. Convergence of profile based estimators. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pages 1843–1847. IEEE, 2005.

[163] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

[164] Bertha Osei-Hwedie. The role of ethnicity in multi-party politics in malawi and zambia. *Journal of Contemporary African Studies*, 16(2):227–247, 1998.

[165] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[166] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

[167] Luca Pappalardo, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. Using big data to study the link between human mobility and socio-economic development. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 871–878. IEEE, 2015.

[168] Dmitri S Pavlichin, Jiantao Jiao, and Tsachy Weissman. Approximate profile maximum likelihood. *arXiv preprint arXiv:1712.07177*, 2017.

[169] Jacob Poushter, Caldwell Bishop, and Hanyu Chwe. Social media use continues to rise in developing countries but plateaus across developed ones. *Pew Research Center*, 22, 2018.

[170] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[171] Michael O Rabin and Dana Scott. Finite automata and their decision problems. *IBM journal of research and development*, 3(2):114–125, 1959.

[172] Aditi Raghunathan, Greg Valiant, and James Zou. Estimating the unseen from multiple populations. *arXiv preprint arXiv:1707.03854*, 2017.

[173] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

[174] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[175] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014.

[176] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[177] Marc Rußwurm and Marco Korner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

[178] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

[179] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays. Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv preprint arXiv:1507.06947*, 2015.

[180] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[181] M. W. Sayers. On the calculation of international roughness index from longitudinal road profile. *Transportation Research Record*, 1501:1–12, 1996.

[182] Robert R Schaller. Moore's law: past, present and future. *IEEE spectrum*, 34(6):52–59, 1997.

[183] Ekrem Serin and Selim Balcisoy. Entropy based sensitivity analysis and visualization of social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 1099–1104. IEEE Computer Society, 2012.

[184] Jitesh Shetty and Jafar Adibi. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, pages 74–81. ACM, 2005.

[185] Zhenwei Shi and Zhengxia Zou. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3623–3634, 2017.

[186] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[187] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[188] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[189] David Skole and Compton Tucker. Tropical deforestation and habitat fragmentation in the amazon: satellite data from 1978 to 1988. *Science*, 260(5116):1905–1910, 1993.

[190] Mingjun Song and Daniel Civco. Road extraction using svm and image segmentation. *Photogrammetric Engineering & Remote Sensing*, 70(12):1365–1371, 2004.

[191] Emma S Spiro. Research opportunities at the intersection of social media and survey data. *Current Opinion in Psychology*, 9:67–71, 2016.

[192] Emma S Spiro, Sean Fitzhugh, Jeannette Sutton, Nicole Pierski, Matt Greczek, and Carter T Butts. Rumoring during extreme events: A case study of deepwater horizon 2010. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 275–283. ACM, 2012.

[193] Kostas Stamatiou, Lukas Kobr, and Nikki Aldeborgh. Settlement detection using convolutional neural networks on the digitalglobe geospatial big data platform. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2066–2069. IEEE, 2018.

[194] Jessica E Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A Alegana, Tomas J Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M Iqbal, et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690, 2017.

[195] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[196] Tavneet Suri and William Jack. The long-run poverty and gender impacts of mobile money. *Science*, 354(6317):1288–1292, 2016.

[197] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014.

[198] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[199] The World Bank. Africa's pulse. 2017.

[200] Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems*, pages 5778–5787, 2017.

[201] Jameson L Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González, and David Lazer. Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, 12(107):20150185, 2015.

[202] Ott Toomet, Siiri Silm, Erki Saluveer, Tiit Tammaru, and Rein Ahas. Ethnic segregation in residence, work, and free-time: Evidence from mobile communication. *University of Tartu*, 2012.

[203] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[204] Frank Tutzauer. Entropy as a measure of centrality in networks characterized by path-transfer flow. *Social networks*, 29(2):249–265, 2007.

[205] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6:1155–1166, 2018.

[206] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 685–694. ACM, 2011.

[207] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011.

[208] Paul Valiant and Gregory Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.

[209] Maarten Vanhoof, Willem Schoors, Anton Van Rompaey, Thomas Ploetz, and Zbigniew Smoreda. Comparing regional patterns of individual movement using corrected mobility entropy. *Journal of Urban Technology*, pages 1–35, 2018.

[210] Hal R Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016.

[211] Shashank Vatedka and Pascal O Vontobel. Pattern maximum likelihood estimation of finite-state discrete-time markov chains. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2094–2098. IEEE, 2016.

[212] Pascal O Vontobel. The bethe approximation of the pattern maximum likelihood distribution. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012.

[213] Aisha Walcott-Bryant, Reginald E Bryant, Michiaki Tatsubori, Daniel Emaasit, Samuel Osebe, John Wamburu, and Simone Fobi. The living roads project: Giving a voice to roads in developing cities. *Transportation Research Board – 96th Annual Meeting*, 2017.

[214] Nils B Weidmann. A closer look at reporting bias in conflict event data. *American Journal of Political Science*, 60(1):206–218, 2016.

[215] Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Heterogeneous mobile phone ownership and usage patterns in kenya. *PloS one*, 7(4):e35319, 2012.

[216] Bruce A Wielicki and Ronald M Welch. Cumulus cloud properties derived using landsat satellite data. *Journal of Climate and Applied Meteorology*, 25(3):261–276, 1986.

[217] Wikipedia contributors. Unbiased estimation of standard deviation — Wikipedia, the free encyclopedia, 2018. [Online; accessed 5-November-2018].

[218] Robin Wilson, Elisabeth zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie, Heather Chamberlain, Christopher Brooks, Christopher Hughes, et al. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 nepal earthquake. *PLoS currents*, 8, 2016.

[219] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *arXiv preprint arXiv:1504.01227*, 2015.

[220] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[221] Guang Xu and Xu Zhong. Real-time wildfire detection and tracking in australia using geostationary satellite: Himawari-8. *Remote Sensing Letters*, 8(11):1052–1061, 2017.

[222] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE international conference on computer vision*, pages 4633–4641, 2015.

[223] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, pages 4559–4566, 2017.

[224] William Chad Young, Joshua E Blumenstock, Emily B Fox, and Tyler H McCormick. Detecting and classifying anomalous behavior in spatiotemporal network data. In *Proceedings of the 2014 KDD workshop on learning about emergencies from social information (KDD-LESI 2014)*, pages 29–33, 2014.

[225] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444. ACM, 2014.

[226] Andrew Zammit-Mangion, Michael Dewar, Visakan Kadirkamanathan, and Guido Sanguinetti. Point process modelling of the afghan war diary. *Proceedings of the National Academy of Sciences*, 109(31):12414–12419, 2012.

[227] Li Zeng, Kate Starbird, and Emma S Spiro. Rumors at the speed of light? modeling the rate of rumor transmission during crisis. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 1969–1978. IEEE, 2016.

[228] Yue Zhang, Qi Liu, and Linfeng Song. Sentence-state lstm for text representation. *arXiv preprint arXiv:1805.02474*, 2018.

[229] Haifeng Zhao, Jasper S Wijnands, Kerry A Nice, Jason Thompson, Gideon DPA Aschwanden, Mark Stevenson, and Jingqiu Guo. Unsupervised deep learning to explore streetscape factors associated with urban cyclist safety. In *Smart Transportation Systems 2019*, pages 155–164. Springer, 2019.

[230] Kun Zhao, Márton Karsai, and Ginestra Bianconi. Entropy of dynamical social networks. *PloS one*, 6(12):e28116, 2011.

[231] Ziliang Zhao, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen, and Ling Yin. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9):1738–1762, 2016.

[232] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 2014.

[233] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems*, pages 1601–1608, 2007.

[234] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.