Estimating poverty maps from aggregated mobile communication networks

Christopher Smith-Clarke

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

February 20, 2021

I, Christopher Smith-Clarke, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Governments and other organisations often rely on data collected by household surveys and censuses to provide estimates of household poverty and identify areas in most need of regeneration and development investment. However, due to the high cost associated with manual data collection and processing, many developing countries conduct such surveys very infrequently, if at all, and only at a coarse level of spatial granularity. Consequently, it becomes difficult for governments and NGOs to determine where and when to intervene. This thesis addresses this problem by examining the feasibility of deriving up to date and high resolution proxy measurements of poverty from an alternative source of data, namely, Call Detail Records (CDRs), which can be used by organisations to help in decision making.

Specifically, we contribute the following:

1. A detailed spatial analysis of economic wealth in two sub-Saharan countries, Senegal and Côte d'Ivoire from which we derive two baseline poverty estimators grounded on concrete usage scenarios.

2. We establish a link between communication patterns and wealth through a simulation-based analysis of information diffusion. We further examine the influence of contextual factors, including data quality issues and economic volatility, on the strength of this relationship.

3. An approach to building wealth prediction models based on features of aggregated CDRs. Features include static and simulation based measures of information access, activity based metrics and econometric inspired metrics. We further perform a comparative analysis of the results of several models in

relation to the baseline predictors.

We conclude that it is possible to produce proxy poverty or wealth indicators from aggregated CDRs that provide a good level of accuracy, particularly where geographical coverage of the mobile phone network is sufficient. The final outcome of this thesis is a method for developing aggregated CDR-based poverty or wealth models that can be readily implemented anywhere in which there is a need for more up to date and/or finer resolution poverty estimates.

Impact Statement

Through an in depth examination of the relationship between aggregated communication patters, as represented by mobile phone data, and poverty/wealth, this work provides a platform from which new research can be undertaken. The methods of spatial aggregation, analysis and modelling detailed in this thesis constitute a process that can be duplicated and adapted to research projects that seek to expand the scope of this work. Such expansion could be in terms of incorporating new countries or time spans, or adapting the process to novel sources of data, both behavioural (input data) and socioeconomic (output data). This work further establishes the limitations that such research may be subject to. The academic impact of this work has also been established by 4 publications directly related to this thesis, as mentioned in Chapter 1.

This work also has the potential to provide significant benefits beyond academic research. Using the established methods to estimate poverty or wealth could provide huge cost savings for Governments and other organisations compared to manual surveying methods. In addition, by providing finer grained estimates the likelihood of identifying smaller pockets of poverty can be increased. The methods can also provide more up to date estimates of poverty, thereby improving the targeting of development funds, as well as enhancing the monitoring of the affects of policy and other influences on poverty. Successfully identifying poverty is the first step towards reducing it, therefore by making poverty identification more accessible to poorer countries in the above ways, this work has the potential to contribute towards the UN millennium goal of reducing global poverty levels. To this end, we have collaborated with UNFPA (United Nations Population Fund) and United

Nations Global Pulse on work that resulted in the publication Smith-Clarke et al. (2014). This research has also been presented by invitation to the UNFPA as part of the Big Data Bootcamp 2015, and from October 2013 - September 2017 was funded by a Google Europe Doctoral Fellowship.

Acknowledgements

It's no secret that pursuing and completing a PhD is hard and it would be impossible without the support and encouragement of many people. I am incredibly lucky to have had the opportunity to study at UCL where I could bump shoulders with many talented researchers. I'd like to thank in particular Neal Lathia and Daniele Quercia for first helping me to believe I could do a PhD and for providing the inspiration for the topic of this thesis. Also thanks to Giovanni Quatronne and Afra Mashhadi for their technical input and feedback and to my co-supervisor Paul Marshal for often helping me see I wasn't doing such a bad job. Special thanks of course to my supervisor Licia Capra for her unwavering and dedicated support and patience and helping to make sure this thesis finally saw the light of day.

I would also like to thank Sabrina Juran of UNFPA, Olivia De Backer, Miguel Luengo-Oroz and René Clausen Nielsen of UN Global Pulse for their helpful feedback on my research papers, and to Enrique Frias-Martinez at Telefonica Research for an interesting three months in Madrid.

This thesis is dedicated to my wife, who's acute sense of empathy often meant she suffered the stresses of PhD life more intensely than myself and without who's support and direction I might still be pretending to be a philosopher.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The United Nations (UN) Sustainable Development Agenda was established in 2015 in order to mobilise efforts to improve many aspects of people's lives across the globe. The agenda is encapsulated in 17 goals, the first being to *end poverty in all its forms everywhere*. Specifically the aim is to eradicate *extreme poverty*, defined as living on less than $1.25 per day, by 2030, and to reduce by half all other forms of poverty[1]. In order to alleviate poverty it must first be identified and measured. As the World Bank states, in addition to keeping the issue on the political agenda, reasons to measure poverty include the targeting of interventions: "institutions, including the World Bank and aid agencies, have limited resources, and would like to know how best to deploy those resources to combat poverty. For this, they need to know where in the world poor people are located." In addition, measuring poverty enables the monitoring and evaluation of the impact of interventions and the effectiveness of institutions designed to reduced it. At both the project and institutional level it is important to be able to determine which are working and which are not Haughton and Khandker (2009). Indeed, it is in these respects that governments, non-governmental organisations (NGOs) and businesses rely on socioeconomic data to guide decision making and to aid in the understanding of socioeconomic processes.

Some policies operate at the national or regional level, for example financial regulation and taxation policies may be designed to increase productivity and eco-

[1] `http://www.un.org/sustainabledevelopment/poverty/`

nomic growth, and thus reduce poverty indirectly. For these kinds of interventions coarse grained data is sufficient to inform the policymaking process and measure the policy's effect. On the other hand, some policies and projects may be designed to affect a more localised area, for example investing in transportation infrastructure, or even more localised, government departments and charities may wish to target highly deprived neighbourhoods for regeneration or social support projects. To effectively inform this kind of undertaking and measure its effects, much more fine grained poverty data is required.

Socioeconomic data disaggregated to smaller areas can reveal a large degree of variation hidden by larger aggregates. For example, it is estimated that 85% of the world's poor live in rural areas Alkire et al. (2014), a fact which perhaps helps explain the dramatic migration of people from rural to urban areas, but which also highlights the continuing importance of reliable estimates that differentiate at a sufficiently fine level of spatial granularity. Deprivation measures at state level will hide discrepancies between rural and urban conditions, as well as sharp differences between urban neighbourhoods. Indeed, a mass urbanisation process is taking place across much of the globe United Nations, Department of Economic and Social Affairs (2012), and yet the potential for raising the standard of living through efficient provision of public services and concentrated economic opportunity is not realised uniformly. A 2008 United Nations report United Nations Human Settlement Program (2008) into the state of the world's cities showed that inequality in urban environments is actually on the rise, as some areas and communities benefit more than others from economic growth and investment in public services with many migrants to the city locating themselves in highly deprived neighbourhoods. The report states that urban inequality has a detrimental effect on citizens' health, education and participation in society and the economy, which in turn leads to social unrest, higher crime levels and the diversion of resources from productive public investment toward security and policing, thus exacerbating the problem further United Nations Human Settlement Program (2008). To reverse this trend deprived areas need to be identified in a timely manner so that intervention and regeneration projects can be

planned and implemented accordingly.

1.1 The problem

Household socioeconomic data and neighbourhood statistics typically requires manual surveying, which can be prohibitively time consuming and expensive. The costs involved prevent comprehensive surveys of the entire population of a country from being carried out and also limit the frequency with which they occur. Instead, households are sparsely sampled and surveys typically only take place every several years. This state of affairs severely limits the spatial and temporal accuracy of poverty data, and in the case of particularly resource limited nations, this kind of data may not be available at all. For example, the situation in Côte d'Ivoire is typical of sub-Saharan countries, where there is socioeconomic data available from 2000 and 2005 that is disaggregated only at the level of 11 subnational regions, as depicted in Figure 1.1. Consequently, no information is available regarding the distribution of wealth within the boundaries of these regions, which hides the severe discrepancies between slum dwellers and those living in wealthy communes in the more densely populated cities. This is in stark contrast to the situation in some wealthier countries, in which data is often available at the small area level (i.e., neighbourhoods of a few thousand residents). For example, in Figure 1.2 we can see the fine spatial granularity of English Index of Multiple Deprivation. However, even here the measures of poverty are not up to date. For example, for the IMD published in 2010 the data sources used to calculate the indices date from between 2000 and 2008. If continuous surveying is infeasible in wealthy countries such as the UK, then it will certainly be beyond the means of developing countries. Furthermore, when severe poverty or deprivation is identified, the longer the time frame in which intervention takes place the more resources are likely to be required to alleviate problems, and in developing countries experiencing the most rapid urbanisation wealth distribution may change significantly year on year. Thus, to meet Sustainable Development Goal 1: *End poverty in all its forms everywhere*, and the aims outlined by the World Bank, it is, among other things, imperative to develop

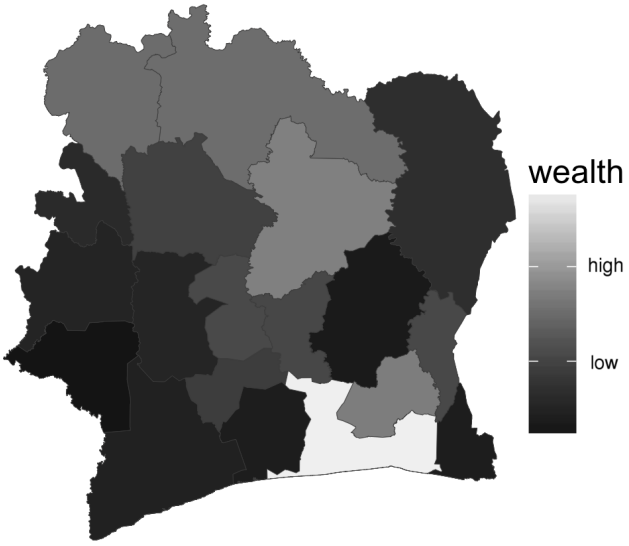


Figure 1.1: DHS Wealth Index in Côte d'Ivoire which is provided for just 11 regions

new, low cost ways to identify areas in need of intervention in a timely and spatially accurate manner.

1.2 Novel approaches to solving the problem

Technological advances and the increasing ubiquity of computing devices in people's lives, along with the rich data they produce, have opened the doors to novel solutions to the above mentioned shortcomings of household surveying. Researchers have begun to exploit a wide variety of different sources of human digital traces in order to explore the relationship between different forms of deprivation, or well-being more broadly. These data sources include content voluntarily published on online social media platforms such Facebook and Twitter, which is particularly amenable to sentiment and topic analysis in order to gauge a population's subjective well-being and also identify markers for unemployment or other socioeconomic characteristics Kramer (2010); Wang et al. (2012); Quercia et al. (2012a,b); Quercia and Saez (2014); Venerandi et al. (2015). Remote sensing via satellite imagery can be utilised to assess the visual impact of society, which in turn may reflect economic status through land usage patterns or intensity of night time lights Elvidge et al. (1997); Doll et al. (2000); Elvidge et al. (2001); Noor et al. (2008); Jean

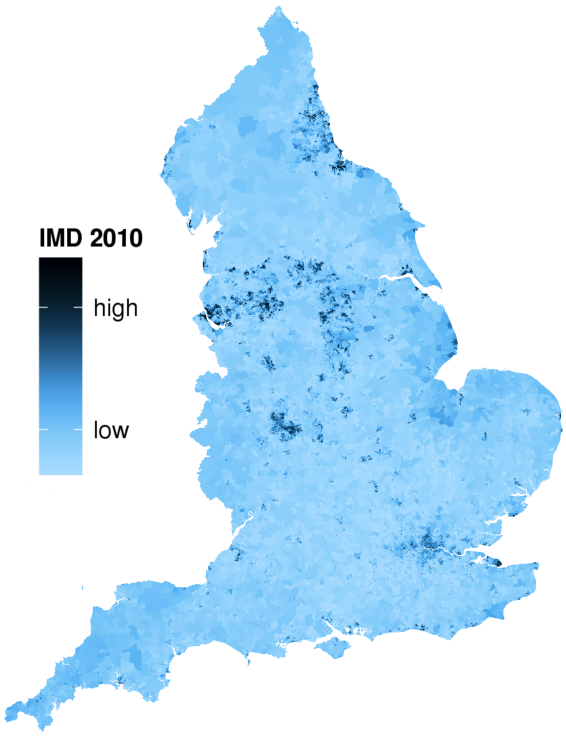


Figure 1.2: Index of Multiple Deprivation in England, which is available for 32844 area. However, even in England the IMD is based partly on out-of-date socioeconomic data.

et al. (2016); Steele et al. (2017). Finally, mobile phone data, or call detail records (CDRs), that also contain geolocation information reveal patterns of communication and mobility Eagle et al. (2010); Soto et al. (2011); Frias-Martinez et al. (2012); Frias-Martinez and Virseda (2012); Blumenstock et al. (2015); Gutierrez et al. (2013); Steele et al. (2017); Mao et al. (2013); Bruckschen et al. (2015).

1.3 The Orange Data for Development Challenges

An initiative that has helped to fuel research in this area is the organisation of open research competitions, known as the D4D (Data for Development) Challenges, and associated conferences around the release of large CDR datasets from telecommunications providers Orange and Sonatel in developing countries, namely Côte d'Ivoire and Senegal[2].

[2]http://www.d4d.orange.com/en/Accueil

Table 1.1: Descriptive statistics of CDR data.

	Senegal	**Côte d'Ivoire**
Country Population	20 m	15 m
Time span (weeks)	52	12
Number of BTS towers	1614	1217
Mean Daily Volume	4.0 m	10.8 m
Mean BTS Distance	236 km	228 km

The release of large data scale CDR datasets as part of these challenges has opened the door to the study of such datasets to researchers who are otherwise not party to private arrangements with telecoms providers. We capitalise on the opportunity this affords for the purposes of this thesis. CDR data from Côte d'Ivoire and Senegal was released as part of the 1st and 2nd D4D Challenge, respectively. The data do not pertain to individual mobile phone users, but rather consist of the hourly total volume (number of calls) and duration of calls between pairs of Base Transceiver Station (BTS) towers, as well as the geographical coordinates of each BTS. The datasets are summarised in Table 1.1. The dataset from Senegal covers a much longer period than that from Côte d'Ivoire (52 weeks verses 12 weeks) however, due to the service provider's larger market share in Côte d'Ivoire, the average daily volume is much larger there, at 1.4 calls per person compared to 0.2 calls per person in Senegal.

1.4 Research Hypothesis

In line with the research agenda identified in Section 1.2, we aim to develop and evaluate effective methods to estimate deprivation in different areas. Specifically, we will mine CDRs from mobile telecommunication providers in order to extract features which can be used to predict wealth. We focus on CDRs as a data source since mobile phones have high adoption rates in all parts of the world and suffer less from biases toward certain user demographics, unlike web based services which tend to have a user base that is concentrated in high income countries and that is more likely to misrepresent certain age, ethnic and income level groups. Indeed,

mobile phone penetration is high enough in many developing countries to make such datasets sufficiently representative of the population[3]. The network operators from which the CDR data is sourced hold a dominant position in their respective markets, with Orange having 48% market share in Côte d'Ivoire[4] and Sonatel commanding 52% of the market in Senegal[5]. As well as high penetration, CDR data also has the advantage of having high spatial and temporal resolution, which lends itself to providing the timely and granular poverty estimates we seek.

When dealing with data derived from individual human behaviour, another important consideration is of course privacy. Much work has been done to create reliable anonymisation methods which aim to prevent reidentification of individuals represented in the dataset. However, these methods may not be enough to appease the end user who is concerned about the use and possible abuse of their data. Tackling this problem directly is not within the scope of this thesis, rather, we sidestep the problem by using only aggregated CDRs, that is, the data represents aggregated activity at a particular location and does not contain any information that could distinguish individual users, even if joined with additional datasets.

> The central hypothesis of this thesis is that the spatially embedded interaction networks inherent in aggregated mobile communication data, contain behavioural patterns that can be mined in order to provide effective features predictive of socioeconomic factors such as poverty rate and average wealth at the areal level.

1.5 Contributions

To test this hypothesis we provide three main contributions:

1. A detailed spatial analysis of economic wealth in two sub-Saharan countries,

[3] http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf

[4] http://www.orange.com/en/group/global-footprint/countries/Group-s-activities-in-Ivory-Coast

[5] http://www.sonatel.com/wp-content/uploads/2017/02/Sonatel_2016_financial_results_EN.pdf

Senegal and Côte d'Ivoire from which we derive two baseline poverty estimators grounded on concrete usage scenarios.

2. We establish a link between communication patterns and wealth through a simulation-based analysis of information diffusion. We further examine the influence of contextual factors, including data quality issues and economic volatility, on the strength of this relationship.

3. An approach to building wealth prediction models based on features of aggregated CDRs. Features include static and simulation based measures of information access, activity based metrics and econometric inspired metrics. We further perform a comparative analysis of the results of several models in relation to the baseline predictors.

1.6 Publications related to this thesis

The validity of this work has been established in four related peer-reviewed publications listed below. The work represented in these papers is substantially my own, with L. Capra providing supervision, and A. Mashhadi providing supervision and repeating analyses on an additional dataset.

- Smith-Clarke, C., Mashhadi, A., Capra, L., *Ubiquitous Sensing for Mapping Poverty in Developing Countries*, Proceedings of Netmob 2013, August 2013

- Smith-Clarke, C., Mashhadi, A., Capra, L., *Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks*, CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 511–520, April 2014

- Smith-Clarke, C., Capra, L., *Beyond the baseline: Establishing the value in mobile phone based poverty estimates*, WWW '16: Proceedings of the 25th International Conference on the World Wide Web, 425-434, April 2016

- Smith-Clarke, C., Capra, L., *Information Diffusion and Economic Development*, ASONAM '17: Proceedings of the 2017 IEEE/ACM International Con-

ference on Advances in Social Networks Analysis and Mining, 475-483, July 2017

1.7 Structure of Thesis

- First we provide a review of related work in Chapter 2. We critically discuss previous research and highlight the areas in which the work presented in this thesis advances the state of the art. We also provide a comparison with similar research that has taken place either in parallel with the author's own work or since.

- In Chapter 3 we provide a detailed analysis of the socioeconomic data we utilise in later chapters and create baseline models against which CDR models can be fairly assessed.

- In Chapter 4 we introduce the CDR datasets and lay the ground work for CDR based models by establishing a link between country-wide communication patterns and wealth via information diffusion simulations.

- In Chapter 5 we then define and examine a number of features derived from CDR data, before building and comprehensively testing predictive models incorporating those features.

- We conclude in Chapter 6 with an overall evaluation of the contributions of this thesis, discuss the limitations faced and propose an agenda for continued research in this topic.

Chapter 2

Related Work

Poverty is usually understood as referring to a person or household subsisting on resources below a certain threshold, where that threshold may be defined in terms of income or consumption. Researchers have also argued for a broader view of poverty that represents the (in)capability of a person to properly function in society (Sen, 1999). In fact, the definition and measurement of poverty is itself a topic of significant debate (Council et al., 1995; Deaton, 2005), the details of which are beyond the scope of the current literature review. What is not controversial however, is the huge difference in wealth between the inhabitants of the richest and poorest nations and even the large differences that exist within the world's poorest countries. Indeed, it should be noted that the aim of this thesis is not to weigh in on the debate about how poverty ought to be defined and measured, but rather it seeks to provide a readily available method for producing a proxy measurement for the general ideas of wealth and poverty. The works discussed in this section all similarly explore novel ways to predict or approximate some sort of socioeconomic metric for an area. To help understand how these metrics relate to each other, Table 2.1 summarises the key types and what they represent.

Until recently, the only viable means of collecting objective socioeconomic statistics pertaining to individuals or households was by in person interviews and data collection. This process is costly and labourious - in the 1990's the advent of laptop computers was celebrated as a way to reduce the time to collect survey data from years to months (Deaton, 1997). In fact, the total time taken from inception to

publication of survey data, is still likely to be upwards of a year. As a concrete example, the Demographic and Health Survey (DHS, which is an exemplar source of socioeconomic survey data for developing countries) manual stipulates a timetable of 15 months from the initial survey design phase, through translation, staff recruitment and training, interviewing and data entry, to final data checking and clean-

Table 2.1: Types of socioeconomic metric used as target variable in the literature

Metric type	Description
Poverty rate	The proportion of people who's income is below the poverty line, which is defined as half the median household income of the population[1]
Asset index	These type of metrics, often referred to as *multidimensional measures of poverty* seek to go beyond simple income based measures by taking into account wealth in the form of asset ownership. The DHS, for example, includes questions regarding the ownership of assets such as mobile phones, computers, vehicles and refrigerators, as well as questions related to living conditions, such as access to electricity, sanitation and material used for flooring. These factors are then combined using Principal Components Analysis into an index representing the level of wealth of each household (DHS, 2012).
Composite	Another multidimensional approach that aggregates many factors into a single measure. As well as wealth, a composite metric usually includes factors such as health, education, housing or environmental conditions and access to essential services. Examples include continuous measures, such as the English IMD (which is used as target variable in a number of works cited below and often referred to simply as *deprivation*) and ordinal measures, such as socioeconomic level (SEL) as discussed in Soto et al. (2011).
Macroeconomic	This type of metric include measures of regional or national economic activity, such as Gross Domestic Product (GDP), as referred to in McClellan et al. (2013).
Subjective well being	In contrast to the objective measures discussed above, proponents of this type of metric acknowledge that non-economic factors are also important in determining overall well-being, and attempt to capture people's self-reported happiness or satisfaction with life. Examples include Bhutan's Gross Domestic Happiness [2].

ing (DHS, 2012). This is not including time taken for data analysis, report writing and publication. In contrast, by plugging in analytic software to APIs that provide access to real time (or near real time) sources of data reflecting people's behaviour and interactions (referred to variously as human digital traces and human exhaust data), after the initial implementation phase, proxy socioeconomic indicators can be produced immediately and continuously. Continuing with the DHS example, the survey manual also provides budget guidelines and clearly indicates that, given the huge human resource requirements, the cost in USD will be several million. Human digital traces on the hand are more often than not already produced for other purposes, such as customer billing for CDRs or as the primary output of users such as for social media platforms. Therefore, the cost of repurposing this data for present purposes is likely to be a small fraction of the cost of manual surveying.

The opening of data APIs by companies such as Twitter, and the advent of open data challenges, such as the Orange D4D series[3], has opened the door for computer scientists to make significant contributions to applied social sciences, by utilising data mining tools and computational methods more generally. Here, the focus of this review is on efforts to exploit human digital traces to develop methods to infer various indicators of human well being. Although research that exploits large scale human digital data to predict or understand socioeconomic factors is a relatively new undertaking, there is already a substantial body of relevant research taking a variety of approaches and operating on a range of different data sources. For expository reasons and in order to help place this thesis within the field, this chapter will be organised according to one salient dimension, namely, the kind of human behavioural or social data source, namely, CDRs, satellite imagery, social media content, transit data and international flows.

2.1 Call Detail Records

Call Detail Records are the records of mobile and fixed line telephone calls produced by telecoms providers primarily for billing and service monitoring. CDRs

[3] http://www.d4d.orange.com/en/Accueil

typically contain the calling and called telephone numbers (usually converted to anonymised identifiers when released for research purposes), call start time and duration, identifiers for the Base Transceiver Station (BTS) towers at which the call entered and exited the exchange, thereby indicating the approximate location of the call parties, and the call type (i.e., voice, SMS, MMS or internet data). There will normally be additional information related specifically to billing and fault monitoring. Despite the proprietary nature of CDRs, this is the most widely studied data type (Bank, 2015). This is no doubt due in large part to the release of datasets as part of the D4D challenges mentioned above, as well as the widespread adoption of mobile phone technology across the globe, in contrast to other technologies and services which may be limited to particular location or demographic. Given this relatively high volume of research we can make a further distinction between works that incorporate individual CDR data and those that use aggregated CDR data. The former potentially possess greater fidelity in terms of behavioural characteristics but also raises privacy concerns which the latter avoid.

2.1.1 Individual CDRs

A prime example of the work that combines individual CDR data (both mobile and fixed line in this case) with socioeconomic data is that of Eagle et al. (2010) who use England's Index of Multiple Deprivation (IMD) to test, at the societal level, theories which expect heterogeneous networks to provide greater access to a diverse range of resources and thus provide individuals with an economic advantage. The authors derive several measures of ego network diversity that aim to capture this effect, including as a function of the Shannon entropy of contacts (see Equation 5.7 in Section 5.2.2) and Burt's measure of Structural Holes (Burt, 2009). A structural hole is an open triad, or a missing link between any two of a node's neighbours, the number of which Burt found correlated with a employees' salary. They find that a composite measure of diversity that includes entropy based measures and structural holes correlates strongly ($r = 0.78$) with deprivation, suggesting that the advantageous effect of a diverse set of contacts operates at the societal level. The results are significant and it would be interesting to see if they could be improved by address-

ing some limitations in the methodology. For example, IMD areas are assigned to the telephone exchange area they most overlap with, rather than being proportionally assigned to all overlapping exchange areas. This simple approach seems to be a likely source of error. Secondly, the network data stems from a period of one month, yet recent research has shown that up to 60% of ties can decay from one month to the next and only around 20% of weak ties[4] remain after a 7 month period (Miritello et al., 2013). This means that, on a month by month basis, the correlation between network diversity and neighbourhood deprivation may change. This suggests two follow up steps: firstly compute the correlation between neighbourhood deprivation and the diversity of a stable network (properties tend to be stable when aggregated by 7 months or more (Miritello et al., 2013)); secondly, given the effect of tie creation and decay on information diffusion, the relation between temporal properties and deprivation ought to be investigated.

Soto et al. (2011) go beyond correlation analysis by combining 6 months of CDRs with socioeconomic data to train a machine learning model to classify areas of a Latin-American city by socioeconomic level (SEL). Specifically, the authors first use feature selection algorithms to determine the most important features from an initial set of 279 derived from individuals' CDR. These include 69 consumption features (such as total number of calls), 192 social features (such as number of contacts) and 18 mobility features (such as distance travelled). Individuals are associated to the cell tower closest to their residence and the parameters of each cell tower are the average of all its associated users. The feature selection process uses the entire dataset, which is a form of 'peeking', that is, it has been determined prior to the training stage which features are relevant to the portion of the data which will be later reserved for testing. Consequently, the models ought to be tested on a further dataset in order to be properly validated. After the feature selection process the data is split into training set (66.6%) and test set (33.4%) and several machine learning techniques are tested. The best achieves an accuracy rate of 80%. However,

[4]The definition of *weak tie* differs from study to study, but in this case two contacts are considered weakly connected if they make less than 10 calls to each other within the observation period.

the dataset is unbalanced: the distribution of 3 SEL classes over the census areas is, A: 12%, B: 59% and C: 29%, and although the final distribution in the dataset is altered when weighted (by area of overlap) averages are computed for Voronoi cells, it is likely that classes remain unbalanced. Furthermore, arguably the most important class to predict correctly is the lowest (i.e., poorest), and the highest recall (or sensitivity) for this class is 68%. Thus, accuracy alone may be a misleading performance measure in this case. Frias-Martinez et al. (2012) and Frias-Martinez and Virseda (2012) take a closer look at the relationship between some of the features used and SEL. They find that mobility features correlate moderately with SEL when aggregated to cell region, including radius of gyration ($r = 0.54$) and number of cell regions traversed ($r = 0.58$). They also find that significant differences exist between mean values of some features when cells are grouped by SEL, including reciprocity of links, distance between contacts and cost of calls. Using a simple multivariate regression model combining all the features they achieved $R^2 = 0.83$, it is therefore quite surprising that when the problem is reformulated as one of classification and more sophisticated methods are applied in Soto et al. (2011) (and later in Frias-martinez et al. (2012)) poorer results are produced. Nevertheless these are important findings, particularly regarding the mobility variables since, unlike communication variables, these features can be derived from other sources of mobility data, such as GPS traces or public transport systems.

The methods developed in Soto et al. (2011) are implemented in a system with a GUI presented and evaluated in Frias-martinez et al. (2012). The system is designed to reduce the number of census areas that need to be manually surveyed (and thus save on costs) by using the SEL of the proportion that is surveyed as training labels and estimating the remainder. Although apparently operating on the same dataset, the highest accuracy quoted in this work is 76% for a 3-class problem and 63% for a 6-class problem. However, these results are somewhat compromised by the fact that the feature selection process incorporates the entire dataset, despite the fact that the use case scenario stipulates that only a proportion of the ground truth data would be available. The problem of unbalanced classes applies here too,

therefore it is difficult to fully evaluate the outcomes.

Frias-Martinez et al. (2013) also investigated whether CDR could not only estimate present socioeconomic indicators, but also forecast them. The authors compared the monthly values of state level employment statistics with consumption and mobility variables derived from a 17 months of CDR in a Latin-American country. They found that a Multivariate Autoregression Moving Average model, which estimates future values of the dependent variable based on previous values of itself and other time-series, offers an improvement over a univariate model that only takes into account previous values of the dependent variable time-series: $R^2 = 0.65$ for predicting employment rate one month ahead and $R^2 = 0.31$ for predicting two months ahead. Visual inspection of the results also shows that the model is fairly accurate at predicting the direction of change (i.e., whether the employment rate will rise or fall) if not the exact value. These results are promising, however it is not clear how significant a benefit is gained from predicting employment rates one or two months in advance, since policy affecting employment tends to be implemented and evaluated over longer time scales. Indeed, quarterly or annual forecasts of socioeconomic data may be more appropriate, but no research has so far investigated the potential for long term forecasts, most likely owing to data limitations.

Blumenstock et al. (2010) present a method for estimating the wealth of an individual mobile phone user in Rwanda. For a sample of 2200 users, the authors collected survey responses pertaining to demographics and asset ownership and from the responses constructed a composite wealth index using principle component analysis. They then compared the wealth index with simple mobile phone behaviour variables derived from the CDRs and billing information of those same users. A linear model predicting expenditure from the CDR variables achieved $R^2 = 0.21$, indicating a moderate strength relationship between expenditure and mobile phone usage. A recent followup study similarly surveyed a sample of users, this time 856, and created a wealth index from the responses (Blumenstock et al., 2015). The modelling process was enhanced by generating thousands of features derived from the CDR data and using elastic net regularisation to remove irrele-

vant features. A cross-validated score of $R^2 = 0.85$ was achieved when predicting wealth from the CDR based model. However, with thousands of features compared to 856 data points, cross validation may not be enough to guarantee that the model has not overfit. Rather, a final validation that is not included in the feature selection step ought to be reserved and final performance scores reported on this unseen portion of data. Nevertheless, these results indicate that there is a strong potential to predict individual wealth from CDRs and in turn predict the aggregated wealth of geographical areas. The approach presented in these works represents an ideal situation for examining relationship between mobile phone use and socioeconomic factors and the results are an important piece of evidence showing that CDRs are a valuable resource in this context. However, by linking individual CDRs with personal demographic surveys this comes at rather a high cost, both in terms of human resources to conduct and process the surveys (albeit this cost is just a fraction of the cost of a DHS survey) and in terms of individual privacy. Because of these constraints it is less feasible to implement this method on a large scale. Instead, what is needed is a method that does not rely on linking data sources pertaining to the same individual.

Gutierrez et al. (2013) present what could be called the 'top-up' model of wealth. In developing countries the prevailing mobile subscription method is pay-as-you-go, as opposed to fixed-term contracts. The reasonable, but as yet unvalidated, hypothesis is that call time credit top-up behaviour reflects the wealth of the phone user (i.e., poorer people are likely to top-up their phone credit in small amounts fairly frequently, whereas wealthier people will top-up infrequently in larger amounts). Applying their model to a dataset of individual call records from Côte d'Ivoire, they derive the wealth indicator proxy and map the average and diversity (standard deviation) of wealth of different regions in the country. They also identify communities in a social network constructed from the call data and measure the diversity of their wealth indicator within communities. An interesting result is that overall diversity of top-up behaviour within a region is in some cases accompanied by low diversity within the communities in that region, suggesting some form

of segregation. However, before the top-up model is validated against some known wealth indicator only speculative conclusions can be drawn.

2.1.2 Aggregated CDRs

Privacy violation (whether perceived or real) is a potential stumbling block for all the above works which require individuals' personal phone records in order to compute the model features. To bypass this problem other work has considered features derived only from aggregated CDRs in which no personally identifiable data is present. For example, Mao et al. (2013) looked for correlations between features of CDRs aggregated to BTS towers in Côte d'Ivoire and economic indicators of ten centres of economic activity. They find that the ratio of total outgoing calls to total incoming and outgoing calls correlates strongly with annual income ($r = 0.80$) and poverty rate ($r = -0.83$). No explanation is proposed for this relationship, but it is similar to a feature, *introversion*, we formulated and tested on a different poverty dataset (see Chapter 5, note that this work and ours were submitted independently to the same venue), which is the ratio of internal flow to external flow. The intuition is that the more isolated (larger introversion ratio) a region, the less opportunity for economic development, and indeed we found a strong negative correlation. The authors also tested a number of other features which showed no correlation. Intriguingly, one such feature that showed no correlation was diversity (the same entropy based measure described above) which we tested ourselves, yet we found a strong correlation. This may be down to the different ground truth datasets used (which highlights the need to consider sensitivity of these methods to data quality, a factor we explore in Chapter 4. Another possible explanation for this discrepancy is that we computed the features for each BTS tower and then aggregated the results for each region, whereas here the authors first aggregate the flow for each region and then compute the variables. Edge weight on a highly aggregated network becomes much more homogeneous compared to a disaggregated network, an effect we discuss further in Chapter 3.

More recently, Bruckschen et al. (2015) have attempted to expand the scope of these efforts to other socioeconomic indicators, such as literacy levels, as well as

increase the spatial resolution at which estimates are produced. The authors create a linear model to predict these indicators from a number of features of mobile call and text communication in Senegal. In many cases good performance is reported, for example $R^2 = 0.87$ for the model fitted to poverty rate. However, the models are fitted at the subnational level with just 14 data points and a process of forward and backward stepwise feature elimination and selection is employed to select the best performing feature set. Subsequently, there appears to be a significant chance the model is overfitted to the data and a robust assessment of model performance on unseen data has not been provided.

2.2 Satellite Imagery

Researchers have also investigated an alternative to directly mining behavioural data or content, which is to use satellite imagery (often referred to as *remote sensing*) to identify the visual signs of economic development. An advantage of this approach is that satellite imagery covering the entire globe is openly available from a number of sources, thereby avoiding the hurdle of the proprietary nature of many other data sources considered here. As early as 1997, the total area lit by Night Time Light (NTL) measured from satellite imagery was shown to correlate with a country's Gross Domestic Product (Elvidge et al., 1997) and later for more countries (Doll et al., 2000; Elvidge et al., 2001). Noor et al. (2008) analysed the correlation between NTL and an asset-based wealth index for administrative regions of several African countries. They found that the mean NTL level (measured as brightness per pixel) correlated strongly with the wealth index (Pearson's $r = 0.64$, Spearman's rank coefficient $\rho = 0.79$) at the administrative level. Although these works demonstrate a clear relationship between NTL and wealth, the utility for targeting the poorest areas or monitoring effects of policy interventions is limited by the geographical scale of the analyses. Furthermore, work which attempts to estimate poverty at the small area level suggests that the relationship does not hold at a finer level of granularity. McClellan et al. (2013) looked at the relationship between NTL and small area poverty levels at two periods in time in Bangladesh. They found

that in 2000 NTL could estimate poverty levels reasonably accurately, but in 2005 the correlation had disappeared. In addition, there was no correlation between the change in NTL intensity and the change in poverty level. The results suggest that penetration of electricity availability has reached saturation, consequently removing the signal previously present.

Greater success has been achieved recently by Jean et al. (2016) who employ convolutional neural networks to improve the accuracy of predictions derived from satellite imagery. Their models exploit not only night time lights but also feature visible in day time, such as roofing material and roads. They test their approach on five African countries and report good R^2 values when comparing predicted wealth to ground truth data (also taken from the DHS survey in each respective country), ranging from 0.55 to 0.75. Of particular note is that they also assessed predictions made by models trained on different countries' data, with most R^2 values ranging between 0.40 and 0.68 in these cases, demonstrating the potential for this approach to be able to provide wealth estimates for countries where no ground truth data is available. However, while seemingly impressive from a scientific point of view, it is not clear whether this level of accuracy is high enough to warrant implementation in practice. Moreover, a potential shortcoming of this work is that, like much of the research already discussed, no baseline has been established against which to compare the results. It may be the case that similar accuracy could be achieved using only population data and readily available spatial data.

Researchers have also combined features derived from satellite imagery and CDR data. Steele et al. (2017) found that separate CDR and remote sensing models performed comparatively well when estimating the DHS wealth index in Bangladesh, and that a combined model offered a modest improvement over the separate models. Two shortcomings of the remote sensing based model as compared to the CDR based model noted were that the currently available spatial resolution of open satellite imagery is too low to distinguish between urban neighbourhoods within a city, and also the temporal resolution is low, with updates to satellite imagery only being made available every few years.

2.3 Social Media

The explosion of user-produced online content in recent years has led researchers to ask whether we can study this content to find new ways to gauge the well-being of populations. Platforms such as Twitter and Facebook encourage users to frequently post status updates, opinions and other such content, as well as responses to other users' posts. This content is ripe for application of Natural Language Processing (NLP) techniques that involve the analysis and quantification of different aspects of unstructured text data, including the classification of the emotion or sentiment contained in the text and determining the topic(s), or what the text is about. In addition, such content is often geo-tagged, meaning that aggregated characteristics of this content can be associated with particular locations. As well as textual content, platforms such as Foursquare and Facebook-Places also contain explicit geo-tagging information, i.e., users recording and sharing visits to certain locations, providing insight into the usage patterns and qualities of particular locations. Finally, online social media platforms generally encourage users to connect with one another, for example as *friends* on Facebook or *followers* on Twitter. These relationships constitute rich social graphs that can be further mined to uncover the interactions and relationships between users and locations.

Kramer (2010) developed a method for approximating Gross National Happiness (GNH), a measure derived from the self-reported satisfaction with life scale (SWLS) (Diener et al., 1985; Pavot et al., 1991), based on the difference in rate of positive and negative words present in Facebook status updates . Although the aim is to provide an unobtrusive measure of subjective well-being at the national level, validation of the work took place at the individual level; a hierarchical regression model is used to fit SWLS scores to individual happiness scores of a sample of users who also filled out a survey. SWLS was found to be a 'significant' predictor of Facebook happiness for this sample, however, the correlation between SWLS and happiness was low ($r = 0.17$). Furthermore, a follow up study of Facebook Gross National Happiness (FGNH) presents contradictory results (Wang et al., 2012): no correlation was found between FGNH and SWLS but instead a moderately strong

correlation ($r = 0.72$) was found between FGNH and the rate of negative words. The authors cite several possible explanations for their results, including limitations in the natural language processing techniques and a misalignment between what is captured in FGNH and SWLS (current mood state and the degree to which aspirations have been met, respectively). In conclusion, although FGNH is not fit to replace GNH, there does appear to be a signal in the data, and given that the model operates at the individual level, a finer level of spatial aggregation ought to be possible. Further research is needed to determine whether Facebook status updates provide a real window into people's state of well-being. A major drawback to this endeavour is of course the risk that users will perceive it to be an encroachment on their privacy.

Work that has begun to investigate the geographical variation in online sentiment includes Quercia et al. (2012a), who found that sentiment expressed in tweets from Twitter users living London, disaggregated by census area, correlates moderately with deprivation in the area ($r = 0.37$). Users from less well-off areas tend to tweet more negatively compared to users from wealthier areas. Note that in this case deprivation is measured using the IMD, which is composed from various socioeconomic indicators rather than self-reports, and which suggests that current mood state is more closely related to current circumstances than it is to overall satisfaction with life. Quercia et al. (2012b) also found that the topic profile of tweets in different areas of London can predict deprivation. A linear regression model fitting topic distributions to IMD achieved $R^2 = 0.49$. This a promising result suggesting that the well-being of communities could be reasonably estimated by monitoring the content of publicly available socially generated content. Moreover, there is plenty of room for improvement since these studies were relatively small scale (just 573 Twitter profiles after filtering) and results were presented for 78 areas of London - fairly low resolution for a city of this size. By scaling up orders of magnitude it ought to be possible to improve the accuracy of well-being estimates and disaggregate at a much finer level of granularity. Stronger confirmation of the relation between topics and deprivation also requires a longitudinal analysis to determine

whether significant changes in topic distributions occur over time. Privacy is less of a concern in the case of Twitter data, since the vast majority of content is already in the public domain.

Moving on from textual content and looking instead at location-based behaviour, Quercia and Saez (2014) studied the land usage data generated by users of Foursquare and extracted metrics that were found to correlate with deprivation in London. Venerandi et al. (2015) extended this work to include two more large urban areas in England and also incorporated data from OpenStreetMap (OSM), a crowd sourced, freely available online mapping service. OSM users record location information explicitly to create accurate and free geographical maps, as opposed to Foursquare users, whose primary aim is to record their personal location history. The authors found that the relative prevalence of certain types of venue, such as 'flea market' or 'embassy', could be associated with the level of deprivation in the neighbourhood, as measured by the IMD. As with Twitter, the data used in these studies is already in the public domain, therefore privacy violation is not a concern. However, this approach has so far only been tested in very large, very densely populated urban regions in England. In rural areas and smaller urban enclaves, where the variety in types of amenity and point-of-interest will be much lower, it is unlikely that the associations found in large urban areas will hold.

Considering the aim of inferring poverty or wealth in developing countries, all online social media platforms suffer the same limitation, namely that the user base is heavily concentrated in richer countries, and moreover, where a user base does exist in developing countries it will not be representative of the population at large.

2.4 Transit data

We have seen how geolocated CDR and social media data have been linked to poverty and well-being. It is no surprise then that other sources of data that explicitly capture the movement of people have also been examined for this purpose. Such sources include automated fare collection (AFC) systems in rail and bus networks that record the number of passengers travelling between locations. Another example

is the GPS traces of taxis and ride sharing services, which can reveal the movement of people around city in a potentially more fine-grained manner, although there is an obvious demographic bias in this case. To the best of our knowledge there has to date been no attempt to date to link taxi GPS traces to poverty or well-being.

An example of research that investigates the relationship between mobility and neighbourhood deprivation is that of Lathia et al. (2012), who use Oyster card (AFC) data from London's underground rail network. The authors first infer important locations of travellers based on their travel history and compute a flow matrix based on visits from resident neighbourhoods to others. First they find that total flow into a station (total number of travellers rather than number of trips) correlates weakly with IMD ($r = 0.21$, $p < .001$), suggesting that deprived areas tend to receive more visitors. Next the authors formulate two metrics representing homophily and heterogeneity of neighbourhoods. Homophily is the degree to which nearby neighbourhoods are similar to each other, and heterogeneity measures the variety in adjacent neighbourhoods, where, in both cases, similarity between neighbourhoods is based on IMD. No correlation was found between deprivation and homophily, and only a very week, negative correlation with heterogeneity ($r = 0.16$, $p < .001$). These are interesting findings, however, given that the correlation coefficients are low, further analysis is needed to confirm that the relationship is not an artefact of the important location inference process or the way flows have been defined.

Smith et al. (2013) also used London AFC data, plugging in a number of mobility based features into linear regression and support vector machine models in order to predict IMD, as well as the constituent factors that make up the IMD separately, such as education levels and crime, among others. The authors achieved an accuracy of up to 80% in some cases, demonstrating the potential for mobility-based features to accurately predict deprivation in an urban environment. However, the validity of the results is limited by the very low geographical coverage of the rail network. Although the rail network provides access to much of the city, only around 5% of census areas contain stations and consequently the mobility data could only be related to this small sample.

2.5 International flow data

Notable related work that does not fit easily into one of the above categories is that of Hristova et al. (2016), who examine several international flow networks including trade, postal, migration and aviation networks, and their relationship to national socioeconomic indicators such as GDP per capita, the Human Development Index and poverty rate. The weighted out-degree (i.e., total outgoing volume) of the postal and trade networks in particular was found to be strongly negatively correlated with poverty, as was the non-weighted out-degree (i.e., the total number of countries to which post and trade is sent). The main limitation for present purposes is that the international flow networks do not reveal any variation within each country. However, the postal network in particular appears to have the potential to be applied within a country's borders if the data were accessible and could provide a picture of economic wealth at the spatial granularity of local sorting offices.

2.6 Summary

Research has shown that it is possible to extract features of large-scale human behavioural and communication data that correlate with socioeconomic factors of geographic areas, although results are mixed and there remains significant room for novel contributions to this new research area. Of the works looked at, in some case the issue of geographical scale has not been solved, that is, the analysis takes place at too course a level of granularity to be of use to policy makers (Doll et al., 2000; Elvidge et al., 2001; Noor et al., 2008; Kramer, 2010; Hristova et al., 2016; Steele et al., 2017), and in others the relationships investigated are ambiguous (Wang et al., 2012; Lathia et al., 2012; Smith et al., 2013). More promising work has shown that it is possible to predict the value of deprivation indicators by combining several features of human behavioural dynamics into a predictive model (Soto et al., 2011; Frias-Martinez et al., 2012; Frias-Martinez and Virseda, 2012; Blumenstock et al., 2010, 2015; Gutierrez et al., 2013). However, results are somewhat inconsistent and their evaluation is incomplete in the sense that it is not clear what improvement is offered over simpler methods. There also remains room for significant improve-

ment in terms of accuracy and the data requirements for training the models. To help place this thesis within the wider literature and to support the choice of source data type, we present a summary of the pros and cons of using each particular data type for the purposes of estimating poverty or related socioeconomic factors in Table 2.2.

Individual CDRs are an ideal data source in many respects. Mobile phones have high penetration rates, making CDR data more representative of the population, even in developing countries. It also has high geographical coverage and high spatial resolution. That is, the BTS towers tend to cover the majority of populated land within a country and BTS tower density increases with population density, meaning smaller areas can be disaggregated in high density locations. CDR data essentially provides a temporally continuous picture of human behaviour, meaning that poverty estimates can be updated continuously. Individual data also has high heterogeneity, or the fine details of individual behaviour make it more likely that distinguishing features indicative of poverty and wealth can be uncovered. However, this detail comes at the cost of raising privacy concerns. Aggregated CDR data, on the other hand, trades in some of the detail of individual CDR data in order to side step these concerns. A significant loss in this aggregation is the ability to characterise mobility patterns, as individuals can no longer be tracked by the BTS towers they connect to. CDR data is of course proprietary, which can make it difficult for researchers to gain access. But despite this, CDR data is the most widely studied data source and indeed, aggregation makes it easier for telecoms companies to take part in data sharing arrangements, for example, as part of their corporate social responsibility agenda.

Satellite imagery, although privacy preserving and open, suffers from having low spatial resolution that prevents sufficient disaggregation in densely populated areas, and low temporal resolution, preventing timely updates to poverty estimates and monitoring of policy effects. Social media data has the benefits of high resolution and heterogeneity, but the low penetration rates and low geographical coverage make this kind of data particularly inadequate for developing countries at present. Transit data similarly suffers from low geographical coverage as public transport

systems tend to be limited to denser areas. Finally, the various international flows cannot be used for subnational poverty estimates, but do have some potential, particularly postal networks, if similar subnational data is made available.

Given the considerations noted above, in this thesis we have chosen to focus on aggregated CDR data as our source of human behavioural information. Aggregated CDRs strike a balance between detail and privacy, and in doing so provide an easier route to wider exploitation. Furthermore, in this work we show that aggregated CDR data can still be used to produce wealth estimates, thereby justifying this trade off.

Table 2.2: Pros and cons of different data sources

Data Type	Pros	Cons	Key Works
Individual CDR	high penetration; high geographical coverage; high temporal and spatial resolution; high heterogeneity	privacy concerns; data is proprietary	Eagle et al. (2010); Soto et al. (2011); Frias-Martinez et al. (2012); Frias-Martinez and Virseda (2012); Blumenstock et al. (2010, 2015); Gutierrez et al. (2013); Steele et al. (2017)
Aggregated CDR	high penetration; high geographical coverage; high temporal and spatial resolution; privacy preserving	low heterogeneity; data is proprietary	Mao et al. (2013); Bruckschen et al. (2015)
Satellite Imagery	complete geographical coverage; privacy preserving; open data	low spatial and temporal resolution	Elvidge et al. (1997); Doll et al. (2000); Elvidge et al. (2001); Noor et al. (2008); Jean et al. (2016); Steele et al. (2017)
Social Media	high temporal and spatial resolution; high heterogeneity; possibly open data	possible privacy concerns; low penetration/high bias; low geographical coverage	Kramer (2010); Wang et al. (2012); Quercia et al. (2012a,b); Quercia and Saez (2014); Venerandi et al. (2015)
Transit	high temporal and spatial resolution; possibly privacy preserving	low geographical coverage, proprietary data	Lathia et al. (2012); Smith et al. (2013)
International Flows	possible high temporal resolution; high geographical coverage	very low spatial resolution	Hristova et al. (2016)

Chapter 3

Baselines

When developing and testing a novel approach to solving a problem, it is important to establish the benefit of the new approach over alternatives. This is typically done by treating an existing approach, or where one is not available, a simplified version of the novel approach, as as a baseline, or benchmark, and making a direct comparison to the novel approach according some defined performance measure. This is standard practice when assessing any proposed scientific advancement, such as a modification to a machine learning algorithm that aims to increase performance or efficiency. It is equally, if not more, important when considering practical applications, when the novel approach may require greater resources to implement compared to existing or simplified approaches and thus by comparing performance an objective value judgement can be made regarding the trade-off between the costs and benefits of adopting the new approach.

Research that seeks to overcome the problem of missing poverty data by exploiting human digital traces, including all of the works discussed in Chapter 2, has thus far progressed by measuring its success against the alternative of having no ground truth (i.e., poverty data, wealth index or well-being) estimates at all, or rather, implicitly making a comparison against a dumb baseline such as a random guess (possibly with values drawn from a specific distribution, as below) or a constant value, both of which will have minimal predictive power. In this light, the level of accuracy achieved may be considered more remarkable than in fact it should be. Although the cost of acquiring and analysing mobile phone data is likely to be

significantly less than that of undertaking a comprehensive household survey, it is nevertheless non-trivial. Therefore, the benefit of this kind of approach over simpler and more readily available means of estimating poverty needs to be established. To this end, in this chapter we perform a detailed analysis of the socioeconomic and population data at hand, from which we extract two realistic baseline estimators that will later be used as performance benchmarks for our CDR based models presented in Chapter 5. In Chapter 5 we will then provide an extensive comparative analysis of the baseline models against CDR-based models, and establish the circumstances under which CDR derived features do indeed add value to poverty prediction models, and to what extent.

In real-life settings, two scenarios can occur: no ground truth at all vs some ground truth. Correspondingly, two approaches can be taken: applying a general model vs retraining on the available ground truth in order to produce a bespoke, and likely more accurate, model. The first approach aims to produce a model fitted to a sample ground truth dataset that can then be utilised in situations where no ground truth data pertaining to poverty or socio-economic status are available, as would be the case for countries in which no recent survey has been undertaken. Research of this kind will produce general models from the study data that can, in principle at least, produce predictions or rankings for other countries or regions without the need for ground truth data from these new regions (Smith-Clarke et al., 2014; Bruckschen et al., 2015; Pokhriyal and Dong, 2015; Mao et al., 2013), although such ground truth would of course still be desirable in order to validate the estimates.

The second approach would follow the same modelling method, but also involve retraining on ground truth data from the region of interest in order to fit a bespoke model for that region. This could be the case if a survey had been undertaken in the past, in which case a model may be fitted and then projected forward in time. Or, if a survey had been conducted recently but only for a subsection of the region (perhaps to cut costs, with the plan to then interpolate the results to unsurveyed locations), in which case the model can be used to make predictions for the remaining unsurveyed locations (Soto et al., 2011; Frias-martinez et al., 2012).

We refer to the two approaches as the general model approach and retrained model approach respectively.

Works taking either approach appear to demonstrate the value to be gained from mining CDRs in terms of the predictive power of poverty estimates. However, they all suffer from the same major limitation; namely, that they have yet to establish a reasonable baseline against which a fair comparison can be made. The implicit assumption is that the best available baseline model would be a random guess, or rather, no prediction at all, and therefore any improvement over this represents a positive result. Yet the reality is that socio-economic data is strongly correlated with population density and, furthermore, often contains a strong degree of spatial autocorrelation. Consequently, we ought to expect a baseline model that takes one or both of these factors into account to perform significantly better than a simple random guess.

In order to measure the *real* added value of mining CDR features to estimate poverty in developing countries, we need to show that our CDR based model out performs alternatives. In particular the aim is not to show that a CDR model can outperform the gold standard of estimating poverty by manual survey, an approach which has a much greater cost, but rather the aim is to show that a CDR based model can outperform readily available, low-cost alternatives. Additionally, a baseline model will need to provide predictions for the same data points and at at same level of spatial granularity at which will be operating later in Chapter 5 when we construct models using CDR features (discussed further in the next section). To this end, we will produce *two* fair baseline models. The first exploits correlations with population population density, corresponding to the general model case and to be used when no ground truth data is available. The second additionally leverages spatial auto-correlation, corresponding to the retrained model case and to be used when partial ground truth data exists.

We next introduce the Demographic and Health Surveys (DHS) data, which is used as ground truth for poverty throughout this thesis, before presenting a detailed spatial analysis of this data and an investigation its relationship with population

density - two factors which will inform the development of the baseline models.

3.1 Socioeconomic Data

Demographic and Health Surveys (DHS) are conducted in several developing countries, usually in collaboration with the national statistical agency and other organisations. Surveys are conducted by interview with household members in order to guarantee a certain level of quality and consistency in survey responses. However, this manual process limits the population coverage that is practically feasible. Subsequently, a household sample process is designed such that the aggregated responses are statistically representative of the population at the largest subnational administrative region. At this level of aggregation there are normally only a handful of regions. In Senegal there are 14 and 11 in Côte d'Ivoire. The household sampling process consists of several stages. First, enumeration areas (EAs) are stratified by an urban or rural designation within each subnational region; then, within each stratum, a certain number of EAs are selected with a probability proportional to their size. EAs normally consist of neighbourhoods in urban areas and villages, or groups of villages in rural areas. Finally, households are randomly selected with uniform probability within each EA selected in the previous stage. The group of selected households within each EA are known as clusters. That is, a cluster is a random sample of households, normally between 15 and 30, from within a single EA. The GPS coordinates of the centroid of each cluster is provided with the DHS in order to enable spatial analysis of the survey data. However, in order to maintain anonymity of survey respondents, cluster coordinates are obfuscated by randomly displacing them up to 2 km for urban clusters and 5 km for rural clusters, with 1% of rural clusters being displaced up to 10 km.

Table 3.1 presents the number of clusters sampled with GPS coordinates recorded for both Senegal and Côte d'Ivoire, together with their total population and surface area. It is evident that Côte d'Ivoire is more sparsely populated, with approximately 15 million people living within $322 km^2$, compared to Senegal, where 20 million live in $197 km^2$. The issue of data sparsity will be addressed in Chapter 5.

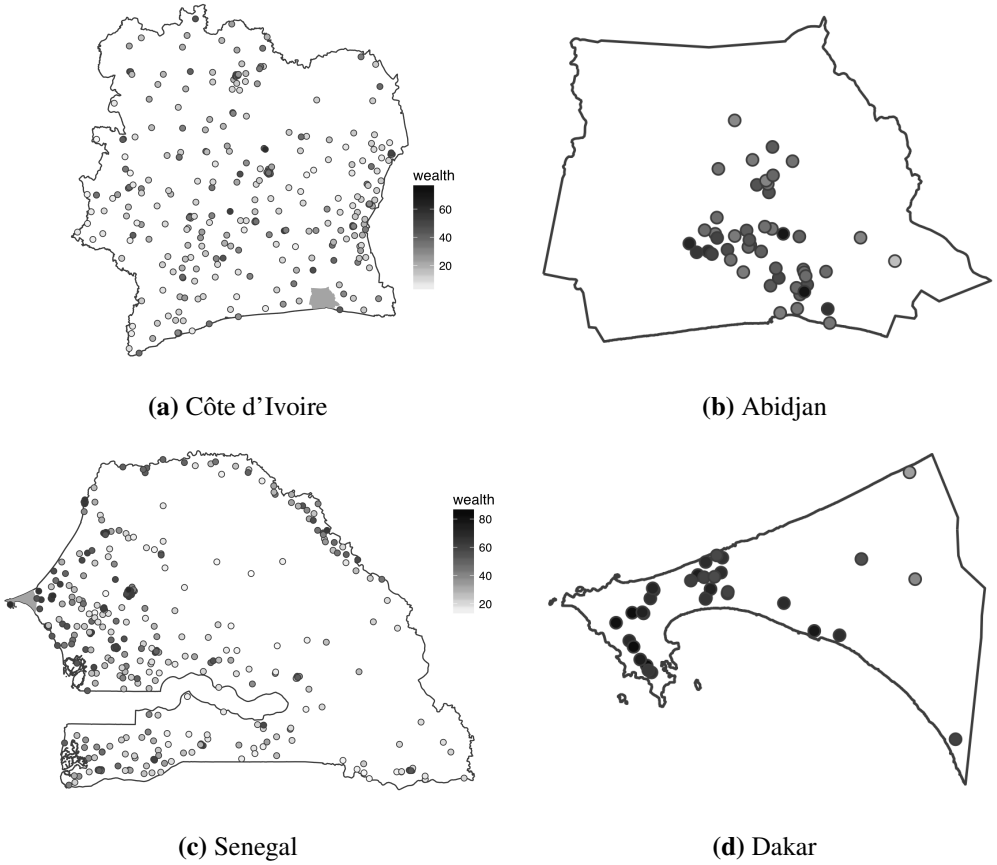
The Senegal DHS took place in 2012-13 and consists of 391 clusters, 6 of which are missing GPS coordinates. The Côte d'Ivoire DHS dates from 2011-12 and contains 351 clusters, 10 of which are missing coordinates. Cluster locations and their relative wealth (as measured by the DHS asset index, described below) are mapped in Figure 3.1. Cluster density closely follows population density, and it is apparent from these figures that more remote areas tend to be poorer and denser areas tend to be wealthier, including in the largest city in each respective country, shown in the zoomed in views on the right. This pattern is analysed in more depth in Section 3.3.

The DHS includes questions regarding the ownership of certain assets, such as mobile phones, computers, vehicles and refrigerators, as well as questions related to living conditions, such as access to electricity, sanitation and material used for flooring. These factors are combined using Principal Components Analysis into an index representing the level of wealth of each household. Note that the indices of Senegal and Côte d'Ivoire are computed separately, therefore the values are not directly comparable. When estimating poverty, we operate at the cluster level rather than that of individual households, we thus aggregate the wealth index by taking the *median* wealth index of households in the cluster to represent the average wealth at that location. Average wealth is conceptually distinct from the notion of poverty rate, which measures the proportion of households classified as poor within each cluster. Average wealth may mask the existence of poverty within the cluster if it coexists with high wealth within the same cluster. Although use of the median will mitigate against the influence of extreme wealth, it could still fail to reflect the existence of poverty in an otherwise wealthy area.

To represent poverty rate from the wealth index data, we can take the percentage of households in the cluster that are among the poorest fifth of households

Table 3.1: DHS and country summary statistics

	Senegal	Côte d'Ivoire
Population	20 million	15 million
Area	197 km^2	322 km^2
Clusters	385	341



(a) Côte d'Ivoire **(b)** Abidjan

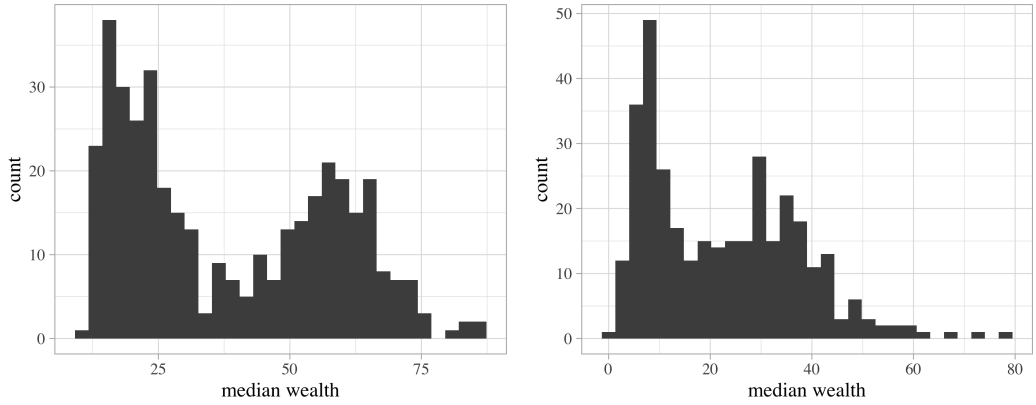
(c) Senegal **(d)** Dakar

Figure 3.1: Median wealth at DHS cluster locations in (a) Côte d'Ivoire (b) Senegal and their respective largest cities (c) Abidjan and (d) Dakar

nationally. This differs from median wealth in that it is invariant to the distribution of wealth among the top four fifths of households within the cluster. In other words, extreme poverty within the cluster is not masked by the coexistence of wealth within the same cluster. Figures 3.2 and 3.3 show the distribution of wealth and poverty rate, respectively, as derived from DHS data for the Senegal and Côte d'Ivoire. Figure 3.4 illustrates the strong correlation between wealth and poverty rate. Given this strong correlation and the fact that the distribution of median wealth is less skewed and closer to a normal distribution compared to poverty rate, we focus our efforts on modelling this attribute rather than duplicating the analysis.

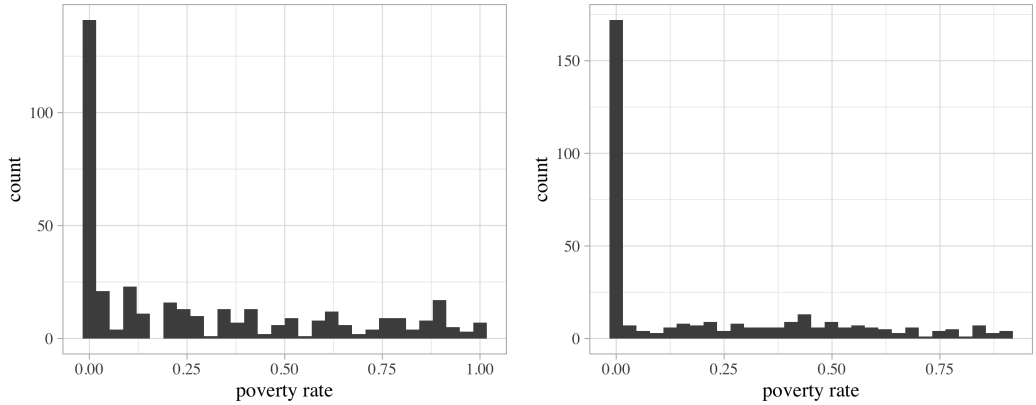
3.2 Wealth and population density

In this section we examine the relationship between wealth and population density, which will later inform the construction of the baseline models. In ad-



(a) Senegal **(b)** Côte d'Ivoire

Figure 3.2: Distributions of median wealth over the DHS survey clusters.

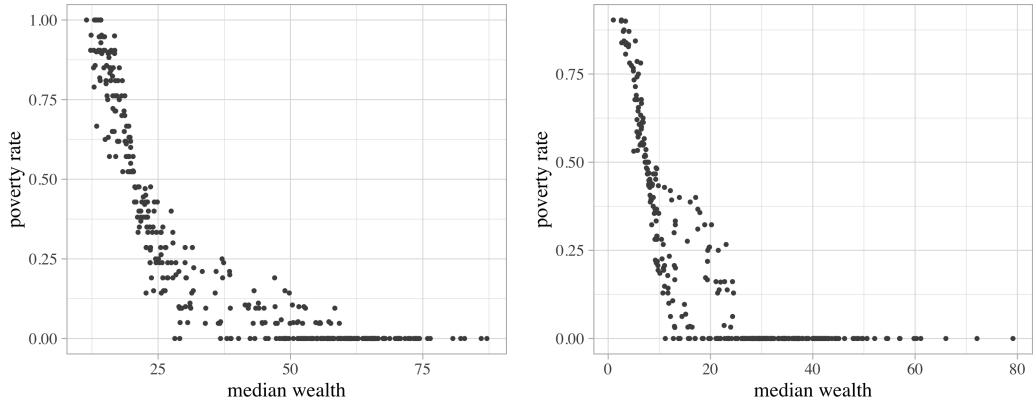


(a) Senegal **(b)** Côte d'Ivoire

Figure 3.3: Distributions of poverty rate over the DHS survey clusters.

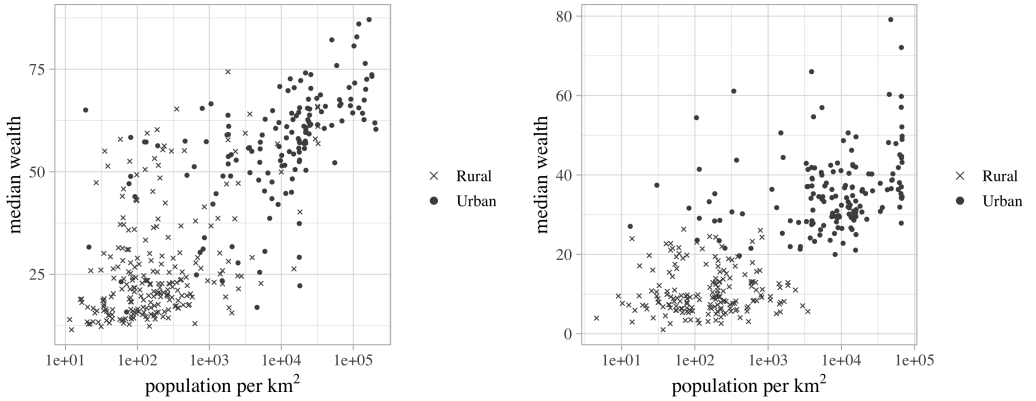
dition to the DHS data, we also obtained gridded population estimates from www.worldpop.org.uk, an organisation that produces accurate and up to date, high resolution population maps. Using this gridded population data we compute the population density of a cluster centroid to be the population with a circle of radius 1 km around the cluster point.

A link between population density and prosperity is often posited, with many mechanisms proposed to explain this relationship, including efficiency of service provision and increased access to diverse sources of information and opportunity (Pan et al., 2013; Gary S. Becker, 1999; Bettencourt et al., 2007). For Senegal and Côte d'Ivoire, this relationship can clearly be seen in Figure 3.5, where wealth (as computed before) is plotted against population density, and where denser areas



(a) Senegal **(b)** Côte d'Ivoire

Figure 3.4: The relationship between median wealth and poverty rate for DHS clusters



(a) Senegal **(b)** Côte d'Ivoire

Figure 3.5: Median wealth in relation to population density.

tend to also be wealthier and have lower concentration of poverty. The Spearman's rank correlation in Senegal is $\rho = 0.72$ and in Côte d'Ivoire it is $\rho = 0.71$, further indicating a fairly strong correlation between population density and wealth in the two countries. We can see a marked division between urban and rural locations, with urban locations tending to be wealthier. Indeed, in Côte d'Ivoire no urban cluster contains a household among the poorest 20%.

3.3 The spatial distribution of wealth

Figure 3.1 shows the average wealth at DHS cluster locations. A degree of spatial clustering of wealth is evident, with wealthier clusters tending to appear in close proximity, although a significant number of exceptions are apparent. These figures

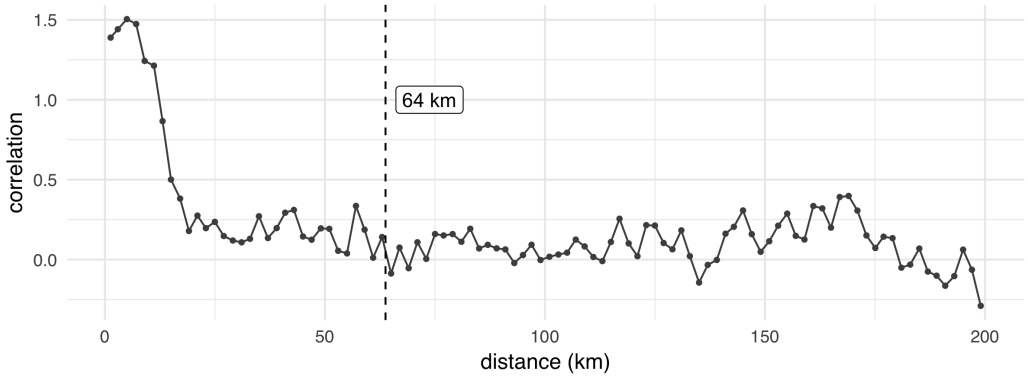
alone also fail to depict the level of clustering at smaller scales.

We therefore quantify the level of spatial clustering further with a correlogram, which measures the similarity of the variable of interest (i.e., wealth) at various distances. The similarity measure used is Moran's I (Moran, 1950):

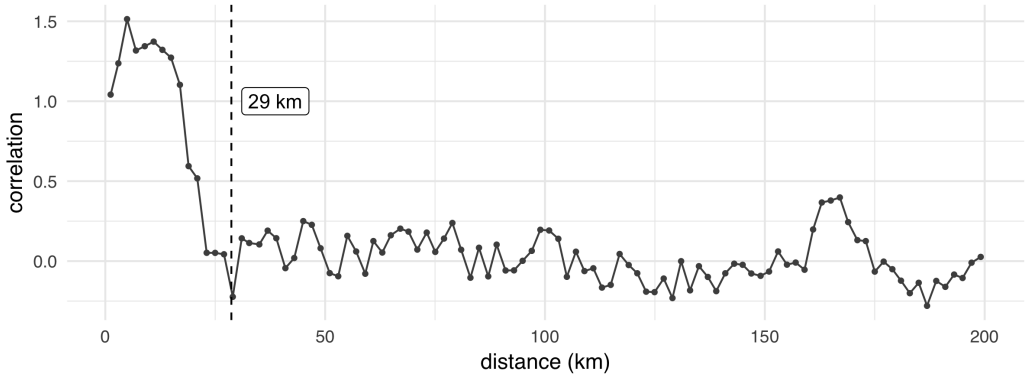
$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2} \tag{3.1}$$

where N is the number of points (cluster centroids), $z_i = (y_i - \bar{y})$ is the deviation from the mean in the quantity of interest (median wealth in our case). The spatial weights matrix w_{ij} is derived from the distance between point pairs, with zeroes on the diagonal (i.e., $w_{ii} = 0$), and W is the sum of all w_{ij}. To produce the correlogram, points pairs are divided into bins according the distance between them, at 2 km increments. Moran's I is then calculated separately for the members of each bin. Positive values of Moran's I indicate the presence of positive spatial autocorrelation and Figure 3.6 depicts the decrease in the strength of spatial autocorrelation of median wealth as the distance between points increases. Note that Moran's I is not guaranteed to fall between 1 and -1.

It is clear from this simple analysis that estimates of wealth at unsampled points derived from nearby sampled points would be significantly more accurate than random guessing. Subsequently, a baseline against which to evaluate predictions from CDR data ought to take proximity to sampled points into account, if these were available. However, it is also clear that, in the case of Senegal and Côte d'Ivoire, many locations are not within range of sampled points for such an approach to be reliable on its own. Furthermore, estimating unsampled locations solely as a function of nearby sample points is likely to miss locations which are outliers relative to their neighbours, and these are arguably among the most important to identify. To establish the extent that this is likely to occur, we measure spatial autocorrelation within the neighbourhood of each point using local Moran's I, or Local Indicators



(a) Senegal



(b) Côte d'Ivoire

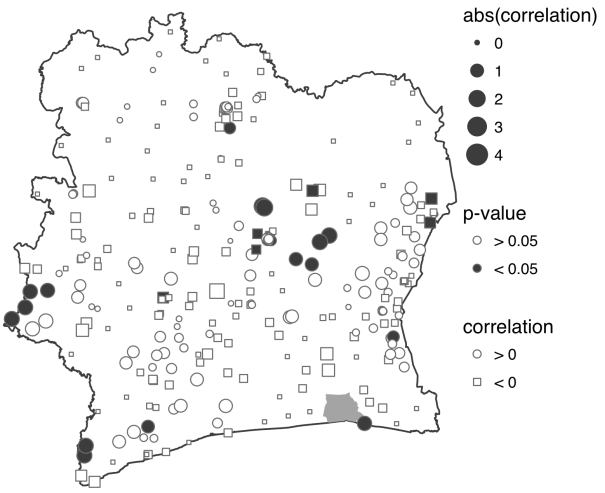
Figure 3.6: Correlograms showing the strength of spatial autocorrelation in wealth according to distance intervals. Distances are binned at 2 km increments with bins containing a minimum of 55 point pairs. Vertical lines indicate the approximate distance at which spatial autocorrelation reduces to a level expected from randomly placed points (fluctuation after this line is not statistically significant and can be considered as noise).

of Spatial Autocorrelation (LISA) (Anselin, 1995):

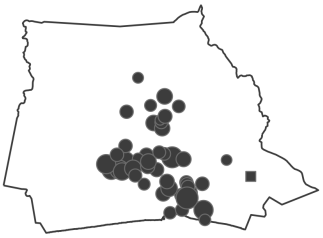
$$I_i = \frac{z_i}{m}\sum_j w_{ij} z_j \tag{3.2}$$

where $m = \sum_i z_i^2 / N$. Figure 3.7 maps the local Moran's I of average wealth in the neighbourhood of each sample point at the spatial scale corresponding to the approximate distance at which the level of spatial clustering reduces to that expected from randomly placed points (i.e., 29 km in Côte d'Ivoire and 64 km in Senegal, indicated by the vertical lines in the correlograms of Figure 3.6). As can be seen, many points have too few neighbours to allow a statistically significant estimate

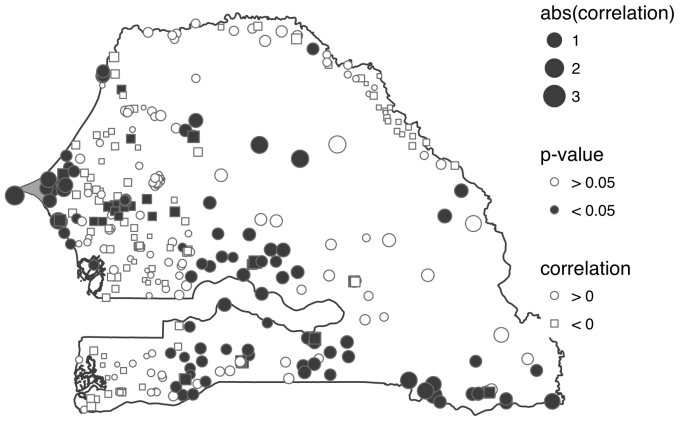
to be computed at the 5% level ($p < .05$), suggesting that poverty estimates derived from proximity alone would perform poorly in these areas. Furthermore, the figures also reveal several sample points negatively correlated with their neighbourhood, which again indicates that relying solely on spatial dependency for estimating unsampled points would be inappropriate in this context.



(a) Côte d'Ivoire



(b) Abidjan



(c) Senegal

(d) Dakar

Figure 3.7: Local Moran's I of wealth at DHS cluster locations in (a) Côte d'Ivoire (b) Senegal and their respective largest cities (c) Abidjan and (d) Dakar. Shape represents the sign of correlation, size of shape represents the magnitude of correlation and solid shapes indicate significance at the 5% level. It can be seen that there exists a number of cluster points that are negatively (square) correlated with their neighbours.

Considering these simple observations, we stipulate that a realistic baseline prediction method ought to take population density into account. In particular, we propose to do so by computing the log of population density, instead of using a binary urban/rural indicator variable, both because the urban/rural designation may not always be readily available, and because it is more appropriate to use a continuous predictor to estimate a continuous outcome.

3.4 The Baseline Models

We now leverage the above observations to construct two simple yet well-grounded baseline models to estimate poverty: one exploits the existing correlation between poverty and population density (as evidenced in Section 3.2); the other expands it, by adding the spatial auto-correlation of poverty (as evidenced in Section 3.3). The former is most suitable when no ground truth poverty data about a country exists; the latter is applicable when partial ground truth data exists instead for a subset of clusters. In order to later assess the predictive power of these new baseline models, relative to a random baseline, we first produce a random baseline ourselves.

3.4.1 Random Baselines

First we created two random baseline models that consist of values drawn from distributions that approximate those exhibited in Figure 3.2. This requires two steps: we first take 5000 random draws from two distinct normal distributions, and concatenate the results to form a vector of length 10000 (the number of draws was chosen simply to be large enough to provide a stable distribution, whilst still being small enough to be fast to compute). The parameters of the normal distributions in the first step were chosen such that the probability density functions of the random vector in the second step visually resemble the density functions of the observed data. For Senegal, these are $\mathcal{N}(20,8)$ and $\mathcal{N}(58,12)$; for Côte d'Ivoire, these are $\mathcal{N}(8,6)$ and $\mathcal{N}(31,10)$.

To measure the predictive performance of these random baseline models we have computed the mean absolute error (MAE), and also Spearman's rank correlation coefficient (ρ) since we are interested in predicting the relative ordering of

locations too. We take 1000 draws (again chosen to be large enough to provide a stable mean) of size n from each respective sample distribution, where n is the number of clusters, and compute the MAE and ρ with each draw, then take the mean over all 1000 draws. Results are shown in Table 3.2. The MAE scores of wealth (23.6 for Senegal and 16.6 for Côte d'Ivoire) provide the benchmark against which to compare later model performance. As expected, the rank correlation of the random predictions is effectively zero. Note that, although random, these baselines are not the most naive predictions since we have injected some knowledge of the underlying distributions. This reflects the more realistic scenario of a researcher having some background knowledge of the underlying distribution of wealth.

Table 3.2: Random baseline metrics

		Wealth
Senegal	MAE	23.6
	Spearman's ρ	0.002
Côte d'Ivoire	MAE	16.6
	Spearman's ρ	-0.001

3.4.2 Baseline Models

The first baseline simply consists of a regression model, where the independent variable is the log of population density for each cluster area, μ, and the predicted response variable $\hat{y}$ will be the median wealth of each area:

$$\hat{y} = \beta \cdot ln(\mu) + \varepsilon. \tag{3.3}$$

For the second baseline, we add a spatially-lagged dependent variable:

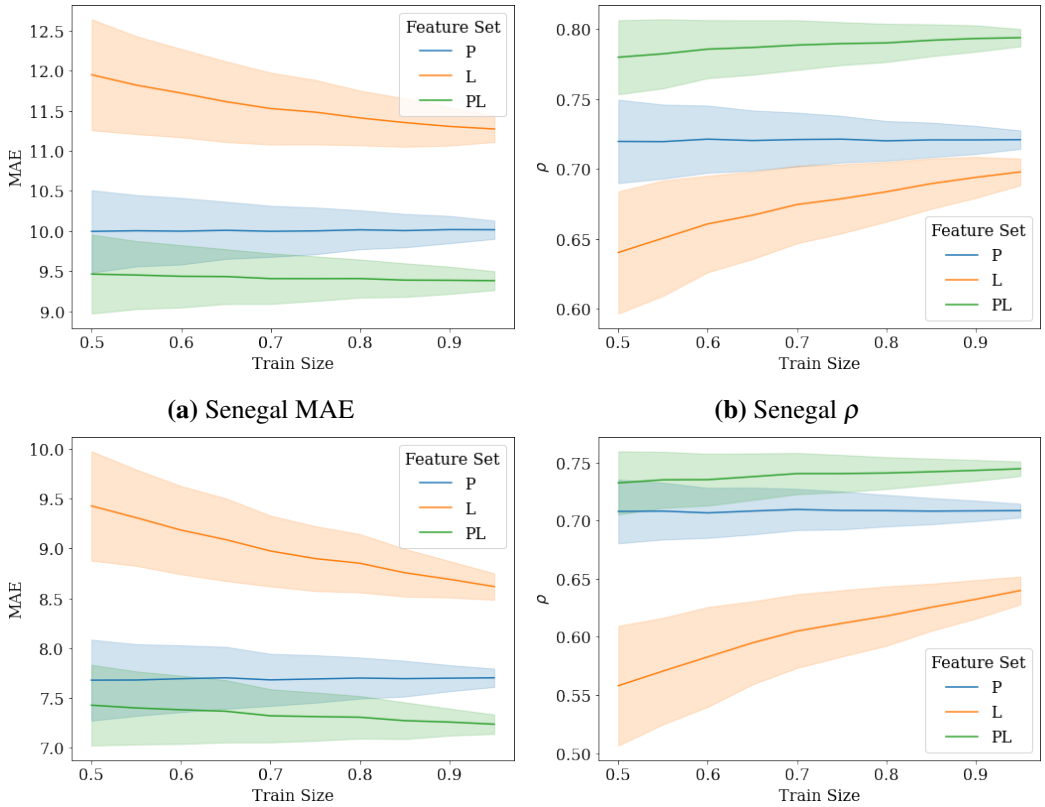
$$z_i = \sum_j w_{ij} y_j, \tag{3.4}$$

where y is the response variable, w is a weight that is inversely proportional to the squared distance between points i and j, and $\sum_j w_{ij} = 1$. This results in the linear model,

$$\hat{y} = \beta_1 \cdot ln(\mu) + \beta_2 \cdot z + \varepsilon. \tag{3.5}$$

By using squared distance to calculate the weights, the effect of distant neighbours on the lagged variable will be negligible for those points with relatively close neighbours, whilst still allowing us to compute a lag for those points with no nearby neighbours.

3.5 Results



(a) Senegal MAE **(b)** Senegal ρ

(c) Côte d'Ivoire MAE **(d)** Côte d'Ivoire ρ

Figure 3.8: Regression test scores for average wealth in Senegal (a, b) and Côte d'Ivoire (c, d). Predictor variables in each model are, P: Population density, L: lag and PL: population density + lag. Bands show the standard deviation.

To obtain a robust measure of predictive performance despite the relatively small number of data points, we varied the proportion of training data and performed 1000 iterations at each proportion, randomly splitting the datasets into training and testing subsets in each iteration. For the spatial-lag baseline, the lagged

variable was computed using only members of the training set in each iteration. Note also that the modelling is always performed separately for Senegal and Côte d'Ivoire - the data from the two countries is not mixed. Figure 3.8 reports the MAE and Spearman's ρ computed with the test data.

The spatial and population density baselines significantly outperform the random baselines in all cases, even with relatively few training points. For example, when estimating wealth in Côte d'Ivoire the random baseline achieved a MAE of 0.210. With 50% training data this reduces to an MAE to 0.116 for the population density baseline, 0.137 for the spatial lag baseline, and 0.110 for the baseline using both population density and spatial lag, the latter representing a 52% reduction in error. As expected, the spatial lag baseline performs less well with fewer training examples, whereas the population density baseline performance metrics remain fairly stable as training size increases. A similar story can be told for baselines in Senegal.

3.6 Discussion

In this chapter, we have found a strong correlation between fine grained poverty data and population density estimates, as well as a significant degree of spatial dependency in the distribution of wealth. We have used these findings to inform the construction of baseline predictive models against which we can fairly compare CDR-based models.

This work is subject to some limitations, owing largely to characteristics of the available data. For our analyses, we have utilised DHS surveys and explored the effect of varying levels of training data. However, it should be noted that the survey clusters themselves represent only a fraction of the census enumeration areas within each country. For example, in Senegal, the survey clusters represent only 4% of a total of 9733 enumeration areas, and on average only 20% of households are surveyed within each selected enumeration area. Consequently, some caution is appropriate when extrapolating the results presented here to the entire country. However, the sampling methodology employed by the DHS surveyors is such that

there is nothing inherently different between the selected enumeration areas and unselected ones, therefore, we would not expect modelling outcomes to differ greatly were we to have access to fully representative data. An important direction for future work would be to perform a comprehensive analysis of a country in which poverty or socio-economic status data is available for every enumeration area.

In addition, we have not explored the effect of the random displacement that is applied to DHS cluster and BTS tower locations. This will introduce a degree of uncertainty in both the predictor and response variables, but we expect this to be largely compensated for by the spatial smoothing that takes place when aggregating predictor variables. Nevertheless, a proper investigation of sensitivity of our modelling approach to such displacement would be prudent.

These results have important implications for the reset of this thesis, and indeed, other related research. Namely, that in order to fairly evaluate the performance of CDR (or other data source) based models, a comparison must be made against the performance of baselines such as those presented here, and moreover, it is not sufficient to compare results only to a random baseline. In the next chapter we turn our attention to the CDR data and to inform the development of our CDR based models we explore the relationship between information diffusion and economic development, before returning to the baseline models in Chapter 5, against which we compare the performance of our CDR based models.

Chapter 4

Information Diffusion and Economic Development

We have so far looked at existing research that seeks to use CDR data to predict wealth in Chapter 2, before overcoming a common shortcoming in these works by establishing baseline model performance in Chapter 3. Prior to presenting our own approach that exploits patterns found in aggregated CDR data to predict wealth in Chapter 5, we first aim to establish a firmer foundation on which to build up our approach, and in particular, answer the question, why should we expect to be able to derive a proxy for wealth or poverty from mobile phone data at all?

Mobile phone ownership and level of usage can reflect the owner's wealth to some extent (mobile phone ownership is one factor among many of which the DHS wealth index is composed), and increased mobile phone adoption can also be a catalyst for economic growth by providing easier access to information at a reduced cost (Aker and Mbiti, 2010). The way in which social network structure mediates access to information has been identified as an important factor in generating individual prosperity. For example Granovetter's strength of weak ties theory (Granovetter, 1973, 2005) and Burt's theory of structural holes (Burt, 1992, 2009) suggest the degree to which personal networks overlap significantly impacts the diffusion of influence and information through a social network. It has further been shown that having a diverse set of contacts is strongly related to living in a less deprived neighbourhood (Eagle et al., 2010).

As well as affecting individual well being, information flow is said to play an important role in determining prosperity and the rate of innovation in cities. Many characteristics of urban areas have been found to scale super linearly with population size, including crime levels, the spread of infectious diseases, and also economic development as measured by Gross Domestic Product (Arbesman et al., 2009; Bettencourt et al., 2007, 2010). This relationship has been attributed to denser social network formation since social networks tend to densify as the number of individuals increases, which in turn increases the capacity of information flow (Bettencourt, 2013; Pan et al., 2013).

These considerations suggest that ranking cities according to population should provide some indication of their relative economic development. Indeed, the analysis of the previous chapter, in which we used population density to define a simple baseline estimator, corroborates this. However, even if such a baseline implicitly captures the effect of information transfer within a city, it ignores the effect of information flowing between cities (and other areas) and instead treats cities as informational silos. Here, we take a wider scope by considering the information flow between cities (and towns and rural areas), and study the relationship between diffusion rates at this scale and wealth, as a means toward answering the questions posed at the beginning of this chapter.

There are two main outcomes of this chapter. First, we construct simulations of country-wide information diffusion processes and present the results, ranking areas according to their access to information. We establish a strong relationship between information diffusion and economic development, represented by the DHS wealth index discussed in the previous chapter. In doing so we lay the ground work for the feature engineering and predictive modelling of Chapter 5. Second, we examine the affect on our results of several contextual factors. We find that the measured strength of association between wealth and access to information is sensitive to the level of spatial aggregation, the coverage of the mobile network, as well the reliability of the ground truth data. These findings highlight important considerations for any interpretation of mobile phone based socioeconomic indicators.

Next we provide some background on information diffusion modelling, before describing the simulation process and discussing its results.

4.1 Information Diffusion

Most previous research investigating the flow of information through networks has taken one of two similar approaches to modelling information diffusion processes (Valente, 1995). The first draws on the field of mathematical epidemiology where models of diffusion were first developed to describe the spread of a disease (Bailey et al., 1975; Monin et al., 1976). These techniques were later applied to the study of the spread of rumours (Daley and Kendall, 1965; Rapoport and Yuan, 1989), news (Deutschmann and Danielson, 1960) and information (Funkhouser and McCombs, 1972). The second approach stems from sociology and focuses more explicitly on how social relationships determine the cascade of information or adoption of innovations (Ryan and Gross, 1943; Griliches, 1957). In a canonical study of the diffusion of innovations, Ryan and Gross (1943) demonstrated the importance of information sharing via interpersonal networks in the adoption among farmers of a newly developed hybrid variety of corn. Similarly, Coleman et al. (1959) found that in a community of physicians peer influence drove the adoption of a new drug more so than positive results from clinical studies. In both cases, the spread of information through social networks paved the way for what would prove to be a beneficial development in each respective community.

4.1.1 Epidemic Models

In the basic propagation models of epidemiology, nodes can take one of three states corresponding to the stages of disease. A person is first susceptible (S) to the disease and can become infected (I) with some probability if exposed to the disease by an infectious contact. That person, or node, is then able to infect their own contacts. Depending on the model, after some time the person can either become recovered (R) and immune, and will be removed from the network (known as the SIR model), or recover but once again become susceptible (SIRS model), or indeed can remain permanently infected (SI model). Early studies of propagation took place on fully

mixed networks in which a node is equally likely to infect any other node. Since then, however, research has considered more realistic models that take into account the structure of social networks (Newman, 2003).

The focus of these works has been primarily on global properties of diffusion processes, such as the *epidemic threshold*, or the minimum transmission probability at which the disease (or information) will spread to a certain fraction of the network. Small-world networks (Watts and Strogatz, 1998) and power law networks, which real-world networks are often found to be (Al Hasan and Zaki, 2011), exhibit strikingly different behaviour in this regard, with the latter often having an epidemic threshold of zero, meaning that an epidemic will always occur with some positive transmission probability.

For example, Wu et al. (2004) study the flow of information through email networks using an SIR model, and compare the results to a modified version in which the transmission probability decays as the distance from the seed node increases. As expected, the decay limits the scope of the spread of information, unlike the original version in which the epidemic threshold is zero as in other scale-free networks. In contrast to the forms of more commonly studied networks, the networks that we study are extremely dense (see Section 4.2), therefore we would expect, as in the case of scale-free networks, that the epidemic threshold would be zero. However, we are not interested so much in global properties such as this, but rather in the behaviour of individual nodes as information propagates.

In this vein, researchers have looked at the importance of nodes in propagating information and the effect that removing the most central nodes has on the diffusion rate, in order to shed light on the resilience of networks (Albert et al., 2000; Callaway et al., 2000), and also on how best to limit the spread of computer viruses via email (Newman et al., 2002). In this work we seek to rank nodes in terms of their influence and ability to acquire information in order understand how this relates to economic development.

Relatedly, Pan et al. (2013) hypothesise that the superlinear scaling of urban characteristics such as wealth and rate of innovation with population size can be

attributed to social tie density, which in turn enables increased flow of information through the population. To test this, they simulated information flow using the SI model in synthetic city-wide social networks and, in support of their hypothesis, found that diffusion rates also scale super-linearly with population size. Similarly, an outcome of the work presented in this chapter is evidence to support the hypothesis that information diffusion is what drives the association between a region's wealth or poverty level, and features of the mobile call graph.

4.1.2 Cascade Models

In contrast to epidemic models, cascade models attempt to capture the decision making process of individuals. Cascade models can be further subdivided into threshold models and independent cascade models. In the linear threshold (LT) model, each node u in the network chooses a threshold $t_u \in [0,1]$, typically drawn from some probability distribution. Every neighbor v of u has a connection weight w_{uv}, and u adopts an innovation from (or is influenced in some other way by) its neighbours if $\sum_{v \in S} w_{uv} > t_u$, where S is the set of nodes that have already adopted the innovation (Watts, 2002).

Independent cascade (IC) models are named so because, unlike the threshold model, the probability that influence propagates from v to u does not depend on the weights of u's other connections, nor on the history of propagation in the network. Rather, if v adopts an innovation, then at the next time step u will adopt with some probability $p_{u,v}$. Moreover, if v fails to influence u at that time step, it will have no further chances (Goldenberg et al., 2001). Gruhl et al. (2004) measure the influence of blog authors on one another by modelling the spread of topics using a variation of the IC model. IC models are also often used in the context of influence maximisation (Kempe et al., 2003) tasks, in which the aim is to find the subset of nodes that can be seeded such that the maximum number of nodes in the network will adopt or be influenced.

Here, we adopt these established diffusion modelling approaches in order to help understand how information flows throughout our countries of interest, and to reveal differences in the time taken for information to arrive in different regions.

4.2 Mobile Call Graphs

Using the D4D CDR data to build our call graphs, we aggregate the total volume between regions over the total period the data covers, giving us a single temporal snapshot in each country. We investigate information diffusion at three levels of geographical aggregation corresponding to different administrative levels. In this setup, nodes of a call graph represent administrative regions and edges represent the total volume of calls between them. We sum calls in both directions so that edge weights $w_{ij} = w_{ji}$, since it is not clear whether the directionality of a call has any bearing on the direction information is passed, and we normalise edge weights by dividing by the maximum edge weight in the network. The aggregated networks each have a density of 1, that is, the networks are completely connected with every region having some level of communication with every other region. This high level of connectivity makes this kind of network unlike the majority of networks studied in relation to information diffusion (Watts, 2002; Kempe et al., 2003; Goldenberg et al., 2001; Gruhl et al., 2004; Pan et al., 2013; Newman, 2003).

The sizes of the resulting call graphs are summarised in Table 4.1. At the third administrative level there are 93 regions in Senegal and 177 in Côte d'Ivoire, and we include all regions as nodes in the call graph in order to more accurately represent the flow of information. However, we present results pertaining only to those regions in which ground truth data was available, which is 92 in Senegal and 138 in Côte d'Ivoire. The same is true of the higher level administrative regions, with the final number shown in parentheses in Table 4.1.

Table 4.1: Number of nodes (regions) of call graphs at different levels of aggregation. The number of regions containing DHS clusters is shown in parentheses.

Adm. level	Senegal	Côte d'Ivoire
Adm. 1	11 (11)	19 (19)
Adm. 2	30 (30)	50 (50)
Adm. 3	93 (92)	174 (138)

In the aggregated networks, since the edge density is 1 the degree distribution is uniform and therefore there is no correlation between a node's degree and its

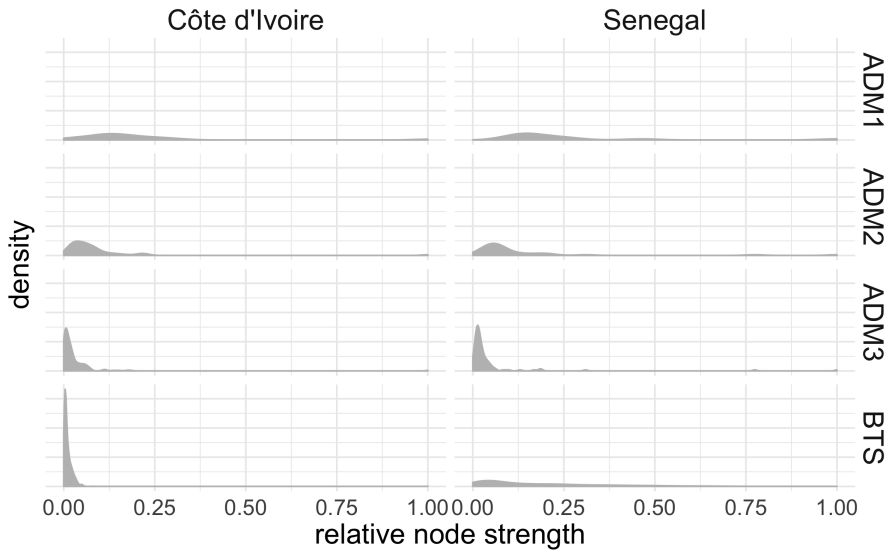
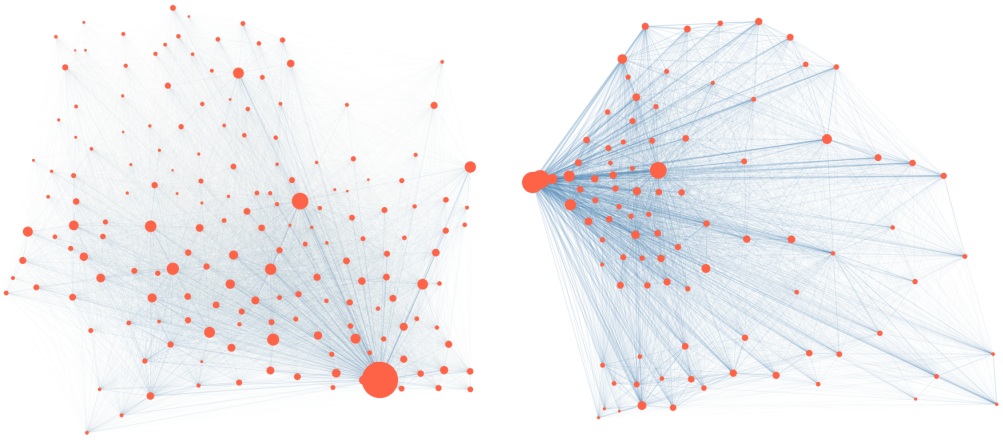


Figure 4.1: Probability density functions of relative node strength (normalised total call volume) for networks at the BTS level and aggregated to the three administrative (ADM) levels

strength (the total edge weight of a node). Topology alone therefore plays no role in determining information diffusion processes. However, the distribution of node strength is highly skewed, owing to the existence of a small number of hubs that account for a large portion of the total traffic on the network. As shown in Figures 4.1 and 4.2, the weight structure of the network is highly heterogeneous and therefore this, rather than topology, will significantly affect diffusion processes across the network.

4.3 Simulation Models

To simulate the flow of information, we experimented with three different models: a simple Susceptible-Infected (SI) model, the Independent Cascade (IC) model and a Linear Threshold (LT) model. For the SI and IC models, at each time step a node is infected by its infected neighbour with some probability relative to the strength of the connection between them, $P(i \rightarrow j) \propto \beta w_{ij}$, where β is a constant controlling the rate of diffusion. For the LT model, a uniform cascade threshold is set for each node, with the node becoming infected if the sum of weights from its infected neighbours exceeded this threshold. For the SI and LT models, an infected node has



(a) Côte d'Ivoire adm. level 3 graph

(b) Senegal adm. level 3 graph

Figure 4.2: To provide a greater intuition into the kind of network under consideration, presented here are call graphs at the third administrative level. The graphs have a geographical layout with nodes positioned at the coordinates of the population weighted centroids of the regions they represent, and with size corresponding to the population of that region. The edge thickness is relative to call volume between nodes (note that node and edge sizes are determined separately for each country's graph). It can be seen that the networks contain a small number of dominant nodes with relatively strong connections.

a chance to infect its neighbours at each time step, whereas for the IC model, the infected node only has one chance to infect its neighbours. Intuitively, the SI and LT models represents the case where the information being transmitted has some long-term value; for example, it could be the adoption of some new technology. The IC model represents the case where the value of the information being transmitted has only temporary value, for example market news related to a certain company or industry, or it could perhaps represent the influence on the sentiment of consumers or producers regarding certain topics. We measure the flow of information through the network by running multiple experiments, with each node playing the role of seed 100 times each. The average time taken for a node to be infected (i.e., receive the information) over all experiments is then used as a measure of access to information at that location. More precisely, we define susceptibility of node v as

$$S_v = 1 / \sum_i t_{v,i}, \tag{4.1}$$

where $t_{v,i}$ denotes the time taken for node v to become infected in experiment i. In other words, we take the inverse of the mean number of steps taken for the node to become infected in experiment in which node v is not the seed.

4.4 Results

We look now at the relationship between susceptibility and economic development, before investigating the effects of various contextual factors.

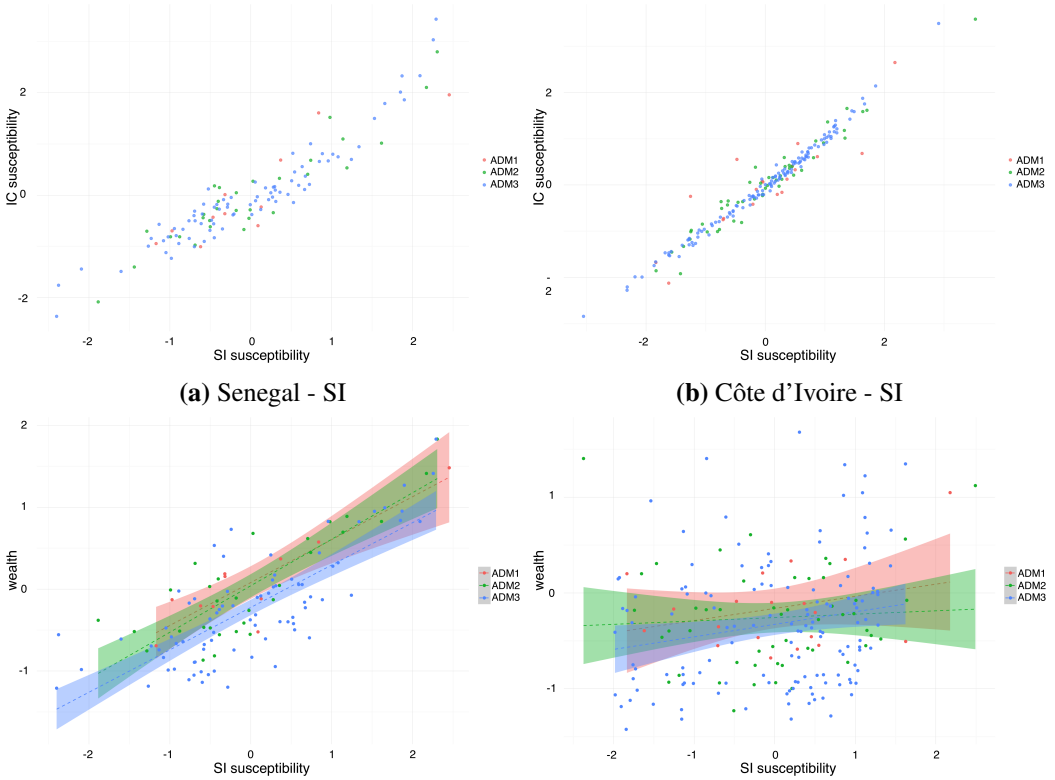
4.4.1 Susceptibility

For the LT model we found that the infection times where largely the same for all nodes and depended only on the choice of threshold and the strength of connections from the seed nodes. Unlike the SI and IC models, the LT model is deterministic, therefore, once the threshold is set sufficiently low for the seed nodes to infect a neighbour, the entire network becomes infected in just 2 or 3 steps. This result shows that for the purposes of distinguishing the role of nodes in information diffusion the LT model is inappropriate for the kind of networks that we are studying, i.e., completely connected. An alternative approach would be to vary the cascade threshold for each node relative to some characteristic. However, we did not explore this option since there were no reasonable candidate variables with which to do this.

For the SI and IC models we investigated the effect of varying β but found that this only has the effect of elongating the distribution of susceptibility, therefore we present here the results for a single value of β.

Susceptibility of nodes as measured in the simulations of both the IC and SI models are very similar, as can be seen in Figures 4.3a and 4.3b. For this reason, going forward we focus on the results of the SI model. To test the hypothesis that access to information is related to economic development we use as a proxy for level of economic development median wealth, as described in Chapter 3, which is derived from the DHS assets index. The number of clusters sampled in Senegal and Côte d'Ivoire is comparable, with 385 (19.25 per million people) in Senegal and 341 (22.7 per million people) in Côte d'Ivoire. The geographical coverage differs significantly however, with 1.95 clusters per km^2 in Senegal and 1.06 clusters per

km^2 in Côte d'Ivoire. We explore these differences further in Section 4.4.2.



(a) Senegal - SI

(b) Côte d'Ivoire - SI

(c) Senegal - IC

(d) Côte d'Ivoire - IC

Figure 4.3: (a) and (b) The Susceptible-Infected model and Independent Cascade model simulations produce similar rankings of areas in terms of susceptibility; (c) and (d) the association between susceptibility and wealth.

In Senegal we find evidence of a strong association between susceptibility and wealth. At the third administrative level we have a Pearson's correlation coefficient of 0.77 (95% CI $[0.67, 0.85]$). However, the relationship appears to be much weaker in Côte d'Ivoire, where the correlation between susceptibility and wealth is 0.31 (95% CI $[0.15, 0.46]$). Figures 4.3c and 4.3d show the association between susceptibility and the median wealth at all three administrative levels. Full results are presented in Table 4.2.

4.4.2 Contextual Factors

4.4.2.1 Representativeness

We can see for Senegal that the higher the administrative level, and correspondingly the greater the level of aggregation, the stronger the correlation between susceptibil-

Table 4.2: Correlation (r) and confidence intervals (CI) between susceptibility in the SI model and wealth at the three administrative levels

	Adm. Level	r	CI
Senegal	1	0.88	[0.60, 0.97]
	2	0.84	[0.68, 0.92]
	3	0.77	[0.67, 0.85]
Côte d'Ivoire	1	0.29	[-0.19, 0.66]
	2	0.17	[-0.11, 0.43]
	3	0.31	[0.15, 0.46]

ity and wealth. Recalling that the DHS assets index is designed to be representative at only the highest (most aggregated) administrative level, a likely explanation is that the strength of the relationship is masked somewhat by the error inherent in the DHS cluster derived ground truth at lower aggregation levels. The correlation remains weak in Côte d'Ivoire at all levels, which at first seems to contradict this hypothesis. However, we note that the number of DHS clusters per region in Côte d'Ivoire is low with a median of 14, 4 and 1 in administrative levels 1, 2, and 3 respectively, compared to 30, 12, and 3 in Senegal. This difference suggests that quality, or rather the representativeness, of ground truth data may be a factor in explaining the poor results in Côte d'Ivoire.

We investigated this aspect further by pruning stepwise the data points (regions) with fewest DHS clusters from the correlation analysis. The results for administrative level 3 are shown in Figures 4.4a and 4.4b. Strikingly, the strength of correlation climbs above 0.93 as we consider only regions with a larger number of DHS survey clusters in Senegal, although as the number of data points, n, decreases, so the confidence intervals tend to widen. For example, if we exclude regions with fewer than 5 clusters the correlation coefficient is 0.91 ($n = 29$, $CI = [0.81, 0.96]$). In Côte d'Ivoire the pattern is similar, albeit less pronounced since n drops rapidly as we prune regions, and consequently the confidence intervals widen. With a minimum of 4 clusters the correlation is 0.58 ($n = 15$, $CI = [0.10, 0.84]$).

4.4.2.2 Volatility

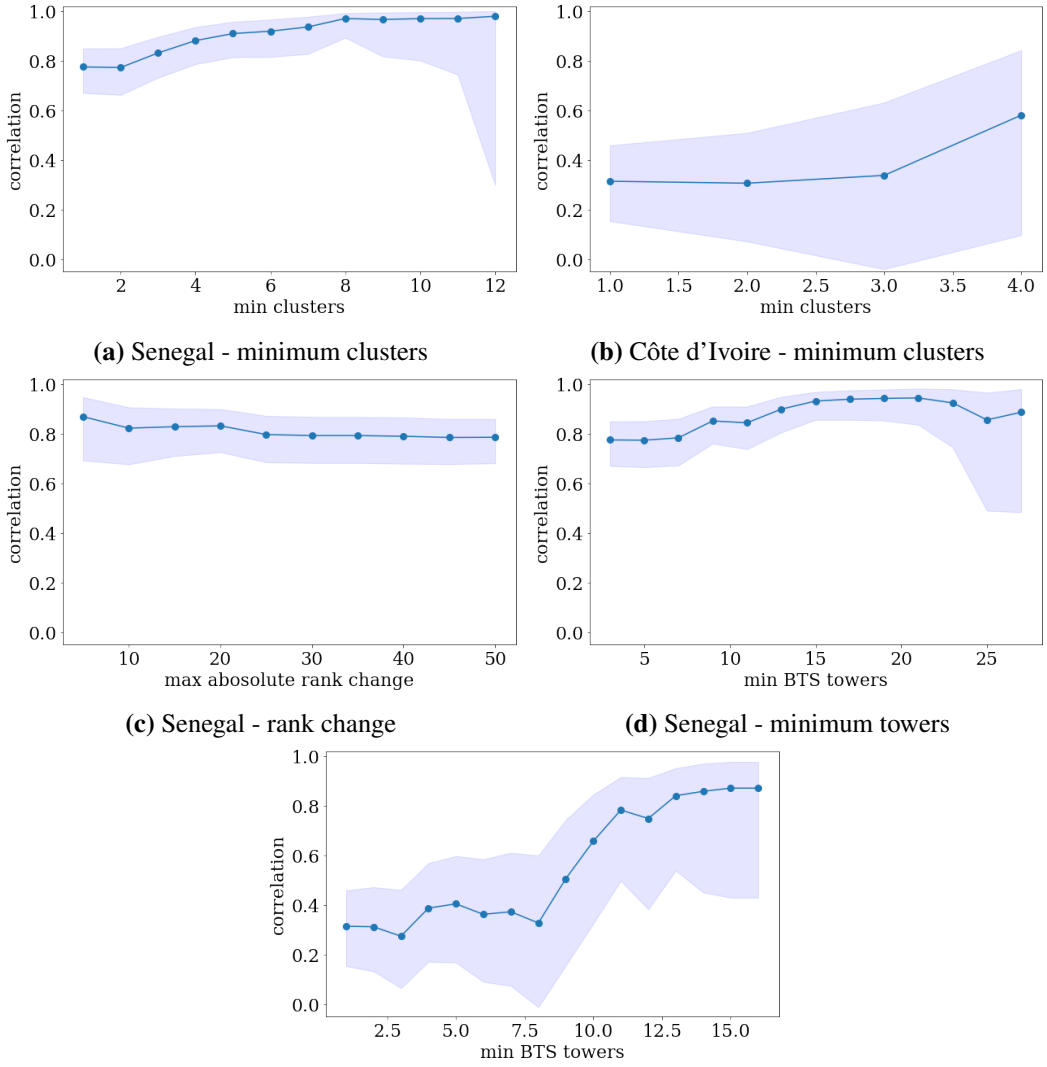
Socioeconomic indicators are naturally dynamic and can change from one year to the next, particularly in sub-Saharan Africa, where economies have experienced both rapid growth and rapid contraction. Considering this, we hypothesise that a further data quality issue is volatility over time in our chosen wealth indicator, which may also affect the strength of the correlation we measure. To test this we compare the 2012 ranking of regions according to the wealth index to the previous DHS survey (2010) and take the absolute rank change. This change could be attributed to either changing circumstances in the region or to inherent variation in the cluster sampling process. Stepwise we prune regions that have the largest absolute rank change so that the dataset contains less volatile regions. This is done only for Senegal as a previous comparable DHS survey was not available for Côte d'Ivoire. As can be seen in Figure 4.4c, there is some evidence that by excluding more volatile regions the measured correlation is higher.

4.4.2.3 Network Coverage

We also investigated whether the coverage of the mobile phone networks could be a factor influencing the relationship between susceptibility and wealth. Poorer coverage in certain parts of a country would mean that the simulation of information diffusion would be less accurate and therefore we might expect weaker correlation. Indeed, we find that there is once again a large discrepancy between the geographical coverage of the mobile networks of Senegal and Côte d'Ivoire. The median number of BTS towers per region at the third administrative level is 11 in Senegal, compared to just 4 in Côte d'Ivoire. As with number of DHS clusters, we prune regions with fewest BTS towers and find that the measured correlation increases, as shown in Figures 4.4d and 4.4e. For example, in Senegal if we consider only regions with a minimum of 13 BTS towers the correlation is 0.90 ($n = 34$, $CI = [0.80, 0.95]$), and likewise in Côte d'Ivoire if we consider only regions with at least 12 BTS towers the correlation is 0.75 ($n = 15$, $CI = [0.38, 0.91]$).

The number of DHS clusters and BTS towers closely follows population density, that is, denser regions tend to have more of each. It is tempting to argue that

the data must therefore be equally representative across all regions. However, such an argument fails to consider the role of geography, that is, that less dense areas also have a much less evenly distributed population, meaning that relatively more DHS clusters or BTS towers may be needed to provide a similar level of coverage as found in denser areas.



(a) Senegal - minimum clusters

(b) Côte d'Ivoire - minimum clusters

(c) Senegal - rank change

(d) Senegal - minimum towers

(e) Côte d'Ivoire - minimum towers

Figure 4.4: Change in correlation between susceptibility and wealth as regions with fewest clusters are removed (a and b), as regions with highest volatility are removed (c), and as regions with fewest BTS towers are removed (d and e)

4.4.2.4 Temporal Coverage

Finally, we investigated the effect of the difference in temporal coverage of our two case studies. Recall that the mobile phone data from Senegal covers a period of 12 months, compared to just 10 weeks in Côte d'Ivoire. To ascertain whether this helps explain the difference in strength of relationship between susceptibility and wealth we subdivided the Senegal data into four sets, with each covering a 3 month period and repeated the simulation experiments on these subsets. We found no substantial difference in the relative edge weights between each 3 month span, only a consistent growth across the whole network. Indeed, the correlation between wealth and susceptibility for each 3 month period is virtually identical. From this we can conclude that a period of the order of 3 months is sufficient to capture the prevailing temporal behaviour in a mobile call graph, and therefore the shorter time span of the Côte d'Ivoire dataset is unlikely to be an important factor in explaining the weaker results in that country.

4.5 Discussion

We have demonstrated a correlation between the flow of information between regions of a country (as revealed by simulations of network diffusion) and the level of economic development of those regions. We found that by using susceptibility, or the average time taken to become 'infected' as a proxy, access to information appears to be related to economic development of a region. The implication for the remainder of this thesis is that by establishing a firm link between access to information and wealth, we have some theoretical grounding from which to begin our hypothesis-led feature engineering process.

Of the two countries under study we have found a strong relationship in one but a weaker relationship in the other. We have conducted further investigations into the effect of contextual factors, the results of which suggest this discrepancy may be partly explained by the differences in coverage of the DHS cluster survey data we used as ground truth, as well as the geographical coverage of the mobile phone networks in each country. Indeed, when focusing on regions with relatively high

mobile network and DHS coverage, the strength of correlation between susceptibility and wealth is found to be significantly higher. These findings have important implications for research into mobile phone based proxies of poverty and similar endeavours. Namely, that by highlighting the effects of data quality and, moreover, by demonstrating increased performance when the data quality issue is removed, we can be much more confident that these methods will work in practice, with the proviso that a minimum level of geographical coverage of the mobile network is met.

Note that we have not attempted to establish a causal link between information flow and economic development, and it may be the case that the social connectivity of a location is reflective of its level of economic development, rather than vice versa, or indeed that they are effects of some third underlying process. Rather, we present these findings in order to both increase the body of evidence establishing the potential for mobile phone based models to help fill the data gaps that exist in many parts of the world, and also to connect the recent advances in using mobile data derived models to predict poverty or economic level with theories that specify the importance of information flow for economic development.

Chapter 5

A Novel Approach to Wealth Prediction with CDRs

We have defined realistic baseline models in Chapter 3 and established a link between the communication patterns revealed by aggregated CDRs and economic development via information diffusion in Chapter 4. We now turn our attention to building and evaluating predictive models that take as inputs features derived from the CDR data. In order to build accurate models of poverty we first need to uncover features of the data that display some relation to the prevailing socioeconomic status of an area. Continuing along the direction taken in Chapter 4, we follow a hypothesis-led approach, focusing on ways in which the data represents behaviour related to economic development, from a simple measure of activity, to more sophisticated measures of network advantage. We then validate each derived feature by examining its correlation with wealth, as well its correlation with other features, before evaluating a number of models trained with different subsets of data and features.

5.1 Hypotheses and Feature Definition

Here we present a number of hypotheses about how economic behaviour may be reflected in the aggregated communication flow patterns, and define features that aim to capture the relationship behind each hypothesis.

5.1.1 Activity

We expect to find that the level of mobile communication activity in an area will reflect its social and economic activity, and thus its level of prosperity. A number of mechanisms have been proposed by which mobile phone adoption could spur economic development, including by reducing the cost of searching for, and accessing information, and by improving the efficiency of supply chain management. Evidence also suggests that mobile phone use is strongly linked to socio-economic status, with early adopters being primarily young, educated, urban males (Aker and Mbiti, 2010).

To capture this relationship, the initial features we compute are simple measures of aggregate activity in terms of call volume (i.e., number of calls) at a node, i (where a node can be a single BTS tower or a spatial aggregation, as described in Section 5.2.1):

- Incoming call volume

$$s_i^{in} = \sum_{j \neq i} w_{ji} \tag{5.1}$$

- Outgoing call volume

$$s_i^{out} = \sum_{j \neq i} w_{ij} \tag{5.2}$$

- Internal call volume (calls to and from the same node)

$$w_{ii} \tag{5.3}$$

- External call volume

$$s_i^{ext} = s_i^{in} + s_i^{out} \tag{5.4}$$

- Total call volume

$$s_i = s^{ext} + w_{ii} \tag{5.5}$$

- Total call volume normalised by population

$$\bar{s}_i = \frac{s_i}{P} \tag{5.6}$$

where w_{ij} is the number of calls from node i to node j. As mobile technology becomes more ubiquitous, and in particular, with mobile phone use rapidly increasing among poorer people (Aker and Mbiti, 2010), we anticipate that the hypothesised link between these simple measures of activity and wealth will erode. This trend motivates the search for more sophisticated metrics that may provide more robust signals of economic well being.

5.1.2 Network Advantage

Our next set of features aims to capture the opportunity for economic development afforded by an advantageous position in the network with respect to the flow of information. In Chapter 4 we found a strong relationship between an area's relative ability to access information, or its *susceptibility*, and its wealth, where susceptibility was computed by taking an average of many hundreds of information diffusion simulations. This method is computationally intensive and scales at $\mathcal{O}(n^k)$, where n is the number of nodes in the network and k is the average degree (in our case $k = n$ so in fact it scales as $\mathcal{O}(n^n)$). Consequently this feature becomes expensive and impractical to compute for networks larger than the relatively small size with which we were operating in Chapter 4 (i.e., networks with up to 138 nodes). For this reason, we opt instead to focus on static network features (i.e., features that do not require computationally expensive simulations to produce) as described below. To help justify this choice we then look at the correlation between these static features and susceptibility.

In studying a social network represented by a fixed-line telephone call dataset, Eagle et al. (2010) showed that the average normalised entropy (in that work referred to as *diversity*) of the social ties of people living in a neighbourhood correlates strongly with the level of socio-economic deprivation (a concept closely related to poverty) in that neighbourhood. In this work we are constrained by the aggregation

of the call records to BTS tower and are unable to look directly at the underlying individual social network. Instead, we hypothesise that the structure of a BTS tower's links will also reflect the poverty level in its location.

We thus extract three measures of a node's integration and importance in the communication network:

- Normalised entropy
$$E(i) = \frac{-\sum_{j \neq i} q_{ij} \log(q_{ij})}{\log(N)} \tag{5.7}$$
- PageRank
$$R(i) = \sum_{j \neq i} \frac{R(j)}{s_j^{ext}} \tag{5.8}$$
- Eigenvector centrality
$$G(i) = \frac{1}{\omega} \sum_{j \neq i} w_{ij} G(j) \tag{5.9}$$

where $q_{ij} = w_{ij}/s_i$ is the fraction of node i's total weight on edge $\langle i, j \rangle$, N is the number of nodes in the graph (equivalent to the degree of node i in a fully connected graph), and where ω is a scaling constant. We exclude node degree as a feature in its own right since the BTS tower network is completely connected and subsequently all node degrees are equal. PageRank (Page et al., 1999; Langville and Meyer, 2005) and eigenvector centrality (Bonacich, 1987) are two recursively defined measures of centrality in which the importance of nodes depends on the strength of connections with other important nodes. Both have previously been found to correlate with a poverty index in Côte d'Ivoire (Mao et al., 2013), and we find that they also strongly correlate with the susceptibility values (Senegal: $\rho = 0.97$, $\rho = 0.93$; Côte d'Ivoire: $\rho = 0.88$, $\rho = 0.95$ for eigenvector centrality and pagerank, respectively, computed on the ADM3 level graph). Conversely, normalised entropy appears to be weakly negatively correlated with susceptibility (Senegal: $\rho = -0.36$; Côte d'Ivoire: $\rho = -0.34$), which suggests that this metric does not capture network advantage in the same manner as centrality measures. This is perhaps not surprising, since it aims to measure the diversity of connections with immediate neighbours only and does not consider the network structure beyond.

5.1.3 Introversion

We hypothesise that an area's level of *introversion* may be a signal of its poverty level. In other words, if an area has relatively low volume of traffic to other areas compared to the traffic within it, the less likely it will be able to benefit from new sources of opportunity arising further afield. The intuition here is that more introverted areas will have access to fewer resources and thus less opportunities for economic development. This is similar in spirit to the theory of open economies, albeit on a different scale, which expects nations that close their borders to international trade to fare less well than those that are more open (Sachs and Warner, 1997). In conjunction with our first hypothesis, that higher activity reflects lower poverty, for two regions with equal activity we would expect the area with lower introversion to have the lower poverty level. It is also related to the idea of network advantage, except that we now explicitly take into account geography and consider only a binary property of flow, that is, whether it is internal or external to the area. A caveat to the above hypothesis is that we may expect denser areas to naturally exhibit higher introversion given that there will be a higher likelihood of communications taking place within the vicinity. However, since density of BTS towers tends to follow population density, the coverage of individual towers in dense areas is smaller, thus mitigating somewhat against the higher likelihood of internal communications.

This metric computes the ratio of internal call volume (source and target are one and the same) to external call volume (source and target are different) of a BTS tower. We calculate the introversion of BTS towers with the following equation:

$$I(i) = \frac{w_{ii}}{s_i^{ext}} \tag{5.10}$$

Intuitively, values of I less than 1 indicate more introverted areas (i.e., internal flow is higher than external flow), and conversely, values greater than 1 indicate more extroverted areas.

5.1.4 Interaction Model Residuals

We next hypothesise that the difference between observed and expected flows between areas reflects the level of social and economic activity in those areas, and thus will be related to poverty. That is, from the residuals between observed and expected flows we aim to capture the restricting effect of poverty on an area's interactions with others. This hypothesis takes a cue from the gravity model of interaction. First introduced in Zipf (1946), gravity models rest on the observation that the size of flow between two areas is proportional to the mass (i.e., population) of those areas, but decays as the distance between them increases. Despite some criticisms (Simini et al., 2012; Yan et al., 2014), the model has been successfully used to describe macro scale interactions (e.g., between cities, and across states), using both road and airline networks (Barrat et al., 2004; Jung and Wang, 2008) and its use has extended to other domains, such as the spreading of infectious diseases (Balcan et al., 2009; Viboud et al., 2006), cargo ship movements (Kaluza et al., 2010), and to model intercity phone calls (Krings et al., 2009).

The simplest form of gravity model has a single scaling parameter:

$$g_{ij} = \beta \frac{P_i P_j}{d_{ij}^2} \tag{5.11}$$

where P_i is the population of area i and d_{ij} is the Euclidean distance between BTS tower locations i and j. The scaling parameter β scales the estimates to bring them in line with observed weights and is fitted to each dataset separately. In general, β depends only on the period of observation. In addition to this simple version, more nuanced models exist to estimate flows, including a 4-parameter version of the gravity model:

$$g_{ij} = \beta_1 \frac{P_i^{\beta_2} P_j^{\beta_3}}{d_{ij}^{\beta_4}}, \tag{5.12}$$

in which scaling coefficients are also fitted to the mass and distance variables. A common problem with the gravity model is that it often fits less well at shorter

distances, hence the existence of a 9-parameter distance-varying version:

$$g_{ij} = \begin{cases} \beta_1 \frac{P_i^{\beta_2} P_j^{\beta_3}}{d_{ij}^{\beta_4}} & \text{if } d_{ij} < \beta_9 \\ \beta_5 \frac{P_i^{\beta_6} P_j^{\beta_7}}{d_{ij}^{\beta_8}} & \text{if } d_{ij} >= \beta_9 \end{cases} \tag{5.13}$$

in which the parameters are allowed to change at some distance threshold, β_9. The more recently introduced radiation model is a parameter free model that aims to overcome some of the inconsistencies of the gravity model and has been shown to reproduce communication volumes between regions more accurately than the gravity model (Simini et al., 2012). One of the main conceptual differences between the gravity model and radiation model is that the latter takes into account the population between the source and target locations as opposed to the distance alone. More precisely, it includes the population with a circle around location i of radius d_{ij}, here denoted with s_{ij}:

$$g_{ij} = g_i \frac{P_i P_j}{(P_i + s_{ij})(P_i + P_j + s_{ij})}, \tag{5.14}$$

where $g_i = P_i \frac{P_c}{P}$ is the population at location i scaled according to the proportion of callers in the total population $\frac{P_c}{P}$. In practice $\frac{P_c}{P}$ acts as a scaling factor that we fit to the data in a similar manner to β in Equation 5.11 so that the flow estimates match the scale of the flows in our data.[1]

For the gravity models, to represent the *mass* at node locations we take an approach similar to that described in Section 5.2.1 in order to account for the fact that the exact spatial coverage of each BTS tower is unknown. That is, we take the mean population of all population raster cells within 30km of node i (where a node can be a single tower or multiple adjacent towers), denoted by H_i, weighted by the inverse squared distance from the node,

$$P_i = \sum_{h \in H_i} P_h / d_{ih}^2. \tag{5.15}$$

[1] In the original presentation of the radiation model in Simini et al. (2012) g_i represented the number of daily commuters at location i.

This effectively gives an estimate of the population density within the node's coverage area. For the radiation model, s_{ij} is similarly calculated overlaying the population raster data with a circle of radius d_{ij} and summing the values of the enclosed raster cells.

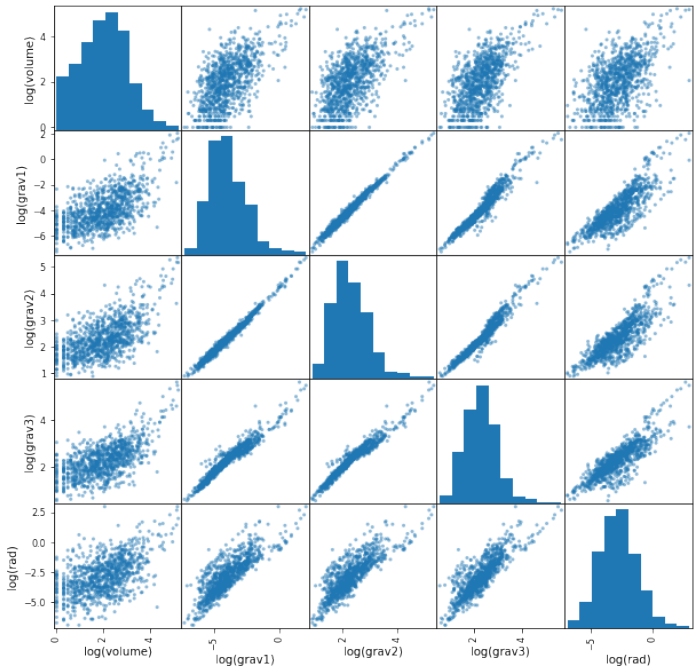
For each of the interaction models we compute the residual flows on each edge as $z_{ij} = w_{ij} - g_{ij}$ (each model contains a scaling factor so we end up with reasonable values for $z_i j$), and to produce features associated with a node i we calculate the mean of the negative and positive residuals separately so that we they do not cancel each other out. Consequently, we have two residual flow features for each model:

$$\begin{aligned} res_i^{-ve} &= \frac{1}{|Z_i^{-ve}|} \sum_{j \in Z_i^{-ve}} z_{ij}, \\ res_i^{+ve} &= \frac{1}{|Z_i^{+ve}|} \sum_{j \in Z_i^{+ve}} z_{ij}, \end{aligned} \tag{5.16}$$

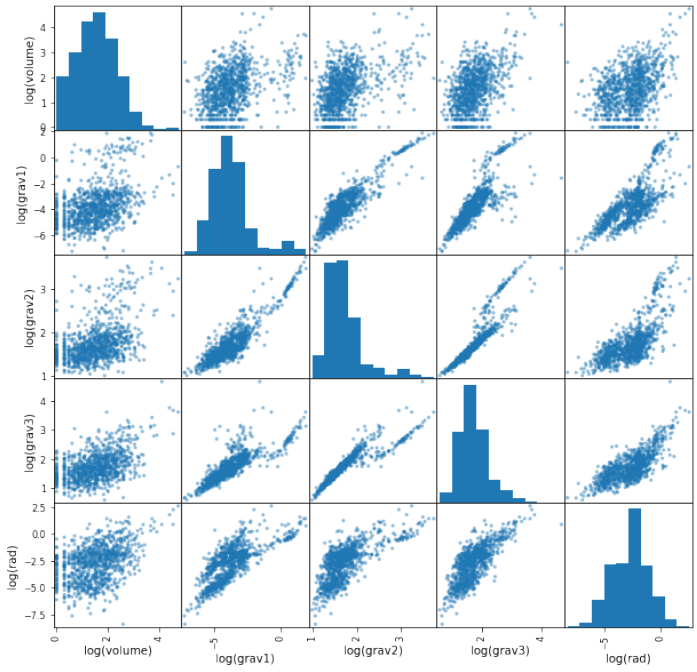
where Z_i is the set of edges connected to i that are negative or positive depending on the superscript.

In addition to the residual flows, we also compute the other features defined in this section on the estimated flows (i.e., introversion, PageRank, eigenvector centrality, and entropy), and take the residuals of these as additional features. The motivation is the same, which is that we surmise that differences between observed and expected values of these features could indicate lower levels of wealth. For example, high entropy gravity residual could indicate that that location is experiencing a less diverse set of interactions with other areas than might be expected.

We experimented with all three gravity model variants and the radiation model. Figure 5.1 shows how well the estimated flows from each model fit the observed volume (top row in each subfigure) as well as how closely the estimates correlate with those of the other models, which will help us to anticipate potential differences in predictive performance when comparing residuals as model features later in Section 5.2.2. It can be seen that in Senegal all model estimates show a moderately strong correlation with observed volume flows, with the simple gravity model, 4 parameter, 9 parameter and radiation model having a Pearson's r of 0.612, 0.619,



(a) Senegal



(b) Côte d'Ivoire

Figure 5.1: Scatter matrices showing the correlation (of logged values) between volume and each of the flow estimates and between each pair of flow estimates. *grav1* is the simple gravity model, *grav2* is the 4 parameter version, *grav3* the 9 parameter version, and *rad* is the radiation model. Note that these plots show a sample of 1000 points.

0.631 and 0.535 respectively. We also see in Senegal that the gravity variants are strongly correlated with each other as is the radiation model, albeit less so. In Côte d'Ivoire the correlation with volume is present but weaker, with an r of 0.409, 0.450, 0.492 and 0.411 respectively. Note also the bimodal distributions in the 1 and 4 parameter gravity models - evidence that they are not able to fit well the full range of distances and motivation for the 9 parameter version. We also see that in Senegal the variants of the gravity model residuals are strongly correlated with each other, whereas in Côte d'Ivoire they are less so. Together with the overall better fit of the gravity models in Senegal, this suggests that the choice of gravity model variant may have less impact when the underlying data is more representative, as it is Senegal. The radiation model fits the observed volumes less well than the gravity models in both countries. However, it does not follow that the residuals of the radiation model will be less predictive of poverty. In fact, by fitting parameters to the observed flows, the gravity model may mask the very signal we hope to uncover. The predictive power of all the residual based features will be examined in the next section. Note that we exclude negative residual features from the radiation model as the majority are null owing to the radiation model's tendency to underestimate flows (i.e., the majority of residuals from the radiation model are positive).

5.2 Method

Here we outline the steps required in order to go from the features as defined above to trained models and wealth predictions.

5.2.1 Spatial Aggregation

In Chapter 3 we created baseline estimators using the DHS clusters as data points. Then in Chapter 4 we looked at the flow of information between regions within the country at three levels of aggregation corresponding to three administrative divisions. Here, in order to compare predictions with the baseline models we again use the DHS clusters as data points, or rather, the grid cell that a cluster finds itself in, as explained below.

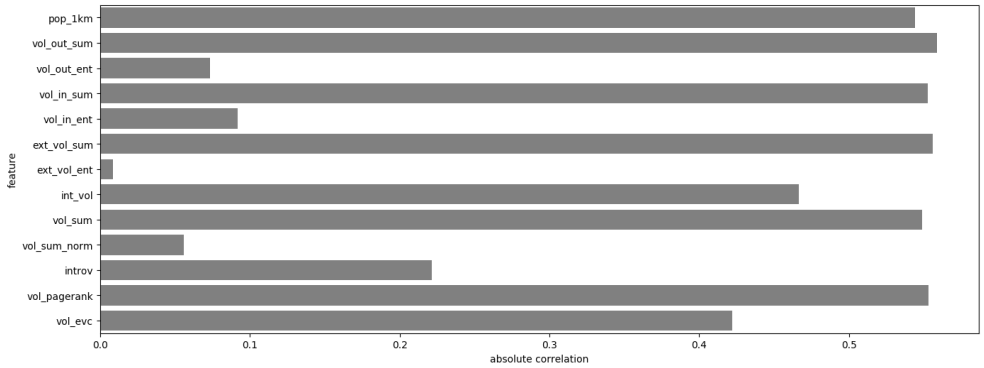
The spatial distribution of BTS towers tends to align with a country's pop-

ulation density distribution and in doing so can be highly skewed. In the most densely populated areas such as city centres, towers can be situated within a few hundred metres of each other, whereas in sparsely populated regions there may be several hundred kilometres between towers. Furthermore, in densely populated areas several BTS towers may have overlapping coverage and the tower that a phone connects to will not always be the closest but will depend on the directionality of the tower cells, the velocity of the caller and other load balancing considerations. For all towers, the maximum range in any given direction is determined by many things, including its design, configuration, the local terrain and climate, and the actual coverage area and load balancing policy of each BTS tower is unknown to the author. To account for this uncertainty in BTS coverage we took two steps. Firstly, we aggregated the call volume of towers within close proximity. To do this we overlaid a 1km hexagonal grid and treated all towers with the same grid cell as a single node within the call graph. The edge weights of the combined towers are summed and this step takes place before the input features are computed. The second step is to use an inverse-distance weighting scheme when aggregating features, which has the effect of creating a smoother feature surface than alternative approaches such as using Voronoi cells. More precisely, the value of a feature, λ_h, at grid cell h is the inverse-squared-distance weighted mean of the feature, λ_n, at all nodes n in the call graph G:

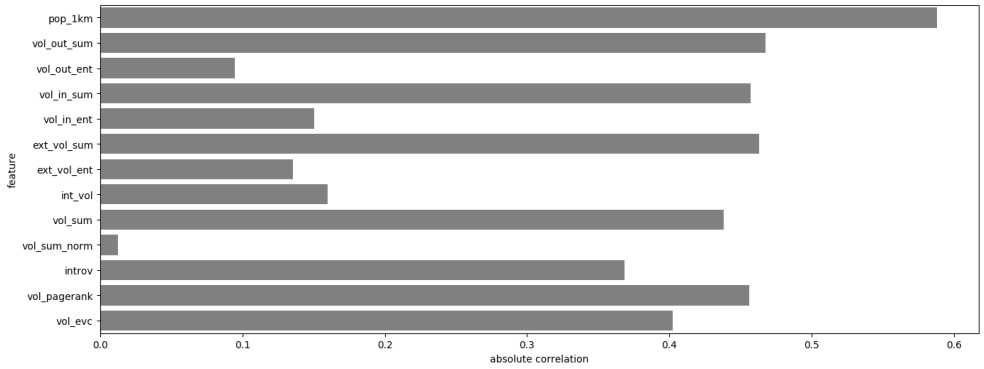
$$\lambda_h = \sum_{n \in V(G)} \frac{1}{d_{hn}^2} \lambda_n, \tag{5.17}$$

where a node could be a single tower or an aggregation within a grid cell, and d_{hn} is the distance from cell h to node n.

We avoid applying a limit on the distance to ensure that all grid cells are assigned a value. Using squared distance as weights also means that, in denser environments many towers (nodes) will be much closer to the assignee grid cell, meaning that the effect of more distant towers on the computed mean will be negligible. We chose the distance weighting approach over Voronoi cells, which is the more common approach in related work, since the latter assumes that call activity at any given location is captured solely by the closest tower - an assumption that we have



(a) Senegal



(b) Côte d'Ivoire

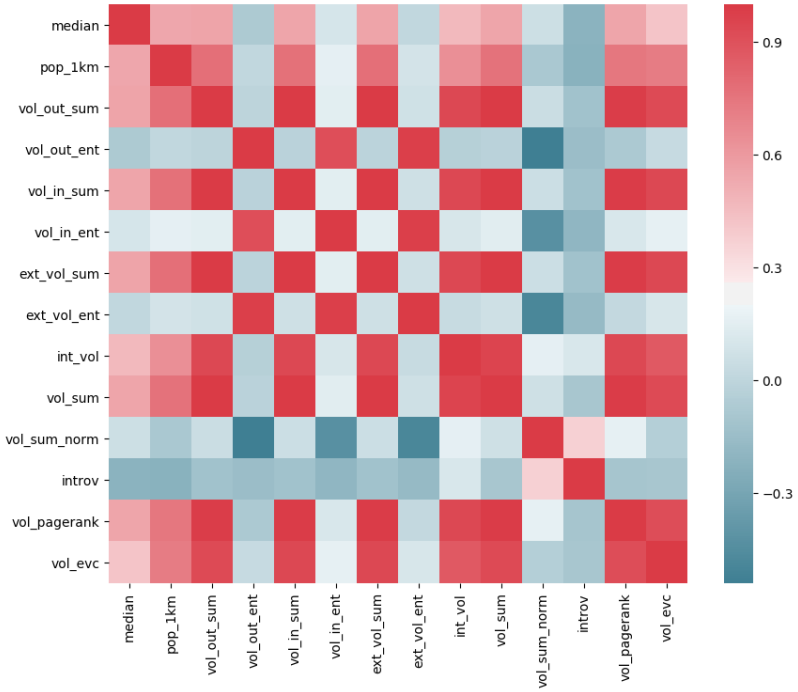
Figure 5.2: Bar plots indicating the correlation between features (excluding residual features) and the target variable, median wealth.

argued above is unrealistic.

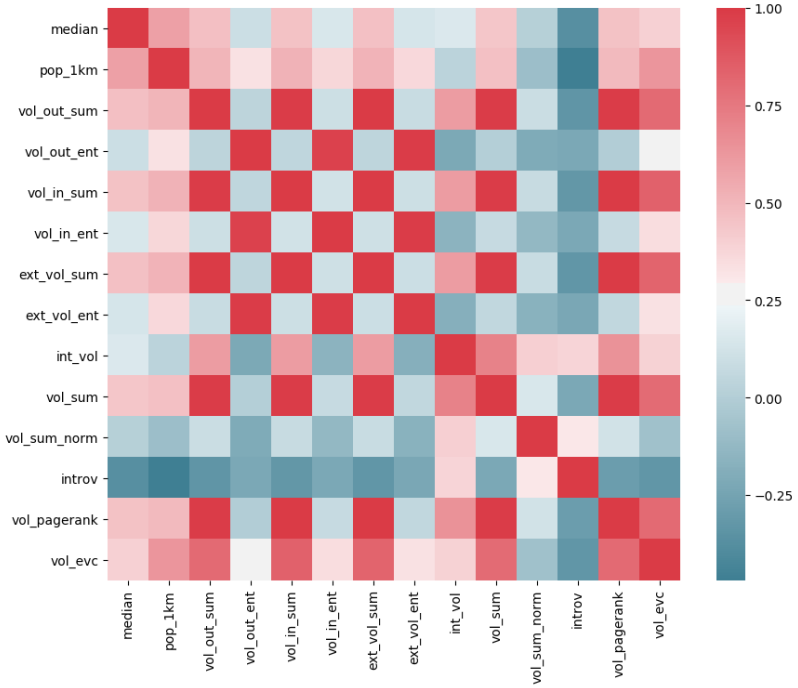
To associate features with DHS cluster points we simply take the values of the grid cell within which the cluster point resides.

5.2.2 Feature Validation

Next, we examine the relationship between each feature and the target variable, as well as the cross correlation among the features. Figures 5.2 and 5.4 show the magnitude of the Spearman's rank correlation coefficient, ρ, between input features and median wealth at cluster locations. Population within a 1km radius is also included as a benchmark, which we know from Chapter 3 is highly correlated with wealth. Figures 5.3 and 5.5 show the cross correlation (ρ) among features. Interaction model residual features are separated from other features in order to keep the correlation matrices to a reasonable size.

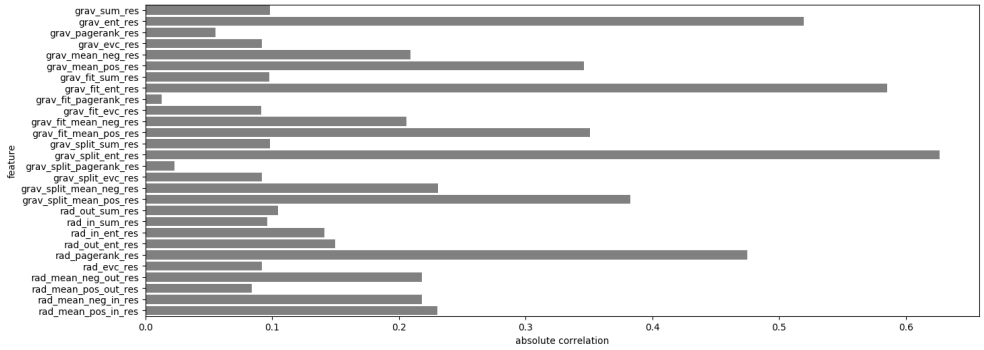


(a) Senegal

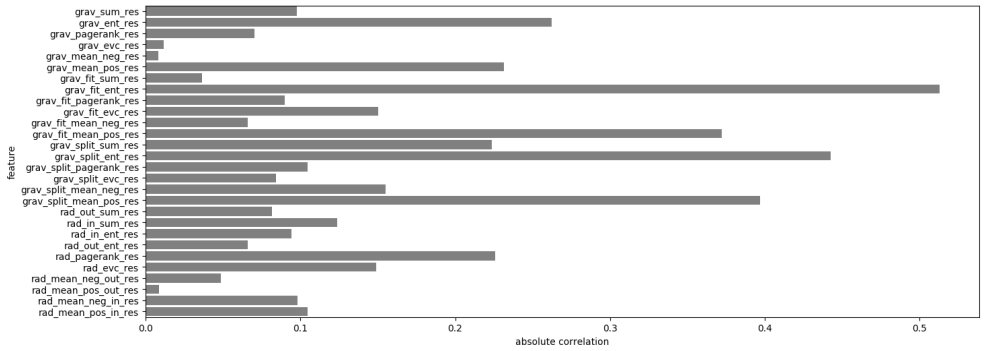


(b) Côte d'Ivoire

Figure 5.3: Correlation matrices indicating the cross correlation between features (excluding residual features).



(a) Senegal



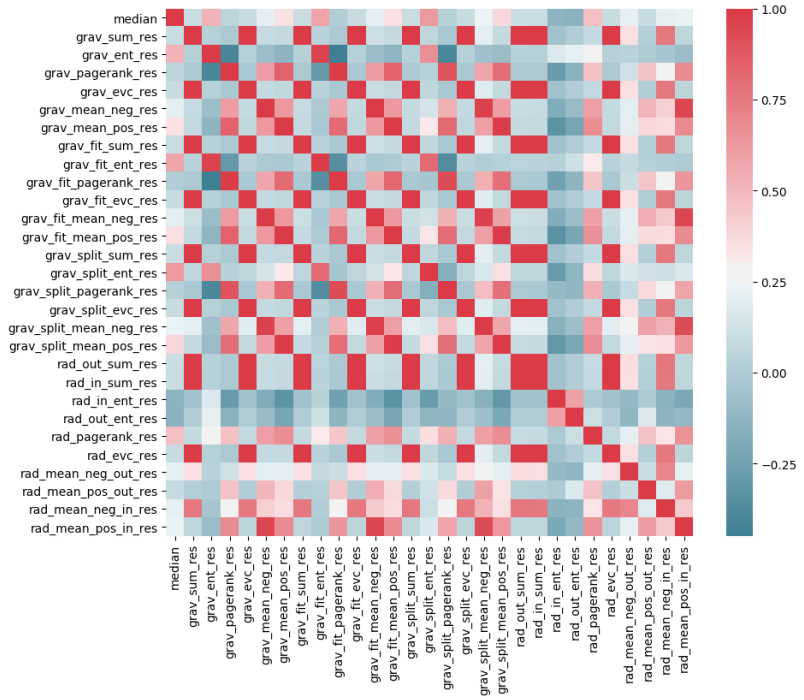
(b) Côte d'Ivoire

Figure 5.4: Bar plots indicating the correlation between residual features and the target variable, median wealth.

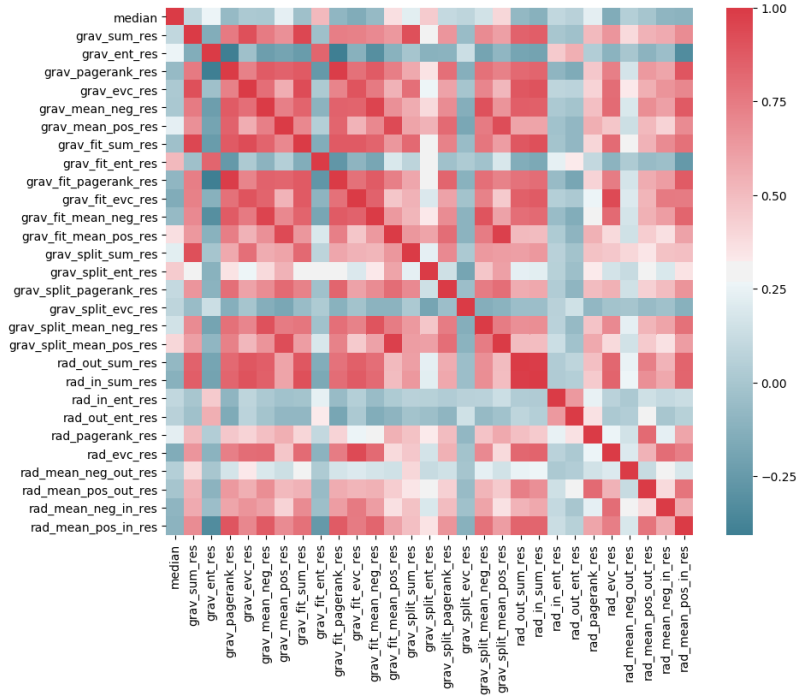
5.2.2.1 Activity

Looking first at activity levels, we see that, somewhat counter intuitively, normalised call volume, or the average volume per person (vol_sum_norm) has no correlation with wealth in either Senegal nor Côte d'Ivoire. This can most likely be explained by population within a radius of 1km (effectively population density) not being an accurate measure of the number of callers contributing to the total call volume for a particular BTS tower. As population density rises, so does the density of BTS towers, and the coverage of individual towers depends on many factors. Ideally, normalised call volume would be computed directly from individual CDRs.

On the other hand we see that other measures of activity have a relatively strong correlation with wealth at cluster locations. Incoming volume (vol_out_sum), outgoing volume (vol_out_sum), external volume (ext_vol_sum) and internal volume (int_vol) all show a similar strength of correlation as population in both Senegal and Côte d'Ivoire, albeit slightly weaker in Côte d'Ivoire. Figure 5.3 shows that these



(a) Senegal



(b) Côte d'Ivoire

Figure 5.5: Correlation matrices indicating the cross correlation between residual features.

activity indicators are also, unsurprisingly, all highly correlated with each other ($\rho > 0.9$), therefore we take only total volume (vol_sum) forward as a feature.

This confirms that aggregated communication activity provides a simple proxy for poverty level; however, as mentioned in the previous section, this relationship may depend in part on the maturity of the mobile telecoms market. Therefore we are particularly interested in the results of the remaining features since these are potentially more robust in the face of market saturation.

5.2.2.2 Network Advantage

We computed normalised entropy across edges on outgoing volume (vol_out_ent), incoming volume (vol_in_ent), external volume (ext_vol_ent) and found only a weak correlation with wealth at cluster locations, and significant only in Côte d'Ivoire. Despite the weak correlation, we retain entropy as a feature in case it turns out to be predictive in combination with other features. We also see that these measures of entropy are highly correlated with each other, therefore we take entropy on external volume (ext_vol_ent) forward alone.

PageRank and eigenvector centrality are computed on the external volume (vol_pagerank and vol_evc, respectively) and we find the these correlate with wealth fairly strongly. This echos the results of Chapter 4 in which we found that a cluster's *susceptibility* to collecting information in simulated experiments also correlates with its wealth, and is in line with previous work which found a link between network advantage and socioeconomic deprivation at the individual level (Eagle et al., 2010), suggesting that similar forces are at play in bestowing greater opportunity to those areas with increased access to sources of information.

5.2.2.3 Introversion

We find that introversion has a fairly weak, but significant, negative correlation with wealth, which, in line with our hypothesis from the previous section, provides some evidence that the more introverted an area the lower its average wealth.

5.2.2.4 Interaction Model Residuals

Turning now to the features derived from interaction model residuals, the most consistent result that can be seen in Figure 5.4 is the fairly strong correlation of residual entropy of the gravity model variants and wealth. This feature captures the diversity

of the residuals from each model, so higher values indicate a more even spread of residual sizes on a node's edges, whereas lower values indicate the presence of one or a few large residuals that dominate others. The correlation is positive, suggesting that the existence of a small number of dominating residuals is associated with lower wealth.

We also see a weak but significant and consistent positive correlation between the mean positive residuals of each gravity model variant and wealth. Positive values of this feature mean that the interaction model tends to underestimate the total volume on a node's edges, so this result suggests that areas with higher than expected call volume are also wealthier. This supports our hypothesis that areas with lower than expected interaction experience lower wealth since the gravity models are fitted to the data. This is further supported by the weaker correlation between the mean negative residuals and wealth. Here, larger magnitude negative residuals mean that volumes tend to be overestimated, so a positive correlation suggests that areas that are overestimated more also tend to be less wealthy.

In general, the pattern of correlations is similar between the different gravity models and countries, no one type of model residual appears to be a better predictor of wealth from this analysis. For the radiation model the pattern is different. Other than PageRank, residuals are computed separately on incoming and outgoing call volumes and only have very weak correlations with wealth or none at all. PageRank on the radiation residuals has a fairly strong correlation in Senegal, but this could be put down to the fact that PageRank tends to be strongly correlated with population density.

5.2.3 Feature and Model Selection

Having pruned a number of features based on strength of correlation with the target variable and cross correlation with other features, we continue a manual feature selection and model comparison process before settling on the final feature set and model type. We purposefully avoid an automated method in order to maintain consistency in the features selected, specifically when considering the interaction model residual based features. That is, to acquire a more easily interpretable feature set

we wish to choose residual features from a single interaction model, whereas an automated feature selection approach would likely result in a mixture of sources. We tested the following sets of features:

- *P* - log of population density alone
- *L* - spatially lagged wealth alone
- *PL* - both log of population density and spatially lagged wealth
- *C* - mobile phone (CDR) based features alone
- *CPL* - all features

We tested the above sets of features separately for each of the following interaction model types. In each case the C features include the residuals from the respective interaction model.

- *gravity* - the single parameter gravity model
- *gravity-fit* - the 4 parameter gravity model
- *gravity-split* - the 9 parameter gravity model
- *radiation* - the radiation model

We compared the following machine learning models:

- *LR* - linear regression model
- *RF* - random forest model

The classic LR model, fitted using ordinary least squares regression, provides more easily interpretable set of feature weights, allowing us to identify the effect of each feature on the model's predictions. Interpretability is important when considering the use case of policy makers using model outputs to inform policy decisions, when they will need to provide clear justifications. However, LR models are known to perform less well in the presence of correlated input variables, and in such cases

the model coefficients also become less easy to interpret. Therefore, we also tested the RF model in order to better establish the predictive power of selected features. We chose Random Forest over alternative machine learning models, such as artificial neural networks, since they are known to perform well on a range of different problems whilst maintaining a reasonable level of interpretability by allowing the inspection of feature importance, and are less sensitive to correlated inputs than the LR model.

To compare model performance we take a staged approach. Firstly, we focus on the C features alone for each model type in order to identify the most predictive set of interaction model residual features - 8 models in total for each country. With the chosen set of C features we then compare different feature sets for each model type in order to determine the added predictive power of the C features over the baseline models introduced in Chapter 3 - a total of 16 models for each country, as summarised in Table 5.1. To see how well they perform with limited data, each model with each set of features is trained with different proportions of training to test data, ranging from 50% to 90%. Taking both Côte d'Ivoire and Senegal together, 160 models are compared in total in the following comparison. For each training proportion, in order to obtain a robust measure of predictive performance, we ran 100 iterations of random train/test splits and took the mean over all iterations. Note that the 100 splits are a subset of the 1000 splits generated for the baseline models, discussed in Section 3.5. We reduced the number of iterations to 100 at this stage in order for the training to complete in reasonable time.

5.3 Results

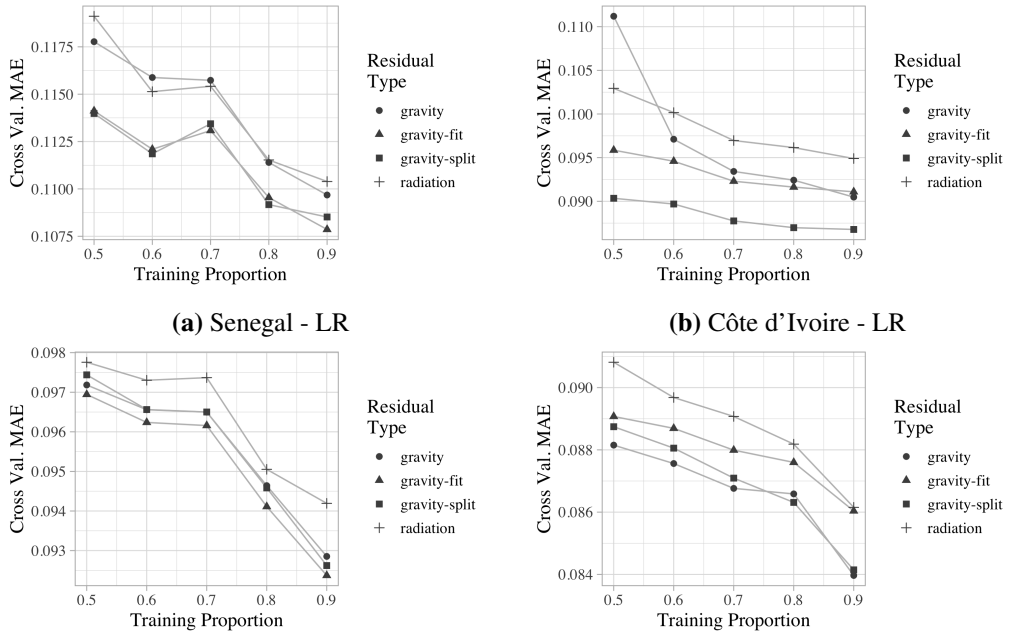
Firstly, in order to determine which residual-type features give the best predictive performance, we compare models that include only CDR features. Figure 5.6 shows that for the gravity and gravity-fit model residuals the results are inconsistent. For example, the gravity residuals are among the worst performing in Côte d'Ivoire with the LR model type and with training proportions below 80%, but among the best performing with RF model type. On the other hand, in the majority of cases

#	Feature Set	Residual Type	Model Type
1	C	gravity	LR
2	C	gravity-fit	LR
3	C	gravity-split	LR
4	C	radiation	LR
5	C	gravity	RF
6	C	gravity-fit	RF
7	C	gravity-split	RF
8	C	radiation	RF
9	P	NA	LR
10	L	NA	LR
11	PL	NA	LR
12	CPL	*best*	LR
13	P	NA	RF
14	L	NA	RF
15	PL	NA	RF
16	CPL	*best*	RF

Table 5.1: A summary of the combinations of different features and model types that are compared. Note that *best* refers to the best residual feature type as determined by the comparison of models 1-8. Each of the 16 combinations is tested with 5 different training set proportions and for each country, making a total of 160 models. Each model is then trained and tested 100 times with random train/test splits to give cross-validated scores.

models with features derived from residuals of the radiation model have relatively poor performance across all training proportions and with both LR and RF model types, and indeed, models with features derived from residuals of the gravity-split model general are among the best performing across the board. For this reason, the remaining analysis of results will focus on models containing only the gravity-split residuals.

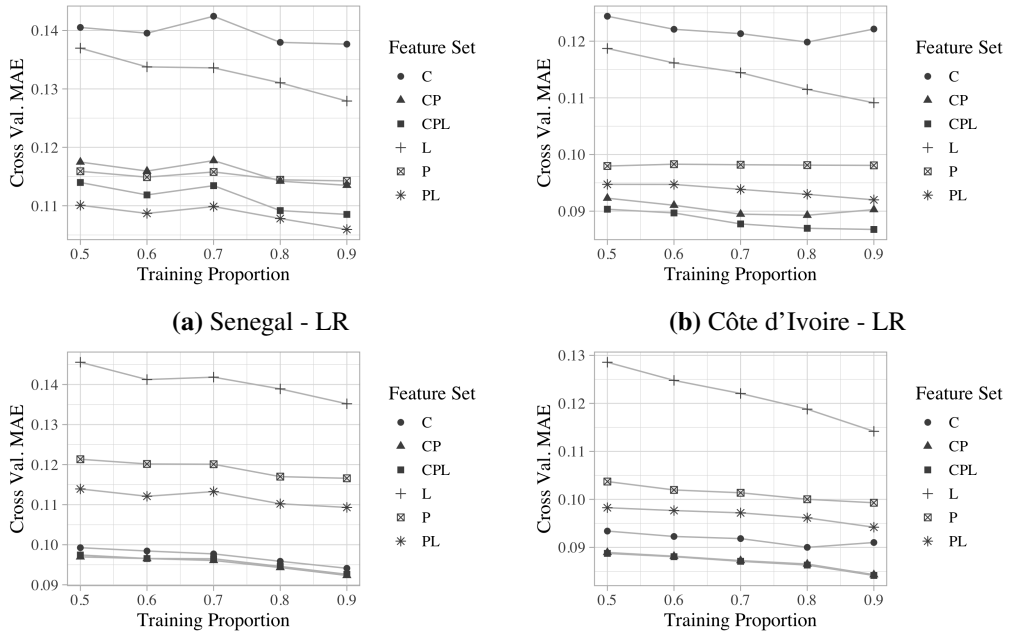
Figure 5.7 shows that for LR models, using C features performs worse than the baseline P, L and PL feature sets, and in Senegal even CPL under performs compared to PL with LR model. But for RF models, CPL is consistently the best performer and even C on its own is comparable and consistently out performs the best baseline feature sets, PL. The inclusion of C features results in poorer predictive accuracy than the baselines when using the LR model, but greater accuracy when used with the RF model, which demonstrates the importance of considering non linear modelling techniques.



(a) Senegal - LR **(b)** Côte d'Ivoire - LR

(c) Senegal - RF **(d)** Côte d'Ivoire - RF

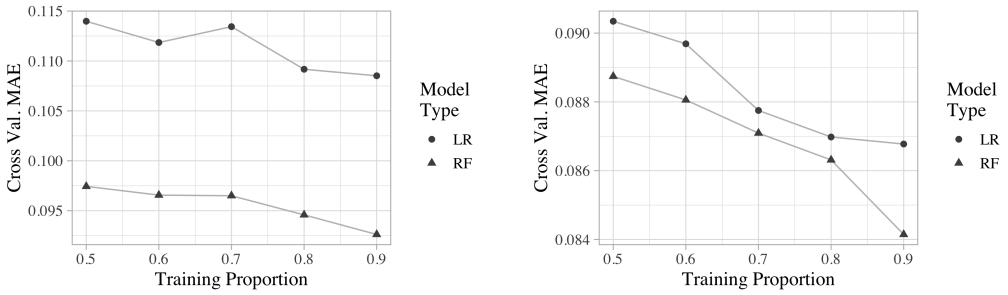
Figure 5.6: Comparison of mean cross-validation errors for different residual types across varying training set proportions and model types. Models trained with gravity-split residuals tend to perform best over all training set proportions.



(a) Senegal - LR **(b)** Côte d'Ivoire - LR

(c) Senegal - RF **(d)** Côte d'Ivoire - RF

Figure 5.7: Comparison of mean cross-validation errors for different feature sets and across varying training set proportions and model type. Models trained with all features tend to perform best over all training set proportions.



(a) Senegal **(b)** Côte d'Ivoire

Figure 5.8: Comparison of mean cross-validation errors for different model types across varying training set proportions. RF model type tends to perform best over all training set proportions.

Focusing on CPL as the best performing feature set, Figure 5.8 shows that the RF model is generally more accurate than the LR over all training proportions. In Senegal, the performance improvement of the RF model over LR is more marked, at approximately 12% on average, and for Côte d'Ivoire the improvement is approximately 2.5% on average.

The headline result here is that the models that include CDR features as predictors outperform those that do not. To quantify this, we look in more detail at the magnitude of gains provided by CDR features. Looking first at Senegal, the MAE of the RF model with CP features is 20.08% lower than with the RF model with P when estimating average wealth with 50% training data, and reaches a 20.80% improvement with 90% training data. We see a similar improvement brought in by spatial lag: comparing RF models with PL and CPL features, we find the addition of CDR features offers a reduction in MAE of between 13.86% and 15.27%.

We see a similar story in Côte d'Ivoire. When predicting average wealth, the RF model with CP features provides an improvement of between 13.49% and 15.09% in MAE over the P baseline. When adding lag, CPL features offers an improvement over PL features of between 9.68% and 10.67%.

5.4 Discussion

We can frame the comparison of results from models trained with the alternative feature sets in terms of concrete situations in which different kinds of data are avail-

able. That is, the general model case and the retrained model case, as described in Chapter 3. The two situations differ in the availability of ground truth poverty or socio-economic status data with which we can create a spatially lagged variable.

In the general model case, where there is no ground truth data available, we can compare baselines that use population density only as an input with models that use CDR features only, and that combine both CDR features and population density. Focusing on the random forest models, which perform comparatively better in the majority of cases, we can see that in both Senegal and Côte d'Ivoire models with CDR features only out performs those with population density only, and the combined model offers a further marginal improvement. These results suggest that organisations wishing to produce the best possible estimates of wealth and poverty in a region for which there is no available socio-economic data could use a pre-trained CDR based model as a reasonable proxy.

In the retrained model case, where there is some ground truth data from which to compute the spatially lagged variable, we can then compare the lag only baseline model, the population density and lag baseline model, the CDR features only model, and the combined model with CDR features, population density and spatial lag. Baseline models with just the lagged variable as input perform worst out of all models, which is somewhat surprising given the high level of spatial autocorrelation present in the data. However, as expected, the performance gap narrows as more training data becomes available. When population density is included, accuracy improves but is still surpassed by models with CDR data only. As before, models combining CDR data with baseline features have the highest performance. These results suggest that organisations wishing to produce the best possible estimates of wealth and poverty in a region for which there is some existing socio-economic data available (but for which it is infeasible to collect more) could use our approach to train a CDR based model to provide a reasonable proxy.

By means of a comparative performance analysis using data from two developing countries (namely, Senegal and Côte d'Ivoire), we have shown that it is possible to predict wealth using models trained on features derived from CDR data, and that

these models offer significantly improved predictive performance when compared to baseline models that do not utilise CDR features. Thus we can be optimistic about the value that our approach can offer to governments and organisations that lack the means to perform a more comprehensive survey of a country's socio-economic status. We have also found that results vary across the two studied countries, with the added value of CDR features being smaller in Côte d'Ivoire compared to Senegal. Continuous testing and refinement ought to be an integral part of any implementation of the methods described in this chapter.

Chapter 6

Conclusions

In this final chapter we first recap and evaluate the main contributions of this thesis (Section 6.1) before discussing their wider impact (Section 6.2). We then outline the main limitations of this work and propose directions for future work (Section 6.3).

6.1 Overall Evaluation of Contributions

Here we recap the main contributions of this thesis and evaluate their significance in light of the work discussed in preceding chapters.

Contribution 1. In Chapter 3 we presented an analysis of the spatial dependency of wealth and its relationship with population density in Senegal and Côte d'Ivoire. We showed that two baseline wealth estimators, each grounded on concrete usage scenarios, can be derived from population density data and sparse poverty data respectively, and moreover, that these baselines provide significantly more predictive power than random baselines. These results highlight a serious limitation of much research in the area, which has thus far failed to compare results of models using CDR features with realistic baselines and therefore failed to properly show the value of such an approach. In doing so this contribution can help ensure governments and NGOs are informed of potential limitations when deciding whether or not to pursue a data driven approach to poverty estimation.

Contribution 2. In Chapter 4 we studied the output of information diffusion simulation models over the BTS network in each country and found that an area's susceptibility, or ability to receive information is strongly correlated with its level

of wealth. We further analysed these results in relation to a number of contextual factors and saw that data quality issues have a significant impact on the strength of correlation between poverty and susceptibility. This has an important implication for this thesis and future work, not least for data providers (both telecoms and socioeconomic) to ensure that more consistent and representative data is made available to researchers. Were these data quality issues resolved we might expect the predictive power of CDR based models to significantly improve.

Contribution 3. In Chapter 5 we presented a number of original hypothesis-based features of aggregated CDRs that can be used as inputs to a poverty prediction model. These include static and simulation based measures of information access, activity based metrics and econometric inspired features. We looked at the correlation of these features with wealth in each country, and at their cross-correlation, in order to downselect a smaller number of important features. These features can be reproduced and used by other researchers in future work. We also built and tested a number of models that incorporate the CDR based features to estimate wealth. We further performed a detailed analysis of the results of this model in relation to the baseline predictors in order to establish their real added value over simpler approaches. In summary, we found that the most successful approach in terms of model accuracy involves training a Random Forest model on a feature set that includes population density, spatially lagged wealth (if available), 9-parameter gravity model residuals, as well as activity and network based features. Further work is required to refine this approach and test in wider variety of circumstances, however, this modelling approach has the potential to be used in practice to provide poverty or wealth estimates in a timely fashion and at fine level of spatial granularity in countries that lack comprehensive survey data.

6.2 Who cares?

This work has the potential to impact the practices of policymakers and NGOs working to improve the living standards of people in countries that lack the resources to manually collect socio-economic data on a frequent basis and at sample

rates that would allow fine spatial disaggregation. Tools built upon these results would be relatively low cost to implement and could provide interpretable results (in contrast to a black-box machine learning approach) to act upon in a timely manner. Furthermore, we enable disaggregation at multiple levels of spatial granularity thus potentially influencing policy implemented at different levels, from neighbourhood to region. Since we use only aggregated CDR data, mobile phone users also benefit by having their privacy protected from the outset, thus removing a barrier to wider adoption of this approach.

In discussion with the United Nations Population Fund (UNFPA) to determine how to put the methodology to actual use, an important need identified is the availability of maps at different levels of spatial granularity, so to provide information as required for different purposes. For example, national governments determining the allocation of a development budget to regional governments would require coarser grained information at the level of the administrative division in question. At the other end of the scale, regeneration or aid projects implemented at the local level for the benefit of small communities would require much finer resolution poverty maps to ensure the most needy areas are targeted. The methodology we have presented provides for both situations, with the ability to aggregate data at multiple levels of granularity, unlike sparsely sampled survey data that must be aggregated to a certain minimum (and often impractically coarse) level in order to achieve statistical significance.

Indeed, tools built upon the methods we have described would be a useful augmentation to socio-economic data collection processes in any country. The cost of producing estimates from passively and automatically collected communication data is negligible compared to that of manual surveying, thus a main barrier to obtaining up to date poverty estimates has been removed. Côte d'Ivoire and Senegal are good examples of countries in which timely and accurate information regarding poverty is severely lacking. In cases such as these, the ability to obtain estimates of poverty levels on a continuous basis would represent a vast improvement. UNFPA has stressed the value that *any* indicative estimates would provide in certain situa-

tions where none are currently available; even if they carried with them a significant level of uncertainty such estimates would still represent a large improvement in many cases. Indeed, novel methods to provide low cost poverty indicators would represent significant value to many governments and NGOSs working to improve people's lives. Limited resources could be allocated in much more efficient manner thereby helping to alleviate some of the detrimental effects of poverty and inequality.

6.3 Limitations and Future Work

By replicating results in two developing economies we have shown that our results are not simply chance correlations. This represents a significant advance towards general application compared to related work. However, difficulty in obtaining CDR data currently prevents us from establishing the global applicability of our work. More examples are required in order to further verify the approach and being able to produce estimates in one country from a model trained in another would also represent a significant advance.

It might be suggested that validity is threatened by variation in adoption rates, but rather, we argue that this is an important factor that will be reflected in the features we derive (others being individual usage, infrastructure, etc.). Consequently, use of our features would not disadvantage groups with low adoption rates, but in fact they would show up as black spots in our models. Most intuitively when measuring activity, that is, low adoption will mean low activity, but also in the other features. For example, an area with lower adoption rates may exhibit higher interaction model residuals, reflecting the fact that it may in turn have relatively lower levels of communication with other areas and would thus be identified much sooner than with traditional methods. To test this argument, we need to combine fine grained adoption rate data with our approach in order to determine the real impact of variations. Conversely, we may expect that this particular signal will weaken as adoption rates increase over time. As mobile phone ownership become more ubiquitous and the cost of usage lowers, we may find that simple measures such as number of

calls made will no longer reliably track wealth. Instead, we may need to focus on more complex features such as gravity residuals and centrality measures that aim to capture signals related to economic and network advantage. Since these types of features reflect the interaction between areas we might expect them to become more robust as adoption rates increase, since, in turn, mobile communication as a medium will become more representative of the relationships between areas.

Despite being the main motivation for this work, the lack of up to date and spatially accurate socio-economic ground truth data also represents a significant hurdle toward a rigorous evaluation of the results. In order to be confident that the features we extract can be used to accurately track poverty in a timely and spatially accurate manner, we initially require knowledge of real poverty rates that also fulfil these constraints. Instead we have a lag of 4 years in Côte d'Ivoire and 3 years in Senegal between the DHS data we use as ground truth and the mobile phone data from which we derive our features. Although this temporal lag will undoubtedly affect the accuracy of predictive models that use CDR data, we would expect their accuracy and utility to increase were this lag removed. Furthermore, as discussed in Section 3.1, aggregated wealth indices may not reveal the existence of extreme poverty if it is in close proximity to wealthier households, and therefore using such an index as a target variable for training models will also limit the ability of those models to identify such cases. Future work should take steps towards overcoming these limitations by acquiring ground truth data that is both more recent and has a more precise level of geo-location. This will allow us to fully investigate the relationship between geographical hierarchies and validate estimates at finer granularity.

CDR data is collected on a continuous basis as users make and receive calls. As such a potentially fruitful extension of this work would be to produce continuous (or regularly updated) estimates, and from these derive forecasts and identify trends, thereby providing early warning of conditions worsening in specific areas. This would could also provide the ability to evaluate the effect of policy and projects in a reasonable time frame (i.e., as soon as changes occur and before policy is due for renewal) by monitoring changes in CDR derived poverty estimates. Relatedly,

more data would allow us to experiment with more sophisticated machine learning models, such as artificial neural networks, which could provide greater accuracy. Recent advances in methods for explaining the predictions of black box models such as SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) could also be investigated and would allow us to maintain the emphasis on interpretability, which is essential for policy making. Finally, the methods presented in this thesis are not specific to wealth or poverty prediction and indeed could be adapted to provide predictions for other similarly sparse socio-economic or health factors.

Bibliography

Aker, J. C. and Mbiti, I. M. (2010). Mobile phones and economic development in africa. *Journal of economic Perspectives*, 24(3):207–32.

Al Hasan, M. and Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics*, pages 243–275. Springer.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *nature*, 406(6794):378–382.

Alkire, S., Chatterjee, M., Conconi, A., Seth, S., and Vaz, A. (2014). Poverty in rural and urban areas: Direct comparisons using the global mpi 2014.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical Analysis*, 27(2):93–115.

Arbesman, S., Kleinberg, J. M., and Strogatz, S. H. (2009). Superlinear scaling for innovation in cities. *Physical Review E*, 79(1):016115.

Bailey, N. T. et al. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

Bank, T. W. (2015). Using big data for the sustainable development goals. Technical report.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–52.

Bettencourt, L. M. (2013). The origins of scaling in cities. *science*, 340(6139):1438–1441.

Bettencourt, L. M., Lobo, J., Strumsky, D., and West, G. B. (2010). Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one*, 5(11):e13541.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., and West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306.

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

Blumenstock, J., Shen, Y., and Eagle, N. (2010). A method for estimating the relationship between phone use and wealth. In *QualMeetsQuant Workshop at the 4th International IEEE/ACM Conference on Information and Communication Technologies and Development*.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Bruckschen, F., Schmid, T., and Zbiranski, T. (2015). Cookbook for a socio-demographic basket. In *D4D Challenge Senegal Sessions Scientific Papers*, Netmob '15.

Burt, R. (1992). *Structural Holes*. Harvard University Press.

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard university press.

Callaway, D. S., Newman, M. E., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468.

Coleman, J., Menzel, H., and Katz, E. (1959). Social processes in physicians' adoption of a new drug. *Journal of Chronic Diseases*, 9(1):1–19.

Council, N. R. et al. (1995). *Measuring poverty: A new approach*. National Academies Press.

Daley, D. and Kendall, D. G. (1965). Stochastic rumours. *IMA Journal of Applied Mathematics*, 1(1):42–55.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *The Review of Economics and Statistics*, 87(1):1–19.

Deutschmann, P. J. and Danielson, W. A. (1960). Diffusion of knowledge of the major news story. *Journalism & Mass Communication Quarterly*, 37(3):345–355.

DHS, M. (2012). Survey organization manual for demographic and health surveys. Technical report.

Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.

Doll, C. N. H., Muller, J.-P., and Elvidge, C. D. (2000). Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. *AMBIO: a Journal of the Human Environment*, 29(3):157–162.

Eagle, N., Macy, M., and Claxton, R. (2010). Supporting online material for: Network diversity and economic development. *Science (New York, N.Y.)*, 328(5981):1029–31.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., and Davis, E. R. (1997). Mapping city lights with nighttime data from the dmsp operational linescan system. *Photogrammetric Engineering and Remote Sensing*, 63(6):727–734.

Elvidge, C. D., Imhoff, M. L., Baugh, K. E., Hobson, V. R., Nelson, I., Safran, J., Dietz, J. B., and Tuttle, B. T. (2001). Night-time lights of the world: 1994–1995. *ISPRS Journal of Photogrammetry and Remote Sensing*, 56(2):81–99.

Frias-Martinez, V., Soguero-Ruiz, C., Frias-Martinez, E., and Josephidou, M. (2013). Forecasting Socioeconomic Trends With Cell Phone Records. In *Proceedings of the 3rd ACM Symposium on Computing for Development (DEV'13)*, New York, New York, USA. ACM Press.

Frias-martinez, V., Soto, V., Virseda, J., and Frias-martinez, E. (2012). Computing Cost-Effective Census Maps From Cell Phone Traces. In *Pervasive Urban Applications (PURBA)*, Newcastle.

Frias-Martinez, V. and Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage. In *Fifth International Conference on Information and Communication Technologies and Development (ICTD '12)*.

Frias-Martinez, V., Virseda-Jerez, J., and Frias-Martinez, E. (2012). On the relation between socio-economic status and physical mobility. *Information Technology for Development*, 18(2):91–106.

Funkhouser, G. R. and McCombs, M. E. (1972). Predicting the diffusion of information to mass audiences†. *Journal of Mathematical Sociology*, 2(1):121–130.

Gary S. Becker, Edward L. Glaeser, K. M. M. (1999). Population and economic growth. *The American Economic Review*, 89(2):145–149.

Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223.

Granovetter, M. (2005). The impact of social structure on economic outcomes. *Journal of economic perspectives*, pages 33–50.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):pp. 1360–1380.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica, Journal of the Econometric Society*, pages 501–522.

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM.

Gutierrez, T., Krings, G., and Blondel, V. D. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*.

Haughton, J. and Khandker, S. (2009). *Handbook on Poverty and Inequality*. World Bank e-Library. World Bank.

Hristova, D., Rutherford, A., Anson, J., Luengo-Oroz, M., and Mascolo, C. (2016). The international postal network and other global flows as proxies for national wellbeing. *PLOS ONE*, 11(6):1–19.

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

Jung, W. and Wang, F. (2008). Gravity model in the Korean highway. *Europhysics Letters*, 81.

Kaluza, P., Kölzsch, A., Gastner, M. T., and Blasius, B. (2010). The complex network of global cargo ship movements. *Journal of the Royal Society, Interface / the Royal Society*, 7(48):1093–103.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

Kramer, A. D. I. (2010). An Unobtrusive Behavioral Model of "Gross National Happiness". In *Proceedings of the 28th ACM CHI*, pages 287–290.

Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003.

Langville, A. N. and Meyer, C. D. (2005). A survey of eigenvector methods for web information retrieval. *SIAM review*, 47(1):135–161.

Lathia, N., Quercia, D., and Crowcroft, J. (2012). The Hidden Image of the City : Sensing Community Well-Being from Urban Mobility. In *Pervasive 2012*, pages 91–98, Newcastle.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Mao, H., Shuai1, X., Ahn, Y.-Y., and Bollen, J. (2013). Mobile communications reveal the regional economy in cˆote d'ivoire. In *Proceedings of the Orange D4D Challenge*.

McClellan, N., Shron, M., Duckworth, D., Georges, M., Mentch, A., Nguyen, T., Lee, M., Korte, T., and Kao, S. (2013). Predicting small-scale poverty measures from night illumination. `https://hackpad.com/`

Predicting-Small-Scale-Poverty-Measures-from-Night-Illumination-f
Accessed: 2013-04-12.

Miritello, G., Lara, R., Cebrian, M., and Moro, E. (2013). Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3.

Monin, J., Benayoun, R., and Sert, B. (1976). *Initiation to the Mathematics of the Processes of Diffusion, Contagion and Propagation*, volume 4. Walter de Gruyter.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E., Forrest, S., and Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101.

Noor, A. M., Alegana, V. A., Gething, P. W., Tatem, A. J., and Snow, R. W. (2008). Using remotely sensed night-time light as a proxy for poverty in africa. *Popul Health Metrics*, 6(1):5.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., and Pentland, A. (2013). Urban characteristics attributable to density-driven tie formation. *Nature communications*, 4:1961.

Pavot, W., Diener, E., Colvin, C. R., and Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality assessment*, 57(1):149–161.

Pokhriyal, N. and Dong, W. (2015). Virtual networks and poverty analysis in senegal. In *D4D Challenge Senegal Sessions Scientific Papers*, Netmob '15.

Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. (2012a). Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 965–968. ACM.

Quercia, D. and Saez, D. (2014). Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *Pervasive Computing, IEEE*, 13(2):30–36.

Quercia, D., Séaghdha, D. Ò., and Crowcroft, J. (2012b). Talk of the city: Our tweets, our community happiness. In *The 6th international AAAI Conference on weblogs and social media, Dublin*.

Rapoport, A. and Yuan, Y. (1989). Some aspects of epidemics and social nets. *The Small World*, pages 327–348.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ryan, B. and Gross, N. C. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural sociology*, 8(1):15.

Sachs, J. D. and Warner, A. M. (1997). Sources of slow growth in african economies. *Journal of African Economies*, 6(3):335–376.

Sen, A. (1999). *Commodities and Capabilities*. Number 9780195650389 in OUP Catalogue. Oxford University Press.

Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, pages 8–12.

Smith, C., Quercia, D., and Capra, L. (2013). Finger on the pulse: identifying deprivation using transit flow analysis. In *The 16th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pages 683–692. ACM.

Smith-Clarke, C., Mashhadi, A., and Capra, L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 511–520, New York, NY, USA. ACM.

Soto, V., Frias-Martinez, V., Virseda, J., and Frias-Martinez, E. (2011). Prediction of socioeconomic levels using cell phone records. In Konstan, J., Conejo, R., Marzo, J., and Oliver, N., editors, *User Modeling, Adaption and Personalization*, volume 6787 of *Lecture Notes in Computer Science*, pages 377–388. Springer Berlin Heidelberg.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690.

United Nations, Department of Economic and Social Affairs, P. D. (2012). World Urbanization Prospects, The 2011 Revision: Highlights. Technical report.

United Nations Human Settlement Program (2008). *State of the World's Cities 2008/2009: Harmonious Cities*. EarthScan.

Valente, T. W. (1995). *Network models of the diffusion of innovations*. Number 303.484 V3.

Venerandi, A., Quattrone, G., Capra, L., Quercia, D., and Saez-Trumper, D. (2015). Measuring urban deprivation from user generated content. In *The 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

Viboud, C., Bjø rnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science (New York, N.Y.)*, 312(5772):447–51.

Wang, N., Kosinski, M., Stillwell, D., and Rust, J. (2012). Can well-being be

measured using facebook status updates? validation of facebook's gross national happiness index. *Social Indicators Research*, pages 1–9.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442.

Wu, F., Huberman, B. A., Adamic, L. A., and Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1):327–335.

Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., and Wang, W.-X. (2014). Universal predictability of mobility patterns in cities. *Journal of The Royal Society Interface*, 11(100):20140834.

Zipf, G. (1946). The P 1 P 2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):677–686.