Detection of Vulnerable Communities in East Africa via Novel Data Streams and Dynamic Stochastic Block Models

Submitted March 2021, in partial fulfillment of
the conditions for the award of the degree **CDT Horizon PhD.**

Madeleine Ellis
14283670

Supervised by James Goulding, Simon Preston

School of Business
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature ____________________

Date ______ / ______ / ______

Abstract

In developing countries it is challenging to collect data on poverty and its associated community health characteristics. Data collection in this context is impractically laborious and resource greedy. Additionally due to the sensitive nature of these themes the data is often unreliable. There is a need for alternative methods of detection of vulnerable communities. However, promising opportunities arise via novel rich data streams such as Call Data Records stemming from the ubiquitous use of mobile phones. Despite the growth of Call Data Record data there has been limited previous application to problems of poverty and development. This thesis makes three main contributions: (i) Methods of collecting ground truth data in Developing areas; (ii) Best practices in application to detect vulnerable regions; (iii) Development of new applications of statistical approaches to the problem via the stochastic block model. This work is focused on Dar es Salaam in Tanzania. Having more reliable and easily accessible truths on these vulnerabilities can have a high potential impact for policy makers and NGOs trying to make positive changes to reduce devastating effects of poverty. This thesis produces comprehensive results to amend the current knowledge gaps, via rigorous fine-grained data collection processes surveying the 452 subwards in Dar es Salaam in relation to poverty and social vulnerability.

Acknowledgements

I would like to thank my supervisors James Goulding and Simon Preston for their support and guidance. James has introduced me to a field of work which inspires, I am excited to continue on this path. Simon joined my PhD team part way through my research project and has been available for all my silly questions since then. I am extremely grateful for the kindness they have both shown me.

I would also like to thank Horizon, Centre for Doctoral Training for providing an opportunity to study a multidisciplinary research project. All the staff in CDT Horizon and N-LAB have made the last four years an absolute pleasure. In particular I would like to thank Gavin Smith and Bertrand Perrat for their support and patience setting me up with all the databases.

Thank you to all the wonderful people I work with. Georgie, thank you for being the first to welcome me into the lab with a cup of tea and all your kindness since then. Gregor thank you for being a wonderful travel buddy. Vanja thank you for making every work day exciting. Rosa thank you for breathing heavily and sharing your snacks. Katie and Rowland thank you for being a constant source of laughter. KG, Joe, Roza and Abi thank you for filling my CDT year with lovely memories. There are too many wonderful people to name everyone, but I am grateful for everyone who has shaped these last four years, I am lucky to have made friends for life through this process. I would also like to thank Nicola Pitchford for her ongoing support and for being a constant source of inspiration during the final two years of my thesis.

Finally I would like to thank my family Mum, Dad, Matt, Chris, Miriam, Penny, Hazel, Slimby, Cobweb, Fiesty and Ronnie you are my everything and I will always be grateful for your support and love.

Professional Partner

Humanitarian Open StreetMap Team are an organisation who provide up-to-date maps to relief organisations responding to natural disasters of political crises. This idea of supporting aid organisations in more efficient responses to the demands of vulnerable people falls perfectly in line with the work in this thesis. I would like to thank Humanitarian Open StreetMap Team for their support with this thesis and for being a constant source of inspirations for my future career path.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The collection of geo-demographic data for understanding the spread of poverty and associated characteristics has typically been undertaken by census and other large scale surveys. Providing the basis for a wide range of activity including city planning, market intelligence and policy, the information is routinely collected by both governments and market intelligence companies within developed countries at significant cost. In less developed countries, this cost means that information is typically collected less frequently and at a lower fidelity if at all. Looking to address this situation, this thesis looks at alternative methods for the collection, derivation and analysis of geo-demographic data. The case-study applications of this work focus on the Tanzanian port city of Dar es Salaam, the former administrative capital and current largest city in East Africa. It is among the fastest-growing cities globally, having doubled in size since 2005 while serving as a significant economic hub and gateway for freight to neighbouring landlocked countries. Owing to rapid urbanisation, 70% of its approximately 6.7 million inhabitants live in vast, sprawling and informal slums (Lands and Ministry, 2000).

1.1 Context

This section presents key details of the context and motivations of this thesis. The deficiencies in reliable ground truth data which are limiting the work of policy makers are described. This section also provides an overview of the case study area Dar es Salaam.

1.1.1 The Problem of Poverty Detection Background

The latest comprehensive data on global poverty in 2013 showed that there were 767 million people estimated to be living below the poverty line (Cuesta et al., 2016). Despite the number of people in poverty falling globally between 2012 and 2013, poverty in Africa is still widespread, and continues to be high relative to all other regions of the world (Cuesta et al., 2016). Communities living in such poverty face a number of societal challenges such as: unemployment, forced labour, lack of safety, health problems and reduced access to education. Without question, poverty and its associated challenges still threaten the lives and well-being of an unacceptably proportion of our population.

The data indicating societal challenges and poverty levels in Africa has been historically extremely sparse, in 1990 only 20 countries in the continent even had data providing measurements of poverty (Channing Arndt, 2016). Household surveys initially provided some insight into wealth distribution, however, these surveys omit a significant proportion of the poorest people, making it an insufficient indicator for poverty (Carr-Hill, 2017). Since then DHS (Demographic and Health Surveys), income and expenditure surveys have been introduced. While this has drastically improved the data situation in these developing countries, there are still massive deficiencies in this data.

Such surveys are often too infrequent, take too long to implement to have much value, and are rapidly out of date (Perez et al., 2016; Espey, 2019).

Surveyed data simply put, is hard to obtain. Being both labour and cost intensive it is therefore scarce (Xie et al., 2015a). The deficiency in reliable data explaining local poverty and societal challenges in developing countries restrains the impact of local policy makers, governments and aid organizations (Lupu and Michelitch, 2018; Perez et al., 2016; Xie et al., 2015a; United Republic of Tanzania, 2017; Evans and Ruane, 2019). This is a significant problem as accurate estimates of population characteristics, such as poverty, remain critical to development, and attainment of UN SDG's (Blumenstock et al., 2015). There has also been serious concerns for the reliability of quantitative data in developing countries for researchers, national statistics on economic production, for example may be off by as much as 50% in Africa (Jerven, 2013). There are potential solutions here, in the form of novel data sources. Satellite data provides a more time efficient approach to investigating the poverty of different areas than traditional surveying (Watmough et al., 2016). High-resolution satellite imagery, is now increasingly inexpensive and reliable (Xie et al., 2015a), and the increase in satellite data availability has contributed to the study of geo-spatial information with broad applications across many areas including the distribution of poverty (Perez et al., 2016; Jean et al., 2016a).

Poverty stricken regions, however, are also the ones which are more likely to have less internal funding, more civil-wars, poor infrastructure and inadequate government resourcing available for comprehensive research such as surveys and satellite data. Hence there are still vast gaps in the collection of reliable data which could be used to describe poverty (Perez et al., 2016). There are, however, increasingly new sources of collecting data on individuals such as mobile phone and internet records which are enabling new approaches to demographic profiling and opening exciting new fields of potential analysis (King, 2011). Data from a communication network of mobile phones and business landlines for example were used to show that communication diversity is a strong indicator for the economic health of communities in the UK (Eagle et al., 2010). A move towards this sort of big data is a vital step in tackling issues of development (Espey, 2019; Holloway and Mengersen, 2018; Vinuesa et al., 2020)

In developing countries there are admittedly fewer sources of big data. However, mobile phone use has become increasingly ubiquitous in these regions, due to a lack of existing landline infrastructure. This is providing a fruitful source of data for researchers (Blumenstock et al., 2015). In regions where resources such as time, labour and money are scarce for such research, this approach creates a method for gathering information on individuals at a fraction of the cost of traditional methods such as surveys and satellite images (Blumenstock et al., 2015). The diversity of individuals relationships is a key indicator of social and economic life, until recently this was not so widely quantifiable. There is now in reach data on networks of people and their behaviours, which is already allowing new insights at population levels (Eagle et al., 2010). Such data streams are predominantly being used to address issues of sustainability and development. Satellite imagery has already been used to detect sites of vulnerability to slavery including: Brick kilns (where debt bondage is common) (Foody et al., 2019) Mobile data for example has been mapped to predict wealth through Rwanda from individual phone subscriber wealth (Blumenstock et al., 2015); a further example is the use of approximately 10 million mobile phone subscribers from multiple operators in Sri Lanka to assess the land use of the regions (Madhawa et al., 2015). Additionally mobile phone data in Tanzania has been used to investigate the spread of poverty (Engelmann et al., 2018).

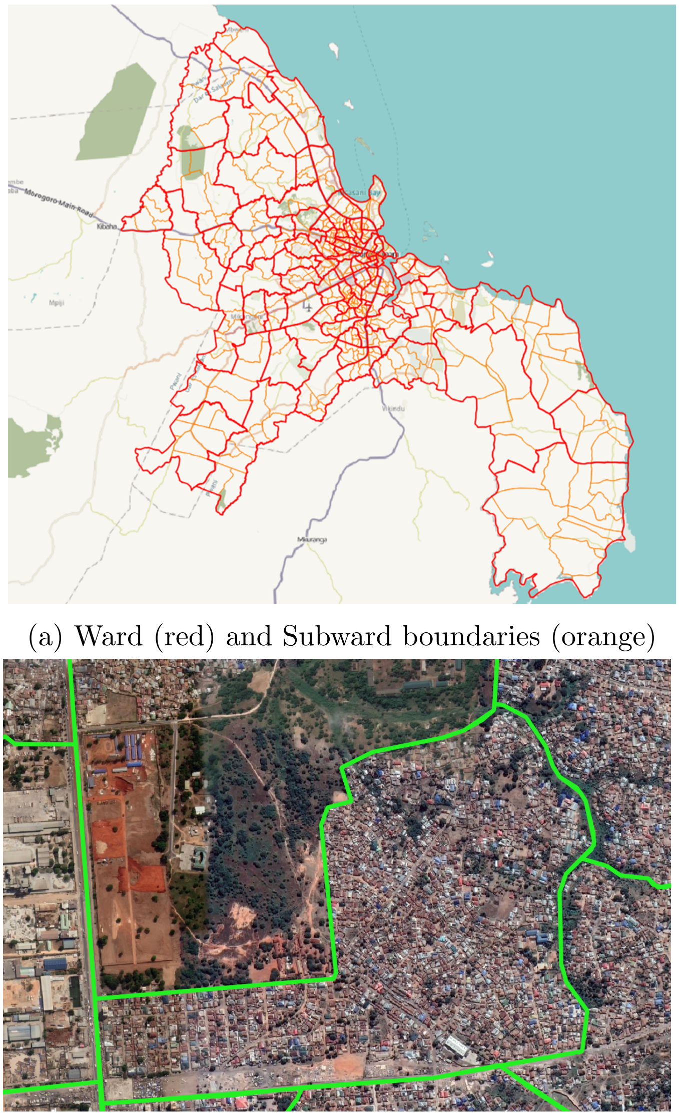
In particular, the ubiquitous use of mobile phones in developing countries generates data sets with huge potential to help organizations who are currently struggling to identify the most vulnerable parts of regions. However, as with all new forms of data, much work remains to be done in finding different ways in its analysis, if we are to obtain useful results (Smith-Clarke and Capra, 2016). Accurate and reliable ground truth information on societal challenges and poverty are essential for policy makers and NGOs to make decisions about resource allocation (Fields, 1989). Take for example flooding, a phenomenon the city of Dar es Salaam is particularly vulnerable. It is well established that poorer households are more likely to be affected by such floods both in terms of direct damage to assets and property and also indirect damage such as loss of labour, education, infrastructure and health (Erman et al., 2019). It is crucial if we are to minimize negative impact, that communities living in such poverty are detected, in order for flood prevention resources to be allocated appropriately. Yet in cities such as Dar es Salaam, which has doubled in size over the last decade, and is comprised of 70% informal settlement, such demographics are not even mapped.

Accurate detection of vulnerable communities has high potential impact. This thesis is looking to contribute directly to this challenge in two ways, first providing high definition new processes to obtain ground truth information on Dar es Salaam (via a comprehensive survey protocol and implementation), but secondly by investigating the new statistical methods of analysing CDR data to detect vulnerable communities in automated low cost and tractable fashion.

1.1.2 Dar es Salaam

The test-bed for this research focuses on the Tanzanian port city of Dar es Salaam, the former administrative capital and current largest city in Tanzania. It is amongst the fastest-growing cities globally, having doubled in size since 2005 while serving as a significant economic hub and gateway for freight to neighbouring landlocked countries. With a population of over 6 million people, it is the largest city in East Africa (Bureau, 2019). The regions of focus of this work are subwards called *Mtaa*, which are the smallest administrative regions of the city (as illustrated in Figure 1.1). Several subwards make up individual wards, with wards themselves belonging to one of the city's three districts - Kinondoni to the north, Ilala in the centre and Temeke to the south. There are 90 wards in Dar es Salaam and 452 subwards, these can be seen in Figure 1.1a. Figure 1.1b illustrates the high diversity in land-use within wards, the image highlights common patterns of diversity across subwards within each ward. This figure shows two internal subwards of the ward Mbagala Kuu, one mostly vegetation and the other predominantly residential slum area. This diversity within wards renders subwards the only credible resolution to analyse risk.

The city has an extremely high population density, thriving activity, overburdened administrative and law enforcement infrastructure, and co-existence of extreme wealth and poverty. For these reasons Dar es Salaam reflects an ideal representative case study for this work, with an opportunity to create high impact, fine grained information on the city which could be utilised by policy makers and NGOs, via advancements of statistical methods.



(a) Ward (red) and Subward boundaries (orange)

(b) Ward Mbagala Kuu, Diversity of Internal Subwards

Figure 1.1: Dar es Salaam

1.2 Research Questions

Can new methods use increasingly ubiquitous and rich novel data streams drawn from the private sector (such as Call Data Records) to support fine grained detection of vulnerable communities in the poorest OECD DAC listed nations?

10.7% of people around the world are living under the poverty line (Bönke et al., 2016). Communities around the world are living in unacceptable conditions, facing a number of different vulnerabilities. Identifying vulnerable communities is essential for governments and aid organisations to make interventions and actions to support such groups. Traditional methods of discovering community demographics such as household surveys are time and resource greedy (Xie et al., 2015b). This makes them difficult to acquire in developing regions. New sources of data streams such as CDR (Call Data Records) are becoming increasingly available and rich source of data in developing countries thanks to the ubiquitous use of mobile phones (Joshua Blumenstock, 2015). Such data has been looked at comprehensively to find communities in the developed world. This thesis examines the less widely used application of such data in developing regions of the world.

Currently there data deficiencies and extensive challenges in the collection of reliable demographic information on the community health of developing regions. Can current data collection methodologies be adapted to create efficient, reliable and cost effective techniques.

There is a lack of observed demographic data on community health (ground truth data) in developing countries (Channing Arndt, 2016). Trying to collect data on communities in the poorest OECD DAC listed nations has a number of problems from weather conditions to limited resources, local government restrictions and everything in between. What adaptions can be made to traditional surveillance methods, and what new methodologies can be implemented to collect large scale and reliable ground truth data?

Can state-of-the-art statistical methods in network analysis be brought to bear on pressing UN SDG problems such as eliminating poverty and community vulnerability in the face of sparse ground-truth datasets?

Despite this increase in data streams such as CDR (Call Data Records) in developing countries, little has been done to examine population interactions, nor apply network analysis to identify the most vulnerable regions. Take for example the stochastic block model, a field of network analysis used to identify communities from the nodes and edges of an observed network. This has successfully been applied to a number of different fields such as epidemiology, political science and social psychology. The stochastic block model for example, has shown great promise for the identification of proteins for biological research (Airoldi et al., 2006b) and been used to link political blogs in the US 2004 elections to liberal or conservative groups (Emmanuel Abbe, 2015). Extending these successes this work aims to test the stochastic block model's ability to model data streams from developing regions, to provide insights into the identification of vulnerable communities. This provides a novel application of such methods for potential social interventions.

In the face of the mass-urbanization and rapid demographic change facing many developing regions, how can existing network analysis models be extended to integrate dynamic data sources?

Data streams such as CDR (Call Data Records) are accompanied with timestamps that can provide dynamic representations of how events change over time. In areas where demographics are rapidly evolving, the use of models which account for such temporal variation could be key in eliciting fruitful insights for policy makers. The stochastic block models are increasingly being considered in dynamic environments. One proposed adaption is the idea of label stitching, combining the static stochastic block model with independent Markov chains for the evolution of node groups through time (Lei Tang, 2012). The goal here is to detect groups based on stable within group interactions over time. This thesis extends this by amending the stochastic block model with Non-homogeneous Poisson Process so that network edges are allowed to change over time while the block structure remains constant. This emerging area of research around dynamic network analysis presents new opportunities to more accurately identify vulnerable communities in developing areas.

1.3 Contributions of the Thesis

Given the above research questions the contributions to knowledge of this thesis are as follows:

- Development of novel survey techniques, to obtain high fidelity ground truths to assess the problem of poverty in the developing world.
- A primary contribution via the knowledge gained from a comprehensive fine-grained data collection process surveying the 452 subwards in Dar es Salaam in relation to poverty and social vulnerability.
- Assessment of the challenges of data collecting in the developing setting are validated and substantiated via semi structured interviews.

- Demonstration of the detection of communities vulnerable to poverty via the stochastic block modelling of CDR data.
- First evidence of the importance of considering dynamic data in the detection of communities vulnerable to poverty.
- Technical adaption of the stochastic block model to infer communities from degree corrected, directed dynamic networks.

1.4 Thesis Structure

Chapter 2

This chapter presents work related to each element of this work. First a review of current data collection approaches used in developing regions is laid out to call attention to their strengths and pitfalls. Furthermore, it is shown that ground truth data on community health is currently deficient in developing regions and yet also essential for activities such as city planning and aid intervention. This message is then re-enforced with a reflection of the data collection methods currently being used in Tanzania, emphasising the need for fine-grained, reliable ground truth data in Dar es Salaam. This review then moves on to discuss alternative data streams such as Call Detail Records, Mobile Money and drone/high resolution satellite imagery which are used to address the problem of poverty detection. Following the theme of big data this chapter goes onto review the stochastic block model a method of community detection in network analysis. The stochastic block model is reviewed from the original model to recent developments and applications.

Chapter 3

This chapter looks at the the novel sources of big data which will be applied to the problem of poverty detection in this thesis. This work leverages existing data streams such as Call Data Records, Mobile Money and Satellite Imagery Data, each which are described in this chapter.

Chapters 4,5

A major contribution of this work has been the knowledge gained from a comprehensive data collection process. These chapters detail the motivations, methods and results of three types of data collection: 1. A grid survey modelled from comparative judgement methodologies applied commonly in the developed world; 2. A street survey modelled from traditional census style surveys commonly applied in developing nations; and 3. Semi structured interviews with local data collection experts and government officials. The processes and challenges of collecting data in this context are reviewed and validated via semi structured interviews. The fine-grained detailed ground truths produced in these chapters create a significant novel contribution of knowledge to this thesis.

Chapter 6

This chapter explores an examination of systematic bias existing within the data collection process. The affluence bias is introduced showing that more affluent areas tend to see themselves as worse off than they are.

Chapter 7

This chapter presents an assessment of theory based indicators of forced labour. Machine learning is used as a highly effective method for revealing new covariates to be leveraged as proxies for the detection of forced labour. Proxies identified are less sensitive than currently theorised covariates and can therefore be collected with fewer challenges and resistance.

Chapter 8

This chapter describes the stochastic block model, and experimental set up of novel data streams used in this chapter. The block model is applied to four separate networks of Call Data Records: SMS records over the weekends, SMS records over the week days, Call Records over the weekends and Call records over the week days. Communities are recovered from these networks by inferring the block structure using a Metropolis-Hastings optimization of the stochastic block model. These recovered communities are then compared with the ground truth information collected in chapters 4 and 5 in order to assess the application of the stochastic block model for fine grained detection of communities vulnerable to poverty.

Chapter 9

The time dependent nature of call data records used in this work is leveraged in this chapter, allowing incorporation of dynamic interactions into the stochastic block model. This chapter starts by describing the dynamic patterns seen in the call record data to highlight the potential benefit of considering this additional network information. The stochastic block model is then amended using Non-Homogeneous Poisson Processes such that network edges are allowed to change though time, while block structures remain constant. This model is then applied to the novel call data record data streams, and the block structure is inferred through Metropolis-Hastings optimization. Finally results are analysed and compared to the static results in chapters 4 and 5 to assess the value of incorporating time dependent data into stochastic block models for the fine grained detection of vulnerable communities.

Chapter 10

Finally the thesis concludes with a summary of the findings throughout this work in response to the original research questions. Key messages of the work are summarised, a personal reflection of the project is presented and future avenues of following research are described.

Chapter 2

Related Work

This section is split into two parts. The first part presents a critical review of existing approaches to collecting data in developing regions in order to highlight the current merits and gaps. The second part presents a review of the stochastic block model, from the original model to recent developments and applications. The stochastic block model will be used in chapter 8 to explore the community structure of subwards in Tanzania based on a network of CDR data.

2.1 Data Collection of vulnerable communities

In 2015 the United nations and 193 countries set out 17 Sustainable Development Goals (SDGs) aiming to make worldwide sustainable change by 2030. Africa is currently considered at risk of not meeting the poverty targets set out in these goals (Cuaresma et al., 2018), and one of the main factors preventing the success of these goals is a lack of reliable data. Developing countries lack the funding and infrastructure needed to create such data, despite their essential role in meeting SGDs (Espey, 2019; Evans and Ruane, 2019; Lupu and Michelitch, 2018).

2.1.1 Survey Methods

One traditional method of data collection for assessing poverty and related issues is by reviewing administrative data (Evans and Ruane, 2019). This can include information collected from facilities such as schools, hospitals and local education facilities. However, obtaining such data is often logistically difficult. Many developing areas are infrastructure-limited and data is often sparse or stored in paper form making access highly problematic (Espey, 2019; Evans and Ruane, 2019; Lupu and Michelitch, 2018). Data collection can also be done through surveying, most commonly implemented through one of the following methods: population censuses, household surveys and internationally standardised surveys (Tortora et al., 2010).

Population censuses contain basic information on education, employment, housing and basic accessibility as indicators of the welfare of regions (Elbers et al., 2003). These censuses are conducted over a 5-10 year basis and are extremely resource and time hungry. Household surveys present another method. Here data is collected with a focus on in-depth understandings of living conditions, and can be tailored to specific requirements of the surveillance (such as incidence of disease, nutrition, rates of childbirth) (Evans and Ruane, 2019; Lupu and Michelitch, 2018). Such surveys can be collected at a city or national level (Grosh and Glewwe, 1998). However, it is highly logistically challenging to survey entire populations of communities in household surveys especially in regions with such constrained resources and infrastructure. Hence sampling is used, this has potential to misrepresent regions containing a diverse range of people (Evans and Ruane, 2019; Lupu and Michelitch, 2018). There exists two predominant, internationally standardised surveys for poverty in the form of the *Demographic and Health Surveys* (DHS) and the *Multiple Indicator Cluster Surveys* (MICS). The DHS approach is to collect data to provide nationally representative samples, giving comparable results across countries (surveys are mostly developed with consistent questions for all countries). DHS's are an extremely important source of data as they provide both national and international data on families such as fertility, mortality, nutrition, health services access, FGM, domestic violence and access to clean water amongst other things. DHS's are usually implemented by national organisations such as the Bureau of Statistics. DHS's tend to be conducted approximately every 5 years with samples of between 5000 and 30,000 households taking part[1]. The comprehensive nature of these surveys and their ability to make international comparisons makes them an extremely important source of data in developing countries.

Another predominant method of data collection in this field is the Multiple Indicator Cluster Surveys (MICS), these are international household surveys. MICS are organised by UNICEF to help countries collect and assess data specifically related to children and women. The first MICS took place in 1995 in over 60 countries, since then MICS take place all over the work and since 2009 these household surveys have taken place every 3 years rather than every 5 years to meet the data demands[2]. MICS and DHS are very similar and continuously work together on improving methodologies and increasing available data.

[1] https://dhsprogram.com

[2] https://mics.unicef.org/surveys

The Harvard Humanitarian Initiative used rolling cross-sectional surveys when dealing with data collection which requires sensitive information. Rolling cross-sectional surveys use snapshots of populations to make inferences about a whole population. Rolling cross-sectional surveys are repeated periodically however unlike household surveys and DHS surveys each snapshot surveys a new set of randomly selected respondents who are not necessarily the same as the original set. Cross-sectional surveys can be conducted using any collection method but one of the most common methods applied to this type of survey is telephone interviews (Johnston and Brady, 2002; Lavrakas, 2008). Rolling cross-sectional surveys have been applied to many different context, some recent applications include the following examples:

- Illustrating malnutrition and associated risk factors among young children in Rural Ethiopia (Chen et al., 2021)
- Illustrating psychological distress, resettlement stress and school engagement in student refugees (Baker et al., 2021)
- Illustrating maternal mortality in low-resource settings (Maru et al., 2016)
- Illustrating the realities of climate change and the association between child poverty, poor governance and natural disasters (Daoud et al., 2016)

Despite these efforts, currently there is a deficiency in reliable, detailed data on community health in developing regions. Following a more detailed review of the data collection procedures currently utilized in the case study region of Tanzania, this chapter will go on to highlight the limitations and challenges of current data collection methods which was the motivation for this thesis.

Surveys in Tanzania

As this work is focused on Tanzania the section examines data collection recently occurring in relation to poverty and community health in this region. Tanzanian National Bureau of Statistics makes use of three main sources of data: the Demographic and Health Survey, Household Budget Survey and National Panel survey. Key observations, methods, strengths and weaknesses from these recent data collections in Tanzania are examined below.

Firstly, this chapter examines the Demographic and Health Survey (DHS). The most recent DHS conducted in Tanzania was the 2015-16 ‘Demographic and Health Survey and Malaria Indicator Survey’. The survey was implemented by the Tanzanian National Bureau of statistics, Office of Chief Government Statistician and Ministry of Heath Zanzibar, funded jointly by the Tanzanian Government, US Agency for International Development, Global Affairs Canada, Irish Aid, United Nations Children’s Fund and United Nations Population Fund (of Health Community Development Gender Elderly et al., 2016).

The survey was conducted between August 2015 and February 2016, with the aim to recover information on population health based on household experiences with nutrition, health care education and other associated characteristics. 12,563 households were successfully interviewed, including 13,266 eligible women aged 15-49 and 3514 men aged 15-49. These reflect response rates of 98% , 97% and 92% respectively from the originally selected households. Households were sampled from rural and urban areas throughout

Tanzania, surveys covered nine zonal areas (Central, Western, Southern Highlands, Southern, South West Highlands, Eastern, Northern, Lake, Zanzibar). Some indicator questions were asked at regional levels but not all regions are covered in this, and areas can be seen in figure 2.1.

While this survey was high-yielding, and gives a good level of detail into the demographics and population health of the nine zones of Tanzania, the areas summarized remain large and generalised. This work nevertheless generated fundamental information previously articulated: 46% of the population of Tanzania are under 15 years of age. 15% of women and 8% of men have no education at all. 23% of Women and 28% of Men have secondary or higher education. Children born to mothers with no education are more likely to die before their 5th birthday (83 vs 60 deaths per 1000 live births). 84% of married women and 99% of married men aged between 15-49 are employed, most likely paid in cash, 42% and 10% not paid for their work. Startlingly, two thirds of women in Tanzania reported problems accessing health care. 58% of women and 40% of men agreed that a husband is justified in beating his wife (for certain reasons) (of Health Community Development Gender Elderly et al., 2016).

The Household Budget Survey (HBS) is examined next. Each of these existing surveillance methods used in Tanzania is examined in detail to gain as much contextual knowledge as possible. These examinations educate, and produce a critical evaluation of current methods which motivates the novel survey techniques suggested in chapters 4 and 5. The HBS was last implemented in 2017-2018 by National Bureau of Statistics in collaboration with the Poverty Eradication Division in the Ministry of Finance and Planning. While informative, the Household Budget Survey's cannot be

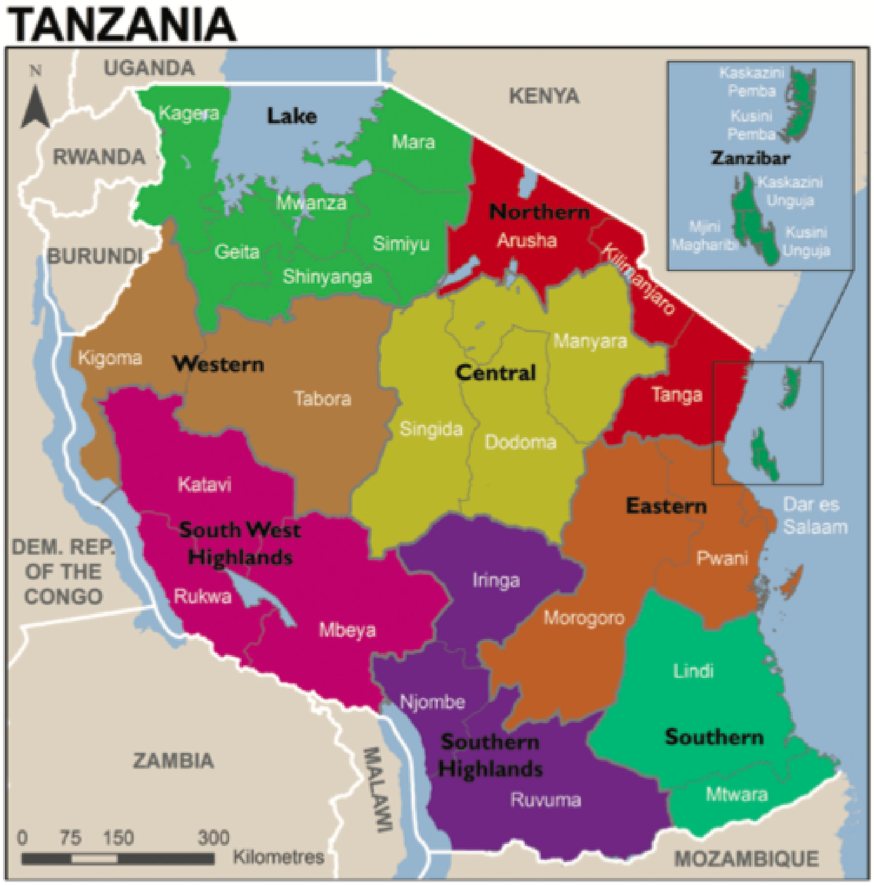


Figure 2.1: Tanzania demographic and Health Survey and Malaria Indicator Survey, Zones and Regions.(of Statistics, 2019)

collected very frequently because of the expense, time and resources needed to complete them, a motivating factor for the methods proposed in this thesis. The most recent HBS before the 2017/2018 survey in Tanzania was implemented in 2011/2012, jointly funded by the Tanzanian Government, World Bank, UN Women, Irish Embassy, United Nations Childrens Fund and Global Affairs Canada. The extent of the required stakeholders emphasises the logistical challenges of such data collection (with the funding here very similar to the Demographic and Health Survey). The aim of the Household Budget Survey is to assess progress in the improvements of living standards for people in terms of poverty and associated characteristics (of Statistics, 2019). The 2017/18 Household Budget Survey in Tanzania took one year and included 9552 households, 9465 of which completed the survey, conducted at a regional level.

With each of the 26 regions of Mainland Tanzania (shown in figure 2.1), included in the HBS, it reflects a more fine-grained survey than the Demographic and Health Survey. However these regions are still vast and contain many diverse districts. Understanding the composition of local, potentially vulnerable communities is simply out of scope. This highlights the need for comprehensive new data focused on fine-grained scales, such as the results collected in this thesis which are carried out over subwards, the smallest administrative geographic region in Tanzania. The regional estimates from the Household Budget Survey nonetheless do create estimates for zonal, urban, rural and national level (of Statistics, 2019).

Poverty is measured in the HBS using information on a number of different household characteristics, such as: ability to buy food and other essentials to live; health care; housing conditions; lighting conditions; toilet facilities;

household size, and proportion of dependants per household. Other indicators of poverty include access to clean water, transport and communication facilities availability, rates of employment, role of women in a household and ownership of personal identification (of Statistics, 2019).

The 2017/18 Household budget survey in Tanzania did produce valuable information needed to assess the progress of poverty prevalence in Tanzania. Some of the key findings include the following: Only 31% of the population has a birth certificate or official notification, 51% of households have modern floors, 29% of households have electricity inside. 43% of households own a radio, 24% own a television. 78% own a mobile phone (note this is up from 57% in 2011), 83% of 7-13 year olds are currently studying, 52% of adults above 15 are employed into agriculture and 26% of the population below the food poverty line. While the HBS is developing and adapting each year, for instance incorporating tablet usage and regional level questions, its real benefit despite its sparseness, is that there is a consistency kept in the questions over the years so development and changes can clearly be seen (of Statistics, 2019).

Finally, let us consider the National Panel survey in order to further contextualise the need for this research. The National Panel survey was last implemented in Tanzania from October 2014 to November 2015, with the National Bureau of Statistics Tanzania and the Office of Chief Government Statistician Zanzibar implementing it jointly. The National Panel survey is focused specifically on providing data over time of a continued sample of households in an attempt to track national development agendas, poverty dynamics and to evaluate impacts of policies and development actions within Tanzania. The National Panel Survey splits this agenda into two themes.

The first theme is focused on the overall improvement of quality of life and wellbeing and reduction (or increase of poverty). The second theme looks at gaining an understanding of success of governance and other actions taken against poverty (United Republic of Tanzania, 2017). Revisiting the same households over time is a valuable advantage as it really allows change and impact to be monitored and reviewed at a household level. (Note: The National Panel survey is jointly funded by European Commission, World Bank, Ministry of Finance and Planning, Gates Foundation and the Tanzanian Statistical Master Plan.)

A beneficial feature of the National Panel Survey is the emphasis placed on pilot surveys. The national panel survey invests considerable efforts completing in field pilot surveys, in order to iteratively improve surveys and ensure they are suitable for the regions under investigation (United Republic of Tanzania, 2017). This is a valuable procedure to complete before large amounts of time and money are spent rolling out the main data collection. This feature is something which has been important to reflect in this work, especially when fine-tuning logistics and questions styles.

Like the Demographic and Health Survey and the Household Budget Survey the National Panel Survey makes use of a number of key indicators for poverty and its associated characteristics in order to present information on the overall poverty of areas. Some of the indicators used in the National Panel Survey include, the role of women in a household, household size, access to piped or protected water sources, access to sanitation facilities and electricity within households. Other indicators include education levels of males and females, health and nutrition of individuals, medical access and satisfaction with health care, personal paperwork ownership, food security and access and use of technology such as personal mobile phone ownership (United Republic of Tanzania, 2017). Investigating the spread of poverty using indicators can provide detailed descriptions of a regions demographics, for this reasons this is something which is investigated in chapter 5.

The most recent National Panel Survey in 2014 created a high volume of essential data on the levels of poverty and its associated characteristics. This data illustrates key information about the households, for example 46% of households has access to clean drinking water in rainy season, 57% in the dry season. 34.5% of households reported worrying about not having enough food, this is an increase from 32.9% in the 2011 wave of the National Panel Survey of these households. 30.4% of births from these households were not attended by any skilled health workers in that 24 month period (United Republic of Tanzania, 2017). Having this information at a household level is a key strength of this survey as comparisons can be made between specific household areas and at regional levels. However as only a small number of households are surveyed in each region there is no information to compare smaller governmental regions such as wards and subward. This emphasizes the gap that this thesis is set against, while surveys exist, they are at levels of granularity which currently make them inapplicable to local interventions.

An additional, well cited survey in Tanzania worth mentioning is that of (Liviga and Mekacha, 1998), who conducted a survey in Dar es Salaam focused on youth migration and poverty. This work considered push and pull factors of young people out of rural areas and into the city in relation to poverty. The hypothesis was that youth migration is both a result and a cause of poverty in migration areas. The push factors (the reasons young people are leaving areas) were: low income, lack of access to social services, poor transport, and unemployment, whereas the pull factors into the city were: prospects of a better life, employment, self-employment opportunities and perceived access to social services. The work reported that movement into cities increases poverty related issues such as crime, unemployment, housing problems, sanitation and increased demand for social services. In particular, the work showed that of 250 traders interviewed in Dar es Salaam who had migrated in to Dar es Salaam from surrounding areas such as Temeke,Ubungo, Kariakoo and Msasani, 46.8% came to Dar for social and economic opportunities, 34% followed a friend or family member and 15.2% came to do solo business and trading, and 4% came for further studies.

In contrast to the national surveys previously detailed, (Liviga and Mekacha, 1998)'s survey was carried out at a fine-grained level looking specifically at Dar es Salaam. However the information collected by this work was aimed at not only exploring where poverty is but also what some of the drivers of poverty are. This survey in both its style (fine grained and with a number of local in field pilots) and its expansion of indicators, informed some of the practises in this thesis. Whilst (Liviga and Mekacha, 1998)'s work did leverage the traditional poverty indicators such as health, technology, electricity and education for example, there is also a strong focus on some of the motivating factors behind poverty and its related characteristics (Liviga and Mekacha, 1998). This is the reason for questions about migration and different forms of informal work (such as street selling) included chapter 5. These questions echo this attempt to get a more detailed picture throughout Dar es Salaam.

Vulnerability Indicators

While vulnerability arises in numerous forms, collecting information on poverty is essential for development to happen. Statistics on poverty can be used by governments and NGO's to launch work and activism done to improve the safety and quality of life of those affected. Not only can information on poverty initiate change, but accurate information is essential for optimal targeting of resources. Collecting data on poverty however, is not simple. Surveys asking direct questions about the level of poverty in an area will be subject to bias from a range of different agendas and perspectives (Lupu and Michelitch, 2018; Ties Boerma and Sommerfelt, 1993). In order to collect estimates of poverty and other sensitive demographic information surveyors therefore often turn to proxies. Such proxies are often easier to collect and create detail information about an area. Commonly used poverty proxies include: • food security, • access to health care, • housing conditions, • lighting conditions, • household size, • proportion of dependants per household, • the role of women in a household, • access to piped or protected water sources, • access to sanitation facilities, • electricity within households. Other indicators include • education levels of males and females • health and nutrition of individuals, • satisfaction of health care and • access and use of technology such as personal mobile phone ownership and ownership of personal identification. (Sundar and Sharma, 2002; Bradshaw et al., 1998; Appleton and Booth, 2001; Roy et al., 2018; Flanagan et al., 2011; Tarozzi and Deaton, 2009; Hoogeveen et al., 2004; Martins, 2007; Kates and Dasgupta, 2007). Many proxies can be seen applied to the three recent surveys in Tanzania mentioned earlier in this chapter (United Republic of Tanzania, 2017; of Statistics, 2019; of Health Community Development Gender Elderly et al., 2016) and are similarly explored in this thesis in chapter 5 to underpin modelling efforts .

2.1.2 Challenges with Ground Truth Methods

Despite the existence of these data sources, they are not without limits and challenges. In fact data indicating poverty levels has been historically poor, throughout lower income countries, and especially so in Africa (Channing Arndt, 2016). Developing countries have notably been held under scrutiny for the reliability of quantitative data, with National statistics on economic production for example being off by as much as 50% in Africa (Jerven, 2013). Only 35% of sub-Saharan countries have data on poverty that have been updated since 2015 (Espey, 2019). While data collection in the developing world has drastically improved over the last few decades, particular in the frequency and detail of collections, there are still significant deficiencies in this area (Perez et al., 2016; Ties Boerma and Sommerfelt, 1993; Lupu and Michelitch, 2018). One key factor that has inhibits the progression of data collection in the developing world is the fact that most of the methodologies used are derived from experiences in developed countries (Lupu and Michelitch, 2018). Data collection in the developed world has very different challenges to those faced in the developing world. It is vital that data collection methods are adapted to suit the area under investigation (Liviga and Mekacha, 1998; Lupu and Michelitch, 2018). A phenomenon that became very apparent during pilot phases employed to prepare for the data collection carried out in this thesis. Chapters 4,5,6 will cover the specific adaptions and challenges in methodology that came about in this work.

When considering other factors relating to vulnerability, such as community health, additional surveying issues appear, such as expense, lack of detail and frequency. One major limitation for data collection in developing regions, and a factor responsible for some of the lack of reliability of data, is the lack of infrastructure. National statistics in developing countries are often compiled on paper, or manually typed into out of date computers unavailable online or collected by teams lacking in training (Espey, 2019).

Many of the problems associated with this lack of infrastructure are routed in the lack of funds available for data collection in such regions. Surveys are cost heavy, with developing countries far less likely to have adequate finances readily available and justifiably applied to data collection and research (Espey, 2019; Ties Boerma and Sommerfelt, 1993; Blumenstock et al., 2015; Perez et al., 2016; Fields, 1989; Xie et al., 2015a).

A further major challenge for data collection in developing countries is scale, when combined with limited resources, it is often logistically challenging for all regions of a country to be covered (of Health Community Development Gender Elderly et al., 2016). Significant proportions of the poorest people are often missed from more rural areas (Carr-Hill, 2017); sample statistics are not necessarily a true reflection of the area; and due to the geographical extent of targeted zones and regions, surveys often miss information regarding the variation of poverty and community health, at fine grained levels such as towns or villages (Ties Boerma and Sommerfelt, 1993; Fields, 1989; Evans and Ruane, 2019; Carr-Hill, 2017; of Health Community Development Gender Elderly et al., 2016). Having data at more fine-grained levels would allow for more potentially high impact decisions such as resource allocation.

Given that poverty and related issues remain highly sensitive topics another pitfall data collection of this type often falls into is *recall biases* (Lupu and Michelitch, 2018; Ties Boerma and Sommerfelt, 1993). Many people reporting data have underlying motivations that must be considered (for example, respondents may believe that describing an area as worse off than it is will result in government improvements being made). Alternatively the sensitive nature of the questions may result in respondents feeling uncomfortable providing answers. The need to paint a more positive picture than is accurate about things such as employment or access to food, due to pride or in order to protect themselves from scrutiny, have been reported (Wilpen and Daniel, 2007; Perry et al., 2013; Lupu and Michelitch, 2018; Ties Boerma and Sommerfelt, 1993).

These challenges of poverty data collection are also exacerbated by rapid urbanisation and dynamic community health. Manual sampling used in data collection is often difficult to replicate and run consistently over the time scales that poverty features are changing. Dar es Salaam for example has seen fast evolving slum areas and uncontrolled growth over the last two decades. Indeed, surveyed areas in developing countries are often changing at a faster rate than they can be effectively surveyed, rendering information on poverty rapidly out of date (United Republic of Tanzania, 2017; Evans and Ruane, 2019; Ties Boerma and Sommerfelt, 1993; Xie et al., 2015a).

There are many common proxies, indicators and data collection techniques regularly being used. However, perhaps surprisingly, there is no universally accepted method for describing the level of poverty in a specific area (Martins, 2007). With methods previously adapted being open to challenges and limitations, it remains difficult to obtain true estimates of poverty, and equally challenging to evaluate the different methods available (Edward and Sumner, 2014). It is clear that there is still a deficiency in reliable data describing local poverty levels in developing countries, due to a number of challenges (Perez et al., 2016; Xie et al., 2015a). There is, however, a notable increase in data being collected on these issues, with the improvements in methodologies and accuracy of data collection (for poverty and associated characteristics) being vital for development (Blumenstock et al., 2015).

Overall there are a number of challenges that are creating the lack of reliable available data in developing countries on community health and poverty: challenges that constrain the impact of local policy makers, governments and aid organisations (Perez et al., 2016; Xie et al., 2015a). Of most relevance to this thesis and its application setting, there remains no fine-grained, local, recent information at a subward level in Dar es Salaam on either poverty, or related issues. This thesis aims to move towards addressing this gap, and in particular through alternative methods for modelling vulnerability in the population.

2.1.3 Alternative Methods of Detecting Poverty

The previous section highlighted that surveyed data is hard to obtain, being labour and cost intensive, and as a consequence scarce (Xie et al., 2015a). Yet accurate estimates of population characteristics such as poverty are well evidenced as critical to development (Espey, 2019; Blumenstock et al., 2015; Fields, 1989). In this thesis, I consider methods of data collection other than traditional surveys.

It is most common for surveying techniques to employ a linear list of short, rapid fire questions, which are presented to respondents for completion.

However, an alternative data collection approach common in social science research is the *bubble sort* method. (Rugg and McGeorge, 1997; Cataldo et al., 1970; Rugg and McGeorge, 1997) More frequently used in western data collections, this forms an online questionnaire platform with an interface similar to traditional Q-sort and pile sort methods. These traditional sort approaches require participants to rank or partition a selection of features into categories. This is a great method for bridging the gap between quantitative and qualitative data research allowing participants to show their opinions in a way easily analyzed numerically by using similarity matrices and standardization grids (Rugg and McGeorge, 1997).

One problem associated with this type of survey is the influence of position bias, for example if a participant has placed cards in a certain box a number of times in a row they may then feel the need to alter their next card in order prevent unbalanced groups. It is evident that it is important to introduce bias checks in this type of data collection (Cataldo et al., 1970; Rugg and McGeorge, 1997). One way to improve the confidence of such sorting surveys is to introduce random labelled cards into the cart set. For example labelling a card ”Put me in category A” and immersing it in the set will help to highlight when people are randomly clicking to sort the cards (Cataldo et al., 1970).

This type of data collection is a very common form of research in the developed world, but is yet to be applied to the context of poverty detection in developing regions. One of the outcomes of this thesis is a bridging of this gap via a Grid surveying approach (seen in chapter 4. The goal of the survey techniques will be to create a map of the affluence of the subward regions of Dar es Salaam. Unlike the traditional techniques this survey will focus on the affluence of the regions, without collecting details on any other demographics.

Another alternative method of detecting poverty and associated characteristics is through the use of earth observation (Watmough et al., 2016; Perez et al., 2016; Jean et al., 2016a; Pinkovskiy and Sala-i Martin, 2016). With high-resolution satellite imagery now increasingly inexpensive and reliable (Xie et al., 2015a), its processing especially via convolution network technique offers a more time efficient solution for the detection of poverty compared with traditional surveys (Watmough et al., 2016). Although satellite data offers fruitful and significantly quicker methods than traditional ones such as surveys, there are still a lot of costs involved particularly with the equipment used (Blumenstock et al., 2015). Poverty stricken regions are also those most likely to have less internal funding, more civil-wars, poor infrastructure and inadequate government resourcing available for either surveys or satellite data; hence there are still vast gaps in the collection of reliable data which could be used to describe poverty (Perez et al., 2016).

There are, however, increasingly new sources of collecting data on localities, via interrogation of anonymised mobile phone and internet records. These are enabling new approaches to demographic profiling and opening an exciting field of potential analysis (King, 2011). Data from a communication network of mobile phones or business landlines, for example, has been used to show that communication diversity is a strong indicator for the economic health of communities in the UK (Eagle et al., 2010). Many therefore see movement towards a more digital ecosystem of data is a key step in meeting sustainability goals and tackling poverty (Espey, 2019).

In developing countries there are fewer sources of big data, yet mobile phone use in particular has become increasingly ubiquitous in these regions, and now provides a fruitful source of data (Blumenstock et al., 2015). In regions where resources such as time, labour and money our scarce for such research. A data driven approach of this fashion creates a method for gathering information on communities at a fraction of the cost of traditional methods such as surveys and satellite images (Blumenstock et al., 2015). The diversity of individuals' relationships is also a key indicator of social and economic life, but until recently this was not so widely quantifiable. We now have data on networks of people and their behaviours which allow us to draw conclusions at population levels (Eagle et al., 2010). Such data streams are now commonly being used to address issues of sustainability and development. In fact, Mobile data, for example, has been mapped to predict wealth across Rwanda, aggregating individual phone subscriber estimated affluence statistics (Blumenstock et al., 2015). A further example is the use of approximately 10 million mobile phone subscribers from multiple operators in Sri Lanka to assess the land use of the regions (Madhawa et al., 2015). The ubiquitous use of mobile phones in developing countries is creating data with a lot of potential for the identification of the most vulnerable parts of regions. Yet, as this is a new form of data there is a lot of work to be done in finding different ways to assess this data for useful results (Smith-Clarke and Capra, 2016).

Attention is also turning to other behavioural data sources such as mass transaction data from credit card and Mobile Financial Services (MFS) (Engelmann et al., 2018; Jean et al., 2016b). Here, derived variables can reveal determinants of human behaviour, rather than geographical features alone.

The use of these data sources and machine learning in helping address Sustainable Development Goals (SDGs) continues to expand (Holloway and Mengersen, 2018; Vinuesa et al., 2020). Applications of machine learning are already providing routes to quicker policy interventions while attempting to overcome human observation biases (Wilpen and Daniel, 2007; Perry et al., 2013). Machine learning methods are being used proactively in assessing vulnerability for example: Identifying cases of inequality (Dalenberg, 2018); simulating the effects of conflict (Saam and Harrer, 1999) and; assessing transmission rates of HIV (Brdar et al., 2016) and Malaria (Wesolowski et al., 2012). Additionally machine learning has been used with satellite imagery to detect sites of vulnerability to slavery including: Brick kilns (where debt bondage is common) (Foody et al., 2019); fish farms (Lechner et al., 2017) and mining sites (Bales, 2012). UN SDG 2 (Ending Hunger), UN SDG 6 (Clean Water and Sanitation) and UN SDG 15 (Life on Land) in particular have benefited from advances in remote sensing and statistical learning (Holloway and Mengersen, 2018).

Whilst the use of novel data streams such as drone and satellite imagery has accelerated interest in this area, challenges remain. As previously mentioned, imagery comes at much expense, it also suffers from resolution limits and impairment by weather and related phenomena (see for example (Smith-Clarke et al., 2014)). Partly, these limitations with imagery emphasise the advantages of interaction data such as CDR, which can be utilized by models in network analysis to infer about the networks. Network analysis of such data has been used for a number of different fields, from epidemiology to political science and social psychology. Such network analysis is less widely used in problems of development, vulnerability and poverty. In particular, one form of network analysis, called *block structure* analysis is able to infer communities from a network of nodes and edges. Given the potential of this approach a core aim of this work is to investigate if this block model can provide insights into identification of communities vulnerable to poverty, providing not only a novel use of the analysis, but also to examine the potential for social interventions that leverage the approach.

2.2 Stochastic Block Model

2.2.1 Networks

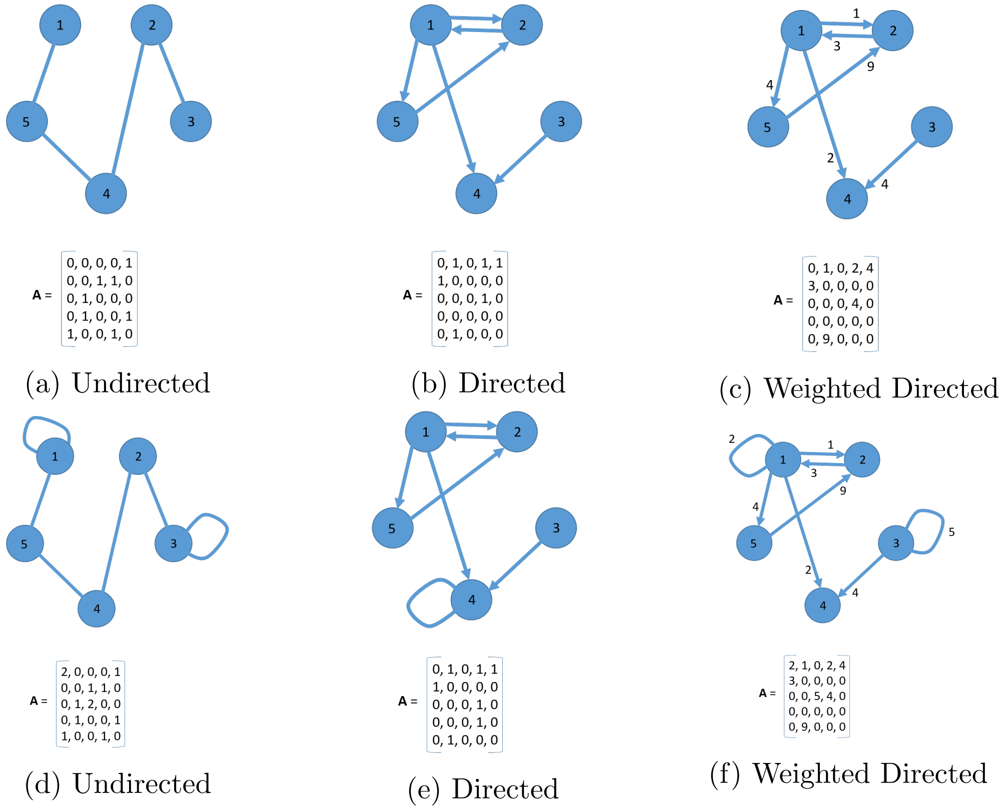
Network analysis provides insight into structural relations in a number of fields such as: physics, biology, computer science, and sociology. The diversity of potential applications of network analysis can be seen in the following examples: (i) Functional Brain Connectivity in Alzheimer's disease. (Supekar et al., 2008); (ii) South Korean politics though citizen blogs (Woo Park and Jankowski, 2008); (iii) Social networks and the survival of wild bottle-nose dolphins (Stanton and Mann, 2012). In order to introduce the specifics of the stochastic block model, this section begins with an introduction of basic network notation.

Networks are a set of objects which are connected together. The objects are known as *nodes* or *vertices* and can represent anything from people, to places or parts of the brain. The connections between these nodes are called *edges*, these could represent anything which connects the nodes, such as types of communication or travel. The *network graph* can be defined as a combination of these nodes and edges such that $G = (N, E)$, where N is the set of nodes and E is the set of edges. The edges can be represented in a matrix known as the *adjacency matrix.* A is the adjacency matrix such that $A_{i,j}$ is the connection from node i to node j. There are three main types of network edges: *Undirected Edges*, *Directed Edges*, and *Weighted Edges*. A graph with undirected edges, does not consider the direction of a connection just its existence, hence $A_{i,j} = A_{j,i}$. In contrast, a graph with directed edges takes into account the direction of connections, as such $A_{i,j}$ is not necessarily the same as $A_{j,i}$. In undirected networks A is symmetric along its major axis, this is not a requirement for directed networks. In a graph with weighted edges there can be more than one edge between nodes, $A_{i,j}$ represents the number of edges from node i to node j.

Additionally, the *degree* of a node is the number of edges connected to that node (*in-degree* refers to the number of incoming edges, and the *out-degree* refers to the number of outgoing edges). Another thing to consider is *self-edges*, otherwise known as *self-loops*, this is the idea of a connection (or multiple connections) from one node back to itself. In networks where self-edges are not allowed $A_{i,i} = 0$. The self-edges of a network are seen on the diagonal of an adjacency graph. Self-edges in a undirected network are by correction written in the adjacency matrix as 2 x the number of edges (so the sum of the degrees of all the vertices is twice the number of edges). This correction is convenient because the various formulas hold whether or not undirected networks have self-edges. In the case of directed networks this is not necessary so the self-edges are represented in the adjacency matrix by the actual number of connection as normal (Newman, 2010). Figure 2.2 illustrates some examples of these edge types with their corresponding adjacency matrices.

Notation

A summary of the notation used in the rest of this review chapter is present below:



$$\mathbf{A} = \begin{bmatrix} 0, 0, 0, 0, 1 \\ 0, 0, 1, 1, 0 \\ 0, 1, 0, 0, 0 \\ 0, 1, 0, 0, 1 \\ 1, 0, 0, 1, 0 \end{bmatrix}$$

(a) Undirected

$$\mathbf{A} = \begin{bmatrix} 0, 1, 0, 1, 1 \\ 1, 0, 0, 0, 0 \\ 0, 0, 0, 1, 0 \\ 0, 0, 0, 0, 0 \\ 0, 1, 0, 0, 0 \end{bmatrix}$$

(b) Directed

$$\mathbf{A} = \begin{bmatrix} 0, 1, 0, 2, 4 \\ 3, 0, 0, 0, 0 \\ 0, 0, 0, 4, 0 \\ 0, 0, 0, 0, 0 \\ 0, 9, 0, 0, 0 \end{bmatrix}$$

(c) Weighted Directed

$$\mathbf{A} = \begin{bmatrix} 2, 0, 0, 0, 1 \\ 0, 0, 1, 1, 0 \\ 0, 1, 2, 0, 0 \\ 0, 1, 0, 0, 1 \\ 1, 0, 0, 1, 0 \end{bmatrix}$$

(d) Undirected

$$\mathbf{A} = \begin{bmatrix} 0, 1, 0, 1, 1 \\ 1, 0, 0, 0, 0 \\ 0, 0, 0, 1, 0 \\ 0, 0, 0, 1, 0 \\ 0, 1, 0, 0, 0 \end{bmatrix}$$

(e) Directed

$$\mathbf{A} = \begin{bmatrix} 2, 1, 0, 2, 4 \\ 3, 0, 0, 0, 0 \\ 0, 0, 5, 4, 0 \\ 0, 0, 0, 0, 0 \\ 0, 9, 0, 0, 0 \end{bmatrix}$$

(f) Weighted Directed

Figure 2.2: Network Edges Type Examples with Corresponding Adjacency Matrices

$E =$ Set of edges

$N =$ Set of n nodes $[1, 2, ..., n]$

$G =$ Graph made up of nodes and edges

$A =$ Adjacency matrix $(n \times n)$, such that $A_{i,j}$ represents edge (edges) from node i and node j

$K =$ Number of block structure groups

$Z =$ Matrix $(K \times K)$ of edges between block groups, such that Z_u, v is the number of edges between block u and block v

$M =$ Matrix $(K \times K)$ of block probabilities, such that $M_{u,v}$ is the probability of an edge from block u to block v

$N_u =$ Number of nodes in block u

$N_{u,v} =$ Number of possible edges from block u to block v

$p_u =$ Probability of node i being in block u

$g =$ Block membership vector such that, $g_i = u$ if node i is in block u

2.2.2 History, Adaptations and Applications

Early History and Applications

This thesis is interested in one particular form of network analysis, the stochastic block model, which is able to infer communities (groups of nodes) from a network of nodes and edges. An essential predecessor to the stochastic block model is the idea of stochastic equivalence (Lorrain and White, 1971). Block modelling is based on the concept of structural equivalence: two nodes are structurally equivalent and hence belong to the same block if they relate to other nodes in the same way (for structurally equivalent nodes i and j, i has the same (and independent) probability of connecting with node k, as j does) (Lorrain and White, 1971, 1977; Holland et al., 1983a; White et al., 1976a; Faust and Wasserman, 1992). Blocks can be thought of as groups of structurally equivalent nodes known as *equivalence classes.* Block modelling was the result of mathematicians trying to find these stochastically equivalent groups. Another important predecessor of the model was (Erdös, 1959). Erdos introduced the concept of generating independent edges based on overall network probabilities. This allowed the idea that an observed network can be modelled as a product of independent Bernoulli trails.

Prior to the named stochastic block model there were a number of publications presenting approaches of dividing networks into communities based on different criteria (Breiger et al., 1975; White et al., 1976b). Combining the idea of these approaches, independent Bernoulli trials and stochastic equivalence, the block model was first established by (Holland et al., 1983b). Around the same time as Holland published the named Stochastic Block Model there was a lot of similar work done in the computer science field where the model was referred to as the *planted partition* model. In fact over time in different fields the model has been referred to with a number of names: Stochastic Block Model, planted partition model and in-homogeneous random graph model (Coja-Oghlan and Lanka, 2010; Abbe, 2017). Following the publication of the origin of the model, there are a number of early developments which made the model what we consider the stochastic block model nowadays. (Wasserman and Anderson, 1987) introduced a development allowing *a posteriroi* block modelling, at the same time (Wang and Wong, 1987) applied the Stochastic block to real directed graphs, they assumed that the block structure is known *a priori.* (Snijders and Nowicki, 1997) studied undirected networks with the assumption that there are two unknown blocks and the probability of an edge between two nodes depending only on the block the nodes are in.

From 1994 to current research, the stochastic block model has been widely used for social network analysis (Wasserman and Galaskiewicz, 1994; Doreian et al., 2005). Community detection is the recovery of groups of nodes from observed network edges, this is an instrumental process in network analysis. The increasingly popular stochastic block model can be used to recover the structure (node groups) of a wide range of types of network. There is a long history of beneficial applications of the model to real world data. The model has been applied to transport network, social media networks, neurological networks and human relations just to name a few.

The stochastic block model has a high level of application in the fields of biology and medicine. Here are some example of a network applications in these fields: • Applying the stochastic block model to single-cell whole-genome sequencing data from breast cancer patients to infer clonal composition (identifying tumor clones) (Myers et al., 2020); • Generating networks of the spread of epidemics using the stochastic block model based on known communities (Yang and Zhang, 2020); • Applying the stochastic block model to neurological networks (connectome) to infer brain atrophy for Alzheimer's diseases and other neurological diseases (Moyer et al., 2015); • Recovering tree and fungal groups with the stochastic block model and species host–parasite interaction networks in forest ecosystem (Mariadassou et al., 2010b). The stochastic block model has been applied to a number of other real world networks in medical and biological research, from global migrations patterns and neural connections in the human brain, to protein interactions and gene regulations (Peixoto, 2018; Airoldi et al., 2008; Jiang et al., 2009).

Another area of study which makes heavy use of stochastic block models is human behaviour fields such as psychology, political science and business. Some examples of application in these study areas are as follows: • The stochastic block model has been applied to friendship networks to infer grades and adolescent health (Airoldi et al., 2008); • Recovery of users political stances based on a network of online blogs via the stochastic block model (Vu et al., 2013). In addition the stochastic block model has been applied to a number of real world networks in these fields, from voting patterns in congress to world trade patterns (Peixoto, 2018; Tallberg, 2004; Zhang et al., 2016).

In fact, the application of the stochastic block model has been extremely wide and includes topics such as social networks of dolphins and fictional characters (Lu and Szymanski, 2019). This is not an exhaustive list of applications, it is important to emphasise that the stochastic model has wide reach and potential. Despite the heavy use of the stochastic block model in such varied applications, to my knowledge it has not yet been used in the context of this work for detecting vulnerable communities.

Basic Stochastic Block Model

Stochastic block models are becoming increasingly prevalent in network analysis. The stochastic block model allows inference of the latent community structure of a network. This inference learning from the network data combines approaches in statistics and machine learning to recover the network's structure. In order to go onto explain the extensions and applications of the stochastic block model here is a simple description and example of the stochastic block model.

Currently the adjacency matrix is binary, that is edges are unweighted and either there is a edge or their isn't. Many of the extensions which will be looked at later in this literature review including weighted graphs, as this can be very useful for real world data. However, note that for this basic explanation the graph is undirected, such that $A_{i,j} = A_{j,i}$ which equals 0 if there is no edge, or equals 1 if there is an edge between the nodes i and j. Further to this, the major axis itself will be made up only of zeros, as for now self-loops are not considered. $A_{i,i} = O$ for all i in N.

In the stochastic block model each of the nodes belongs to one of K groups. The number of groups K is less than the total number of nodes, at this point only fixed values of K are considered. Each node's group membership is unknown and each node can only be a part of one group. Now there is vector g the *block membership vector*, which is a vector of length n such that $g_i = u$ if node i is in block u.

There is also a $K \times K$ matrix $Z =$ matrix of edges between groups. Where $Z_{u,v}$ is the number of edges between group u and group v. $Z_{u,u}$ shows the number of edges between group u and itself. As A is undirected, Z is symmetric. In the stochastic block model the most important concept to be aware of is that the generation of the graph G and its edges depends on the group membership of the nodes. There is also a $K \times K$ block matrix M such that $M_{u,v} \epsilon [0, 1]$, $1 \leq u \leq v \leq K$ is the probability of an edge in between group u and group v. As G is undirected, M is symmetric.

The probability $P(A_{i,j} = 1)$ is assumed to follow a Bernoulli distribution on the block matrix given g_i and g_j. $P(A_{i,j} = 1)$ is independent of $P(A_{k,l} = 1)$ for $(i, j) \neq (k, l)$. In other words each edge is conditionally dependent given g_i and g_j. It follows that the total number of edges between block u and block v are Binomially distributed with a mean that is the product of the total number of potential edges available between the two groups and $Z_{u,v}$. The total number of potential edges between two groups is dependent on the number of nodes in each of the groups. In the stochastic block model the core idea is that the probability of an edge occurring between two nodes in a graph depends on only the group each node is a part of. Basically two nodes i and j share the same probability of connecting with another node r if both i and j are in the same block.

Letting $N_{u,v}$ be the number of possible edges from block u to block v. Looking at the most basic form of the stochastic block model the following assumptions are made: the graph G is undirected and unweighted; there are no self-loops; edges between blocks $Z_{u,v}$ are independent binomial random variables. The likelihood function for the stochastic block model is as follows.

$$L(G \mid M, g) = \prod_{u \leq v}^{K} (M_{u,v})^{Z_{u,v}} (1 - M_{u,v})^{(N_{u,v} - Z_{u,v})} \tag{2.1}$$

Then because $Z_{u,v}$ are independent and there are no constraints on $M_{u,v}$, the maximum likelihood estimate for M can obtained by maximizing each of the marginal likelihood functions:

$$L_{u,v}(G \mid M_{u,v}, g) = (M_{u,v})^{Z_{u,v}} (1 - M_{u,v})^{(N_{u,v} - Z_{u,v})} \tag{2.2}$$

Hence, with respect to $M_{u,v}$ the maximum occurs at:

$$\widehat{M}_{u,v} = Z_{u,v}/N_{u,v} \tag{2.3}$$

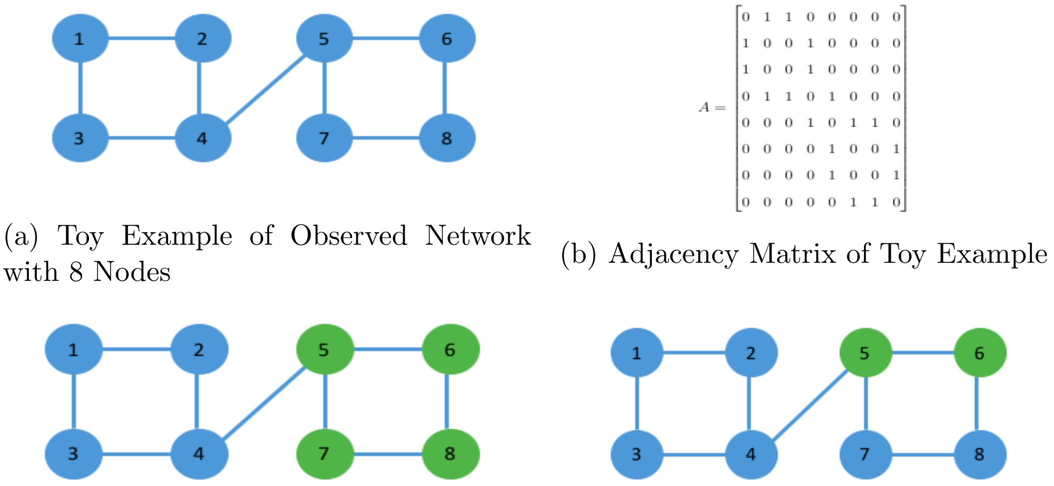
Therefore the maximum likelihood estimate of $\widehat{M}$ is made up of the maximum likelihood estimates $\widehat{M}_{u,v}$. $\widehat{M}$ can be substituted into 2.1 to create the objective function:

$$L(G \mid g) = (Z_{u,v}/N_{u,v})^{Z_{u,v}}(1 - (Z_{u,v}/N_{u,v}))^{(N_{u,v}-Z_{u,v})} \tag{2.4}$$

The objective function is large for *good* group assignments and small for *bad* ones. This is the most basic form of the stochastic block model. It is worth noting that while in this thesis the focus is on the inference of parameters M and g based on an observed network, the stochastic block model can be used for either generation or inference. Generation means that each node has a fixed community membership, which determines with which probability an edge exists to other nodes. Inference however, aim to discover the block structure which maximize the likelihood of an observed network. This thesis is focused on the inference of the stochastic block model in order to recover community membership, given a real world observed network. The latent block structure g follows the multinominal distribution with probabilities p. $L(g|p) = \prod_{u=1}^{K} p_u^{n_u}$ Here p is the vector $(p_1, p_2, ...p_K)^T$, showing the block assignment probabilities. $P(g_{i,u} = 1) = p_u =$ probability of node i being in block u. Hence $\sum_{u=1}^{K} p_u = 1$.

Toy Example

To clarify what has been explained up to this point I illustrate the stochastic block model as described above using the following example. Note in this



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

(a) Toy Example of Observed Network with 8 Nodes

(b) Adjacency Matrix of Toy Example

(c) Toy Example Partition One

(d) Toy Example Partition Two

Figure 2.3: Toy Example of an observed network split into two different partitions.

example edges are assumed to be undirected, unweighted and there are no-self loops. The number of groups K is considered fixed, with $K = 2$. This toy example will illustrate how different block structures affect the likelihood of an observed network. The network used in this example and its associated adjacency matrix can be seen in figure 2.3. Figures 2.3c and 2.3d show two example partitions for the two blocks *blue* and *green*, their respective block matrices (using the maximum likelihood estimates $\widehat{M}_{u,v} = Z_{u,v}/N_{u,v}$) can be seen below:

Partition One:

$$\widehat{M} = \begin{bmatrix} 6/15 & 2/12 \\ 2/12 & 1/1 \end{bmatrix}$$

Partition Two:

$$\widehat{M} = \begin{bmatrix} 4/6 & 1/16 \\ 1/16 & 4/6 \end{bmatrix}$$

The maximum likelihood estimate for the block matrix $\widehat{M}_{u,v}$ for each partition can now be substituted into the block model likelihood function 2.1. Hence for the objective function for the two partitions based on the observed adjacency matrix is 8.31×10^{-10}, 2.71×10^{-7} respectively. Since the latter is larger, partition two is a more likely block structure of the graph using the stochastic block model. In general to infer the block structure from an observed network, the model parameters $M_{u,v}$ and g are maximised, for small graphs this can be done using numeric methods over all possible partitions. This becomes more complicated for larger networks, later in this chapter a review of inference methods is discussed.

Poisson Stochastic Block Model

The previous section concerned a block model for unweighted graphs. The *Poisson Stochastic Block Model* (Karrer and Newman, 2011a) in this section is an extension to weighted graphs. In the Poisson Stochastic Block Model the number of expected edges between two nodes is independently Poisson distributed based on only the blocks they are in. The edges in graph G are now undirected, weighted and self-loops are allowed. Note from here consider the number of groups K to be fixed.

Recalling the following notation: Block membership vector g such that, $g_i = u$ if node i is in block u; Adjacency matrix A such that, $A_{i,j}$ is the number of observed edges from node i to node j. Here ω_{g_i,g_j} is introduced, this it the expected number of edges between the groups that nodes i and j respectively belong to. The probability of the graph is now:

$$P(G \mid \omega, g) = \prod_{i<j} \frac{(\omega_{g_i,g_j})^{(A_{i,j})}}{A_{i,j}!} e^{-\omega_{g_i,g_j}} \tag{2.5}$$

With n_u as the number of nodes in block u and $n_{u,v}$ as the number of edges from block u to block v, this can be simplified to:

$$P(G \mid \omega, g) = \prod_{i<j} \frac{1}{A_{i,j}!} \prod_{u,v} (\omega_{u,v})^{n_{u,v}} e^{-n_u n_v \omega_{u,u}} \tag{2.6}$$

The goal now is to maximise this probability based on the unknown model parameters ω and g. It is easier to maximise the logarithm of this probability (which will maximise at the same point). Neglecting constant terms, and terms independent of parameters ω and g, logarithm of the probability to be maximised is shown below:

$$\ln(P(\mathbf{G} \mid \omega, \mathbf{g} = \sum_{u,v} (n_{u,v} log \omega_{u,v} - n_u n_v \omega_{u,u}) \tag{2.7}$$

Maximizing with respect to ω by simple differentiation and neglecting the constants and terms which are independent of the parameters and group assignments, gives the maximum-likelihood estimate:

$$\widehat{\omega} = \frac{n_{u,v}}{n_u n_v} \tag{2.8}$$

Substituting maximum likelihood values $\widehat{\omega}$ this into equation (3)

$$log(P(G \mid g)) = \sum_{u,v} (n_{u,v} log \frac{n_{u,v}}{n_u n_v} - n_u n_v \frac{n_{u,v}}{n_u n_v}) \tag{2.9}$$

Simplifying

$$log(P(G \mid g)) = \sum_{u,v} (n_{u,v} log \frac{n_{u,v}}{n_u n_v} - n_{u,v}) \tag{2.10}$$

Equation 2.10 is an objective function, which is large for *good* group assignments and small for *bad* ones. The Poisson extension of the block model is important for this thesis as the observed network analysed in chapters 8 and 9 contains weighted edges.

How is Stochastic Block Model Assessed

This section now reviews some methods that can be used to compare block structure results to ground truth information to evaluate their success, starting with the *Rand Index*, introduced by (Hubert and Arabie, 1985). The Rand Index is a method of measuring methods of data clustering against each other. The Rand Index is able to determine the accuracy of clustering even when labels are not used as long as both clusters are split into the same number of groups. The Rand Index is calculated as shown below:

$$RI = \frac{a+b}{\binom{n}{2}} \tag{2.11}$$

Where $\binom{n}{2}$ is the binomial coefficient giving number of unordered pairs in a set of n elements, a is the number of times a pair of elements belongs to the same group across two different clustering results and b is the number of times a pair of elements are in different groups across two different clustering results. A Rand Index produces a value between 0 and 1, 1 representing two partitions are the same and 0 representing no agreement on any pair of points in the two partitions. The adjusted Rand Index score is an extension to this, again the closer two community partitions, the greater their adjusted Rand Index, however, the adjusted calculation considers all cluster pairs in contrast to the traditional index, which only considers whether a pair of elements are in the same cluster or in different clusters. Hence it is possible to get a low ARI and a high RI. A larger number of clusters has a higher chance that a pair of items in both partitions are in different clusters. The ARI is calculated as below from a contingency table where a_i is the sum of rows , b_j is the sum of columns, $n_{i,j}$ is the elements of the contingency table and n is the number of elements:

$$ARI = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]\backslash\binom{n}{2}}{0.5[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}]\backslash\binom{n}{2}} \tag{2.12}$$

Normalized Mutual Information is similar criterion for evaluating network community detection algorithms. Normalized information is able to compare community structures even when the two different partitions have a different number of blocks. For the Normalized Mutual Information allow n to be the total number of nodes, $n_{u,v}$ to be the number of nodes in block u in the inferred block structure and in block v in the true block structure. The joint probability that a randomly selected node is in u in the inferred block structure and v in the true block structure is $p(X = u, Y = v) = \frac{n_{uv}}{n}$. Using this joint probability over the random variables X and Y, the Normalized Mutual Information is as follows:

$$NMI(X,Y) = \frac{2MI(X,Y)}{H(X) + H(Y)} \tag{2.13}$$

Where $MI(X,Y)$ is the mutual information and $H(X)$ is the entropy of random variable X.

Note that entropy ($H(X)$) is measuring the expected uncertainty in discrete random variable X, with probability mass function $p(x)$, such that:

$$H(X) = -\sum_x p(x) log p(x) \tag{2.14}$$

Additionally note given X and Y are two random variables jointly distributed according to the probability mass function $p(x,y)$, then their mutual information ($MI(X,Y)$) is:

$$MI(X,Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{2.15}$$

The Normalized Information score is 1 if two block structures are identical and 0 if they are uncorrelated. Normalized Mutual Information index has become the most popular evaluation method (Liu et al., 2019; Funke and Becker, 2019). Despite the popularity of methods such as ARI and NMI, these methods are often criticized for being unable to evaluate the importance of particular groups (Liu et al., 2019). This is important in the context of this work because detecting high risk regions is much more important than distinguishing between low and very low risk communities. To combat this it is important to evaluate the percentage of high risk communities in each detected block. More traditional cluster assessments can also be used to evaluate community detection methods. The detected block structure can be compared with ground truth clusters using recall, precision, accuracy and ANOVA analysis. (Vu et al., 2013; Airoldi et al., 2008; Liu et al., 2019). The NMI score is used in chapters 8 and 9 as a criterion to assess block structures inferred from observed networks of CDR data against ground truth structures collected in chapters 4 and 5.

Degree Correction

One problem with the stochastic block model is that it is not flexible enough to realize patterns which reflect the complexity of many real-world networks. (Karrer and Newman, 2011a) makes the analogy of drawing a straight line on naturally curved data, the line will miss important features and trends of the data. Equally fitting the traditional block model will provide poor results for many complex real-world networks. In fact node properties in general tend to be neglected when recovering community stucture using the block model (Doreian et al., 2005). The particular issue here is that when the stochastic block model is in its standard form it ignores the variation in the node degrees in real-world networks. In standard stochastic block model the expected degree would be the same for all nodes within the same group. It is argued that the stochastic block model is not applicable for real-world networks because real world networks are likely to have nodes within the same communities that have highly inhomogeneous degrees (Karrer and Newman, 2011a; Yan et al., 2014).

The degree of a node will be the sum of the independent Poisson variables, hence the fitted the model will tend to split high and low degree nodes into different blocks. Traditionally the model assumes that edges are places randomly inside each group, hence nodes in the same group usually have similar degrees. This resistance to putting nodes with different degrees in the same block leads to problems when reflected in networks in the real world which are often highly skewed.

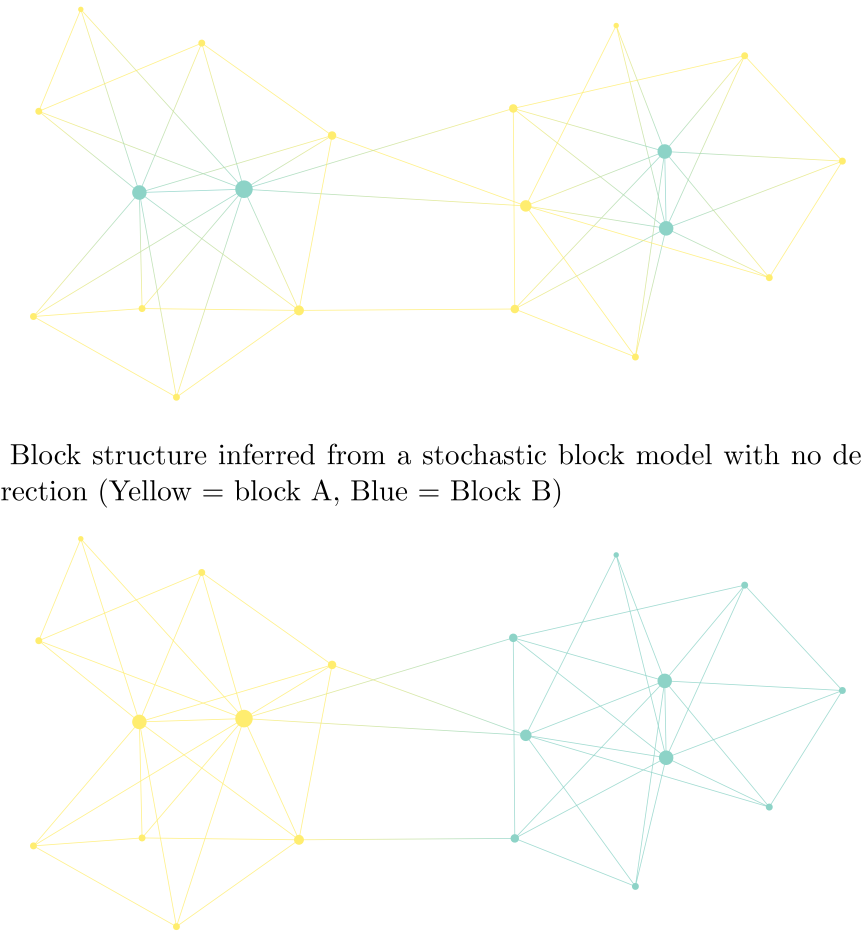
To address these concerns (Karrer and Newman, 2011a) introduces a new parameter θ_i. The graph now depends not only on the expected number of edges between groups and block communities but also on θ_i. The parameter θ_i manages the expected degrees of each node. Given degree correction the graph probability of the stochastic block model becomes:

$$P(G \mid \omega, g) = \prod_{i<j} \frac{(\theta_i \theta_j \omega_{g_i,g_j})^{(}A_{i,j})}{A_{i,j}!} e^{-\theta_i \theta_j \omega_{g_i,g_j}} \tag{2.16}$$

For each block m $\sum_{i \in m} \theta_i = 1$ Hence, θ_i is the probability that an edge connected to the community to which i belongs lands on i itself. The maximum likelihood estimate of θ_i becomes a ratio of the degree of a node to the sum of node degrees in the respective group. The inclusion of node degree variation has been shown to improve scoring criterion (such as NMI) of block structures recovered with the stochastic block model against real world targets (Karrer and Newman, 2011a; Aicher et al., 2014; Peixoto, 2012; Karrer and Newman, 2011a).

This degree corrected stochastic block model underpins the application work in chapters [8, 9]. The model in equation 2.16 is applied to networks made up of different subsets of CDR data in Tanzania to recover community structures. These community structures are then compared with different vulnerability features to investigate the potential use of the stochastic block model in recovering at risk communities.

To visualize the effects of the degree corrected version of the stochastic block model take the following example (illustrated in Figure 2.4). The graph network is an undirected, unweighted and self-loops are omitted. The graph is made up of 20 nodes, 4 of which have a higher degree than the others. The size of the nodes in the graphs below reflects the degree of that node. The block structure of this assigned network has been inferred twice (Using an MCMC optimization algorithm, described in detail in chapter 8). For this example the number of blocks K has been fixed at 2 to best illustrate the skewed results degree variation can create. The first block structure has been inferred (as seen in Figure 2.4a) using equation the Poisson stochastic block model without degree correction described in equation 2.10. The second block structure has been inferred (as seen in Figure 2.4b) using Degree



(a) Block structure inferred from a stochastic block model with no degree correction (Yellow = block A, Blue = Block B)

(b) Block structure inferred from a stochastic block model with no degree correction (Yellow = block A, Blue = Block B)

Figure 2.4: Illustrative example comparing the stochastic block model with and without degree correction

Corrected Poisson stochastic block model described in equation 2.16. In the block structure inferred without degree correction the four nodes with the highest degrees have all been assigned to the one block, and the low degrees to the other block. The block structure inferred with degree correction however has separated the high degree nodes, each block node contains a range of node degrees.

While (Karrer and Newman, 2011a) is the most cited and one of the earliest degree corrected extensions of the stochastic block model, a number of others have also made similar approaches to explore various methods allowing degree inconsistency. (Mørup and Hansen, 2009) and (Reichardt et al., 2011) both allow the probability of an edge to depend on node attributes as well as their group membership. Consideration of degree inhomogeneity within community detection methods and network analysis is an active field in the literature. The degree-corrected model has been shown to improve NMI scores providing improved recoveries of real world block structures (Aicher et al., 2014; Peixoto, 2012; Karrer and Newman, 2011a).

Degree correction will improve the results of networks with highly inhomogeneous degree distributions and have little affect on the performance of networks with little degree variation. The graph likelihoods of the two models will only differ when the degrees vary. Many real world networks have such degree variation hence the importance of the degree corrected stochastic block model. To argue this importance (Karrer and Newman, 2011a) uses two empirical example networks. Firstly, *karate club* network from (Zachary, 1977), this is a social network using members of the karate club as nodes and friendships as edges. The network is built of 34 members from a University in USA. The clubs members are split into two known communities called fractions due to an internal dispute. The traditional block model fails to recover these communities, instead grouping the members into one of low or high degree blocks. In contrast the degree corrected block model successful partitions all but one member into the correct fraction groups.

The second example used is a network of political blogs assembled by (Adamic and Glance, 2005), (Karrer and Newman, 2011a) uses blogs (personal or group web diaries) as nodes and web links between them as edges. The blogs are about US politics and covers a screenshot of the network for a single day in 2005. The blog communities which are known are the political leanings of the blogs. The network has 1222 nodes, is undirected and has high degree variability. Again the degree corrected model outperforms the traditional model which groups all the high degree blogs together. The degree corrected model has a normalized mutual information of 0.72 with the true labellings of the blogs, a significant improvement from the 0.0001 of the uncorrected model. Overall (Karrer and Newman, 2011a) successfully illustrates that the degree corrected model is useful when looking for community detection of complex networks with varying node degrees. Overall the degree corrected stochastic model successfully allows for nodes in blocks to have different expected degrees which reflects many real world networks. The CDR data used in this thesis creates a network of nodes (subward) and edges (CDR connection), the subward nodes in this network are shown in chapters 8 and 9 to have highly varied degrees, for this reason this thesis utilized the degree corrected stochastic block model.

Inference of Parameters, Community Detection

It was mentioned earlier that there are two options for the block model either (i) a network of nodes and edges can be generated given the block structure parameters or, (ii) the block structure parameter can be inferred given an observed network. This work is focused on the inference of block structures, in order to recover ground truth information from observed CDR data. Early studies of *a posteriori* block models, with unknown block structure parameters use numerical estimation methods to maximize the likelihood function. For small graphs and allowing K to be fixed to two groups (Snijders and Nowicki, 1997) proposed direct maximisation and the expectation maximization(EM) algorithm to infer the parameters and thus block structure. Direct maximization involves computing the likelihood function itself and maximising it whereas EM algorithm is an iterative way of approximating the maximum likelihood function. (Dempster et al., 1977) introduced the EM algorithm a method of calculating maximum likelihood estimates in situations where data is missing. This can be applied to the stochastic block model so that g the block membership of nodes can be considered as missing as it is unobserved (Snijders and Nowicki, 1997). However, using the EM algorithm for maximum likelihood estimation of model parameters despite being faster than direct methods, is still not practical unless n is small (($n < 30$)) (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001). There are two main inference approaches to address this, Bayesian inference and point optimisation.

Bayesian inference for stoachastic block models can be split into two main areas, Markov chain Monte Carlo (MCMC) algorithm and variational expected maximization (VEM). MCMC and VEM are both suitable for large networks (McDaid et al., 2013; Airoldi et al., 2008). In literature MCMC is often preferred for algorithmic simplicity while VEM tends to be used for computational efficiency. Given the many adaptions and variants of these methods in relation to the stochastic block model in the last decade both MCMC and VEM are shown to be powerful methods of recovering the model parameters(Lee and Wilkinson, 2019).

The MCMC method alters the membership of each node randomly, and accepts or rejects the assignment with a probability given as a function of the difference of relations at each move. The process moves are reversible and nodes can be classified to different communities over time. Eventually the algorithm after a period of equilibrium will produce partitions with desirable properties (McDaid et al., 2013; Mariadassou et al., 2010a). A simple regular Gibbs sampler was the first MCMC method applied to the block model by (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001). Variants of both Gibbs and more broadly MCMC are used for community detection throughout the literature including but not limited to (McDaid et al., 2013; Tallberg, 2004; Funke and Becker, 2019; Karrer and Newman, 2011a; Lu and Szymanski, 2019). Alternatively the VEM method is split into two steps, first the E-step, this is where the SBM is maximized with variational parameters and then the M-step where the maximization is with respect to the original parameters of the SBM. This is then repeated until convergence occurs and produces estimates for the unknown parameters (Airoldi et al., 2008; Jiang et al., 2009; Vu et al., 2013; Mariadassou et al., 2010a).

Despite the vast amount of work done on Bayesian inference it is arguable that the most popular inference approach is to focus on the most likely set of parameters. Point estimation looks for the peak optimal estimate of the likelihood function. A large benefit of this approach in this work is the simplicity of the concepts. The overall goal for this work is to produce impactful results which can be shared with policy makers and NGOs. In order for results of this work to be meaningful it is vital that the methods are translatable to people who are not from a mathematical background. Simply put the optimization method looks for the most likely values of the parameters. This is done by maximizing the likelihood function of the network graph. This method is easy to comprehend, works well and can be applied to large systems (Peixoto, 2014a; Doreian et al., 2005; Lu and Szymanski, 2019; Peixoto, 2013). This thesis utilizes point estimation through a MCMC optimization algorithm, full details of the optimization approach can be seen in chapter 8.

Extensions of Stochastic Block Model

With the block model being one of the most frequently used techniques in social network analysis it is no surprise that over the last 40 years there have been a number of extensions and adaptions made to the model to solve different problems. (Funke and Becker, 2019) and (Lee and Wilkinson, 2019) present detailed descriptions and comparisons of these extensions and their benefits. This section highlights some of the most significant developments in addition to the degree corrected work discussed in detail above. Before the methods of block structure recovery were decided in this work a thorough review of current model adaption was considered, the extensions mentioned here are not specifically used in this thesis.

An early but key extension was the idea that the model holds for directed networks, hence the matrix of probabilities between blocks does not need to be symmetric. The probability of a edge being present between nodes 1 and 2 can be different from the probability that there is an edge between nodes 2 and 1 (Holland and Leinhardt, 1981; Holland et al., 1983a). Considering the direction of edges in block structure analysis can dramatically improve the results because the model is able to make use of more information (Wang, 1987).

An addition to the Poisson stochastic block model discussed earlier there is an number of approaches to account for edge weights (Peixoto, 2018) looked at the treatment of different kinds of edges such as bounded, unbounded, signed, unsigned, discrete and continuous. (Aicher et al., 2014) introduced a weighted stochastic block model with edge weights drawn from exponential distributions. Including edge weights gives a higher level of information about the networks and as a result can improve results and is beneficial in a number of real world applications (Aicher et al., 2014; Wang, 1987). Another way to increase the level of information about the networks in the model is to consider node information (Tallberg, 2004; Zhang et al., 2019, 2016). Node information can be considered as observable covariates on nodes such as gender, age or income for example.

Another noteworthy development of the stochastic block model is the concept of soft and hard constraints. Up to this point this review has assumed that each node belongs exclusively to one block, this is known as hard constraints. In contrast there are soft constraints, where nodes can belong to more than one block. There are two distinct ways in which the soft constraint version of the model has been developed: Overlapping groups where each node can be a full member of multiple different blocks, and mixed membership groups where each node has partial assignment to a number of groups (Airoldi et al., 2006a, 2008; Peixoto, 2015). The inclusion of soft constraints can be useful in real world applications, think about the social groups of a person for example. It is highly probable that a person is part of a sports team and a music club, the person won't match exclusively the behaviour of either group in a hard constraint model, so is likely to be misplaced (Lee and Wilkinson, 2019; Funke and Becker, 2019).

A final noteworthy extension to the stochastic block model is the work done to include resolution. Smaller blocks can often be missed and become part or larger blocks when looking at large networks. To solve this problem a nested model was introduced which looks at a networks structure at multiple levels (Peixoto, 2014c). This method is referred to as the hierarchical stochastic block model and provides a multilevel hierarchical description of the network.

Dynamic Expansions

It is evident that there are vast developments of the stochastic block model, most existing efforts have been in relation to static networks. In reality many real world static networks are a snap-shot of dynamic networks. Dynamic networks allow edges between nodes to vary with time, they provide a richer level of information on a network. This section now moves to look at the more recent development of the model which accounts for these dynamic networks. There are two ways to look at the dynamic approach: (i) Detecting static communities from observed dynamic networks and; (ii) Allowing the communities themselves to also change over time (this idea that nodes can change their block membership through time is sometimes referred to as label stitching).

This idea of label stitching between two successive time steps has been presented in literature to provide improved statistical accuracy of community detection of dynamic networks (Yang et al., 2011; Matias and Miele, 2017; Yang et al., 2011). These methods looked at discrete time intervals made up of aggregated edges over the time intervals. They presented the idea that nodes could change the block they are in next dependent on the block they are currently in. Unlike the static model, the block structure of each node at any time t is dependent on the block structure at time $t-1$ though a transition matrix which shows the change of block probabilities over time. This method of label stitching was extended to account for binary or weighted edges (Matias and Miele, 2017). Another approach was made by (Xu and Hero, 2013) who used a state space model to describe the evolution of connectivity matrices over time. This approach however only considered networks with unweighted edges. Another similar method is presented by (Wilson et al., 2016), here the nodes of the network were able to change throughout time by relying on the maximum likelihood estimators derived at different points in time, the points these parameter estimate changes are used to indicate a change in node block structure. This approach was shown to successfully recover the dynamic changes in U.S. Senate co-voting networks (Wilson et al., 2016). All these methods allow the block structure itself to evolved over time, this is by far the most explored approach to dynamic stochastic block models.

Alternatively, this section now considers the adaptions of the model where the observed networks are dynamic but the block structure being inferred is static. Simultaneously, (Matias et al., 2015, 2018) and (Corneli et al., 2016) introduce a dynamic version of the stochastic block model based on non-homogeneous Poisson processes. They assume that edges between nodes are counted by a non-homogeneous Poisson process and the intensity of this process depends only on the block membership of the nodes. (Matias et al., 2015, 2018) uses a variational EM algorithm to approximate the likelihood of the model, whereas (Corneli et al., 2016) presents an exact integrated classification likelihood criterion, relying on a greedy search. These methods present the closest work to the dynamic stochastic block model used in chapter 9, however in this work point estimation is used to recover the community parameters and the degree corrected version of the block model is incorporated. Full details of the dynamic block model can be seen in chapter 9.

2.3 Summary

This chapter has presented a review of the literature relevant to this thesis via the following two section: Firstly, a review of current data collection approaches used in developing regions calling attention to their benefits and limits; Secondly, a review of the stochastic block model from its origins to recent developments and applications. This chapter has highlighted the need for novel fine-grained data collection techniques which can provide efficient, reliable and cost effective ground truths about developing regions. Another research gap introduced in this chapter is the detection of communities vulnerable to poverty via the novel application of CDR to both static and dynamic stochastic block modelling approaches.

Chapter 3

Novel Big Data Sources

This section consists of a summary of the data sources used in this work. The two main sources of data are Call Data Records Data (a network of call and SMS interactions) and Ground Truth Data (fine grained data collected describing the social economic health of subwards in Dar es Salaam, discussed in chapters 4,5,6). These data sources provide the basis for the construction and assessment of both the static and dynamic block model investigative work. Additionally the Mobile Money and Landuse data sets are described which provide additional information on the city. Alongside the main data sets, these data sources are used in order to investigate the comparison of theory based proxy models versus machine learnt models.

3.1 Call Data Records Data

Call Data Records Data used in this thesis refers to transactional data shared by a mobile network operator provider in Tanzania. Every time a phone call is made, an SMS is sent, or any other network event occurs information is logged, generating CDR data. Tigo is the network provider who generated the data used here. They are one of the largest network providers in Tanzania and rapidly growing. Despite often being infrastructurally poor, Tanzania is actually data rich. Information such as calling cell ID, called cell ID, calling tower ID, called tower ID and timestamp amongst other information is logged as part of the CDR records. Network providers use this information for billing and network optimization purposes. Tables 3.2 and 3.1 specify some of the relevant information which was provided by the Tanzanian Network provider. The data used as part of this thesis covers a total of 450.2m call and SMS events for 330k mobile phone subscribers taking place across the Dar es Salaam region of Tanzania over a period of 122 days in the autumn of 2014. This CDR data provides information on individuals interactions and thus all the cell ID's are anonymised for sensitivity of information. Given events in this data have associated the timestamps and geographic locations (Tower IDs), a spatial and temporal network is generated and studied in this thesis. It is worth noting that as this is real world data there is noise, for example towers breaking and undergoing repair. The data is cleaned as part of the feature engineering used within various models in this work, see methods sections in chapters 8 and 6 for details. Due to the sensitive nature of this data, the anonymised data is not made publicly available, it was provided through a partnership with the mobile phone operator in Tanzania. The engineered features described in the methods section of this thesis are however made available.

Table 3.3 describes the amount of data available for use in this work. In the blue section you can see the complete data provided. The yellow section then shows the number of interactions available for the 122 day period for which there is have both Call and SMS data. Given that the data provided also has information from outside of the area of study (Dar es Salaam), the yellow section shows the further data reduction to city relevant data. First by considering only connections coming out of towers in the Dar es Salaam

Field Name	**Description**	**Type**
Tower ID	Unique identifier for each tower	INTEGER
Location	Longitude and latitude location of tower	GEOM POINT
Partynumber	The entity who is the focus of the datapoint (for outgoing SMS this is the callingpartynumber, for incoming calledpartynumber	TEXT
Callingpartynumber	Encrypted phone number of person sending the text	TEXT
Calledpartynumber	Encrypted phone number of person receiving the text	TEXT
Callingcellid	Cell cgi from the SMS's point of origination. This can be used for the location of the tower	TEXT
Calledcellid	Cell cgi from the SMS's destination. This can be used for the location of the tower	TEXT
Timestamp	Timestamp of when the connection took place. YYYY-MM-DD HH:MM:SS	TIMESTAMP
Sms length	Length of the SMS message	INTEGER

Table 3.1: SMS Data Summary Features

Field Name	**Description**	**Type**
Tower ID	Unique identifier for each tower	INTEGER
Location	Longitude and latitude location of tower	GEOM POINT
Partynumber	The entity who is the focus of the datapoint (for outgoing call this is the callingpartynumber, for incoming calledpartynumber	TEXT
Callingpartynumber	Encrypted phone number of person sending the call	TEXT
Calledpartynumber	Encrypted phone number of person receiving the call	TEXT
Callingcellid	Cell cgi from the calls's point of origination. This can be used for the location of the tower	TEXT
Calledcellid	Cell cgi from the call's destination. This can be used for the location of the tower	TEXT
Timestamp	Timestamp of when the connection took place. YYYY-MM-DD HH:MM:SS	TIMESTAMP
Call length	Length of call in seconds	INTEGER

Table 3.2: Call Data Summary Features

	Period	**Tuples**	**Size**
SMS	2014:08:01, 2014:12:19	772,166,648	424 GB
Calls	2014:01:01, 2014:12:31	800,157,047	555 GB
	Period	**Out of Dar**	**Closed Circuit**
SMS	2014:08:01, 2014:12:19	333,348,829	206,916,093
Calls	2014:08:01, 2014:12:19	451,326,435	129,470,673

Table 3.3: CDR Data Available.

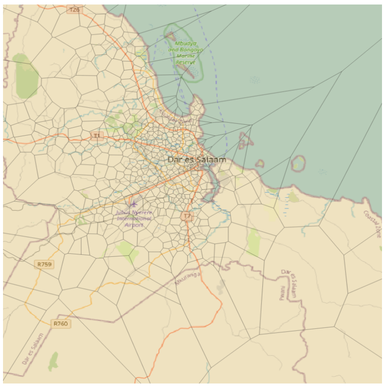


Figure 3.1: Dar es Salaam Voronoi cells

region, then summarising the number of connections in the Dar es Salaam closed circuit where connections are only included if they are both from and to the study region. There are 565 towers in Dar es Salaam, Figure 3.1 is a map showing the voronoi cells of the towers, exact tower locations are not published in this work for privacy reasons. It can be seen that the city centre has a much denser placement of towers than the more rural outskirts. Towers each have a mean average of 229152 calls and 366223 SMS connections going on over the 122 day period with a range of (277,856136) and (279,1184516) calls and SMS respectively. Per day these averages become 1878, 3002 respectively. This can be seen illustrated in the box plots in Figure 3.2. The range of tower activity though the city can be seen in the heatmap in Figure 3.3. Darker colours represent a higher activity level, greens represent SMS and reds represent calls. In summary there is two sparse and dynamic networks here. There are 565 nodes representing the cell towers. 206916093 time stamped edges representing

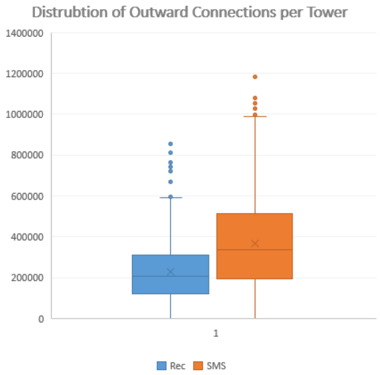


Figure 3.2: Average Total Connections per Tower

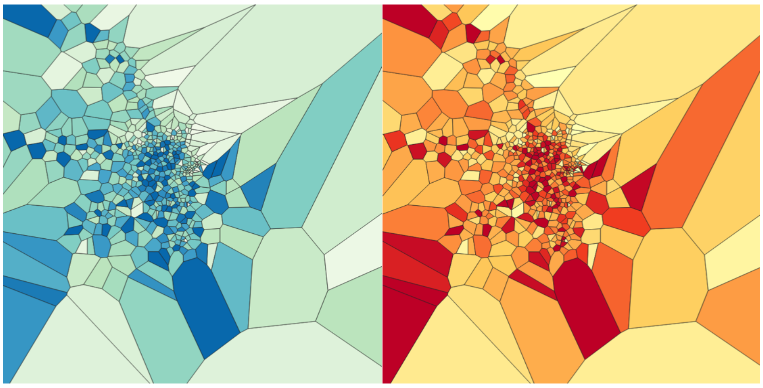


Figure 3.3: Spread of Activity. Left: SMS, Right: Calls Darker Colour represent more calls

SMS connections and finally 129470673 time stamped edges representing call connections.

3.2 M-Money Data

M-Money is an umbrella term for a range of services offered by network operators, which include “sending and receiving money, making savings deposits, bill payments, making non-cash payments and transferring money from one’s mobile phone account to bank accounts and vice versa”. Similar to the CDR data, the M-Money data has been made available to us for research by a large Tanzanian telecommunications provider. The data set contains a sample of the company’s M-Money transaction records for regular mobile phone users (subscribers), businesses (agents) and the network operator itself. Specifically, this work extracts 47.6m M-Money records of approximately 147k customers for the same 122 day period as covered by the CDR data in 2014.

Each record contains a number of attributes collected by the network operator as part of the day-to-day M-Money provision. Those attributes include: • SIM identifier: anonymized identifier for the handset • Date: timestamp of when the transaction occurred • Transaction amount: total monetary amount for the transaction, including service charge • Event type: the category of the good/ service purchased via the transaction • Subtype: a categorization of the business which provided the good/service featured in the transaction • Error code: indicator of transaction success/failure, denoting the cause if the latter • User type: account type of the individual invoking the transaction (e.g. subscriber/agent)

Only users based in Dar es Salaam, who also made use of CDR services were included in this work.

3.3 Landuse Data

A set of 24 inputs were generated from data collected via satellite imagery. These inputs were was augmented with over 750k manual demographic and environmental annotations, including land-use estimates. Annotations were collected as part of the Dar Ramani Huria project (Eichleay et al., 2016) with local community members creating highly accurate maps of Dar es Salaam. An example of the images and annotations present in this dataset can be found in (Torres et al., 2017) having been shared by the authors. The 24 features extracted from this dataset included the total area (in km^2) of the subward, land-use measure (consisting of the residential area, slum, urban, industrial and unused, with equivalent percentages of the subwards that each made), and distances that the subward was from the coast, the central business district, the port and predominant slum and industrial zones.

Ethics and Privacy

Ethical approval for this work has been obtained from the Nottingham University Business School ethical review committee, application reference No. 201819072. In addition, a variety of privacy and ethical limitations have been considered and overcome as follows. **Data Access** There are privacy and security implications involved in sharing CDR data. Raw mobile network data sets are inherently proprietary and private. In order for this type of work to continue, it has been essential to build trust and confidence with network providers. Providing clear research plans and understanding of the

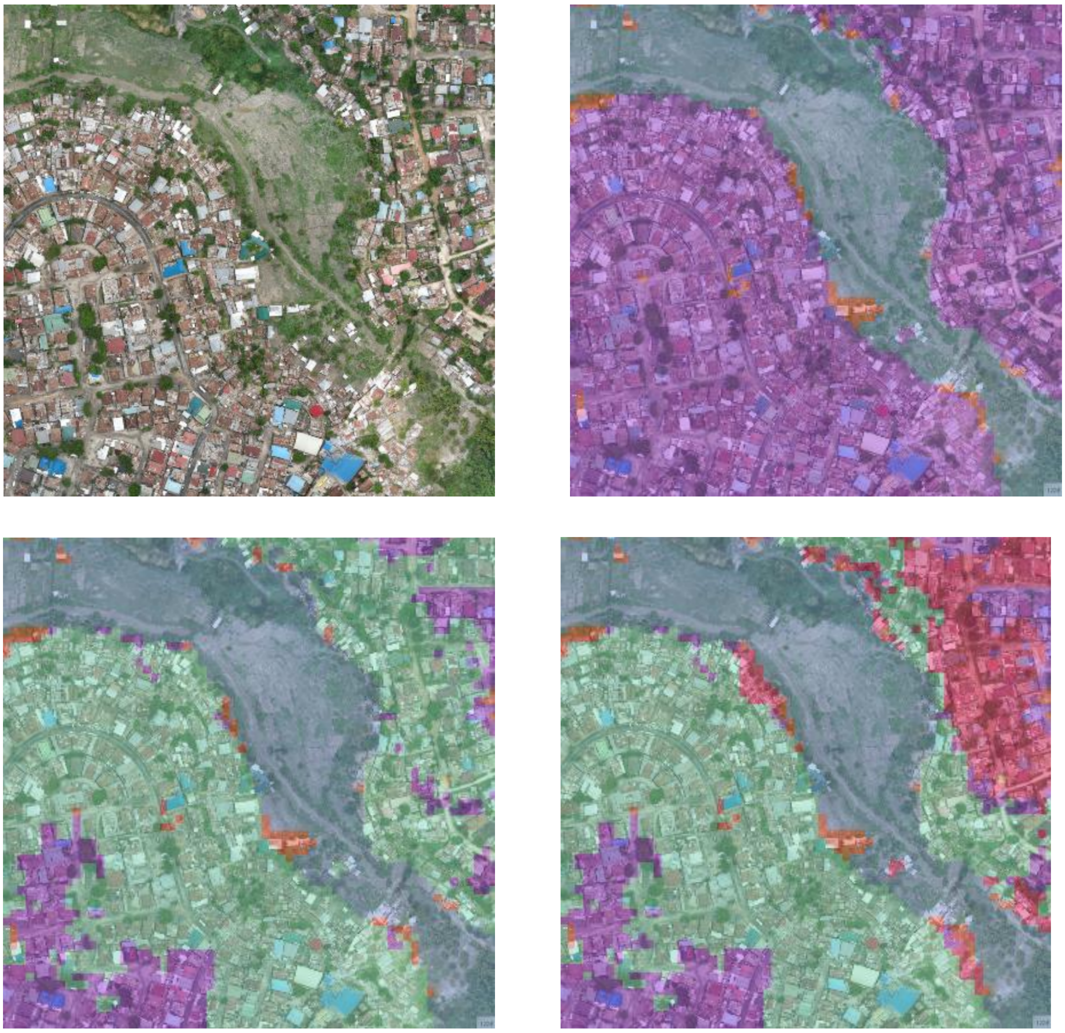


Figure 3.4: Raw satellite image and 6,9 and 10 Land Use classes Respectively

data magnitude requirements have been essential in preparing for future access negotiations with partners. **Data Privacy** A number of actions have taken place in this work to protect the privacy of individuals. The CDR data has been anonymised as mentioned above. Additionally the data is aggregated before analysis (see chapters E, 8, 9 for details on aggregation). Analysing hundreds or thousands of user's data simultaneously is an effective strategy for ensuring further protection of individuals privacy. This process makes the recovery of individual subscribers even harder. Beyond individual privacy there are also concerns for the network operators themselves, including damage of public perception in case of data leaks such as the location of the cell towers. This is another reason the analysis has been done at a Subward level to protect the specific locations of individual cell towers. Both anonymisation and aggregation have also been applied to the survey data collection. Interviewees and participants have all been kept anonymous and data has been collected on a subwards level. Each subward has a range of participants so no individuals can been identified by their subward. Chapter 5 goes into more details on the consent procedures of the data collection. **Ethical Concerns for Intervention** Beyond individual privacy considerations there are also ethical concerns as certain findings can lead to (unintended) structural discrimination in the design of policy interventions. For example using results from CDR data to influence policies can discriminate against people who do not use mobile phones. Additionally Poorly designed interventions for development or safety could lead to the neglect of certain groups or areas. However given the server lack of fine grained data on affluence and related issues in Tanzania (Discussed in detail in chapter 2.1) this work is aiming to support the work being done to improve the level of information policies and interventions rely on.

Summary

This section has provided a short summary of the data sources used in this work. This work benefits from real world data in Dar es Salaam. From CDR to comprehensive ground truths, the work is able to utilize information reflecting a reality in Dar es Salaam to produce meaningful results. The data collection is a substantial contribution of this work and is thus explained in detail in chapters 4,5,6.

Chapter 4

Comparative Judgement Methods to Improve Estimates of Poverty Across Dar es Salaam

It is clear from chapter 2 that data on poverty and its associated characteristics is not only deficient in developing countries, but also essential for policy makers and NGOs to make impact. Looking to address this situation, this chapter develops alternative methods for deriving geo-demographic data, with a focused application to Tanzania in Africa.

As part of this goal I carried out an extensive exercise to develop ground-truths for fine-grained geographical regions in Dar es Salaam Tanzania, jointly funded by the RCUK and the Gates Foundation. This immediately raised challenges in how best to collect ground truth surveys in a context where traditional household surveying is highly impractical due to poor infrastructure, cost and logistical obstacles. In such contexts new surveying mechanisms are required. The ground-truth data collection process was made up of three parts. 1) Grid Survey - Focused on obtaining an overall map of affluence across the city of

Dar es Salaam. 2) Street Survey - Focused on obtaining indicators of both poverty and vulnerability to a range of social issues. 3) Validation Interviews - Focused on eliciting, exploring, and dissecting the challenges of data collection in developing countries. All data collection studies for this thesis have been performed at a Subward level, the smallest governmental geographic regions in Tanzania. There are 452 subwards within 90 wards, in Dar es Salaam, all of which were covered in the studies.

The following three chapters look at the process and results of these three data collection studies. The overall goal of these chapters is to provide knowledge and methodological information which is necessary to underpin a number of potential projects investigating poverty in developing countries as well as the statistical work in chapters 8, 9. The methodological contributions made in this chapter are as follows: • Novel application of comparative judgement to development setting. • Advances in adaptions of 1st work survey techniques in development setting. • Review of challenges of data collection in development setting. • Assessment of theory based indicators of poverty. The knowledge contributions made in this chapter are: • Derivation of fine grained ground truths in Dar es Salaam from both the grid and street survey. The technical contribution made in this chapter is • Systematic reduction of bias in data collection analysis.

Motivation

The motivation behind this piece of work was to create an efficient, reliable and cost effective description of the levels of poverty within Dar es Salaam. Survey data is hard to obtain, being both labour and cost intensive (amongst other issues) and is therefore scarce in developing areas (Xie et al., 2015a). A new method of data collection is required, not only due to these logistical problems implementing traditional surveys, but also due to the well-established response biases inherent to household surveying (Randall and Coast, 2015; Lynn and Clarke, 2002; Kalton and Schuman, 1982). In the developed world there is a much higher variation of data collection methods applied to research, whereas the majority of data collection in developing countries is done using traditional census survey methods. To address the issues of time, cost, response bias and resources required, this work introduces an alternate method to traditional census and household survey methods to create ground truths efficiently. A further goal of this work is to bridge the gap between qualitative and quantitative data, and explore a portfolio of analytical data methods to gather peoples opinions on poverty, in a way which opens up various options for numerical analysis. Comparative judgement models, such as the Bradley-Terry model (Bradley and Terry, 1952), offer a promising solution, leveraging local knowledge, elicited via comparisons of different subwards affluence. Both the data produced in this chapter, and the survey techniques have the potential to underpin a number of different impactful projects.

Method

The Grid Survey consists of comparative judgements of subwards in Dar es Salaam, Tanzania. Comparative judgement offers a way to address the lack of official data and rapid changes in the city, providing access to informed and up-to-date opinions from local citizens. Comparative judgement methods contrast sharply with traditional surveying approaches, in which a respondent might be asked to indicate the affluence level of an area, or their own household income, based upon some arbitrary scale. During the grid survey, participants were given pairs of subwards and asked to choose the richest area, this avoids arbitrary definitions and elicits clear opinions between areas.

The Bradley-Terry model can be used to rank groups of objects by modelling from pairwise comparisons between them (Bradley and Terry, 1952). This model is used in a number of fields including the following: animals abilities to outfight each other (Stuart-Fox et al., 2006); superiority of sports teams (Cattelan et al., 2012; Phelan and Whelan, 2018); ranking chess players (Caron and Doucet, 2012); and for educational assessment (Pollit, 2012). This thesis presents a novel application of the model to subward affluence in Dar es Salaam. The Bradley-Terry model is used infer the deprivation of subwards based on the pairwise comparisons produced in the grid survey. Comparative judgement methods have the advantage that data can be collected in workshops, creating vast amounts of data in limited time.

This work is done using the Bradley-Terry Model however it is worth noting the data could have also been applied to other sorting algorithms. There are many sorting algorithms such as Heapsort (Williams, 1964) and Quick Sort (Hoare, 1961). Heapsort works by dividing its input into a 'sorted' region and an 'unsorted' regions, the algorithm then iteratively reduces the unsorted region by extracting the largest element from it and placing it into the sorted region. Quick Sort works by selecting a 'pivot' element and partitioning the other elements into sub arrays, these other elements are then swapped around the pivot according to whether they are less or greater than the pivot. Quick Sort algorithm has shown to be three times faster than Heapsort (Hoare, 1961). Although such sorting methods could produce a simple ranking order of subward deprivation, the decision to use the Bradley-Terry model was based on the following reasoning; firstly, the

Bradley-Terry model allows not only areas to be ranked, but deprivation levels in each subward to be estimated, secondly the Bradley-Terry model is able to tolerate some errors and disagreements in the data (Bradley and Terry, 1952). Additionally, the model has been widely used in recent research (as seen in the examples above) this presents an opportunity to develop work in an active research area.

As previously mentioned, Dar es Salaam is split into three regions which are made up of 90 wards, each of which is made up of subwards. Subwards are the smallest government official regions of the city. All data collection in this thesis was done at a subward level, the importance of which is detailed below.

It is key to understand the governmental structure and organisation of areas being statistically analysed prior to any data collection. In Dar es Salaam each ward is presided over by a ward officer and each subward correspondingly by a subward officer. Subward officers, however, represent individuals that the community can go to with any issues they need to raise about the area they live in. In contrast, ward officers work more directly with higher government officials on city planning. Figure 4.1 shows the wards and subwards regions of Dar es Salaam. This mapping work was underdone by HOT (Humanitarian OpenStreetMapping Team). A map of subwards had to be physically created, by asking citizens in the respective subwards who their subward officer is and drawing lines on a map that reflect these answers. Surprisingly, and due to the information structures of these relationships, subwards have not previously been officially designated and mapped. This lack of formal organisation highlights the importance of dedicating time to understanding the context of an area before attempting to collect data.

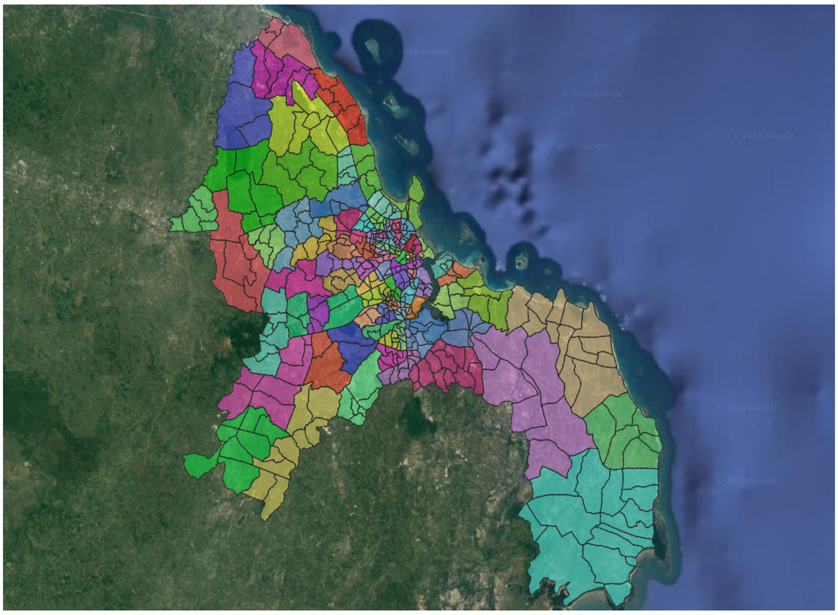


Figure 4.1: Subwards Dar es Salaam, (Colours: Wards, Outlines: Subwards)

Wards are, on average $18.8km^2$, whereas subwards are on average $3.6km^2$. The subwards are the smallest geographic areas used for governance (informal or otherwise) in Dar es Salaam. Most other data collected in Tanzania regarding poverty has previously corresponded to district levels, with no details on the poverty variation of wards let alone subwards. Figure 4.2 shows an example of the diversity which can be seen within wards using the ward Mbagala Kuu as an example, and highlights the important contribution of this work in using subward level granularities for the first time.

Another example of this inter ward variation can be seen from the Charambe ward in the Temeke District. Charambe is made up of 7 subwards each with its own characteristics, and the range of variation is illustrated in Figure 4.3. Breaking down these areas further emphasises the dangers of aggregating statistics at ward levels in such a diverse city: In figure 4.4 we can see an industrial area of the ward; Figures 4.5 and 4.6 both show residential areas of the ward; Figure 4.5 shows an area with spacious, structured living

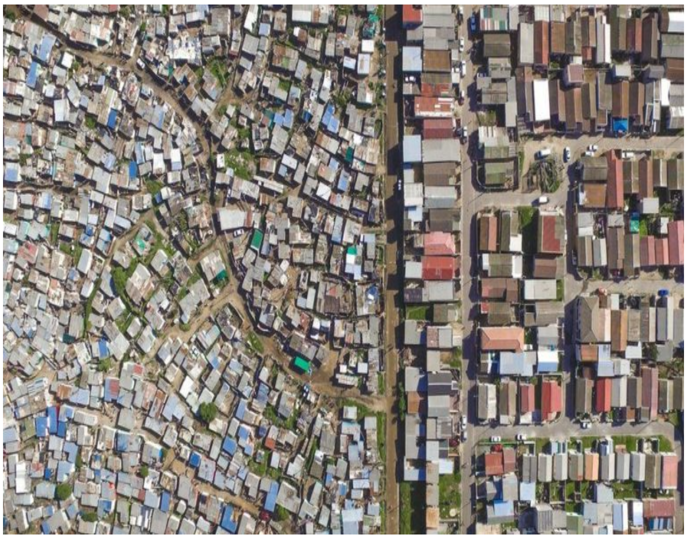


Figure 4.2: Mbagala Kuu Ward, illustrating an example of diversity within wards

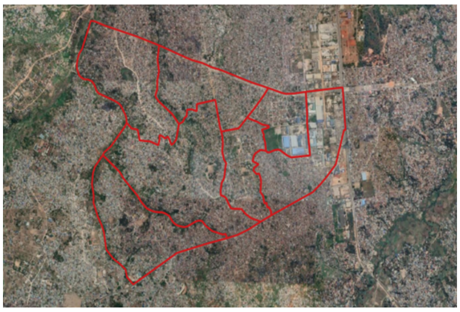


Figure 4.3: Subwards in Charambe Ward in the Temeke District

arrangements; and Figure 4.6 shows an unstructured, dense slum area. Looking at poverty estimates at a ward level would have missed these details, with respondents forced to average out their answers in order to account for existence of both affluent and deprived areas in a ward. Alternatively people may have answered based on instinct of the most notable area in the ward (for example if there is a particularly nice shopping area in a ward, as is the case in the north of the city, respondents may have classed that ward as well off, even if slum areas proliferate the rest of the ward).

Although sorting surveys such as comparative judgements in the past have

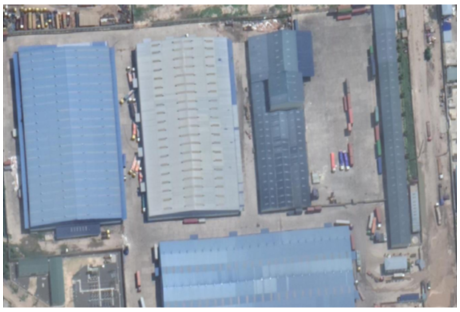


Figure 4.4: Industrial area in Charambe Ward

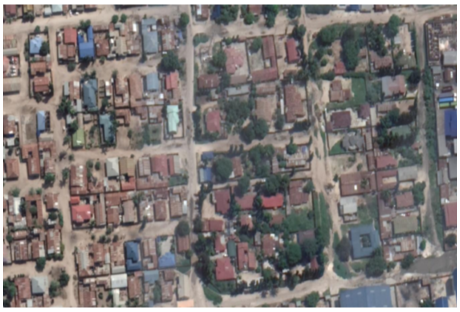


Figure 4.5: Affluent Residential area in Charambe Ward

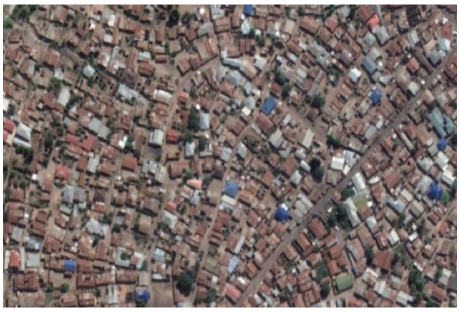
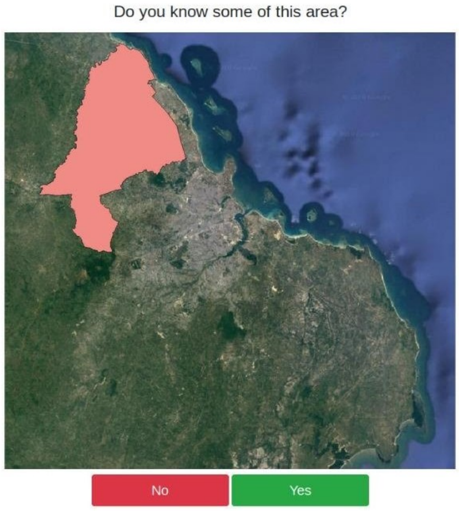


Figure 4.6: Slum Residential area in Charambe Ward

been traditionally carried out on paper, it is more common now for this type of survey to be completed via electronic devices, reducing the chance of human error imputing data afterwards and speeding up the process of data analysis. Additionally, as there are 452 subwards, any analogue implementation of a sorting approach would require a large amount of physical paraphernalia per respondent, unfeasibly adding to costs and, most importantly logistics. As such the grid survey was carried out via laptops, leveraging a web interface through which pairwise comparisons could be made.

The first step for participants was to register their details including basic information such as preferred name, profession and age. Each participant was also assigned a participation number. This information was collected to ensure a fair mix of participant types were used and enables participant trend analysis in the future. This registration page provided participants with the option to skip their personal details, and all pairwise comparison results were kept anonymous in this work.

Prior to completing the grid survey participants were asked to indicate areas they did and didn't know in order to ensure they only made pairwise comparisons of areas they were familiar with. This was an important part of the process to reduce the chance of random orderings in the results. Participants were first shown highlighted large zones of the city and asked to confirm if they knew the area, an example being shown in Figure 4.7. If they knew the area, participants were then shown individual wards in that zone, again to determine if they knew those wards. As well as a highlighted ward, participants were also provided with the ward names on screen for contextualisation, an example of this is shown in Figure 4.8. Participants were then shown the next zone and the process was continued for each of the

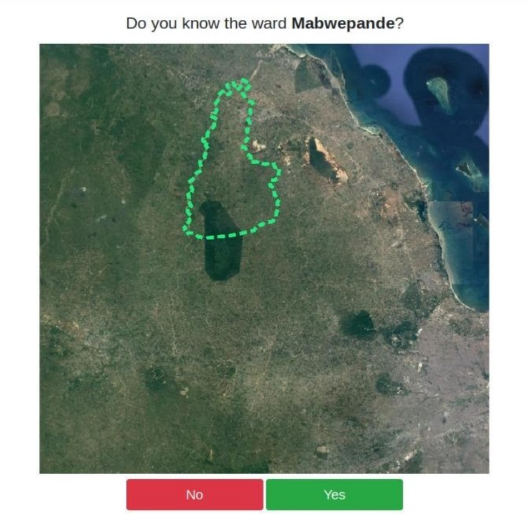


Do you know some of this area?

Figure 4.7: Grid Survey Instructions - Area Check. Participants shown a red area and asked if they know that area.

7 zones (containing 90 wards). Identifying if they knew each subward individually was prohibitive due to the number of them. However, throughout the pairwise comparison process, respondents were able to skip subwards and class them as "unknown" at any time. After registering and selecting known areas participants were able to start their comparisons. For each pairwise comparison, two subwards were presented using their name, and the name of parent ward, visually presented in a map (an example can be seen in Figure 4.9).

Participants were asked to click the subward they believed to be richest and then confirm this choice. If participants were not familiar with one of the subwards they were able to click "Don't know" and the subward was then consequently removed from their comparison list. Equally, if they did know both subwards but were unable to distinguish a choice for the richest one, they could skip the comparison and move onto the next set. Finally there was also an option to click 'previous' if they want to change their mind on a comparison.



Do you know the ward **Mabwepande**?

Figure 4.8: Grid Survey Instructions - Ward Check

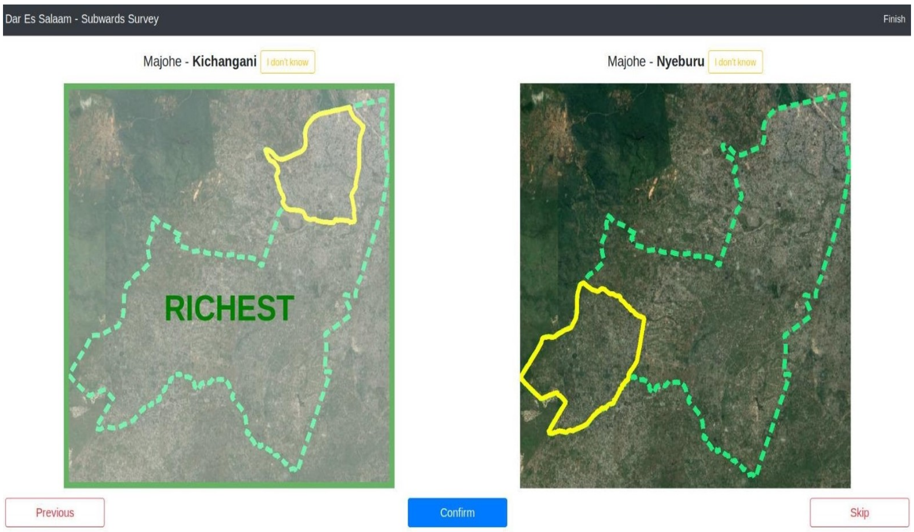


Figure 4.9: Grid Survey Instructions - pairwise comparison Example

Prior to conducting the survey in Dar es Salaam pilot versions were undertaken in order to determine appropriate session lengths, the goal was to collect as much data as possible without making the sessions too long. (Cognisant that if sessions were too long participants were more likely to tire and make random selections.) Two hours was deemed an appropriate session length based on pilot trails. To this end pairwise comparisons were also shown to each participant in random orders, again in case respondents tired towards the end of the session. In addition, repeat comparisons were included in the survey to increase the reliability of the data produced. Another key decision made in preparation for this grid survey was the choice of how to represent each subward for comparison. Initially showing various pictures of the subwards was considered. However, this runs the risk of pushing participants views based on limited features within subwards. For example, considering two buildings from two different subwards, participants are likely to feel pushed to select the subward with the highest quality building as the richest area, despite such images not necessarily being representative of subward affluence. In order to prevent this bias, I decided to show participants a satellite map image of each subward, its position within a ward and its name, providing participants with a clear image of the areas they were comparing, without any leading inputs.

A key distinction in doing a comparative judgement survey rather than a traditional census survey is that there is more opportunity to be particular with participant selection. Where normally participants would be made up of a random selection of household occupants in an area, this survey style presented an opportunity to hand-pick experts with knowledge of the area. When selecting households you are often interviewing random members of the public who have little experience academically assessing the areas they are in.

Household participants are also likely to hold strong views on the area they are in based on the daily problems they are facing. In the context of this work another problem with household participants is that there is little movement particularly in the more rural subwards, thus making comparisons between subwards problematic.

For the grid survey I decided to build a network of local experts to make up the participant team. Here *local experts* can be thought of as people who have high levels of knowledge of the quality and demographics of subwards across Dar es Salaam. The participants gathered were made up of local experts such as taxi drivers and local researchers from the universities, reflecting a network of experts built up for this study over a six months period leading up to the grid survey. The experts came from a variety of sources, some of whom I met through FOSS4G (a conference for open source projects relating to geospatial technologies). Others I connected with through my professional partner, *Humanitarian OpenStreetMapping team*, who have a strong base in Dar es Salaam. Many of the local experts, however, were connected to the project though word of mouth, following the significant amount of time I invested talking to people at the Universities about this work to build up interest. A final source of participants which was very useful were hotel taxi ranks. In Tanzania there a number of different transport options from Boda bodas (motor bike taxis) to Bajaj (three wheeled taxis), many of which are unofficial and unregistered. However many of the hotels and shopping centres have official taxi drivers working their ranks. Performing a snowball recruitment strategy, speaking to people in taxi ranks such as these (See figure B.1a) proved an effective way to engage new, experienced participants.

Once participants were engaged it was important to assess their suitability to the project, the next stage of recruitment was to give them more details on the project and find out more about their level of expertise on the area. Through individual virtual conversations with each participant I determined: how long they had lived/worked in the region; how well they knew the city; and their interest in the project was. During this process it was also important to make sure the local experts were made up of a mixture of ages, genders and professions in order to keep results as unbiased as possible. Building this network of top-quality participants and organising their participation in the workshops took substantial organisation, and was only made possible by regular contact with them to keep up the engagement (predominantly undertaken over the WhatsApp messaging platform).

4.1 Implementation

The data itself was collected via a series of small workshops (examples of which can be seen in Figures 4.10,4.11). Workshop numbers were carefully considered to balance efficiency of the data collection while maintaining the quality. Workshops were limited to a maximum of 15 people, partly due to the limited resource of laptops available, meaning no-one was waiting around for a laptop during sessions. Additionally, smaller workshop sizes allowed for more active support for each participant. This method of data collection is not currently used in this context, so many of the participants didn't have prior understanding of how to complete the grid survey, or use the laptops. It was therefore important to make sure each participant had the support they needed to complete the work. Each workshop group also had a minimum of three research assistants available to support the participants. To ensure participants fully understood what they needed to do, each session started with

Figure 4.10: Example of Data Collection Street Survey - Grid Survey Workshop

a short presentation confirming the task. Each participant was also provided with a worksheet with screenshots of the task explaining the steps. (This worksheet can be seen in Appendix A). All instructions were given in both English and Swahili. In addition to the three research assistants available to support participants, each workshop also had a translator available.

The location of the workshops was also an important factor to consider, each location needed to meet the following criteria: *1) Easily accessible for participants with minimal or no cost. 2) Private space for 15 people to use laptops and work without distraction. 3) Safe, for participants to work on laptops without risk to their personal safety. 4) Access to internet. 5) Affordable, venue hire and food costs needed to meet the financial restrictions of the project.* After speaking with the participants to find which areas of the city were most convenient for them, appropriate venues such as restaurants,

Figure 4.11: Example 2 of Data Collection Street Survey - Grid Survey Workshop

shopping centres and hotels were booked accordingly. [1]

Although one of the major benefits of the comparative judgement approach was the reduction of cost from traditional methods, it was not without its own expenses and resource requirements. As the grid survey was completed on laptops, this necessitated access to 15 laptops which were able to be transported to, and around, Dar es Salaam. This grid survey study was able to leverage resources in the form of laptops and study server from N/LAB, UoN, so the only cost requirements were for incentive rewards for participants. However, future studies must consider such expenses.

When building the participant network there was a strong emphasis on finding people who were interested in the project itself and to engender a sense of local community ownership. As all participants were local experts they predominantly relayed satisfaction in being part of a potentially

[1] Funding Acknowledgement: EPSRC: Horizon Centre for Doctoral Training (EP/L015463), Rights Lab at University of Nottingham, N/LAB EPSRC grant: Neodemographics (EP/L021080/1, 201)

impactful project. Additionally all participants in the grid survey were offered the opportunity to attend training to be involved in later stages of this research (See the street survey, detailed in 9), where there was an opportunity to earn money via involvement. Despite these future project motivations, respondents were still giving up a proportion of their day to travel to workshops and spend a couple of hours inputting their responses into the survey, in appreciation of this, each participant was provided with a free meal and drink after completing the survey. This was budgeted at 10,000 Tanzanian Shilling (roughly £3) per person (sufficient to buy a good meal and drink in Tanzania). In total (excluding cost of flights to Tanzania) this project was budgeted at £550 to run the survey.

4.2 Results

The data was collected over the course of the last two weeks in August 2018. I engaged 174 participants in the project, completing the grid survey in 17 sessions. The sessions took place in Kariakoo, Quality Centre, Mimani City, Kibi Complex and Kiagamboni waterfront. Figures 4.10, 4.11 and B illustrate examples of the workshop sessions.

During the field work the 174 participants provided a total of 75,457 comparisons. Of these, 10,826 (14.6%) were ties (skipped), where a judge was unable to distinguish between the quality of the two wards. Due to the experimental set up, some comparisons could include subwards which judges indicated they did not know. However, if this lack of knowledge was repealed after a subward had already been involved in a comparison, those earlier comparison were retroactively removed. This accounted for 4,362 (5.8%) of judgements, these comparisons were eliminated from the data set. Some judgements were recorded twice, but at different times. Each subward received between 40 and 539 rankings (mean 288). Figure 4.12 shows the map of the subwards illustrating the proportion of participants who knew each area. Figure 4.13 shows the spread of how well-known each subwards was across Dar es Salaam. The proportion of participants who knew each subward was between 0.25 and 0.98 and an average of 0.66. Many of the lesser known regions were in the more rural parts of the city. Figure 4.12 illustrated that many of the subwards with a high known proportion (green) are in the centre of the city. Whereas many of the subwards with lower known proportion (more red areas) are on the outskirts of the city. Minindo subward in Somangila ward for example is in a rural area away from the city centre, and has the lowest proportion, 0.25. Equally Uwanja wa Nyani subward in the Msongola ward is another rural area of Dar es Salaam, with a proportion of 0.29. In contrast, Mnazi Mmojo subward in Manzese ward is a suburban area between the university and airport and has a proportion of 0.98, and Msimbazi Bondeni subward in Mchikichini ward is in the business district and has a proportion of 0.79. Overall, participants knowledge proved easily sufficient for all areas, resulting in a mean of 288 comparisons per subward (the selective process finding local experts to underpin this citizen science approach, enabled this high level of data coverage).

The results produced in the grid survey were then used to derive a map of affluence. This was achieved through a comparative judgement model known as Bradley-Terry Model (Bradley and Terry, 1952). The Bradley-Terry model starts with a set of N objects (452 subwards in this case), whose relative qualities (in this case affluence levels), $\lambda_i \in R$ $(i = 1, \ldots, N)$ are to be inferred from the outcomes of a set of pairwise comparisons (the grid survey). All qualities λ_i are assumed independent quantities for each object.

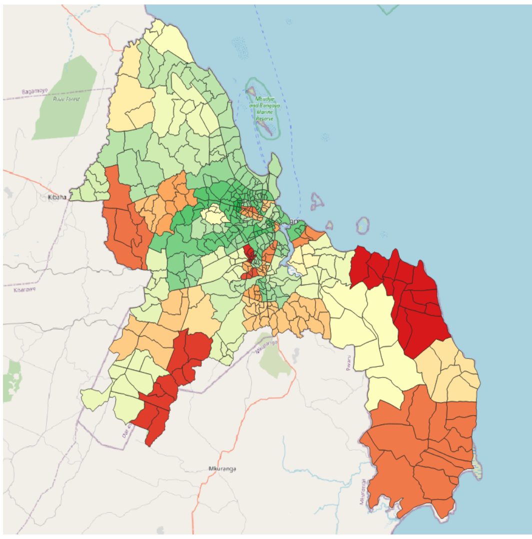


Figure 4.12: Proportion of Participants who knew each subward (Green: high proportion, Red: low proportion

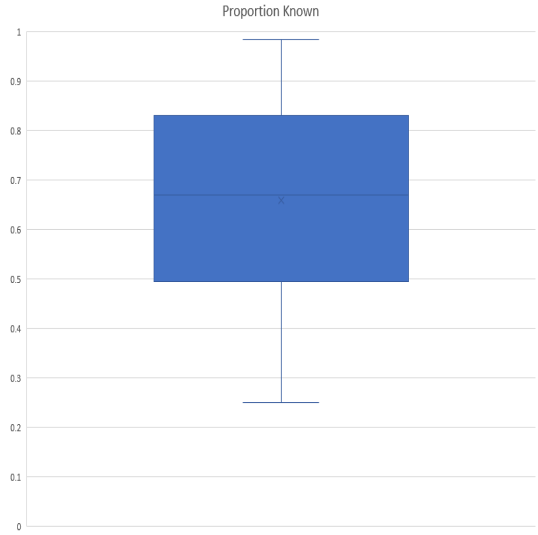


Figure 4.13: Spread of how well known each subward in the grid survey was. (Proportion Known)

For a comparison between object i and object j, the outcome is modelled as

$$Y \sim \text{Bernoulli}(\pi_{ij}); \tag{4.1}$$

in which $Y = 1$ indicates that i is selected over j and $Y = 0$ indicates that j is selected over i; and π_{ij} is the probability, dependent on the qualities λ_i and λ_j of objects i and j respectively, that i is preferred over j. The Bradley–Terry model assumes

$$\pi_{ij} = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)} \quad \iff \quad \text{logit}(\pi_{ij}) = \lambda_i - \lambda_j, \tag{4.2}$$

where $\text{logit}(\pi) = \log(\pi) - \log(1 - \pi)$.

Given K independent judgements (in this case 75,457), the likelihood function for the qualities $(\lambda_1, ..., \lambda_N)$ is the product

$$\pi(\boldsymbol{y}; \lambda_1, \ldots \lambda_N) = \prod_{k=1}^{K} \pi_{i_k,j_k}^{y_k} \left(1 - \pi_{i_k,j_k}\right)^{1-y_k}, \tag{4.3}$$

where $\boldsymbol{y} = (y_1, ..., y_K)$ (Bradley and Terry, 1952).

The maximum likelihood of this likelihood function produces the estimates for all qualities λ_i. This generates a level of deprivation, $\lambda \in R$, for each subward, where the mean subward has deprivation 0. This is due to the necessary constraint that $\sum_i \lambda_i = 0$, so the solution is non-identifiable. A large negative value of λ implies a very deprived subward and a large positive value implies a very affluent subward. Note this has been implemented using `BradleyTerry2` R package (Turner and Firth, 2012).

This ordering of affluence over Dar es Salaam from these λ can be seen in Figure 4.14, where the darker red colours represent more deprivation and poverty, and

Subward	**Ward**	**Subward_ID**	**Lambda**	**Prop_Known**
Idrisa	Magomeni	1	0.63625817	0.88829787
Dossi	Magomeni	2	0.61814458	0.88829787
Makuti 'A'	Magomeni	3	0.56038758	0.88829787
Makuti 'B'	Magomeni	4	0.56931183	0.88829787
Sunna	Magomeni	5	0.19287858	0.88829787
Sisi Kwa Sisi	Makurumla	6	-0.5094609	0.43085106
Kagera	Makurumla	7	-0.2342597	0.43085106

Table 4.1: Highlight of grid survey results, showing seven subwards and their corresponding affluence estimates λ_i as well the proportion of participants who knew these subwards.

the lighter yellow colours represent more affluence.

Following the creation of this map I shared the results with experts (at *Humanitarian OpenStreetMapping Team*) undertaking other field work, responses were highly confirmatory with key features such as the business district and harbour corroborated as rightly showing high affluence, while the known slums were shown to be deprived. All results; and corresponding λ_i affluence estimates, are listed in Appendix C. Table 4.1 shows a small highlight of seven subwards and their corresponding affluence estimates λ_i as well as the proportion of participants who knew these subwards.

Figure 4.15 shows a cluster of (red) subwards over Tandale (a predominantly inner city slum) surrounded by lighter more affluent subwards. This is an example of the fidelity of this work. Tandale represents an area of Dar es Salaam with levels of poverty, with lots of over crowding and informal settlement. Tandale is, however, also surrounded by university areas and adjacent to several more attractive shopping areas. The fine grained nature of this survey was able to detect and highlight this small area of deprivation, amidst highly variant surrounding contexts.

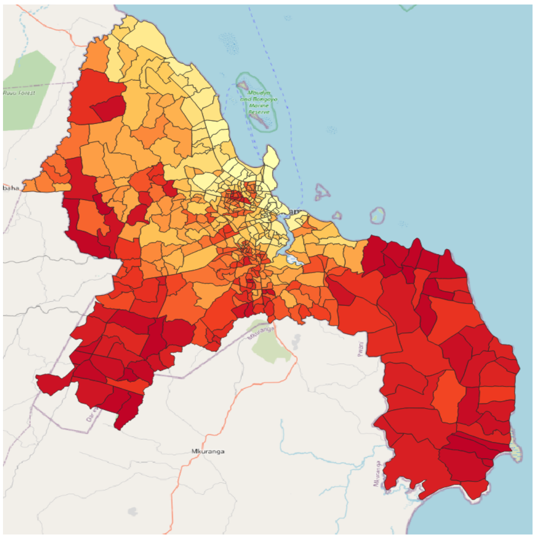


Figure 4.14: Ordering of affluence over Dar es Salaam from λ values produced using the comparative judgement data and Bradley-Terry Model,(the darker red colours represent more deprivation and poverty, and the lighter yellow colours represent more affluence).

A fundamental goal of this study was to create data, information and methods which can be used to underpin a future research, and potentially enable ongoing impact towards attainment of UN SDGs. This is already reflected in the use of the grid survey implementation described above. Additionally the grid survey comparison data has been further used to extend the Bradley-Terry model to incorporate spatial structure into the Bradley-Terry model (Seymour et al., 2020), a direct consequence of the approach established in this study.

The levels of deprivation λ, were also subsequently clustered using k means algorithm (MacQueen et al., 1967; Lloyd, 1982), in order to create categorical ordered results that can be used as alternate targets for detecting communities vulnerable to poverty. These results are presented in Figures 4.16, with clusters shown from red (most deprived), to green (most affluent). These categorical results are useful as an alternative type of ground truth target.

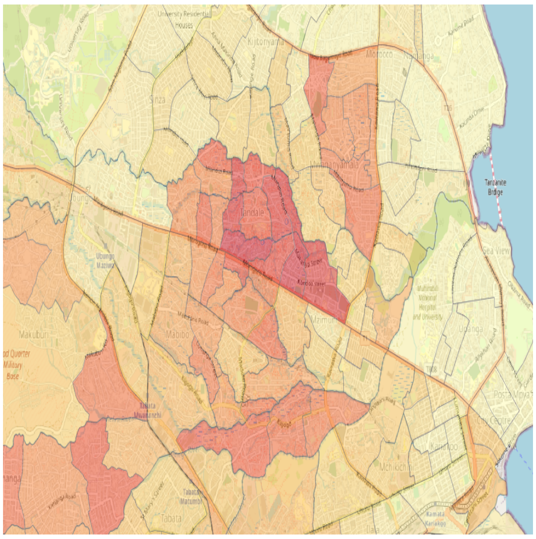
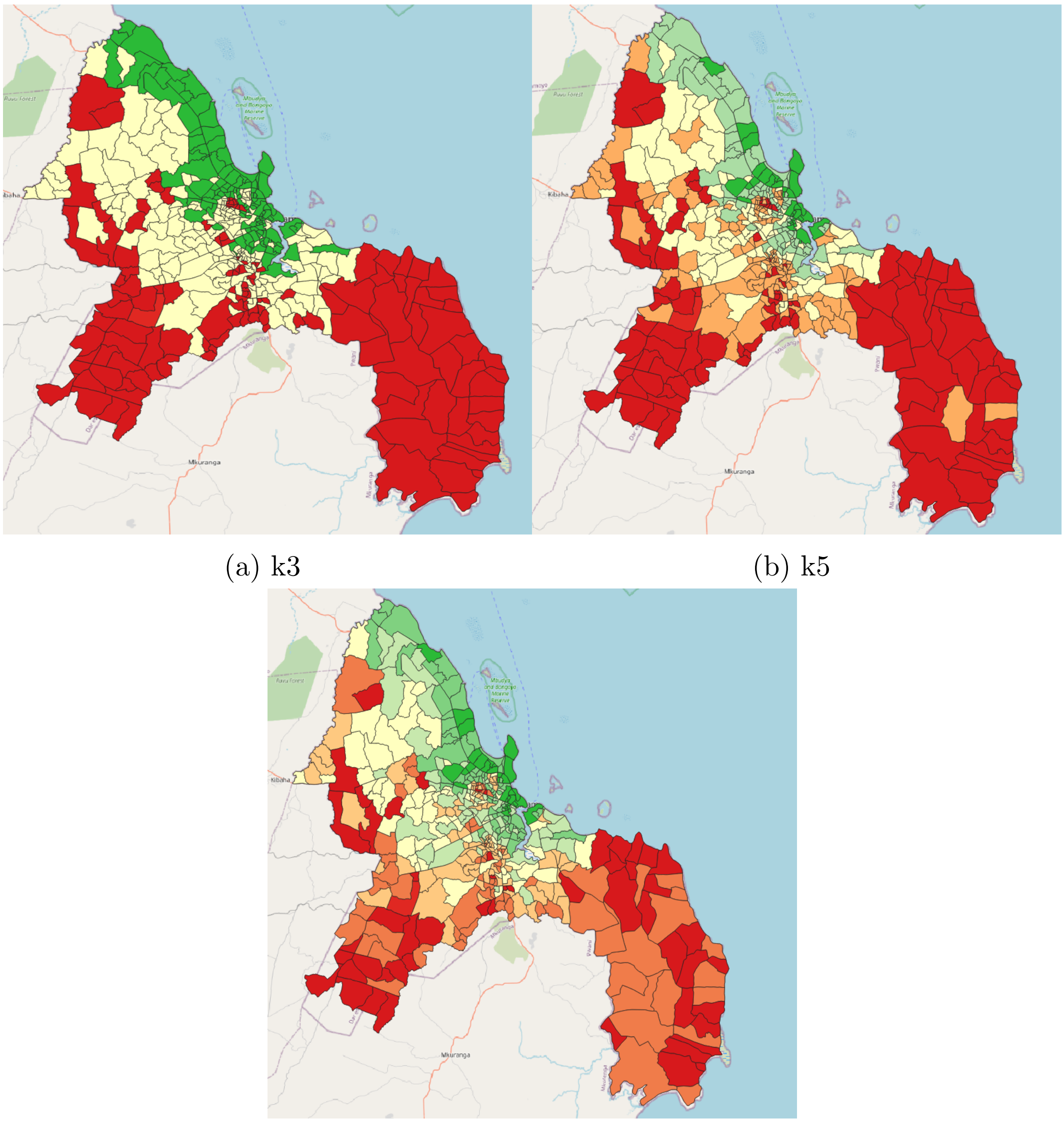


Figure 4.15: Tandale

4.3 Discussion

Determining the level of deprivation in each individual subward is key to designing policies and strategies to alleviate the problems of poverty, especially in the face of limited resources, yet traditional household surveys are simply not viable (Randall and Coast, 2015). Comparative judgement offers a novel way to address the deficiency in data, providing fine grained and up-to-date opinions from local experts. Given the novelty of the comparative judgement methods presented in this chapter, in addition to the data gained there were several methodological lessons learnt from the study. Below is an evaluation of what has been discovered:

- *Efficiency* This grid survey has proved successful in efficiently producing cost effective data. In comparison to traditional surveys, the data produced was collected in a significantly reduced time frame (a matter of weeks), at a fine-grained comprehensive scale at a significantly reduced price. Leveraging local knowledge, elicited via comparisons of different areas' affluence both simplified logistics and circumvented biases inherent to household surveys. The main draw



(a) k3

(b) k5

(c) k7

Figure 4.16: K means clustering of affluence estimates Λ of subwards from the comparative judgement data

backs of this method was the effort and resources needed to create a cohort of reliable local experts, and the ability to only collect one feature at a time. Despite these limitations, the overall process of data collection was more logistically feasible than traditional methods.

- *Participant Assumptions* When planning the grid survey workshop it was important to keep the number of participants in each workshop relatively small to ensure each participant had enough support to complete the work. During the grid survey data collection this resulted in additional benefits, due to assumptions made about participants which were incorrect. Some of the participants, for example, did not know how to use a laptop, or were unable to read maps. This is well illustrated in one example of a participant who had been a taxi driver in Dar es Salaam for over 30 years, he had high levels of knowledge on the city but had never used a laptop or seen a map of his city. Yet, when asked to consider each of the subwards by name he could provide extensive levels of information on the subward and make informed pairwise comparison. He loved seeing the physical maps of the city and was very interested in the work (see figures B.1a, B.1b in Appendix B for images illustrating this situation). Given the methodologies this study introduces is specifically targeted at developing contexts (where maps and technology are not ubiquitously used) this study emphasises the need to not make assumptions about participant experiences and to provide as much support as possible to facilitate the work.

- *Incentives* Having the emphasis of this work on the research itself rather than participant rewards was beneficial. Participants were not taking part for financial gain, having a free meal as part of the work showed gratitude and was important, but there was a clear atmosphere of passion for the city and the research. Engaging local experts who already had a vested interest in the development of the city made it much easier to engage them in the project than traditional participants such as random households and citizens who are often busy and disinterested. This produced a confidence in the results they contributed as they wanted to put the effort in. This is reflected by fact all of the grid survey participants registered interested to take part in the street survey.

4.4 Summary

The grid survey method introduced in this chapter has produced an efficient cost effective method for assessing the description of poverty through fine grained regions. The study successful met the following goals: 1) Novel application of grid style surveying to a development setting. 2) Advances in adaptions of 1st work survey techniques in development setting. 3) The collection of fine grained ground truths in Dar es Salaam. Surveying based around the comparative judgement method is reliable, more affordable and quicker than traditional surveys. It can be conducted at fine grained scales, with rapidly applicable uses (already underpinning other work in Tanzania (Seymour et al., 2020)). There have been many lessons learnt from this work; care needs to be taken to not make assumptions only applicable to western settings and the results benefit highly from local experts with knowledge of rural areas.

Chapter 5

Efficient Collection of Demographic Data In Dar es Salaam via a Street Survey

Motivation

The collection of geo-demographic data for understanding cities has typically been undertaken by census and other large scale surveys. Providing the basis for a wide range of activity including city planning, market intelligence and policy the information is routinely collected by both governments and market intelligence companies within developed countries at significant cost. In less developed countries, this cost means that information is typically collected less frequently and at a lower fidelity if at all. One goal of this thesis is to develop methods and knowledge to narrow the lack of fine grained demographic information in developing areas. Although the grid survey, (detailed in 4) and its comparative judgement approach successfully provides a reliable and cost effective description of the levels of poverty within Dar es Salaam, it does not provided detailed descriptions of the subward characteristics. An underlying goal of this work is to investigate the covariates which can be used to model social deprivations, without the need of expensive surveying. However to simulate such models prior to explanatory analysis, accurate ground truths must be found as such an extensive survey to develop ground-truths for fine-grained geographical regions in Dar es Salaam in Tanzania was undertaken. The street survey is similar to traditional household surveys, asking for detailed accounts of socio-demographic information on a range of topics relating to each subward and its residents. There are two novel distinctions, however, between the street survey and standard household surveys. 1) In the street survey participants are questioned about their subward, rather than personal circumstance, to reflect the absolute need to respect sensitivity and the goal of eliminating response bias. 2) The street survey is carried out over each and every one of the smallest governmental areas rather than taking random samples to represent large regions or zones. This was done to uncover a novel level of fine grained information. The street survey aims to underpin a number of projects (such as the investigation of systematic bias and the assessment of theory based indicators of forced labour as seen in chapter 6, as well as the detection of vulnerable communities via stochastic block models seen in chapter 8 and chapter 9) via new knowledge of Dar es Salaam and novel data collection methods. Given the size of the file a full copy of the street survey results are available via the following link rather than the appendix https://www.maddyellis.co.uk/new-page

Brief summary of activity time scales

The following chapter goes into detail describing the survey procedures and initial results, these activities happened over the following time scale. In the last two weeks of August 2018 the grid survey was performed, the in-field street survey pilots were carried out in Dar es Salaam and 190 local facilitators were recruited. In April 2019 the 190 local facilitators attended a training day for the upcoming street survey. During May 2019 the local facilitators conducted surveys for the first 176 subwards. During August 2019 the street survey was concluded when the remaining 276 subwards were completed. By the end of August 2019 a total of 3668 surveys had been collected from each of the subwards.

Initial Steps

In addition to the literature studied, (see chapter 2.1) prior to the creation of this street survey I attended community meetings (See Figures B.1d,B.2a,B.2b in appendix B) in Dar es Salaam, in order to connect and collaborate with in-country experts. I also met with data collection experts in the UK and Tanzania, subward officers and local residents in order to fully understand the context of the work prior to planning the survey. These initial meetings enabled this extensive survey to be created in direct collaboration with in-country translators, data collection experts active locally (*Humanitarian OpenStreetMap Team*) and modern-slavery domain experts (*Rights Lab, University of Nottingham*). Meeting with modern-slavery and data collection experts in the UK helped to ensure covariates theorised in the literature were covered in survey questions. The questions selected covered a wide range of SDG related indicators, from education and social vulnerability, to local medical facilities and road conditions. It was important to explore a range of topics in the survey questions, in part because a broad scope would help identify which ground truth features can be detected using the application of big data to the block model. Additionally, because of the unique scale of this fine-grained survey, resulting data has the potential to underpin a number of external projects, as

Figure 5.1: Street Survey Training Day

well as the studies it informs within this thesis.

Facilitators in this study were people who live in Dar es Salaam and were trained to collected survey data from participants in relation to this project. There were two main stages in the recruitment of facilitators for this study: 1) Engagement recruitment from the grid survey participants. 2) Recruitment from data collection experts active locally (*Humanitarian OpenStreetMap Team*). This recruitment process created a team of 190 facilitators who had prior knowledge of the project and varying levels of experience collecting data. In April 2019, a training program in survey data collection methods was provided, where I hosted the facilitator team (see Figure 5.1) for an initial training day. This training day provided opportunity to also allocate all facilitators into teams to organise future communication for training, work allocations and stipends. Each team had a team leader whom I would pass key communications through. Establishing an organisational structure of this fashion was essential given the size of the team, but also encouraged co-ownership of the work.

Overall, this initiative was underpinned by an iterative process of survey development with subject and area experts. Initial prototyping, five in-field pilot surveys and numerous survey development meetings with facilitators, survey experts and other specialists were carried out from August 2018 until May 2019, at which point the final street survey commenced. Figures B.2c and B.2d in appendix B show examples of meetings carried out following pilot surveys. This iterative process provided essential feedback to improve and develop the final street survey method.

Initial Reflections

This section details the initial feedback and key messages resulting from the pilot surveys, meetings and training sessions, which led to the final street survey structure. The key initial reflections were as follows:

- The survey was piloted using both paper surveys and a mobile app version. Facilitators found walking around with lots of papers drew excessive attention and made respondents less comfortable participating whereas utilising an app in contrast felt more discrete. *Open data kit* [1] is an open piece of software which can be used for data collection surveys, and which offers several benefits. In particular, it is quicker and clearer for users. There is no need to manually input data (reducing labour time, effort and error). Questions can be translated into multiple languages. There are numerous options for things such as filtering - having, for example, specific questions appear based on the answer to previous questions. The app can be installed with wifi or simply via USB, then the questionnaire can be filled out multiple

[1] https://getodk.org

times without wifi or data while offline. Results are then automatically downloaded when connected to the internet.

- One thing which emerged from the pilots and specifically review meetings was that local community members within developing regions are becoming disillusioned, and ultimately apathetic to surveys which are representing large parasitic organisations that bring inferences of change. Citizens reported that they invest time in the surveys, yet do not see the changes that they seek actually happening. This makes them reluctant to do more. However, by taking a more gentle and honest approach (for example having local researchers as survey co-owners, rather than westerners who don't live locally) and taking the time to describe the project while having a more grateful attitude increased the likelihood of positive participation. Sitting down with people and saying, 'thank you for helping me let me get you a cold coke cola' was a simple gesture, yet one that helped to maintain a positive relationship with people in the community. The need for community engagement is proving increasingly valuable in this kind of work, and for this study encouraged a *Gratitude beverage stipend* for each facilitator to buy participants soft drinks.

- Facilitators tended to find it easier to approach younger male participants as they were more confident and easier to engage. Obtaining a range of gender and age group opinions is, however, essential to achieve unbiased results. To encourage this diversity it became clear that taking time to talk to the more confident participants and ask them to introduce other potential participants of different age groups and gender would be beneficial.

- The safety of facilitators is of the upmost importance in data collection work. Due to the sensitive nature of some of the topics being investigated it was important to put various safety measures in place. Things which help to keep participants safe include: • Facilitators travelling and working in pairs. • Careful considered wording of questions, avoiding questions on personal experience and circumstances. • In country team leader support system available. • Clear written permission from the subward officer (people do not feel comfortable talking unless they know sanction has been provided by the subward officer). • Travel stipend to ensure facilitators have enough funds to get to and from their target subwards.

• During the piloting phase, one group of facilitators was found to be producing unreliable submissions. Some individuals where rushing surveys, skipping questions and not fully explaining the project to participants. Some facilitators were even found to be fabricating responses, filling them in from home prior to submission rather than visiting subwards and engaging with participants. In order to prevent this kind of behaviour these are some of different quality checks which can be incorporated: • Automatic GPS readings on survey submissions. • Automatic start and finish time of survey submissions. • Facilitator pairs planned carefully to mix research experience, allow more experienced reliable data collectors to support newer facilitators. • Clear instructions and regular contact with facilitators. Although the remaining facilitator team proved to be very reliable during the pilots, some facilitators had either more confidence or experience collecting data. Facilitators confidence engaging with participants has a direct effect on their ability to communicate the project and survey questions. For this reason it was deemed essential to ensure no subward had data collected by only one facilitator. Each subward results needed to be produced by multiple facilitators.

- Location of subwards and name of subward officers cannot be assumed as known. The physical location of subward boundaries is not commonly known, with people simply interpreting subwards as households which sit in the purview of subward officers (Mjumbee). Subward officers are the community leaders who communicate residents problems to the government. In some areas these officers are elected by the community in other areas officers take the role after a relative passes away. As such subward geometry cannot be assumed to be known by the general public particularly on a map. It is therefore important to ensure participants know questions are in relation to areas covered by subward officers. Equally, facilitators emphasized that it would be useful to go around with a map, in part to ensure they are in the correct place, but also so they can show people where they are on the map and where their subward is. This was not only useful for context, but generally also invoked much interest from participants.

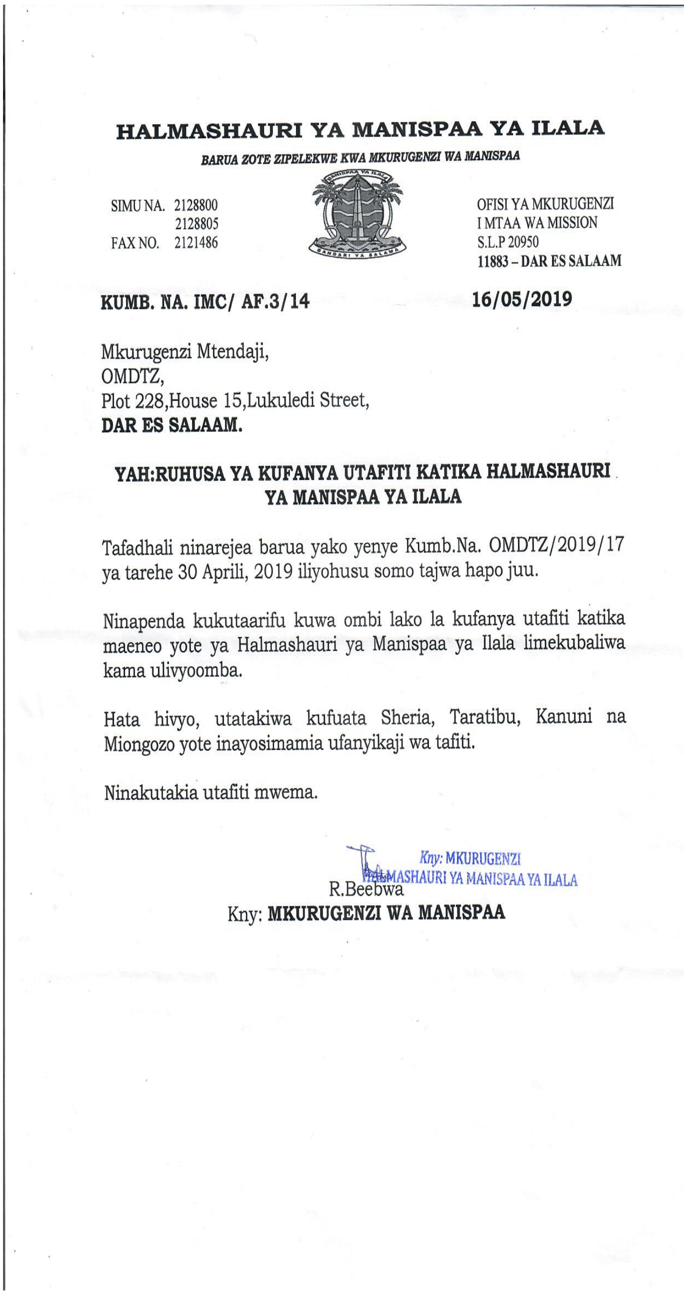
Having multiple pilot surveys and regular feedback from facilitators and experts was essential in the development of this survey. The street survey covered a range of socio-demographic, infrastructural and environmental factors relating to each subward and its residents (a full copy of the survey can be seen in appendix D).[2] Due to the highly sensitive nature of some of these topics, particularly those around illegal activities such as forced labour, much diligence, time and ethical consideration was put into the process of framing the questions. Meetings with experts in data collection was an essential part of this framing. Having the pilots in-country allowed specific area knowledge to be found, such as the legal necessity for written subward

[2] Survey questions where initially established in liaison with UoN Rights Lab and then further iterated with the local surveying team

officer approvals and the age structures in school systems. The initial reflections listed here reflect key highlights from the information learned through this iterative process. Although future studies will benefit from some of the generalizable insights produced, area specific pilots should always be utilised to assess how a survey is received in a specific geography.

Method

Facilitators completed at least 8 surveys in each subward. The facilitators spoke to a mixture of people living and working in each subward covering different age groups and genders. The survey questions ask the participants their perspectives on conditions in that subward in comparison in comparison to the rest of Dar es Salaam. The questions cover a number of different indicators of poverty and vulnerability as well as demographic information and behaviours such as phone ownership, family dynamics, forced labour, education levels, housing conditions, environmental hazards and facility and transport availability amongst others (again a full copy of the survey can be seen in appendix D. The survey questions are made up of over 70 features split into three sections. A section on bias, checking things such as age and gender of the participants to encourage fair results. The main section investigates indicators for vulnerability and poverty. The final component is a participant confidence section. This includes questions which refer to the participants knowledge of the area, including questions such as *Do you live in this area, how long have you lived in this area, do you work in this area?* The results of these questions allow derivation, for each participant, a confidence score from 1 - 12 which can then be used to assess confidence in their responses.



HALMASHAURI YA MANISPAA YA ILALA

BARUA ZOTE ZIPELEKWE KWA MKURUGENZI WA MANISPAA

SIMU NA. 2128800
2128805
FAX NO. 2121486

OFISI YA MKURUGENZI
I MTAA WA MISSION
S.L.P 20950
11883 – DAR ES SALAAM

KUMB. NA. IMC/ AF.3/14 **16/05/2019**

Mkurugenzi Mtendaji,
OMDTZ,
Plot 228,House 15,Lukuledi Street,
DAR ES SALAAM.

YAH:RUHUSA YA KUFANYA UTAFITI KATIKA HALMASHAURI YA MANISPAA YA ILALA

Tafadhali ninarejea barua yako yenye Kumb.Na. OMDTZ/2019/17 ya tarehe 30 Aprili, 2019 iliyohusu somo tajwa hapo juu.

Ninapenda kukutaarifu kuwa ombi lako la kufanya utafiti katika maeneo yote ya Halmashauri ya Manispaa ya Ilala limekubaliwa kama ulivyoomba.

Hata hivyo, utatakiwa kufuata Sheria, Taratibu, Kanuni na Miongozo yote inayosimamia ufanyikaji wa tafiti.

Ninakutakia utafiti mwema.

Kny: MKURUGENZI
HALMASHAURI YA MANISPAA YA ILALA

R.Beebwa
Kny: **MKURUGENZI WA MANISPAA**

Figure 5.2: Example Ward Officer Permission Letter

It was a significant challenge to ensure questions were dually comprehensive whilst fitting the target environment and cultural context. Participants were strictly questioned about their subward, rather than personal circumstance, to reflect the absolute need to respect sensitivity and reduce response bias. If participants did not wish to answer a question, no response was recorded, and support structures were put in place, supported by local partner *Humanitatian OpenStreetMapping Team* in case of respondent vulnerability. Although the survey includes a range of question styles such as *open* questions and *select multiple* questions, most questions generally took a categorical or Likert scale form, as the following examples demonstrate:

- *Overcrowding is a problem in this area (Msongamano ni tatizo kwenye mtaa huu?): 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree'*

- *Which type of roads are most common in this subward? (Ni aina gani ya barabara zimetawala katika mtaa huu?): 1. Tarmac 2. Wide, flat dirt track 3. Wide, uneven dirt track 4. Narrow, uneven dirt track*

- *There are people in this subward being forced to work against their will. (Baadhi ya watu katika mtaa huu hufanyishwa kazi bila ya hiari yao) : 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree*

Letters of data collection approval were collected from government officials from each region before data collection started. An example approval letter from a ward officer can be seen in Figure 5.2. In addition to the approval letters the facilitators all had a copy of an information letter about the work in case any participants or officials had any future questions, this can be seen in appendix H.

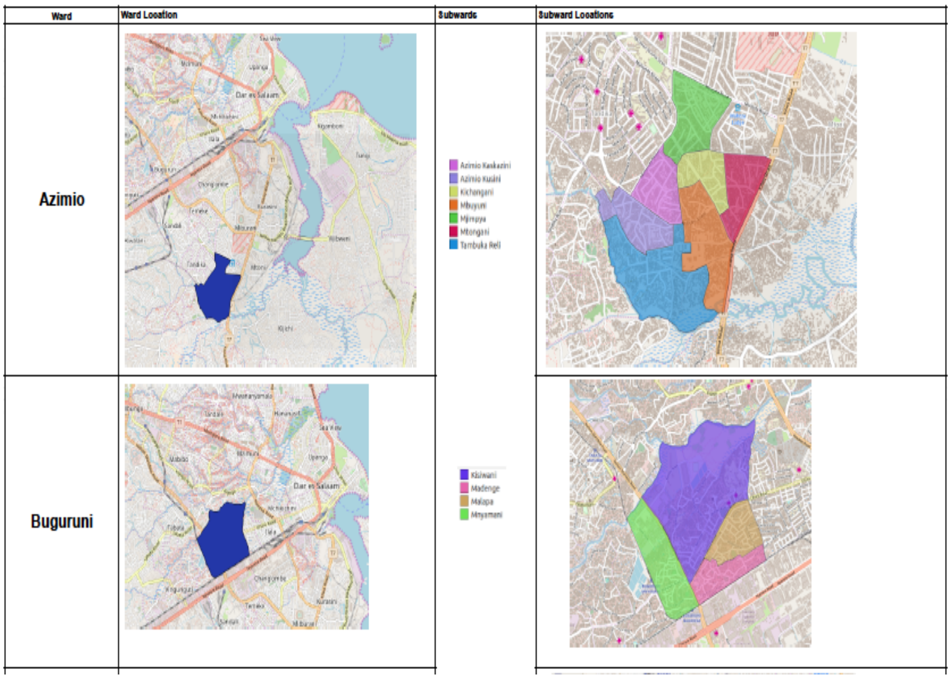


Figure 5.3: Example Facilitator Subward, Ward Maps. Dark Blue showing the subwards, transparent colours showing the subwards.

Each facilitator collected data via the mobile-based surveying app *KoBo Toolbox*[3], collecting street-survey responses from at least 4 individuals per subward. Each subward was scheduled to have at least 8 responses from participants via at least 2 different facilitators. The facilitators spoke to a mixture of people living and working in each subward covering different age groups and genders. Facilitators were given booklets containing team leaders contact information, ward and subward maps (See Figure 5.3), street survey instruction sheets (See appendix F) and a hints and tips document (See appendix G). Facilitators were also given a travel stipend and participant gratitude beverage stipend before each excursion, then a further work stipend after their required surveys were submitted, checked and approved. Given the quantity of facilitators and surveys to be completed (190 facilitators completing a total of 3668 surveys) regular communication and organising

[3] Kobo Toolbox is a data collection tool (https://www.kobotoolbox.org) built using ODK (Open Data Kit) ecosystems.

was a key part of this process.

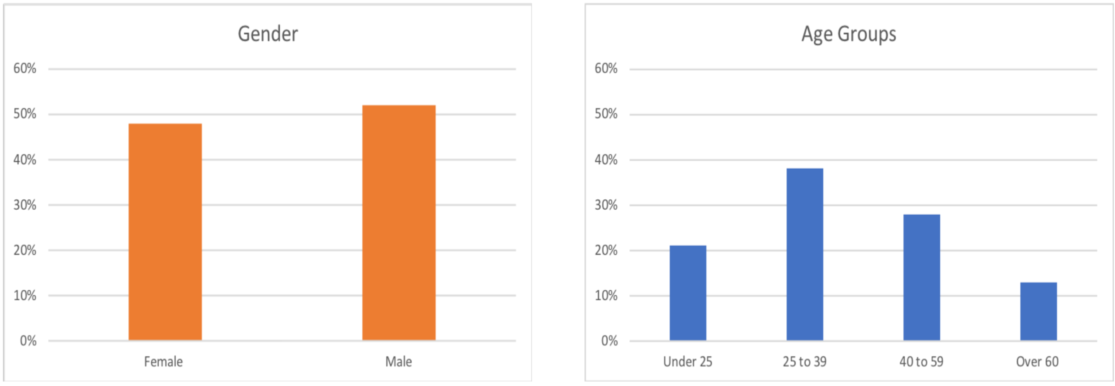
5.1 Summary Results

Following pilot phases, the final street survey was conducted between May 2019 and August 2019, with 190 facilitators submitting a total of 3668 surveys from 451 subwards. During the process of this survey one subward was dissolved into other subwards (administratively by Dar es Salaam City Council). As the defined subward areas depend on subward leadership and subward leadership changes, it is not uncommon for subwards to dissolve or split over time. Each subward has successfully obtained a minimum of 8 responses. With travel stipends, participant drink stipends and facilitator work stipends the survey cost a total of £5450. Although this is considerably more than the £550 costs to run the grid survey, this still reflects a relatively cost effective total for such an extensive survey.

This section details a selection of summary statistics from the survey results, there is over 70 features in this survey so a complete summary is neglected but it is important to note that the street survey is extensive and has the potential to underpin many future projects. Figures B.3a - B.6c in appendix B show a sample of photos of facilitators collecting data, and Figures B.6d - **??** in appendix B show a sample of 14 of the 3668 photos taken as part of the survey by participants.

Participant Information

Confidence 86% of participants scored 12/12 on the confidence questions. (Confidence questions such as: *Do you live here, how long have you lived here, do you work here, do you have relatives here*). 12% scored 11/12 and



(a) Street Survey Participant Genders

(b) Street Survey Participant Ages

Figure 5.4: Street Survey Participant Information

the remaining 2% scored 8 or over. When training the facilitators the need to engage participants who had high levels of knowledge of the area was emphasized, and this has been reflected in the results. There was no need to remove any submissions from the result data set because all participants indicated good working knowledge of their subward.

Age and Gender Following training advice, facilitators engaged a diverse range of participants. Of the engaged participants, 48% were female and 52% were male as can be seen in Figure gender 5.4a. Facilitators also engaged a range of age groups, participants where made up of the following groups: 21% Aged under 25, 38% aged between 25 and 39, 28% aged 40 to 59 and 13% aged 60 and over. This is illustrated in Figure 5.4b. This age and gender diversity is consistent throughout the subward submissions.

Summary Information

Underage Marriage The map in Figure 5.5 shows red in the areas which at least one participant holds the opinion that underage marriage does happen in their subward. There are a number of other questions on marriage in the survey, eliciting the existence of arranged marriages, average age of marriage and youngest age of marriage (Some of which is as low as 13 in certain areas).

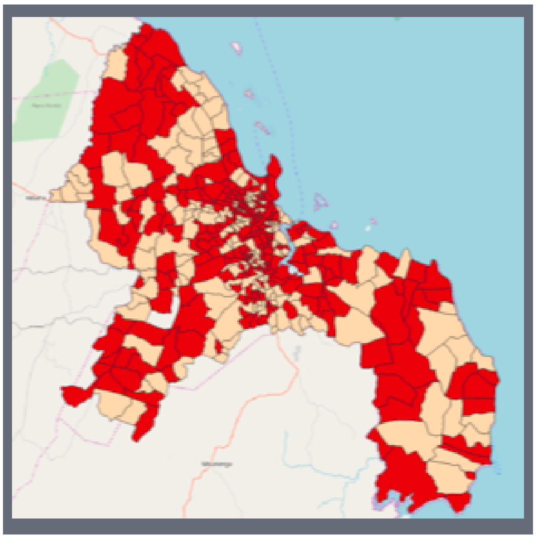


Figure 5.5: Street Survey Results Underage Marriage. Red represents the areas which at least one participant holds the opinion that underage marriage does happen in their subward

31% of participants said that marriages below the age of 18 occur in their subward.

Housing Arrangements

The survey has also indicated the spread of perceived types of living arrangements in Dar es Salaam. In equal majorities most people are either living with close family (46%) or extended families (45%). There are very few people living on their own or in shared accommodation (9%), most individual households were found near the business district in the centre of Dar es Salaam.

ID Ownership

Figure 5.6 shows a map of the areas which have the highest likelihood of people not owning personal identification documents. Figures 5.8 and 5.7 show the distribution across the city of likelihood of ID ownership from teenage females and males respectively. These maps are on the following scale: white (Highly

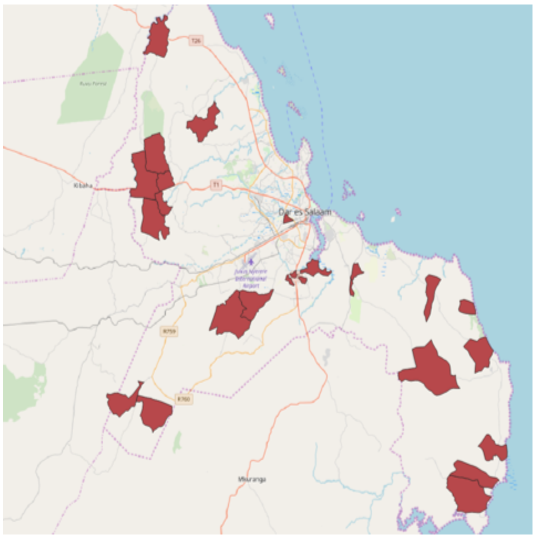


Figure 5.6: Street Survey Results Maps ID Ownership. Red areas show the places people are least likely to own their own ID.

likely they do not own personal identification) to red (Highly likely they do own personal identification). Young females are less likely than males to own personal identification in the more rural areas of Dar es Salaam such as Temeke. In Temeke 36% of subwards indicated that young males are unlikely to have a phone, while this figure is 72% for young females).

Subward Average Observations

Taking the average response over subwards produces overall risk maps to different survey features in an area. The maps in Figure 5.9 give an example of the subward average response to six different survey features. These survey questions are as follows:

- *Theft and Violence is a problem in this subward? (Msongamano ni tatizo kwenye mtaa huu?): 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree'*

- *There is good access to medical care in this subward? (Kuna ufikiaji mzuri*

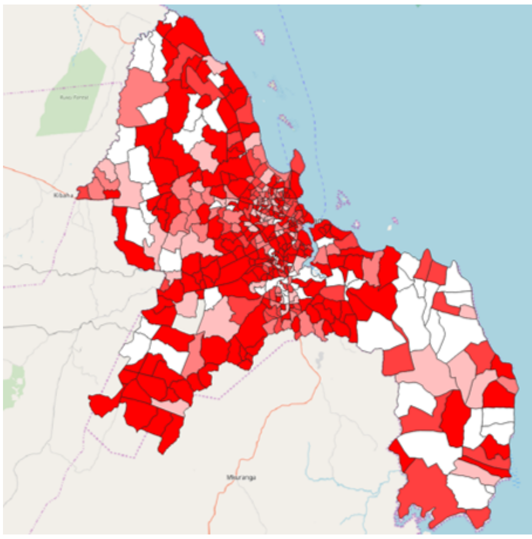


Figure 5.7: Street Survey Results ID Ownership Teenage Males. White represents areas where it is highly likely teenage males do not own their own personal identification, and red represents areas where it is highly likely they do own personal identification

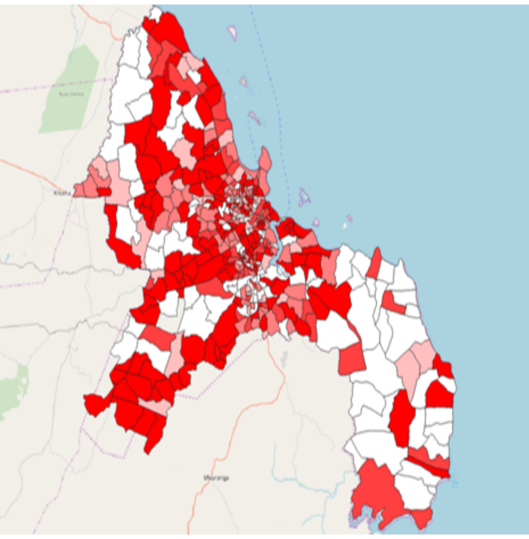


Figure 5.8: Street Survey Results ID Ownership Teenage Females. White represents areas where it is highly likely teenage females do not own their own personal identification, and red represents areas where it is highly likely they do own personal identification

wa matibabu katika eneo hili la chini?): 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree'

- *There are people in this subward being forced to marry against their will. (Kuna watu katika subward hii wanalazimishwa kuoa bila mapenzi yao.): 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree'*
- *I feel safe in this subward at night (Ninahisi salama katika eneo hili la usiku): 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree'*
- *There are people in this subward being forced to work against their will. (Baadhi ya watu katika mtaa huu hufanyishwa kazi bila ya hiari yao) : 1. Strongly Disagree. 2. Disagree 3. Neither 4. Agree 5. Strongly Agree*

Other interesting insights revealed by the survey include the links between variables. The following features from the street survey were all found to be highly correlated: • Secondary school attendance and high mobile phone usage by young people • High prevalence of forced labour and high pay in a subward (although this seems counter intuitive, there is more opportunity for exploitation in industrial areas, forced labour is often masked by employment opportunities) • Subward overcrowding, litter and poverty • Violence and perceptions of night time safety • Unemployment and poverty. These sorts of links can be useful ways of exploring alternative methods of ground truth data collection. For example, given the high correlation between over crowding and poverty, overcrowding, which is readily detected from satellite imagery might be leveraged as an indicator of poverty. This can be done at great scale and at much lower cost and resource requirements than traditional surveys. Figures 5.10 - 5.13 show example subwards with both the highest and lowest recording of poverty from the street survey. Figures 5.10 and 5.11 show a highly deprived area from the street survey, providing an example of the observable nature of overcrowding in earth observation data. Figures 5.12 and 5.13 however are

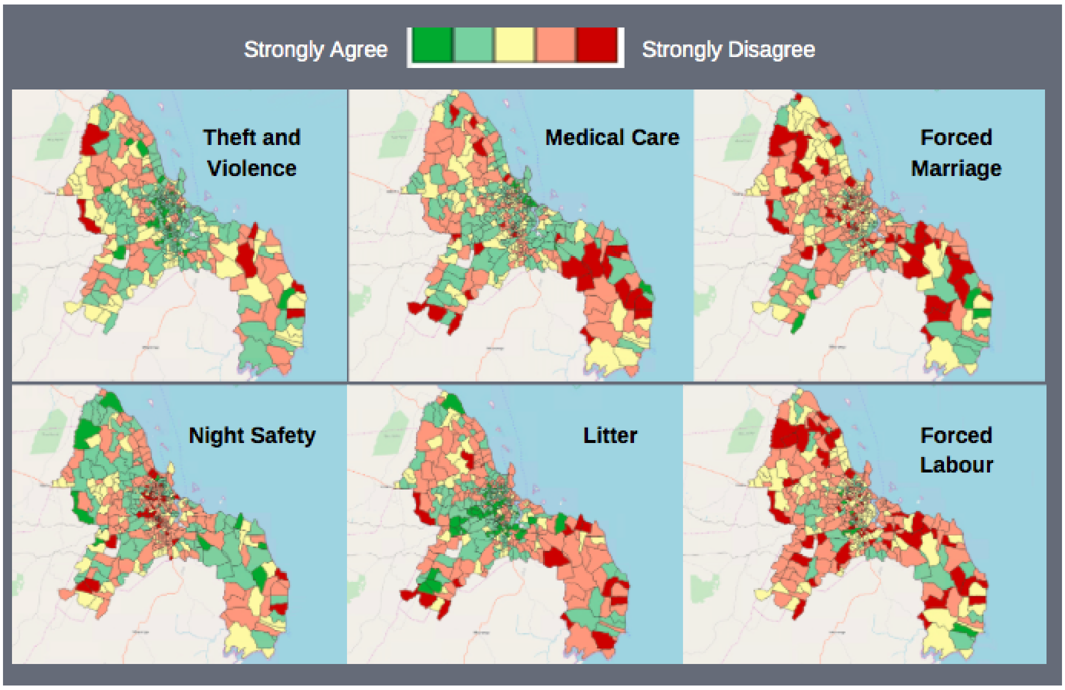


Figure 5.9: Street Survey Results Maps Dar es Salaam. Red represents the issue is a problem for the area and green represents that the issue is not a problem for the area.

much more spacious and have been recorded as highly affluent in the street survey. Given the quantity of questions in this survey this is a small glimpse at the extensive knowledge learnt from the street survey and the potential work which can be done with it. [4]

5.2 Proxies

It became apparent from both pilot surveys and facilitators feedback that participants are more willing to talk about observable demographic issues such as overcrowding, rather than personal and sensitive issues such as forced labour. Questions directly asking participants about their perceptions of forced labour and poverty for example were met with hesitation. There are

[4] Given the size of the file a full copy of the street survey results are available via the following link rather than the appendix https://www.maddyellis.co.uk/new-page

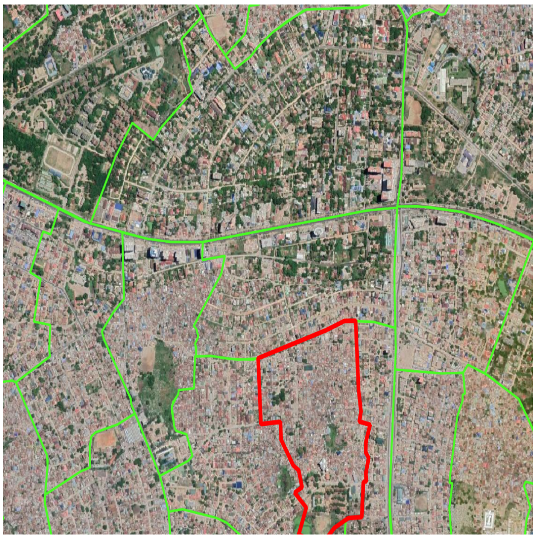


Figure 5.10: Example of subward with a high poverty score

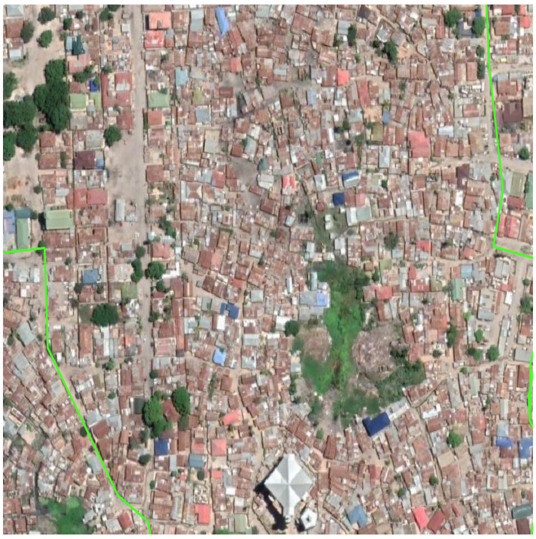


Figure 5.11: Zoomed in example of subward with a high poverty score

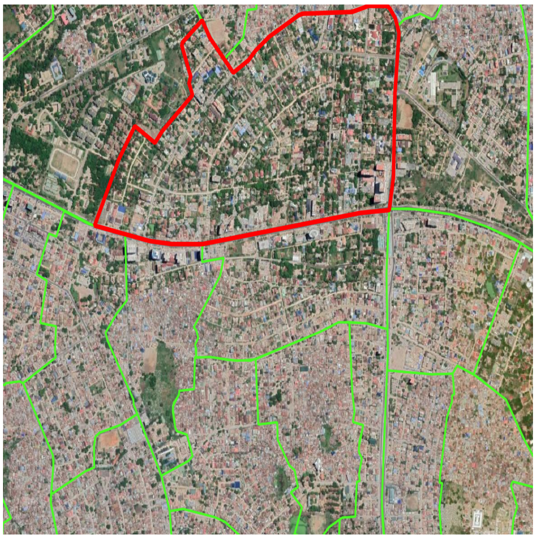


Figure 5.12: Example of subward with a low poverty score

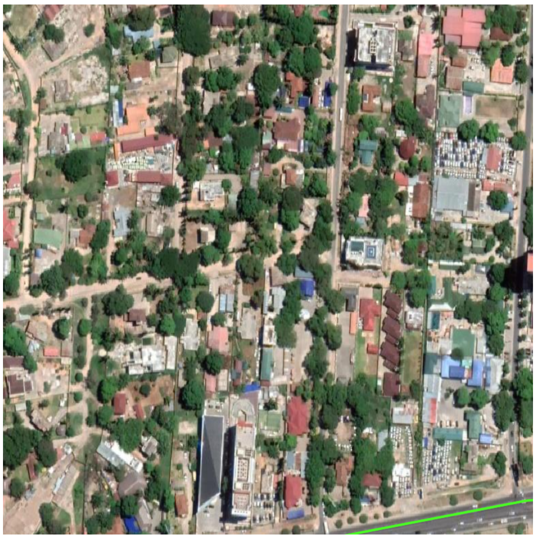


Figure 5.13: Zoomed in example of subward with a low poverty score

potential fears of incrimination with subward/ward officers for saying the "wrong thing". This not only creates issues surrounding privacy and ethics, but increases risk of skew in the results. Further, facilitators felt that some participants might think the results of the survey would be able to push immediate improvements in the subward around these issues. This is also likely to skew results given citizens uniformly want to see their subward improve. Two topics in this survey that facilitators felt might be particularly impacted by such concerns are poverty and forced labour. Facilitators fed back that they felt nervous directly asking questions on forced labour and obtained responses to such questions quicker than normal, and allowing less participant deliberation. Surveying sensitive topics like forced labour is often difficult, and one of the main reasons true predictions on forced labour estimates are hard to come by (Belser et al., 2005; Moreau, 2018; Ruwanpura and Rai, 2004; Bales, 2012). It is for this reason that proxies are often used in literature surrounding issues such as modern slavery and forced marriage. Proxies ask less sensitive demographics information in order to create approximations for sensitive demographic issues. Here we present proxies score for both poverty and forced labour. These proxy scores are created from a linear combination of demographic features collected in the street survey based on theoretical literature. The proxy scores in this thesis are average scores of the questions listed in the following paragraphs. This is done to motivate theory based proxies for analysis in chapter 6.

Forced Labour Proxy

Demographic features used to infer a forced labour proxy are drawn here from my interrogation of the following literature: (International Labour Organization, 2017; International Labour Office, 2012; Ruwanpura and Rai,

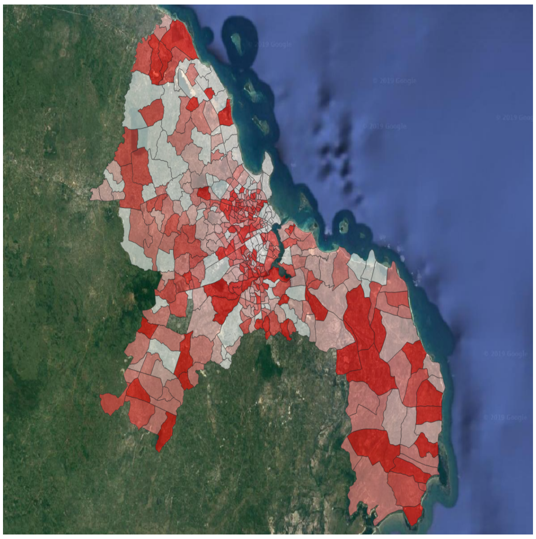


Figure 5.14: Proxy Map Forced Labour, darker colours represent a higher risk of forced labour

2004; Morgan and Olsen, 2014; Europol, 2016). Features used in proxy model: - *ID documentation* - *Mobile money usage* - *Mobile Usage* - *Work Formality* - *Street lighting* - *Under aged marriage* - *Theft and violence* - *Medical care* - *Overcrowding* - *Litter* - *Poverty* - *Unemployment1* - *Unemployment2* - *Arranged marriage* - *Road quality* - *Building quality* - *Teenagers in school* - *Children in school.* Figure 5.14 illustrates the results of the forced labour proxy, the darker red the subwards indicating high risk to forced labour.

Poverty Proxy

Based on the theoretical literature presented in chapter 2.1 the features used for a poverty proxy model are as follows: - *Perceptions of day safety* - *Perceptions of night safety* - *Pay* - *Medical care* - *Theft and violence* - *Poverty* - *Unemployment1* - *Unemployment2* - *Perceptions of arranged marriage* - *Perception of forced labour* - *Road quality* - *Building quality* - *Street lighting* - *Work formality* - *Age start work* - *Teenagers in school* -

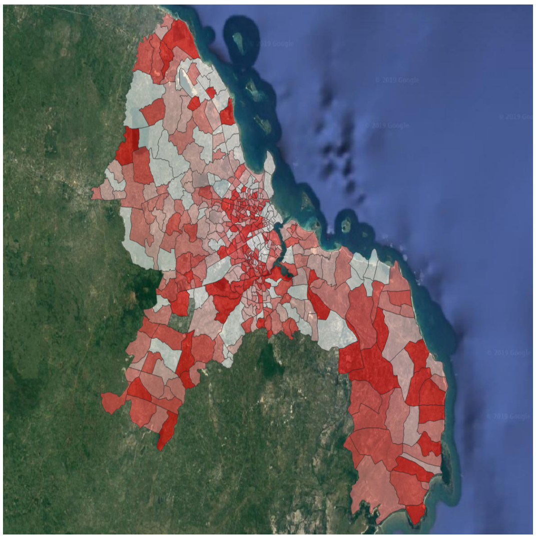


Figure 5.15: Proxy Map Poverty, darker colours represent a higher risk of poverty.

Children in school - Mobile money usage - Mobile usage - Under age marriage - Id documentation - Overcrowding - Litter Figure 5.15 illustrates the results of the forced labour proxy, the darker red the subwards indicate less affluence. The proxy scores for forced labour and poverty are used in the investigation of systematic bias and the assessment of theory based indicators of forced labour as seen in chapter 6, as well as the detection of vulnerable communities via stochastic block models seen in 8.

5.3 Validation Interviews

The street survey represents a comprehensive data collection method which has despite its inherent logistical challenges produced a magnitude of new knowledge on Dar es Salaam for this thesis. The methods underpinning this survey are further evaluated in this section with semi structured validation interviews, but what is clear is that the implementation has successfully produced fruitful results. From underage marriages to household structures and mobile usage, the quantity of data generated here has the potential to underpin a great number of future projects.

Motivation

Collecting ground truth information for this work created a number of unplanned incidents and observations, from challenges with government official cooperation, to flooding and privacy barriers. To further investigate some of the issues relating to data collection in developing regions interviews were conducted with local data collection experts and government officials. These interviews were semi-structured and investigated the difficulties and challenges of collecting ground truth data in developing areas. Interviews took place in the last week of August 2019, and the insights revealed are an accompanying contribution, alongside the datasets produced and modelling extensions that were built upon them (see chapters 6, 8 and 9)

Process

Government Official Interviews Three government officials were interviewed; One ward officer and two subward officers. These interviews were focused on their experience of having data collected in their ward/subward.

Local Data Collection Expert Interviews Four local data collection experts were interviewed, these local experts had both taken part in the street survey and also worked collecting data in Tanzania for a number of years. The Local data collection experts were interviewed on the challenges collecting data in developing areas.

In both the government official interviews and local data collection expert interviews a translator was present and the interviewee signed an adult consent form. The form verified they understood the purpose of the study which had been explained to them and consented to take part and be referred to anonymously in any publications. This form can be seen in appendix I.

5.4 Results and Discussion

Ward Officer Views on Data Collection in their region

Quantity of Data Collection Of the three officers interviewed one was from a rural area of Dar es Salaam and the other two where from urban areas. Their views on the quantity of data collected in their ward were very different. The officers in the urban areas both fed back that they had a lot of people coming to collect data in their regions, city planners, university researchers and a number of organisation researchers such as the world bank. The officers from the rural area however had a different experience, explaining that no one (other than their experience with this project) had previously collected data from them. During the data collection in this project, it was observed there are far more barriers to collecting data in more rural areas. Issues such as impacts of flooding, participant engagement and understanding, transport and access to the region were all exacerbated in the more rural areas. This emphasises the need for community collaboration and careful planning in the preparation of fine-grained surveys. Having more knowledge on more rural areas will help to overcome some of these barriers and ensure areas are not missed from data collection.

Advice for future data collections in their regions' The strongest message that came from all three ward officers was that the most problematic things about data collection was 1) Lack of change 2) Lack of officer involvement and community awareness. When talking about the lack of change, one officer mentioned that the people in their region are tired of participating in surveys because they have done so for years and seen no improvements for their problems. Another said people do not trust data collectors because again they have contributed their time and seen not changes. This officer told a story of a project years ago, where many people gave their time to explain problems of theft and violence in the region. Since then no-one has heard from that team, and they have seen no new safety measure put in place. Although significant changes cannot happen overnight it is clear that data collection needs to be introduced carefully in order gain participant trust and not provide false hope.

The lack of ward officer involvement and community awareness was by far the leading conversation which all interviewees. The feedback on this topic can be split into three categories. Community meetings, Subward Guides and Officer data. Community meetings are gatherings held regularly (often weekly or biweekly) in subwards. These meetings provide a chance for speakers to present any news to the subward and an opportunity for residents to voice their views. Although community meetings are open to everyone in the community, they are mostly attended by active members of the community and news is filtered through them to other residents. One interviewee mentioned the community meetings were viewed positively because the citizens trust them, people are comfortable speaking in these meetings and see many outcomes and chances from these meetings. The interviewee suggested that integrating data collection into these meetings would add a level of trust between projects and citizens. Another officer mentioned that if data collection was done through community meetings there would additionally be less confusion around the project, and eliminate the need to explain a projects' remit and requirements many times, with the whole community being introduced to it in concert. This officer said:

'We must prepare maybe a community meeting, so that we can introduce the thing which we plan to do, so that they the citizen can understand what we are about to do, about the data collection'

A third officer also felt community meetings would be an ideal way to conduct the interviews, adding that whenever community meetings were hosted there are street leaders who use megaphones to relay the meeting messages to the rest of the residents. They added that citizens will not want to answer questions, if they are not aware of the projects, and their purpose. When conducting my data collection there were some participants who felt nervous engaging with facilitators, despite permission letters as they wanted to know how there leaders 'felt'. I engaged in community meetings for context only rather than disseminating the goal of the project, it is clear a deeper collaboration between pre-existing community structures and leaders would benefit future work.

The second category of this discussion was the advantages of facilitators being paired with subward guides (citizens of subwards who support the work of subward officers). When discussing their experience with past data collectors one interviewee said:

'They are coming from lots of different places and they do not know this area so they must have the (local) members here to lead them in different places in the subward.'

It was suggested that having a local from the subward to both liaise, contextualize and guide the data collectors, introducing them to citizens and introducing the project could be highly advantageous. When asked how data collectors could encourage citizen engagement another officer said:

'The best ways is taking the people who are originally here, to take us to them'

When planning the street survey I knew it was important to have in country data collectors, however this advice to include more local knowledge in the process is not only entirely sensible but collaborating with people already within communities would additionally help to build trust and encourage engagement further.

The final category of discussion centred around pre-existing subward officers. Both subward officers interviewed felt that as part of their job citizens come to them to relay their thoughts and problems in the subward. As such they felt they had the most generalized knowledge on the subward and its surrounding issues, and that a rapid way to improve overall information about a subward was to speak directly to subward officers. One of the subward officers went on to explain that as their subward was part of a larger ward, they had no knowledge of the permission letter we had from their ward officer. This lack of awareness clearly made them feel uncomfortable with the data collection occurring in their ward, despite the existence of the official permission letter. It was clear that both subward officers felt they needed to be more informed in this data process. In contrast to this, the ward officer discussed the issues they had retrieving data from subward officers themselves, reporting that subward officers often do not have computers and so data is either kept on paper or simply left unrecorded and having to be relayed verbally. For both ward and subward officers it was clearly important they were aware of what was happening in their subward; this is something which must be reflected in future work, especially given the internal logistical challenges local organisational infrastructures currently face.

Local Data Collection Expert Interviews

The local data collection expert interviews focused on the four interviewees experiences of data collection processes collected in Dar es Salaam, reflecting on both the street survey as well as other data collection work they have been party to in the area. The feedback from these interviews is heavily linked to the overall feedback from facilitators on the street survey. As such below is a summary of interview feedback followed by overall facilitator feedback.

- *Participants* When discussing the difficulties engaging participants, 3 out of the 4 interviewees mentioned that engaging women in the survey was harder than men. One interviewee said:

  'It will take you a lot of time to make them (women participants) understand and they are not happy to speak if someone is listening'

  Another point raised was that men do not like to give personal information, there is a concern that they will get in trouble with officials if they say the wrong thing. One interviewee added that young men were the easiest to engage in surveys because they often had the most education and outside exposure. All of the interviewees said further concurred that this was the easiest demographic to engage, many of the respondents actively wanted to be involved in this kind of work (and offer suggestions of other people in the area to engage with and talk to). It was clear from these interviews also that the primary challenge with participants is privacy, many people (and in particular older generation) are unkeen to share personal information and experiences. This supports the decision to shift focus of the street survey to participant perceptions on areas, rather than personal circumstances.

- *Engagement* A challenge brought up by all four interviewees was the issue of understanding. Many respondents either didn't understand the project when described, or simply ignored the facilitators. Facilitators all said it consumed a lot of time 'selling' the project to get people involved. This echoes what was learnt from the ward officer interviews, that a formal project introduction through community meetings would improve the efficiency of future data collection.

- *Question Style* All of the interviewees mentioned that they would, in future, avoid questions with percentages. One interviewee mentioned that they would re-phrase these questions with options such as 'Most', 'Half' because not everyone understood percentages. One of the interviewees emphasized their preference for surveys which had multiple choice answers, due to the increased efficiency they had experienced. However, and in contrast, the other three interviews all talked about how participant engaged well with open questions. They felt open questions gave participants the opportunity to tell their story: for example if they want more security in their area, they were able to reflect upon multiple problems and specific factors they wished to be changed and improved. Such information can be extremely useful. As the street survey aims to underpin a number of analytical projects, and has to balance 1) numeric data, 2) detailed information on an area and 3) survey length it makes sense to include a mixture of open and closed questions in the future.

- *Subward Connections* The interviewees brought up a couple of issues in relation to the connections with subwards and officers. One thing mentioned that supports the views gathered in the officer interviews is that many subwards are becoming saturated with data collection. There is clearly many people coming in and out of subwards looking for data without producing tangible results. This is creating a mistrust and disengagement. One of the local data collection experts suggested that as well as an open and honest introduction, something which benefits citizen engagement, would be to shift the focus to university student-based experiences. Citizens are much happier taking part in a students survey to support their learning rather than taking the time to complete a survey which they feel is full of false promises. This indicates the ongoing benefits of integrating university students as a valuable resource in future street surveys. Another thing mentioned in relationship to the subwards was the cooperation of officers. It is clear that there is a huge variety of subward officers, some who are very cooperative (and others who are less so and indeed some looking for ways to earn money from data collectors). This is reflected in this street survey, where some of our facilitators were arrested by subward officials, despite having an official ward permission letter. The facilitators were only released when a payment was made to the officers. Eventualities such as this, which cannot be ruled out in developing contexts, reflect the need for a strong facilitator support system and supports once again the idea of introducing the project through community meetings before data collectors are sent in to work.

The data collection methods created a number of unplanned incidents; from challenges with government official cooperation, to flooding and privacy barriers. To evaluate the street survey process, in addition to the interviewees direct comments, the following points summarise the issues and facilitators feedback collated over the study period:

- *Legal Situation* Once one is in possession of formal permission letters from ward officials it is legal to collect data in Dar es Salaam, Tanzania. In reality, however, local authorities such as subward officers want to be involved and informed when data is collected in their subward. As such, during the street survey, two facilitators were wrongfully held in a subward office by officials. In this instance the situation was resolved thanks to the careful support system put in place for facilitators and the collaboration with Humanitarian OpenStreetMapping Team. This incident was a strong reminder of the importance of careful planning both in terms of facilitator safety and local authority collaboration. As was suggested in the interviews future surveys would benefit from community meeting introductions so everyone is clear on the project before data is collected.

- *Demographic Indicators* One of the most successful elements of the street survey was the shift of focus from participant experiences to their views on overall behaviours within their subward. Asking 'Are you employed' is a lot more accusatory and personal than 'What proportion of people in this subward are unemployed'. This shift had two main benefits. 1) Facilitators felt safer engaging with participants in the field. 2) Participants were more willing to take part in the survey and answer the questions. This is a key element of the street survey and something which is highly recommended for future work.

- *Data Collection Saturation* Again, this was a reflection of the quantity of organisations, charities and research groups (often initiated from western intervention) collecting data within the subwards over the years. All making implicit promises of change without, in actuality, long term impact occurring creating the sceptical attitude of citizens towards new data collection projects. From the interviews the main methods of combating this in future work would be to introduce the project properly with realistic aims and involve local researchers as much as possible. Facilitators did feel that approaching people with a soft beverage and being grateful for their time did improve the engagement of participants.

- *Weather* Tanzania has extreme weather, from humid heat to rainy season and floods and as such, the weather brings many challenges to data collection. Throughout this survey there were a number of issues related to floods and heat, facilitators had to adjust travel plans and postpone field work. When planning future work it will be important to plan the survey to be adaptable for unplanned weather interruptions.

- *Safety* Another unplanned incident in this survey was a facilitator road accident. While travelling to a subward a facilitator fell off a motor bike taxi. Luckily in this instance the facilitator was unharmed and the in country support system worked well, he was quickly collected and returned home. This issue did emphasise again the need for the in country support system given there is only so much support which can be provided virtually. In future work it will be essential to construct in country support systems to ensure the safety and well being of facilitators and this is particularly relevant again in developing contexts.

- *Scale* Given the goal of this work was to collected detailed ground truths on 452 subwards, and each subward needed to be visited by different facilitators, the logistics of this project were complex. With the added issues mentioned above such as floods, accidents and other issues, data collection of this scale needs to be carefully monitored and supervised. Although this survey's preparation and collection remained faster than traditional methods, it is clear a significant proportion of time should be spent ensuring careful planning. Another issue related to scale is the quantity of incoming data which needs to be checked and verified. Given the scope of the work, there were hundreds of participant responses submitted daily, the use of digital surveying methods and an app speeds up the transfer of data. There remains, however, a need to review and check each response stringently. The survey was an iterative process, responses were continually checked for markers such as location and speed. Once data is checked and facilitators have done an allocated amount of work, payment is sent to facilitators and they are given the opportunity to feedback their experiences and challenges collecting the data. Having in country support teams collate this information helps speed up this process and was found to be an essential component.

- *Question Style* As mentioned in the interview analysis, questions including percentages were difficult for some participants to understand. Much like the maps in the grid survey, it is important not to make assumptions and use 'western style' survey questions. Piloting surveys reduces some of these issues, and in future work it is recommended to work with proportions such as 'Most of' and 'All of', as opposed to percentage-based questions.

5.5 Summary

The street survey produced data collection results in Dar es Salaam despite the challenges of implementing such efforts in developing regions. The shift in focus from participant experiences to area views has helped to overcome issues of sensitivity and privacy. The results produced in this work are fine grained and have the potential to underpin a range of future projects. This study has successfully met the following goals: • Advances in adaptions of 1st work survey techniques in development setting, • The collection of fine grained ground truth knowledge in Dar es Salaam.

This chapter has also evaluated the data collection carried out in the street survey, meeting the following goal *Review of challenges of data collection in development setting*. The street survey is the first census style street survey on poverty and related issues to be completed in Dar es Salaam at a subward level covering all regions. At £5450, this survey cost considerably more than the comparative judgement based grid survey, but considerably less than traditional household surveying methods. The efficiency, focus on subwards rather than individuals and fine-grained scale of this work are all aspects of this survey which have successfully produced comprehensive and accessible data. The methods used in the street survey can be replicated in other developing regions to underpin a number of potential projects. Future works might consider adapting the street survey slightly to collaborate with local community meetings, to properly introduce the project. Careful planning and in country field pilots tests are an essential part of the survey process, and the survey needs to be adaptable to unplanned incidence such as floods. The collaboration with ward officers, local experts and facilitators has allowed a meaningful evaluation of the data collection which can be utilized to improve future projects.

Chapter 6

Examination of Systematic Bias

6.1 Examination of Systematic Inclusion of Bias

The street survey and grid survey have both produced estimates for subward poverty levels throughout Dar es Salaam. Given their different strategies there is perhaps unsurprisingly, a disparity between these results. However, given the comprehensive participant selection, high levels of participant subward knowledge, quantity of data collected and validation from substituent experts, the comparative judgement results from the grid survey, reflecting $> 70,000$ pairwise comparisons reflect a 'ground truth' for affluence levels across the extent of this work. The grid survey is also less likely to be affected by personal bias, with participants scoring the whole of Dar es Salaam rather than answering specific questions about an area they live in.

There are many potential reasons why the street survey might produce differing results for the spread of poverty across Dar es Salaam, but systematic biases are well evidenced as impacting such approaches. Having parallel surveys, one which can be confidently considered as a 'ground truth' provides investigative opportunity. Here I consider the hypothesis that the street survey is introducing such biases which are skewing the results. If these biases can be related to easily observable characteristics of subwards or participants, this is something which could then be accounted for in future studies to produce more reliable results. Observable characteristics related to participants might be aspects such as age or gender of participants (for example women could feel more negatively towards the affluence of a subward based on personal experiences of safety). Alternatively, older generations could have stronger positive or negative feelings towards a subward based on how it has changed over the years. Observable characteristics related to the subward could be issues such as building or road quality, poor qualities may be pushing negative views on subwards where quality of life is actually quite good. One interesting observation in the street survey data was that people living in more affluent areas seemed more likely to downgrade their views of their subward. Treating the grid survey lambda results (affluence estimates) as the overall ground truth for poverty in Dar es Salaam Figure 6.1 shows a plot of the difference between the poverty proxy scores against the ground truth of the lambda scores (with all scores having been standardised). Although there are limits to the extent to which respondents from less affluent areas can downgrade themselves (due to the lower bound of the likert scale), this Figure 6.1 still infers that people living in more affluent areas are holding disproportionately negative views of their area in comparison to the rest of Dar es Salaam. This is further illustrated in an alternative plot in Figure 6.2 where the poverty proxy scores are plotted directly against the ground truth of the lambda scores. Again it can be seen that areas recorded as more affluent in the grid survey are downgrading themselves in the poverty proxy scores. The following sections investigates

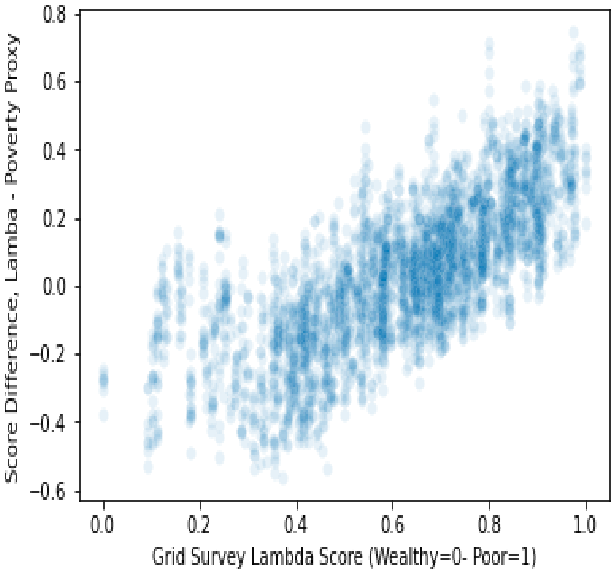


Figure 6.1: Bias score λ_D against poverty λ_G)

the hypothesis that there is an affluence bias occurring in this data.

Method

The target variable of this section of work is the *difference* (λ_D) between the standardised grid (λ_G) and street proxy scores for poverty (λ_S), such that $\lambda_D = \lambda_G - \lambda_S$. Recall that λ_S is the proxy score for poverty calculated from the average of the street survey features related to poverty listed in chapter 5.2. λ_G is the Bradley-Terry estimate of deprivation as seen in chapter 4. Both λ_S and λ_G are translated and scaled to lie within [0,1]. The difference score λ_D has values between -1 and 1, negative values representing views which are more pessimistic of an areas affluence than the reality, and positive values representing views which are more optimistic than the reality. A difference score of zero represents an accurate response in the street survey proxy poverty score. From this point on the target variable λ_D is defined as bias score.

In order to investigate potential observable sources of bias within the street

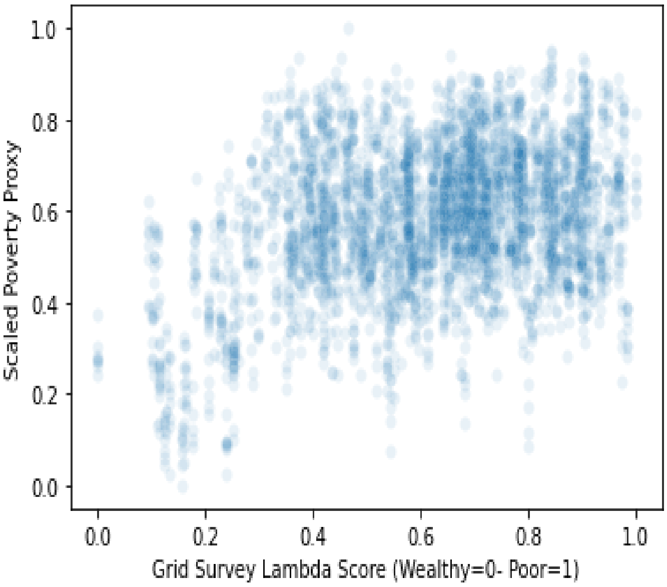


Figure 6.2: Scaled Proxy Poverty λ_S against Grid Survey Poverty λ_G)

survey, the first consideration is the quantity of data. Each participant answered over 60 questions about their subward, creating over 120 individual features per survey response. These features are first reduced to those which can be easily observable, such as participant gender, subward overcrowding, litter and road types. This step has been done because the overall goal of this section of work is to uncover features which can be easily observed and accounted for when performing future street surveys. For this reason, features such as perceptions of forced labour and arranged marriage which are sensitive and would require the street survey itself to uncover have been removed from the data set. Principal Component Analysis (PCA) (Jolliffe, 2002) and Partial Least Squares Regression (PLS) (Wold et al., 1993) are initially considered to reduce the number of input variables when recovering the bias score. Following this lasso regression (Santosa and Symes, 1986) is used to further investigate the feature importance explaining the bias.

Results

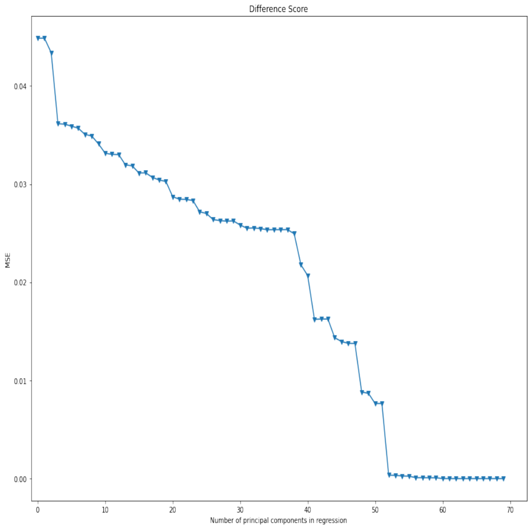


Figure 6.3: Principal Component Analysis with simple regression MSE

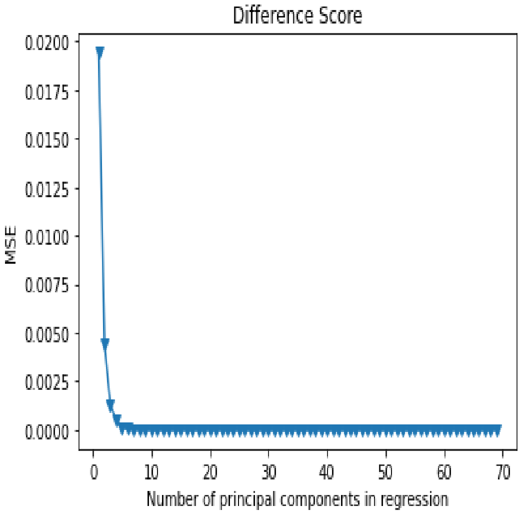


Figure 6.4: Partial Least Squares Regression MSE

PCA and PLS are used here to convert highly correlated variables to a set of independent variables via linear transformations as a technique for variable reduction. There are 70 standardised input variables used in this analysis from the street survey, so PCA and PLS can have a maximum of 70 components (latent variables). Figures 6.3 and 6.4 show the MSE results for the bias score given simple regression and different numbers of components for PCA and PLS respectively. It is clear from these figures that PLS is outperforming PCA. PCA requires over 50 components to reach a minimum MSE, whereas PLS is achieving this in only 5 components, and producing a MSE for the bias score of 0.0012. Although PLS is doing a good job in terms of predictive accuracy the overall goal of this section of work is to explore features responsible for the bias. PLS makes such analysis extremely hard, and continues to provide challenges in evaluation of important variables (there is a lack of methods to clearly unpack the story behind the variables (Tran et al., 2014)).

In order to explore the causes of bias the next step was to perform lasso regression. Lasso regression removes input variables that do not contribute much to the prediction task this provides an automatic feature selection. Lasso regression removed 43 of the 70 input variables. The remaining 27 input variables achieve an R-squared value of 0.945. Figure 6.5 shows the feature importance of these 27 inputs in the regression model. The feature scores indicated are the regression coefficients of the standardised inputs, which can be safely be interpreted as indicative of the strength of influence each variable has in the optimized model.

From Figure 6.5 it is evident that the main cause of bias is *'lambda_scaled_rev'* (which is λ_G, affluence levels). More affluent areas tend to see themselves as

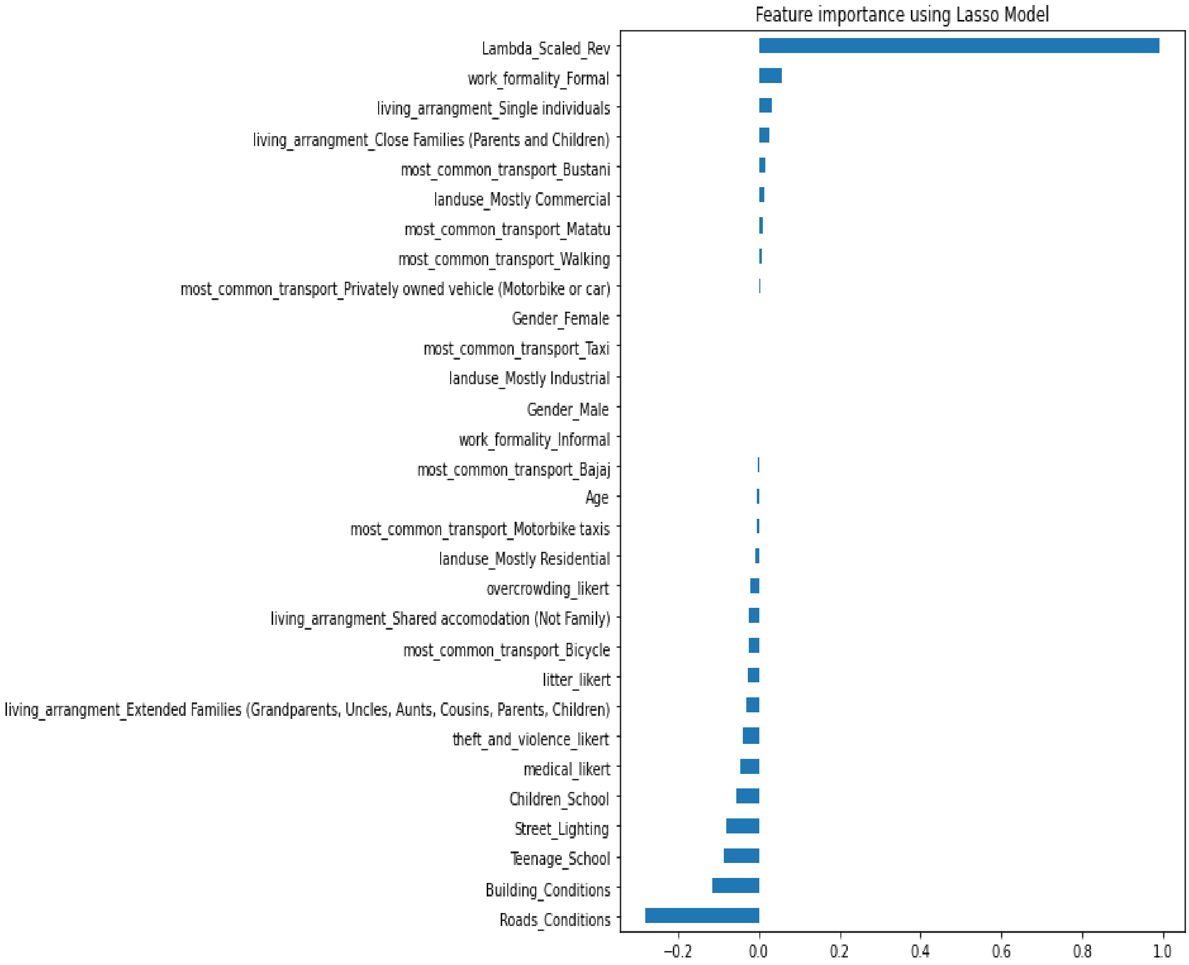


Figure 6.5: Feature importance for bias score, illustrating that the main cause of bias is *'lambda_scaled_rev'* (which is λ_G, affluence levels)

worse off than they are in comparison to the rest of Dar es Salaam ($Bvalue = 0.89$). This 'affluence bias' makes sense because people are likely to spend their time in similar areas to those that they live in. For example, someone living in an affluent area is also likely to visit other affluent areas for work or social reasons. They will be less likely to spend a lot of time in the more deprived areas. When comparing their subward to other subwards they will therefore likely be considering other affluent areas they are familiar with, rather than comparing themselves with more deprived areas which they may never have visited. The affluence bias seen here could be accounted for when collecting future data, incorporating a bias correction could improve the results of street survey analysis.

Two other notable features in this bias regression are road and building quality. The results show that seeing poor quality buildings and roads push participants to downgrade subwards to worse than they actually are ($Bvalue = 0.22, 0.29$ respectively). This shows that building and road quality are not necessarily the best indicator of poverty per se. An area may appear run down, but this may not reflect the overall affluence and quality of life in that area. Road and building quality is something which again, could be observed on-the-ground or via satellite or drone imagery. These features are easy to assess and could therefore easily be recorded and incorporated in bias corrections of future models.

Although it was hypothesised that the age and gender of participants might contribute to biases in the street survey, results do not indicate this ($Bvalue = 0.004, 0.02$ respectively). Both age and gender have very little impact on the bias in the street survey. Despite other research showing differences in survey responses from participant gender (Olsen and Cox, 2001;

Flores-Macias and Lawson, 2008; Haselton, 2003; Huddy et al., 1997), this work finds no notable bias created from participants' genders. One reason for this could be the shift in focus of the street survey questions, most traditional surveys ask participants to answer questions based on personal experiences, whereas this street survey focused on participants perceptions of an area as a whole. For example, rather than asking 'Have you experienced theft of violence in this subward' the street survey asks 'Are theft and violence a problem in this subward'. This shift moves the focus onto the subward rather than the participant themselves. This lack of participant bias in terms of age and gender strengthens evidence in support of the decision to focus future street surveys on participants perceptions rather than experiences.

6.2 Discussion

An overwhelming takeaway from the bias study is that the people living in more affluent areas are downgrading themselves and that response bias exists. Other insights revealed are that, despite the literature, there is no evidence to suggest the age and gender of participants engenders a bias in results. Poor quality of observable infrastructure such as roads and buildings, however, is pushing participants to downgrade areas to worse than they actually are. While these results account for pair-wise linear interactions, there may be more complex non linear interactions occurring. As such, there is potential here both to systematically account for observable bias in future work and also further investigate the non linear relations creating bias in the results.

The grid survey results contrast sharply, and are taken to be the most reliable ground truths for poverty levels specifically in Dar es Salaam. However, this did not replace the need for the wider breadth of the street survey. The street survey although more vulnerable to bias, produced more detailed information on each subward, allowing for wider analyses. Surveys are an essential part of development and often carried out in resource limited environments. This work therefore suggests that street survey and grid survey approaches should be used together to create reliable, efficient and detailed ground truths in developing areas.

6.3 Summary

This chapter has examined the systematic inclusion of bias in data collection analysis. A key influence of bias found in this study is the affluence bias, people living in more affluent areas are downgrading themselves disproportionately. This section of work has also shown that machine learning is a highly effective method for revealing new covariates to be leveraged as proxies for the detection of forced labour. Further to this, proxies have been identified, which are less sensitive than currently theorised covariates and can, therefore, be collected with fewer challenges and resistance. The key proxies identified are pay, household structure and area of origin, alongside confirmation of already theorised covariates such as education and lack of identification documents.

Chapter 7

Assessment of Theory Based Indicators of Forced Labour

7.1 Machine Learning vs Proxy Models

I now return to assessment of potential proxy models, detailed in chapter 5, which may remove reliance on sensitive questions such as those relating to forced labour. Recent figures estimate the number of people living in slavery to be as high as 40.3 million globally (International Labour Organization, 2017). Crimes such as forced labour, human trafficking and domestic servitude inhibit economic development, decimate some of the planet's most ecologically vital environments, and reflect a serious threat to sustained development (Landman and Silverman, 2019).

Quantifying modern slavery remains of crucial importance to the design of effective policy interventions, with the UN stating "without good data on where slaves are, how they become slaves and what happens to them, anti-slavery policy will remain guesswork"(Cockayne, 2015). While the global community has rallied around a target to end slavery by 2030 (UN SDG 8.7), progress is severely hampered by a lack of fine-grained data concerning the issue at local levels in particular. The lack of data is understandable, given the subjects hidden and highly illegal nature, which renders traditional data collection approaches both dangerous and impractical (Bales, 2012; Moreau, 2018). Traditional data-collection techniques, such as household surveys, cannot quickly be brought to bear. By definition, modern-slavery is challenging to assess; data is hidden; questioning surrounding the issue is extremely sensitive; ethical considerations abound; and due to criminal involvement, the safety of both survey facilitators and respondents is at risk.

Statistical techniques such as Multiple Systems Estimation have been used as substitutes to generate broad-brush national figures by leveraging structure between parallel lists of slavery events, recorded by multiple organisations (Silverman, 2020). However, given their data requirements, such techniques are inapplicable at sub-regional levels. Modellers are left with theoretical ‘proxies’ when assessing risk - hypothesised co-varying factors such as poverty, crime, and gender inequality (Bales, 2006; International Labour Office, 2012; Larsen and Durgana, 2017; Ruwanpura and Rai, 2004).

However, despite firm qualitative foundations, the validity of such proxies remains broadly untested empirically, especially at sub-regional levels. National figures have drawn valuable attention to the issue, but estimates are generally broad-brush and predominantly unvalidated. Figures at sub-regional zones are very rarely examined. Moreover, proxy indicators tend to conflate many different types of slavery (such as forced labour; domestic servitude; sex trafficking; debt bondage; and illegal child labour), as well as different geographical settings (urban slavery has vastly different characteristics compared to rural or coastal contexts). As such, existing models currently remain blunt instruments, linearly compressing hypothesised proxy indicators from qualitative fields (such as human security and crime prevention theories (Larsen and Durgana, 2017)) before linearly projecting them. It is here that machine learning methodologies hold potential to contribute.

This section challenges current practice, comparing 'theoretical' models established from qualitative research, with a non-linear machine learning approach to predict slavery risk. For the first time, incidences of modern slavery are modelled at hyper-local levels across the 451 subwards of Dar es Salaam. A multi-view supervised learning framework is used below to investigate a range of model classes and evaluation metrics, validating against 3668 ground-truth data points derived from the street survey.

Predictors are drawn from 3 distinct sources: • non-sensitive survey questions reflecting a range of demographic, social and environmental factors for each subward • behavioural features derived from mass transactional data shared by a Mobile Network Operator (cell tower activity; mobile money) • geospatial features derived from drone imagery (land-use; infrastructure). Models are optimised and tested via nested cross-validation, with best-of-class models undergoing a series of variable importance analyses (SHAP) to identify new hypotheses for co-varying factors reflecting risk to inhabitants.

Method

This section aims to: (1) provide empirical evidence of the theorised proxy indicators of the risk of forced labour at sub-national levels, addressing the current paucity of information and (2) demonstrate the utility of ML to isolate alternate 'non-sensitive' proxy indicators collectable at medium to large scale. Modelling of forced labour in subwards is underpinned by key datasets comprising of:

1. The street survey, and specifically the questions eliciting: (i) ground truth measurements for risk of forced labour; (ii) the further demographic, infrastructure and behavioural survey question indicators, obtained via non-sensitive questions. The distinction between (i) and (ii) is of key significance, with the former being significantly logistically challenging and prohibitively expensive to acquire at scale due to the hidden nature of SDG 8.7, and serious risk of facilitator and respondent safety.
2. Transactional data shared by a mobile network operator provider (Described in 3)
3. Drone imagery and crowd-sourced annotated geospatial features (Described in 3)

More details on the novel big data sources can be seen in chapter 3 and more details on the street survey collection can be seen at the start of chapter 5. These data sources provide the basis for the construction of both (a) a deductive theory-based model, based solely on qualitatively proposed indicators and (b) a range of inductive machine learning models, conferring a set of alternative proxy indicators for consideration. This new approach to SDG8.7 is grounded in empirical establishment of proxy indicators, and non-linear relationships between, as discussed in the experimental setup below.

Feature Engineering

Street Survey Forced Labour Target

A ground-truth risk score for forced labour was extracted for each subward, taking the median response from each respondent in that *Mtaa* (subward). Of the 452 subwards surveyed, only 8 indicated the highest possible level of risk (with every participant returning a 'Strongly Agree' response, conferring high confidence). To generate labels for the binary classification experimental setup, these forced-labour risk scores y_i were discretised into two classes - subwards where $y_i \leq 2$ classified as High Risk (corresponding broadly to *strongly agree/agree*) and subwards with $y_i > 2$ classified as Low Risk. This resulted in a set of 61 positive and 390 negative data-points, with one ward nonassignable. For the 3-class formulation of the slavery-prediction problem, scores were discretised using boundaries $y \leq 2$ (*High Risk*), $2 < y \leq 3$ (*Medium Risk*) and $y > 3$ (Low Risk), producing class counts of 61, 157 and 233 respectively.

Street Survey Features

Input features in this set are drawn from non-slavery (and hence non-sensitive) survey questions. Initial filtering removed any questions directly investigating modern-slavery to prevent potential information leak. Subsequently, 144 features covering: local infrastructure (e.g. building quality, road conditions, healthcare facilities, street lighting); sociodemographics (e.g. age levels, employment, pay levels, school attendance, family living arrangements, the origin of residents); safety (e.g. daytime safety; night safety; crime; poverty); and behaviour (e.g. mobility; vehicle usage; mobile phone usage; weekend activity; etc.) were derived. Most survey features were encoded as the median response from all participants surveyed in a subward, with a variety of categorical items one-hot encoded (for example, the most popular form of transport in the region: taxi, motorbike, walking, cycling, bus, etc.). Note one-hot encoding is one method of converting data to prepare it for analysis and get a better prediction. In one-hot, each categorical value is converted into a new categorical column and a binary value of 1 or 0 is assign to those columns. For a full list of the available variables, please see the supplementary material in appendix E.

Mobile Features

A range of variables was extracted from the CDR, and MFS, resulting in 32 input features reflecting behavioural aggregates for residents of each subward. A CDR is logged every time a network event such as sending an SMS or making a phone call takes place. As such, they allow insight into both micro- and macro-patterns of human interaction, while allowing for the preservation of individual anonymity through spatial and temporal aggregation. Residential behaviour was deduced from these records through the calculation of the mode Base Transceiver Station (BTS) favoured by users between 10 pm and 6 am. By georeferencing the locations of these mode BTS, an anonymised dataset corresponding to users 'resident' in each subward was extracted. Features pertaining to residents at each BTS were then aggregated, ensuring strict privacy for analysis purposes. Features included the number of cell phone users in a subward (reflecting population), uptake of mobile money, the mobility of residents, the mean call distance, measures of activity at different points in the day, along with estimates of income, and spend from mobile money transactions. With estimates of as much as 25% of GDP reported to be passing through MFS in some East African countries, the data provides a representative indication of economic activity in the region. The full list of features is seen in appendix E.

Geospatial Features

A set of 24 inputs was generated from data collected via drone imagery, which was augmented with over 750k manual demographic and environmental annotations, including land-use estimates. Annotations were collected as part of the *Dar Ramani Huria* project (Eichleay et al., 2016) with local community members creating highly accurate maps of Dar es Salaam. An example of the images and annotations present in this dataset can be found in (Torres et al., 2017) having been shared by the authors. The 23 features extracted from this dataset included the total area (in km^2) of the subward, land-use measure (consisting of the residential area, slum, urban, industrial and unused, with equivalent percentages of the subwards that each made), and the distances that the subwards were from the coast, the central business district, the port and predominant slum and industrial zones. Again, a full list of these inputs and their descriptions can be seen in the supplementary material in appendix E.

Experimental Setup

Experiments were run to consider the ability of previously unassessed covariates to predict forced labour risk (measured based on sensitive survey questions) in both a linear and non-linear capacity. In this work, forced labour risk was considered as either a 2-class or 3-class variable. As detailed in the feature engineering, the considered covariates included those derived from the re-purposing of the transaction (CDR and MFS) and Earth Observation (drone imagery) data. Compared to the sensitive survey questions, these potential indicators are significantly easier to acquire at medium (non-sensitive survey questions) to large (re-purposed data) scale. The experiments focused on the models' F_1 score [1] (harmonic mean of the precision and recall) in order to assess the utility of the models overall and

[1] For multi-class models macro averaging was used to compute F_1 scores.

SHAP values[2] (Lundberg and Lee, 2017a) to assess the contribution/importance of each covariate used within each model. The F_1 score was selected instead of accuracy due to the unbalanced nature of the problem, the F_1 metric is also commonly used in machine learning analysis.

In order to evaluate previously unconsidered covariates, predictive models were formulated using each data set type (non-sensitive survey-based features; CDR and MFS based features; and drone imagery-based features). Experiments were grouped this way to reflect acquisition reality - i.e. if some non-sensitive survey features were available, all of the set could be. This led to three separate experiments, within which nine different machine learning model classes were compared: • a baseline Stratified Majority Classifier (BASE); • 3 linear models: Ridge Regression (RIDGE), Logistic Regression (LR) and Linear-SVM (SVM-L); • and 5 non-linear models: Kernel-based non-linear SVM (SVM-K), Random Forest (RF), AdaBoost (ADA), k-Nearest Neighbour (kNN), and Deep Neural Net (MLP)[3]. The experiments were first undertaken for the 2-class problem before being replicated for the 3-class problem.

For a single experiment (a type of input feature and either the 2 or 3-class output) all 9 models were trained and tested using a nested stratified cross-validation[4] procedure in order to tune each model's meta-parameters via a grid search[5]. The grid search parameters can be seen in figure 7.1. This ensured

[2] Via the Kernel SHAP implementation: `https://github.com/slundberg/shap`.

[3] In all cases sklearn implementations were used. For Deep Nets, this meant the MLPClassifier.

[4] 5 folds were used in both the inner and outer cross-validation stages.

[5] The following meta-parameters were considered, with exact values listed in the supplementary material in appendix E: Ridge: α. LR: penalty, C. kNN: number of neighbours, weights, metric. SVM-K: kernel, γ. ADA: learning rate, subsample percentage. RF: max features, max depth, min samples split, min samples per leaf, use of bootstrapping. MLP: hidden layer number and size, activation function, solver, α and learning rate.

the reported values were a generalised measure of the evaluation metric and SHAP values, guarding against traditional and procedural over-fitting during the meta-parameter selection for each machine learning model type. Selection of a final model across these model classes was not part of the nested cross-validation procedure as consideration of the performance of the different model classes (linear vs non-linear) and the different explanations (as indicated by the model's SHAP values) indicated by the different non-linear learners were a key goal of the experiments.

To compare the novel covariates and non-linear relationships uncovered by the machine learnt model, a baseline **theory-based proxy model** had to be constructed. While many proxies for slavery are either sensitive personal accounts, case reports (Belser et al., 2005) or countrywide features such as whether a country has ratified the ILO conventions on forced labour (Ruwanpura and Rai, 2004), there are also several community-based proxies and indicators in use to assess the vulnerability to slavery. As detailed previously (chapters 2.1 and 5), several 'pull' factors also underpin established vulnerability models [6] and can be broadly categorised as: Basic Needs, Inequality, Disenfranchisement, Violence and Conflict and Governance. Based on conversations with modern-slavery experts these categories informed the theory-based model design via corresponding survey questions on: poverty; unsanitary and overcrowded living conditions; lack of education; low income, and lack of employment opportunities; retention of identification documents; isolation due to lack of mobile phone or transport coverage; and safety, theft and violence (Europol, 2016; International Labour Office, 2012; International Labour Organization, 2017; Mishra, 2001; Morgan and Olsen, 2014). Corresponding features from the study's datasets that

[6] https://www.globalslaveryindex.org/2018/methodology/vulnerability/

Table 7.1: Parameters for model grid searches:

Parameter Name	**Parameters**
General Parameters	
Random State	17
strategy	stratified
Ridge	
alpha	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Logistic	
penalty	l1, l2
C	100, 10, 1.0, 0.1, 0.001
KNN	
Number Neighbors	10 integers between 1 and 22
weights	uniform, distance
metric	euclidean, manhattan
SVM	
kernel	poly, rbf, sigmoid
gamma	scale
Tree/Random Forest	
number estimators	5, 10, 20, 30, 50, 75, 100, 150, 250, 500
max features	auto, sqrt
min samples split	2 ,5 , 10
min samples leaf	1 ,2, 4
bootstrap	True, False
Boosting	
learning rate	0.001, 0.01, 0.1
sub sample	0.5 , 0.7 , 1.0
MLP	
Hidden Layer Sizes	(50, 50, 50), (50, 100, 50), (100,)
Activation	tanh, relu, logistic
Solver	sgd, adam, lbfgs
alpha	0.0001, 0.05
learning rate	constant, adaptive

conformed to these theoretically identified factors were extracted to form the input features to the theory-based logistic regression model (reflecting the linear models typically used within such literature). The included features and their dataset source, grouped by their theorised proxy theme, are fully detailed in the supplementary material in appendix E.

Results

2-class results

The first section of Table 7.2 reports the results for the binary prediction problem for all models when predicting using the three datasets. A boxplot is additionally included in Figure 7.1a for the non-sensitive survey dataset. Immediately striking is how poorly the theory-based proxy model, designed from prior hypotheses taken from the literature, performs against other models (perhaps reflecting the focus on push rather than pull factors for modern-slavery in social science literature). While this is particularly true when compared to survey datasets, a slight increase in performance is also present when using the full potential set of either only Geospatial or only Mobile features, despite the theory-based proxy model utilising a mix of geospatial, mobile and survey features. Further, within the machine learning set, use of non-linear models consistently outperforms linear approaches resulting in F_1 scores which demonstrate real-world utility.

Results in Table 7.2 also provide strong evidence that a range of previously unconsidered features have a predictive relationship with the forced labour risk. Results comparing best-performing linear/non-linear models (also illustrated in Figure 7.1) highlight that these relationship are highly likely to be non-

Table 7.2: F_1 score results for optimised models across all input datasets. The best performing model in each group is highlighted in bold.

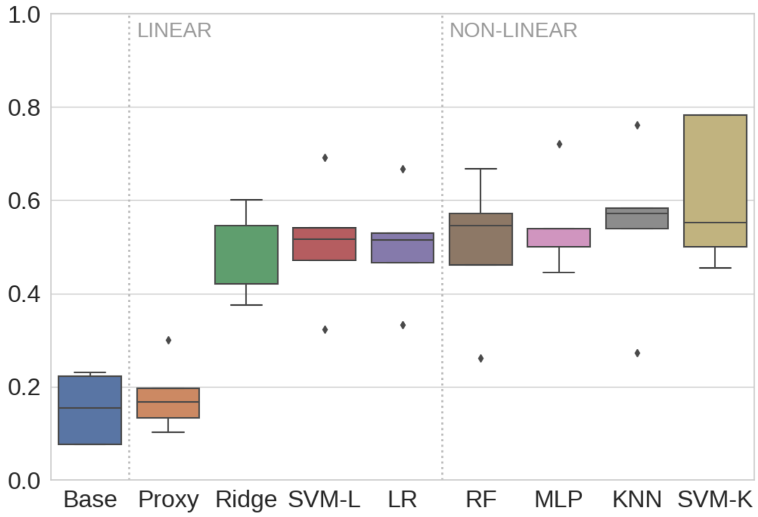
	Base	**Proxy**	**Ridge**	**SVM-L**	**LR**	**RF**	**ADA**	**MLP**	**KNN**	**SVM-K**	**Best Linear**	**Non-Linear**
BINARY MODELS												
SURVEY	0.152	0.180	0.497	**0.508**	0.502	0.501	0.490	0.548	0.546	**0.614**	0.508	**0.614**
MOBILE	0.152	0.180	0.143	0.172	0.146	0.131	0.029	0.214	0.206	**0.241**	0.180	**0.206**
GEOSPATIAL	0.152	0.180	0.143	**0.172**	0.146	0.131	0.029	0.214	**0.206**	0.241	0.172	**0.241**
3-CLASS MODELS												
SURVEY	0.227	0.395	0.618	0.636	**0.641**	0.659	0.636	0.611	**0.710**	0.699	0.641	**0.710**
MOBILE	0.293	0.395	0.276	0.290	**0.328**	0.342	0.290	0.318	0.340	**0.346**	0.328	**0.346**
GEOSPATIAL	0.293	0.395	0.277	0.284	**0.287**	**0.344**	0.323	0.335	0.325	0.311	0.287	**0.344**

linear. This is statistically significant[7] ($p < 0.05$) for the non-sensitive survey features and the importance of these features are considered in the variable importance section below providing an avenue for consideration in the context of the problem domain and theory. For the Mobile and Geospatial data, the results show a less marked, but still consistent, improvement over theory-based models, with a similar trend towards non-linear classifiers. This indicates that further exploitation and feature engineering may be beneficial given the significantly reduced acquisition cost of these features.

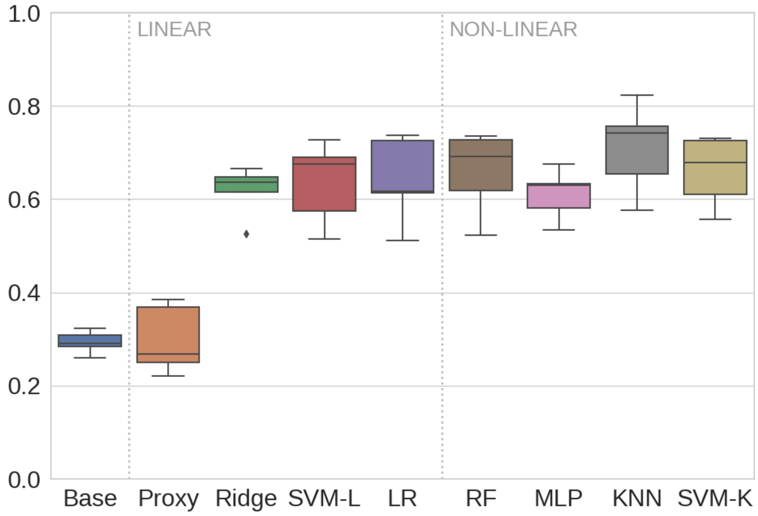
3-class results

The second section of Table 7.2 reports all results for the experiments for the 3-class problem, detailing the Macro-F_1 scores over each experimental run. A boxplot is included in Figure 7.1b for the non-sensitive survey dataset. Somewhat notably, the absolute values of the Macro F_1 score for the 3-class problem are higher than the F_1 score for the 2-class problem. This, however, is not completely unexpected as the 3-class classification avoids forcing the neutral responses of "neither agree nor disagree". This prevents subwards being arbitrarily interpreted as low risk in this work. Since the subwards that

[7] Based on conducting *corrected t-tests* as described in (Nadeau and Bengio, 2003, pg 251) which accounts for the data reuse inherent in cross-validation evaluations.



(a) Binary Model (High/Low risk)



(b) 3-class Model (High/Med/Low risk)

Figure 7.1: F_1 scores for optimised SURVEY based models.

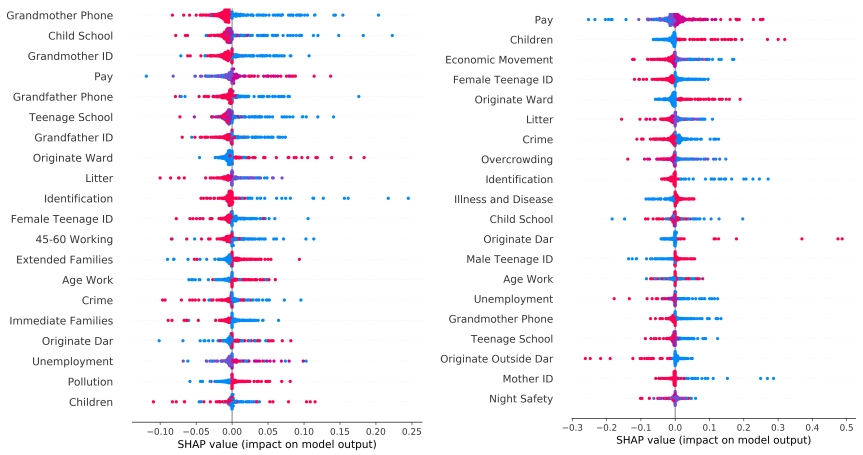
fall into this middle class share characteristics of both classes, and are likely to include respondents reticent to answer sensitive questions, a three-class system is preferred as it enables better per class predictions as highlighted by the macro F_1 score.

Analysis

Overall, results for the 3-class problem remain consistent with those of the 2-class design. Overall classification scores highly encourage the use of non-sensitive data to infer risk of forced labour. Machine learning models again outperform the theory-based proxy model fails to perform. In this case, however, rather than performing slightly better than the theory-based proxy model, the Mobile and Geospatial models perform slightly worse. Considering that models, based only on features from re-purposed data and not survey data, perform even as well as the theory-based proxy model, which incorporates features from across the datasets, indicates that there are likely features of utility in these datasets that warrant further exploration. Finally, similar to the 2-class problem non-linear models show increased performance over non-linear counterparts, with a statistically significant difference in the case of models based on non-sensitive survey features ($p < 0.05$).

2-class Variable Importance

Experimental results indicate that previously unconsidered covariates, modelled in non-linear fashion, can yield higher levels of predictive performance than current theory alone suggests. To further understand which of these features were important in explaining risk of forced labour,



(a) KNN 2 Class

(b) SVM 2 Class

Figure 7.2: Results (top 20 features) for best performing 2-class SURVEY based models. Red indicates high feature values, blue low feature values.

variable importance analysis was undertaken (an approach that has until recently been commonly neglected in applied AI research (Lipton, 2018)). To this end, SHAP values are leveraged, which measure the contribution of each feature to the overall prediction (given all features in the model) over and above a mean/majority predictor (Lundberg and Lee, 2017a). Given the significant improvements made within the models including non-sensitive questions (the SURVEY dataset) this work focuses on these, considering the covariates that are indicated as important in the two best performing non-linear models (kNN and Non-linear SVM). Two are considered in order to check for stability in the features reported as highly predictive.

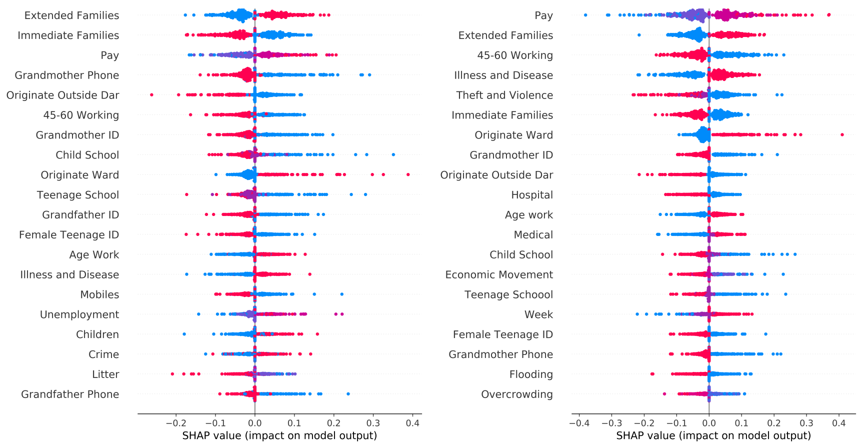
SHAP Values break down a prediction to show the impact of each feature. SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction that would be made if that feature took some baseline value. The SHAP values of all features sum up to explain why prediction are different from the baseline. The SHAP value figures in this chapter have been implemented using `SHAP` python package (Lundberg and Lee, 2017b). The resulting figures have three main things to considered: **Feature Importance** Features are ranked in descending order according to the overall influence they have on the predictions. **Impact** The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. **Original Value** The colour shows whether that variable is high (red) or low (blue) for that observation. Take for example figure 7.2 where pay if red and the SHAP value is high, this is showing that a high pay score is influencing a higher risk in slavery. In contrast the grandmother phone feature is showing that low scores of this feature (blue) are matched with high SHAP values, showing that less grandmothers having phones is a predictor for risk of slavery.

Figure 7.2a shows the SHAP values for the kNN model. Considering features theorised to be important, the results highlight that a low level of both children and teenagers in school is a strong pull feature for the risk of vulnerability to forced labour. Similarly, the level of identification documentation ownership was negatively correlated with risk. A high level of litter has a high, positive impact on the risk of vulnerability to forced labour. Considering non-theorised covariates, it can be seen that it is not just overall low ID ownership acting as a pull factor but low ID ownership for teenage girls and older men and women specifically. Furthermore, high levels of well-paid residents have a high, positive impact on the risk of vulnerability to forced labour. Additionally, high levels of people originating from either specifically that subward or other areas internal to the city, a majority of households being extended rather than immediate families, and local pollution as a more significant concern than local crime all also have a high and positive impact on the risk of vulnerability to forced labour.

Considering the non-linear SVM, 2-class model using categorical survey inputs (Figure 7.2b), it can be seen that many of the features agree with the KNN results. Pay, place of origin, ID ownership (specifically with the elderly), levels of litter, children and teenagers in school, all agree with the KNN results described above. The consistency in the impact of these variables in the non-linear models is a good indicator that these features need to be considered in future modelling. Additionally, the SVM SHAP analysis has shown, that in line with theory, high levels of overcrowding are exhibiting high, positive impacts on risk. In addition to prevailing theory and in line with the suggested proxy of residents origin, it can be seen that the level of economic movement is negatively correlated with risk. Areas of high pay and incoming migration link to forced labour risk is of particular note. While disingenuous offers of high pay have been widely theorized as a common tactic used by slave-masters to lure victims, this study provides evidence that risk can be generated without such agency - and simply that existence of urban environments attracts vulnerable individuals in the first place.

3-class Variable Importance

When moving to the multi-class analysis (equivalent SHAP plots shown in Figure 7.3), many of the feature importance remain consistent. Pay and origin, for example, remain of the upmost importance. Equally, lack of schooling for children and teenagers, lack ownership of IDs (particularly of the females and elderly), lack of mobiles and having extended family within households among others all remain essential pull factors for forced labour risk. However, particularly in the SVM, results begin to differ a bit, with increased impact from features relating to health and medical facilities. A high level of illness and disease has a high and positive impact on the risk of vulnerability to forced labour. Similarly, access to medical facilities is negatively correlated with the



(a) KNN 2 Class

(b) SVM 2 Class

Figure 7.3: Results (top 20 features) for best performing 3-class SURVEY models. Red indicates high feature values, blue low feature values.

target variable.

7.2 Discussion

The results in the proxy analysis in this chapter provide both empirical evidence regarding theorised proxy indicators as well as highlighting the utilization of machine learning techniques to isolate *alternate* proxy indicators. Established theoretical covariates, including a lack of identification documentation, lack of schooling and poor living conditions, were corroborated as being robust indicators of the risk of forced labour.

Forced labour remains an extremely sensitive area of research. During initial street survey pilots and review processes, it became evident that many of the traditional covariate questions for forced labour were extremely sensitive. However, people were much more willing to answer detailed questions about the subwards they lived in than their circumstances. High pay rates in a subward were found to be a significant proxy for increased slavery risk. Forced labour is known to affect all economic levels without being restricted to those living in poverty (Morgan and Olsen, 2014; Ruwanpura and Rai, 2004). The proximity of slums to sites of manual labour provides opportunities for exploitation, affecting the influence of levels of pay. Additionally, area of origin, a lack of economic movement and household family structure have emerged as strong proxies.

This work presents opportunities for practitioners to investigate these new proxy indicators, highlighting a number of pull factors, not only useful in the generation of vulnerability heat maps, but also to design risk reduction interventions. Current covariates appear insufficient partly due to the dynamic and varied nature of forced labour. It is therefore crucial to have an iterative loop between theoretical covariates and data-driven proxies. The results indicate unfortunately that there is no substitute for surveys within the currently available big data derived proxy features without extensive further investigation.

Modern slavery varies widely and will continue to exhibit different characteristic in different areas (Ruwanpura and Rai, 2004). It is a complex and commonly hidden problem, and one would, therefore, expect the pull factors increasing vulnerability to forced labour in an area to have complex interactions. It makes sense therefore, to expect proxy features to fit a non-linear model better. The results have supported this hypothesis. Non-linear models consistently outperformed equivalent linear models. The research emphasised the need for multidisciplinary research across technical fields and the modern-slavery domain.

Although the street survey produced rich, fine-grained ground truths, and the reduced sensitivity in questions made it easier to engage a broad range of survey respondents; the process was not without its challenges. The survey data collection required running multiple adaptive pilot surveys, obtaining approval letters to survey each ward region, organising training and the logistics of hundreds of facilitators travelling to and from the 452 subwards along with handling several unforeseen issues including floods, difficulties with local authorities and road accidents, highlighting the heavy resource nature of manual sampling in both time and money. Similar to other sampling efforts, it is difficult to replicate the data collection process, particularly considering the intricate hyper-local scale of data collection at subward level.

These challenges highlight the need for alternative and more targeted data collection methods. Despite the big data presented in this paper being outperformed by the survey data, the results are still encouraging. Big data can provide a significant improvement from baseline results. Mobile features, in particular, showed a promising influence on the models. A beneficial follow on to this work would be to expand the feature engineering of the big data sets in order to find potentially more meaningful proxy features.

7.3 Summary

Machine learning models demonstrated the possibility to predict vulnerability to forced labour with high levels of accuracy, showing significant improvements over theorised covariates alone. The importance of using non-linear models to account for the complex relationships occurring in the proxy covariates in particular has been highlighted. Traditional linear statistics used in slavery analysis are often not suited to detect such subtleties. However, the results indicate that there is no single substitute for surveys within the currently available big data derived proxy features. This work therefore provides the grounds for promising new research on feature engineering of transactional data to support the creation of meaningful proxies to support and identify target areas for survey data collection. In particular, I hope this work can provide useful insight for policymakers seeking to identify high vulnerability areas requiring action.

Chapter 8

Detecting Vulnerable Communities via the Stochastic Block Model

This section of research considers an application of the Stochastic Block Model to CDR data in Tanzania. The hypothesis underpinning this work is that the stochastic block model can be applied to CDR data to detect vulnerable communities. The block model is applied to four separate networks of Call Data Records: SMS records over the weekends, SMS records over the week days, Call Records over the weekends and Call Records over the week days. Communities are recovered from these networks by inferring the block structure using a point optimization of the stochastic block model. These recovered communities are then compared with the ground truth information collected in chapters 4,5 in order to assess the application of the stochastic block model for fine grained detection of communities vulnerable to poverty.

Notation

g = Block structure vector, such that $g_i = u$ if node i is in block u.

ω = Expected number of edges between the blocks such that, $\omega_{u,v}$ is the expected number of edges from the block u to block v.

G = Graph made up of the edges and nodes in the network, where n = the number of edges in G.

A = Adjacency Matrix (Size n x n), such that $A_{i,j}$ is the number of observed edges from node i to node j.

N = The total number of edges in a network.

$n_{u,v}$ = The number of observed nodes from block u to block v, given a block structure.

θ = Parameter to control the expected degree of the nodes.

d_i^+ = The total number of out degrees from node i.

d_i^- = The total number of in degrees attached to node i.

$D_{g_i}^+$ = $\sum_{i \in g_i} d_i^+$ which is the sum of all the out degrees in block g_i.

$D_{g_i}^-$ = $\sum_{i \in g_i} d_i^-$ which is the sum of all the in degrees in block g_i.

n_t = The sum of all out edges starting in block t such that $n_t = \sum_u n_{t,u}$.

K = The total number of blocks.

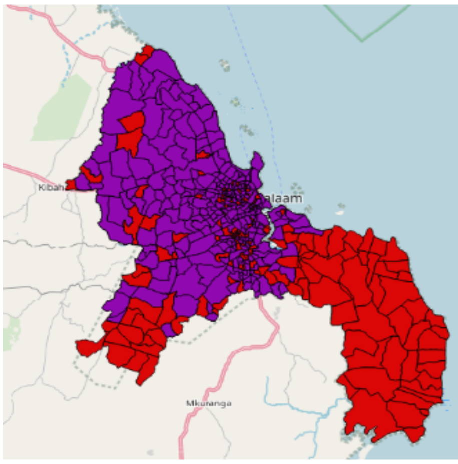
f_t^i = The fraction of neighbours of node i which belong to block t.

8.1 Experimental Set up

Data Cleaning

Despite often being infrastructurally poor, Tanzania is actually data rich. This work uses Call and SMS records provided by a leading network provider in Tanzania. Details on this data are included in the data section 3 at the start of this thesis. Initially the CDR data is aggregated to create two tables representing the calls and SMS interaction between towers which are within Dar es Salaam. The results are then filtered to remove outlying observations, if a tower had no interactions in or out for a period of time it is considered to be under repair and removed from the data. The cleaned data is then arranged into an adjacency matrix with each element representing the weighted interactions between nodes (cell towers). The network is then aggregated over the subward regions showing the number of connections (SMS and REC) to and from each subward. The network graph modelled by the adjacency matrix is weighted, directed and contains self-loops. All edges going outside of Dar es Salaam are taken out of the network to create a closed network of Dar es Salaam. Any subwards which do not have towers in appear red in the Figure 8.1 below, these subwards are not represented as nodes in the network.

The majority of the subwards without towers are in an area called Temeke which at the time of the Tigo data (2014) was governed differently to the rest of Dar es Salaam which might explain why they have used a different network provider for the majority of their regions. This was not as clear on a ward level because each ward is much more sparse than the subwards so have a few towers in. When looking at a subward level its clear that Temeke lacks CDR data so in order to avoid skewing the results, subwards in the region of Temeke are also removed from the network. This observation has been another benefit of working at a fine-grained subward level. After cleaning there are 556 towers in the data set and 289 subwards.



-

Figure 8.1: Subward Regions of Dar es Salaam for which CDR Data is Available, (Subwards without CDR towers: Red. Subwards with CDR towers: Purple)

After cleaning the remaining 289 subwards have on average 1.958 towers each, Figure 8.2 shows that there is a range of tower densities per subward. This is not averaged out in the model because the density of towers is representative of the density of cell activity, areas of high population require more cell coverage. There are more people using mobile phones in the centre of Dar es Salaam than in the rural areas. Choosing to aggregate at a subward level is just one of many options. I could, for example, have aggregated instead target variables to a tower level or accounted for the percentage of tower coverage within each subward. There are a few key reasons for aggregating at a subward level here. Firstly the target variable ground truths are at a subward level and I want to work at this governmental spacial level in order for the results to be useful for policy changers and organisations aiming to make changes in the vulnerable areas. Secondly, chapter 9 moves to a dynamic stochastic block model where the interaction times are an important addition to the network adjacency matrix, this renders percentage coverage no longer suitable. Keeping the aggregation level consistent between chapters 8 and 9 provides

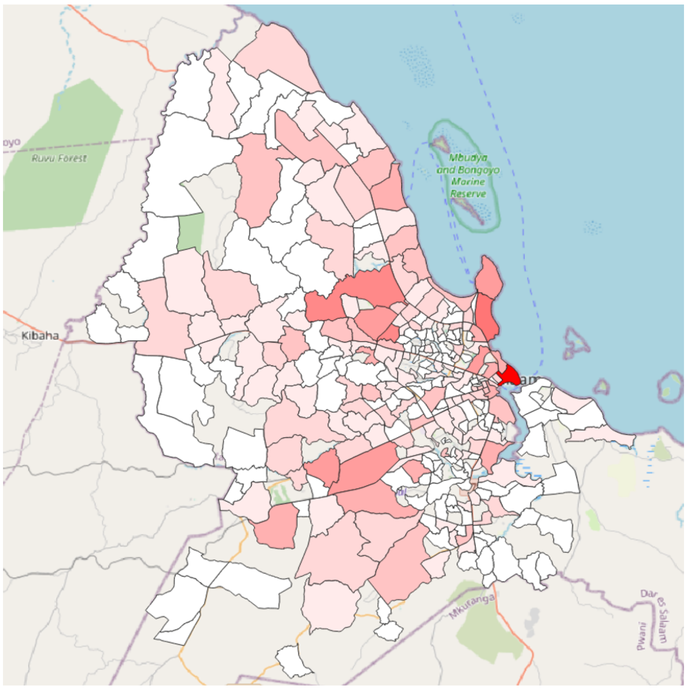


Figure 8.2: Number of Towers in Subwards Dar es Salaam. (White = 1, Dark Red = 14)

consistency for result comparisons. Additionally, it is beneficial to work at a subward level because the goal of the work is to provide information which could be used to underpin a number of projects outside of this work to increase impact. Additionally, the tower regions are not publishable geographically due to privacy concerns from the network provider, whereas the subward regions are the smallest governmental and commonly acknowledged regions.

Experimental Set up

The overall goal of this work is to apply a block model to CDR data such that the model produces blocks that are meaningful related to the social economic health of regions of Dar es Salaam. Within this there are various considerations about what are the input and target features of the model. The input features refer to the make up of the observed network inputted into the model for

Trial	**Edges**
1	Total number of SMS between subwards over all week days
2	Total number of SMS between subwards over all weekends
3	Total number of Calls between subwards over all week days
4	Total number of Calls between subwards over all weekends

Table 8.1: Experimental Trials for Stochastic Block Model

example looking at subward nodes with SMS or call edges. Target features refer to the different social economic health ground truth information gathered from the surveys, such as poverty and unemployment levels. Table 8.1 shows experimental set up of the total input features considered in this investigation.

SMS and Call data are looked at separately based on the hypothesis that different communities utilize the various services differently. For example, affluent areas are likely to have more mobility and therefore may require phone calls to allow for longer conversations. Short SMS messages might not be adequate to support social and professional relationships.

Another input feature which has been varied is the days. I have separated week days and weekends into different categories. This is to account for the expected alterations in daily activities. For example it may be that less affluent areas have less of a distinction work-wise between the weekend and week days. This could be due to the fact that much of the work in these areas will be informal and less limited to working hours ie fruit seller, street sellers and tuktuk drivers.

I have summed all the week and weekend days over the total 122 day period. I have chosen to do this rather than averaging to keep all the data in the model, this keeps the results consistent with chapter 9. In chapter 9 I will be looking at the event series of connections over time for all week and weekend days. Additionally I use a log transformation of the edge weights such that $A_{i,j}^{new} = log(A_{i,j}^{previous} + 1)$ to account for the large ranges within the adjacency matrix (Keene, 1995). This work aims to find which of these input variables can provide the most successful community vulnerability detection.

These experimental runs are individually tested against a range ground truths. All target variables in this analysis are created from the data collected in the street and grid survey in Dar es Salaam. There are three types of target variables used here; direct question targets, proxy targets and finally a grid survey target. The direct question targets are questions lifted directly from the street survey. Participants answers the questions, each equation is averaged at a subward level. Questions with more than three options (such as 5-scale likert questions) are grouped into high, medium and low risk groups. These results have aggregated into three groups to account for participant reluctance to answer on questions with extreme responses, which is a long acknowledged reaction to surveys (Guilford, 1954). Proxy targets are the linear combination of street survey features from literature based proxies of forced labour and poverty. The grid survey target is a 3 class, k-means clustering of the grid survey poverty rankings. A full list of target features can be seen in 8.2. Each of these target variables are individually compared to the stochastic block model results to determine which ground truths are best detected by the model. More information about the target features and the collection process can be seen in chapters 4,5.

8.2 Model Specifics

Poisson Stochastic Block Model

The Poisson model described and used in this chapter has been introduced

Grid Suvey Features	Poverty
Proxy Features	Poverty, Forced Labour
Direct Street Survey Questions	Overcrowding, Road Quality, Night Safety, Unemployment, Economic Movement, Medical Access, Theft and Violence, Arranged Marriages, Forced Labour

Table 8.2: Target Features Stochastic Block Model

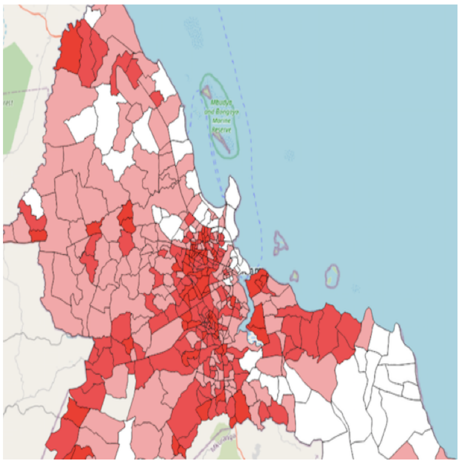
by (Karrer and Newman, 2011b), chapter 2.2 provides details on this model and its history. To model the number of connections between subwards, the number of expected edges between two nodes is modelled to be independently Poisson distributed based on only the blocks they are in. Hence the Poisson Block Model can be written as :

$$P(\mathbf{G} \mid \omega, \mathbf{g}) = \prod_{i<j} \frac{(\omega_{g_i,g_j})^{(A_{i,j})}}{A_{i,j}!} e^{-\omega_{g_i,g_j}} \prod_{i} \frac{(\frac{1}{2}\omega_{g_i,g_j})^{(\frac{A_{i,j}}{2})}}{\frac{A_{i,j}}{2}!} e^{-\frac{1}{2}\omega_{g_i,g_j}} \tag{8.1}$$

Equation (8.1) is the probability of graph $\mathbf{G}$ given the block structure and expected number of edges between blocks for an un-directed weight graph with self-loops. Here $A_{i,j} = A_{j,i}$ and $\omega_{u,v} = \omega_{v,u}$ where $A_{i,i}$ represents 2 times the number of edges from i to i. This is the probability which will be maximized to infer an optimal block structure. It is important to use the Poisson model because of a potential 83521 connections between each of the 289 subwards, there is only 102 connection with no calls between them. An unweighted model would lose the information showing the different number of connections between subwards.

Degree Corrected Poisson Stochastic Block Model

Before inferring the block structure it is important to consider the specifics of the network available from the CDR data. There is a vast variation in the



-

Figure 8.3: Subwards in Dar es Salaam, illustrating the variety of average daily outward connections (Red: Higher average number of daily outward connections

out degree of connections between the different subwards as seen in Figure 8.4. This figure is describing features directly from the data providing evidence for the choice to model with degree correction. Many of the subwards with high outward degrees are in the centre of Dar es Salaam, as seen in Figure 8.3. Without degree correction all the subwards in the centre of the city with the highest outward degree would be blocked together. For this reason a degree correction Poisson stochastic block model as shown in equation (8.2) (Karrer and Newman, 2011b) is used:

$$P(\mathbf{G} \mid \omega, \mathbf{g}, \theta) = \prod_{i<j} \frac{(\theta_i\theta_j\omega_{g_i,g_j})^{(A_{i,j})}}{A_{i,j}!} e^{-\theta_i\theta_j\omega_{g_i,g_j}} x \prod_{i} \frac{(\frac{1}{2}\theta_i^2\omega_{g_i,g_j})^{(\frac{A_{i,j}}{2})}}{\frac{A_{i,j}}{2}!} e^{-\frac{1}{2}\theta_i^2\omega_{g_i,g_j}} \tag{8.2}$$

Here there is a new parameter θ which controls the expected degree of the nodes such that the expected number of edges between node i and node j now equals $\theta_i\theta_j\omega_{g_i,g_j}$. It is assumed that $\sum_{i\in g_i} \theta_i = 1$ for all blocks. Figure 8.4 shows that all data trials being used have significant degree variation between nodes. The number of edges between subwards are particularly varied in the

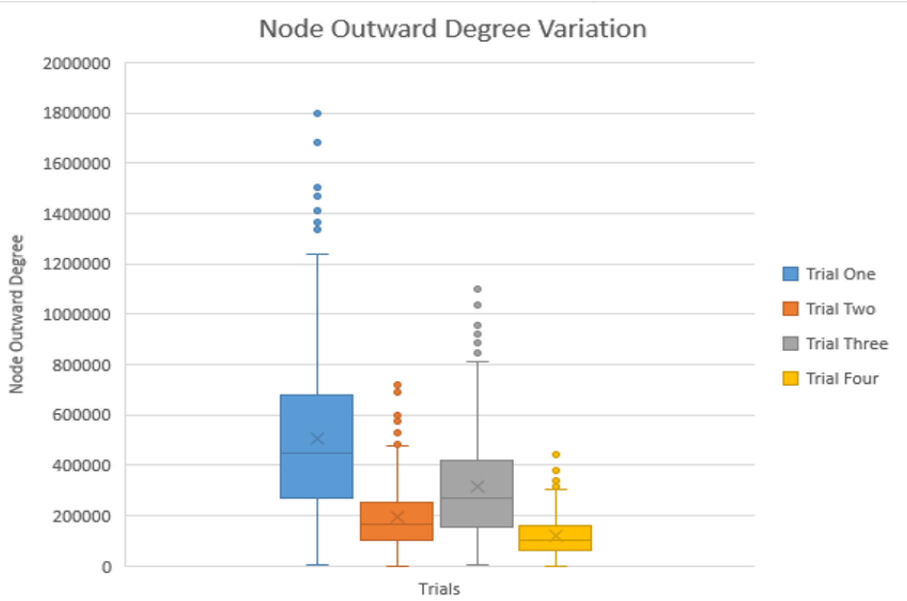


Figure 8.4: Data Summary, SMS and Call out degree variation between subward (Trial One = Total number of SMS between subwards over all week days; Trial Two = Total number of SMS between subwards over all weekends; Trial Three = Total number of Calls between subwards over all week days; Trial Four = Total number of Calls between subwards over all)

case where edges are represented by the number of week day SMS connections.

Directed Degree Corrected Poisson Stochastic Block Model

Another thing to consider before inferring block structure is that the connections are directed, there is a calling subward and a called subward for each connection (edge). From this point the directed case is used where self loops are allowed. $A_{i,j}$ is not necessarily equal to $A_{j,i}$ and $\omega_{u,v}$ is not necessarily equal to $\omega_{v,u}$. It is worth noting that the self-edges in an undirected network are by correction written in the adjacency matrix as 2 x the number of edges (so the sum of the degrees of all the vertices is twice the number of edges). This correction is convenient because the various formulas hold whether or not undirected networks have self-edges. For example $A_{i,j} = A_{j_i}$ and so appears in the matrix twice. Now directed edges are considered this is not the case so it is not necessary for the leading diagonal to be represented with doubled counts, instead $A_{i,i}$ = the actual number of edges from node i back to itself. The majority of work done on stochastic block model looks at either directed edges without self loops or, undirected edges with (or without self loops). The following is a modification of the undirected stochastic block model in equation (8.2) by (Karrer and Newman, 2011b) to account for directed edges:

$$P(\mathbf{G} \mid \omega, \mathbf{g}) = \prod_{i,j} \frac{(\omega_{g_i,g_j})^{(A_{i,j})}}{A_{i,j}!} e^{-\omega_{g_i,g_j}} \tag{8.3}$$

Before adding the degree correction it is important to consider direction within the degree correction parameter. Where before θ was the overall degree correction, now θ^+ and θ^- are used to account for the outward and inward degrees of nodes. Given this (8.3) becomes:

$$P(\mathbf{G} \mid \omega, \mathbf{g}, \theta_i^+, \theta_j^-) = \prod_{i,j} \frac{(\theta_i^+, \theta_j^- \omega_{g_i,g_j})^{(A_{i,j})}}{A_{i,j}!} e^{-\theta_i^+, \theta_j^- \omega_{g_i,g_j}} \tag{8.4}$$

Using the assumption $\sum_{i \in g_i} \theta^+ i = \sum_{i \in g_i} \theta^- i = 1$ so $\omega_{u,v}$ will correspond to the average number of directed edges between groups. From here the number of blocks K, is assumed know, I choose to fix K to equal 3. This is done for consistency with the number of groups in the ground truth targets. This also has the benefit of reducing the complexity of the model which is important here as it makes the work easier to communicate with non technical audiences which is essential given the context of the work. Given this information the conditional probability (8.4) can be simplified as:

$$P(\mathbf{G} \mid \omega, \mathbf{g}, \theta_i^+, \theta_j^-) = \prod_i (\theta_i^+)^{d_i^+} (\theta_i^-)^{d_i^-} \prod_{i,j} \frac{1}{A_{i,j}!} \prod_{u,v} \omega_{u,v}^{n_{u,v}} e^{-\omega_{u,v}} \tag{8.5}$$

in which $n_{u,v}$ is the total number of edges between groups u and v. d_i^+ is the total amount of out degrees from node i and d_i^- is the total number of in degrees attached to node i. To maximize this, first the logarithm of this probability is taken and then it is simplified and constants can be removed.

$$\ln(P(\mathbf{G} \mid \omega, \mathbf{g}, \theta^+, \theta^-)) = \sum_i d_i^+ \ln \theta_i^+ + \sum_i d_i^- \ln \theta_i^- + \sum_{u,v} (n_{u,v} \ln(\omega_{u,v}) - \omega_{u,v}) \tag{8.6}$$

The constants are removed because regardless of block state or degree correction they will be the same, they will make no difference when maximizing this logarithm probability. From (8.6) the maximum likelihood values of the parameters $\omega, \theta^+, \theta^-$ can be derived. First consider:

$$\frac{d}{d\omega_{u,v}}(\ln(P(\mathbf{G} \mid \omega, \mathbf{g}, \theta^+, \theta^-))) = \frac{n_{u,v}}{\omega_{u,v}} - 1 \tag{8.7}$$

Then setting the derivative function to zero :

$$(\frac{n_{u,v}}{\omega_{u,v}} - 1) = 0 \tag{8.8}$$

This can be rearranged such that:

$$\widehat{\omega}_{u,v} = n_{u,v} \tag{8.9}$$

As θ^+, θ^- are constrained by $\sum_{i \in g_i} \theta_i^+ = \sum_{i \in g_i} \theta_i^- = 1$
Likelihood estimators for θ^+, θ^- can be found by solving $\frac{d}{d\theta_i^+}$, $\frac{d}{d\theta_i^-}$ via Lagrange Multipliers. The number of Lagrange Multipliers λ_{g_i} needed is proportional to the number of blocks. The system that needs solved is as follows:

$$\frac{d_i^+}{\theta_i^+} = \lambda_{g_i} \tag{8.10}$$

$$\frac{\sum_{i\in g_i} d_i^+}{\lambda_{g_i}} = 1 \tag{8.11}$$

Hence,

$$\lambda_{g_i} = \sum_{i\in g_i} d_i^+ \tag{8.12}$$

Substitute this into (8.10) and rearrange to give the likelihood estimator:

$$\widehat{\theta_i^+} = \frac{d_i^+}{D_{g_i}^+} \tag{8.13}$$

Where $D_{g_i}^+ = \sum_{i\in g_i} d_i^+$ which is the sum of outward degrees in the block which node i is in. By the same logic and given $D_{g_i}^- = \sum_{i\in g_i} d_i^-$ there is now the following estimators: $\omega, \theta^+, \theta^-$ such that $\widehat{\omega} = n_{u,v}$, $\widehat{\theta_i^+} = \frac{d_i^+}{D_{g_i}^+}$ and $\widehat{\theta_i^-} = \frac{d_i^-}{D_{g_i}^-}$. Estimates can now be substituted into (8.6) giving the maximum as:

$$\ln(P(\mathbf{G} \mid \omega, \mathbf{g}, \theta^+, \theta^-)) = \sum_i d_i^+ \ln\frac{d_i^+}{D_{g_i}^+} + \sum_i d_i^- \ln\frac{d_i^-}{D_{g_i}^-} + \sum_{u,v}(n_{u,v}\ln n_{u,v} - n_{u,v}) \tag{8.14}$$

Allowing N to be the total number of edges in the network this can be rearranged to give.

$$\ln(P(\mathbf{G} \mid \omega, \mathbf{g}, \theta^+, \theta^-)) = \sum_i d_i^+ \ln\frac{d_i^+}{D_{g_i}^+} + \sum_i d_i^- \ln\frac{d_i^-}{D_{g_i}^-} + \sum_{u,v}(n_{u,v}\ln n_{u,v}) - N \tag{8.15}$$

The first two terms of this are manipulate given the fact that $D_u^+ = \sum_{i\in g_i} d^+ =$

$\sum_v n_{u,v}$ and $D_u^- = \sum_{i \in g_i} d^- = \sum_u n_{u,v}$ to write:

$$\begin{aligned}
&\sum_i d_i^+ \ln \frac{d_i^+}{D_{g_i}^+} + \sum_i d_i^- \ln \frac{d_i^-}{D_{g_i}^-} \\
&= \sum_i d_i^+ \ln d_i^+ + \sum_i d_i^- \ln d_i^- - \sum_u (D_u^+ \ln D_u^+ + D_u^- \ln D_u^-) \\
&= \sum_i d_i^+ \ln d_i^+ + \sum_i d_i^- \ln d_i^- - \sum_u (\sum_v n_{u,v} \ln D_u^+ + \sum_u n_{u,v} \ln D_u^-) \\
&= \sum_i d_i^+ \ln d_i^+ + \sum_i d_i^- \ln d_i^- - \sum_{u,v} n_{u,v} \ln D_u^+ D_v^-
\end{aligned} \tag{8.16}$$

Substituting this back into (8.15) and ignoring the constant terms gives:

$$L(\mathbf{G} \mid \mathbf{g}) = \sum_{u,v} (n_{u,v} \ln \frac{n_{u,v}}{D_u^+ D_v^-}) \tag{8.17}$$

The larger this objective function is the better the block structure g is. This allows us to infer the optimum block structure by optimizing this function. The goal now is to optimize this log-likelihood function with respect to g. This is done using the Metropolis-Hastings MCMC algorithm for optimization, a method of point estimation known to work well for block models (Peixoto, 2014a, 2013). After running initial investigatory results using the python package *Graph$_T$ool* the final results in this chapter have been implemented using python code written from scratch during this PhD.

Inferring Block Structure with Metropolis-Hastings Optimization

The Metropolis-Hastings method for optimization is an optimized Markov chain Monte Carlo algorithm proposed to reduce the partition space preventing it from growing exponentially with the number of blocks. (Peixoto, 2014a) Firstly every node is given a random block assignment, then for a given number of iterations a node tries to move from one block to another with a probability conditional on an adjacent nodes block assignment. Each node move is restrict so that no blocks can be empty and no new blocks can be formed. The conditional probability of moving a node from block u to block v depending on the block assignment t belonging to a neighbouring node is referred to as the proposal function which is as follows:

$$p(u \to v \mid t) = \frac{n_{t,v} + \epsilon}{n_t + \epsilon K} \tag{8.18}$$

Here n_t is the sum of all out edges starting in block t such that $n_t = \sum_u n_{t,u}$, K is the total number of blocks and ϵ is a free parameter. Free parameter ϵ is such that $\epsilon > 0$ allows the possibility of any potential block structure to be reached from any block structure starting point in a finite number of iterations. Smaller values of ϵ increases the chances that an accepted move in the iteration increases the log-likelihood function, moves are more likely to move the nodes into their correct blocks. Larger values of ϵ however make it more likely for unproductive moves to be accepted, thus taking more computational resource. Larger ϵ values however allow for more randomness so can help to get out of local optimums. Relatively small values of ϵ will reach a convergence point quicker than larger values. I choose the average number of edges between nodes as ϵ, as ϵ is a free parameter naturally scaled to the size of the network this choice has been made empirically. The idea behind (8.18) is that if there are two blocks u and v with a large number of edges between them, then a node with a lot of neighbours in block v is likely to be in block u. Using the proposal function (8.18) and the objective function (8.15) which is now defined as L, the Metropolis-Hastings acceptance probability is:

$$min\left[e^{(\Delta L)} \frac{\sum_t f_t^i p(v \to u|t)}{\sum_t f_t^i p(u \to v|t)}, 1\right] \tag{8.19}$$

Where f_t^i is the fraction of neighbours of node i which belong to block t. It is important to note that $p(v \to u \mid t)$ is calculated after node i is moved from block u to block v. $p(u \to v \mid t)$ however is calculated before this move. The number of iterations needed depends on $(NumberofNodes)^2$ and the initial starting point of the block structure. This optimization acceptance probability is known to increase the likelihood without getting too stuck in local optima (Peixoto, 2014a, 2013).

Illustrative Example

This illustrative example starts with the following information: a set of 10 nodes $[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$; an initial block structure vector $g = [0, 0, 0, 0, 1, 1, 2, 2, 2, 2]$ (showing the block membership of each node); and an initial block matrix ω 8.5 (showing the expected number of edges between the blocks). Using this initial information and *GraphTools* (Peixoto, 2014b) (a python package for stochastic block models) a network is generated from the degree corrected stochastic block model. The generated network is directed and contains self-loops.

Figure 8.5 shows the initial block matrix ω with the expected number of edges between the 3 blocks. Figure 8.6 shows the generated network of edges from the initial block structure. The network is generated from the degree corrected stochastic block model. Here the block structure is successfully recovered using Metropolis-Hastings optimization of the objective function for directed degree corrected block models (8.15). The adjacency matrix of the generated network is inputted into the Metropolis-Hastings algorithm for optimization 20 times, each starting with a different random block structure.

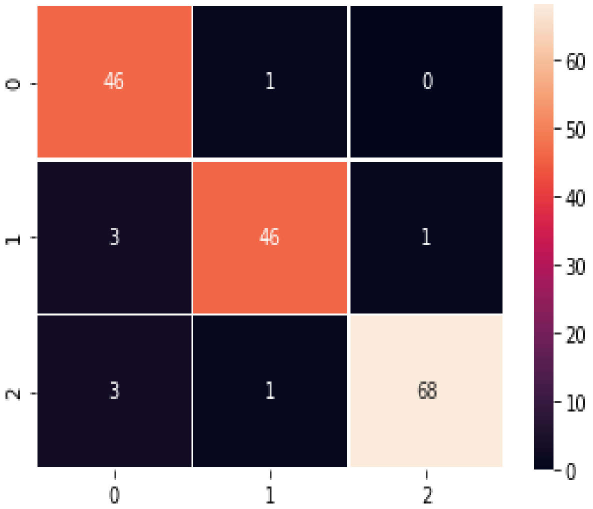


Figure 8.5: Expected Number of Edges between Blocks set for the Illustrative Example

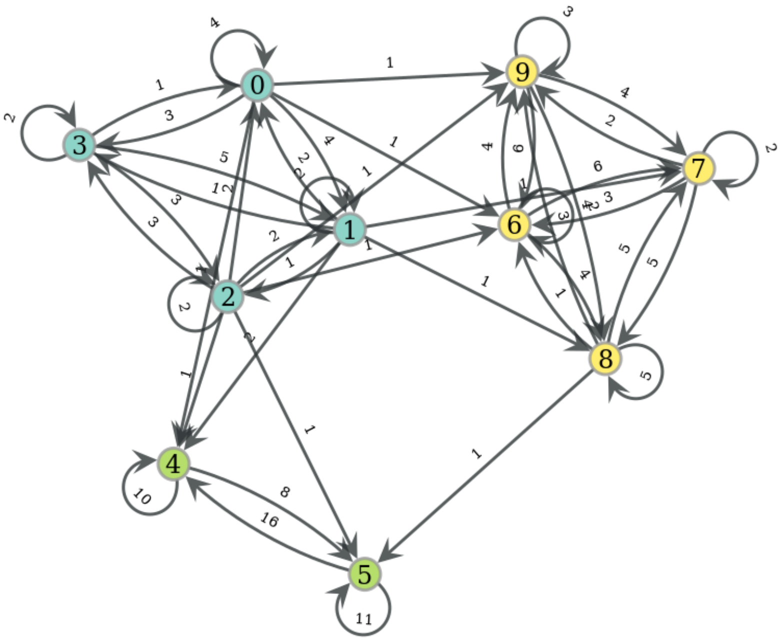


Figure 8.6: A generated network from the degree corrected stochastic block model, coloured by block structure.

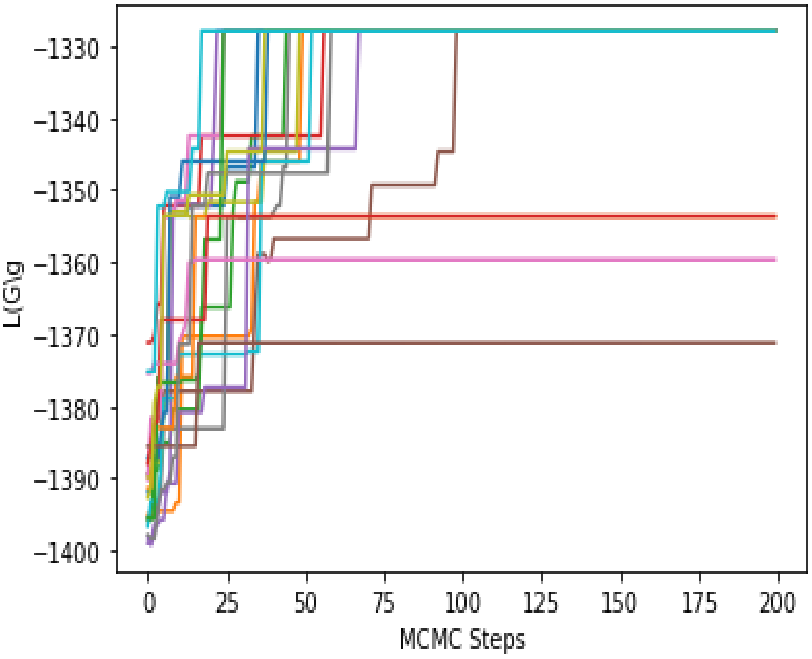
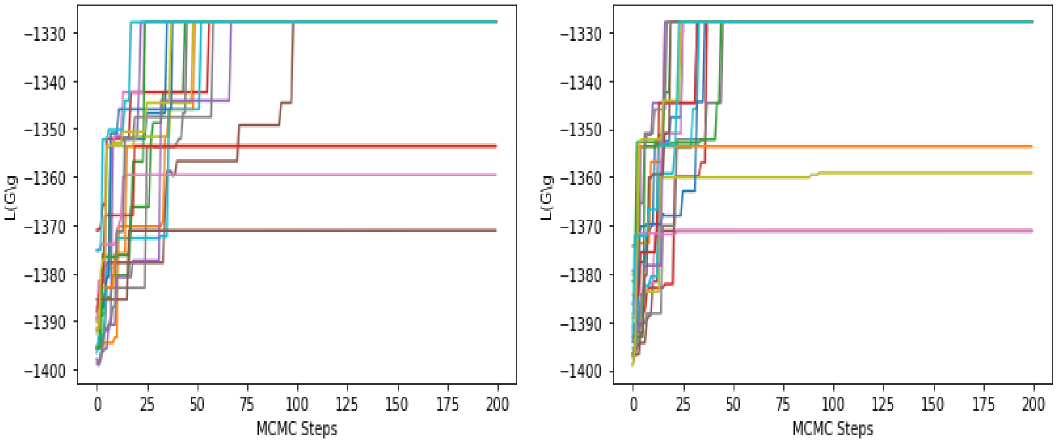


Figure 8.7: 20 separate cases of MCMC optimization algorithm to maximize the likelihood of the stochastic block model with regard to different block structures g.

Of the 20 MCMC cases 17 recover the original block structure of the generated network, the other 4 get trapped at local optima, which emphasises the importance of running multiple cases. Another important thing to note is that while the block structure has successfully been recovered, the labels of the blocks will differ. The process is simply recovering the partition space of the nodes. For example both $g = [0, 0, 0, 0, 2, 2, 1, 1, 1, 1]$ and $g = [1, 1, 1, 1, 2, 2, 0, 0, 0, 0]$ are finishing points of different cases which have both successfully recovered the block structure of the nodes from the original generation parameter $g = [0, 0, 0, 0, 1, 1, 2, 2, 2, 2]$ and have the same $L(G \mid g)$ value. In other words the model is invariant to the particular labelling of different blocks.

It is worth noting that while Metropolis-Hastings is usually referred to as a method for posterior simulation, in the context of this thesis it is used for



(a) 20 MCMC optimization cases with ϵ = 20

(b) 20 MCMC optimization cases with ϵ = 0.1

Figure 8.8: 20 separate cases of MCMC optimization algorithm with different ϵ to maximize the likelihood of the stochastic block model with regard to different block structures g.

optimization. The Metropolis-Hastings is used as a sampler with independent runs where the best value over all runs is taken (in terms of log posterior). The decision to use this method was taken as the collaborators of the project were familiar with this optimization procedure. Given the potential ground impact of the results it was important to make all results as communicable as possible.

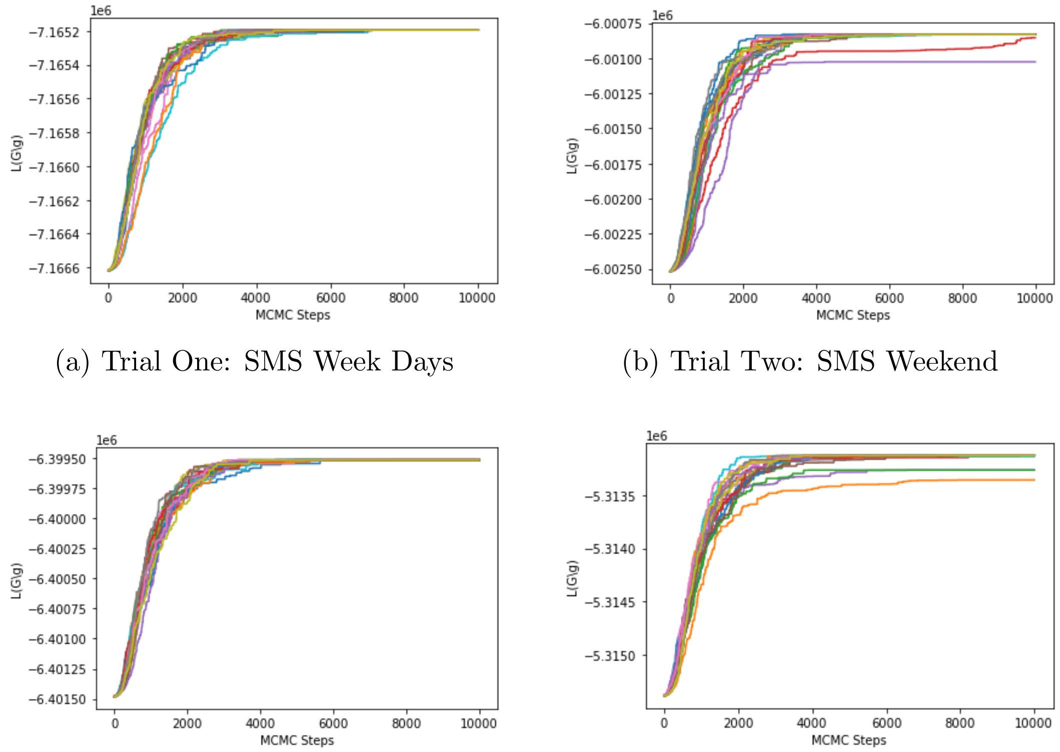
Effects of changing parameter ϵ

We now illustrate the affects of changing the algorithm parameter ϵ in the Metropolis-Hastings algorithm. The size of this parameter is relative to the number of edge. Using the sample network generated in the example above, the average number of edges between nodes is 0.3825.

Figure 8.8 shows the trace of 20 cases and starting points of the algorithm with $\epsilon = 20$, $\epsilon = 0.1$ respectively. The smaller value of ϵ takes fewer steps for the optimization, because it is less likely too accept unproductive moves and thus can save computational resources. Here both extremes of choice for ϵ large and small each successfully recover the block structure on the majority of the cases. On a larger network you would see the increased randomness from larger ϵ values allowing the algorithm to avoid local optima. It is important to balance the need for randomness with the desire to reduce computational cost, for this reason I chose to use the average number of edges between nodes of a network as ϵ going forward.

8.3 Applying the block model to Novel Data Streams Results

In order to balance the performance of results and computational cost, the optimization algorithm is performed with 20 different cases each with a different random starting points g over 10,000 steps. The number of blocks is set as $K = 3$ and ϵ is set to be average number of connections in the observed adjacency matrix. This set up was used for the four data trials described in table 8.1. For example in trial one the model notation refers to the following information: the set of nodes is the 289 subward; the adjacency matrix A represents the log transformation of the total number of SMS connections between subwards over all week days, such that $A_{i,j} = log(numberofweekdaySMSconnectionsfromsubwarditosubwardj + 1)$; the block matrix ω represents the expected number of edges between subward; and g is the block structure vector, such that g_i is the block assignment of subward i.



(a) Trial One: SMS Week Days

(b) Trial Two: SMS Weekend

(c) Trial Three: Call Week Day

(d) Trial Four: Call Weekend

Figure 8.9: 20 cases of MCMC optimization for each of the 4 data trial options

Figure 8.9 shows the MCMC optimization steps over 20 cases (each starting with a different random g) for each of the four data trials. Trial One, with an adjacency matrix of week day SMS connections between subwards finished with 19 runs on a maximum finishing point and the other 1 at a local optimum. Trial Two, with an adjacency matrix of weekend SMS connections finished with 17 runs on a maximum finishing point and the other 3 at 2 different local optimum points. Trial three, with an adjacency matrix of week day call connections finished with 17 runs on a maximum finishing point and the other 3 at a single local optimum. Trial four, with an adjacency matrix of weekend call connections finished with 16 runs on a maximum finishing point and the other 4 at 3 local optimum points.

The following results summarise the Normalized Mutual Information score

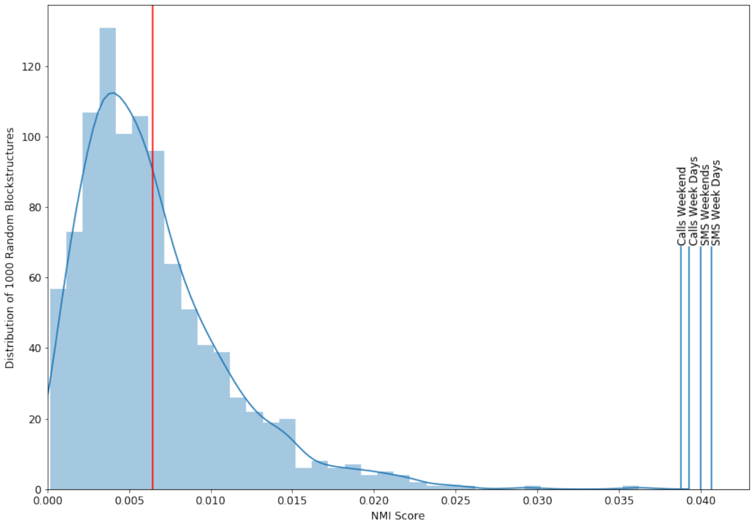


Figure 8.10: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the grid survey, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

(NMI) of the block model trials against the target values. Recall from chapter 2.2 that NMI is the method of measuring the quality of a generated clustering in comparison to a ground truth clustering. This is a widespread score measure for stochastic block modules. NMI values range between 0 (no mutual information) to 1 (perfect correlation).

Grid Survey Poverty

The NMI scores between the grid survey ground target and the data trials are as follows: Trial One: 0.0407, Trial Two: 0.0401, Trial Three: 0.0393, Trial Four: 0.0388. At first look these results seem disappointingly low, however it is unrealistic to assume a model will produce perfect results uncovering the ground truth structure. However, we can consider whether the application of the stochastic block model to novel data streams has produced results which are better than random block structures.

To do so, results are evaluated in comparison to 1000 random block structures. For each target variable, subwards are randomly blocked into three groups one thousand times. The thousand random structures are then scored with the NMI measure against the ground truth targets. The distribution of these random NMI results are then plotted, with their averages highlighted. The CDR call and SMS trial block model results are then labelled on the plot to assess if an improvement from random has been achieved.
results are evaluated in comparison to 1000 random block structures. For each target variable, subwards are randomly blocked into three groups one thousand times. The thousand random structures are then scored with the NMI measure against the ground truth targets. The distribution of these random NMI results are then plotted, with their averages highlighted. The CDR call and SMS trial block model results are then labelled on the plot to assess if an improvement from random has been achieved.

Figure 8.10 shows the first example of the NMI plots. All four CDR trials have produced NMI results well above the random average of 0.0064. Both SMS trials have done slightly better at recovering the ground truth block structure than the call trials. The success of these results detecting the areas of poverty in Dar es Salaam is particularly important as the grid survey has produced the most reliable estimation of poverty spread across the city.

Proxy Targets

Proxy Poverty

The NMI scores between the proxy poverty target (defined above in the experimental set up and in chapter 5) and the data trials are as follows: Trial

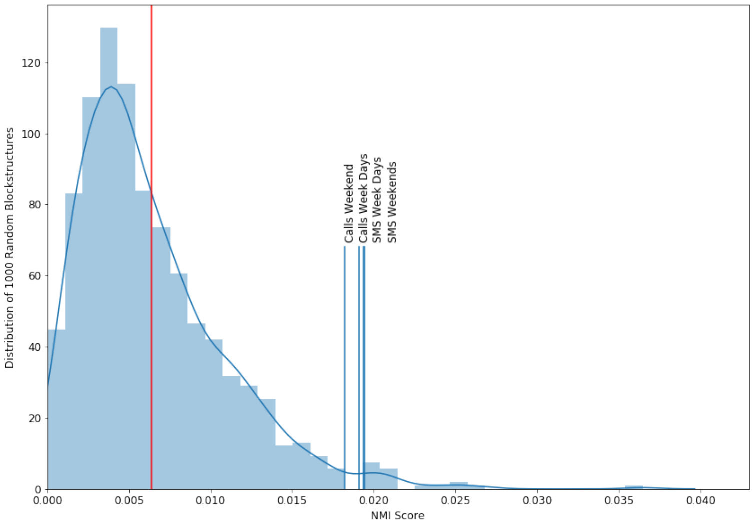


Figure 8.11: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the proxy poverty scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

One: 0.0205, Trial Two: 0.0216, Trial Three: 0.0247, Trial Four: 0.0198. Again the trial results can be seen to be well above the random average of 0.0065 in Figure 8.12. Although the difference from the random average is slightly reduced here, the trial results are still well into the tail end of the random distribution.

Proxy Forced Labour

Chapter 5 explained the extensive challenges researchers face trying to estimate and detect levels of forced labour. The hidden nature of the issue exasperates data collection challenges. As such any information on this area is highly valuable. The NMI scores between the proxy forced labour target and the data trials are as follows: Trial One: 0.01941, Trial Two: 0.01936, Trial Three: 0.01912, Trial Four: 0.01820. Figure 8.11 shows a random average of 0.0063, with each of the trials above this line.

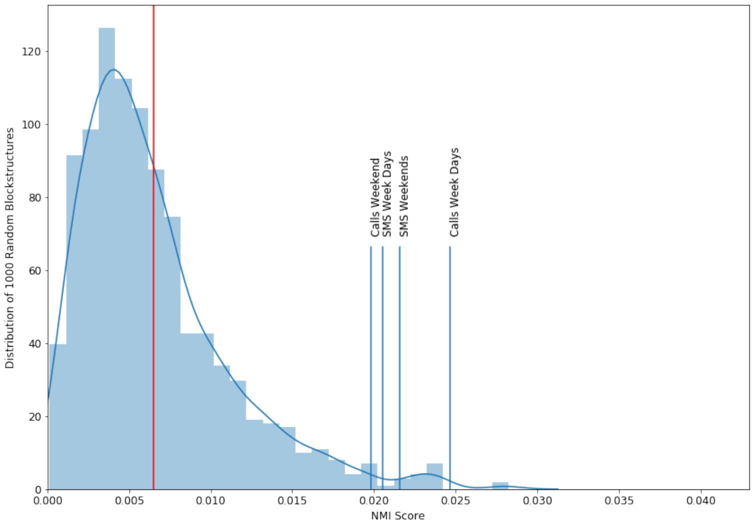


Figure 8.12: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the proxy forced labour scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

Street Survey Direct Features

Overcrowding

The NMI scores between the overcrowding ground truth target and the data trials are as follows: Trial One: 0.0083, Trial Two: 0.0078, Trial Three: 0.0089, Trial Four: 0.0078. The trial results shown in Figure 8.13 are only slightly above the random average here and within the body of the distribution that rises from completely random block assignments. Although application of stochastic block models do not produce useful results detecting overcrowded subwards, overcrowding is not a sensitive topic, this could easily be observed using satellite imagery or traditional survey methods.

Road Quality

The NMI scores between the road quality ground truth target and the data trials are as follows: Trial One: 0.0175, Trial Two: 0.0197, Trial Three: 0.0192,

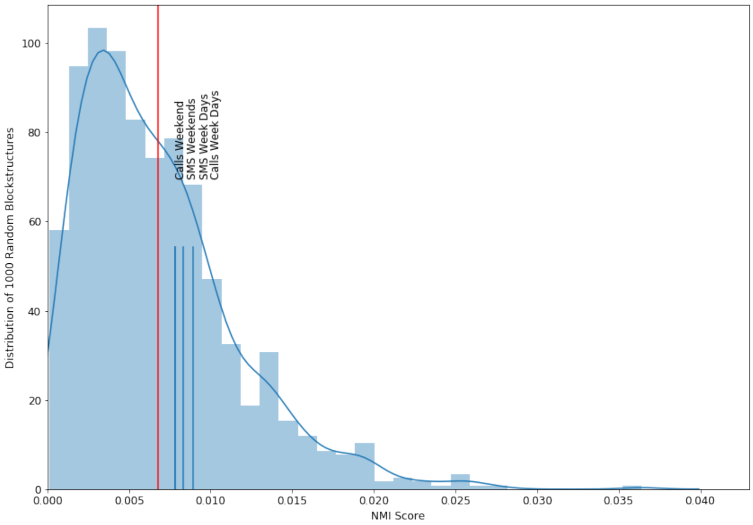


Figure 8.13: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey overcrowding scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

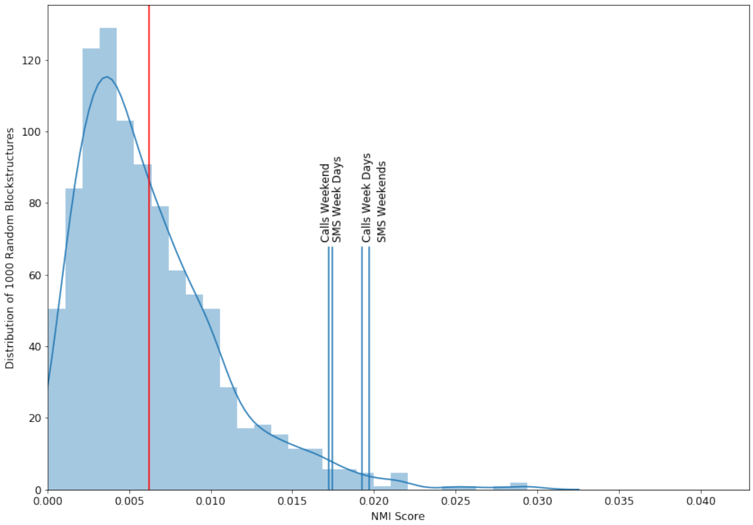


Figure 8.14: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey road quality scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

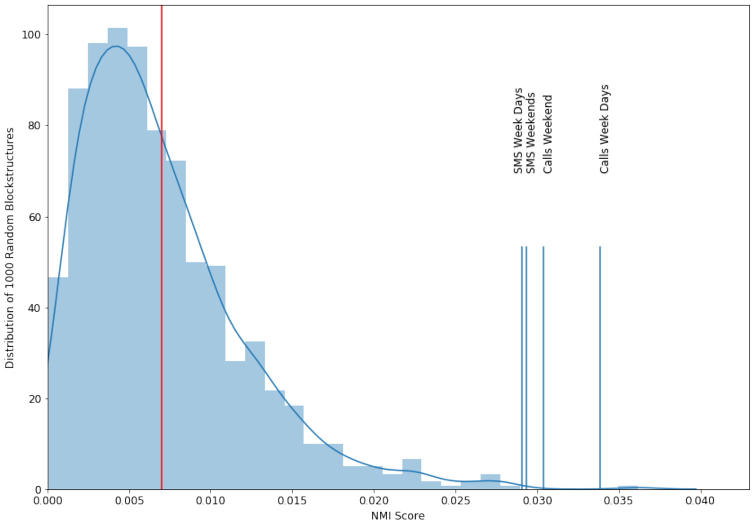


Figure 8.15: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey night safety scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

Trial Four: 0.0172. The trial results shown in Figure 8.14 are well above the random average of 0.0062.

Night Safety

The NMI scores between the night safety ground truth target and the data trials are as follows: Trial One: 0.02907, Trial Two: 0.0294, Trial Three: 0.0338, Trial Four: 0.0304. How safe people feel at night in an area is a good reflection of what the area is like. Knowing where people feel most unsafe can help governments and other aid organisation with limited resources best allocate where to put more measure in place to keep people safe. For this reason the improvement of the trials compared with the random average 0.0070 as shown in Figure 8.15 indicated this is a potentially useful feature.

Unemployment

The NMI scores between the unemployment ground truth target and the data

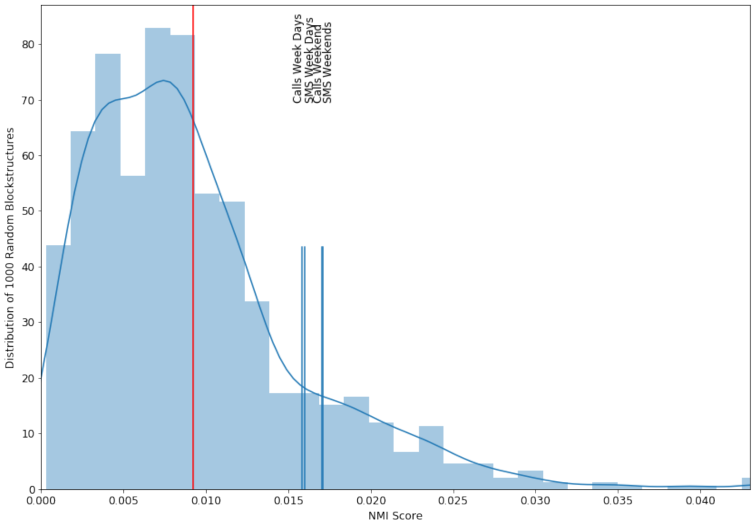


Figure 8.16: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey unemployment scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

trials are as follows: Trial One: 0.0159, Trial Two: 0.0171, Trial Three: 0.0158, Trial Four: 0.0170. Figure 8.16 shows these trial results to be above the random average 0.009, although not as much as many of the other results.

Economic Movement

The NMI scores between the economic movement ground truth target and the data trials are as follows: Trial One: 0.01623, Trial Two: 0.0173, Trial Three: 0.0162, Trial Four: 0.0174. Figure 8.17 shows these trial results to be above the random average 0.006, although again not as much as many of the other results.

Medical

The NMI scores between the medical availability ground truth target and the data trials are as follows: Trial One: 0.0071, Trial Two: 0.0071, Trial Three: 0.0061, Trial Four: 0.0060. Figure 8.18 shows these trial results make no real

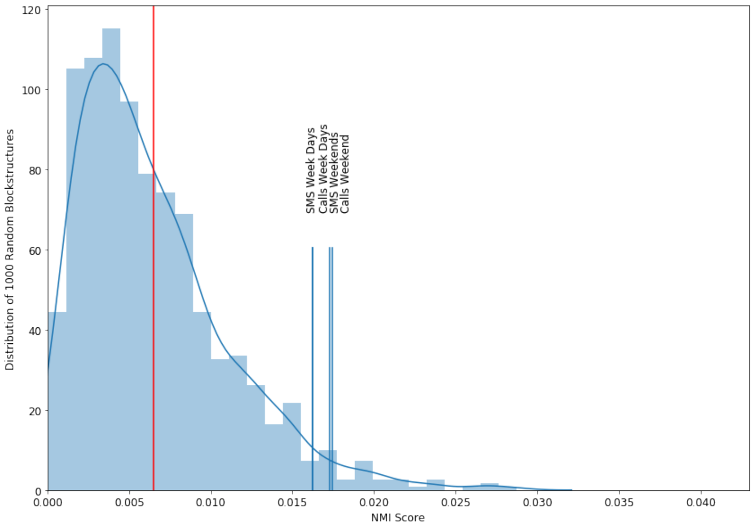


Figure 8.17: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey economic movement scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

improvement from the random average 0.006. Applying this model to the CDR data does not produce any useful results for this target feature.

Theft and Violence

The NMI scores between the theft and violence ground truth target and the data trials are as follows: Trial One: 0.0147, Trial Two: 0.0149, Trial Three: 0.0169, Trial Four: 0.0146. Figure 8.19 shows these trial results are greater than the random average 0.006.

Pay

The NMI scores between the pay levels ground truth target and the data trials are as follows: Trial One: 0.0120, Trial Two: 0.0110, Trial Three: 0.0098, Trial Four: 0.0123. Figure 8.20 shows these trial results are only slightly above the random average 0.007.

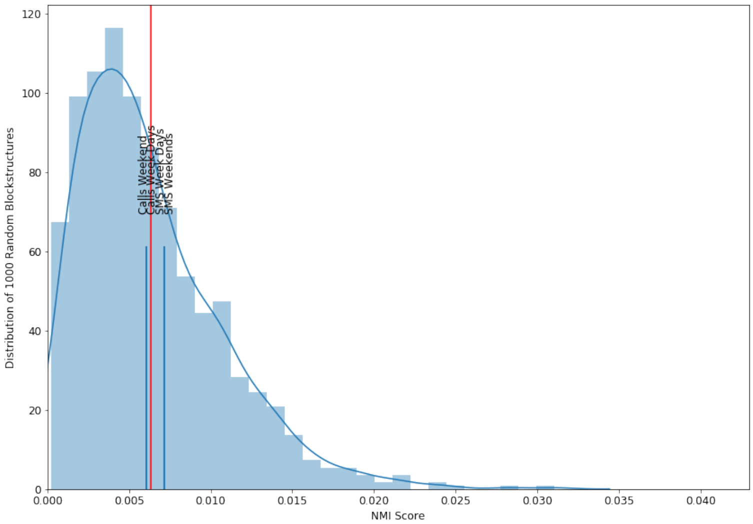


Figure 8.18: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey medical scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

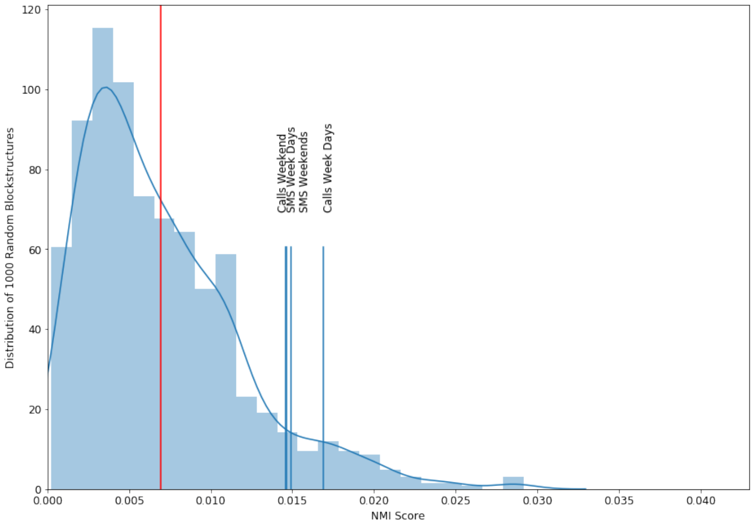


Figure 8.19: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey theft and violence scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

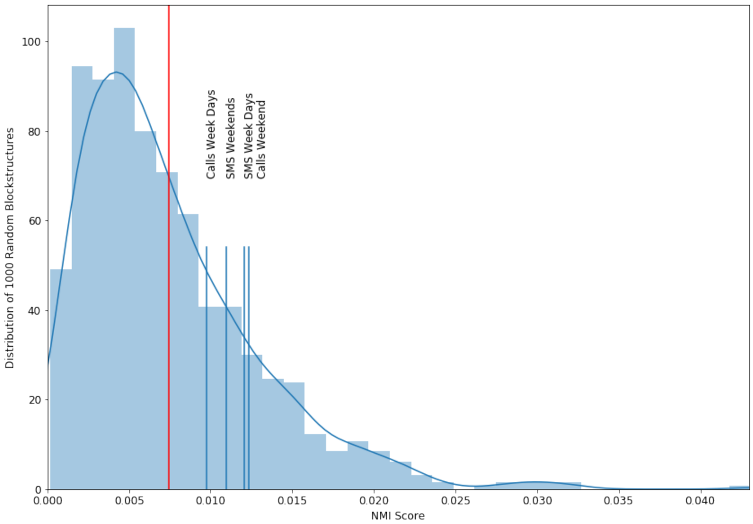


Figure 8.20: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey pay scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

Arranged Marriage

The NMI scores between the arranged marriages ground truth target and the data trials are as follows: Trial One: 0.0340, Trial Two: 0.0356, Trial Three: 0.0307, Trial Four: 0.0334. Figure 8.21 shows these trial results are well above the random average 0.006 into the tail of the random distribution. Again this is a useful target feature to be showing improvements from random results because arranged marriages often correlate with issues of gender inequality. These are also sensitive topics to discuss in traditional surveys. In these results the SMS trials are outperforming the call trials and the weekend trials are outperforming week days.

Forced Labour

The NMI scores between the forced labour ground truth target and the data trials are as follows: Trial One: 0.0258, Trial Two: 0.0276, Trial Three: 0.0194, Trial Four: 0.0256. Figure 8.22 shows these trial results are above the random

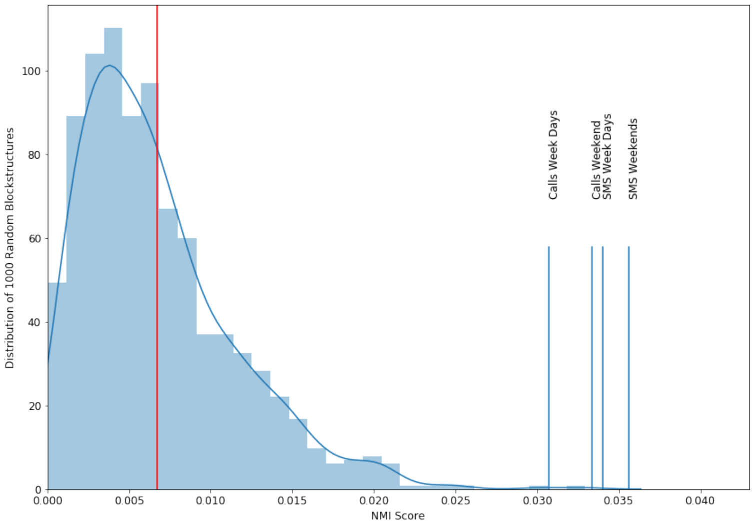


Figure 8.21: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey arranged marriage scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

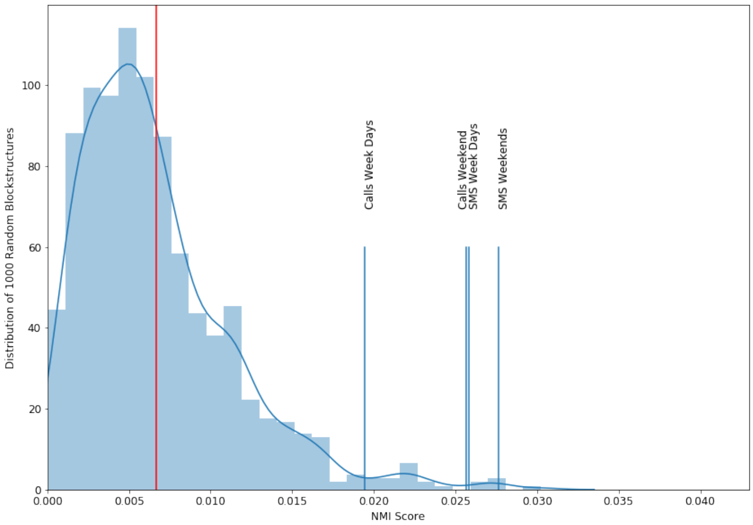


Figure 8.22: The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey forced labour scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

average 0.006. As mentioned in the forced labour proxy results, forced labour is extremely challenging to collect data on so any new indications are useful.

8.4 Discussion

Overall the application to novel data streams of the stochastic block model have produced block structures that meaningfully reflect poverty and related issues when compared with random block structures. The most successful target features were grid poverty, night safety, arranged marriage and forced labour. Interestingly these four targets are all features which are sensitive and difficult to detected with traditional surveys, so identifying these areas could help governments and NGO's allocate limited resources to make positive changes.

Of the four trial options neither CDR type (Call vs SMS), or time (Week vs Weekend) notably outperformed all others across the board. Different targets had different patterns, which makes sense as different target features are likely to match up to different behaviours. For example in the grid results, both SMS trials outperformed the call trials and both the week day trials outperformed the weekend trials. Alternatively, for night safety, both call trials outperformed the SMS trials. For the arranged marriage trials the both SMS trials again out performed the call trials, but this time the weekend trials out performed the week day trials.

The highest improvement compared with random block structures was achieved by the grid survey results. This is a positive aspect of the results as the grid survey is the most reliable ground truth collected in this work and has a high potential for impact. Being able to detect more deprived areas can

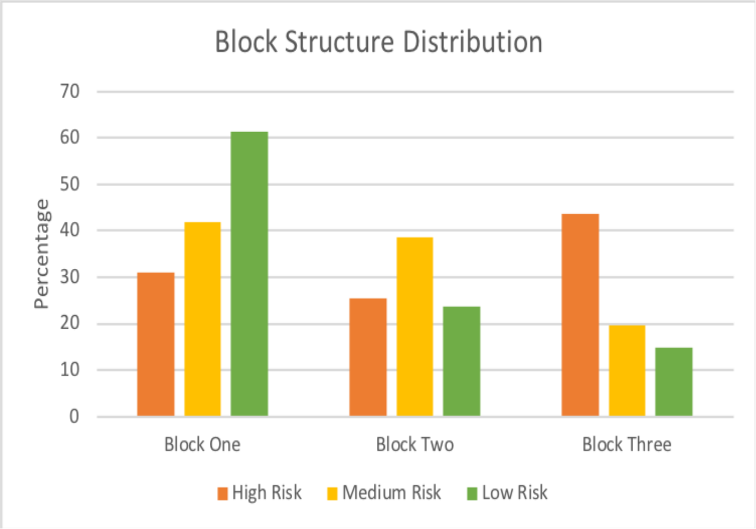


Figure 8.23: The percentage of each level of risk distributed throughout the blocks produced by the stochastic block model for the grid survey.

better allocate limited resources for future surveillance.

Despite the success of these results almost all making improvements compared with the random block structures, the underlying goal of this work is to detect the most vulnerable communities, and it is more important to detect all of the highest risk areas than it is to distinguish between medium and low risk. Take for example the grid structure results, despite the promising improvement compared with random blocks, when looking at the distribution of risk through the blocks the results are less encouraging. Figure 8.23 shows the percentage of each level of risk distributed throughout the blocks. Although block three contains over 40% of the high risk subwards, if I was to allocate resources to only this subward I would still be missing over 50% of the highest risk areas.

The results produced in this chapter are a good initial start, however on further inspection it is clear the block structures produced here would not be adequate enough to advice action in the real world. Surveys are expensive and time consuming, so finding efficient ways to highlight 'at risk' areas to be surveyed is an important goal.

One thing potentially holding these results back is that they currently do not incorporate the time at which interactions are taking place. The next chapter investigates the hypothesis that a dynamic approach to the block model will produce stronger results.

8.5 Summary

In this chapter a novel application of the stochastic block model has been performed in relation to the detection of poverty and related characteristics in the developing world. The block model has been applied to 4 separate networks of Call Record Data; SMS records over the weekends; SMS records over the week days; Call Records over the weekends; and Call records over the week days. Communities structures have been recovered from these networks by inferring the block structure using a Metropolis-Hastings optimization of the stochastic block model. These recovered communities have then been compared with the ground truth collected in the grid and street survey. The stochastic block model has shown promising improvement compared with random block assignments across many of the target features. Overall the most successful target features were grid poverty, night safety, arranged marriage and forced labour.

Chapter 9

Detecting Vulnerable Communities via a Dynamic Stochastic Block Model.

Chapter8 has shown the stochastic block model to provide promising improvement compared with random block assignments in recovering target ground truths. This chapter moves from a static graph to a dynamic one, in order to potentially improve the results by incorporating the time at which interactions are taking place. In this chapter not only are the number of interactions between nodes important but also when the interactions occur over time. The word dynamic in relation to stochastic block models tends to mean one of two things. 1) How do nodes change between blocks through time. 2) How can edges which change through time be used to infer a block structure which remains constant. More about these dynamic approaches can be found in chapter 2.2. This chapter is focused on inferring constant block structures from networks with edges that change over time.

The nature of the CDR data allows behaviour to be modelled over time, the hypothesis is that this inclusion of dynamic events will help improve the detection of vulnerable communities. This hypothesis is based on the idea that the time of day has different effect on different peoples behaviour. Aggregated histograms (shown in the Figures 9.1 and 9.2) show the average number of CDR calls and SMS going out of subwards with difference affluence (the affluence is considered known in these figures through grid survey results described in chapter 4) at different times of day and days of the week. These graphs illustrates different behaviour patterns. Overall the wealthy areas appear to make less outward CDR interactions, this could be because they have access to smart phones and data. On all days between around 19:30 and 23:00 less affluent subwards have a spike in activity while in more affluent areas activity reduces. This could potentially be because the working day is more varied in less affluent areas were informal jobs occur, for example tuktuk drivers, street sellers and manual labourers could be working and thus using their phones later in the day particularly when the heat is reduced, whereas more affluent areas are more likely to have a 9-5 working schedule. Regardless of these theories what is clear is that the time of connection activities varies in subwards of different affluence.

The experimental set up used in chapter 8 is replicated in this chapter. The only addition is that the inputs matrices are matrices of vectors such that $A_{i,j}$ is the vector of edges from node i to node j. $A_{i,j}(t)$ is the number of edges from node i to node j at time t, the time is split into half hour slots. The dynamic block model in this chapter is aiming to infer block structures from the dynamic networks though optimization. These recovered block structures are then compared to the ground truth targets described in chapter 8. This chapter focused specifically on the four ground truth targets most successfully recovered via the static block model in the previous chapter: street survey night safety, arranged marriages, forced labour and grid survey poverty. This chapter now aims to determine if a dynamic adaption to the block model can improve the NMI scores of the recovered block structures from the results presented in chapter 8 and increase the percentage of at risk subwards contained in individual blocks.

The dynamic stochastic block model used in this chapter was introduced by (Corneli et al., 2016), the dynamic expansion is based on Non-Homogeneous Poisson processes. It is assumes that edges between nodes are counted by a Non-Homogeneous Poisson process and the intensity function of this process depends only on the block membership of the nodes. Chapter 2.2 provides details on the history of this model and alternative models. The Non-Homogeneous Poisson process method is adapted in the chapter to include degree correction.

9.1 Model Specifics

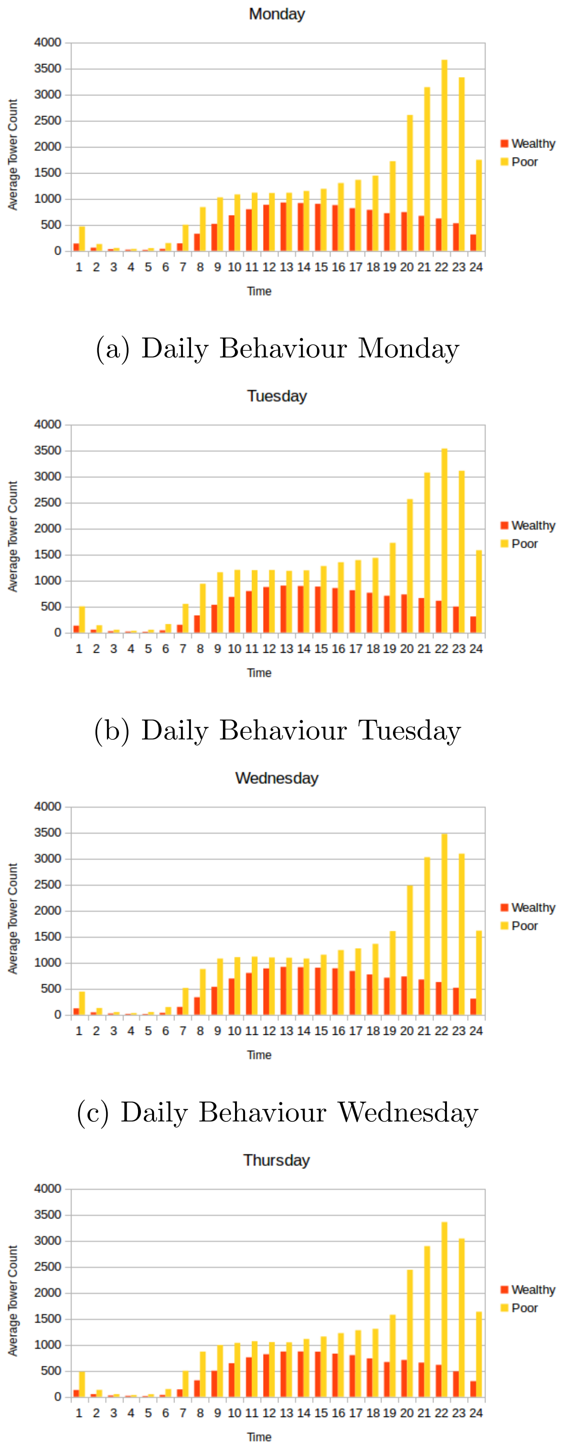
Notation

In addition to the notation provided in chapter 8 here is a list of notion used in this chapter:

A = Adjacency Matrix (size n x n), which is a matrix of vectors such that $A_{i,j}(t)$ is the counting process that gives the number of edges from node i to node j.

λ = Matrix of intensity functions between blocks (size k x k), such that $\lambda_{u,v}$ is the intensity function of the counting process from block u to block v.

w = Time interval such that $w = t_w - t_{w-1}$



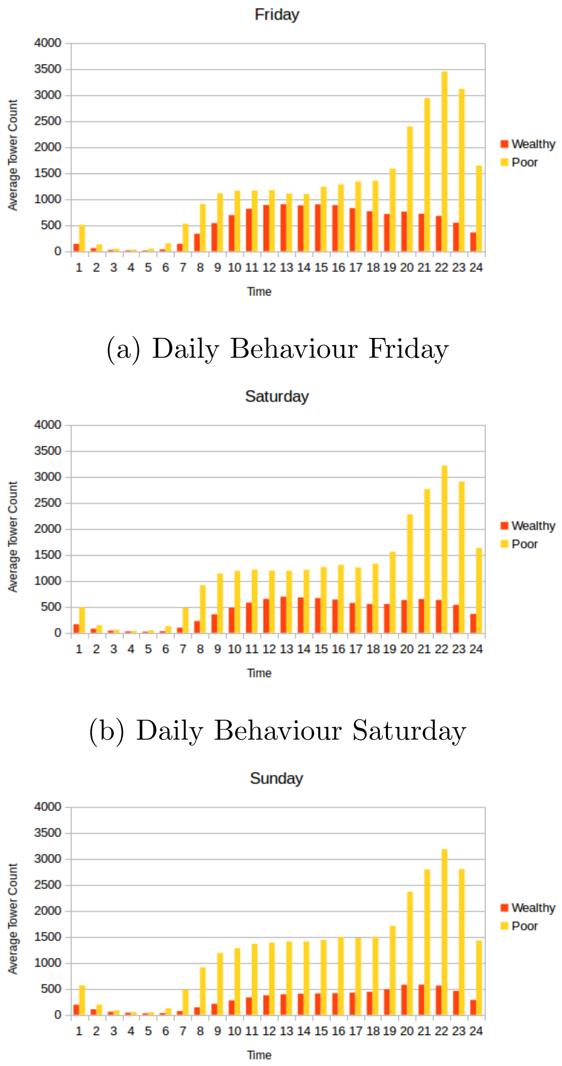
(a) Daily Behaviour Monday

(b) Daily Behaviour Tuesday

(c) Daily Behaviour Wednesday

(d) Daily Behaviour Thursday

Figure 9.1: Average number of CDR calls and SMS going out of subwards with difference affluence levels at different times of day and days of the week (Monday - Thursday).



(a) Daily Behaviour Friday

(b) Daily Behaviour Saturday

(c) Daily Behaviour Sunday

Figure 9.2: Average number of CDR calls and SMS going out of subwards with difference affluence levels at different times of day and days of the week (Friday - Sunday).

$X_{i,j,w}$ = The number of edges from node i to node j in time interval w, such that $X_{i,j,w} = A_{i,j}(w) - A_{i,j}(w-1)$.

G = The graph of edges and nodes over the time [0,T].

$n_{u,v,w}$ = The total number of edges between block u and block v at time interval w.

n_u = The number of nodes in block u, this is fixed and does not change with time.

$\theta^{-}_{w,i}$ = Degree correction parameter at time interval w for the inward degree of node i.

$\theta^{+}_{w,i}$ = Degree correction parameter at time interval w for the onward degree of node i

$d^{+}_{i,w}$ = The outward degree of node i during time interval w.

$d^{-}_{i,w}$ = The inward degree of node i during time interval w.

$D^{+}_{g_i,w} = \sum_{i \in g_i} d^{+}_{i,w}$ which is the sum of all the outward degrees in block g_i during interval w.

$D^{+}_{g_i,w} = \sum_{i \in g_i} d^{-}_{i,w}$ which is the sum of all the outward degrees in block g_i during interval w.

Non-Homogeneous Poisson Processes

The adjacency matrix previously made up of the number of edges between nodes is now replaced by a matrix of vectors, describing the number of connections between edges over time. To view the Non-Homogeneous Poisson Process, first consider one vector within the adjacency matrix, showing the counting process that gives the number of edges from node i to node j.

Assuming this process is a Non-Homogeneous Poisson counting process with intensity function $\lambda_{i,j}$, that a single edge can occur instantaneously in the time interval $[0, T]$, and $A_{i,j}(0) = 0$. Then $A_{i,j}(b) - A_{i,j}(a)$ is the number of edges from node i to node j between time a and time b, given $0 \leq a < b \leq T$. Given all these assumptions the probability of the number of edges between times a and b from node i to node j is as follows (Corneli et al., 2016):

$$P(A_{i,j}(b) - A_{i,j}(a) \mid \lambda_{i,j}) = \frac{(\int_a^b \lambda_{i,j}(t)dt)^{A_{i,j}(b)-A_{i,j}(a)}}{(A_{i,j}(b) - A_{i,j}(a))!} \exp - \int_a^b \lambda_{i,j}(t)dt \quad (9.1)$$

Directed Stochastic Block Model

From chapter 8 recall that the conditional graph probability of the directed stochastic block model is as follows:

$$P(G \mid \omega, g) = \prod_{i,j} \frac{(\omega_{g_i,g_j})^{(A_{i,j})}}{A_{i,j}!} e^{-\omega_{g_i,g_j}} \quad (9.2)$$

The probability of there being an edge between two nodes depends on the blocks they are in.

Dynamic Directed Stochastic Block Model

Combining the Non-Homogeneous Poisson Processes and the Directed Stochastic Block Model, the main assumption of the dynamic model is that the intensity function $\lambda_{i,j}(t)$ depends only on the blocks of each of those nodes, g_i and g_j. Given g, each counting process $A_{i,j}(t)$ is assumed independent. If node i and node c are both in block u and node j and node d

are both in block v, it follows that $\lambda_{i,j} = \lambda_{c,d}$. Therefore the intensity functions can be referred to in relation to their respective blocks (for example λ_{g_i,g_j}). Combining the assumptions in the stochastic block model and the Non-Homogeneous Poisson Processes for $0 \leq a < b \leq T$ the probability of all edges between nodes in the time interval (a,b) is as follows (Note this is as seen in (Corneli et al., 2016), but edited to account for directed edges):

$$P(A(b) - A(a) \mid \lambda, g) = \prod_{i,j} \frac{(\int_a^b \lambda_{g_i,g_j}(t)dt)^{(A_{i,j}(b)-A_{i,j}(a))}}{(A_{i,j}(b) - A_{i,j}(a))!} e^{-\int_a^b \lambda_{g_i,g_j}(t)dt} \tag{9.3}$$

Given the probability of all the edges in individual time intervals, the next aim is to define a probability of the entire graph in the total period of time [0,T]. For this allow the total time interval [0,T] to be split up into W individual intervals. Giving $0 = t_0 \leq t_1... \leq tw - 1 \leq t_W = T$ and the intervals $1, 2, ...W - 1, W$, such that $w = t_w - t_{w-1}$. Given this, there is the notation $X_{i,j,w}$ for the number of edges from node i to node j in time interval w, such that $X_{i,j,w} = A_{i,j}(w) - A_{i,j}(w-1)$. There is also a random vector $X_{i,j}$, such that $X_{i,j} = X_{i,j,1},...,X_{i,j,W}$. (It is worth noting that $A_{i,j}$ in chapter 8 is the same as $A_{i,j}(T)$ in this chapter which is $\sum_w X_{i,j,w}$ in this chapter). Given that increments of a Poisson process are independent and using equation (9.1) the joint probability of $X_{i,j}$ can be written as follows:

$$P(X_{i,j} \mid \lambda_{i,j}) = \prod_{w=1}^{W} \frac{(\int_w \lambda_{i,j}(t)dt)^{X_{i,j,w}}}{X_{i,j,w}!} e^{-\int_w \lambda_{i,j}(t)dt} \tag{9.4}$$

Where $\int_w \lambda_{i,j}(t)dt = \int_{t_{w-1}}^{t_w} \lambda_{i,j}(t)dt$ for all intervals w. Next allowing the block model assumption that the intensity functions between nodes depends only on the blocks of those nodes as before. This assumptions allows equation (9.4) to be written as:

$$P(X_{i,j} \mid \lambda_{g_i,g_j}) = \prod_{w=1}^{W} \frac{(\int_w \lambda_{g_i,g_j}(t)dt)^{(X_{i,j,w})}}{(X_{i,j,w}!)} e^{-\int_w \lambda_{g_i,g_j}(t)dt} \tag{9.5}$$

We now alter the notation for simplicity such that $\Lambda^w_{g_i,g_j} = \int_w \lambda_{g_i,g_j}(t)dt$. Given this there is $\Lambda_{g_i,g_j} = (\Lambda^1_{g_i,g_j}, \Lambda^2_{g_i,g_j}.....\Lambda^w_{g_i,g_j})^T$ and Λ is a k x k matrix of each vector Λ_{g_i,g_j}. Then writing equation 9.5 over all graph edges is as follows:

$$\begin{aligned} P(G \mid \Lambda, g) &= \prod_{i,j} P(X_{i,j} \mid \lambda_{g_i,g_j}, g_i, g_j) \\ &= \prod_{i,j} \prod_{w=1}^{W} \frac{(\int_w \lambda_{g_i,g_j}(t)dt)^{(X_{i,j,w})}}{(X_{i,j,w}!)} e^{-(\int_w \lambda_{g_i,g_j}(t)dt)} \\ &= \prod_{i,j} \prod_{w=1}^{W} \frac{(\Lambda^w_{g_i,g_j})^{(X_{i,j,w})}}{(X_{i,j,w})!} e^{-(\Lambda^w_{g_i,g_j})} \\ &= \prod_{i,j} \prod_{w=1}^{W} \frac{1}{(X_{i,j,w})!} \prod_{u,v} \prod_{w=1}^{W} (\Lambda^w_{u,v})^{n_{u,v,w}} e^{-n_u n_v (\Lambda^w_{u,v})} \end{aligned} \tag{9.6}$$

Where $n_{u,v,w}$ is the total number of edges between block u and block v in time interval w and n_u is the number of nodes in block u. This is the probability which needs to be maximized over the different grid structures in order to infer an optimal community structure. However, before going into the details of this optimization it is important to account for the varying degrees of nodes in the network over time, this is done by adapting the dynamic model to account for degree correction applying methods introduced by (Karrer and Newman, 2011b). Figure 9.3 shows an example using trial one (this is a network of SMS connections over weekdays between subwards as described in chapter 8) of how varied the node degrees can be over individual time slots. This figure is describing features directly from the data providing evidence for the choice to model with degree correction. This is particularly apparent during the

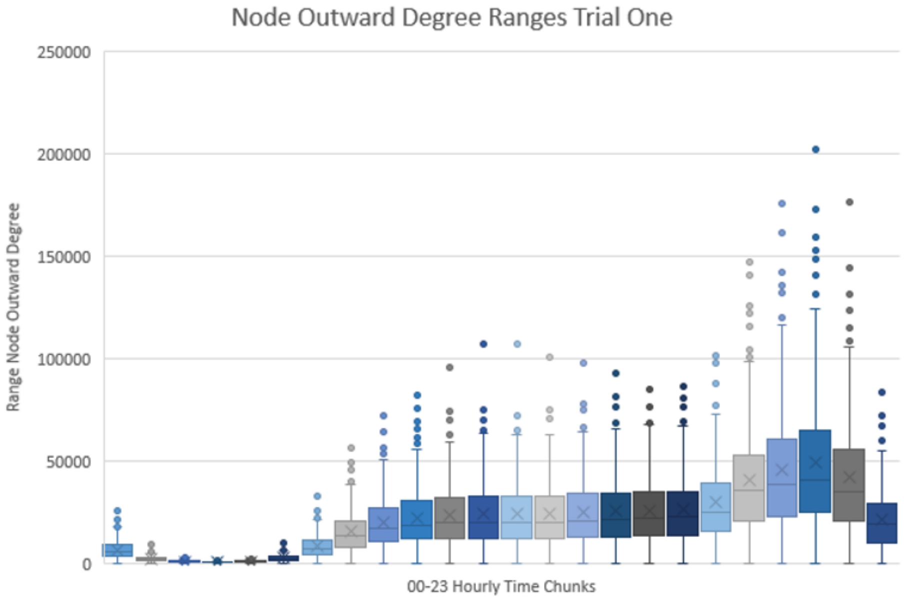


Figure 9.3: Outward node degrees over time.

day between the hours of 6am and midnight. There is less degree variation overnight, this could be because most people sleep at night and therefore are less active on their phones.

Dynamic Degree Corrected Directed Stochastic Block Model

In the degree corrected block model (Karrer and Newman, 2011b) new parameters are introduced to control the expected degrees of each node. Here the following parameters are introduced: $\theta^{+} = (\theta^{+}_{1}, ...\theta^{+}_{w-1}, \theta^{+}_{w})$ and $\theta^{-} = (\theta^{-}_{1}, ...\theta^{-}_{w-1}, \theta^{-}_{w})$ which are vectors of length w, which is the number of time intervals within the network. This idea of degree correction was detailed in chapter 8. Each element θ^{+}_{w} is a vector of parameter length n such that $\theta^{+}_{w} = (\theta^{+}_{w,i}, ..., \theta^{+}_{w,n-1}, \theta^{+}_{w,n})$ to account for the outward degrees of nodes over time. Likewise $\theta^{-}_{w} = (\theta^{-}_{w,i}, ..., \theta^{-}_{w,n-1}, \theta^{-}_{w,n})$ to account for the inward degrees of nodes over time. That is $\theta^{-}_{w,i}$ is the parameter at time w to account for the inward degree of node i and $\theta^{+}_{w,i}$ is the parameter at time w to account for the outward degree of node i. It can be assumed by definition that $\sum_{i \in g_i} \theta^+_{w,i} = \sum_{i \in g_i} \theta^-_{w,i} = 1$ for each value of w. The rate function of between nodes depends on the blocks they are in and the degree correction parameters. Again the number of blocks K is assumed fixed and known. Given all this information the conditional probability can be written as follows:

$$P(G \mid \Lambda, g, \theta^+, \theta^-) = \prod_{i,j} \prod_{w=1}^{W} \frac{(\theta^+_{w,i} \theta^-_{w,j} \Lambda^w_{g_i,g_j})^{(X_{i,j,w})}}{(X_{i,j,w})!} e^{-(\theta^+_{w,i} \theta^-_{w,j} \Lambda^w_{g_i,g_j})} \tag{9.7}$$

Given the degree correction parameters are introduced as a multiplicative constant which is absorbed into Λ and thus $\sum_{i \in g_i} \theta^+_{w,i} = \sum_{i \in g_i} \theta^-_{w,i} = 1$ for all groups u equation 9.7 can now be simplified to give:

$$\begin{gathered} P(G \mid \Lambda, g, \theta^+, \theta^-) = \\ \prod_{i,j} \prod_{w=1}^{W} \frac{1}{(X_{i,j,w})!} \prod_{i} \prod_{w=1}^{W} (\theta^+_{i,w})^{d^+_{i,w}} (\theta^-_{j,w})^{d^-_{i,w}} \prod_{u,v} \prod_{w=1}^{W} (\Lambda^w_{u,v})^{n_{u,v,w}} e^{-(\Lambda^w_{u,v})} \end{gathered} \tag{9.8}$$

Where $d^+_{i,w}$ and $d^-_{i,w}$are the outward and inward degrees of node i at time w and $n_{u,v,w}$ again is the number of edges from block u to block v in time interval w. Making the assumptions that rate functions are consistent, $\Lambda^w_{u,v}$ can be estimated by maximizing the log likelihood.

$$\begin{gathered} \ln(P(G \mid \Lambda, g, \theta^+, \theta^-)) = \\ \sum_{i} \sum_{w} d^+_{i,w} \ln \theta^+_{i,w} + \sum_{i} \sum_{w} d^-_{i,w} \ln \theta^-_{i,w} + \sum_{u,v} \sum_{w} n_{u,v,w} \ln \Lambda^w_{u,v} - \Lambda^w_{u,v} \end{gathered} \tag{9.9}$$

Constants are ignored in equation (9.9) because they will not change dependent on the parameters $\omega, g, \theta^+, \theta^-$ so will not affect the maximum likelihood. Given the log likelihood equation (9.9) the following maximum likelihood values of the parameters $\Lambda, \theta^+, \theta^-$ can be derived.

$$\begin{aligned}
\widehat{\Lambda^w_{u,v}} &= n_{u,v,w} \\
\widehat{\theta^+_{i,w}} &= \frac{d^+_{i,w}}{D^+_{g_i,w}} \\
\widehat{\theta^-_{i,w}} &= \frac{d^-_{i,w}}{D^-_{g_i,w}}
\end{aligned} \tag{9.10}$$

We now substitute these estimators into the log likelihood equation (9.9) to give the following:

$$\ln(P(G \mid \Lambda, g, \theta^+, \theta^-)) =$$
$$\sum_i \sum_w d^+_{i,w} \ln \frac{d^+_{i,w}}{D^+_{g_i,w}} + \sum_i \sum_w d^-_{i,w} \ln \frac{d^-_{i,w}}{D^-_{g_i,w}} + \sum_{u,v} \sum_w n_{u,v,w} \ln n_{u,v,w} - n_{u,v,w} \tag{9.11}$$

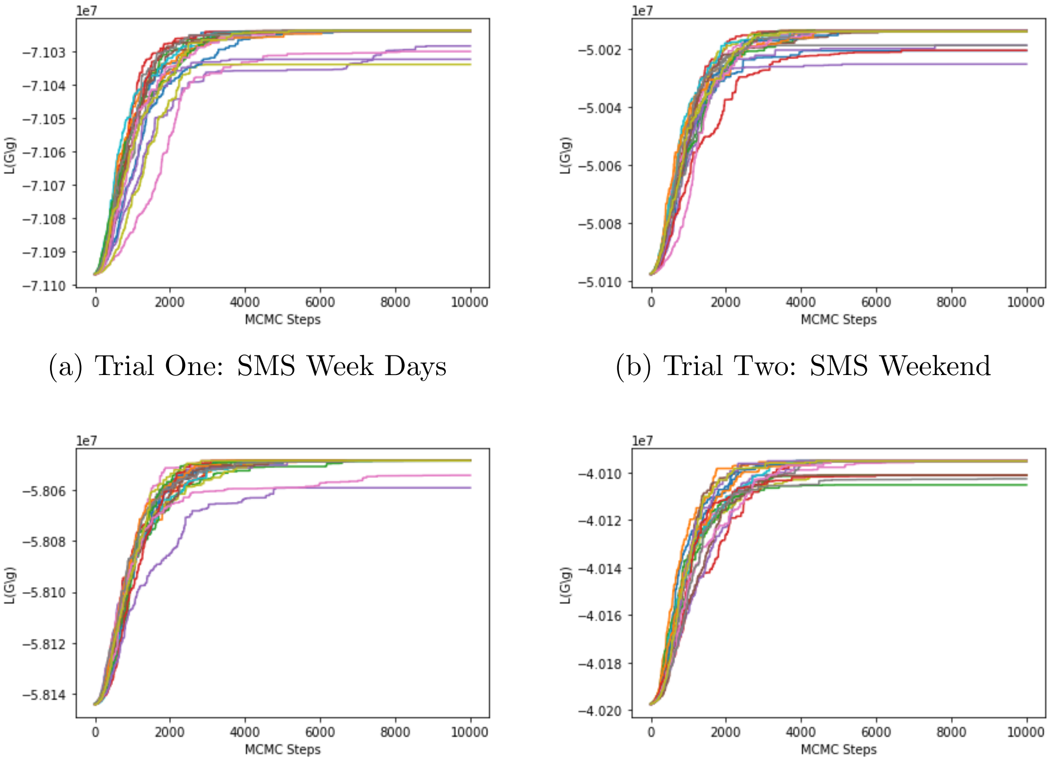
We then rearrange the first two terms as follows:

$$
\begin{aligned}
&\sum_i \sum_w d^+_{i,w} \ln \frac{d^+_{i,w}}{D^+_{g_i,w}} + \sum_i \sum_w d^-_{i,w} \ln \frac{d^-_{i,w}}{D^-_{g_i,w}} + \\
&= \sum_i \sum_w d^+_{i,w} \ln d^+_{i,w} - \sum_i \sum_w d^+_{i,w} \ln D^+_{g_i,w} + \\
&\sum_i \sum_w d^-_{i,w} \ln d^-_{i,w} - \sum_i \sum_w d^-_{i,w} \ln D^-_{g_i,w} \\
&= \sum_i \sum_w d^+_{i,w} \ln d^+_{i,w} + \sum_i \sum_w d^-_{i,w} \ln d^-_{i,w} - \\
&(\sum_u \sum_w D^+_{u,w} \ln D^+_{u,w} + \sum_u \sum_w D^-_{u,w} \ln D^-_{u,w}) \\
&= \sum_i \sum_w d^+_{i,w} \ln d^+_{i,w} + \sum_i \sum_w d^-_{i,w} \ln d^-_{i,w} - \\
&(\sum_u \sum_w (\sum_v n_{u,v,w} \ln D^+_{v,w} + (\sum_u n_{u,v,w} \ln D^-_{u,w}) \\
&= \sum_i \sum_w d^+_{i,w} \ln d^+_{i,w} + \sum_i \sum_w d^-_{i,w} \ln d^-_{i,w} - \sum_{u,v} \sum_w n_{u,v,w} \ln D^+_{u,w} D^-_{v,w}
\end{aligned}
\tag{9.12}
$$

Substituting this back into the equation (9.11) and ignoring the constant terms gives:

$$
L(G \mid g) = \sum_{u,v} \sum_w n_{u,v,w} \ln \frac{n_{u,v,w}}{D^+_{u,w} D^-_{v,w}} \tag{9.13}
$$

This is the objective function which needs to be maximized over g in order to infer the best possible community structure of the dynamic network. Given that $D^-_{v,w}$, $D^+_{u,w}$ and $n_{u,v,w}$ can all be derived from a block structure g and an observed network, in order to recover this block structure the Metropolis-Hastings optimization method described in chapter 8 is used with the objective function equation (9.11). All results in this chapter have been implemented using python code written from scratch during this PhD.



(a) Trial One: SMS Week Days

(b) Trial Two: SMS Weekend

(c) Trial Three: Call Week Day

(d) Trial Four: Call Weekend

Figure 9.4: 20 MCMC optimization cases each starting with different random block structures g over the four observe network trials.

9.2 Applying Dynamic Block Model to Novel Data Streams Results

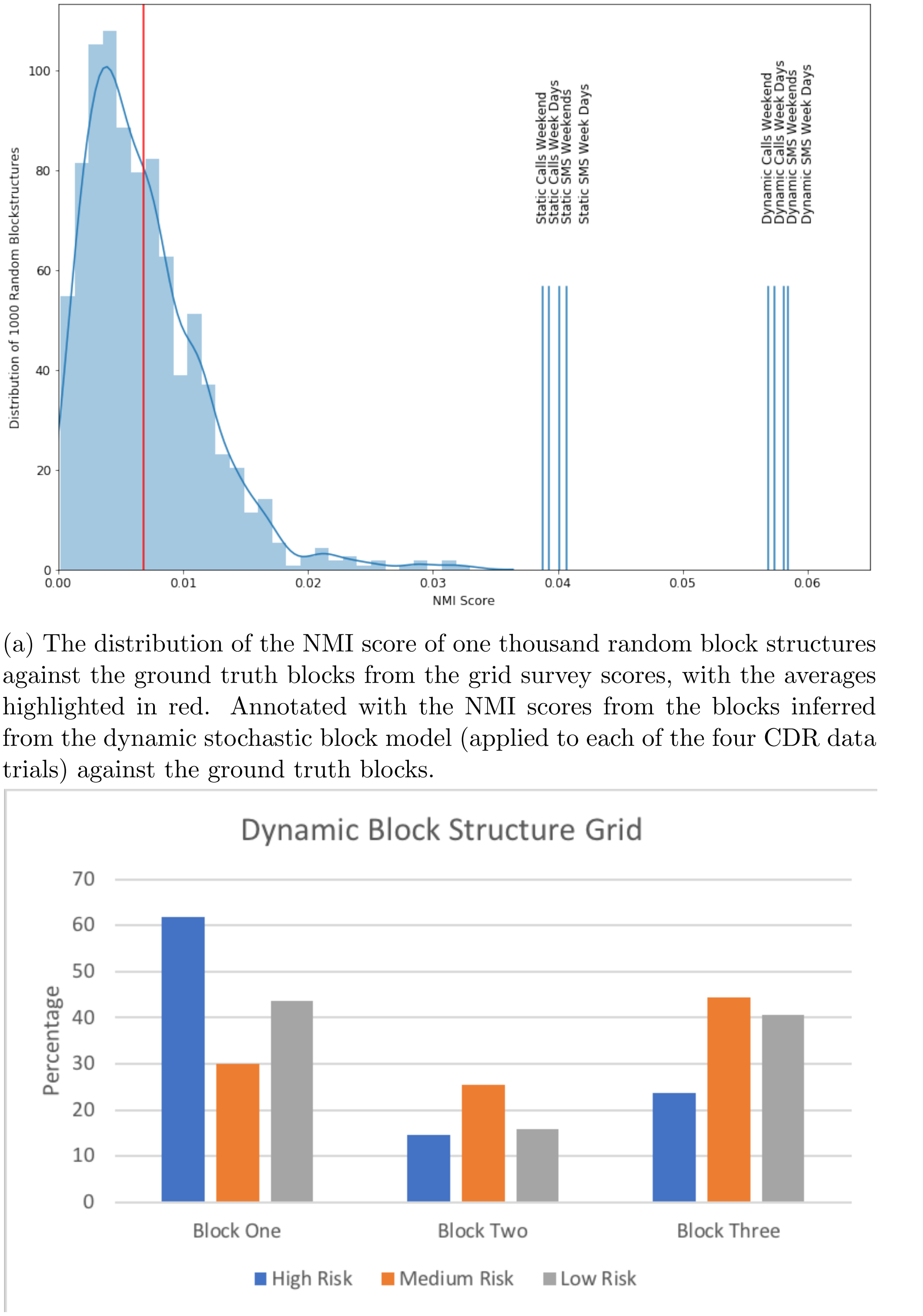
The stochastic block model has been applied to the four observed network trials of CDR describe in chapter 8. In this chapter $X_{i,j,w}$ is the number of CDR connections (related to each trial) between subwards. The time intervals w are half hour intervals from 12am to 12pm. Figure 9.4 shows 20 cases of the MCMC optimization algorithm each starting with different random block structure g. Trial One, with an input of week day sms connections between subwards finished with 15 runs on a maximum finishing point and the other 5 at a local optima. Trial Two, with an input of weekend SMS connections finished with 14 runs on a maximum finishing point and the other 6 at a local optima. Trial three, with an input of week day call connections finished with 17 runs on a maximum finishing point and the other 3 at a local optima. Trial four, with an input of weekend call connections finished with 15 runs on a maximum finishing point and the other 5 at a local optima.

Following is a comparison of results for the dynamic version of the block model versus the static version in terms of (i) NMI score, and (ii) percentage of high risk subwards assigned to the same block.

Grid

The NMI scores between the grid survey ground target and the data trials are as follows: Trial One: 0.0584, Trial Two: 0.0581, Trial Three: 0.0573, Trial Four: 0.0569. Not only is this well above the random average but this is also notably above the static scores (Trial One: 0.0407, Trial Two: 0.0401, Trial Three: 0.0393, Trial Four: 0.0388). These results can be seen in Figure 9.5a. The SMS and week day trials have again produced the best results. In this case it is clear that the dynamic adaption of the block model has improved the results. Given the grid results were used to create the histograms (Figures 9.1a-9.2c) which motivated this dynamic expansion, it makes sense that accounting for interaction times has improved the results.

The most successful NMI input for this target was dynamic week day calls, Figure 9.5b shows the distribution of risk levels within blocks for this trial. The high risk subwards are less split between the blocks in the results produced by the dynamic model. In the static model, the block with the highest percentage of high risk subwards contained only 40% of the total high risk subwards. Figure 9.5b shows that this has increased by over 20%. One of the blocks



(a) The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the grid survey scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the dynamic stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

(b) The percentage of each level of risk distributed throughout the blocks produced by the dynamic stochastic block model for the grid survey.

Figure 9.5: Dynamic degree corrected stochastic block model results for grid survey affluence

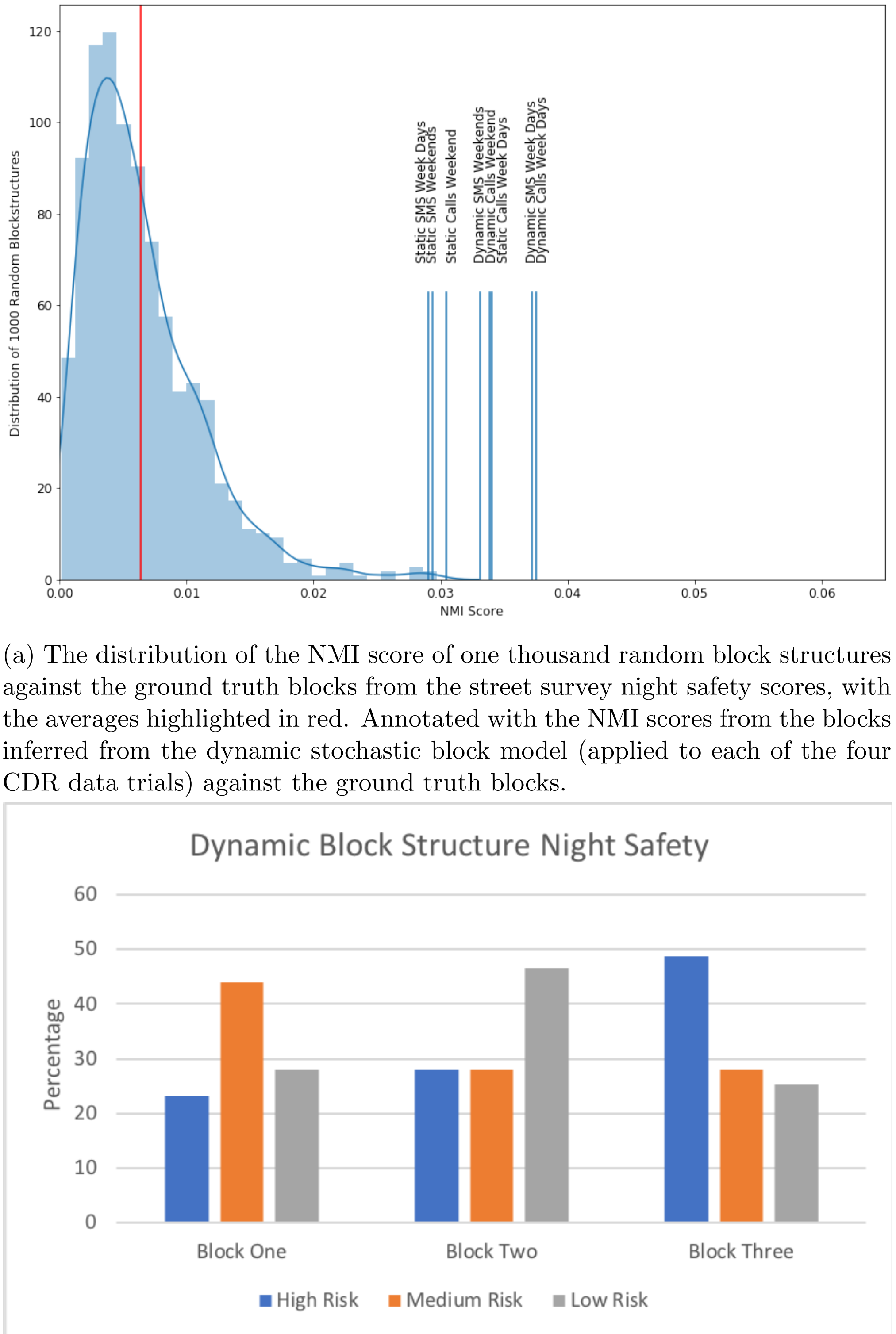
now contains over 60% of the total high risk subwards, which is a notable improvement. Being able to allocate 60% of the subwards at highest risk to deprivation could greatly benefit government groups and NGO's trying to detect risk and allocate limited resources. However, it is also worth noting that the block containing 60% of the high risk subwards is a large block, it also contains 30% and over 40% of the medium and low risk subward. As such, in the context of this work, surveying subwards based on this block structure would not greatly reduce the overall number of subwards being investigated.

Night Safety

The NMI scores between the night safety ground truth target and the data trials are as follows: Trial One: 0.0372, Trial Two: 0.0331, Trial Three: 0.0375, Trial Four: 0.0341. There results can be seen in comparison with the static results (Trial One: 0.02907, Trial Two: 0.0294, Trial Three: 0.0338, Trial Four: 0.0304) in Figure 9.6a. Both week day trials (trial one and three), have reached the highest improvement from the random average. The weekend trials have improved from the static weekend trials, however not to the degree that they are above the static week day trials. Overall the dynamic model extension has improved the results, but not as notably as for the grid survey target.

The most successful NMI input for this target was again dynamic week day calls, Figure 9.6b shows the distribution of risk levels within blocks for this trial. Block three contains the largest percentage of high risk subwards at just under 50%. Although the dynamic model has improved the results slightly, over 50% of the high risks subwards are split between the other two blocks and would therefore be missed if block three was used as the 'high risk block'.

Arranged Marriage



(a) The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey night safety scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the dynamic stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

(b) The percentage of each level of risk distributed throughout the blocks produced by the dynamic stochastic block model for the street survey night safety

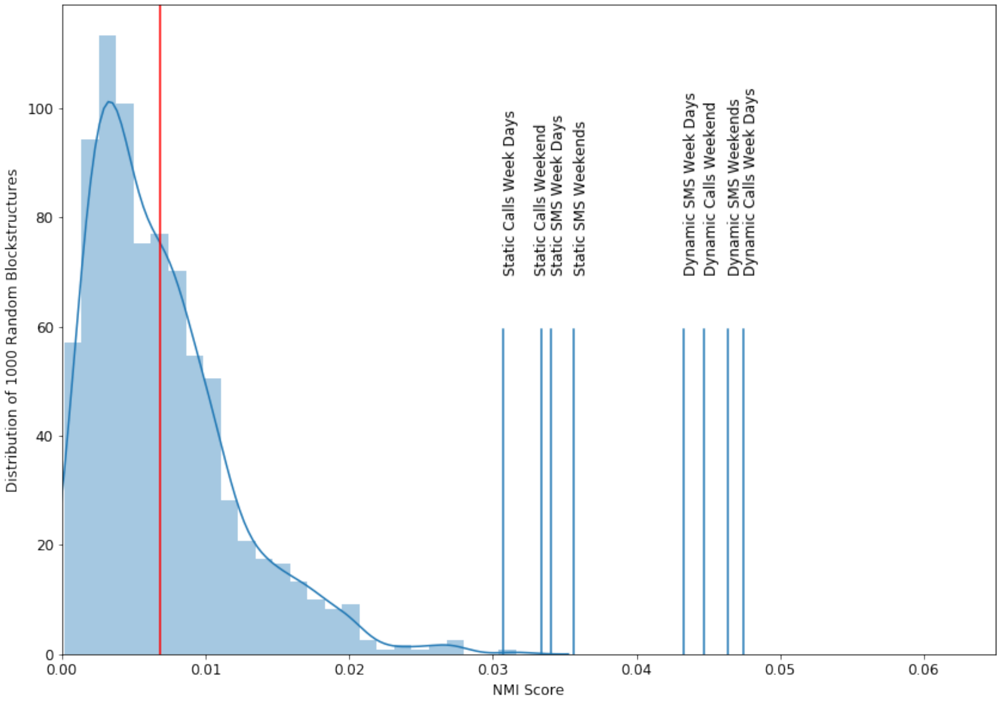
Figure 9.6: Dynamic degree corrected stochastic block model results for street survey night safety

The NMI scores between the arranged marriages ground truth target and the data trials are as follows: Trial One: 0.0433, Trial Two: 0.0464, Trial Three: 0.0475, Trial Four: 0.0447. Like in the grid results, the dynamic has notably improved the NMI scores of the block structures. The improvements can be seen in Figure 9.7a, although the improvements are slightly less than the grid results, the dynamic expansion has clearly improved all four trials. The most successful trial for this target is the week day call data.

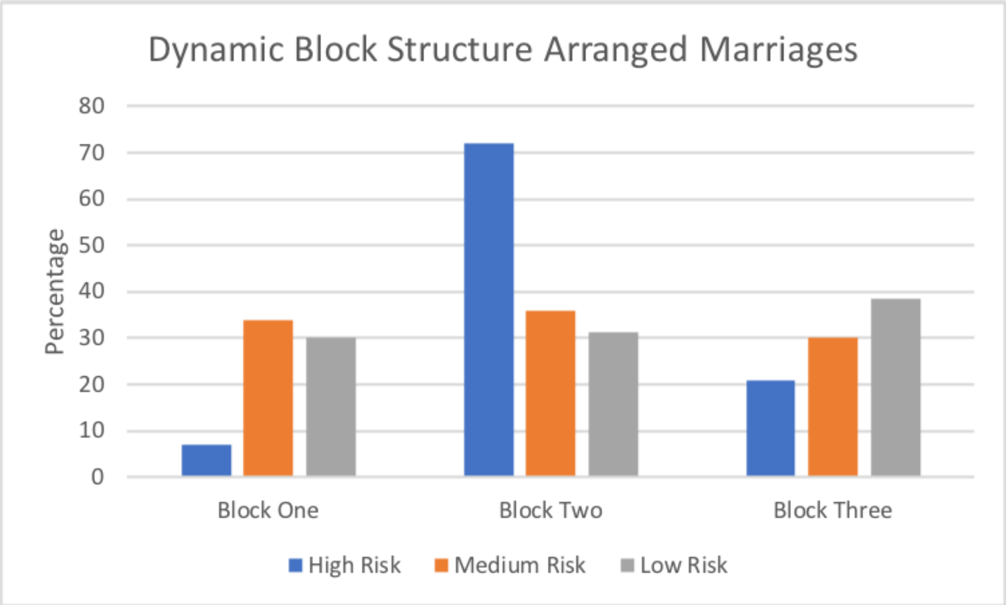
The most successful NMI input was again dynamic week day calls, Figure 9.7b shows the distribution of risk levels within blocks for this trial. Although the NMI scores are higher in the grid results, the arranged marriage target has the overall highest percentage of high risk subward in one block. Block one has less than 10% of the high risk subward, block three has just over 20% and block two has over 70% of the total high risk subwards. Arranged marriages are an extremely sensitive topic and highly linked with issues of gender equality and exploitation. The arranged marriages in this target feature relate to marriages arranged by the family for financial reasons. This is a very challenging topic for data collectors to discuss in surveys because extreme views co-exist in Tanzania on the morality of financial marriages of young girls. Having a reliable indication of where to further investigate these issues could be very impactful for NGOs and government groups.

Forced Labour

The dynamic extension has not improved the NMI scores for the forced labour targets. The results seen in Figure 9.8a are all still above the random average, however of the four trials only one (calls at weekends) improved with the inclusion of dynamic data. Overall the dynamic approach has not improved

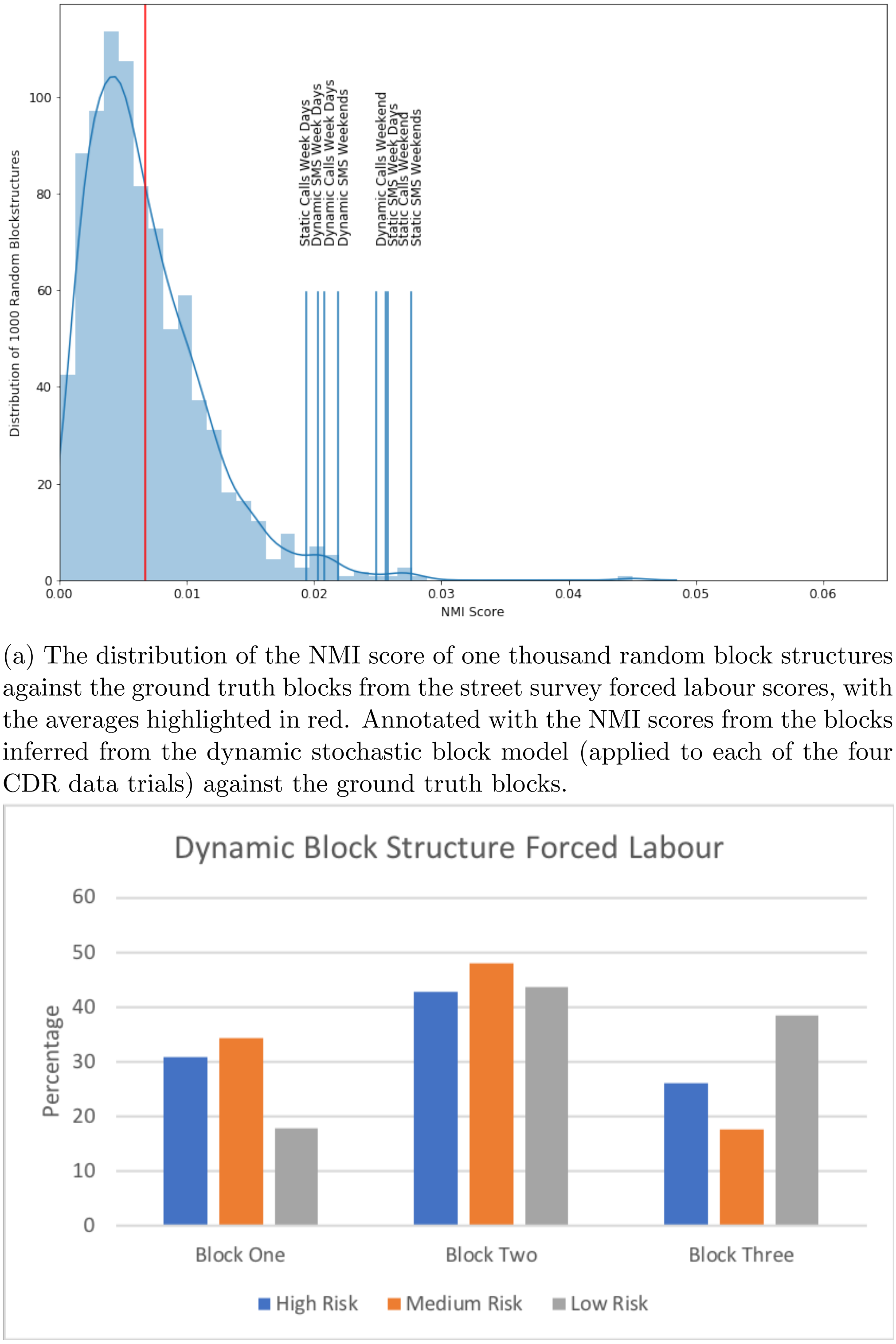


(a) The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey arranged marriage scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the dynamic stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.



(b) The percentage of each level of risk distributed throughout the blocks produced by the dynamic stochastic block model for the street survey arranged marriage

Figure 9.7: Dynamic degree corrected stochastic block model results for street survey arranged marriage



(a) The distribution of the NMI score of one thousand random block structures against the ground truth blocks from the street survey forced labour scores, with the averages highlighted in red. Annotated with the NMI scores from the blocks inferred from the dynamic stochastic block model (applied to each of the four CDR data trials) against the ground truth blocks.

(b) The percentage of each level of risk distributed throughout the blocks produced by the dynamic stochastic block model for the street survey forced labour

Figure 9.8: Dynamic degree corrected stochastic block model results for street survey forced labour

the results for this target features.

The most successful trial for this target is the static sms weekend input. As expected the distribution of risk throughout the blocks is quite spread here. Figure 9.8b shows that the largest percentage of high risk subwards recovered in a block is block two with less than 50% of the total high risk subwards. Forced labour has not been recovered as successfully as the other features.

9.3 Discussion

The goal of this chapter was to determine if the inclusion of dynamic events will help improve the detection of vulnerable communities. The results show that a dynamic approach can improve the detection of vulnerable communities. Between all the results static and dynamic, the highest improvements from the random average NMI scores where achieved by the following inputs and targets: • Dynamic Call SMS Week Day with Grid Targets • Dynamic Call SMS Week Day with arranged marriage targets. The greatest improvements from static results to dynamic results were achieved again by • Dynamic Call SMS Week Day with Grid Targets and • Dynamic Call SMS Week Day with arranged marriage targets. There is a clear indication that the inclusion of dynamic events is beneficial for some target features, this is not the case for all features. Of the remaining eight features investigated in chapter 8, two ('overcrowding' and 'theft and violence') improved with the inclusion of dynamic data, the other six either had reduced or similar NMI scores. Overall the dynamic block model is useful for some features but not others, for targets such as overall deprivation and arranged marriage in particular the dynamic approach has been notably fruitful.

Despite the success of both the static and dynamic block model (which can be seen in Figures 9.5a and 9.7a) given the delicate context of this work the results are not performing to a satisfactory level. Take for example the target arranged marriages. If resources were to be allocated based on these results, 30% of the highest risk subwards would be missed (See Figure 9.7b). There is work to be done to investigate and improve these methods.

One interesting option for future work is to include geographical influence. For example, it makes sense that people are likely to communicate with people local to them. Equally it might be expected that people with a higher level of wealth are more likely to maintain communication with people who are geographically further away from them due to their ability to travel more. To account for this geographic effect future work could include the distance of each interaction in the adjacency matrix counts.

A current limitation of this work is the quantity of CDR data. In this work, a significant proportion of the subwards in Dar es Salaam were removed because the mobile provider did not have adequate coverage. This district of the city was covered by a different governmental unit who contracted an alternative mobile provider at the time the data was shared with my research team. Further to this I recognise that there is a discrepancy in the dates of the datasets used in this work. The survey data has been recently collected, while the mobile data was collected in 2014. Dar es Salaam is a rapidly growing city and areas could have changed in this time period. This work would benefit from up-to-date data with comprehensive coverage. CDR data is not freely available from most providers, accessing new data and up to date CDR data presents a challenge for future research on this topic. Additionally, technology is moving on rapidly and more people are using data-reliant services such as WhatsApp to send messages and make calls. This transition will have an impact on CDR call and SMS behaviours. Future work on this topic should investigate the opportunities to access new and up-to-date sources of interaction networks.

Considering on balance the success and limitations of the dynamic block model, my overall recommendation is that the most beneficial route of future work would be to investigate simpler relationships between CDR entropy and ground truth features. While future work improving the stochastic block model is interesting, a potentially more valuable follow up to this work would be to look at simple cluster sequence mining from event sequence data. There are three main reasons for this. Firstly, the motivation of the dynamic extension was the observable difference in daily CDR events between subwards with high affluence and low deprivation. The patterns found here clearly indicate a link between the times and quantities of CDR events and subward affluence. Secondly, although the stochastic block model results have in many cases successfully improved the results from random clustering, in the context of the real world application I do not feel an improvement from random is enough to support a change of practice. Finally, given the results are not overwhelmingly accurate, a simpler approach would have the benefit of uncomplicated communication. The dynamic block model would be quite complicated to explain to a non-mathematical audience, the overall goal of this work is to provide potentially impactful methods and results. Partners who would be acting on these detections of high risk subwards are more likely to trust a method which can be clearly communicated and understood.

9.4 Summary

Overall, this work has successfully shown that a dynamic approach to the stochastic block model can improve the detection of at risk subwards. This chapter has used the dynamic block model introduced by (Corneli et al., 2016) and the degree corrected block model introduced by (Karrer and Newman, 2011b). The technical contribution of this chapter has been the combination of these two methods for a degree corrected dynamic stochastic block model. The dynamic block model has also been implemented with CDR data and the block structure was inferred through Metropolis-Hastings optimization. The novel application of recovering community health has been assessed, showing particular promise for the targets of arranged marriages and subward deprivation. Various future avenues of research have been discussed, both in terms of the block model and alternative lines of work.

Chapter 10

Conclusions

10.1 Summary

The work of policy makers and NGOs in attaining UN SDGs in developing regions is inhibited by data deficiencies and extensive challenges in the collection of reliable demographic information on community health. In light of these challenges new methods of data collection are required. Comparative judgement methods are a promising solution, and the grid survey introduced in this work leverages local knowledge elicited via comparisons of the affluence of different subwards. Such models can both simplify logistics and circumvent biases inherent to household surveys. The grid survey method presented in this work is both quicker and more affordable than traditional survey methods, and can be conducted at a uniquely fine grained scale. The collaboration with local experts this method introduced also hold potential to increase the reliability of this outputs, leveraging the knowledge of local populations while avoiding response bias.

A primary contribution of this work is made via the knowledge gained from the study of Dar es Salaam, and novel data and insights accrued. During the grid survey workshops, 174 participants made a total of 75,457 comparisons leading to an average of 288 rankings per subward, over the 452 subwards in Dar es Salaam. These results have been validated by regional experts, and reflect high fidelity ground truths with potential future use by researchers and policy makers alike and are already underpinning other work in Tanzania.

In addition to the grid survey, the street survey approach presented in this work has also contributed to the development of novel survey techniques. The street survey made a focus shift away from household survey, eschewing participants personal experiences in preference of participants perceptions of subwards as a whole. This shift has helped to avoid personal bias of participants commonly found in traditional surveys (such as reticence to honestly categorise household income). The literature review carried out in this thesis highlighted the deficiencies in fine grained, detailed data on poverty and social vulnerability in Dar es Salaam, the street survey moves towards filling this gap by introducing data from 3668 surveys, each with over 70 features attached. Given the wide berth of novel survey methods applied in this work, a validation process was also performed via semi-structured interviews aiming to validate and substantiate the methodological outcomes of the data collection processes. Key outcomes of this validation process were the importance of collaboration, adaptability and in-field pilot testing. This validation process and efforts to provide a meaningful evaluation of the data collection methods, hoped to educate future projects aiming to utilize and replicate the street and grid survey techniques.

Collection of data, of course, had focused analytical purpose. An initial analysis of the street survey data suggested the need for examination of systematic inclusion of bias in ground truth data. Fixed effects models indicated that bias does exist. Notably participants living in more affluent areas are more likely to disproportionately downgrade themselves, and while other biases were noted, this must be accounted for in future surveying across developing regions. Further to this, the street survey results have also shown that machine learning is a highly effective method for revealing new covariates to be leveraged as proxies for the detection of forced labour. The key proxies identified were pay, household structure and area of origin alongside a confirmation of theorised covariates such as education and lack of identification documents. Machine learning models were used to predict vulnerability to forced labour with high levels of accuracy, showing significant improvements over theorised covariates alone.

The results from these data collection processes have also been utilized in this work as ground truth targets to illustrate the potential uses of increasingly ubiquitous and rich novel data streams drawn from the private sector (such as Call Data Records) to support fine grained detection of vulnerable communities in developing countries. A novel application of the stochastic block model was performed on observed CDR networks to recover block structures. These recovered block structures have been assessed against the ground truths in order to illustrate the promising combination of CDR data and statistical analysis in the detection of poverty and its related characteristic in the developing world. The affluence levels from the grid survey in particular have shown notable improvements from random block structures. The NMI score between the grid survey ground target and the stochastic block model results was 0.0347 above the random average.

The final section of this thesis made use of the dynamic nature of the CDR data sources, implementing a new degree corrected dynamic stochastic block model and performing a novel application of the methods to recover community health structures. This work demonstrates evidence of the importance of considering dynamic data in the detection of communities vulnerable to poverty. The greatest improvements of NMI scores from the static to dynamic approach was achieved for the targets of overall deprivation and arranged marriage. In the static model the block with the highest percentage of high risk subwards contained only 40% of the total high risk subwards. This was increased by over 20% in the dynamic approach with one of the blocks containing over 60% of the total high risk subwards a notable improvement. Being able to allocate 60% of the subwards at highest risk to deprivation could greatly benefit governments groups and NGOs trying to detect risk and allocate limited resources.

10.2 Reflections

Having more reliable and easily accessible estimates of poverty and related vulnerabilities has a high potential importance for policy makers and NGOs trying to make positive changes to reduce the devastating effects of poverty. The comprehensive knowledge on Dar es Salaam produced in this thesis has moved towards amending the current data gaps, via rigorous fine-grained data collection processes. The results produced in this study have potential to benefit future research by providing real world data which can be utilized to produce meaningful results. The survey techniques themselves can also be seen as a test bed for future work in other developing regions.

In addition to the wide range of potential future research underpinned by the data collection methods and results produced in this work, various avenues of future research have also been introduced via the analysis of the ground truth data. This work provides the grounds for promising new research on feature engineering of transactional data to support the creation of meaningful proxies to support and identify target areas for survey data collection. Future work adapting the stochastic block model has also been highlighted as a potential avenue for new research following this work, in particular the inclusion of geographical influence on network edges. Overall, this thesis indicates that while there is no single substitute for surveys, an efficient grid survey methodology can, however, be used in the future to direct limited resources and future, targeted data collection. This multi-disciplinary bridge of research between analytical work and in-field data collection has proven a fascinating area of investigation with lots of potential research. It is my hope that this crucial and nascent field of work can continue to grow, to help empower both policy makers and NGOs aiming to meet UN SDGs.

Bibliography

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Adamic, L. and Glance, N. (2005). Proceedings of the www-2005 workshop on the weblogging ecosystem.

Aicher, C., Jacobs, A. Z., and Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2006a). Mixed membership stochastic block models for relational data with application to protein-protein interactions. *Proceedings of the international biometrics society annual meeting*, Issue(June):1–34.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2006b). Mixed membership stochastic block models for relational data with application to protein-protein interactions. proceedings of the international biometrics society annual meeting.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed Membership Stochastic Blockmodels. *Jmlr*, 9(2008):1981–2014.

Appleton, S. and Booth, D. (2001). Combining participatory and survey-based approaches to poverty monitoring and analysis. In *Background paper for Uganda workshop*, volume 30.

Baker, J. R., Silove, D., Horswood, D., Al-Shammari, A., Mohsin, M., Rees, S., and Eapen, V. (2021). Psychological distress, resettlement stress, and lower school engagement among arabic-speaking refugee parents in sydney, australia: A cross-sectional cohort study. *PLoS medicine*, 18(7):e1003512.

Bales, K. (2012). *Disposable People: New Slavery in the Global Economy, Updated with a New Preface.* Univ of California Press.

Bales, K. B. (2006). Testing a theory of modern slavery.

Belser, P., De Cock, M., and Mehran, F. (2005). Ilo minimum estimate of forced labour in the world. Technical report, ILO.

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

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Bradshaw, J., Gordon, D., Levitas, R., Middleton, S., Pantazis, C., Payne, S., and Townsend, P. (1998). Perceptions of poverty and social exclusion. *Bristol: Townsend Centre for International Poverty Research.*

Brdar, S., Gavric, K., Culibrk, D., and Crnojevic, V. (2016). Unveiling spatial epidemiology of HIV with mobile phone data. *Sci. Rep.*

Breiger, R. L., Boorman, S. A., and Arabie, P. (1975). An algorithm for blocking relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of mathematical psychology*, 12:328–383.

Bureau, T. (2019). Tanzania in figures 2019. *Tanzania Bureau of Statistics*.

Bönke, T., Chattopadhyay, S., Shaohua, Lakner, C. W. D. M. E. G. A. G. C., Lawson-Remer, T., Leary, M. K., Massari, R., Montes, J., Newhouse, D., and Stace Nicholson Espen Beer Prydz Maika Schmidt José Cuesta, Mario Negre, A. S. (2016). Taking on equality world bank.

Caron, F. and Doucet, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196.

Carr-Hill, R. (2017). Improving population and poverty estimates with citizen surveys: Evidence from east africa. *World Development*, 93:249 – 259.

Cataldo, E. F., Johnson, R. M., Kellstedt, L. A., and Milbrath, L. W. (1970). Card sorting as a technique for survey interviewing. *Public Opinion Quarterly*, 34(2):202–215.

Cattelan, M., Varin, C., and Firth, D. (2012). Dynamic Bradley-Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150.

Channing Arndt, Andy McKay, F. T. (2016). *Growth and poverty in Sub-Saharan Africa*. Oxform University Press, United Stated of America, 198 Madison Avenue, New York, NY 10016.

Chen, D., McKune, S. L., Singh, N., Yousuf Hassen, J., Gebreyes, W., Manary, M. J., Bardosh, K., Yang, Y., Diaz, N., Mohammed, A., et al. (2021). Campylobacter colonization, environmental enteric dysfunction, stunting, and associated risk factors among young children in rural ethiopia: A cross-sectional study from the campylobacter genomics and environmental enteric dysfunction (caged) project. *Frontiers in public health*, 8:1043.

Cockayne, J. (2015). Why we need a global partnership to end modern slavery. Technical report, United Nations University.

Coja-Oghlan, A. and Lanka, A. (2010). Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics*, 23(4):1682–1714.

Corneli, M., Latouche, P., and Rossi, F. (2016). Block modelling in dynamic networks with non-homogeneous poisson processes and exact icl. *Social Network Analysis and Mining*, 6(1):55.

Cuaresma, J. C., Fengler, W., Kharas, H., Bekhtiar, K., Brottrager, M., and Hofer, M. (2018). Will the sustainable development goals be fulfilled? assessing present and future global poverty. *Palgrave Communications*, 4(1):1–8.

Cuesta, J., Negre, M., Bönke, T., Chattopadhyay, S., Chen, S., Durbin, W., Genoni, M. E., Goyal, A., Lakner, C., Lawson-Remer, T., andRenzo Massari, M. K. L., Montes, J., Newhouse, D., Nicholson, S., Prydz, E. B., Schmidt, M., , and Silwal, A. (2016). Poverty and shared prosperity.

Dalenberg, D. (2018). Preventing discrimination in the automated targeting of job advertisements. *Comput. Law Secur. Rev.*, 34:615–627.

Daoud, A., Halleröd, B., and Guha-Sapir, D. (2016). What is the association between absolute child poverty, poor governance, and natural disasters? a global comparison of some of the realities of climate change. *PLoS one*, 11(4):e0153296.

Dempster, A., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38.

Doreian, P., Batagelj, V., and Ferligoj, A. (2005). *Generalized blockmodeling*, volume 25. Cambridge university press.

Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981):1029–1031.

Edward, P. and Sumner, A. (2014). Estimating the scale and geography of global poverty now and in the future: how much difference do method and assumptions make? *World Development*, 58:67–82.

Eichleay, M., Mercer, S., Murashani, J., and Evens, E. (2016). Using unmanned aerial vehicles for development: perspectives from citizens and government officials in tanzania.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Emmanuel Abbe, C. S. (2015). Recovering communities in the general stochastic block model without knowing the parameter.

Engelmann, G., Smith, G., and Goulding, J. (2018). The unbanked and poverty: Predicting area-level socio-economic vulnerability from m-money transactions. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1357–1366.

Erdös, P. (1959). Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38.

Erman, A., Tariverdi, M., Obolensky, M., Chen, X., Vincent, R. C., Malgioglio, S., Rentschler, J., Hallegatte, S., and Yoshida, N. (2019). *Wading out the storm: The role of poverty in exposure, vulnerability and resilience to floods in Dar Es Salaam.* The World Bank.

Espey, J. (2019). Sustainable development will falter without data. *Nature*, 571(7765):299–300.

Europol (2016). Situation report: Trafficking in human beings in the eu. Technical report, Europol.

Evans, J. and Ruane, S. (2019). *Data in society: challenging statistics in an age of globalisation*. Policy Press.

Faust, K. and Wasserman, S. (1992). Blockmodels: Interpretation and evaluation. *Social Networks*, 14(1):5 – 61. Special Issue on Blockmodels.

Fields, G. S. (1989). Changes in poverty and inequality in developing countries. *The World Bank Research Observer*, 4(2):167–185.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., and Lewis, B. (2011). A social vulnerability index for disaster management. *Journal of homeland security and emergency management*, 8(1).

Flores-Macias, F. and Lawson, C. (2008). Effects of interviewer gender on survey responses: Findings from a household survey in mexico. *International journal of public opinion research*, 20(1):100–110.

Foody, G. M., Ling, F., Boyd, D. S., Li, X., and Wardlaw, J. (2019). Earth observation and machine learning to meet sustainable development goal 8.7: Mapping sites associated with slavery from space. *Remote Sens.*, 11.

Funke, T. and Becker, T. (2019). Stochastic block models: A comparison of variants and inference methods. *PloS one*, 14(4):e0215296.

Grosh, M. E. and Glewwe, P. (1998). Data watch: the world bank's living standards measurement study household surveys. *Journal of Economic Perspectives*, 12(1):187–196.

Guilford, J. P. (1954). Psychometric methods.

Haselton, M. G. (2003). The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events. *Journal of Research in Personality*, 37(1):34–47.

Hoare, C. A. R. (1961). Algorithm 64: quicksort. *Communications of the ACM*, 4(7):321.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983a). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983b). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Holland, P. W. and Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs: Rejoinder. *Journal of the American Statistical Association*, 76(373):62.

Holloway, J. and Mengersen, K. (2018). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sens.*, 10.

Hoogeveen, J., Tesliuc, E., Vakis, R., Dercon, S., et al. (2004). A guide to the analysis of risk, vulnerability and vulnerable groups. *World Bank. Washington, DC. Available on line at http://siteresources. worldbank. org/INTSRM/Publications/20316319/RVA. pdf. Processed.*

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Huddy, L., Billig, J., Bracciodieta, J., Hoeffler, L., Moynihan, P. J., and Pugliani, P. (1997). The effect of interviewer gender on the survey response. *Political Behavior*, 19(3):197–220.

International Labour Office (2012). Ilo indicators of forced labour. Technical report, ILO.

International Labour Organization (2017). Global estimates of modern slavery: Forced labour and forced marriage.

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

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

Jerven, M. (2013). *,Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It.* Cornell Univ. Press.

Jiang, Q., Zhang, Y., and Sun, M. (2009). Community detection on weighted networks: A variational bayesian method. In *Asian Conference on Machine Learning*, pages 176–190. Springer.

Johnston, R. and Brady, H. E. (2002). The rolling cross-section design. *Electoral Studies*, 21(2):283–295.

Jolliffe, I. T. (2002). Springer series in statistics.

Joshua Blumenstock, G. C. (2015). Predicting poverty and wealth from mobile phone metadata.

Kalton, G. and Schuman, H. (1982). The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society. Series A (General)*, 145(1):42–73.

Karrer, B. and Newman, M. E. (2011a). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

Karrer, B. and Newman, M. E. (2011b). Stochastic blockmodels and community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(1):1–10.

Kates, R. W. and Dasgupta, P. (2007). African poverty: A grand challenge for sustainability science. *Proceedings of the National Academy of Sciences*, 104(43):16747–16750.

Keene, O. N. (1995). The log transformation is special. *Statistics in medicine*, 14(8):811–819.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721.

Landman, T. and Silverman, B. W. (2019). Globalization and modern slavery. *Politics Gov.*, 7(4).

Lands and Ministry, H. S. D. (2000). National human settlements development policy, ministry of lands and human settlements development. dar es salaam.

Larsen, J. J. and Durgana, D. P. (2017). Measuring vulnerability and estimating prevalence of modern slavery. *Chance*, 30(3):21–29.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods.* Sage publications.

Lechner, A., McIntyre, N., Raymond, C., Witt, K., Scott, M., and Rifkin, W. (2017). Challenges of integrated modelling in mining regions to address social, environmental and economic impacts. *Environ. Model. Softw*, pages 268–281.

Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1).

Lei Tang, Huan Liu, J. Z. (2012). Identifying evolving groups in dynamic multimode networks. ieee transactions on knowledge and data engineerin.

Lipton, Z. (2018). The mythos of model interpretability. *Commun. ACM*, 61:36–43.

Liu, X., Cheng, H.-M., and Zhang, Z.-Y. (2019). Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering*.

Liviga, A. J. and Mekacha, R. D. (1998). Youth migration and poverty alleviation: A case study of petty traders (wamachinga) in dar es salaam.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80.

Lorrain, F. and White, H. C. (1977). Structural equivalence of individuals in social networks. In *Social Networks*, pages 67–98. Elsevier.

Lu, X. and Szymanski, B. K. (2019). A regularized stochastic block model for the robust community detection in complex networks. *Scientific reports*, 9(1):1–9.

Lundberg, S. M. and Lee, S.-I. (2017a). A Unified Approach to Interpreting Model Predictions.

Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural*

Information Processing Systems 30, pages 4765–4774. Curran Associates, Inc.

Lupu, N. and Michelitch, K. (2018). Advances in survey methods for the developing world. *Annual Review of Political Science*, 21:195–214.

Lynn, P. and Clarke, P. (2002). Separating refusal bias and non-contact bias: evidence from uk national surveys. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(3):319–333.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Madhawa, K., Lokanathan, S., Maldeniya, D., and Samarajiva, R. (2015). Using mobile network big data for land use classification cprsouth 2015.

Mariadassou, M., Robin, S., and Vacher, C. (2010a). Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics*, 4(2):715–742.

Mariadassou, M., Robin, S., Vacher, C., et al. (2010b). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, 4(2):715–742.

Martins, J. H. (2007). Household budgets as a social indicator of poverty and inequality in south africa. *Social Indicators Research*, 81(2):203–221.

Maru, S., Rajeev, S., Pokhrel, R., Poudyal, A., Mehta, P., Bista, D., Borgatta, L., and Maru, D. (2016). Determinants of institutional birth among women in rural nepal: a mixed-methods cross-sectional study. *BMC pregnancy and childbirth*, 16(1):1–8.

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

Matias, C., Rebafka, T., and Villers, F. (2015). Estimation and clustering in a semiparametric Poisson process stochastic block model for longitudinal networks. working paper or preprint.

Matias, C., Rebafka, T., and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680.

McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013). Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics and Data Analysis*, 60(1):12–31.

Mishra, L. (2001). A perspective plan to eliminate forced labour in india.

Moreau, M.-A. (2018). Ilo convention 29 forced labour convention, 1930 (no. 29). In *International and European Labour Law*, pages 1049–1073.

Morgan, J. and Olsen, W. (2014). Forced and unfree labour: An analysis. *Int Crit Theor.*, 4(1):21–37.

Mørup, M. and Hansen, L. K. (2009). Learning latent structure in complex networks.

Moyer, D., Gutman, B., Prasad, G., ver Steeg, G., and Thompson, P. (2015). Mixed Membership Stochastic Blockmodels for the Human Connectome. *Bayesian and grAphical Models for Biomedical Imaging*, pages 1–12.

Myers, M. A., Zaccaria, S., and Raphael, B. J. (2020). Identifying tumor clones in sparse single-cell mutation data. *Bioinformatics*, 36(Supplement_1):i186–i193.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Mach. Learn*, 52(3):239–281.

Newman, M. (2010). Networks: An introduction.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087.

of Health Community Development Gender Elderly, M., Children Ministry of Health Zanzibar National Bureau of Statistics, O. o. t. C. G. S., and ICF (2016). Tanzania demographic and health survey and malaria indicator survey (tdhs-mis) 2015-16.

of Statistics, T. N. B. (2019). The 2017-18 household budget survey: Key indicators report.

Olsen, R. A. and Cox, C. M. (2001). The influence of gender on the perception and response to investment risk: The case of professional investors. *The journal of psychology and financial markets*, 2(1):29–36.

Peixoto, T. P. (2012). Entropy of stochastic blockmodel ensembles. *Physical Review E*, 85(5):056122.

Peixoto, T. P. (2013). Parsimonious module inference in large networks. *Physical review letters*, 110(14):148701.

Peixoto, T. P. (2014a). Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1):012804.

Peixoto, T. P. (2014b). The graph-tool python library. *figshare*.

Peixoto, T. P. (2014c). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047.

Peixoto, T. P. (2015). Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1):011033.

Peixoto, T. P. (2018). Nonparametric weighted stochastic block models. *Physical Review E*, 97(1):012306.

Perez, A., Ganguli, S., Ermon, S., Azzari, G., Burke, M., and Lobell, D. (2016). Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty. Technical report, Stanford University.

Perry, W., McInnis, B., Price, C. C., Smith, S. C., and Hollywood, J. S. (2013). Predictive policing: The role of crime forecasting in law enforcement operations.

Phelan, G. C. and Whelan, J. T. (2018). Hierarchical Bayesian Bradley-Terry for applications in major league baseball. *Mathematics for Applications*, 7(1):71–84.

Pinkovskiy, M. and Sala-i Martin, X. (2016). Lights, camera… income! illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, 131(2):579–631.

Pollit, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19:281–300.

Randall, S. and Coast, E. (2015). Poverty in african households: the limits of survey and census representations. *The Journal of Development Studies*, 51(2):162–177.

Reichardt, J., Alamino, R., and Saad, D. (2011). The interplay between microscopic and mesoscopic structures in complex networks. *PloS one*, 6(8):e21282.

Roy, D., Palavalli, B., Menon, N., King, R., Pfeffer, K., Lees, M., and Sloot, P. M. (2018). Survey-based socio-economic data from slums in bangalore, india. *Scientific data*, 5:170200.

Rugg, G. and McGeorge, P. (1997). The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93.

Ruwanpura, K. N. and Rai, P. (2004). Forced labour: Definition, indicators and measurement. Technical report, ILO.

Saam, N. J. and Harrer, A. (1999). Simulating norms, social inequality, and functional change in artificial societies. *J. Artificial Soc.Social Simul.*, 2.

Santosa, F. and Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms.

Seymour, R. G., Sirl, D., Preston, S., Dryden, I. L., Ellis, M. J. A., Perrat, B., and Goulding, J. (2020). The Bayesian spatial Bradley–Terry model: Urban deprivation modeling in Tanzania.

Silverman, B. (2020). Multiple-systems analysis for the quantification of modern slavery: classical and bayesian approaches. *J R STAT SOC A*.

Smith-Clarke, C. and Capra, L. (2016). Beyond the baseline: Establishing the value in mobile phone based poverty estimates. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 425–434, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Smith-Clarke, C., Mashhadi, A., and Capra, L. (2014). Poverty on the cheap. In *CHI 2014, One of a CHInd*, Toronto, CA.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Stanton, M. A. and Mann, J. (2012). Early social networks predict survival in wild bottlenose dolphins. *PloS one*, 7(10):e47508.

Stuart-Fox, D. M., Firth, D., Moussalli, A., and Whiting, M. J. (2006). Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271.

Sundar, R. and Sharma, A. (2002). Morbidity and utilisation of healthcare services: a survey of urban poor in delhi and chennai. *Economic and Political Weekly*, pages 4729–4740.

Supekar, K., Menon, V., Rubin, D., Musen, M., and Greicius, M. D. (2008). Network analysis of intrinsic functional brain connectivity in alzheimer's disease. *PLoS Comput Biol*, 4(6):e1000100.

Tallberg, C. (2004). A bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1):1–23.

Tarozzi, A. and Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics*, 91(4):773–792.

Ties Boerma, J. and Sommerfelt, A. E. (1993). Demographic and health surveys (dhs: contributions and limitations. *World health statistics quarterly 1993; 46 (4): 222-226*.

Torres, M. T., Perrat, B., Iliffe, M., Goulding, J., and Valstar, M. (2017).

Automatic pixel-level land-use prediction using deep convolutional neural networks.

Tortora, R. D., Srinivasan, R., and Esipova, N. (2010). The gallup world poll. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, pages 535–543.

Tran, T. N., Afanador, N. L., Buydens, L. M., and Blanchet, L. (2014). Interpretation of variable importance in partial least squares with significance multivariate correlation (smc). *Chemometrics and Intelligent Laboratory Systems*, 138:153–160.

Turner, H. and Firth, D. (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software, Articles*, 48(9):1–21.

United Republic of Tanzania, N. B. o. S. (2017). Tanzania national panel survey wave 4, 2014 – 2015.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Nerini, F. F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nat.*, 11(233).

Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013). Model-based clustering of large networks. *The annals of applied statistics*, 7(2):1010.

Wang (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Wasserman, S. and Anderson, C. (1987). Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9(1):1 – 36.

Wasserman, S. and Galaskiewicz, J. (1994). Advances in social network analysis: Research in the social and behavioral sciences, volume 171. sage publications.

Watmough, G. R., Atkinson, P. M., Saikia, A., and Hutton, C. W. (2016). Understanding the evidence base for povertyâ environment relationships using remotely sensed satellite data: An example from assam, india. *World Development*, 78:188 – 203.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Sci.*, 338.

White, H. C., Boorman, S. A., and Breiger, R. L. (1976a). Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *American Journal of Sociology*, 81(4):730–780.

White, H. C., Boorman, S. A., and Breiger, R. L. (1976b). Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4):730–780.

Williams, J. W. J. (1964). Algorithm 232: heapsort. *Commun. ACM*, 7:347–348.

Wilpen, G. and Daniel, N. (2007). Detecting and preventing emerging epidemics of crime. *Adv. Dis. Surveillance*, 4(13).

Wilson, J. D., Stevens, N. T., and Woodall, W. H. (2016). Modeling and detecting change in temporal networks via a dynamic degree corrected stochastic block model. *arXiv preprint arXiv:1605.04049*.

Wold, S., Lindgren, F., and Geladi, P. (1993). The kernel algorithm for pls.

Woo Park, H. and Jankowski, N. W. (2008). A hyperlink network analysis of citizen blogs in south korean politics. *Javnost-The Public*, 15(2):57–74.

Xie, M., Jean, N., Burke, M., Lobell, D. B., and Ermon, S. (2015a). Transfer learning from deep features for remote sensing and poverty mapping. *CoRR*, abs/1510.00098.

Xie, M., Jean, N., Burke, M., Lobell, D. B., and Ermon, S. (2015b). Transfer learning from deep features for remote sensing and poverty mapping.

Xu, K. S. and Hero, A. O. (2013). Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 201–210. Springer.

Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P., and Zhu, Y. (2014). Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(5):P05007.

Yang, J.-X. and Zhang, Y. (2020). Epidemic spreading of evolving community structure. *Chaos, Solitons & Fractals*, 140:110101.

Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine learning*, 82(2):157–189.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473.

Zhang, Y., Chen, K., Sampson, A., Hwang, K., and Luna, B. (2019). Node features adjusted stochastic block model. *Journal of Computational and Graphical Statistics*, 28(2):362–373.

Zhang, Y., Levina, E., Zhu, J., et al. (2016). Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178.

Appendix A

Grid Survey Instruction Worksheet

N/LAB

Ground Truth Grid Survey

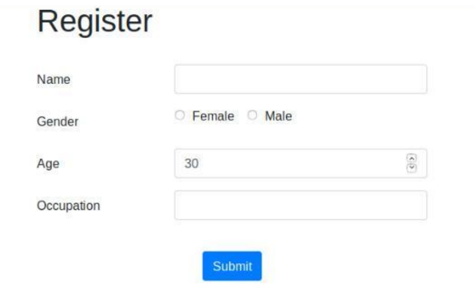
Dar es Salaam

Thank you, for taking part in our grid survey! This survey is part of a fine-grained ground truth data collection we are doing of geographical regions in Dar es Salaam. We are collecting demographic truths, to underpin a number of projects undertaken by the N/LAB at the University of Nottingham, England.

We are so grateful for your time.

How to do the Survey

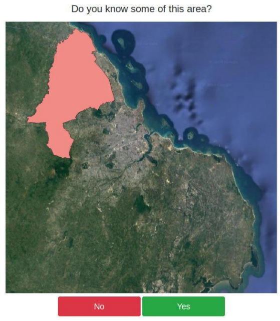
1. Register



Register

Name	
Gender	○ Female ○ Male
Age	30
Occupation	

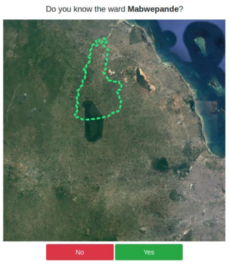
2. Do you know some of this clustered area?



Do you know some of this area?

If you do not know the cluster, you can move on to the next cluster

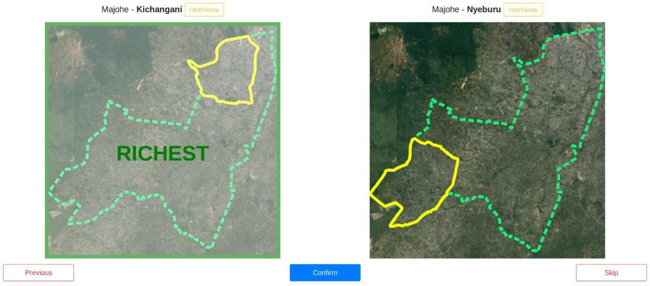
If you do know the cluster you will then be asked to select the wards you know.



Do you know the ward **Mabwepande**?

3. Now that you have selected all the wards you know, the ranking begins!

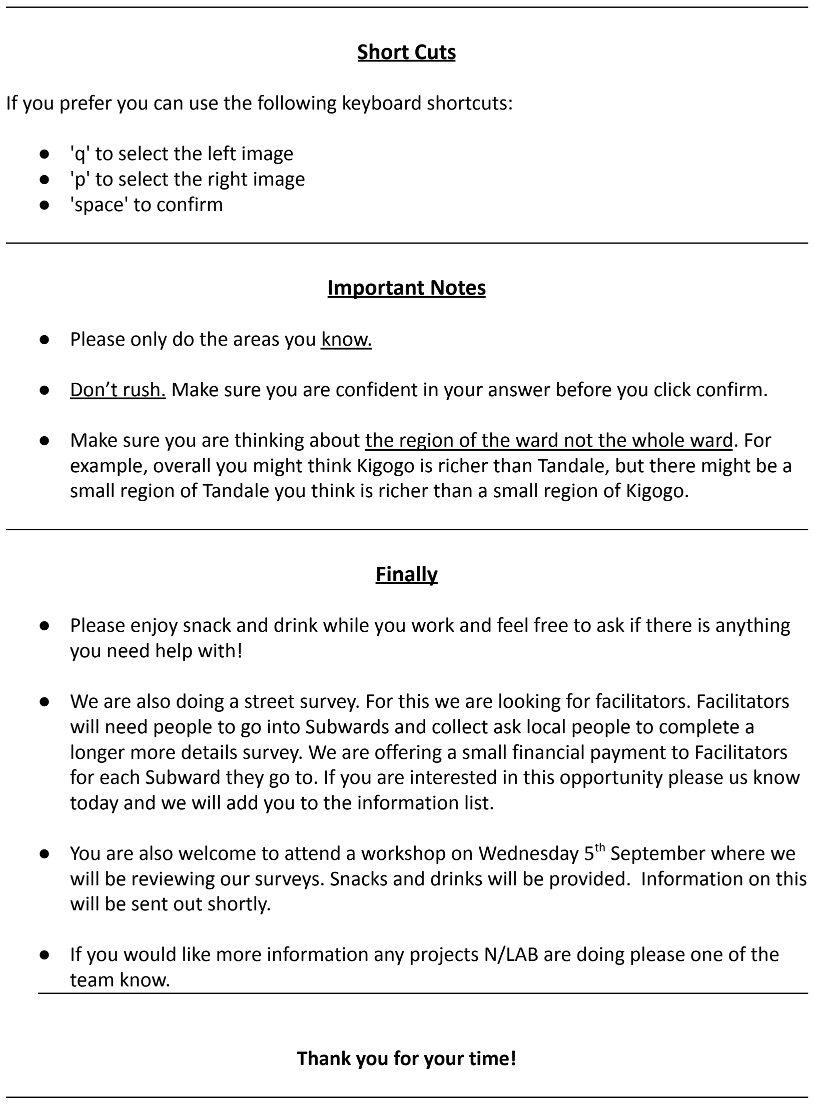
4. Click the richest and then select ‘Confirm’.



Majohe - **Kichangani** | Majohe - **Nyeburu**

5. As you go along if you do not know one of the regions you can click ‘I don’t know’. If you know them both but you cannot decide which is richer you can click ‘Skip’. If you know them both and have decided which is a richer area, you can click ‘Confirm’.

6. At the end of the 2 hour session, click ‘Finish’ at the top right.



Short Cuts

If you prefer you can use the following keyboard shortcuts:

- 'q' to select the left image
- 'p' to select the right image
- 'space' to confirm

Important Notes

- Please only do the areas you know.
- Don’t rush. Make sure you are confident in your answer before you click confirm.
- Make sure you are thinking about the region of the ward not the whole ward. For example, overall you might think Kigogo is richer than Tandale, but there might be a small region of Tandale you think is richer than a small region of Kigogo.

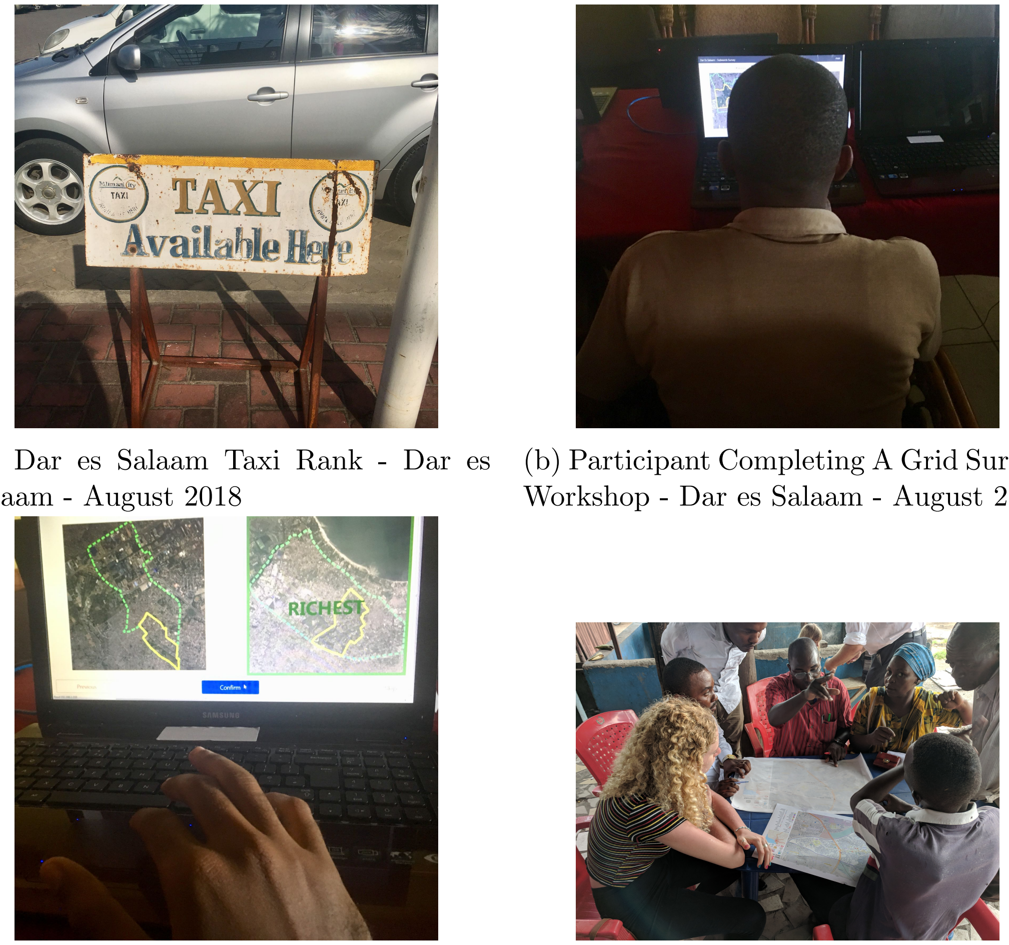
Finally

- Please enjoy snack and drink while you work and feel free to ask if there is anything you need help with!
- We are also doing a street survey. For this we are looking for facilitators. Facilitators will need people to go into Subwards and collect ask local people to complete a longer more details survey. We are offering a small financial payment to Facilitators for each Subward they go to. If you are interested in this opportunity please us know today and we will add you to the information list.
- You are also welcome to attend a workshop on Wednesday 5th September where we will be reviewing our surveys. Snacks and drinks will be provided. Information on this will be sent out shortly.
- If you would like more information any projects N/LAB are doing please one of the team know.

Thank you for your time!

Appendix B

Project Photos



(a) Dar es Salaam Taxi Rank - Dar es Salaam - August 2018

(b) Participant Completing A Grid Survey Workshop - Dar es Salaam - August 2018

(c) Example Screen View for Grid Survey - Dar es Salaam - August 2018.

(d) Community Meeting - Dar es Salaam - August 2018.

Figure B.1: Dar es Salaam Survey Photos

(a) Community Meeting - Dar es Salaam - August 2018

(b) Community Meeting - Dar es Salaam - August 2018

(c) Example Pilot Survey Review Meeting - Dar es Salaam - August 2018

(d) Example Pilot Survey Review Meeting - Dar es Salaam - August 2018

Figure B.2: Dar es Salaam Survey Photos

(a) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(b) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(c) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(d) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

Figure B.3: Dar es Salaam Survey Photos

(a) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(b) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(c) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(d) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

Figure B.4: Dar es Salaam Survey Photos

(a) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(b) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(c) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(d) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

Figure B.5: Dar es Salaam Survey Photos

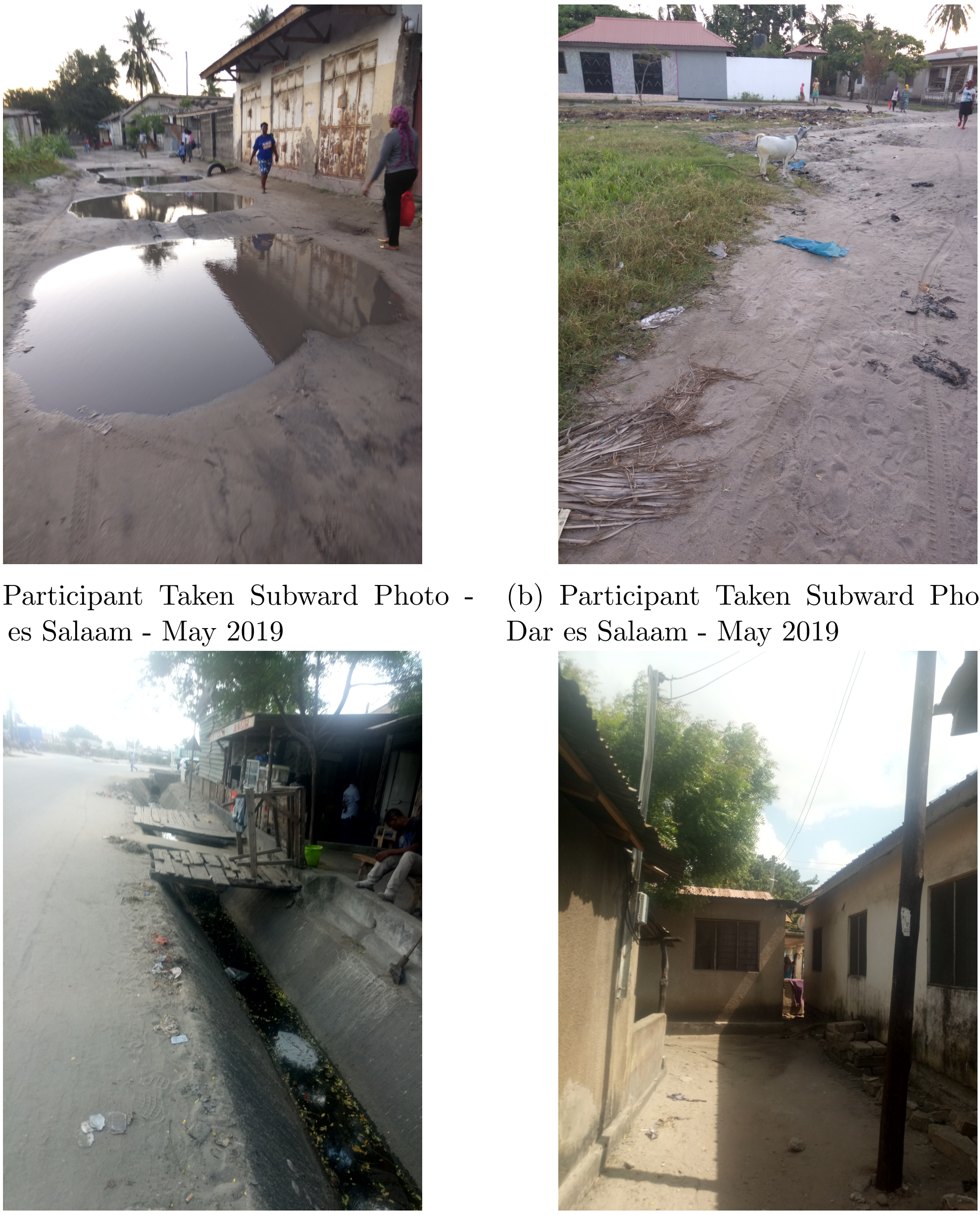
(a) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(b) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(c) Street Survey In-Field Data Collection - Dar es Salaam - May 2019

(d) Participant Taken Subward Photo - Dar es Salaam - May 2019

Figure B.6: Dar es Salaam Survey Photos



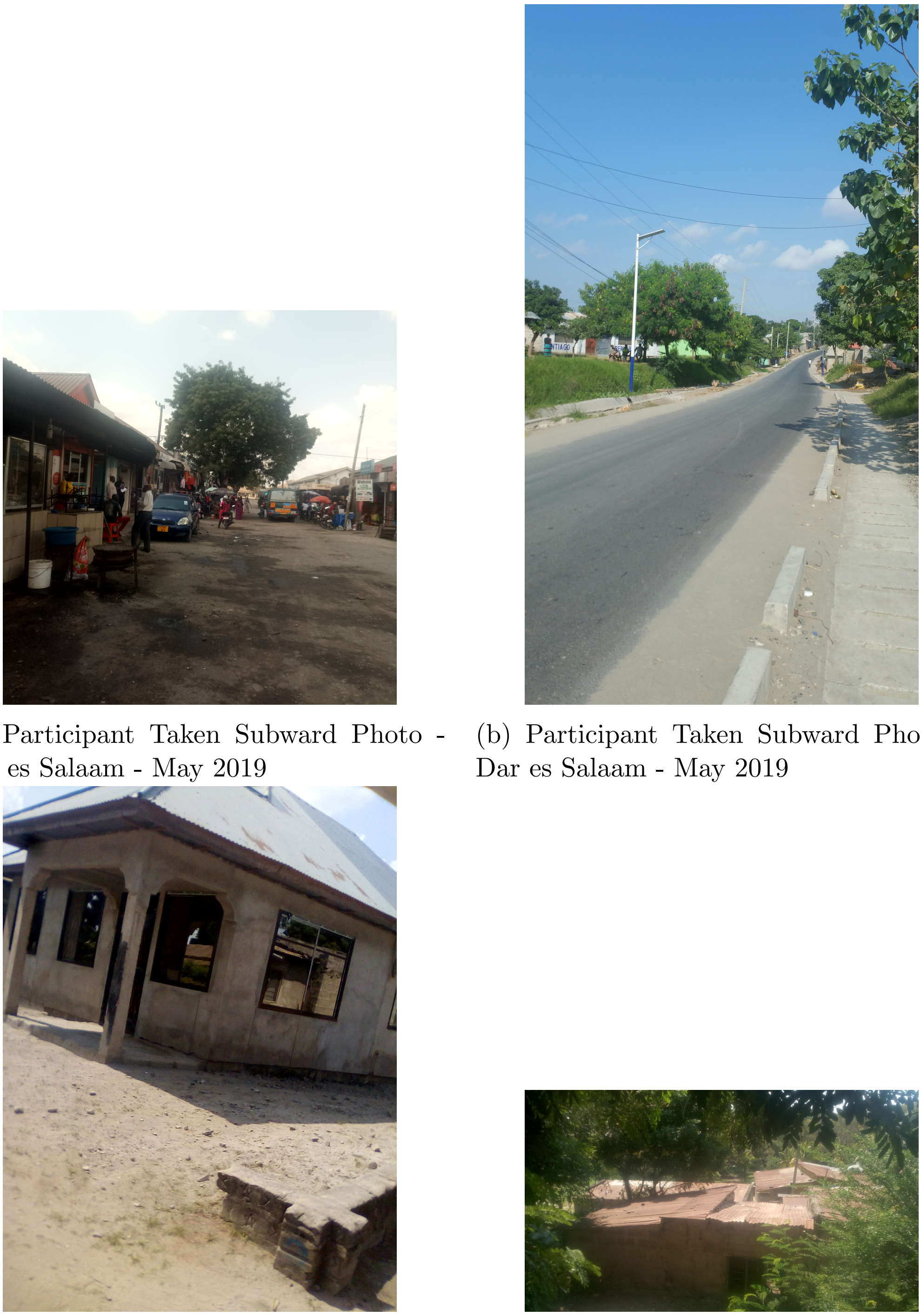
(a) Participant Taken Subward Photo - Dar es Salaam - May 2019

(b) Participant Taken Subward Photo - Dar es Salaam - May 2019

(c) Participant Taken Subward Photo - Dar es Salaam - May 2019

(d) Participant Taken Subward Photo - Dar es Salaam - May 2019

Figure B.7: Dar es Salaam Survey Photos



(a) Participant Taken Subward Photo - Dar es Salaam - May 2019

(b) Participant Taken Subward Photo - Dar es Salaam - May 2019

(c) Participant Taken Subward Photo - Dar es Salaam - May 2019

(d) Participant Taken Subward Photo - Dar es Salaam - May 2019

Figure B.8: Dar es Salaam Survey Photos

Appendix C

Grid Survey Results

Table C.1: Grid Survey Results

subward.name	ward.name	subward.id	rank	prop.known
Idrisa	Magomeni	1	93	0.888298
Dossi	Magomeni	2	95	0.888298
Makuti 'A'	Magomeni	3	104	0.888298
Makuti 'B'	Magomeni	4	96	0.888298
Sunna	Magomeni	5	174	0.888298
Sisi Kwa Sisi	Makurumla	6	327	0.430851
Kagera	Makurumla	7	259	0.430851
Kimamba	Makurumla	8	155	0.430851
Mianzini	Makurumla	9	293	0.430851
Kwajongo	Makurumla	10	250	0.430851
Kilimahewa	Makurumla	11	226	0.430851
Vigaeni	Ndugumbi	12	379	0.37234
Makanya	Ndugumbi	13	402	0.37234
Kagera Mikoroshini	Ndugumbi	14	352	0.37234

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mpakani	Ndugumbi	15	410	0.37234
Kwa Mkunduge	Tandale	16	380	0.81383
Kwa Tumbo	Tandale	17	371	0.81383
Sokoni	Tandale	18	321	0.81383
Kwa Pakacha	Tandale	19	382	0.81383
Muhalitani	Tandale	20	378	0.81383
Kwa Mtogole	Tandale	21	338	0.81383
Mwinjuma	Mwananyamala	22	289	0.851064
Kambangwa	Mwananyamala	23	264	0.851064
Msisiri B	Mwananyamala	24	241	0.851064
Kwa Kopa	Mwananyamala	25	303	0.851064
Bwawani	Mwananyamala	26	207	0.851064
Msisiri ’A’	Mwananyamala	27	242	0.851064
Msolomi	Mwananyamala	28	243	0.851064
Masaki	Msasani	29	1	0.765957
Oysterbay	Msasani	30	4	0.765957
Makangira	Msasani	31	71	0.765957
Bonde la Mpunga	Msasani	32	56	0.765957
Mikoroshini	Msasani	33	48	0.765957
Kinondoni Mjini	Kinondoni	34	16	0.882979
Kinondoni Shamba	Kinondoni	35	40	0.882979
Ada Estate	Kinondoni	36	6	0.882979
Kumbukumbu	Kinondoni	37	20	0.882979
Makumbusho	Mzimuni	38	49	0.484043

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Idrissa	Mzimuni	39	130	0.484043
Mwinyimkuu	Mzimuni	40	82	0.484043
Mtambani	Mzimuni	41	98	0.484043
Kigogo Mkwajuni	Kigogo	42	312	0.824468
Kigogo Kati	Kigogo	43	290	0.824468
Mbuyuni	Kigogo	44	269	0.824468
Mabibo	Mabibo	45	122	0.898936
Jitegemee	Mabibo	46	181	0.898936
Kanuni	Mabibo	47	220	0.898936
Matokeo	Mabibo	48	227	0.898936
Azimio	Mabibo	49	271	0.898936
Mabibo Farasi	Mabibo	50	153	0.898936
Madizini	Manzese	51	270	0.984043
Mwembeni	Manzese	52	222	0.984043
Mnazi Mmoja	Manzese	53	163	0.984043
Kilimani	Manzese	54	233	0.984043
Mvuleni	Manzese	55	308	0.984043
Muungano	Manzese	56	254	0.984043
Uzuri	Manzese	57	236	0.984043
Chakula Bora	Manzese	58	315	0.984043
NHC	Ubungo	59	32	0.904255
Kisiwani	Ubungo	60	51	0.904255
Chuo Kikuu	Ubungo	61	18	0.904255
Kibo	Ubungo	62	92	0.904255

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Msewe	Ubungo	63	65	0.904255
Kiluvya	Kibamba	64	221	0.702128
Gogoni	Kibamba	65	246	0.702128
Hondogo	Kibamba	66	299	0.702128
Kibwegere	Kibamba	67	262	0.702128
Kibamba	Kibamba	68	169	0.702128
Matosa	Goba	69	170	0.712766
Kulangwa	Goba	70	232	0.712766
Kinzudi	Goba	71	190	0.712766
Goba	Goba	72	145	0.712766
Ukwamani	Kawe	73	68	0.781915
Mzimuni	Kawe	74	97	0.781915
Mbezi Beach ’A’	Kawe	75	19	0.781915
Mbezi Beach ’B’	Kawe	76	15	0.781915
Tegeta	Kunduchi	77	58	0.760638
Pwani	Kunduchi	78	70	0.760638
Ununio	Kunduchi	79	29	0.760638
Kondo	Kunduchi	80	52	0.760638
Mtongani	Kunduchi	81	44	0.760638
Kilongamiwa	Kunduchi	82	63	0.760638
Maputo	Mbweni	83	74	0.62766
Mbweni	Mbweni	84	47	0.62766
Mpiji	Mbweni	85	99	0.62766
Basihaya	Bunju	86	115	0.739362

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Boko	Bunju	87	101	0.739362
Dovya	Bunju	88	114	0.739362
Kilungule	Bunju	89	146	0.739362
Bunju 'A'	Bunju	90	62	0.739362
Makoka	Makuburi	91	184	0.611702
Kajima	Makuburi	92	217	0.611702
Kibangu	Makuburi	93	180	0.611702
Makuburi	Makuburi	94	108	0.611702
Mwongozo	Makuburi	95	277	0.611702
NHC	Mburahati	96	132	0.760638
Barafu	Mburahati	97	203	0.760638
Kisiwani	Mburahati	98	274	0.760638
Mbuyuni	Makumbusho	99	133	0.893617
Makumbusho	Makumbusho	100	45	0.893617
Minazini	Makumbusho	101	111	0.893617
Mchangani	Makumbusho	102	139	0.893617
Kisiwani	Makumbusho	103	103	0.893617
Sinza 'C'	Sinza	104	42	0.904255
Sinza 'A'	Sinza	105	34	0.904255
Sinza 'B'	Sinza	106	73	0.904255
Sinza 'E'	Sinza	107	50	0.904255
Sinza 'D'	Sinza	108	88	0.904255
Mwenge	Kijitonyama	109	28	0.898936
Mpakani 'A'	Kijitonyama	110	59	0.898936

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Kijitonyama	Kijitonyama	111	36	0.898936
Bwawani	Kijitonyama	112	102	0.898936
Alimaua 'A'	Kijitonyama	113	105	0.898936
Aliamua 'B'	Kijitonyama	114	94	0.898936
Mpakani 'B'	Kijitonyama	115	81	0.898936
Baruti	Kimara	116	138	0.914894
Kimara Baruti	Kimara	117	120	0.914894
Kilungule 'A'	Kimara	118	215	0.914894
Kilungule 'B'	Kimara	119	201	0.914894
Mavurunza	Kimara	120	209	0.914894
Golani	Kimara	121	173	0.914894
Mikocheni 'B'	Mikocheni	122	2	0.81383
TPDC	Mikocheni	123	5	0.81383
Ally Hassan Mwinyi	Mikocheni	124	11	0.81383
Regent Estate	Mikocheni	125	12	0.81383
Darajani	Mikocheni	126	31	0.81383
Mikocheni 'A'	Mikocheni	127	27	0.81383
Mshikamano	Mbezi	128	234	0.765957
Msakuzi	Mbezi	129	196	0.765957
Mpiji Magohe	Mbezi	130	263	0.765957
Msumi	Mbezi	131	191	0.765957
Makabe	Mbezi	132	177	0.765957
Mbezi Luisi	Mbezi	133	156	0.765957
Hananasif	Hananasifu	134	69	0.505319

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mkunguni A	Hananasifu	135	109	0.505319
Mkunguni B	Hananasifu	136	39	0.505319
Kisutu	Hananasifu	137	38	0.505319
Saranga	Saranga	138	414	0.430851
Upendo	Saranga	139	388	0.430851
Stopover	Saranga	140	267	0.430851
Kimara B	Saranga	141	305	0.430851
King'ongo	Saranga	142	356	0.430851
Michungwani	Saranga	143	358	0.430851
Matangini	Saranga	144	406	0.430851
Mpakani	Kwembe	145	449	0.367021
King'azi	Kwembe	146	417	0.367021
Kwembe	Kwembe	147	282	0.367021
Kisopwa	Kwembe	148	439	0.367021
Luguruni	Kwembe	149	409	0.367021
Kwa Yusufu	Msigani	150	224	0.484043
Msigani	Msigani	151	216	0.484043
Temboni	Msigani	152	172	0.484043
Malamba Mawili	Msigani	153	206	0.484043
Jogoo	Mbezi Juu	154	55	0.728723
Ndumbwi	Mbezi Juu	155	67	0.728723
Mbezi Kati	Mbezi Juu	156	89	0.728723
Mbezi Mtoni	Mbezi Juu	157	87	0.728723
Mbezi Juu	Mbezi Juu	158	54	0.728723

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mbuyuni	Makongo	159	183	0.829787
Changanyikeni	Makongo	160	135	0.829787
Makongo	Makongo	161	80	0.829787
Mlalakuwa	Makongo	162	168	0.829787
Bunju ’B’	Mabwepande	163	218	0.574468
Mabwepande	Mabwepande	164	361	0.574468
Mbopo	Mabwepande	165	425	0.574468
Madala	Wazo	166	131	0.648936
Mivumoni	Wazo	167	175	0.648936
Kisanga	Wazo	168	171	0.648936
Wazo	Wazo	169	113	0.648936
Salasala	Wazo	170	72	0.648936
Kilimahewa	Wazo	171	160	0.648936
Mongo la Ndege	Ukonga	172	205	0.808511
Mwembe Madafu	Ukonga	173	154	0.808511
Mazizini	Ukonga	174	199	0.808511
Markaz	Ukonga	175	223	0.808511
Kinyamwezi	Pugu	176	318	0.670213
Kigogo Fresh	Pugu	177	364	0.670213
Bombani	Pugu	178	353	0.670213
Kichangani	Pugu	179	325	0.670213
Pugu Station	Pugu	180	340	0.670213
Bangulo	Pugu	181	349	0.670213
Yange yange	Msongola	182	424	0.297872

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mbondole	Msongola	183	445	0.297872
Mvuleni	Msongola	184	387	0.297872
Kitonga	Msongola	185	415	0.297872
Kidole	Msongola	186	435	0.297872
Mkera	Msongola	187	451	0.297872
Sangara	Msongola	188	355	0.297872
Uwanja wa Nyani	Msongola	189	446	0.297872
Kiboga	Msongola	190	442	0.297872
Mandela	Tabata	191	123	0.888298
Matumbi	Tabata	192	194	0.888298
Msimbazi	Tabata	193	142	0.888298
Tenge	Tabata	194	137	0.888298
Tabata Kisiwani	Tabata	195	166	0.888298
Tabata	Tabata	196	107	0.888298
Bonyokwa	Kinyerezi	197	214	0.867021
Kinyerezi	Kinyerezi	198	134	0.867021
Kifuru	Kinyerezi	199	202	0.867021
Shariff Shamba	Ilala	200	100	0.93617
Karume	Ilala	201	43	0.93617
Kasulu	Ilala	202	85	0.93617
Mafuriko	Ilala	203	147	0.93617
Msimbazi Bondeni	Mchikichini	204	124	0.792553
Ilala Kota	Mchikichini	205	106	0.792553
Mission Quarter	Mchikichini	206	77	0.792553

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Miembeni	Vingunguti	207	313	0.819149
Mtakuja	Vingunguti	208	330	0.819149
Kombo	Vingunguti	209	346	0.819149
Mtambani	Vingunguti	210	329	0.819149
Mogo	Kipawa	211	211	0.664894
Karakata	Kipawa	212	261	0.664894
Kipunguni	Kipawa	213	136	0.664894
Kisiwani	Buguruni	214	150	0.888298
Malapa	Buguruni	215	161	0.888298
Madenge	Buguruni	216	212	0.888298
Mnyamani	Buguruni	217	158	0.888298
Kariakoo Kaskazini	Kariakoo	218	17	0.925532
Kariakoo Mashariki	Kariakoo	219	23	0.925532
Kariakoo Magharibi	Kariakoo	220	26	0.925532
Mtambani	Jangwani	221	117	0.914894
Ukombozi	Jangwani	222	86	0.914894
Mnazi Mmoja	Jangwani	223	30	0.914894
Mtambani A	Jangwani	224	75	0.914894
Gerezani Mashariki	Gerezani	225	33	0.893617
Gerezani Magharibi	Gerezani	226	24	0.893617
Mtendeni	Kisutu	227	22	0.840426
Kisutu	Kisutu	228	10	0.840426
Kitumbini	Mchafukoge	229	128	0.441489
Mchafukoge	Mchafukoge	230	151	0.441489

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Kitonga	Upanga Mashariki	231	3	0.760638
Kibasila	Upanga Mashariki	232	7	0.760638
Mfaume	Upanga Magharibi	233	13	0.718085
Fire	Upanga Magharibi	234	21	0.718085
Charambe	Upanga Magharibi	235	14	0.718085
Sea View	Kivukoni	236	8	0.861702
Kivukoni	Kivukoni	237	9	0.861702
Kiwalani	Kiwalani	238	197	0.648936
Minazi Mirefu	Kiwalani	239	164	0.648936
Yombo	Kiwalani	240	302	0.648936
Kigilagila	Kiwalani	241	287	0.648936
Migombani	Segerea	242	178	0.882979
Ugombolwa	Segerea	243	193	0.882979
Amani	Segerea	244	213	0.882979
Liwiti	Segerea	245	179	0.882979
Segerea	Segerea	246	127	0.882979
Kitunda kati	Kitunda	247	260	0.707447
Mzinga	Kitunda	248	198	0.707447
Kimwani	Chanika	249	430	0.654255
Tunguni	Chanika	250	431	0.654255
Vikongolo	Chanika	251	396	0.654255
Yongwe	Chanika	252	399	0.654255
Lukooni	Chanika	253	422	0.654255
Lubakaya	Chanika	254	436	0.654255

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Zingiziwa	Chanika	255	412	0.654255
Nzasa	Chanika	256	420	0.654255
Kipunguni 'B'	Kivule	257	324	0.547872
Kivule	Kivule	258	291	0.547872
Ulongoni	Gongolamboto	259	231	0.845745
Gongolamboto	Gongolamboto	260	208	0.845745
Guruka Kwalala	Gongolamboto	261	252	0.845745
Mji Mpya	Majohe	262	403	0.505319
Kichangani	Majohe	263	362	0.505319
Kivule	Majohe	264	375	0.505319
Zavala	Majohe	265	433	0.505319
Kigezi	Majohe	266	376	0.505319
Nyeburu	Majohe	267	397	0.505319
Mgeule	Majohe	268	398	0.505319
Kisukuru	Kimanga	269	188	0.760638
Tembomgwaza	Kimanga	270	294	0.760638
Kimanga	Kimanga	271	237	0.760638
Darajani	Kimanga	272	322	0.760638
Ferry	Kigamboni	273	25	0.888298
Tuamoyo	Kigamboni	274	90	0.888298
Kigamboni	Kigamboni	275	46	0.888298
Vijibweni	Vijibweni	276	118	0.595745
Kisiwani	Vijibweni	277	141	0.595745
Mkwajuni	Vijibweni	278	167	0.595745

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Kibene	Vijibweni	279	187	0.595745
Uvumba	Kibada	280	309	0.648936
Kifurukwe	Kibada	281	314	0.648936
Kiziza	Kibada	282	255	0.648936
Nyakwale	Kibada	283	316	0.648936
Sokoni	Kibada	284	296	0.648936
Kichangani	Kibada	285	189	0.648936
Kigogo	Kisarawe II	286	360	0.617021
Vumilia Ukooni	Kisarawe II	287	401	0.617021
Mwasonga	Kisarawe II	288	373	0.617021
Mkamba	Kisarawe II	289	389	0.617021
Mwaninga	Kisarawe II	290	368	0.617021
Lingato	Kisarawe II	291	357	0.617021
Kichangani	Kisarawe II	292	416	0.617021
Dege	Somangila	293	369	0.25
Malimbika	Somangila	294	452	0.25
Mwera	Somangila	295	450	0.25
Kizani	Somangila	296	438	0.25
Mbwamaji	Somangila	297	437	0.25
Bamba	Somangila	298	444	0.25
Visikini	Somangila	299	440	0.25
Mkwajuni	Somangila	300	432	0.25
Kichangani	Somangila	301	381	0.25
Shirikisho	Somangila	302	447	0.25

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Sara	Somangila	303	394	0.25
Minondo	Somangila	304	354	0.25
Mwanzo Mgumu	Somangila	305	441	0.25
Kijaka	Kimbiji	306	423	0.553191
Ngombanya	Kimbiji	307	411	0.553191
Kizito Huonjwa	Kimbiji	308	386	0.553191
Mikenge	Kimbiji	309	407	0.553191
Kwa Chale	Kimbiji	310	383	0.553191
Golani	Kimbiji	311	408	0.553191
Bughudadi	Mbagala	312	342	0.845745
Kizinga	Mbagala	313	347	0.845745
Mangaya	Mbagala	314	306	0.845745
Mbagala	Mbagala	315	192	0.845745
Moringe	Mbagala	316	266	0.845745
Msufini	Chamazi	317	335	0.670213
Mwembe Bamia	Chamazi	318	366	0.670213
Kiponza	Chamazi	319	393	0.670213
Rufu	Chamazi	320	377	0.670213
Magengeni	Chamazi	321	367	0.670213
Kisewe	Chamazi	322	317	0.670213
Yombo Vituka	Yombo Vituka	323	310	0.659574
Machimbo	Yombo Vituka	324	248	0.659574
Sigara	Yombo Vituka	325	273	0.659574
Kilungule	Charambe	326	244	0.526596

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Kwazomboko	Charambe	327	363	0.526596
Kurasini mjimpya	Charambe	328	159	0.526596
Rangi Tatu	Charambe	329	204	0.526596
Nzasa 'B'	Charambe	330	341	0.526596
Nzasa 'A'	Charambe	331	372	0.526596
Majimatitu B	Charambe	332	284	0.526596
Vikunai	Toangoma	333	328	0.494681
Changanyikeni	Toangoma	334	286	0.494681
Mikwambe	Toangoma	335	280	0.494681
Mwapemba	Toangoma	336	301	0.494681
Ponde	Toangoma	337	333	0.494681
Masuliza	Toangoma	338	343	0.494681
Malela	Toangoma	339	256	0.494681
Toangoma	Toangoma	340	298	0.494681
Masaki	Toangoma	341	162	0.494681
Goroka	Toangoma	342	304	0.494681
Kongowe	Toangoma	343	331	0.494681
Mzinga	Toangoma	344	374	0.494681
Keko Machungwa	Miburani	345	76	0.393617
Keko Juu	Miburani	346	110	0.393617
Uwanja wa Taifa	Miburani	347	37	0.393617
Miburani	Miburani	348	165	0.393617
Wailes	Miburani	349	143	0.393617
Maganga	Temeke	350	251	0.87234

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Njaro	Temeke	351	121	0.87234
Temeke	Temeke	352	119	0.87234
Matumbi	Temeke	353	186	0.87234
Bustani	Mtoni	354	275	0.670213
Sabasaba	Mtoni	355	79	0.670213
Mtoni	Mtoni	356	279	0.670213
Relini	Mtoni	357	323	0.670213
Keko Mwanga 'B'	Keko	358	64	0.829787
Keko Mwanga 'A'	Keko	359	53	0.829787
Keko Magurumbasi 'A'	Keko	360	61	0.829787
Keko Magurumbasi 'B'	Keko	361	144	0.829787
Keko 'B'	Keko	362	66	0.829787
Mivinjeni	Kurasini	363	35	0.808511
Kiungani	Kurasini	364	91	0.808511
Kurasini	Kurasini	365	41	0.808511
Minazini	Kurasini	366	57	0.808511
Shimo la Udongo	Kurasini	367	126	0.808511
Mjimpya	Azimio	368	240	0.462766
Kichangani	Azimio	369	350	0.462766
Mtongani	Azimio	370	337	0.462766
Mbuyuni	Azimio	371	229	0.462766
Tambuka Reli	Azimio	372	265	0.462766
Azimio Kusini	Azimio	373	292	0.462766
Azimio Kaskazini	Azimio	374	257	0.462766

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mabatini	Tandika	375	276	0.797872
Maguruwe	Tandika	376	300	0.797872
Tamla	Tandika	377	307	0.797872
Tandika	Tandika	378	195	0.797872
Kilimahewa	Tandika	379	319	0.797872
Nyambwera	Tandika	380	288	0.797872
Kimbunga	Sandali	381	345	0.276596
Kisiwani	Sandali	382	239	0.276596
Mamboleo B	Sandali	383	285	0.276596
Usalama	Sandali	384	83	0.276596
Veterinary	Sandali	385	140	0.276596
Mkwinda	Sandali	386	311	0.276596
Sandali	Sandali	387	268	0.276596
Mwembeladu	Sandali	388	245	0.276596
Mwembemnofu	Sandali	389	200	0.276596
Mamboleo A	Sandali	390	258	0.276596
Tindwa	Sandali	391	283	0.276596
Mpogo	Sandali	392	336	0.276596
Chang'ombe A	Chang'ombe	393	60	0.81383
Toroli	Chang'ombe	394	112	0.81383
Bora	Chang'ombe	395	129	0.81383
Chang'ombe B	Chang'ombe	396	84	0.81383
Kizuiani	Mbagala Kuu	397	210	0.803191
Makuka	Mbagala Kuu	398	228	0.803191

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mbagala Kuu	Mbagala Kuu	399	152	0.803191
Kichemchem	Mbagala Kuu	400	230	0.803191
Kibondemaji 'B'	Mbagala Kuu	401	278	0.803191
Makangarawe	Makangarawe	402	339	0.324468
Uwazi	Makangarawe	403	348	0.324468
Dovya	Makangarawe	404	334	0.324468
Msakala	Makangarawe	405	400	0.324468
Kichangani	Pemba Mnazi	406	390	0.37234
Muhimbili	Pemba Mnazi	407	326	0.37234
Potea	Pemba Mnazi	408	429	0.37234
Kwa Morris	Pemba Mnazi	409	344	0.37234
Puna Centre	Pemba Mnazi	410	404	0.37234
Kibungo	Pemba Mnazi	411	421	0.37234
Chambewa	Pemba Mnazi	412	385	0.37234
Mahenge	Pemba Mnazi	413	443	0.37234
Buyuni	Pemba Mnazi	414	448	0.37234
Mti Mweupe	Pemba Mnazi	415	434	0.37234
Gulubwida	Pemba Mnazi	416	427	0.37234
Pemba Senta	Pemba Mnazi	417	391	0.37234
Songani Centre	Pemba Mnazi	418	392	0.37234
Nyange	Pemba Mnazi	419	405	0.37234
Tundwi Centre	Pemba Mnazi	420	395	0.37234
Maweni	Mjimwema	421	125	0.648936
Mjimwema	Mjimwema	422	78	0.648936

Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Kibugumo	Mjimwema	423	182	0.648936
Ungindoni	Mjimwema	424	148	0.648936
Magogoni	Tungi	425	149	0.398936
Tungi	Tungi	426	272	0.398936
Muungano	Tungi	427	219	0.398936
Misheni	Kijichi	428	238	0.718085
Mtoni Kijichi	Kijichi	429	116	0.718085
Mgeninani	Kijichi	430	185	0.718085
Butiama	Kijichi	431	176	0.718085
Mwanamtoti	Kijichi	432	249	0.718085
Majimatitu 'A'	Mianzini	433	413	0.505319
Machinjioni	Mianzini	434	426	0.505319
Mianzini	Mianzini	435	332	0.505319
Majimatitu	Mianzini	436	418	0.505319
Kibonde Maji 'A'	Mianzini	437	419	0.505319
Mchikichini	Mianzini	438	297	0.505319
Kimbangulile	Mianzini	439	384	0.505319
Mponda	Mianzini	440	359	0.505319
Barabara ya Mwinyi	Kiburugwa	441	365	0.43617
Kingungi	Kiburugwa	442	281	0.43617
Juhudi	Kiburugwa	443	320	0.43617
Kwa Nyoka	Kiburugwa	444	428	0.43617
Kiburugwa No.3	Kiburugwa	445	370	0.43617
Kiburugwa	Kiburugwa	446	351	0.43617

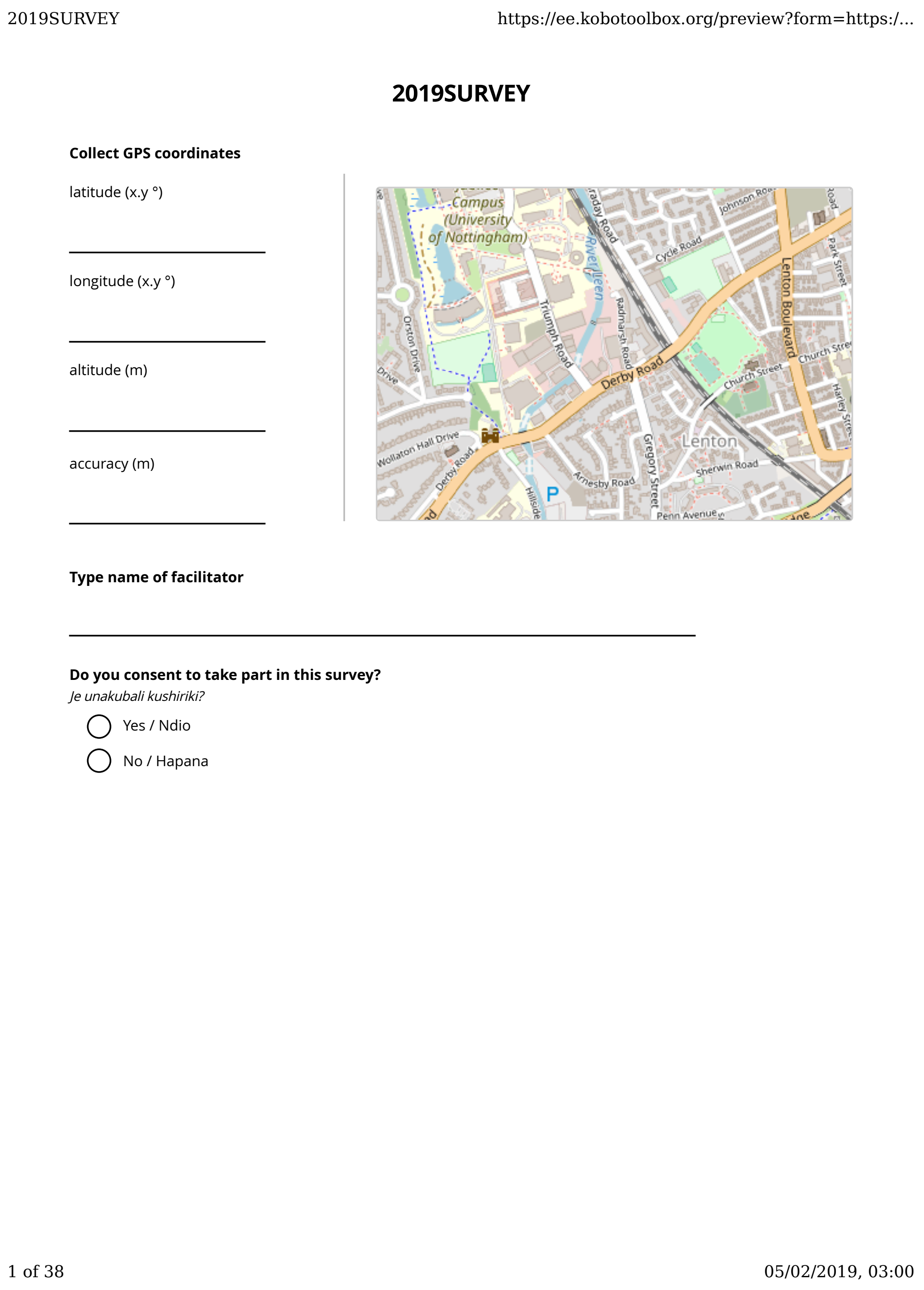
Continued on next page

Table C.1 – continued from previous page

subward.name	ward.name	subward.id	rank	prop.known
Mashine ya Maji	Buza	447	225	0.739362
Mjimpya	Buza	448	253	0.739362
Buza	Buza	449	295	0.739362
Kilakala	Kilakala	450	235	0.452128
Kigunga	Kilakala	451	247	0.452128
Barabara ya Mwinyi	Kilakala	452	157	0.452128

Appendix D

Street Survey



2019SURVEY

Collect GPS coordinates

latitude (x.y °)

______

longitude (x.y °)

______

altitude (m)

______

accuracy (m)

______

Campus (University of Nottingham)
River Leen
Triumph Road
Radmarsh Road
Cycle Road
Johnson Road
Lenton Boulevard
Park Street
Church Street
Harley Street
Derby Road
Orston Drive
Wollaton Hall Drive
Arnesby Road
Gregory Street
Lenton
Sherwin Road
Penn Avenue
Hillside
P

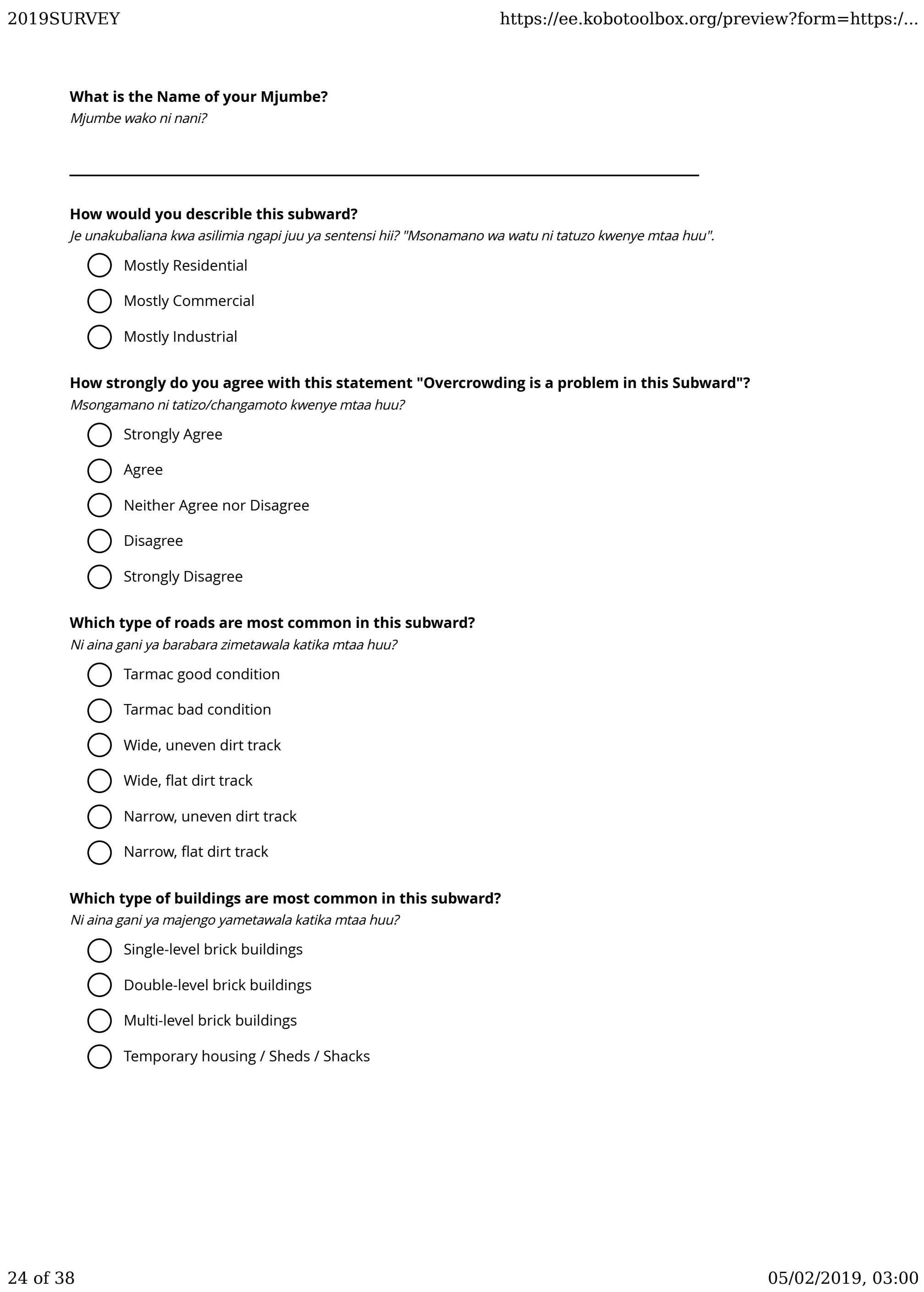
Type name of facilitator

______

Do you consent to take part in this survey?
Je unakubali kushiriki?

○ Yes / Ndio

○ No / Hapana



What is the Name of your Mjumbe?
Mjumbe wako ni nani?

____

How would you describe this subward?
Je unakubaliana kwa asilimia ngapi juu ya sentensi hii? "Msonamano wa watu ni tatuzo kwenye mtaa huu".

○ Mostly Residential

○ Mostly Commercial

○ Mostly Industrial

How strongly do you agree with this statement "Overcrowding is a problem in this Subward"?
Msongamano ni tatizo/changamoto kwenye mtaa huu?

○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

Which type of roads are most common in this subward?
Ni aina gani ya barabara zimetawala katika mtaa huu?

○ Tarmac good condition

○ Tarmac bad condition

○ Wide, uneven dirt track

○ Wide, flat dirt track

○ Narrow, uneven dirt track

○ Narrow, flat dirt track

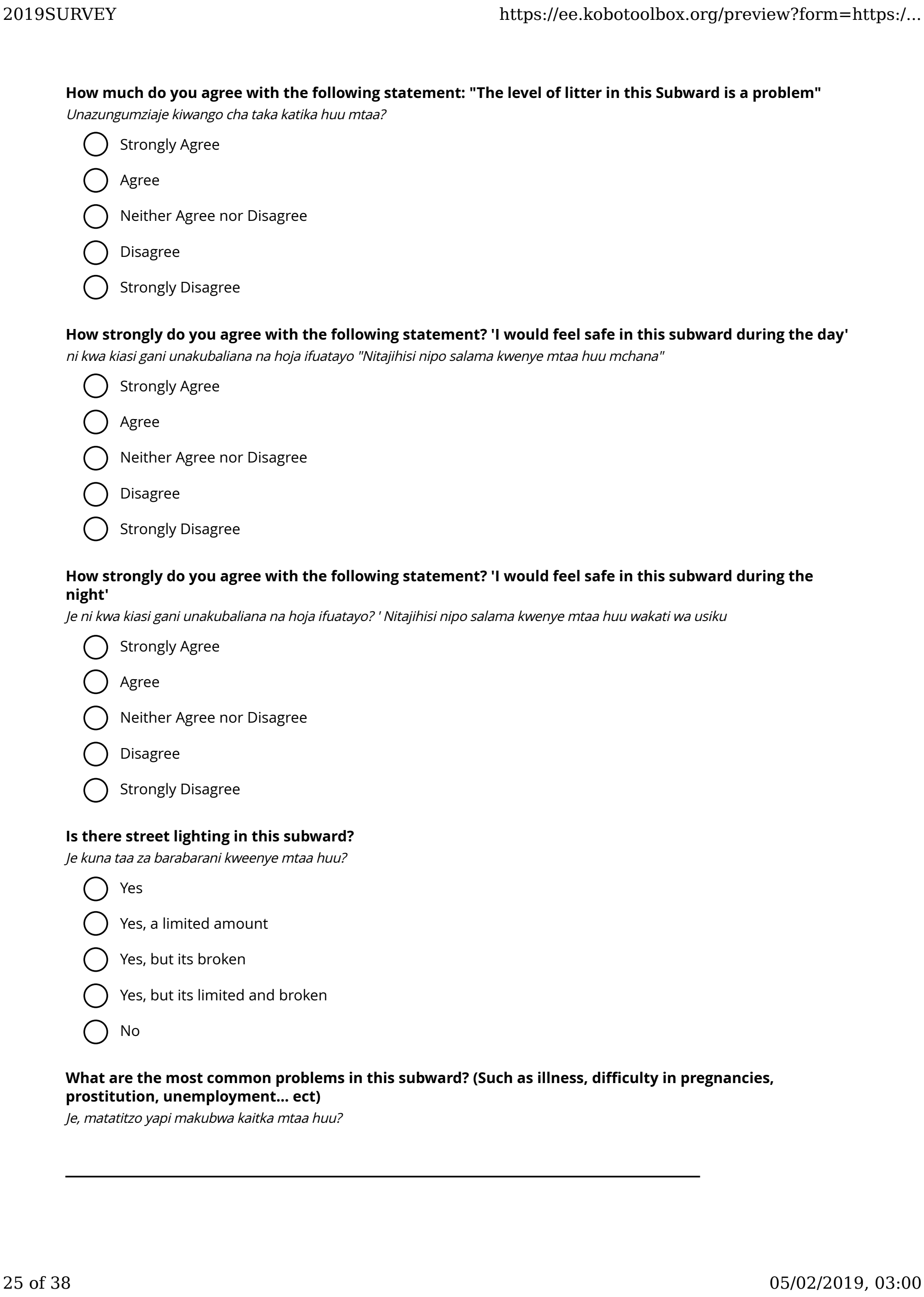
Which type of buildings are most common in this subward?
Ni aina gani ya majengo yametawala katika mtaa huu?

○ Single-level brick buildings

○ Double-level brick buildings

○ Multi-level brick buildings

○ Temporary housing / Sheds / Shacks



How much do you agree with the following statement: "The level of litter in this Subward is a problem"
Unazungumziaje kiwango cha taka katika huu mtaa?

○ Strongly Agree
○ Agree
○ Neither Agree nor Disagree
○ Disagree
○ Strongly Disagree

How strongly do you agree with the following statement? 'I would feel safe in this subward during the day'
ni kwa kiasi gani unakubaliana na hoja ifuatayo "Nitajihisi nipo salama kwenye mtaa huu mchana"

○ Strongly Agree
○ Agree
○ Neither Agree nor Disagree
○ Disagree
○ Strongly Disagree

How strongly do you agree with the following statement? 'I would feel safe in this subward during the night'
Je ni kwa kiasi gani unakubaliana na hoja ifuatayo? ' Nitajihisi nipo salama kwenye mtaa huu wakati wa usiku

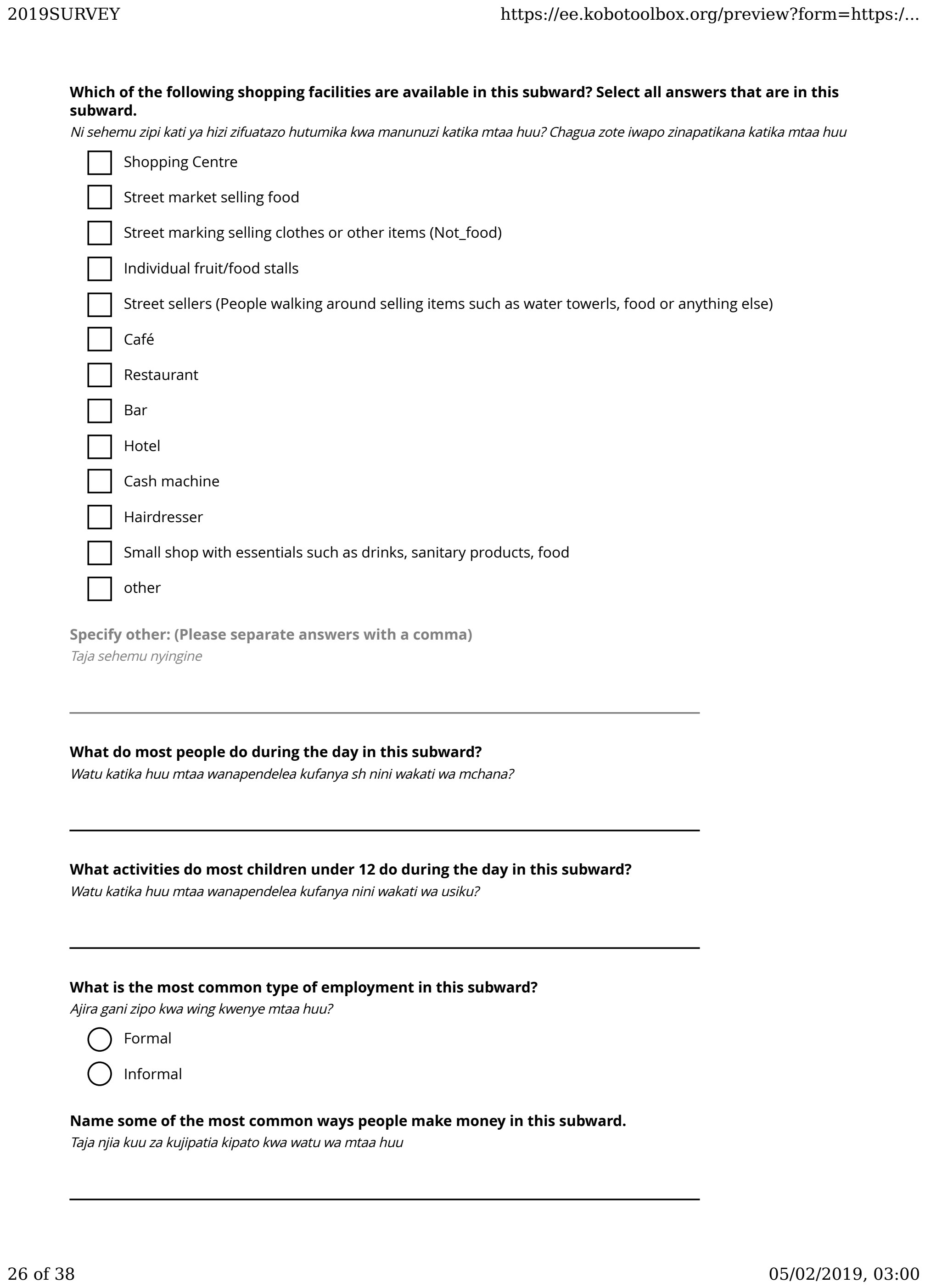
○ Strongly Agree
○ Agree
○ Neither Agree nor Disagree
○ Disagree
○ Strongly Disagree

Is there street lighting in this subward?
Je kuna taa za barabarani kweenye mtaa huu?

○ Yes
○ Yes, a limited amount
○ Yes, but its broken
○ Yes, but its limited and broken
○ No

What are the most common problems in this subward? (Such as illness, difficulty in pregnancies, prostitution, unemployment... ect)
Je, matatitzo yapi makubwa kaitka mtaa huu?

____



Which of the following shopping facilities are available in this subward? Select all answers that are in this subward.
Ni sehemu zipi kati ya hizi zifuatazo hutumika kwa manunuzi katika mtaa huu? Chagua zote iwapo zinapatikana katika mtaa huu

- □ Shopping Centre
- □ Street market selling food
- □ Street marking selling clothes or other items (Not_food)
- □ Individual fruit/food stalls
- □ Street sellers (People walking around selling items such as water towerls, food or anything else)
- □ Café
- □ Restaurant
- □ Bar
- □ Hotel
- □ Cash machine
- □ Hairdresser
- □ Small shop with essentials such as drinks, sanitary products, food
- □ other

Specify other: (Please separate answers with a comma)
Taja sehemu nyingine

____

What do most people do during the day in this subward?
Watu katika huu mtaa wanapendelea kufanya sh nini wakati wa mchana?

____

What activities do most children under 12 do during the day in this subward?
Watu katika huu mtaa wanapendelea kufanya nini wakati wa usiku?

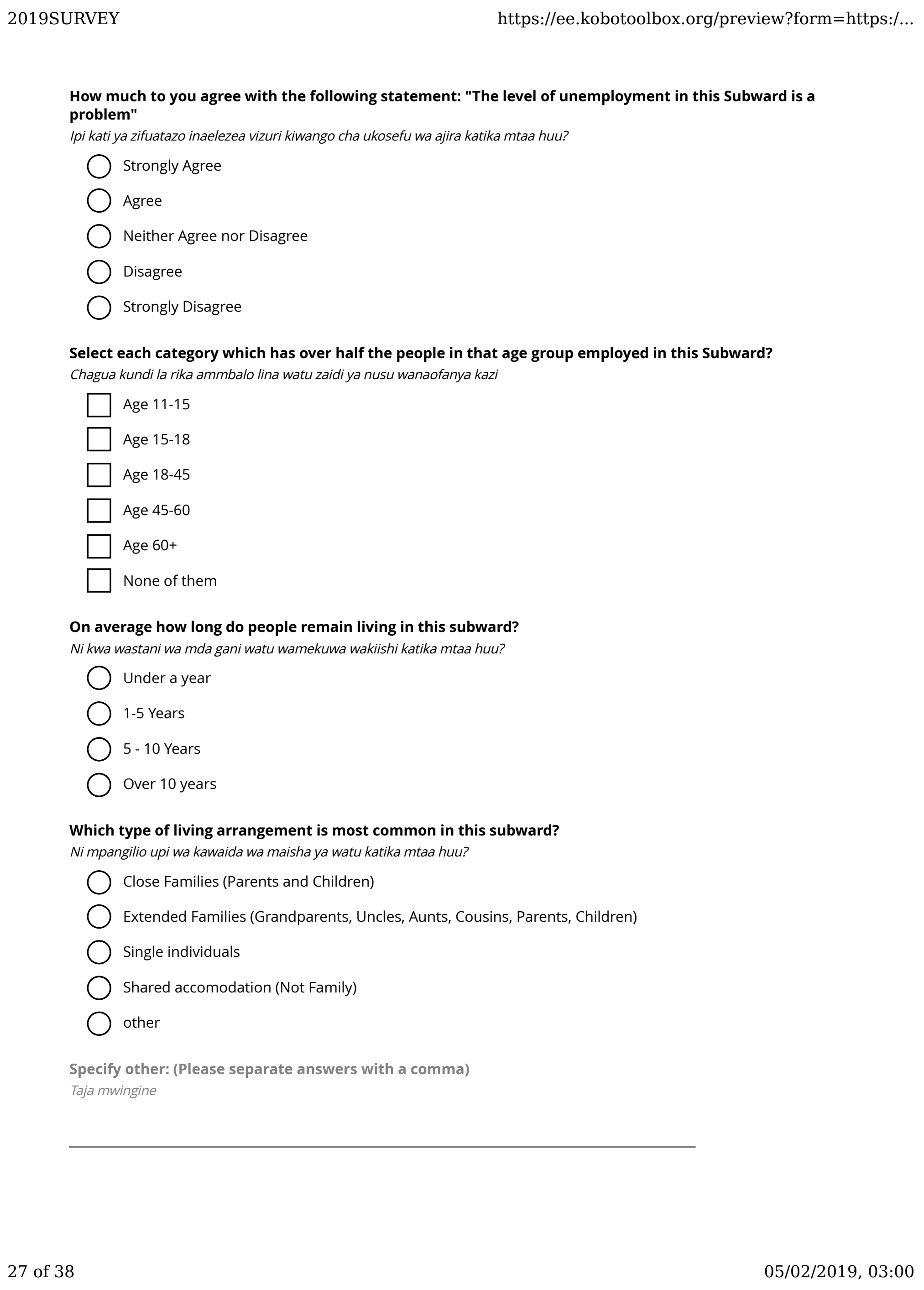
____

What is the most common type of employment in this subward?
Ajira gani zipo kwa wing kwenye mtaa huu?

- ○ Formal
- ○ Informal

Name some of the most common ways people make money in this subward.
Taja njia kuu za kujipatia kipato kwa watu wa mtaa huu

____



2019SURVEY https://ee.kobotoolbox.org/preview?form=https:/...

How much to you agree with the following statement: "The level of unemployment in this Subward is a problem"
Ipi kati ya zifuatazo inaelezea vizuri kiwango cha ukosefu wa ajira katika mtaa huu?

○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

Select each category which has over half the people in that age group employed in this Subward?
Chagua kundi la rika ammbalo lina watu zaidi ya nusu wanaofanya kazi

□ Age 11-15

□ Age 15-18

□ Age 18-45

□ Age 45-60

□ Age 60+

□ None of them

On average how long do people remain living in this subward?
Ni kwa wastani wa mda gani watu wamekuwa wakiishi katika mtaa huu?

○ Under a year

○ 1-5 Years

○ 5 - 10 Years

○ Over 10 years

Which type of living arrangement is most common in this subward?
Ni mpangilio upi wa kawaida wa maisha ya watu katika mtaa huu?

○ Close Families (Parents and Children)

○ Extended Families (Grandparents, Uncles, Aunts, Cousins, Parents, Children)

○ Single individuals

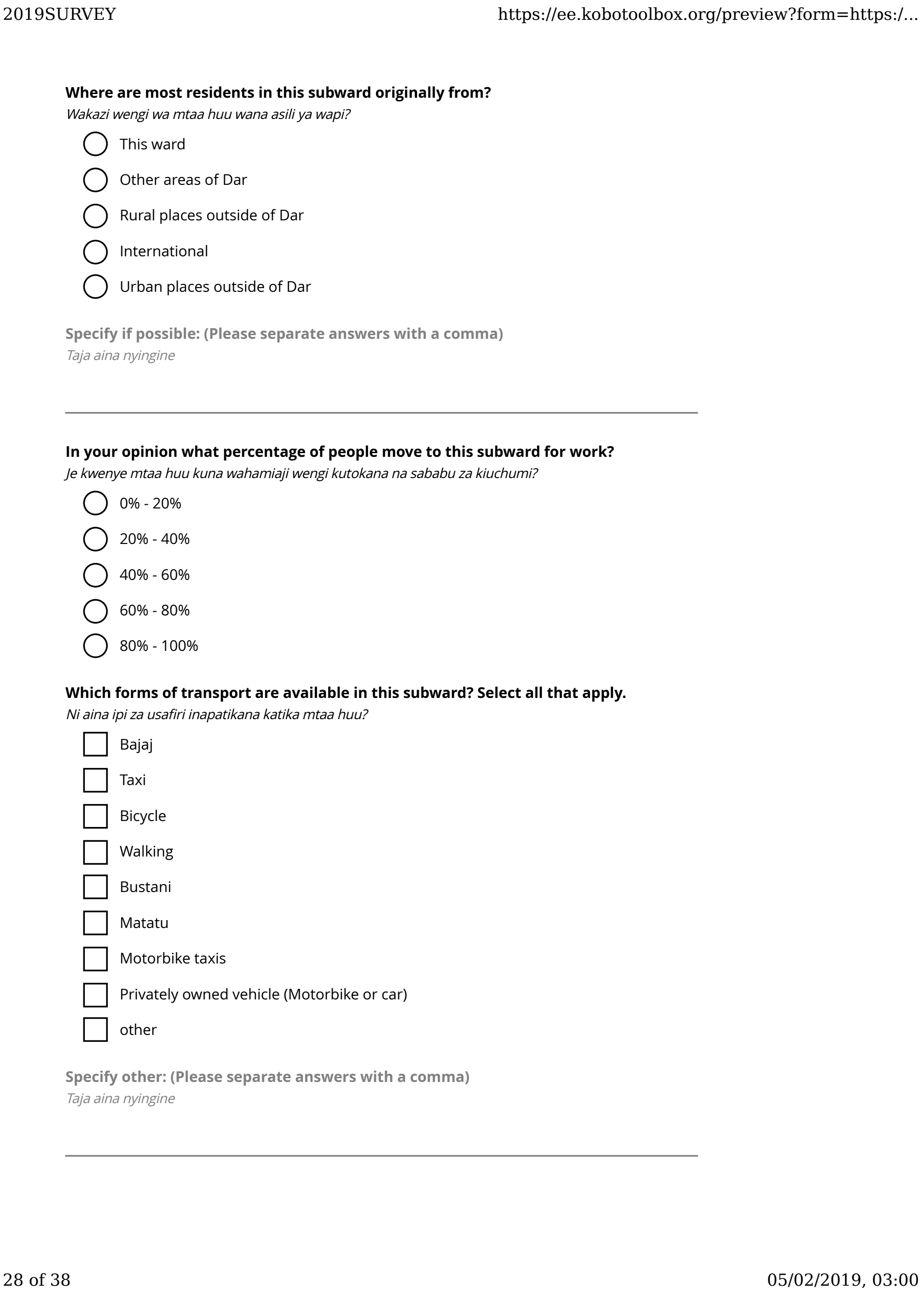
○ Shared accomodation (Not Family)

○ other

Specify other: (Please separate answers with a comma)
Taja mwingine

____

27 of 38 05/02/2019, 03:00



Where are most residents in this subward originally from?
Wakazi wengi wa mtaa huu wana asili ya wapi?

○ This ward
○ Other areas of Dar
○ Rural places outside of Dar
○ International
○ Urban places outside of Dar

Specify if possible: (Please separate answers with a comma)
Taja aina nyingine

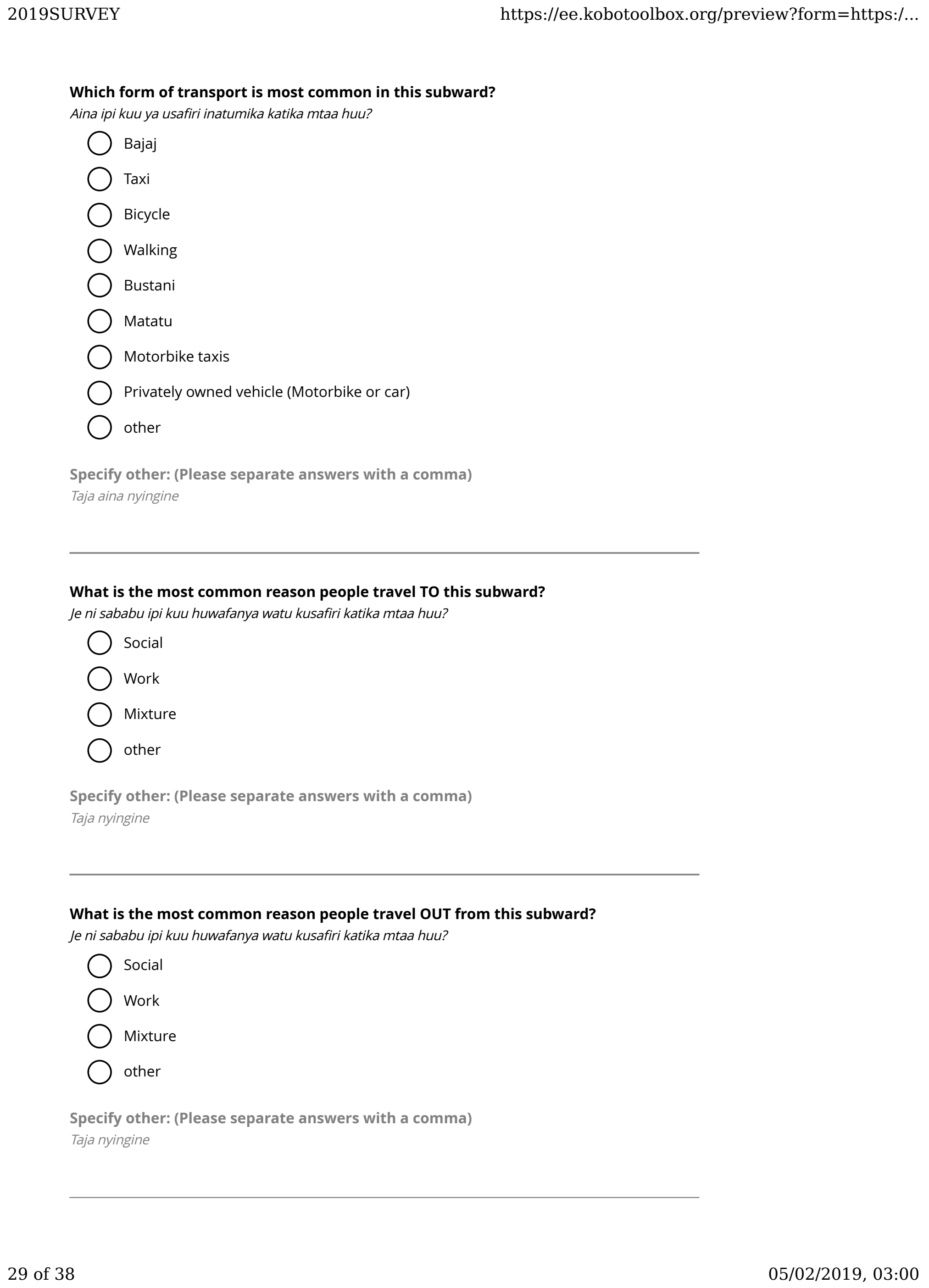
In your opinion what percentage of people move to this subward for work?
Je kwenye mtaa huu kuna wahamiaji wengi kutokana na sababu za kiuchumi?

○ 0% - 20%
○ 20% - 40%
○ 40% - 60%
○ 60% - 80%
○ 80% - 100%

Which forms of transport are available in this subward? Select all that apply.
Ni aina ipi za usafiri inapatikana katika mtaa huu?

□ Bajaj
□ Taxi
□ Bicycle
□ Walking
□ Bustani
□ Matatu
□ Motorbike taxis
□ Privately owned vehicle (Motorbike or car)
□ other

Specify other: (Please separate answers with a comma)
Taja aina nyingine



Which form of transport is most common in this subward?
Aina ipi kuu ya usafiri inatumika katika mtaa huu?

○ Bajaj

○ Taxi

○ Bicycle

○ Walking

○ Bustani

○ Matatu

○ Motorbike taxis

○ Privately owned vehicle (Motorbike or car)

○ other

Specify other: (Please separate answers with a comma)
Taja aina nyingine

What is the most common reason people travel TO this subward?
Je ni sababu ipi kuu huwafanya watu kusafiri katika mtaa huu?

○ Social

○ Work

○ Mixture

○ other

Specify other: (Please separate answers with a comma)
Taja nyingine

What is the most common reason people travel OUT from this subward?
Je ni sababu ipi kuu huwafanya watu kusafiri katika mtaa huu?

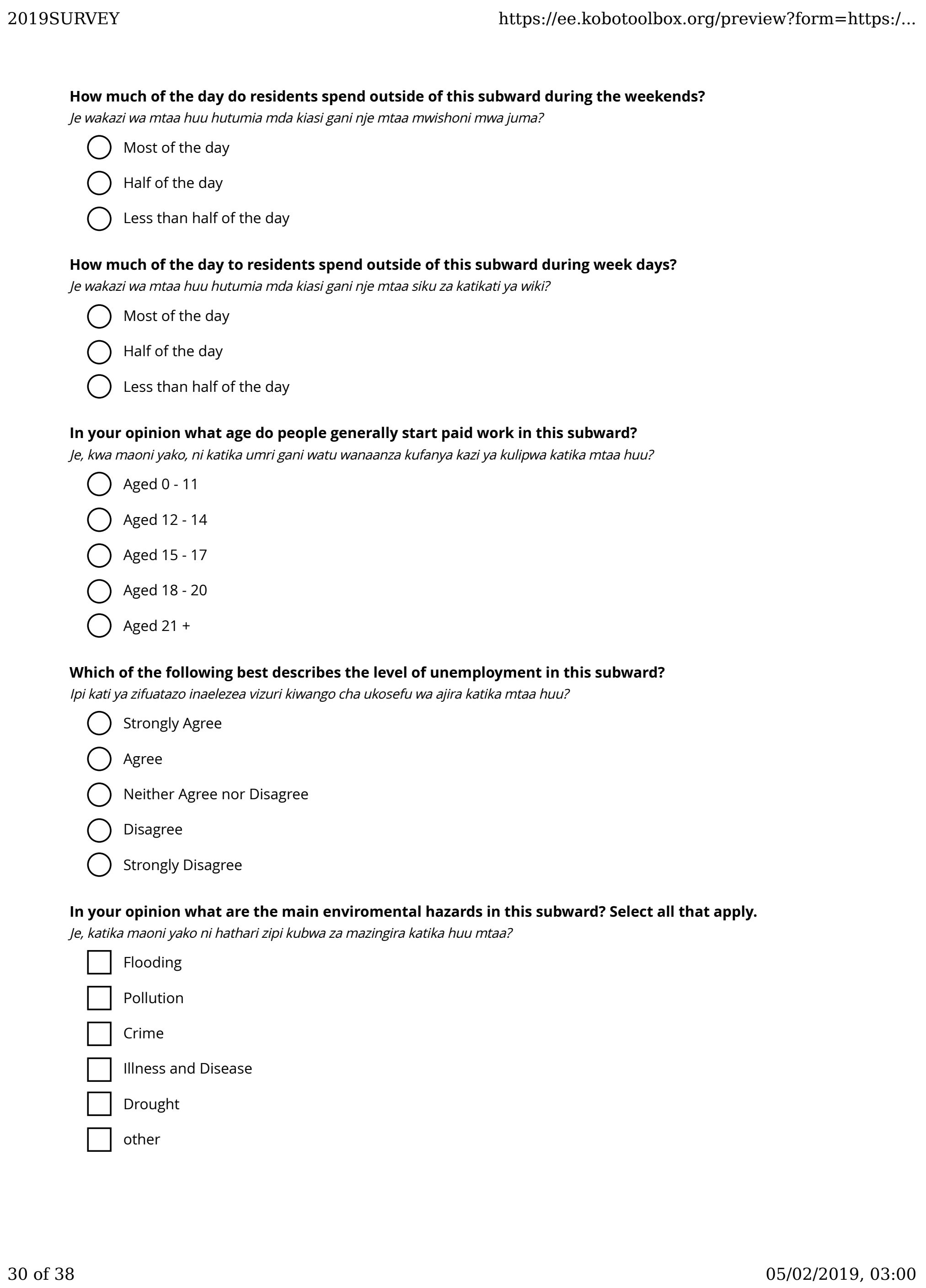
○ Social

○ Work

○ Mixture

○ other

Specify other: (Please separate answers with a comma)
Taja nyingine



How much of the day do residents spend outside of this subward during the weekends?
Je wakazi wa mtaa huu hutumia mda kiasi gani nje mtaa mwishoni mwa juma?

○ Most of the day
○ Half of the day
○ Less than half of the day

How much of the day to residents spend outside of this subward during week days?
Je wakazi wa mtaa huu hutumia mda kiasi gani nje mtaa siku za katikati ya wiki?

○ Most of the day
○ Half of the day
○ Less than half of the day

In your opinion what age do people generally start paid work in this subward?
Je, kwa maoni yako, ni katika umri gani watu wanaanza kufanya kazi ya kulipwa katika mtaa huu?

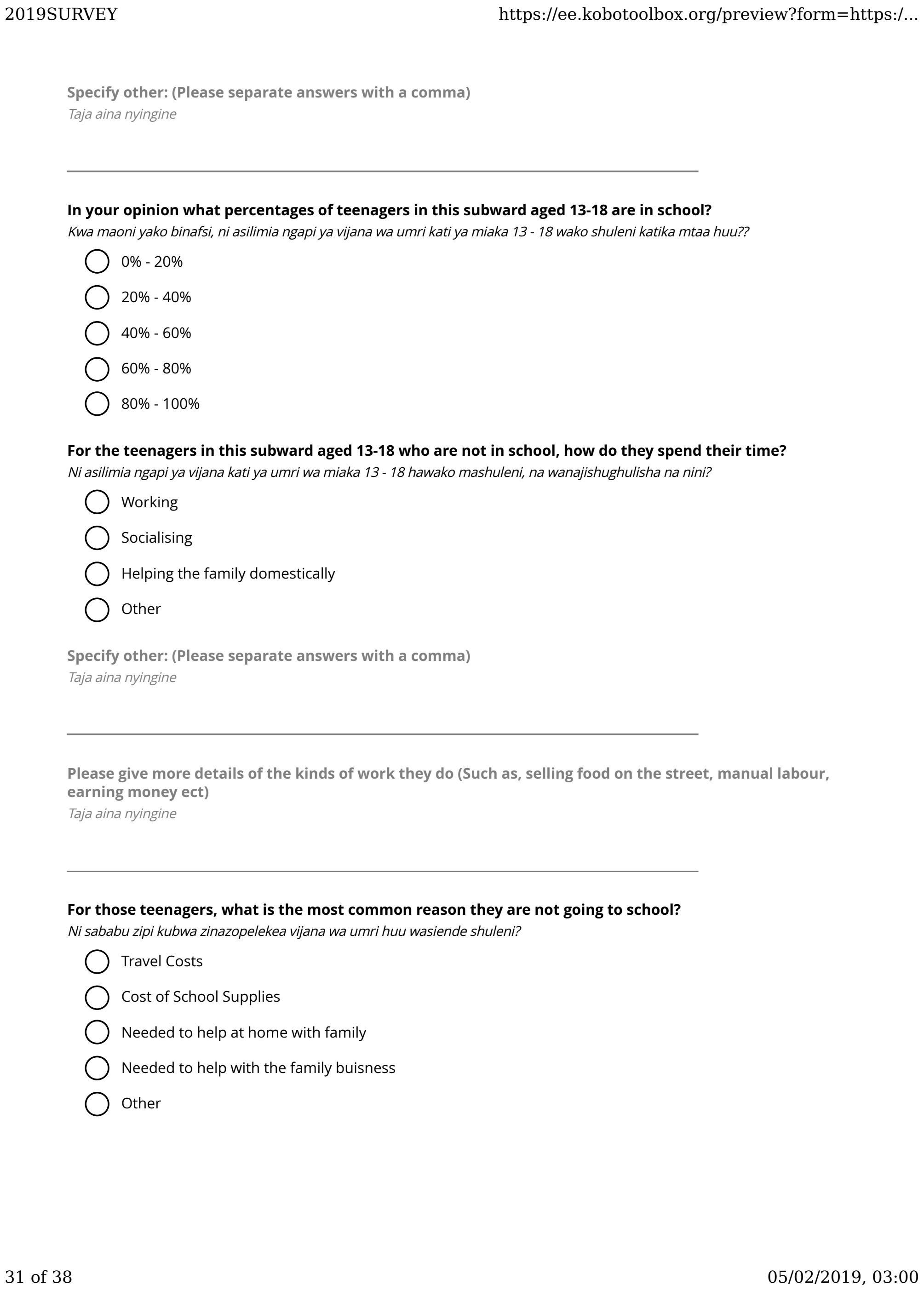
○ Aged 0 - 11
○ Aged 12 - 14
○ Aged 15 - 17
○ Aged 18 - 20
○ Aged 21 +

Which of the following best describes the level of unemployment in this subward?
Ipi kati ya zifuatazo inaelezea vizuri kiwango cha ukosefu wa ajira katika mtaa huu?

○ Strongly Agree
○ Agree
○ Neither Agree nor Disagree
○ Disagree
○ Strongly Disagree

In your opinion what are the main enviromental hazards in this subward? Select all that apply.
Je, katika maoni yako ni hathari zipi kubwa za mazingira katika huu mtaa?

□ Flooding
□ Pollution
□ Crime
□ Illness and Disease
□ Drought
□ other



Specify other: (Please separate answers with a comma)
Taja aina nyingine

____

In your opinion what percentages of teenagers in this subward aged 13-18 are in school?
Kwa maoni yako binafsi, ni asilimia ngapi ya vijana wa umri kati ya miaka 13 - 18 wako shuleni katika mtaa huu??

○ 0% - 20%
○ 20% - 40%
○ 40% - 60%
○ 60% - 80%
○ 80% - 100%

For the teenagers in this subward aged 13-18 who are not in school, how do they spend their time?
Ni asilimia ngapi ya vijana kati ya umri wa miaka 13 - 18 hawako mashuleni, na wanajishughulisha na nini?

○ Working
○ Socialising
○ Helping the family domestically
○ Other

Specify other: (Please separate answers with a comma)
Taja aina nyingine

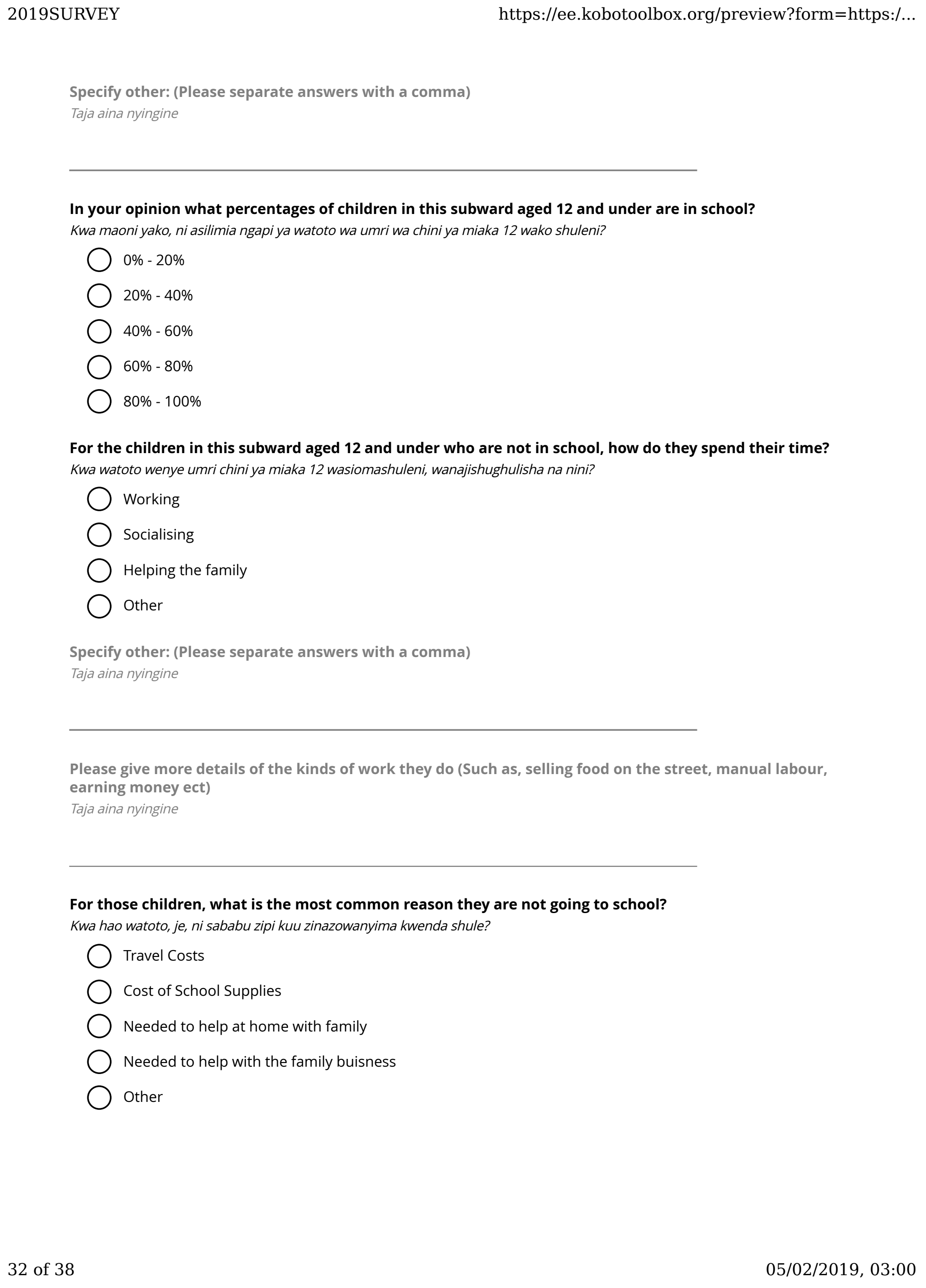
____

Please give more details of the kinds of work they do (Such as, selling food on the street, manual labour, earning money ect)
Taja aina nyingine

____

For those teenagers, what is the most common reason they are not going to school?
Ni sababu zipi kubwa zinazopelekea vijana wa umri huu wasiende shuleni?

○ Travel Costs
○ Cost of School Supplies
○ Needed to help at home with family
○ Needed to help with the family buisness
○ Other



Specify other: (Please separate answers with a comma)
Taja aina nyingine

____

In your opinion what percentages of children in this subward aged 12 and under are in school?
Kwa maoni yako, ni asilimia ngapi ya watoto wa umri wa chini ya miaka 12 wako shuleni?

○ 0% - 20%

○ 20% - 40%

○ 40% - 60%

○ 60% - 80%

○ 80% - 100%

For the children in this subward aged 12 and under who are not in school, how do they spend their time?
Kwa watoto wenye umri chini ya miaka 12 wasiomashuleni, wanajishughulisha na nini?

○ Working

○ Socialising

○ Helping the family

○ Other

Specify other: (Please separate answers with a comma)
Taja aina nyingine

____

Please give more details of the kinds of work they do (Such as, selling food on the street, manual labour, earning money ect)
Taja aina nyingine

____

For those children, what is the most common reason they are not going to school?
Kwa hao watoto, je, ni sababu zipi kuu zinazowanyima kwenda shule?

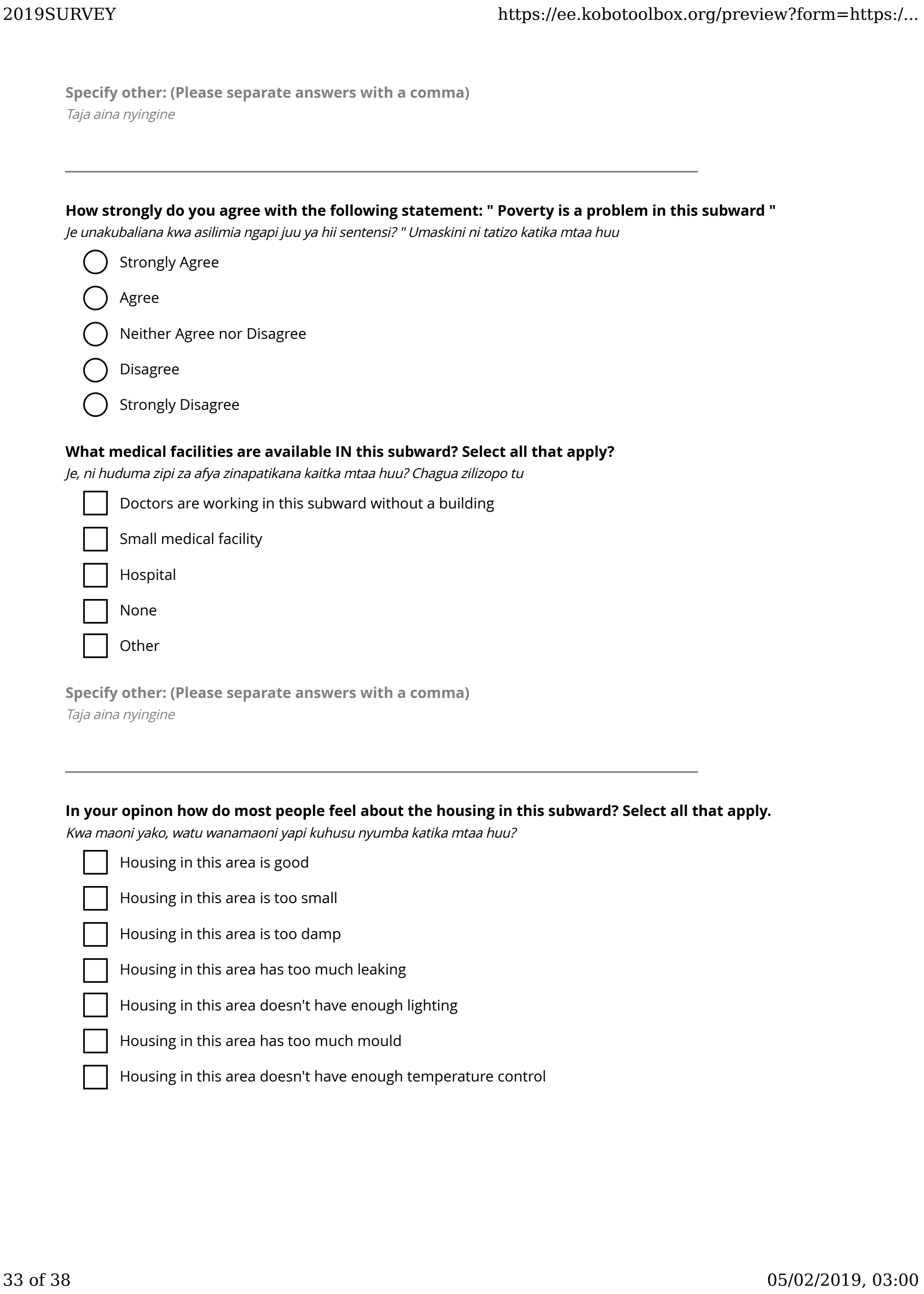
○ Travel Costs

○ Cost of School Supplies

○ Needed to help at home with family

○ Needed to help with the family buisness

○ Other



Specify other: (Please separate answers with a comma)
Taja aina nyingine

____

How strongly do you agree with the following statement: " Poverty is a problem in this subward "
Je unakubaliana kwa asilimia ngapi juu ya hii sentensi? " Umaskini ni tatizo katika mtaa huu

○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

What medical facilities are available IN this subward? Select all that apply?
Je, ni huduma zipi za afya zinapatikana kaitka mtaa huu? Chagua zilizopo tu

□ Doctors are working in this subward without a building

□ Small medical facility

□ Hospital

□ None

□ Other

Specify other: (Please separate answers with a comma)
Taja aina nyingine

____

In your opinon how do most people feel about the housing in this subward? Select all that apply.
Kwa maoni yako, watu wanamaoni yapi kuhusu nyumba katika mtaa huu?

□ Housing in this area is good

□ Housing in this area is too small

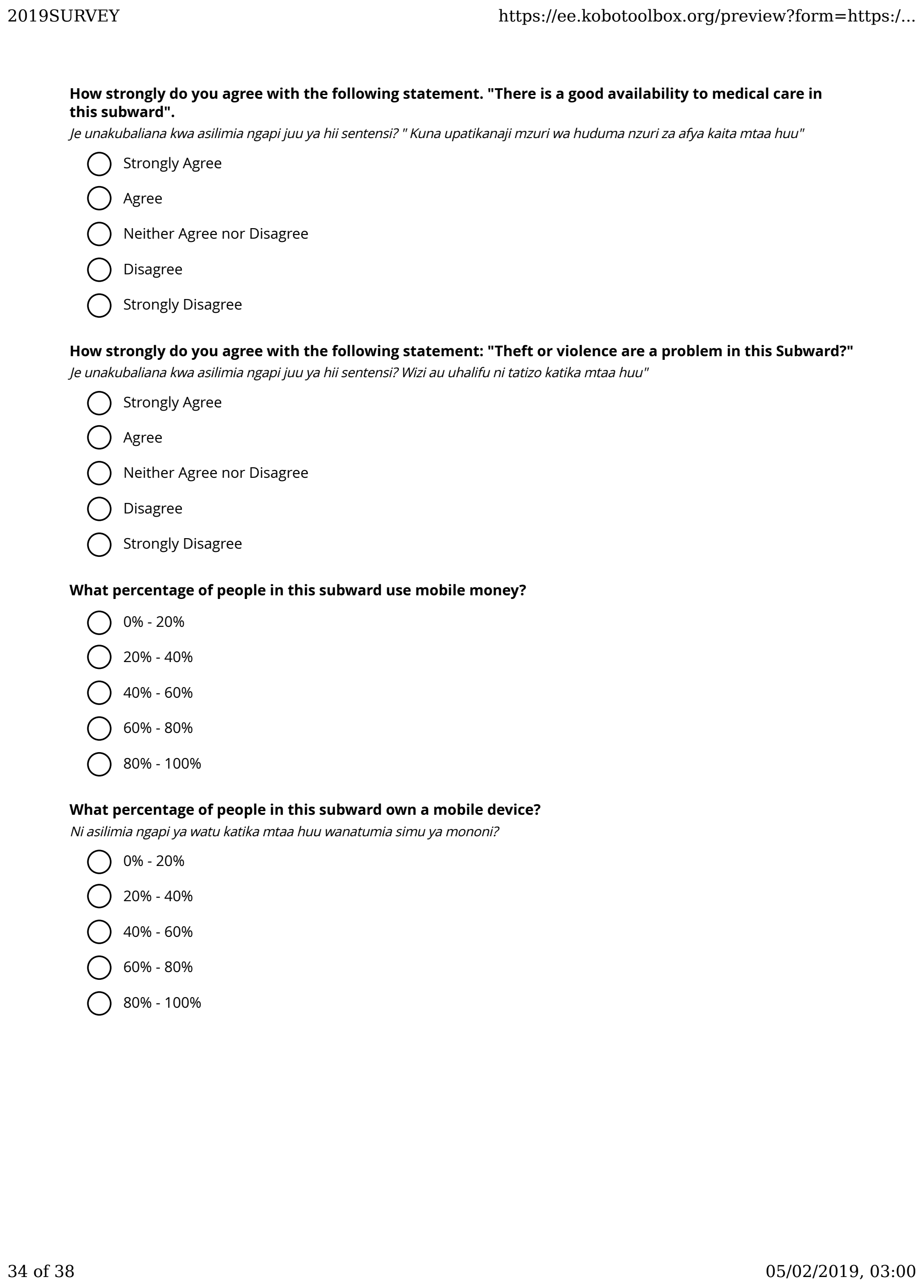
□ Housing in this area is too damp

□ Housing in this area has too much leaking

□ Housing in this area doesn't have enough lighting

□ Housing in this area has too much mould

□ Housing in this area doesn't have enough temperature control



2019SURVEY https://ee.kobotoolbox.org/preview?form=https:/...

How strongly do you agree with the following statement. "There is a good availability to medical care in this subward".

Je unakubaliana kwa asilimia ngapi juu ya hii sentensi? " Kuna upatikanaji mzuri wa huduma nzuri za afya kaita mtaa huu"

○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

How strongly do you agree with the following statement: "Theft or violence are a problem in this Subward?"

Je unakubaliana kwa asilimia ngapi juu ya hii sentensi? Wizi au uhalifu ni tatizo katika mtaa huu"

○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

What percentage of people in this subward use mobile money?

○ 0% - 20%

○ 20% - 40%

○ 40% - 60%

○ 60% - 80%

○ 80% - 100%

What percentage of people in this subward own a mobile device?

Ni asilimia ngapi ya watu katika mtaa huu wanatumia simu ya mononi?

○ 0% - 20%

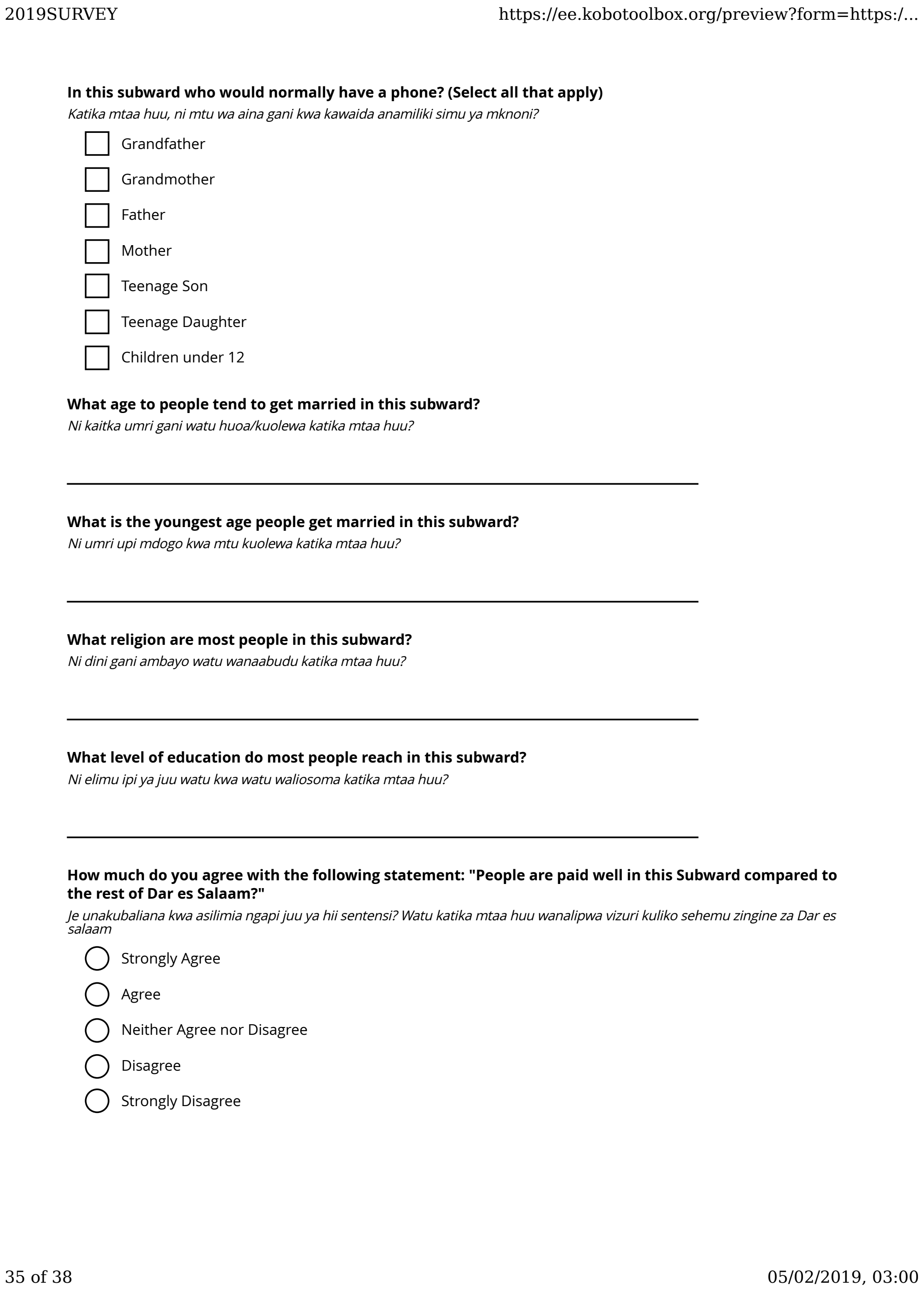
○ 20% - 40%

○ 40% - 60%

○ 60% - 80%

○ 80% - 100%

34 of 38 05/02/2019, 03:00



2019SURVEY https://ee.kobotoolbox.org/preview?form=https:/...

In this subward who would normally have a phone? (Select all that apply)
Katika mtaa huu, ni mtu wa aina gani kwa kawaida anamiliki simu ya mknoni?

□ Grandfather

□ Grandmother

□ Father

□ Mother

□ Teenage Son

□ Teenage Daughter

□ Children under 12

What age to people tend to get married in this subward?
Ni kaitka umri gani watu huoa/kuolewa katika mtaa huu?

____

What is the youngest age people get married in this subward?
Ni umri upi mdogo kwa mtu kuolewa katika mtaa huu?

____

What religion are most people in this subward?
Ni dini gani ambayo watu wanaabudu katika mtaa huu?

____

What level of education do most people reach in this subward?
Ni elimu ipi ya juu watu kwa watu waliosoma katika mtaa huu?

____

How much do you agree with the following statement: "People are paid well in this Subward compared to the rest of Dar es Salaam?"
Je unakubaliana kwa asilimia ngapi juu ya hii sentensi? Watu katika mtaa huu wanalipwa vizuri kuliko sehemu zingine za Dar es salaam

○ Strongly Agree

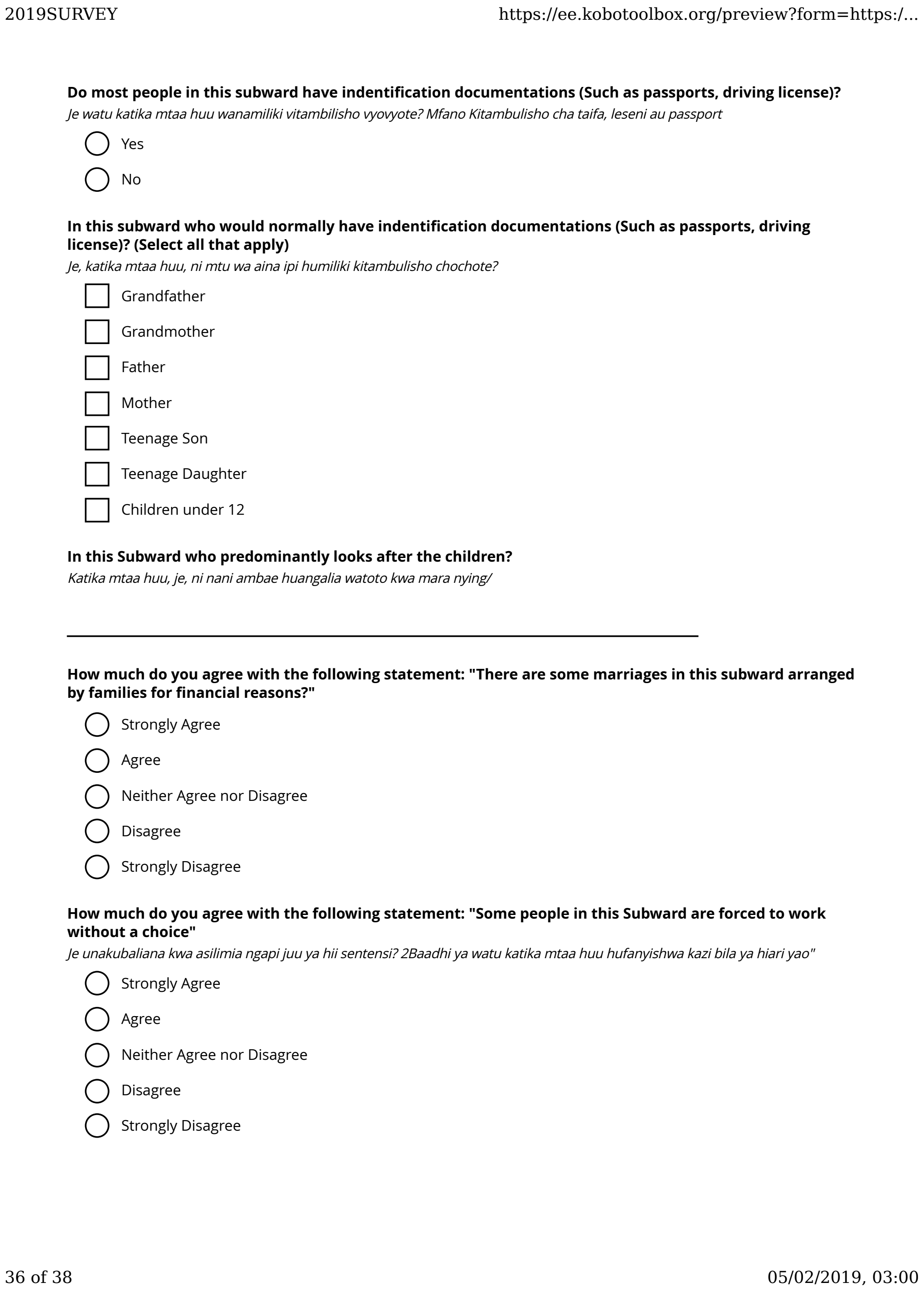
○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

35 of 38 05/02/2019, 03:00



Do most people in this subward have indentification documentations (Such as passports, driving license)?

Je watu katika mtaa huu wanamiliki vitambilisho vyovyote? Mfano Kitambulisho cha taifa, leseni au passport

○ Yes

○ No

In this subward who would normally have indentification documentations (Such as passports, driving license)? (Select all that apply)

Je, katika mtaa huu, ni mtu wa aina ipi humiliki kitambulisho chochote?

□ Grandfather

□ Grandmother

□ Father

□ Mother

□ Teenage Son

□ Teenage Daughter

□ Children under 12

In this Subward who predominantly looks after the children?

Katika mtaa huu, je, ni nani ambae huangalia watoto kwa mara nying/

____

How much do you agree with the following statement: "There are some marriages in this subward arranged by families for financial reasons?"

○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree

How much do you agree with the following statement: "Some people in this Subward are forced to work without a choice"

Je unakubaliana kwa asilimia ngapi juu ya hii sentensi? 2Baadhi ya watu katika mtaa huu hufanyishwa kazi bila ya hiari yao"

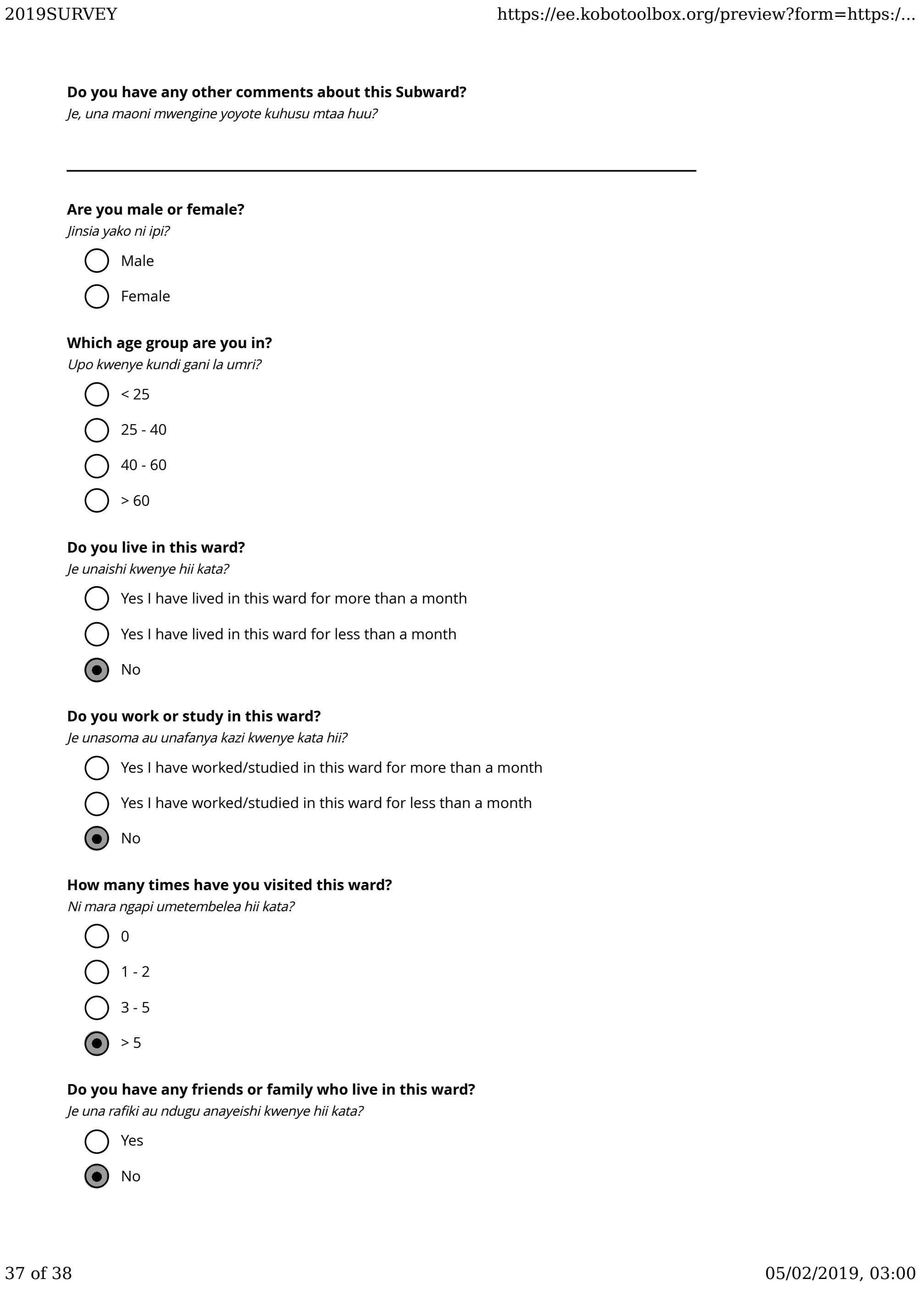
○ Strongly Agree

○ Agree

○ Neither Agree nor Disagree

○ Disagree

○ Strongly Disagree



2019SURVEY https://ee.kobotoolbox.org/preview?form=https:/...

Do you have any other comments about this Subward?
Je, una maoni mwengine yoyote kuhusu mtaa huu?

____

Are you male or female?
Jinsia yako ni ipi?

○ Male

○ Female

Which age group are you in?
Upo kwenye kundi gani la umri?

○ < 25

○ 25 - 40

○ 40 - 60

○ > 60

Do you live in this ward?
Je unaishi kwenye hii kata?

○ Yes I have lived in this ward for more than a month

○ Yes I have lived in this ward for less than a month

● No

Do you work or study in this ward?
Je unasoma au unafanya kazi kwenye kata hii?

○ Yes I have worked/studied in this ward for more than a month

○ Yes I have worked/studied in this ward for less than a month

● No

How many times have you visited this ward?
Ni mara ngapi umetembelea hii kata?

○ 0

○ 1 - 2

○ 3 - 5

● > 5

Do you have any friends or family who live in this ward?
Je una rafiki au ndugu anayeishi kwenye hii kata?

○ Yes

● No

37 of 38 05/02/2019, 03:00

2019SURVEY https://ee.kobotoolbox.org/preview?form=https:/...

Please take a photo of what this Subward looks like.
Tafadhali piga picha ya kuonesha uhalisia wa mtaa huu

Click here to upload file. (< 5MB)

38 of 38 05/02/2019, 03:00

Appendix E

Machine Learning vs Proxy Tables

1

Table 1: Parameters for model grid searches:

Parameter Name	Parameters
General Parameters	
Random State	17
strategy	stratified
Ridge	
alpha	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Logistic	
penalty	l1, l2
C	100, 10, 1.0, 0.1, 0.001
KNN	
Number Neighbors	10 integers between 1 and 22
weights	uniform, distance
metric	euclidean, manhattan
SVM	
kernel	poly, rbf, sigmoid
gamma	scale
Tree/Random Forest	
number estimators	5, 10, 20, 30, 50, 75, 100, 150, 250, 500
max features	auto, sqrt
min samples split	2 ,5 , 10
min samples leaf	1 ,2, 4
bootstrap	True, False
Boosting	
learning rate	0.001, 0.01, 0.1
sub sample	0.5 , 0.7 , 1.0
MLP	
Hidden Layer Sizes	(50, 50, 50), (50, 100, 50), (100,)
Activation	tanh, relu, logistic
Solver	sgd, adam, lbfgs
alpha	0.0001, 0.05
learning rate	constant, adaptive

Table 2: Features in each Candidate Input Dataset:

Feature	**Source**	**Description**
SURVEY Feature Set		
Overcrowding	Survey	Question: How strongly do you agree with this statement Overcrowding is a problem in this subward? Answer: 5 Scale Likert
Road Quality	Survey	Question: Which type of roads are most common in this subward? Answer: 6 suggested road qualities
Building Quality	Survey	Question: Which type of buildings are most common in this subward? Answer: 4 suggested road qualities
Litter	Survey	Question: How strongly do you agree with the following statement? The level of litter in this subward is a problem? Answer: 5 Scale Likert
Day Safety	Survey	Question: How strongly do you agree with the following statement? I would feel safe in this subward during the day Answer: 5 Scale Likert
Night Safety	Survey	Question: How strongly do you agree with the following statement? I would feel safe in this subward during the night Answer: 5 Scale Likert
Street Lighting	Survey	Question: Is there street lighting in this subward? Answer: 5 suggested conditions
Unemployment1	Survey	Question: How strongly do you agree with the following statement? The level of unemployment in this Subward is a problem Answer: 5 Scale Likert
Remain time	Survey	Question: On average how long do people remain living in this subward? Answer: 4 suggested times
Economic Movement	Survey	Question: In your opinion what percentage of people move to this subward for work? Answer: 5 suggested conditions
Weekend	Survey	Question: How much of the day to residents spend outside of this subward during the weekends? Answer: 3 suggested conditions
Week	Survey	Question: How much of the day to residents spend outside of this subward during the week? Answer: 3 suggested conditions
Age Work	Survey	Question: In your opinion what age to people generally start paid work in this subward? Answer: 5 suggested conditions
Unemployment2	Survey	Question: How strongly do you agree with the following statement: The level of unemployment in this subward is high compared to the rest of Dar? Answer: 5 Scale Likert
Poverty	Survey	Question: How strongly do you agree with the following statement: Poverty is a problem in this subward compared to the rest of Dar? Answer: 5 Scale Likert
Medical	Survey	Question: How strongly do you agree with the following statement: There is a good availability to medical care in this subward? Answer: 5 Scale Likert
Theft and Violence	Survey	Question: How strongly do you agree with the following statement: Theft of violence are a problem in this Subward? Answer: 5 Scale Likert
Mobile Money Usage	Survey	Question: What percentage of people in this subward use mobile money? Answer: 5 suggested conditions
Mobile Usage	Survey	Question: What percentage of people in this own a mobile device? Answer: 5 suggested conditions
Pay	Survey	Question: How strongly do you agree with the following statement: People are paid well in this Subward compared to the rest of Dar es Salaam? Answer: 5 Scale Likert
Child School	Survey	Question: In your opinion, what percentages of children in this subward aged 12 and under are in school? Answer: 5 suggested conditions
Teenage School	Survey	Question: In your opinion, what percentages of teenagers in this subward aged 13-19 are in school? Answer: 5 suggested conditions
Landuse	Survey	Question: How would you describe this subward? Answer: 3 suggested conditions
Work Formality	Survey	Question: What is the most common type of employment in this subward? Answer: 2 suggested conditions
Living Arrangement	Survey	Question: Which type of living arrangement is most common in this subward? Answer: 4 suggested conditions
Originate	Survey	Question: Where are most resident in this subward originally from? Answer: 5 suggested conditions
Most Common Transport	Survey	Question: Which form of transport is most common in this subward? Answer: 6 suggested conditions
Travel in	Survey	Question: What is the most common reason people travel TO this subward? Answer: 3 suggested conditions
Travel out	Survey	Question: What is the most common reason people travel OUT of this subward? Answer: 3 suggested conditions
Activities Children	Survey	Question: For the children in this subward aged 12 and under who are not in school, how do they spend their time? Answer: 3 suggested conditions

Table 3: Features in each Candidate Input Dataset Continued:

Feature	Source	Description
SURVEY Feature Set continued		
Activities Teens	Survey	Question: For the teenagers in this subward aged 13-18 who are not in school, how do they spend their time? Answer: 3 suggested conditions
Reasons Children	Survey	Question: For the children in this subward aged 12 and under, what is the most common reason they are not going to school? Answer: 5 suggested conditions
Reasons Teenagers	Survey	Question: For those teenagers, what is the most common reason they are not going to school? Answer: 5 suggested conditions
Identity Documentation	Survey	Question: Do most people in this subward have identification documentations (Such as passports, driving license)? Answer: Yes or No
Who Identity	Survey	Question: In this subward who would normally have identification documentations? Answer: 7 suggested conditions
Who Phone	Survey	Question: In this subward who would normally have a phone? Answer: 7 suggested conditions
Medical Facilities	Survey	Question: What medical facilities are available in this subward? Answer: 4 suggested conditions
Environmental Hazards	Survey	Question: In your opinion what are the main environmental hazards in this subward? Answer: 5 suggested conditions
Age Employed	Survey	Question: Select each category which has over half the people in that age group employed in this subward? Answer: 5 suggested conditions
Forced Labour	Survey	Question: How strongly do you agree with the following statement: There are people in this subward being forced to work against their will. Answer: 5 Scale Likert
MOBILE Feature Set		
Number of CDR users	CDR	Number of CDR users in a subward
Number of MFS users	MFS	Number of MFS users in a subward
Frequent BTS	CDR	Number of BTS that account for 80% of a users network events
Activity at ToD	CDR	Level of call and sms activity during the day (7:00-18:00), evening (18:00-23:00) and night (23:00-7:00)
Resident density	CDR	Number of CDR users in a subward relative to its area size
Individual daily distance	CDR	Average distance travelled by subward residents
Average trip distance	CDR	Average length of an individual trip, a part of the overall daily distance
Interaction distance	CDR	Average distance of inbound/outbound network events of residents
Hybrid in/out weekday	CDR	Number of trips into and out of a respective subward
CDR events ward	CDR	Number of call and sms events in a subward during the study period
Network event density	CDR	Number of CDR events in a subward relative to its area size
Income TZS	MFS	Overall influx of mobile money into a subward
Spending uptake	MFS	The total spend in an area divided by local MFS uptake (number of MFS users divided by the number of CDR users in an area).
Percent low/med/high income	MFS	Percent of residents in an area with low (ntile 1-3), medium (ntile 4-7) or high income (ntile 8)
Overall TZS P2P	MFS	Average amount of P2P transfers for a subward resident
GEOSPATIAL Feature Sets		
Area	Official	Area of subward as calculated from official subward shape files
Area Urban	Drone	Area of subward designated with ‘urban’ pixels by World Bank
Area Industrial	Drone	Area of subward designated with ‘industrial’ pixels
Area Residential	Drone	Area designated with ‘formal’ or ‘informal’ residential pixels
Area Slum	Drone	Area designated with ‘slum’ pixels
Area Unused	Drone	Area designated with ‘barren’, ‘water’ or ‘vegetation’ pixels
Percentage Urban	Area	Percentage of subward pixels designated as ‘urban’ by World Bank
Percent Industrial	Drone	Percentage designated with ‘industrial’ pixels
Percent Residential	Drone	Percentage designated with ‘formal’ or ‘informal’ residential pixels
Percent Slum	Drone	Percentage designated with ‘slum’ pixels
Percent Unused	Drone	Percentage designated with ‘barren’, ‘water’ or ‘vegetation’ pixels
Slum/Industry Ratio	Drone	Area Slum / Area Industry
Residential/Industry Ratio	Drone	Area Residential / Area Industry
District	Official	The district subward belongs to.
Coastal	Drone	Designation if the subward is a coastal zone
Port	Drone	Designation if the subward is part of Dar es Salaam Port
Harbour	Drone	Designation if the subward is part of the inland waterway
Distance to CBD	Drone	Distance of subward centroid to the ‘Posta’ area of Dar
Distance to slum	Drone	Distance to any subward with greater than 50% slums
Distance to industry	Drone	Distance to any subward with greater than 50% slums
Distance to coast	Drone	Distance of the subward centroid to the coastline
Distance to port	Drone	Distance of the subward centroid to any port subward
Distance to harbour	Drone	Description

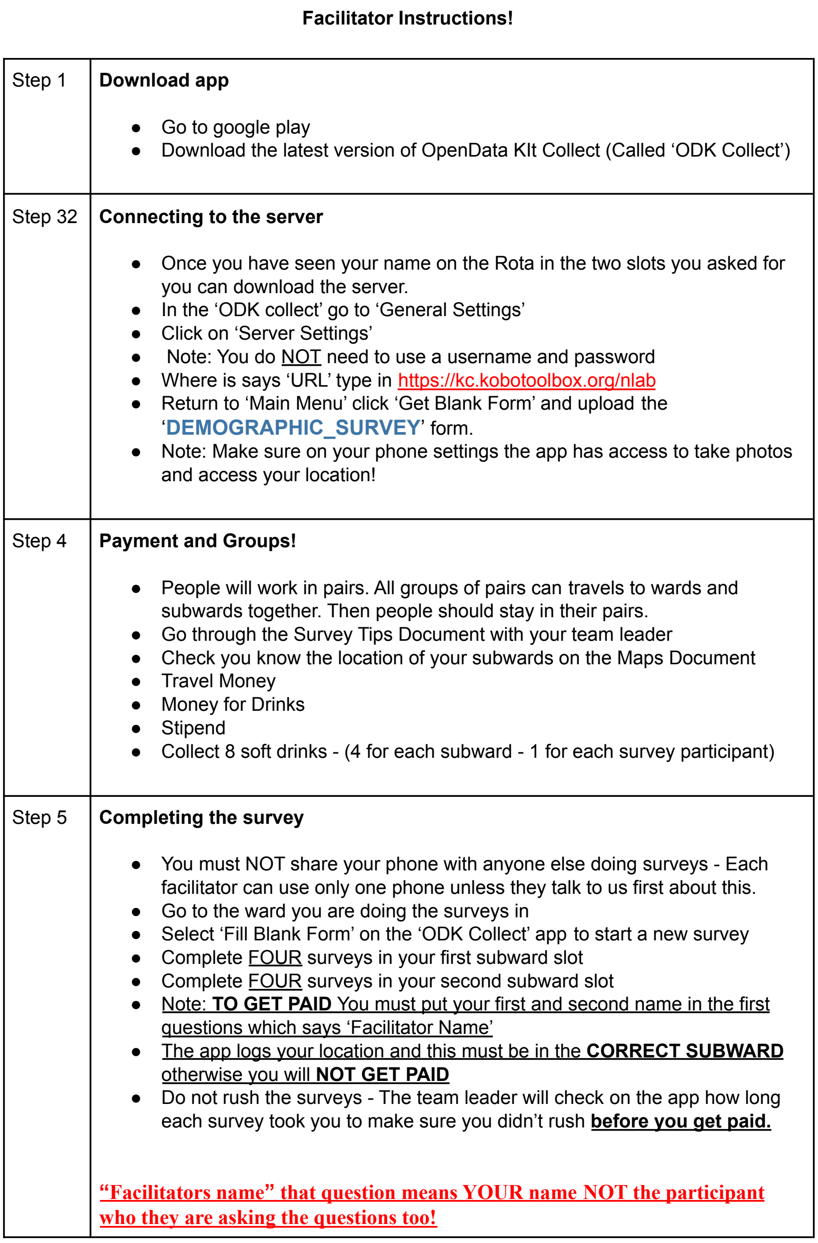
4

Table 4: Features in Theory Driven Proxy Model:

Feature	Source
Theory Driven Proxy Feature Set	
Living Conditions	
Overcrowding, Road / Building Quality	Survey
Resident Density	CDR Data
Slum Area Coverage	Landuse Data
Violence	
Theft and Violence Rates	Survey
Isolation and Restriction	
Amount of mobile money Users	Mobile Money Data
Amount of mobile phone Users	CDR Data
Distance from Central Business District	Landuse Data
Mobile / Mobile Money Usage, ID Overnership	Survey
Education	
% of Teenagers / Children in School	Survey
Employment and Finances	
Income	Mobile Money Data
Unemployment	Survey
Distance from Central Business District	Landuse Data

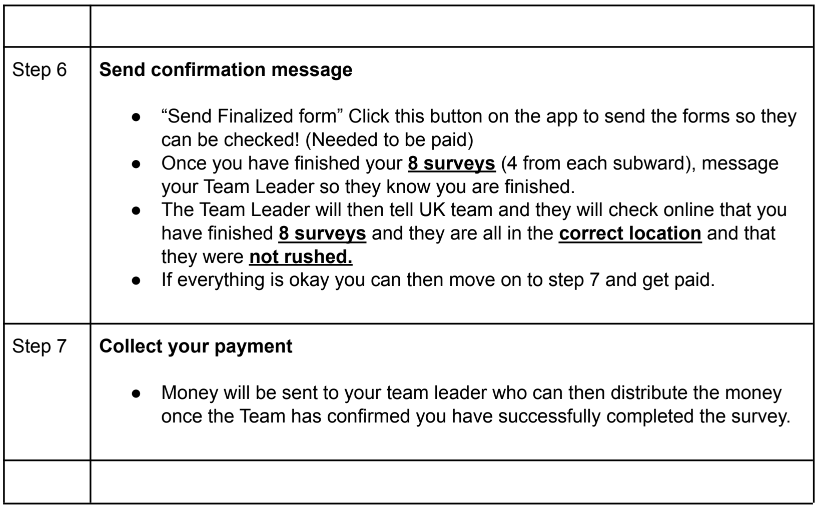
Appendix F

Street Survey Instructions



Facilitator Instructions!

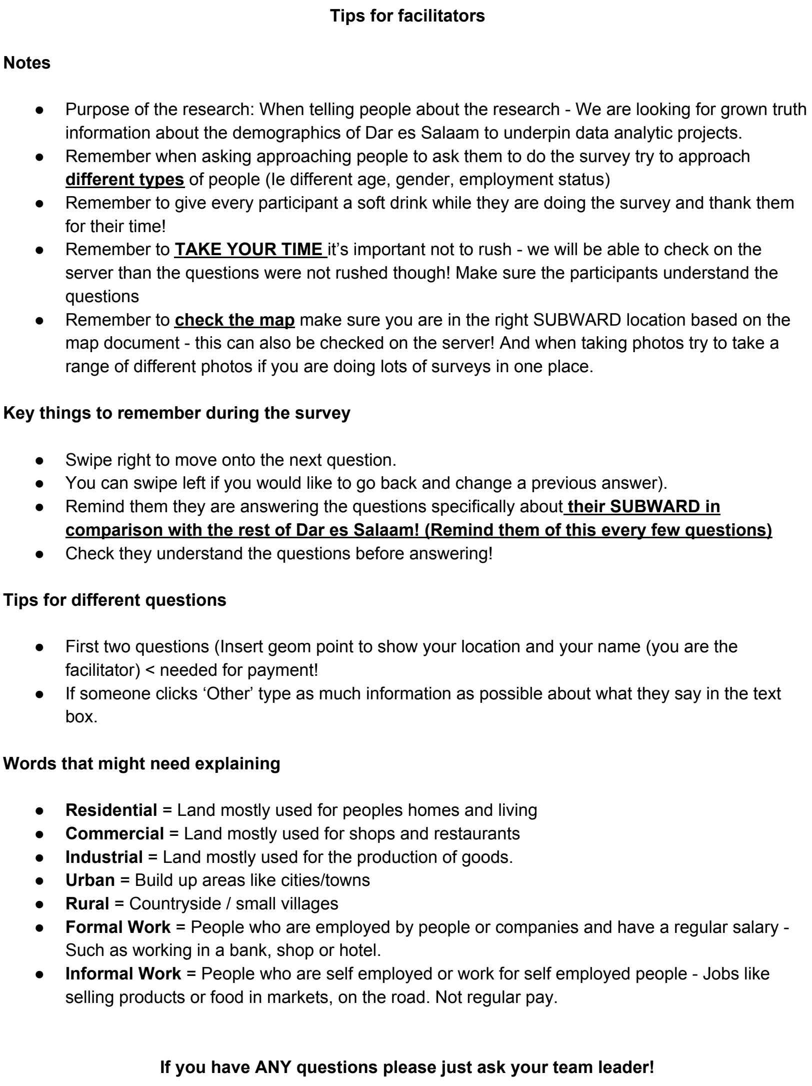
Step 1	**Download app** ● Go to google play ● Download the latest version of OpenData KIt Collect (Called 'ODK Collect')
Step 32	**Connecting to the server** ● Once you have seen your name on the Rota in the two slots you asked for you can download the server. ● In the 'ODK collect' go to 'General Settings' ● Click on 'Server Settings' ● Note: You do NOT need to use a username and password ● Where is says 'URL' type in https://kc.kobotoolbox.org/nlab ● Return to 'Main Menu' click 'Get Blank Form' and upload the '**DEMOGRAPHIC_SURVEY**' form. ● Note: Make sure on your phone settings the app has access to take photos and access your location!
Step 4	**Payment and Groups!** ● People will work in pairs. All groups of pairs can travels to wards and subwards together. Then people should stay in their pairs. ● Go through the Survey Tips Document with your team leader ● Check you know the location of your subwards on the Maps Document ● Travel Money ● Money for Drinks ● Stipend ● Collect 8 soft drinks - (4 for each subward - 1 for each survey participant)
Step 5	**Completing the survey** ● You must NOT share your phone with anyone else doing surveys - Each facilitator can use only one phone unless they talk to us first about this. ● Go to the ward you are doing the surveys in ● Select 'Fill Blank Form' on the 'ODK Collect' app to start a new survey ● Complete FOUR surveys in your first subward slot ● Complete FOUR surveys in your second subward slot ● Note: **TO GET PAID** You must put your first and second name in the first questions which says 'Facilitator Name' ● The app logs your location and this must be in the **CORRECT SUBWARD** otherwise you will **NOT GET PAID** ● Do not rush the surveys - The team leader will check on the app how long each survey took you to make sure you didn't rush **before you get paid.** **"Facilitators name" that question means YOUR name NOT the participant who they are asking the questions too!**



Step 6	**Send confirmation message** ● "Send Finalized form" Click this button on the app to send the forms so they can be checked! (Needed to be paid) ● Once you have finished your **8 surveys** (4 from each subward), message your Team Leader so they know you are finished. ● The Team Leader will then tell UK team and they will check online that you have finished **8 surveys** and they are all in the **correct location** and that they were **not rushed.** ● If everything is okay you can then move on to step 7 and get paid.
Step 7	**Collect your payment** ● Money will be sent to your team leader who can then distribute the money once the Team has confirmed you have successfully completed the survey.

Appendix G

Street Survey Facilitator Reminders



Tips for facilitators

Notes

- Purpose of the research: When telling people about the research - We are looking for grown truth information about the demographics of Dar es Salaam to underpin data analytic projects.
- Remember when asking approaching people to ask them to do the survey try to approach **different types** of people (Ie different age, gender, employment status)
- Remember to give every participant a soft drink while they are doing the survey and thank them for their time!
- Remember to **TAKE YOUR TIME** it's important not to rush - we will be able to check on the server than the questions were not rushed though! Make sure the participants understand the questions
- Remember to **check the map** make sure you are in the right SUBWARD location based on the map document - this can also be checked on the server! And when taking photos try to take a range of different photos if you are doing lots of surveys in one place.

Key things to remember during the survey

- Swipe right to move onto the next question.
- You can swipe left if you would like to go back and change a previous answer).
- Remind them they are answering the questions specifically about **their SUBWARD in comparison with the rest of Dar es Salaam! (Remind them of this every few questions)**
- Check they understand the questions before answering!

Tips for different questions

- First two questions (Insert geom point to show your location and your name (you are the facilitator) < needed for payment!
- If someone clicks 'Other' type as much information as possible about what they say in the text box.

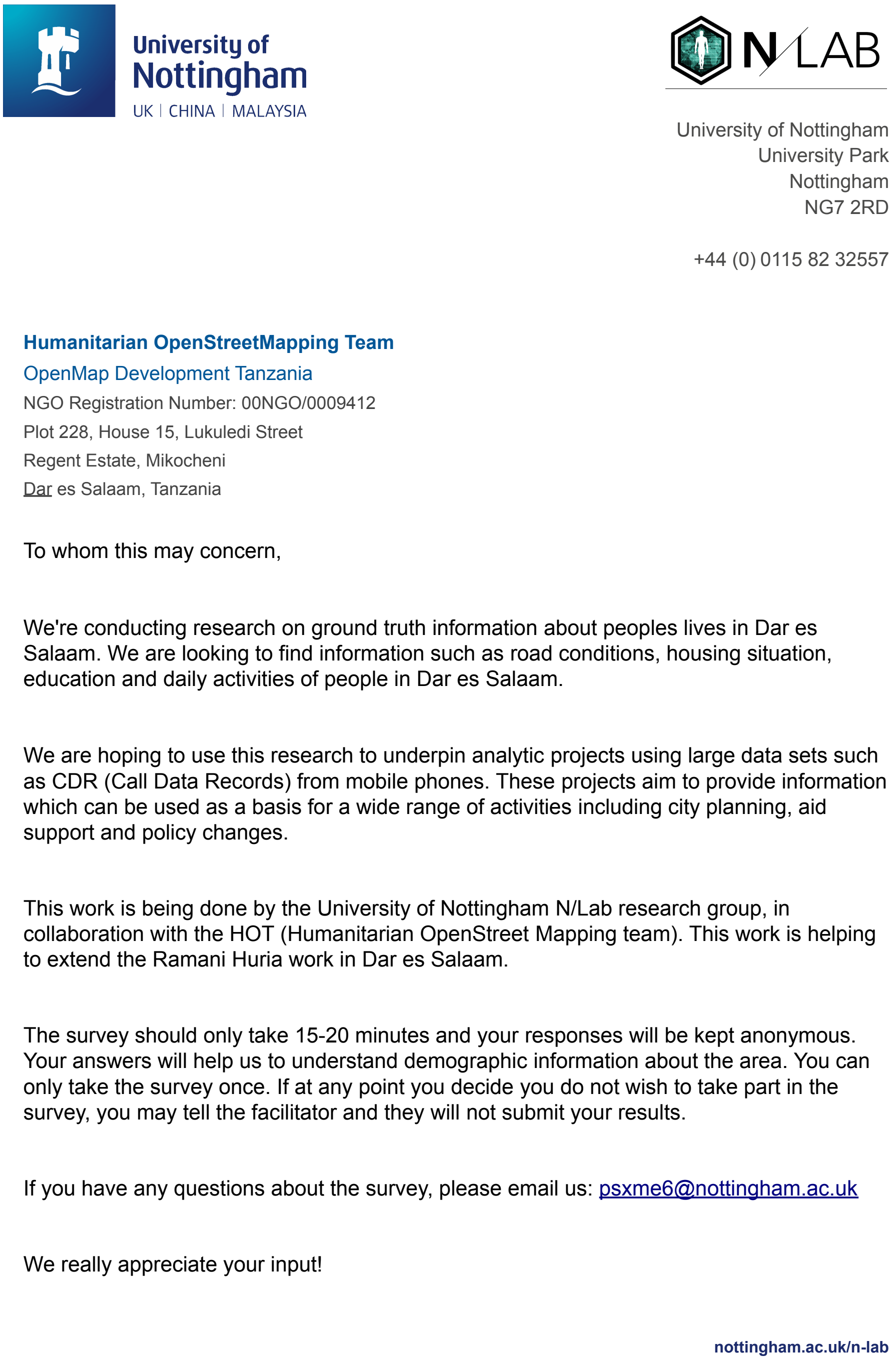
Words that might need explaining

- **Residential** = Land mostly used for peoples homes and living
- **Commercial** = Land mostly used for shops and restaurants
- **Industrial** = Land mostly used for the production of goods.
- **Urban** = Build up areas like cities/towns
- **Rural** = Countryside / small villages
- **Formal Work** = People who are employed by people or companies and have a regular salary - Such as working in a bank, shop or hotel.
- **Informal Work** = People who are self employed or work for self employed people - Jobs like selling products or food in markets, on the road. Not regular pay.

If you have ANY questions please just ask your team leader!

Appendix H

Street Survey Information Letter



University of Nottingham
University Park
Nottingham
NG7 2RD

+44 (0) 0115 82 32557

Humanitarian OpenStreetMapping Team
OpenMap Development Tanzania
NGO Registration Number: 00NGO/0009412
Plot 228, House 15, Lukuledi Street
Regent Estate, Mikocheni
Dar es Salaam, Tanzania

To whom this may concern,

We're conducting research on ground truth information about peoples lives in Dar es Salaam. We are looking to find information such as road conditions, housing situation, education and daily activities of people in Dar es Salaam.

We are hoping to use this research to underpin analytic projects using large data sets such as CDR (Call Data Records) from mobile phones. These projects aim to provide information which can be used as a basis for a wide range of activities including city planning, aid support and policy changes.

This work is being done by the University of Nottingham N/Lab research group, in collaboration with the HOT (Humanitarian OpenStreet Mapping team). This work is helping to extend the Ramani Huria work in Dar es Salaam.

The survey should only take 15-20 minutes and your responses will be kept anonymous. Your answers will help us to understand demographic information about the area. You can only take the survey once. If at any point you decide you do not wish to take part in the survey, you may tell the facilitator and they will not submit your results.

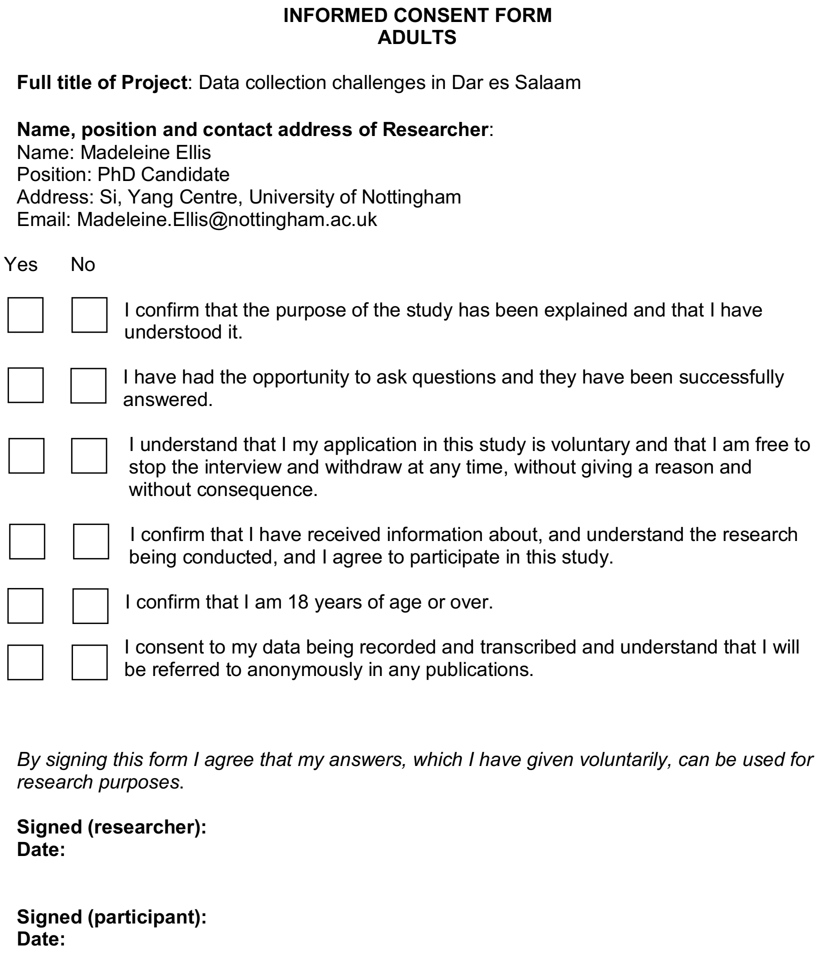
If you have any questions about the survey, please email us: psxme6@nottingham.ac.uk

We really appreciate your input!

nottingham.ac.uk/n-lab

Appendix I

Adult Consent Form



INFORMED CONSENT FORM
ADULTS

Full title of Project: Data collection challenges in Dar es Salaam

Name, position and contact address of Researcher:
Name: Madeleine Ellis
Position: PhD Candidate
Address: Si, Yang Centre, University of Nottingham
Email: Madeleine.Ellis@nottingham.ac.uk

Yes No

□ □ I confirm that the purpose of the study has been explained and that I have understood it.

□ □ I have had the opportunity to ask questions and they have been successfully answered.

□ □ I understand that I my application in this study is voluntary and that I am free to stop the interview and withdraw at any time, without giving a reason and without consequence.

□ □ I confirm that I have received information about, and understand the research being conducted, and I agree to participate in this study.

□ □ I confirm that I am 18 years of age or over.

□ □ I consent to my data being recorded and transcribed and understand that I will be referred to anonymously in any publications.

By signing this form I agree that my answers, which I have given voluntarily, can be used for research purposes.

Signed (researcher):
Date:

Signed (participant):
Date: