LU(S)TI in the Global South: An Empirical Analysis of Land Use and Socio-Economic – Transport Interaction in Tanzania using Mobile Network Data

Submitted November 2019, in partial fulfillment of
the conditions for the award of the degree **PhD Digital Economy.**

Gregor Engelmann
4238057

Supervised by James Goulding, Bertrand Perrat, David Golightly

Nottingham University
Business School

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

Signature _____

Date _____ / _____ / _____

Abstract

The majority of rural-urban migration is filtered through slums: informally established, unplanned, and unrecognised by the government, scientists have a minimal understanding of the 200,000 that exist worldwide, never mind enough insight into the millions of individuals living there. This limited understanding often coincides with a more general absence of data in traditional urban planning approaches, leading to most cities seeing development, positive or otherwise, preceding planning.

Wesolowski and Eagle (2010) highlighted the key need to use models of human mobility to help guide effective spatial planning policies. Previous research has shown that thinking about the built environment alone cannot account for individual differences in behaviour, and that we must also consider factors such as socio-economic circumstance and context (which are far more likely to contain explanatory value than the geographies of points of interest, such as home and work locations of individuals alone). However, this remains a very difficult topic to study. Emerging economies are often characterised by institutions struggling to keep even demographic data streams up to date. Combined with ineffective data collection strategies, it is often realistic to expect stakeholders to retain any overview of the dynamics of urban systems. This gap causes many issues, but particularly in East Africa: expense and logistics restrict ability to deploy sensor technologies; fast-changing environments reduce the utility of traditional household and census surveying; and even when raw data exists there are distinct skill gaps for data analysis.

To address this, this thesis extends nascent work, and systematically investigates the use of Call Detail Records (CDR) and Mobile Financial Service (MFS) transaction logs to model mobility, demographics, land use and their interplay. Data used was automatically generated as part of day-to-day operations of a major Tanzanian Mobile Network Operator. As part of this thesis, three empirical analyses are carried out to test the boundaries of inferring activity-based land use, predicting cell tower coverage level socio-economic levels and generating mobility metrics in the form of Origin-Destination matrices and synthetic daily activity plans for the Tanzanian port city of Dar Es Salaam. Further, shortcomings of CDR and MFS data, and ways to overcome these, are identified.

Empirical chapters form the basis for the identification of factors from the spatial dimension focused on assessing the impact of the built environment, socio-economic circumstance and mobility behaviour allowing for the extension of traditional land use-transport interaction (LUTI) models, through the inclusion of socio-economic characteristics. This culminates in a new empirical LU(S)TI analysis for a sub-Saharan context. The metropolitan area of the port city of Dar es Salaam, Tanzania, is a pertinent case study area as it is facing similar challenges to many other fast-growing metropolitan areas in emerging economies globally.

List of Publications

A number of papers have been published for the work of this thesis.

Engelmann, G., Goulding, J., Smith, G. (2018). The unbanked and poverty: Predicting area- level socio-economic status from M-Money transactions. Paper presented at IEEE Big Data, Seattle.

Engelmann, G. Franco, P., Pages, A.M., Tusting, R. (2018). A comparative analysis of telematics and mobile network phone data for evaluating travel behaviour. Paper presented at UTSG Annual Conference, London.

Engelmann, G., Goulding, J., Smith, G., Golightly, D. (2017). Estimating population behaviour to describe activity-based land-use in emerging economies using mobile phone event series. Paper presented at NetMob, Milan.

Engelmann, G., Goulding, J., Golightly, D. (2017). Estimating activity-based land-use through unsupervised learning from mobile phone event series in emerging economies. Paper presented at GISRUK, Manchester.

Franco, P., McCormick, E., Van Leeuwen, K., Johnston, R., Engelmann, G. (2017). Multimodal activity based models to support flexible demand mobility services. Paper presented at 24th ITS World Congress 2017, Montreal.

Goulding, J., Engelmann, G., Gavin, S., Iliffe, M. (2014). Best Practices and Methodology for OD Matrix Creation from CDR data, N/Lab, University of Nottingham.

Acknowledgements

While this thesis is my own work, it would not have been possible without the support, guidance and mentoring of my parents, colleagues, peers and friends.

I would especially like to thank my PhD supervisors James Goulding, David Golightly and Bertrand Perrat, my internal assessor Gavin Smith and my colleagues and staff at the University of Nottingham, the N-Lab and the Horizon Centre for Doctoral Training and here particularly Alex Ottway, Georgiana Avram, Roza Vasileva, Vanja Ljevar, Maddy Ellis, Rosa Ellen and Alfie Cameron. They have supported, suffered and put up with me throughout my years of research culminating in this thesis.

Furthermore, I would like to thank the Tanzanian Mobile Network Operator, Dar Ramani Huria and the World Bank for providing the datasets used as part of this thesis. Without data, this thesis would have not been possible.

Additionally, I would like to thank my external partner, the late Transport Systems Catapult, which is now a part of the newly formed Connected Places Catapult, and here specifically Robin North, Kristoff Van Leeuwen, Patrizia Franco, Chris Rushton and Ryan Johnston for feedback on both my research and the collaboration during my Industrial Placement.

A special thank you has to Joe Nash for introducing me to Hackathons at my first welcome fare at the University of Nottingham fresh of an undergraduate degree in International

Relations. Hackathons and my engagement with Major League Hacking, the global student hacker community, have fundamentally contributed to my interest in technology and data analytics as a catalyst for change that motivated me to undertake this thesis research.

Lastly, a sizeable chunk of various parts of this thesis was written in pubs and while travelling and I would like to acknowledge the Room with a Brew and the various Star Alliance airlines for making the task of writing on flights and in airport lounges an enjoyable exercise.

Contents

List of Tables

List of Figures

Acronyms

BTS Base Transceiver Station.

CDR Call Detail Record.

CFA Confirmatory Factor Analysis.

CFI Confirmatory Fit Index.

D4D Data for Development.

DHS Demographic and Health Survey.

GIS Geographic Information System.

GPS Global Positioning System.

GSMA Groupe Spcial Mobile.

LAU Location Area Update.

LDA Latent Dirichlet allocation.

LSOA Lower Super Output Area.

LSS Location Sharing Services.

LUTI Land Use – Transport Interaction.

MFS Mobile Financial Services.

MND Mobile Network Data.

MNO Mobile Network Operator.

MODLE Mobility on Demand Laboratory Environment.

MPI Multidimensional Poverty Index.

MSC Mobile Switching Centre.

NHTS National Household Travel Survey.

NMF Non-negative matrix factorization.

NSI National Statistical Institute.

NTL Night Time Light.

OD Origin-Destination.

OOB Out-of-bag.

PCA Principal Component Analysis.

PDP Partial Dependency Plot.

POI Points of Interest.

RSI Road Side Interview.

RSS Residential Self Selection.

SEL Socio-economic Level.

SEM Structural Equation Modelling.

SVM Support Vector Machine.

TCRA Tanzania Telecommunications Regulatory Authority.

TMZ Traffic Management Zone.

TZS Tanzanian Shilling.

VGI Volunteered Geographic Information.

Chapter 1

Introduction

Emerging economies are often characterised by governments and institutions struggling to keep official statistics up to date. Ineffective or absent data collection and governance strategies lead to a lack of current, fine-grained and reliable data on everything from land use and mobility to development and socio-economic statistics. Devarajan (2013) [93] referred to this as the 'Statistical Tragedy' affecting emerging economies throughout the world and in Sub-Saharan Africa in particular.

> "How would you feel if you were on an airplane and the pilot made the following announcement: "This is your captain speaking. I'm happy to report that all of our engines checked fine, we have just climbed to 36,000 feet, will soon reach our cruising speed, and should get to our destination right on time.... I think. You see, the airline has not invested enough in our flight instruments over the past 40 years. Some of them are obsolete, some are inaccurate and some are just plain broken. So, to be honest with you, I'm not sure how good the engines really are. And I can only estimate our altitude, speed and location. Apart from that, sit back, relax and enjoy the ride." This is, in a nutshell, the story of statistics in Africa." [137]

The majority of official statistics in emerging economies are estimated with actual numbers on metrics such as maternal mortality or malaria only being collected in less than 20% of cases if at all [237].

While data availability has significantly improved with the introduction of Demographic and Health Survey (DHS) programs in the mid-1980s and later monitoring systems for the Millennium Development Goals since 2000 and Sustainable Development Goals since 2015 collecting up to date, fine-grained and reliable data remains a persistent challenge.

This thesis examines how mobile phone generated 'Big Data' might bridge this gap through: generating insights on land use; socio-economics; mobility at scale, and the analysis of their interactions. While the potential for Big Data to improve official (development) statistics have received significant attention in previous years, the majority of this (empirical) work has focused on countries in the Global North, East Asia and South America, which was most likely predicated by data set availability [121]. The unfortunate side effect is increasing inequality in our understanding of urban contexts and urban phenomena between emerging economies in the Global South and developed economies elsewhere. While some of the research is transferable, a lack of specific insight into urban phenomena and dynamics in Africa remains. This lack of empirical work is particularly notable in East Africa, where conventional sensing technologies have not been able to overcome the inability of manual surveying to capture insights in fast-changing and particularly informal areas. Additionally, there is a gap in the usage of machine learning techniques to improve official statistics over more traditional statistical analysis techniques.

Geographically, the investigations of this thesis focus on the Tanzanian port city of Dar es Salaam metropolitan area, the extent of which is shown in Figure 1.1. Dar es Salaam is made up of the three municipalities – Ilala, Kinondoni and Temeke, which are akin to Medium Super Output Areas in the UK. These are then further divided into 90 wards, similar to Lower Super Output Areas in the UK, with 26 in Ilala, 34 in Kinondoni and 30 in Temeke. The data set used as part of this study also contains information for the rest of the country. Mobile phone subscribers behaviour in other geographic locations

were included to support the rationale and findings of this thesis for reasons which will be expanded upon in the following chapters.

The remaining sections of this chapter introduce the geographic scope, research domain, questions and contributions of this thesis, lastly describing the thesis structure.

1.1 Geographic Scope

Dar es Salaam is the former capital of Tanzania and was the first urban centre in Tanzania to be named 'city' on December 9, 1961, [72]. The city population has doubled in the last decade, bringing it to an estimated 4.3-5 million people with an expected rise in population numbers by 85% by 2025 [2]. According to recent estimates, 70% of the city's residents live in informal slums that are outside the scope of official statistics. Dar es Salaam is recognised as the second fastest-growing city in Africa after Kampala in Uganda and the ninth-fastest-growing globally [62, 300]. It spans an area of approximately $1590km^2$ and accounts for approximately 10% of the total Tanzania mainland population according to the last census conducted in 2012. While Dar es Salaam has lost the status of national capital to Dodoma (with some government bodies still being in the process of moving) the city plays an integral part in Tanzania's economy, accounting for almost 40% of the countries GDP. It also remains home to most of the national government agencies, and as a result, the central government retains considerable influence in city governance. Tanzania, and particularly Dar es Salaam with its extensive port facilities, are of considerable importance to economic development in East Africa as it is a transport hub for landlocked neighbours including Burundi, Rwanda, and Zambia. As such the city continued to invest heavily in infrastructure and was successful in attracting external funding for several infrastructure investments including a $300million investment by the International Development Association to fund a new open-access railway, and funding to build a new terminal for Tanzania's largest Airport Julius Nyerere located in Dar es Salaam, and for the Mfugale overpass at the junction of the Mandela Expressway and Julius Nyerere Road.

Mobile phones were first introduced in Tanzania in the mid-1990s with a single network offering by Millicom's Tigo. In 1995, there were 2200 mobile phone subscribers in Tanzania. Similar to many other emerging economies, mobile phone adoption in Tanzania grew rapidly to 10.4 million by 2008, and over 30 million by 2013. Current mobile phone penetration and connection numbers vary, with some reporting 25.3 million unique subscribers, 16.1 million mobile internet subscribers and 25.3% smartphone penetration [136] with the Tanzania Telecommunications Regulatory Authority (TCRA) [328] reporting over 39 million mobile connections within Tanzania indicating, that over 73% of adults in Tanzania own at least one feature phone or smartphone [245, 250, 275].

The Mobile Network Operator (MNO) that provided the data within this work has approximately 19.6 million active users, which accounts for 40% of the Tanzanian population. Studies in this thesis are based on a 20% sample of Mobile Network Data (MND) in the form of mobile financial transaction logs and Call Detail Record (CDR) logs for calls for one year, and SMS and internet data for six months.

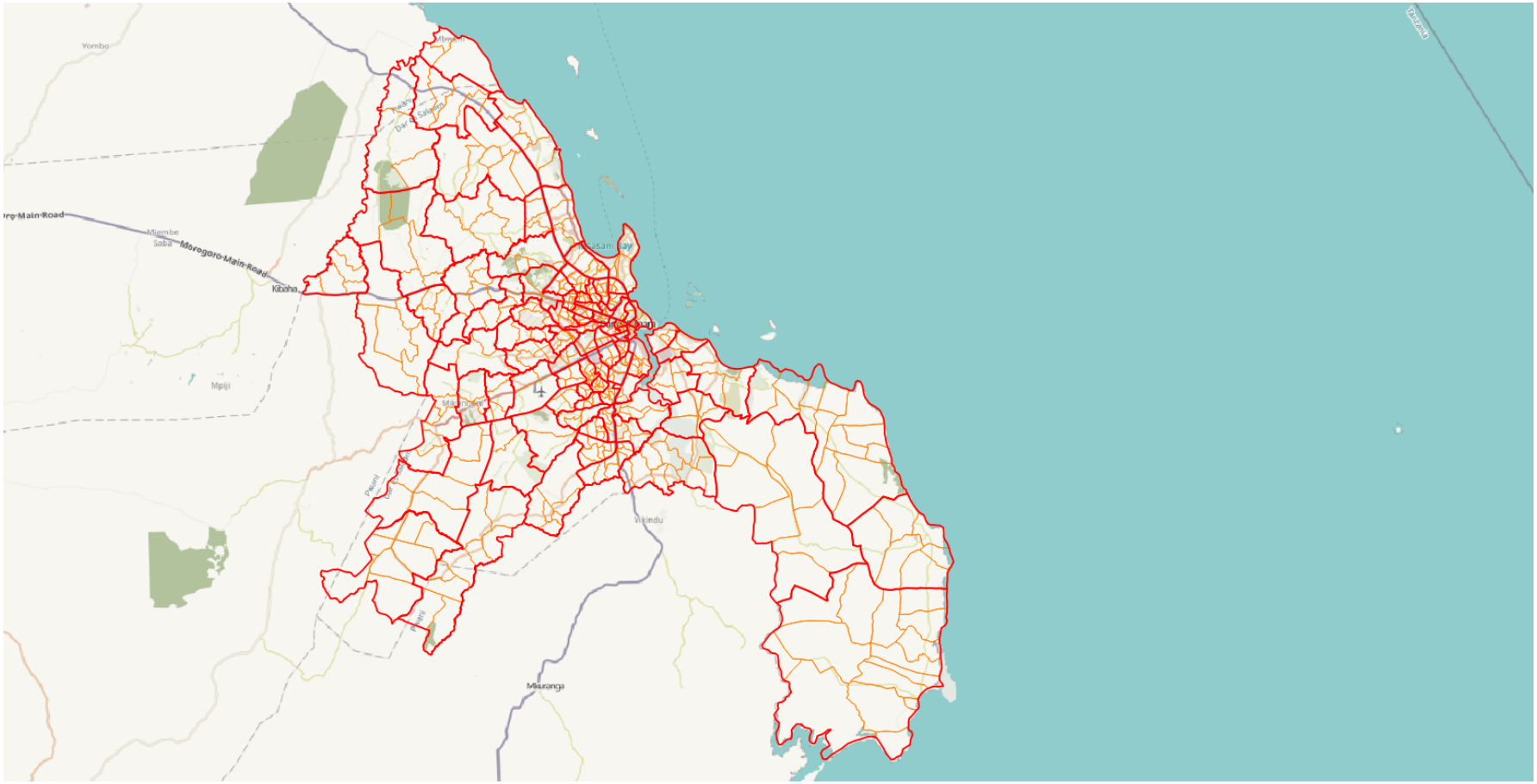


Figure 1.1: Extent of the Dar es Salaam, Tanzania, metropolitan area overlaid with administrative district (red) and ward (yellow) boundaries

1.2 Rural-Urban Migration

Estimations suggest that the global population will increase to ten billion by 2050 with the majority living in urban areas and with the majority increase taking place in emerging economies [356]. This increase in population numbers coincides with an increasing rural-urban migration leading to the majority of the global population living in formal and informal urban and peri-urban areas. The emergence of slums is an inevitable result of this shift, representing areas which are informally established, unplanned, and unrecognised by the government.

High levels of congestion and poor infrastructure conditions, in particular, have been recognised as significant challenges within Dar es Salaam. In recent years, these problems have been exacerbated by frequent flooding and the rise of unplanned settlements. Flooding has become an increasing threat to the low-lying coastal city of Dar es Salaam with rising sea levels and more intense rainfalls. The rise in unplanned settlements is partly a by-product of a housing shortage estimated at 3 million units with a projected increase of an additional 200,000 housing units each year [300].

Both formal and informal urban areas within emerging economies commonly face frequent land use changes. These changes often coincide with an absence of coordinated land management, unclear land use planning approaches and complex land tenure systems. This has led to many urban areas around the world where development precedes planning, giving rise to the emergence of vast slums and informal areas in the first place. Major metropolitan areas in emerging economies such as Dar es Salaam in Tanzania or Kumasi in Ghana are at the centre of countries' "socio-economic development, and yet have poor transport infrastructure and services, in terms of street connectivity, motorability and ease of mobility within neighborhoods particularly in emerging residential and peri-urban neighborhoods" [281, p.565].

This issue is exacerbated by settlement in the peripheries of cities, contributing to urban

sprawl, a phenomenon that is not unique to emerging economies. In the United States as well as elsewhere in the Global North, 'sprawling' low-density residential areas with commercial strip development and dependence on extensive automobile use are commonplace [178]. Overcoming the 'Statistical Tragedy' is essential to generating insights to guide effective infrastructure investments and land management and land use planning that is tailored to local contexts rather than simply copied from existing implementations in the Global North [193, 223].

The majority of the rural-urban migration in emerging economies is channelled through slums, areas which are informal, unplanned and unrecognised by the government. This pattern is particularly pertinent in Dar es Salaam, as according to recent estimates, 70% of the cities residents live in informal slums that are outside the scope of data collection for official statistics. This issue will likely be exacerbated in the future, as Dar es Salaam is the second-fastest-growing city in Africa after Kampala in Uganda and ninth-fastest-growing globally [62, 300].

Understanding those migratory patterns is integral to understanding the growth of urban areas. Without understanding the relationship between urban growth and mobility to guide transport infrastructure and service design, these urban spaces risk becoming dysfunctional and unlivable [281]. The prevention of which is enshrined within goal 11 of the United Nations SDGs [338]:

> "Make cities and human settlements inclusive, safe, resilient and sustainable"

Sub-goal 11.2 pays specific attention to the role of mobility in ensuring the overall success of goal 11:

> "By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons."

Addressing those goals is at the core of land management and land use planning and

> "correctly administered, it is an important tool for promoting investment, development, environmental improvements and quality of life." [339, p. 40]

As Chapter 2 will discuss, much of the land management and land use planning research over the last three decades has focused on the relationship between land use and mobility as well as land use determinants and travel time in the analysis of Land Use – Transport Interaction (LUTI) [241, 320]. Biased by 'urban sprawl', much of the research indicates a strong effect of land use on urban mobility, as residents in peri-urban areas tend to rely on extensive automobile use compared to those in dense urban centres. The latter assumption of reduced reliance in urban centres is frequently based on findings in areas with efficient integration of public transport within a city's land management and land use planning approach. These findings have significant impacts on the design of cities in the developed world and have been driving integrated land management and land use planning in developed economies for years [320, 352].

Urban areas in developed countries generally have high levels of transport accessibility and development tends only to commence once transport access and utilities such as water and electricity are in place unlike in emerging economies, where development often precedes planning [281, 320]. However, research and insight has mostly focused on countries in the Global North and is much more limited in emerging economies. There is a particularly limited understanding of the approximately 200,000 slums worldwide and the millions of individuals living within them [356].

The world has seen a rapid increase in the spread of Internet Communication Technologies with the proliferation of cheap mobile devices and massive investments in telecommunication and broadband infrastructure [127, 319]. Chief among these developments has been the drastic reduction in the price of mobile phones:

> "In 10 short years, what was once an object of luxury and privilege, the mobile

> phone, has become a basic necessity in Africa." Paul Kagame, President of Rwanda [1].

The increasing proliferation of mobile phones, in particular, has led to them becoming a proxy for data sources about human movement, generating vast amounts of data on human behaviour at scale and at low cost, effectively turning the mobile phone subscribers operating them into what Goodchild (2007) coined 'citizens as sensors'. MND generated by all phones, both smart and feature, such as CDR is a promising alternative to closing data gaps left by official data collection strategies while addressing some of the shortcomings of traditional and often costly sensor-based data collection (discussed in more detail in §5.2.1) to generate up to date, fine-grained data. While the potential for Big Data to improve official (development) statistics have received significant attention in previous years, the majority of this (empirical) work has focused on countries in the global, East Asia and South America, which was most likely predicated by data set availability. The unfortunate side effect is increasing inequality in our understanding of urban contexts, urban phenomena and the 'alternative' relationship between emerging economies in the Global South and developed economies elsewhere.

CDR data, in particular, has received extensive attention in fields such as epidemiology [53, 217, 231, 327, 332, 344, 355], transport [12, 26, 33, 40, 60, 58, 63, 125, 142, 176, 177, 185, 209, 228, 335, 347, 359, 363], and urban planning [113, 114, 213, 226, 232, 233, 234, 260, 273, 315, 316, 336, 341, 369] as they are collected automatically and at scale by the network operator for day-to-day operations such as network management and billing purposes. CDR data is available quickly, often within minutes of the network event, allowing for near real-time data collection [177, 293] and monitoring of changes in human activity over a prolonged period. The combination of CDR data with new machine learning methods has recently been proposed as a way to obtain this data without the expense required by traditional census and household survey methods. While the clear potential exists, the challenges of re-purposing such data with high utility in this context

[1] During a speech at Connect Africa Summit, October 29, 2007

have seen limited study.

1.3 Thesis Motivation

While each country and its cities are unique, all of them are united by a need for accurate and fine-grained data to design adequate urban spaces and provide public services to their populace. This is of particular importance among emerging economies in the Global South. Traditional methods of data collection such as census and household surveys are unable to cope with rapid changes in the urban fabric driven by increasing rural-urban migration and the emergence of informal slums. A pattern that is particularly pertinent in Dar es Salaam, as according to recent estimates, a mere 30% of the city is formally planned, with the remaining 70 % of the city being *de-facto* classed as slums with little or no reliable, fine-grained data on land use, socio-economics or mobility or ways to collect it effectively [173]. This issue will likely be exacerbated in the future, owing to its rapid growth.

The rapid growth creates multiple gaps in the already ineffective data collection and governance strategies and capacities of the responsible statistical and data-providing institutions. New data streams such as MND and, much more recently, drone imagery, which can be collected quickly, at scale and low cost, have the potential to close or at least lessen the gaps in data collection and governance.

1.4 Data in Dar es Salaam

Similar to many other emerging economies around the world, Tanzania is facing parts of the 'Statistical Tragedy'. In a report on the state of urbanisation in Dar es Salaam, Baker (2011) [27, p.12] highlights, this as:

> "accessing data, maps, and climate projections was problematic. Information is scattered across many different agencies, departments, organizations, and research institutions, with some reluctant to share data."

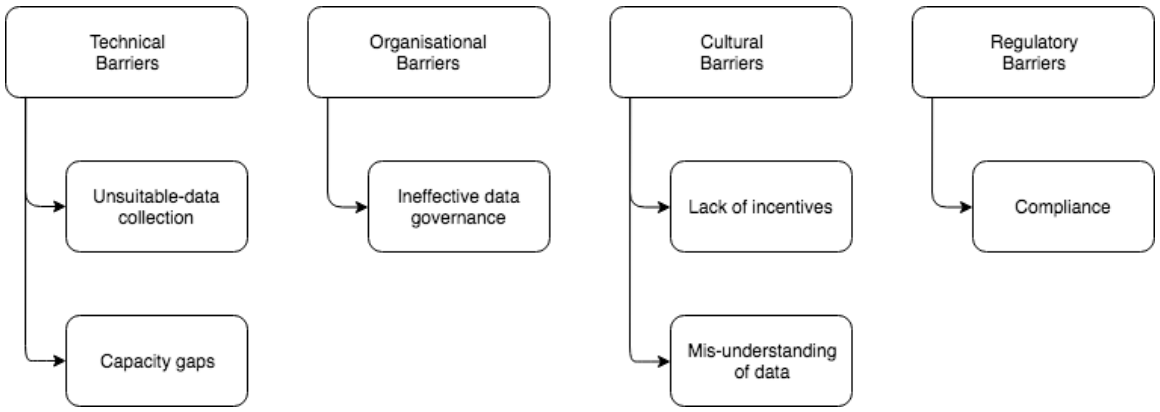


Figure 1.2: Technical, organisational, cultural and regulatory barriers affecting effective data collection and governance, and data-driven planning and decision making [342]

This issue is further exacerbated through barriers spanning several different domains that were identified as part of preliminary research by Roza Vasileva (2019) [342] at the University of Nottingham (see Figure 1.2):

1. **Technical barriers:**

   (a) Unsuitable data-collection approaches: Census data, which is the primary source of numerous official statistics were last collected in 2012 and can be reliably used by neither government nor other organisations and communities for proper planning and service provision. More detailed household surveys only provide data for extremely small sub-samples of the population while censuses are coarse and rapidly out-of-date, and often conducted but once a decade [93, 232].

   (b) Capacity gaps: In addition to unsuitable data collection and governance approaches, the coordination between ministries within Tanzania is minimal as staff with the necessary technical capabilities to analyse collected data are often overwhelmed, while others are struggling to contribute effectively in light of the challenges faced by the city of Dar es Salaam [27]. Outside Ministries, the technical capabilities are often even lower. From the author's experience (see Figure 1.3), the majority of ward leaders, for example, are struggling with the interpretation of maps. The lack in capability creates gaps between the collection of geospatial data and the ability use it effectively within decision-

Figure 1.3: Mapping workshop with local ward officials facilitated by Dar Ramani Huria and Humanitarian Open Street Map

making processes as well as general frustration when being confronted with the information presented in ways that appear confusing [362, 381].

2. **Organisational barriers:** The process of requesting data is time-consuming as highlighted earlier [27], which undermines the timeliness of evidence-based planning and decision-making. The city's complex governance structure further exacerbates these coordination challenges, as political tensions hinder the smooth facilitation of data sharing.

3. **Cultural barriers:** Traditionally, communities may have been incentivised to display weak (development, economic, etc.) performance in order to be eligible for additional development funding, resulting in the non-disclosure or release of manipulated data. While a core premise behind the opening and sharing of data is the fostering of accountability, the prior incentives lead to a cultural climate in which this step is not well-received [227, 307]. In other cases, decision-makers may not understand the value of data, making the collection of it an afterthought to any activity. This misunderstanding of data tends to strive in the absence of technical capabilities outlined earlier.

4. **Regulatory barriers:** The new Tanzania Statistics Act, which has been enacted in Tanzania in 2015 (with further amendments passed recently,) requires non-governmental stakeholders within the country to obtain permits for data collection. The Act regulates how non-governmental organisations can collect data and how data should be used publicly. It clarifies the mission of the National Bureau of Statistics of Tanzania "to produce quality official statistics and services that meet needs of national and international stakeholders for evidence based planning and decision making" and enshrines the NBS vision "to become a one-stop centre for offical statistics [published] in Tanzania" [256, p.22]. The data shared through the NBS portal covers geospatial data, census data, and socio-economic data, among others. Complying with official regulations to obtain a permit for data collection is difficult, however. Several organisations have raised concerns that the Act may have harmful implications for anyone who works with data in Tanzania, especially media and non-governmental organisations [309]. They argue that the Act hinders the collection and analysis of data, as well as the publishing of new statistics and reports.

1.5 Land Use – Transport Interaction (LUTI)

The importance of incorporating land use within the analysis of transport patterns has been recognised since the 1960s with initial work by Forrester (1969) [123] in his Theory of Urban Interaction. The initial spatial Land use – transport interaction model was an aspatial model to study the interaction between population, employment and housing.
In a seminal work on slum analysis, Wesolowski and Eagle (2010) [356] have echoed the need to understand human mobility to better aid land management and land use planning policies.

In an inverse relationship, however, research has shown that the built environment alone cannot account for individual differences in mobility behaviour [25]. Instead, socio-economic circumstances and more subjective aspects such as culture, attitudes and pref-

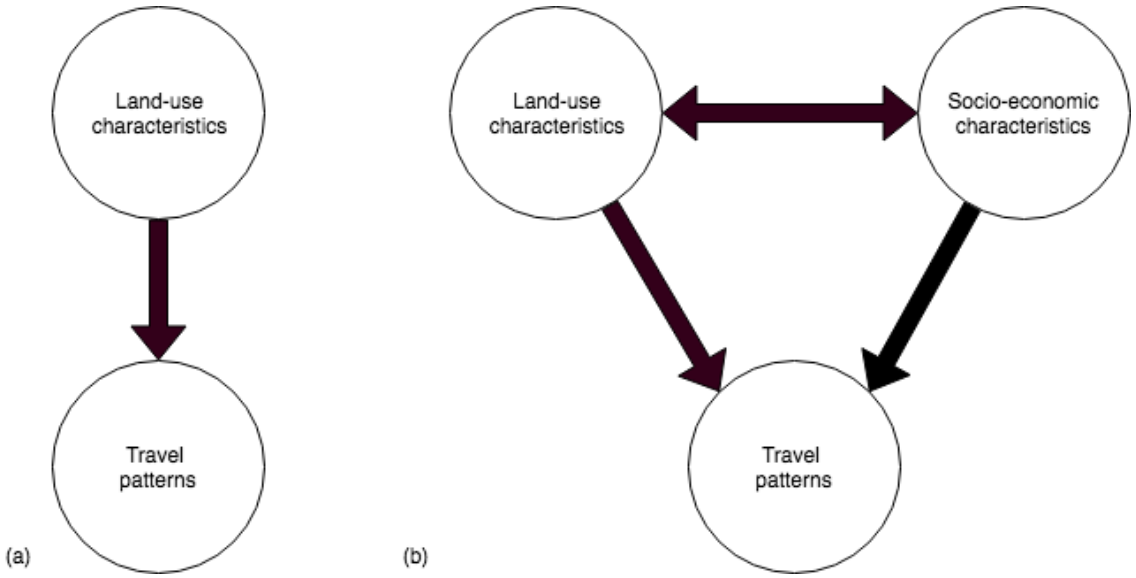


Figure 1.4: From Stead (2001)[320] defined the relationship between urban form and mobility as (a) 'traditional' cause and effect relationship and (b) 'alternative' interdependent relationship.

erences have been recognised to have additional explanatory value over the built environment and the physical location of Points of Interest (POI) and geographies such as home and work alone [197, 320].

Through generating both land use, mobility variables and additional socio-economic factors, the opportunity exists to move beyond the traditional LUTI relationship analysis as defined by Stead (2001) [320] and highlighted in Figure 1.4. In doing so, the opportunity exists to empirically examine the alternative relationship in Dar es Salaam as a pertinent example of a fast-evolving urban space with high-levels of informal living in the Global South. Structural Equation Modelling (SEM) will allow for the analysis of multiple variables across several directions of influence [340].

1.6 Research Objectives

As is fully explored in the following chapter reviewing the literature, there are clear gaps in the analysis of the traditional LUTI relationship in general, and difficulties investigating it empirically in the emerging economies in particular:

1. Traditionally, mobility, land use, and socio-economic factors have been analysed in isolation due to the absence of data and often incompatible levels of aggregation. Figure 1.4 highlights the traditional cause-effect relationship that has been the focus of much research and the alternative relationship accounting for both land use and socio-economic characteristics that has seen much more limited study.

2. There is a scarcity in empirical studies on both the traditional and the alternative relationship in emerging economies. Those few studies analysing either the traditional or the alternative relationship focus on India [8], Iran [117], China [218, 258, 377], Thailand [343] with Ghana as the only African country on the list [281].

3. There is a general lack of work using data generated through non-traditional data sources, such as household interviews, census surveys, fixed-sensor infrastructure, or satellite-imagery generated within African countries.

Accordingly, the research question of the thesis is:

> ***Is there value in using CDR and Mobile Financial Services (MFS) data for generating insight into urban land use, socio-economic levels and mobility trends, and can those be used to analyse the alternative LUTI relationship in an emerging economy context?***

This question will be addressed through the analysis of CDR and MFS data generated by a Tanzanian MNO for the city of Dar es Salaam, ultimately creating methodological contributions in the analysis of those features using MND and empirical findings for both Dar es Salaam, and the alternative relationship for the city. To research and validate these aims, a series of research objectives guide this thesis, namely:

Research Objective 1: to examine whether differences in activity-based land use and density can be distinguished from behavioural patterns contained within CDR data;

Research Objective 2: the investigation of small area Socio-economic Level (SEL) classification using CDR and MFS data through Supervised machine learning, and

subsequent analysis of features used for classification to understand the main determinants behind classification results;

Research Objective 3: exploration of synthetic daily activity plans based on the previously-evidenced assumption that the majority of human movement is predictable, and the generation of transient Origin-Destination (OD) matrices to understand travel and mobility patterns for Dar es Salaam;

Research Objective 4: analysis of the alternative land use – transport interaction accounting for socio-economic characteristics for Dar es Salaam using variables identified from CDR and MFS data through Research Objective 1-3;

Research Objective 5: identification of shortcomings of both CDR and MFS data, and potential solutions to address those.

These objectives allow for a thorough investigation of the concepts and practices of MND data analysis in the respective fields, while suitably defining the scope and methods deployed in the analysis chapters of this thesis.

1.7 Thesis Contributions

Dar es Salaam is among the most pertinent examples of rapidly growing cities globally. Increasing rural–urban migration, population growth, and the resulting urbanisation are among the main challenges countries globally and in emerging markets, in particular, need to solve. Conventional data sources such as manually collected surveys, satellite imagery and sensor-based infrastructure are limited in their ability to produce up to date, fine-grained data dynamically and at scale. This thesis argues that MND, using new machine learning methods over traditional statistical approaches, could meet this gap. As such, machine learning techniques are used to both generate insights into land use (Chapter 3), socio-economics (Chapter 4) and mobility patterns (Chapter 5) for the metropolitan area of Dar es Salaam; to analyse the relationship between them (Chapter 6) to better understand the city; and to identify limitations of CDR data analysis and ways to address

those (Chapter 7).

Due to aforementioned issues of increasing rural–urban migration, population growth, and the resulting urbanisation; and barriers to data collection in Dar es Salaam (§1.4) that are common throughout the Global South, up to date and fine-grained insights generated dynamically and at scale required for effective planning are seldom available. At the same time, however, Sub-Saharan Africa is home to the most successful Mobile Financial Services offerings and very high rates of mobile phone subscriber penetration leading to the generations of mass amounts Mobile Phone generated 'big data' that has the potential to supplement or even replace more conventional methods of data collection in emerging economies. It is this gap, where this thesis aims to make its contributions.
In summary, the contributions of this thesis are as follows:

- Discussion on how data generated through mobile phone usage can be used to create novel insights into land use, socio-economic and mobility patterns and their interaction in lieu of traditional demographic data.
- Empirical analysis of the land use and socio-economic – transport interaction for the metropolitan area of Dar es Salaam.
- Discussion of transferability of CDR/MFS methodology and avenues to overcome limitations in the analysis of such data.

1.8 Thesis Structure

The remainder of this thesis is structured as follows:

Chapter 2 describes the MND data used as part of this thesis, reviews the relevant literature on evidence-based planning and decision-making approaches in LUTI in emerging economies, and the factors within the different dimensions used to analyse the traditional and alternative relationship and further defines the problem statement that this thesis addresses, expanding upon the rationale for the above-

mentioned problem statement, research question, aims and objectives;

Chapter 3 reviews the relevant existing work in tracking activity-based land use, traditional sources of data used and develops activity-based land use features for Dar es Salaam using the techniques of dimension reduction and unsupervised clustering from SMS and call-derived activity signatures;

Chapter 4 reviews the relevant existing work in tracking fine-grained socio-economics and traditional sources of data used, and derives features on basic usage, regularity, diversity, activity and spatial behavior from both CDR and MFS data for SEL prediction using supervised classification, and exploratory analysis to understand the main determinants behind the classification results;

Chapter 5 reviews the relevant existing work in deriving OD matrices and compares traditional transient approaches with a more synthetic daily activity plan approach;

Chapter6 analyses the alternative relationship within Dar es Salaam using variables generated as part of this thesis using Structural Equation Modelling;

Chapter 7 discusses the individual-level, Base Transceiver Station (BTS)-level, population-level and real-world usability limitations, and ways to address these, identified through the analytical work in chapters 3-5 of this thesis;

Chapter 8 provides concluding remarks and proposes future research directions, against which work can be evaluated and furthered.

Figure 1.5 shows the main structure of the thesis.

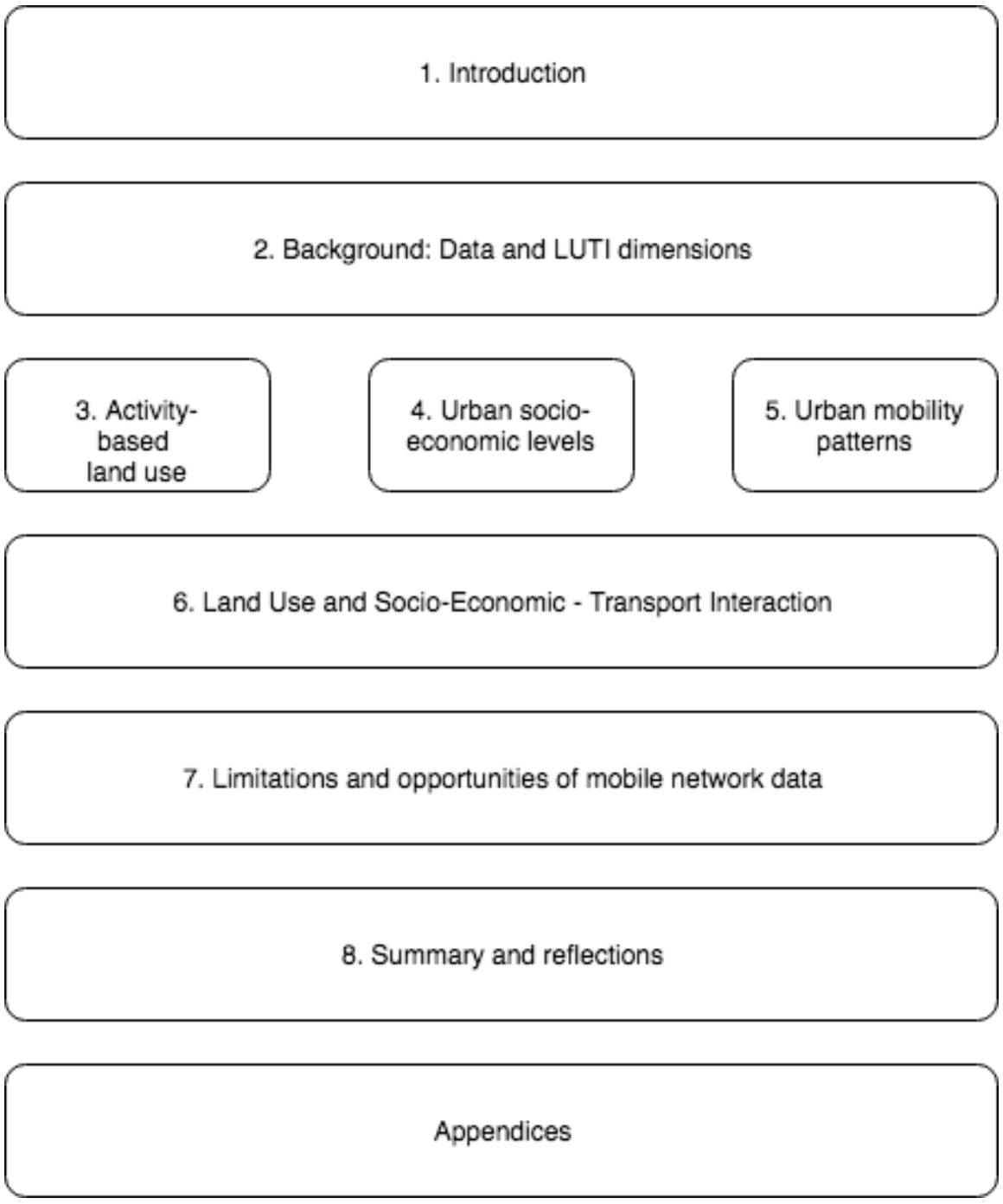


Figure 1.5: Thesis Structure

1.9 Chapter Summary

This chapter lays the foundations for this thesis by introducing the context, motivation and outline of this program of research. The aims and objectives of the research have been described. The structure of the various analyses and case studies are illustrated using a research overview diagram in Figure 1.5.

Chapter 2

Background: Data and LU(S)TI Dimensions

2.1 Chapter Introduction

Before the examination of existing literature on LUTI in emerging economies and reviews of the dimensions and factors used to model mobility behaviour, this chapter begins by providing a brief overview of major datasets that paved the way for the scientific analysis of mobile phone datasets. Second, this chapter describes CDR and MFS data in more detail by providing an overview of the data's structure, and generation. The final part of this chapter introduces the independent spatial, socio-economic and personality dimensions and commonly-used mobility behaviour variables they seek to explain.

2.1.1 Enabling CDR Data Research

The following section provides an overview of datasets within the field of CDR data analysis. While not an exhaustive list, it includes some of the key ones that have shaped this emerging field of research.

MIT Media Lab Reality Mining Project

MIT Media Lab Reality Mining project dataset that tracked BTS locations of a 100 students equipped with Nokia 6600 feature phones over nine months. The project was aimed at studying community dynamics and interaction with the data set later published as open data [104].

2012-2015 D4D Challenge

The Data for Development (D4D) challenge organised by Orange was the first major initiative to make CDR data available to external researchers. It was launched in collaboration with Orange, University of Louvain, Belgium, and MIT.

> "The goal of the challenge is to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Ivory Coast [and Senegale [89]] population." [46]

Work using these datasets has focussed on country-level trends of patterns of communication and mobility, social network structure within communication networks and epidemiology [248, 45]. Two editions of the challenge were run, each focussing on a different country.

Côte d'Ivoire 2012-2013 [46] Côte d'Ivoire has a population of approximately 20 million. At the time of the challenge, Orange Côte I'Voire had an estimated 25% market penetration with 5 million mobile phone subscribers. 2.5 billion calls and SMS CDR's generated over 150 days from December 1st 2011 until April 28th 2012 were used to generate the four datasets used as part of the challenge. The CDR's only account for Orange customers and interactions between them. The study period encompassed 3600 hours worth of recording with 100 hours missing for technical reasons.

SET1 BTS-BTS traffic: Hourly counts of Erlang and number of calls aggregated to different administrative levels in Côte I'Voire

SET2 Individual trajectories: High Spatial Resolution Data: High-resolution trajectories for 50,000 randomly sampled users for 14 days at BTS level

SET3 Individual Trajectories: Long Term Data: 50,000 users for entire period with reduced spatial resolution at sub-prefecture level ($n = 255$) with some sub-prefectures not having any BTS

SET4 Communication Subgraphs: for 5000 randomly selected users

The first edition of the challenge received 260 applications with more than 83 papers using the data sets being produced [89, 214].

Senegal 2014-2015 [89] The data for the second edition was provided by Sonatel of the Orange Group. At the time of the challenge, Sonatel had approximately 9 million mobile phone subscribers. CDR's generated between January 1st, and December 31st 2013 were used to generate three datasets. Customers with less than 75% active days and more than 1000 interactions per week were excluded from the analysis.

SET1 BTS-BTS traffic: Hourly counts for 1666 BTS in Senegal

SET2 Individual trajectories: High Spatial Resolution Data: High-resolution trajectories for 14 days for originally 9 million mobile phone subscribers at BTS level in addition to behavioral indicators for 300k users

SET3 Individual Trajectories: Long Term Data: Mobility traces for originally 9 million mobile phone subscribers for the entire period with reduced spatial resolution at *arrondissement* level ($n = 255$) in addition to behavioral indicators for 146,352 randomly sampled users

Behavioural indicators in SET2 and SET3 had been generated using the Bandicoot toolkit, an open-source Python kit for the analysis of MND [88], and include some of those generated for the classification of SEL in Chapter 4.

2012 Cairo Transport App Challenge

Based on the success of the first edition of the D4D challenge on Côte d'Ivoire, Vodafone initiated the Cairo Transport App Challenge in collaboration with the World Bank, Vodafone and IBM research. As part of the challenge, IBM was to use its AllAboard solution developed as part of the D4D challenge to analyse mobility patterns in Cairo [40, 94]. Legal issues led to the eventual failure of the Cairo Transport App Challenge, however, at the behest of the Egyptian National Telecoms Regulatory authority. The Authority requested, that all CDR's be retained within Egyptian state borders and only be accessed by Egyptian nationals and that the AllAboard platform be run on Egyptian servers [214].

2013 Telefonica's 'Datathon for Social Good'

Telefonica Dynamic Insights organised the Datathon for Social Good in conjunction with the ODI and the MIT Human Dynamics Laboratory as part of the European Campus Party. It included "Geo-localised open data sets in transportation, hospital admission and emergency services location; non-localised Twitter data sets; and anonymised and aggregated data from Telefnicas UK mobile network including calculations of footfall for the London Metropolitan Area over the course of 3 weeks" [262].

2015 Telecom Italia Big Data Challenge

The Telecom Italia Big Data Challenge was organised by Telecom Italia in collaboration with Politecnico di Milano, MIT Media Lab, Trento RISE and EIT ICT Labs [29, 329]. In total, more than 650 teams from over 100 universities participated with the data later released under the Open Database License. Similar to the D4D challenge, two editions were run based on the success of the initial challenge. At the time of the challenge, Telecom Italia had approximately 34% market penetration within Italy.

Challenge 1: November 1st 2013 to January 1st 2014 The challenge datasets included CDR, social pulse (geo-located tweets), weather, precipitation, electricity and news data for the city of Milan and the Province of Trentino for 61 days from November to

December 2013. As data was received from multiple different agencies and companies, it was spatially aggregated into $235m \times 235m$ grid cells in WGS84 (EPSG:4326) to account for different spatial granularities of the datasets. The issue of different spatial granularities and the need for aggregation is also discussed in more detail in §7.5.2. Milan was chosen as a study site as it is a major economic hub in Italy without any scheduled major events (such as the Fiere and Milan Design week) that could have introduced anomalous patterns into the data throughout the study period. The CDR data was pre-aggregated into 10-minute intervals, and included metrics on received and sent SMS, incoming and outgoing call and data connections. Similar to the D4D datasets, only Telecom Italia customers and interactions between those were included in the datasets.

Challenge 2: In addition to sets from Challenge 1, "the second edition also provides private mobility data (trips performed by customers of some car security and insurance companies), demographic data from Telecom Italia (e.g., gender, age-range and living area) and detailed Italian companies' information (e.g., number of employees, size and locations)" [29, p.14]. Beyond the additional datasets, it was expanded to include the cities of Bari, Milan, Naples, Rome, Turin, Venice and Palermo.

2.2 Mobile Network Data

This thesis makes use of datasets comprising CDR data for calls, SMS and Mobile Data usage, and MFS transaction from a Tanzanian MNO for 2014. While MFS and CDR's comprising call events were available for the entire year of 2014, Mobile Data usage was only available from January 1st until July 5th and SMS records from August 1st until December 19th. Due to the unavailability of data for a period with overlaps across all four event categories, periods with the SMS data were chosen over those with mobile data usage due to the relatively low penetration of smartphones compared to feature ones across all sectors of society at the time the data was collected for analysis in Chapters 3 to 5. The following section will provide a brief description of CDR and MFS data structure and creation.

2.2.1 CDR Data

CDR's are automatically generated for every network event by MNO's for billing, network management and maintenance purposes. Those range from tracking handsets within the network to understanding network performance [228]. They are generated whenever a network event takes place, allowing for the capture of the individualistic, spatial and temporal behaviour of users. They capture insight into both micro- and macro patterns of human interaction while allowing for the preservation of individual anonymity through spatial and temporal aggregation. Network events generating data include

- **Active events**
  - **Connection events** when a handset is turned on or off, or losses or regains connection
  - **Call events** when a phone call is placed or received and when the handset is moving between cells during a call. In the case of an intra-network call, one CDR per user is created, which is indicated by a change in called-party number and calling party number with corresponding charging times between both CDR's. The charging time corresponds with the time when the receiver answered the phone.
  - **Text events** when a text message is sent or received.
- **Passive events**
  - **Time-based events** when no event is generated over a fixed time period such as 3 hours.
  - **Movement events** when a handset moves between Location areas or when switching between 2G/3G/4G bands. With call events or data usage, 'handovers' or LAU can take place. Both describe a change in BTS associated with a network event due to movement and change in coverage. A LAU occurs when a call is transferred between different BTS while the user is moving. In the data set used as part of this study, LAU's were identified when the subsequent

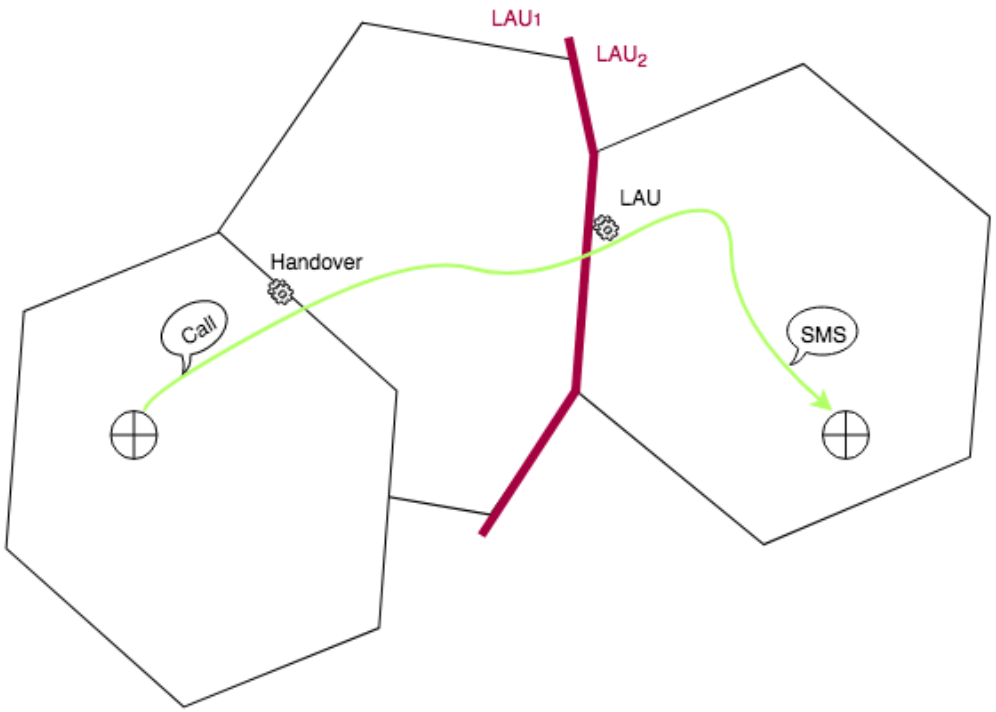


Figure 2.1: Example of location management of a mobile phone: a phone call, handover between different BTS and a LAU [209]

record for a user had a different BTS, but an identical called party and a wait time of zero. A location area is a 'supercell', which is made up of multiple BTS coverage cells. In the case of phone calls, it can be assumed that the BTS associated with a CDR corresponds with the origin BTS for the network event. Becker *et al.* (2011) [34] have shown, that handover patterns are consistent across different routes, speeds, directions, handset models and weather conditions. Figure 2.1 shows an example trajectory, including handovers and an LAU during a phone call. Whether separate CDR's for handover or LAU events are generated depends on the location management preferences of the network operator.

In its simplest form, each CDR log contains a wide range of meta-data including:

- **Timestamp** - when a network event occurred.
- **Temporal Mobile Subscriber Identity** - is used for security reasons to obscure the identify of a subscriber

Table 2.1: Structure of a simplified CDR depicting an anonymised identifier for the caller and the recipient of the call, a cell coverage identifier, which can be linked to the BTS used for the call, a timestamp of when the connection was terminated, the time from placing the call until the recipient answered the phone, and the duration of the call itself.

Calling party	Called Party	BTS ID	Timestamp	Waittime	Callduration
asjhldhfsf	jdshfewuip	12038097523	13-02-2014 00:01:12	0:00:10	0:05:12

- **Mobile Station International Subscriber Directory Number** - The encoded phone number. For privacy purposes, the mobile phone number of the caller and called user are anonymised by the network provider using the secure SHA-3 algorithm in line with Groupe Spcial Mobile (GSMA) guidelines [149].

- **'CallingcellID'** - is a unique sector antenna ID that is used to link a CDR to a BTS. There can be multiple callingcellIDs associated with a single BTS. Network events within the CDR are directly linked to BTS locations through a callingcellID associated with each event. This cell tower is formally called a BTS, which has a unique identifier, and latitude and longitude. As no signal strength information is contained within the database, the CDR's used as part of this study are of a more coarse-grained resolution and associated with BTS rather than individual handset triangulations.

- **'Waitduration'** and **'Callduration'** - are unique to call records and are used to calculate Erlang and handovers

The structure of a sample CDR can be found in Table 2.1. More specific information for the different types of network events such as a wait duration between a user dialling a number and the phone being answered and the call duration can also be included. A CDR record only contains meta-data about the network event and never any of the content transmitted as part of the interaction.

The data sets used as part of this thesis only contain data from active call and SMS events, which are a subset of the mobility data that is available in cellular networks.

2.2.2 Mobile Money Data

MFS is an umbrella term for a range of services offered by MNO, which include "sending and receiving money, making savings deposits, bill payments, making non-cash payments and transferring money from ones mobile phone account to bank accounts and vice versa" [250, p.4]. Similar to CDR's being recorded whenever an active or passive network event takes place as discussed in §2.2.1, a MFS transaction record is generated whenever a transaction takes place. Those transactions types include:

- **Account management** such as changing the pin, a statement showing the last five transactions, and checking the remaining balance
- **Money management** which includes deposits, withdrawals, and transfers to savings accounts
- **Payments** for money transfers to others using Peer-to-Peer and bill payments

Each record contained several attributes collected by the network operator as part of the day-to-day MFS provision. Those attributes include:

- **SIM identifier**: anonymised 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)

Each transaction could have either no error code (success) or one of the multiple error codes (some of which indicate a successful transaction) attached. The full taxonomy of error codes is explained in more detail in Appendix B.

Similar to the CDR data discussed above, the MFS data set has been made available by a sizeable Tanzanian MNO. The datasets represent a random sample from the entire subscriber base incorporating millions of mobile phone subscribers within the geographic boundaries of Tanzania. The MFS data set contains transaction records for regular mobile phone subscribers, businesses or agents and the network operator itself.

2.3 Land use – Transport Interaction

Much of the LUTI research over the last three decades has focused on the relationship between land use and mobility behaviour only [25, 68, 69, 124, 198, 241, 320]. Most of these studies found a correlation between dense development, mixed land use, accessible design and a reduction in vehicle trips, travel distances and the increasing attractiveness of public transit, walking and cycling. Due to differences in geographical settings and urbanisation patterns, however, conclusions about the relative importance of the built environment in determining mobility patterns may differ across empirical settings [340].

> "European cities have a historic city centre, sometimes dating from the Middle Ages. Their narrow and winding streets discourage intensive car use, unlike the grid-like street pattern of North-American cities. [...] Moreover, the spatial scale of cities in Europe is smaller than North-American cities. Combined with a spatial planning tradition that favors compact developments, European cities are more suitable for walking and cycling. Differences in culturally defined norms and values may influence travel behavior as well. For instance, Americans may be more inclined to move house in response to employment changes than Europeans." [340, p.340]

> "These cities cannot rely only on the evidence of those in developed economies,

> given large differences between them, such as infrastructure, socio-economic conditions and administrative capacities. Thus, this study identifies the main determinants of urban mobility in India and draws lessons for promoting effective transportation policies fit for Indian and other cities in developing countries." [8, p.107]

Initially, aggregated models were employed to assess the impact of land use characteristics on collective mobility behaviour in large geographical zones using aggregated data (e.g. [124]) but this approach has fallen out of favour due to an alleged oversimplification of the complex interactions of LUTI [198]. Additionally, research has shown that the built environment alone cannot account for individual differences in mobility behaviour – instead, economic circumstances and more subjective factors such as culture, preferences and attitudes have been recognised to have additional explanatory value [25, 198, 320].

The following sections will provide an overview of the geographical setting of these studies with a particular focus on work carried out in emerging economies and the different dimensions that were being investigated as causal factors for mobility behaviour.

2.3.1 Geographical Setting

Predicated by the availability of necessary data sources LUTI studies have traditionally focused on North America [25, 68, 69, 85] with some limited focus on other Western countries such as the UK [320], Austria [311], Italy [106] and the Netherlands [96, 285, 340]. The result is a scarcity in empirical studies on urban areas in emerging economies predicated by limited availability of fine-grained, accurate and up to date data in the Global South.

Among the first to analyse the relationship between land use and mobility patterns in an emerging economy were Vichiensan *et al.* (2007) [343], who found a strong relationship between the two when analysing urban railway development in Bangkok, Thailand using transportation forecasting models. They noted that the full set of necessary data for

their analysis was not available, and some of the factors were at incompatible levels of aggregation, requiring a switch to a coarser level of analysis.

Ahmad and de Oliveira (2016) [8] analysed modal choice, and out-of-pocket travel expenditure as a proxy for traditional metrics such as distance or time travelled using household survey data for the 98 largest cities in India. They found a difference in mass transit usage between small and medium-sized cities (preference for private motorised transport) and large cities (preference for public transit in favour of non-motorized transportation). The prior difference most likely due to the absence of sufficient transit infrastructure, the lack of non-motorized transportation due to city size in the latter.

Focusing on commuting patterns in Shirza, Iran, Etminani-Ghasrodashti and Ardeshiri (2016) [117] found that Residential Self Selection (RSS) has a significant impact on mobility behaviours, while the influence of street density, accessibility and other design measures was mixed.

Lin *et al.* (2017) [218] investigated the relationship between socio-economic characteristics and home/work location on commuting times in Beijing using multiple linear regression analysis. They found a strong relationship between the decentralisation of employment, RSS and commuting times with the extent varying depending on the extent of sub-urbanisation of the different employment sectors. Their research also suggests a correlation between income and education factors and commuting behaviour.

Another study focusing on China used CDR data as well as passive signalling data from China mobile combined with a POI dataset and Traffic Management Zone (TMZ) stats [258] to analyse urban OD flows in Hangzhou, China for a week in September 2015. They used an OD spatial econometric model with auto-correlation to analyse the relationship between socio-economic factors, the physical location of POIs, transit accessibility, and commuting distance on OD trip counts. They found a positive correlation between per-

manent population numbers, the number of crucial POI physically located in the area (used as a proxy variable for land use) and transport accessibility (measured as road lane length), and OD counts, while travel time was found to be negatively correlated with counts.

Recognising a lack of research in the analysis of the interaction between the dimensions of the built environment and mobility, Kandt (2018) [193] compared the cities of Sao Paulo, Istanbul and Mumbai with a particular focus on motorized transportation. Data was collected by Ipsos MORI through a sample of 1000 household surveys in all three cities. The survey covered socio-demographic "data on the usual socio-demographic variables gender, age, household size, socio-economic status, educational qualification and economic activity as well as household vehicle ownership" [193, p.727] as well as regular trips and their purpose and mode. Multinominal logit models of mode choice and Ordinary Least Square regression models of trip duration were used. The study found, that each city requires different policy interventions and priorities - Sao Paulo requires increasing public transit accessibility in Favelas, Istanbul should reduce car ownership while in Mumbai the car fills a gap left by a lack in support of public transit services.

Zhang *et al.* (2018) [377] focused on the prevalence of mass transit usage among the elderly in 274 rural and urban neighbourhoods in Zhongshan, China. They used National Household Travel Survey (NHTS), TMZ, land use and population data in their analysis. The binominal regression model used highlighted a strong relationship between the accessibility of an area, gender, age and positive attitudes toward public transit and mass transit uptake.

The only LUTI study on an African country was Poku-Boansi and Cobbinah (2018) [281], who investigated the impact of land use on mobility behaviour in the urban area of Kumasi, Ghana while providing a review of research into the link between urban travel, transport accessibility and land use in Ghana. They found a weak effect of land use on urban travel as areas experiencing a rapid change in land use had poor accessibility with

overall poor conditions of transit services, infrastructure and high levels of congestion.

2.3.2 Exogenous, Causal Factors

Each factor can be classified across one of three dimensions – spatial, socio-economic, personality [340]. Spatial and socio-economic factors are regarded as 'objective' factors, with those on the personality dimension being regarded as 'subjective' factors giving rise to an objective-subjective divide within the research community [303, 345]:

> "focusing only on spatial and infrastructural characteristics tends to disregard the reality of individual perception, evaluation and decision. On the other hand, too much focus on the individual can obscure the fact that an individuals travel behavior is still linked to objective factors such as urban form and infrastructure." as cited in [199, p.246]

Spatial dimension

As living, working, shopping and recreation are spatially separated activities, they induce the need to travel. The need to travel does not derive its utility from the trip itself, but rather from the inherent need to reach activity locations. As a result, the spatial dimension was identified as key to understanding the drivers behind mobility patterns [1] early on with initial work in the late 60s by Forrester (1969) [123] in his Theory of Urban Interaction. Since the 1990s, the spatial dimension has increasingly focused on assessing the impact of the built environment through three latent constructs first described by Cervero and Kockelman (1997) [69] as the three D's that were later expanded by Ewing and Cervero (2010) [119] to include distance to transit and destination accessibility.

Density factors include population density, employment density, and job accessibility [69]. Most studies have found a positive correlation between increases in density, public mass transit and activity-based mobility uptake, and a reduction in motorised trip numbers and car ownership due to reduced distances supporting the hypothesis

[1] For a comprehensive review of the built environment-travel literature until 2009 see [119] - focus on environmental changes, i.e. urban core to peri-urban and related impact on daily mobility patterns

that higher density increases the likelihood of activity locations being within reach [8, 178, 219]. Ultimately, however, the relationship depends on the type of density as industrial employment density results in shorter trips, whereas an increase in commercial and residential density results in the opposite [124].

Diversity factors encompass the mixture of activity locations available within an area and are commonly defined through an entropy index [2] to quantify the degree of balance across land use types [69, 70, 124]. Low values are associated with single-use environments, whereas a higher value indicates the availability of more diverse land use types such as commercial, business, industrial and residential [69, 119, 198]. The effects of diversity are similar to the effects of higher densities in terms of reduced motorised trip numbers, increase in public mass transit uptake and increasing activity-based mobility (e.g. walking and cycling) [69, 124].

Design factors include street characteristics, pedestrian and sidewalk coverage, average block size, residential parking and accessibility [69, 205, 320]. Accessibility, the distance to the nearest public transit point, in particular, was found to drive increasing activity-based mobility [69], increasing public transit uptake [198], and reducing vehicle miles travelled [119, 205]. While the overall distance and number of trips per tour decreased with high accessibility, it resulted in an increased average number of tours [205]. Similarly, Meurs and Haijer (2001) [241] found, that design characteristics have more of an influence on mode choice for shopping, social and recreational trips than work travel.

Socio-economic dimension

As mentioned above, daily mobility demand is derived from the desire to reach activity locations. The focus here is on individuals as one of three possible sets of 'actors' within the activity system with the other sets being firms and institutions [340]. Existing research has shown that the overall impact of the built environment as an explanatory dimension to mobility behaviour is reduced when accounting for socio-economic characteristics of

[2] as opposed to the lesser-used Herfindahl index

the study population as it cannot account for the means that individuals have at their disposal [320, 353]. The most commonly used factors include:

Household composition generally defined as the number of adults in a household [285, 340]. The number of working adults in a household was found to have an impact on car dependence and trip distance [69, 205]. Single people and couples were found to favour public mass transit over private motorised transport during mode choice while distances and travel times for work trips may be longer for work, than for non-work due to an absence of childcare responsibilities [320, 340].

Age of a household member or average age of residents in an area. Older people using private motorised transport were found to be more likely to have shorter trips [320]. The number of children, on the other hand, can have similar impacts on mobility behaviour to working adults [205].

Gender split, as women are more likely to use public transit [304]. As a result of the slower speed associated with public transit, they are likely to travel shorter distances than men [320]. Stead (2001) [320], however, noted that this is perhaps due to a generally lower income and employment in different job sectors, while car use is potentially higher for non-work trips than car use of men.

Car ownership as either endogenous (explained by socio-economic variables), or exogenous (explaining mobility behaviour) [124, 340]

Education level, Employment status and Income are sometimes interwoven as proxies for each other as higher levels of education open opportunities to work in more senior positions with better levels of income [340]. Studies have shown a correlation between high private motorised transport use, long trip distances and commuting times among highly educated, employed, high-income groups [205, 320].

Life cycle stages as key personal or family events such as house moves, changes in education or profession, or marriage [36]

Those socio-economic and demographic factors are sometimes combined into latent constructs such as household responsibility or social status.

Personality dimension

Traditionally, mobility was considered as derived demand with mobility behaviour assumed to be based on a cost-benefit analysis of different mobility options dictated by the built environment to reach activity locations [351]. Existing research has shown, however, that homogeneous groups with similar objective factors, show differences in 'subjective' factors such as attitudes and preferences toward mode choice, neighbourhood characteristics and lifestyles that can have an impact on mobility behaviour, *de facto* overruling traditional assumptions of derived demand [282, 303, 345, 351].

Kitamura *et al.* (1997) [198] were among the first to incorporate lifestyles as a factor in the analysis when investigating the relative explanatory power of land use, socio-economic and lifestyle factors for the San Francisco Bay Area. Their findings, that attitudes and lifestyles explain that highest level of variation, was later confirmed by Bagley and Mokhatarian (2002) [25] as well as Wee (2002) [351], who found that attitudes, preferences and lifestyles have a more significant impact on RSS and by extension travel demand than the spatial dimension alone. Past RSS choices were found to be an adequate predictor for current mobility behaviour in what Beige and Axhausen (2012) [36] termed 'state dependency', which can in itself result in a dissonance between actually selected and desired resident neighbourhood [345].

> "modal changes at the new residential location are not determined by the changed built environment but are an expression of a more adequate realisation of travel-related preferences which already existed latently at the previous residential location, but was not fully implemented due to constraints such as local accessibilities." [199].

While the above mentioned research has shown those factors to have an influence in certain circumstances, they were not considered as part of this thesis research but are instead

mentioned to provided a fuller background on factors and dimensions considered within land use and socio-economic – transport interaction research.

2.3.3 Mobility Factors

There are a factors that are commonly used to model mobility patterns to be explained through the exogenous dimensions discussed earlier:

Modal choice difference for either work and shopping trips was found to be different depending on the geographical setting. In the Pudget Sound Area, USA density explained much of the modal choice while diversity decreased in explanatory value once socio-economic variation was taken into account [124]. In Austria, on the other hand, socio-economic factors had significantly more explanatory value on mode choice than either factor within the spatial dimension [311].

Trip distance or vehicle miles travelled was used as a principle mobility factor by Krizek (2003) [205] and Stead (2001) [320] as it can be regarded as a proxy for several environmental impacts including energy consumption and emissions.

Out of pocket travel expenditure has been used as a proxy for more traditional mobility metrics such as distance or time travelled by Ahmad and de Oliveira (2016) [8] for mobility analysis in India.

Trip frequency, trip counts or OD travel flow measures the number of inbound-outbound trips for an area or between areas [219]. Ni *et al.* (2018) [258] were the first to explore CDR-derived OD flows at an aggregated level using econometric analysis.

2.4 Chapter Summary

This chapter described key data sets that enabled CDR data to be examined as a novel source of data for understanding human behaviour on a micro- and macro level. It further provided an overview of the common structure of CDR and MFS data as well as a

description of when they are created.

The chapter also provided an overview of existing research on land use and socio-economic – transport interaction within the Global South, highlighting a gap in wealth of research compared to the Global North. The final part of the chapter outlined the three dimensions used to explain mobility behaviour as part of land use and socio-economic – transport interaction research. Specifically, this focused on commonly used factors within the spatial dimensions defined through the three D's of density, diversity and design, socio-economic, and personality dimensions.

The following chapters examine the value in using CDR data to generate insights into the spatial dimension (in chapter 3), the socio-economic dimension (in chapter 4), and mobility behaviour (in chapter 5); the analysis of the interaction between land use, the socio-economic dimension and mobility patterns (in chapter 6); and the limitations of CDR and MFS data and ways to address those identified through the empirical analysis in Chapters 3-6. Chapters 3-5 will include their own literature review specific to CDR and MFS data analysis, while Chapter 6 will include a review of existing studies using SEM for land use – transport interaction analysis.

Chapter 3

Tracking Activity-Based Land Use

3.1 Chapter Introduction

The previous chapter examined the characteristics of CDR and MFS data, the geographical context of existing LUTI research, and introduced the different dimensions used to explain mobility behaviour. This chapter generates density and diversity factors explicitly from the spatial dimension through the analysis activity-based land use using automatically generated CDR data. The following research objective guides this chapter:

Research Objective 1: to examine whether differences in activity-based land use and density can be distinguished from behavioural patterns contained within CDR data.

Traditionally, land use encompasses land characteristics, ownership characteristics, and the socio-economic and activity-based use of the land itself [17]. As discussed in §2.3.2 the spatial dimension within the LUTI relationship is commonly analysed through the three latent constructs comprising of 3D's - density, diversity and design [69]. While density and design fall under land cover, the biophysical state of the land, diversity is associated with human behaviour and activity. The activity-based land use of an area is a proxy indicator for diversity within the area. High levels of diversity and density are generally associated with reduced mobility-demand due to the availability of nearby activity loca-

tions [69, 119, 124, 198]. Successful urban planning requires an understanding of land use within the area itself and by extension, accurate measurements to assess both progress and success of interventions.

This chapter takes an unsupervised approach to answering the research question as they have the advantage of not requiring any pre-existing zoning data, thus increasing their utility for land use classification in areas with poor ground truth data. Before detailing this approach traditional data sources are discussed in Section 3.2.1, ultimately highlighting numerous shortcomings which motivate the examination of alternate data sources. This is followed by a discussion of existing approaches to conducting Land Use Analysis using Mobile Network Data in Section 3.2.3. The research approach is discussed in Section 3.3 after a discussion of existing approaches (§3.2.3).

3.2 Literature Review

Traditionally, techniques to monitor land use have focused on manual surveys and satellite imagery [172, 368]. The following section will highlight key works in the analysis of land use using such data sources.

3.2.1 Traditional Data Sources

The majority of previous work on land use classification has been conducted using official statistical data and satellite imagery with more recent approaches focusing on the potential of social media data.

Official statistics

Statistical sources are the most widely used source of data for the tracking of ownership characteristics, socio-economic and activity-based land use. Statistical sources include spatially-referenced population data collected in the form of census surveys and population registers. Population registers are becoming more commonplace in European

countries such as the Netherlands, Finland, Sweden and are more routinely updated than census survey responses. In rare cases, additional data sets such as social surveys, hospital patient registers, tax records and other administrative and industry collected customer data, including loyalty card and customer survey data is used [213].

Ineffective or absent data collection and governance strategies, the 'Statistical Tragedy', contribute to a lack of up to date, fine-grained and reliable data, however. Effective data collection and governance requires technical know-how and infrastructure for dissemination, collection, processing and quality control. Emerging economies also

> "face the additional challenges born from a lack of prior measurement of housing or infrastructure, significant population heterogeneity in literacy levels and languages spoken, the presence and location of nomadic people" [232, p.47] and presence of informal slums.

Generally, census surveys are only carried out every couple of years, with the gap increasing rapidly in emerging economies around the world. In Tanzania, census surveys were carried out in 1988, 2002 and 2012 [1]. The result is a lack of temporal and spatial resolution that makes official statistics unsuitable for capturing changes in land use dynamically.

Satellite Imagery

An alternative data source allowing for the collection of land use and land cover data at different temporal and spatial scale is satellite imagery. Satellite imagery are images of the earth surface recorded in different spatial (recording size), spectral (wavelength), temporal and radiometric (bit-depth and brightness) resolution [2]. Advantages of satellite imagery include global coverage, a high revisiting capability and relative ease of access, making it the most pertinent data source for tracking land cover change over time [115]. Yuan *et al.* (2005) [372] for example used reflective spectral bands of satellite imagery to identify land characteristics and classify the study area by land cover types including

[1] https://unstats.un.org/unsd/demographic-social/census/censusdates/

[2] https://www.sciencedirect.com/topics/earth-and-planetary-sciences/spatial-resolution

urban land, agriculture, forest, grass, water, and wetland. Similarly, Aburas *et al.* (2015) [1] have used satellite imagery to identify remote land cover features, including the normalised difference vegetation index in Malaysia. Land cover information can be useful for filtered area weighting during spatial interpolation and apportioning as will be discussed in more detail in §7.5.2.

Through a combination of spectral and texture features contained in the imagery, satellite imagery can be used to identify a range of design and density factors including building locations, building density, road conditions, and proximity to the physical location of POI such as schools and hospitals [373]. Research by Wu *et al.* (2009) [368] has shown, that design and density factors alone are insufficient to reliably differentiate between different land use types such as office, industrial, transportation unless additional data sets are used. Instead, they proposed using additional data sets on contextual properties, while others have chosen to use census data [240], Geographic Information System (GIS) data sets [172, 371] or expert knowledge [370]. Hu and Wang (2013) [172] used aerial imagery taken from a single-engine aircraft, land parcel boundary and land use layer data for land use classification in Austin, Texas. They found building footprint and height to be the most important factors for land use classification. Their study relies on pre-existing ground reference data, which may be incorrect or missing in other countries.

While satellite imagery has shown much promise as a data source for land cover classification, its sole focus on biophysical characteristics, density and design have limited its utility for the identification of diversity features from the spatial dimension (§2.3.2) of a city.

> "Another reason for the unsatisfactory accuracies is the definition of the land-use scheme. Remote-sensing data mainly depict urban biophysical characteristics, while the land use definition is based on the human use perspective. There is a mismatch between urban land cover and urban land use, so the diversification of the utilised characteristic of civic land use and the human

> operation is no doubt the main reason for the low accuracy" [172, p.800]

Empirical evidence has shown these approaches to be time-consuming, expensive, infrequent, and quickly outdated however.

3.2.2 Social Media

The proliferation of Global Positioning System (GPS) equipped mobile devices and basic cartographic design skills have allowed for the increasing generation of geospatial data. The generation of this data was previously restricted to professional with specialist equipment [51, 127, 143, 319]. By the millennia,

> "the cost of entry into map-making had fallen to no more than the cost of a simple PC, and the role of the expert had been replaced by GPS, mapping software, and other technologies" [144, p.232-233].

The proliferation of GPS equipped mobile devices, and the increasing pervasiveness of the internet has given rise to the development of Location Sharing Services (LSS) such as Twitter, Facebook, Foursquare and Flickr. The services generate vast amounts of geolocalised user-generated content or Volunteered Geographic Information (VGI). Official land use categories are strongly related to patterns of human activity. Human activity data generated through LSS are thus a promising avenue for overcoming limitations of official statistics and satellite imagery [130, 221, 261, 276, 295, 330, 350]..

Foursquare is a search-and discover based LSS. Its purpose is to provide recommendations on the physical locations of POI and local businesses such as restaurants, cafe's, business.

Twitter is a microblogging and social networking LSS. Using Twitter data for 49 days, Fras-Martnez *et al.* (2018) [130] sought to both determine land uses for specific parts of Manhattan and urban POI with high activity levels validated against official land use and landmark statistics. Garca-Palomares *et al.* (2018) [132] sought to analyse the link between land use and daily dynamics of the city of Madrid using 'typical

Twitter activity profiles', TMZ land use and Cadastral data. The approach heavily relied on pre-existing land use information only available at the TMZ level. They found that while the approach had some utility that other techniques, particularly remote sensing through satellite imagery, are "more reliable and easier to apply in order to obtain land use maps" [132, p.311].

Flickr is an image and video hosting LSS. Focussing more on landscape descriptions, Wartmann *et al.* (2018) [350] used on-site interviews, text data from hiking blogs and Flickr tags to classify 10 study sites across Switzerland. Terroso-Saenz and Munoz (2020) [330] combine user-centred text (Flickr) and venue-centred (Foursquare) data to train a Random Forest classifier with land use categories based on the Foursquare hierarchy for New York and San Francisco. Random Forest was used as the base model for the construction of a multi-label classifier for the association of discussed topics (generated through Latent Dirichlet allocation (LDA)) with pre-defined land use labels as it outperformed Logistic Regression and Support Vector Machine (SVM) for the classification task at hand. Land use areas were identified as topics from point-based data contained in the user-centred text corpus using *HDBSCAN*. They found that the utility of this approach diminishes as the document density decreases toward peri-urban 'outskirts'.

Challenges of using social media data for land use classification include: a lack of interoperability due to the vernacularity of the generated land use labels; and the under-utilisation of available data sources as the focus is generally on user-centred text LSS rather than a combination of different data sets. Unless land use labels are derived from specific pre-defined taxonomy of categories such as those used in Foursquare [3], VGI is inherently vernacular affecting their utility for coherent classification [51, 144].

"On the one hand, most proposals use different information retrieval and topic modelling techniques in order to analyse the textual information contained in VGI documents (e.g., the text written by a user in a tweet or the tags

[3] https://developer.foursquare.com/docs/resources/categories

> describing a photo on Flicker). Next, the outcome of these techniques is used to label the regions of the target city describing their usage. This procedure tends to generate ad-hoc labels which actually hamper the interoperability of the generated land-use maps." [330]

Furthermore, VGI is much more likely to be generated in dense urban areas and those frequented by tourists, affecting its utility for land use classification in peri-urban and rural areas. The always-recording nature of many LSS also raises concerns regarding security, confidentiality and personal privacy [32, 111].

3.2.3 Land Use Analysis Using Mobile Network Data

A potential alternative, however, exists in the analysis of automatically generated MND in the form of CDR transaction logs. Work on the remote identification of land use areas using CDR data at the BTS level has been undertaken in several countries. MND based land use detection generally follows a three-step process of signature construction, area division and clustering. Signatures for subsequent classification and clustering are derived from CDR event series and can be represent as time series of different temporal granularity [233, 273, 315, 316, 336, 341, 369], different communication variables [234, 232, 260] or matrices [113, 114, 213, 226, 234]. Area division approaches for study region identification, and visualisation of land use maps have been on the grid-level [113, 114, 213, 273, 336] and polygon-level [226, 232, 233, 234, 260, 315, 316, 341, 369]. Previous classification approaches can be divided according to the applied technique: semi-structured, supervised, and unsupervised classification discussed in more detail below. An overview of studies using MND for land use classification can is shown in Table 3.2.3.

Semi-structured approach

In Semi-structured learning, a small set of known POI is used to train a land use classifier. Soto and Frias-Martinez (2011) [316] used semi-structured fuzzy c-means clustering in their analysis of land use in Madrid. They used network event activity counts for one month from October 1st to 31st, 2009 in 5-minute intervals in the form of Erlangs

averaged to different signature levels: one day of the entire study period, one weekday, one weekend and a week pattern. The areas are divided into polygons using Voronoi tesselation. They found that the use of fuzzy c-means allows for the representation of land use areas in terms of the balance across land use types.

Merging one month of CDR with Foursquare data, Noulas *et al.* (2013) [260] generated a range of communication-based variables rather than time series signatures for land use classification in Madrid and Barcelona. Study areas were dynamically generated using *DBSCAN* for merging of BTS near one another before being divided into polygons using Voronoi tesselation. They compared logistic regression, SVM, a multi-layer perceptron, logistic model trees and *DMNBText* for classification of areas based on Foursquare hierarchy derived labels. While binary class accuracy was at 65%, additional classes led to a sharp drop in performance to below 50% once more than four classes were used.

Also using fuzzy c-means clustering, Pei *et al.* (2014) [273] classified land use areas in the Singapore metropolitan area. Using Erlang for one week from March 28th to April 3rd, 2011 for 5500 BTS in Singapore, they constructed normalised signatures for a 4-day mode - general weekday consisting of Monday to Thursday, Friday, Saturday and Sunday. Areas were aggregated to $200m \times 200m$ grids using inverse distance weighting. Classification accuracy versus ground reference labels from the Urban Redevelopment Authority in Singapore was 58.03% at best. The authors suspect that this error is due to more considerable heterogeneity in land use.

Combining SET1 of the D4D challenge Senegal (§2.1.1) with POI data sets obtained from OpenStreetMap and Facebooks Graph API, Mao *et al.* (2017) [233] examined Commercial/ Business/ Industrial, and residential areas in Dakar, Senegal. They generated average hourly time series for a week from 1 year of hourly BTS counts for 1666BTS in Senegal from January 1st to December 31st, 2013 in addition to a 'spatial residual' variable. Areas were represented at the Polygon-level following Voronoi tesselation and

Simple Area Weighting. Areas were classified as either Commercial/ Business/ Industrial, or residential based on the ranking of Non-negative matrix factorization (NMF) vectors.

Supervised approach

Supervised approaches classify areas by land use using classifiers trained on pre-labelled ground truth datasets. Existing research using supervised approaches are scarce due to the lack of accessible ground truth data in many countries around the world.

Toole *et al.* (2012) [336] used Zoning labels from the *MassGIS* aggregated to 5 categories (residential, commercial, industrial, parks, other) to train their land use classifier. They used CDR for the Boston Metropolitan area for 600k subscribers for three weeks to generate average hourly time series without accounting for weekday/weekend differences. Those are fundamentally different due to the absence of 7-day working weeks [63]. Random Forest was used for classification of time series with areas as interpolated as $200m \times 200m$ grid cells due to the difference in spatial granularity from triangulated point-based CDR and tract-level zoning labels in the *MassGIS* data. They found that outdated zoning data used for building models was the prime reason for the misclassification of the majority of areas.

Vanhoof *et al.* (2017) [341] used CDR data for 154 days from May 13th to October 14th, 2007 to classify land use in various French cities using French Urban Area Zoning data, which is updated every 5 to 10 years. They generated four average one-day scenarios of non-summer weekday, non-summer weekend, summer weekday and summer weekend. Areas were divided into polygons using Voronoi tesselation and Spatial autocorrelation with Voronoi circumference used as an additional variable to the four generated scenarios. Three (Random Forest, Boosting trees and elastic net penalised logistic regression) classifiers were used to classify urban areas.

Unsupervised approach

Both supervised and semi-structured approaches have shown only average accuracy due to inaccurate ground truth data used to train classifiers, and high heterogeneity in land use. Unsupervised approaches have the advantage of not requiring any pre-existing zoning data, increasing their utility for land use classification in areas with poor ground truth data. Instead, land use categories can be presented once clusters are formed, and categorised spaces can then be compared to known locations of specific land use categories for validation. This enables a data-driven classification of the traditional aspect of diversity based on summaries of emergent groupings. It is of note that such an approach in general does not guarantee clusters will distinguish or classify regions based on emergent features of utility in the problem domain. However, in Section 3.4 it is shown empirically that at least in the case of land use classification the approach does lead to clusters that are distinguished by interpretable features with high utility.

Among the first to apply an unsupervised approach for land use classification was Soto and Frias-Martinez [315], who applied K-means to Erlang data from Madrid in a similar process to their other paper the same year [316].

Lenormand *et al.* (2015) [213] used CDR data for 55 days days from September to November 2009 for land use classification in the Spanish cities of Madrid, Barcelona, Valencia, Seville and Bilbao using a community detection approach. Signatures were generated in the form of hourly time series for weekdays with areas represented on Voronoi interpolated with $500m \times 500m$ grid cells. The classification was performed using a community detection algorithm on the Pearson correlation matrix between cell activities.

Madhawa *et al* (2009) [226] used the principle components from hourly time series for CDR data for Colombo, Sri Lanka, generated through Principal Component Analysis (PCA) for classification. Areas were divided into polygons using Voronoi tesselation of BTS and clustered using K-means.

Xing *et al.* (2018) [369] is following the same approach as [131] using traffic volume and BTS time series on mobile data usage for four weeks from May 18th to June 14th 2016 in a north-eastern Chinese city. BTS under 100m were merged and subsequently represented as polygons generated through Voronoi tesselation resulting in 1143 analysis areas ($n = 3489$ pre-merging). Three BTS-level signatures in the form of hourly time series for weekday-weekend, median week and weekday-weekend median were generated for clustering using K-means++. Validation was undertaken using a distribution-based approach, with POI assigned to Voronoi polygons. Here, NMF was used for quantification of the land use mixture.

Using K-means, Manley and Dennett (2018) [232] generate variables on regional activity and regional interactions from D4D Senegal (§2.1.1) data to classify polygon-level land use by 11 classes. The regional activity was based on activity generated in the area akin to time series used by other studies described above, with regional interactions measured as the proportion of calls between an origin and a destination region. A building use data set from the National Agency for Statistics and Demography of Senegal is used to assess density and land use mixture within the study region. The 11 generated classes were predominantly based on the relative density of certain land use categories such as industrial, hotel and military.

Table 3.1: Overview of research on land use classification using Mobile Network Data

Reference	Data Source	Sample Size	Time Period	Signature Construction	Region identification	Area division	Clustering Method	Target Labels	Region
Soto and Frias-Martinez (2011) [315]	CDR	100m CDR	1 month	5min Erlang Time series (Total—total week-day/weekend—Daily)	Voronoi tesselation	Polygon-level	K-means	Ad-hoc labels	Madrid
Soto and Frias-Martinez (2011) [316]	CDR	100m CDR	1 month	5min Erlang Time series (Total—total week-day/weekend—Daily)	Voronoi tesselation	Polygon-level	Fuzzy C-means	Ad-hoc labels	Madrid
Toole *et al.* (2012) [336]	CDR & *MassGIS* Zoning labels	600k subscribers	3 weeks	Average hourly Time series	Grid interpolation of CDR (triangulated) and Labels (tract)	200m Grid-level	Random Forest	5 *MassGIS* Zoning labels	Greater Boston Area

Noulas *et al.* (2013) [260]	CDR & Foursquare	12m subscribers	1 Month	User communication entropy, outgoing tower entropy, remote communications, weekend calls, night time call volume, durations, user return times	DBSCAN for nearby BTS aggregation & Voronoi Tesselation	Polygon-level	Logistic Regression & SVM &Multi-layer Perceptron & Logistic Model Trees & *DMNBText*	Foursquare hierarchy	Madrid and Barcelona
Pei *et al.* (2014) [273]	Erlang & URA Zoning labels	5500BTS	1 week	Vector of normalized pattern and Erlang	IDW	200m Grid-level	Fuzzy C-means	5 URA Land use types	Singapore Metro Area
Lenormand *et al.* (2015) [213]	CDR	unknown	55 days	Pearson correlation matrix	Voronoi tesselation	500m Grid-level	Community Detection	Ad-hoc labels	Madrid, Barcelona, Valencia, Seville & Bilbao
Madhawa *et al.* (2015) [226]	CDR	10m subscribers	1 month	Hourly time series derived PCA	Voronoi tesselation	Polygon-level	k-means	Ad-hoc labels	Colombo, Sri Lanka

Mao *et al.* (2016) [234]	D4D & Facebook	477 BTS	365 days	Hourly time series derived NMF matrices, Call density, Connectivity, PageRank	Voronoi Tesselation	Polygon-level	Jaccard Coefficient	Ten-class land use	Dakar, Senegal
Engelmann *et al.* (2017) [113, 114]	CDR	415k subscribers, 433.6m CDR, BTS	122 days	Hourly time series derived NMF matrices	Voronoi tesselation	500m Grid-level	k-means	Ad-hoc labels	Dar Es Salaam, Tanzania
Mao *et al.* (2017) [233]	D4D, OSM & Facebook	488 BTS	365 days	Hourly time series (168) &spatial residual	Voronoi tesselation & Simple Area Weighting	Polygon-level	NMF	2 Ad-hoc classes	Dakar, Senegal
Vanhoof *et al.* (2017) [341]	CDR & ZAUER TMZ	18m subscribers	154 days	Hourly time series (weekdays summer/non-summer — weekends summer/non-summer), Voronoi circumference	Vornoi Tesselation & Spatial autocorrelation	Polygon level	Random Forest, Boosting Trees, Elastic-Net penalized logistic regression	Downsampling of ZAUER labels (9 to 6)	France
Xing *et al.* (2018) [369]	CDR & POI	Unknown	4 weeks	Hourly time series	Voronoi tesselation	Polygon-level	K-means++	Ad-hoc labels	North-eastern Chinese City

Manley and Dennett (2018) [232]	D4D, OSM & ANSD	Erlang	7 months	Regional activity & Regional Interactions	Voronoi Tesselation	Polygon-Level	K-means	GIS derived 11-class Ad-hoc labels	Dakar, Senegal

3.3 Research approach

The analysis of land use focused on BTS located in the metropolitan area of the Tanzanian port city of Dar es Salaam. It goes beyond other unsupervised studies by considering activity signature (§3.3.2) activity for weekdays and weekends while accounting for heterogeneity (§7.4.3) in usage patterns between those. Unlike earlier research by Madhawa *et al* (2009) [226] and Soto and Frias-Martinez [315], signatures were constructed based on the network events rather than unique number of users, and Erlang respectively. An unsupervised approach was chosen for this research, as while some form of ground truth data that will be discussed in more detail in §4.3.1 was available, this is rarely the case in other emerging economies making the approach less generalisable otherwise.
A core dataset of 565 BTS from Dar es Salaam, Tanzania, was identified. The analysis followed a three-step process similar to other approaches discussed in the previous section and by Xing *et al.* (2018) [369], who published following the completion and publishing of research (same as K-means, Manley and Dennett (2018) [232]) described in this chapter:

1. **Signature Construction:** Raw CDR event series are aggregated into different activity signatures representing hourly network event counts for each BTS. As part of this step, BTS in very close spatial proximity where the location of a user cannot be distinguished are merged. Three signatures are generated for detection of BTS with unclear operational status. Feature scaling was used to standardise time series for comparison. NMF and PCA were applied to the time series to extract matrices of latent factors and principal components.

2. **Clustering:** K-Means was applied to the NMF matrices to cluster areas with similar underlying behavioural patterns.

3. **Area division:** Polygons representing Voronoi shapes around BTS were generated using Voronoi tesselation. Labelled Voronoi polygons are interpolated with a grid to protect individual and commercial privacy.

3.3.1 Data Description

The CDR data used as part of this study covers a total of 433.6 million CDR records for call and SMS network events generated by 415k mobile phone subscribers across the metropolitan area of Dar es Salaam over a period of 122 days in the autumn of 2014 [4]. As discussed in §2.2.1 a raw CDR record is automatically created for each network event (§2.2.1) and includes a range of attributes including: timestamp of when the event occured; an anonymised mobile phone subscriber ID; call duration; and a BTS identifier.

3.3.2 Signature Construction

First, time series representing hourly activity at each BTS were generated to represent activity signatures. Second, time series were feature scaled for comparative purposes. Third, the obtained relative hourly time series were averaged per hour of the day resulting in activity profiles or 'signatures' for each BTS.

Timeseries

A series of network events were extracted for each BTS located in the Dar es Salaam Metropolitan area. The event series for each BTS were aggregated as hourly time series with each bin representing the number of network events generated each day at every hour. Timeseries account for the internal structure of longitudinally collected data including diurnal patterns of day/night, working day/weekend and seasonal variation between summer and non-summer months [63, 341].

Feature scaling

The network event counts within the per BTS time series differed quite significantly across the study area, making a direct comparison and clustering difficult. Per BTS event counts for signature2 ranged from 4424 to over 2.4 million network events. In order to compensate for the differences, time series were standardised before signature construction prior to

[4] Due to both individual and commercial privacy, the anonymised data used as part of this study is not publicly available, and was provided through a partnership with a major Tanzanian MNO

clustering to match shape rather than magnitude [290]. There are three common methods for standardisation:

- Standardisation uses the mean (μ) and standard deviation (σ) to calculate Z-scores
- Mean Normalisation uses the mean, min and max to redistribute values around a mean μ of zero
- Min-Max Scaling uses the minimum and maximum values to standardise data between zero and one.

Mean normalisation and standardisation to z-scores was not possible in the present case, as NMF does not allow for the input matrix to contain any negative values. Instead, Min-Max scaling was used:

$$y_i = \frac{x_i - x_i^{min}}{x_i^{max} - x_i^{min}}$$

v_i is the normalised version of the original timeseries x for BTS i. x_i^{min} is the minimum and x_i^{max} the maximum activity count for any given BTS in the sample respectively.

Signatures

Each BTS signature is represented through a feature vector of feature-scaled timeseries v_i. Signature 2 for example can be represented as:

$$v_i(1), v_i(2), ..., v_i(168)$$

The combination of all signatures results in a matrix $\mathrm{V_{m \times n}}$ with m=565 as the number of BTS and n=168 as the length of a signature. Some time series showed a uniform distribution of activity during day time, while others showed a higher network activity pattern either at evening times only or at both morning and evening times. In total, three sets of signatures representing different temporal granularities were generated per BTS for both analysis and outlier detection:

- **Signature1** time series hours were averaged into a single day (n=24). One BTS was excluded from the analysis as its signature only contained a single recording.

- **Signature2** time series hours were averaged into a single week (n=168). BTS with fewer than 72 recordings in its signature were classed as outliers. 72 was chosen as the cutoff as it is equivalent to 3 weekdays worth of data to account for BTS being switched off throughout the weekend with some tolerance for Friday evening and Monday morning. As a result, an additional BTS with only 73 out of 168 recordings made between midnight and 6 am was excluded.

- **Signature3** represents the original time series for 122days (n=2928). BTS with fewer than 1296 recordings were treated as outliers. 1296 recordings were chosen as the cutoff as it is equivalent to 4 days per week worth of data to account for towers being switched off throughout the weekend with some tolerance for Friday evening and Monday morning for the entire study period similar to the approach for Signature2. This lead to the exclusion of an additional 13 BTS.

Table 3.2: Aggregate statistics of activity signatures prior and post outlier removal

Signature	Temporal granularity	min prior	min post	max prior	max post	avg prior	avg post	σ prior	σ post
Signature1	24=24x1	1	16	24	24	23.918	23.959	1.063	0.446
Signature2	168=24x7	1	115	168	168	166.758	167.218	9.352	4.811
Signature3	2928=24x122	1	1210	2928	2928	2778.71	2838.467	439.508	247.934

The removal of outliers had a drastic effect on the overall signature matrix V, as can be seen in Table 3.3.2. The σ decreased by 58.04% (Signature1), 48.56% (Signature2) and 43.59% (Signature3) respectively. Signature2 is used for subsequent classification of land use. In contrast to signature1, it captures differences in weekday and weekend usage patterns and was used to understand common behaviour within the area surrounding the BTS over the course of a week to address the problem. Additionally, it is of a reduced dimension compared to Signature3. In order to further address the curse of high

dimensionality [138], reduce noise, and the overall feature space of the covariance matrix V, factorisation is required [4].

3.3.3 Factorisation

Feature extraction and dimension reduction techniques can be used to uncover hidden structures in large data sets. Techniques such as PCA, LDA and NMF have previously been applied in fields such as facial recognition [210], document analysis [210, 134] and bioinformatics [102]. In this research, NMF and PCA based on the scikit-learn Python library implementation were used to decompose the input matrix V containing the signatures and identify latent features occurring in weekly usage behaviour. Both NMF and PCA factorise an input matrix V into two smaller matrices W and H with k dimensions:

$$V \simeq W_{\mathrm{n\times r}} H_{\mathrm{r\times m}}$$

with W as the weight matrix, H as the basis vectors, and r as the number latent features (NMF) or principal components (PCA) to extract. When multiplied together, the basis vectors represent an approximation of the input matrix V. Each row of the matrix W indicates the strength of association between the input items and latent features. Descriptions of the corresponding topics can be generated by ordering columns and selecting top-ranked latent features or principal components. If the input is an item-value matrix instead of a value-item matrix, the interpretations of W and H are reversed. NMF and PCA differ in the constraints placed on the weight matrix W and the basis vectors H [210].

PCA is an unsupervised data decomposition approach for feature learning. "PCA constrains the columns of W to be orthonormal and the rows of H to be orthogonal to each other" [210, p.789]. With PCA, matrix V is approximated through a linear combination of all the basis vectors H with negative values and subtractions allowed within the resulting principal components. As a result, many principal components may lack intuitive meaning, as was the case with the resulting components as can be seen in Figure 3.3a.

NMF uses a group of unsupervised algorithms based on linear algebra to perform dimension reduction. While LDA is a probabilistic model for expressing uncertainty about a range of topics associated with each item [44], NMF is a deterministic algorithm arriving at single representations [41, 210]. NMF requires all values in the matrices W and H to be positive numbers. In contrast to PCA, NMF learns a parts-based representation of the input signatures. Through the non-negativity constraints, latent factors of NMF represent parts-based representations. NMF was chosen over LDA due to its deterministic nature. The previously generated covariance matrix of signatures $V_{550BTS \times 168 \text{ Observations}}$ was used as the input matrix.

3.3.4 Area Division

Each of the 550 BTS located in the metropolitan area of Dar es Salaam was represented on the polygon-level. The coverage area of each BTS is approximated through Voronoi polygons generated using Voronoi Tesselation in Qgis, resulting in an average study area size of approximately $6.23km^2$. Voronoi polygons were assumed to represent the service area of a given BTS and thus the area described by a BTS's signature. In order to protect individual privacy (§7.6.2) and commercial interests (§7.6.4) of the MNO that provided the data areas were further interpolated with a $1km \times 1km$ grid shown in Figure 3.1 following clustering discussed below.

3.3.5 Clustering

With these latent features and principal components in hand and each time series from Signature2 projected into the lower dimensional space they represent, K-means clustering technique was applied to identify clusters with similar underlying behavioural patterns. As discussed in §3.2.3, the use of supervised and semi-structured approaches requires the availability of (accurate and up to date) ground truth data for the training of classifiers. An unsupervised approach was chosen for this research, as while some form of ground truth data that will be discussed in more detail in §4.3.1 was available, this is rarely the case in other emerging economies making the approach less generalisable otherwise. Here,

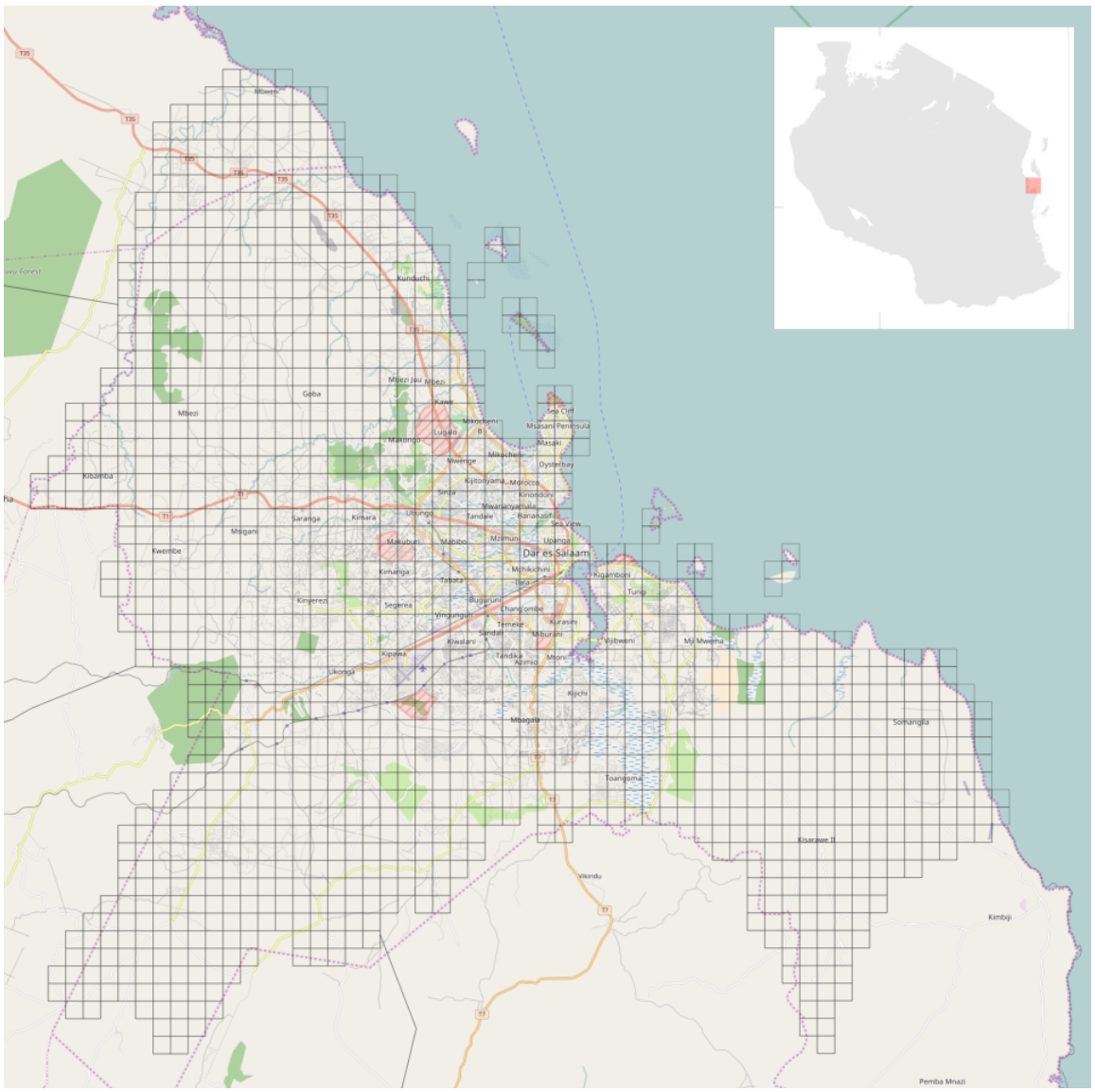


Figure 3.1: $1km \times 1km$ grip map of Dar es Salaam used to protect individual privacy and commercial interests

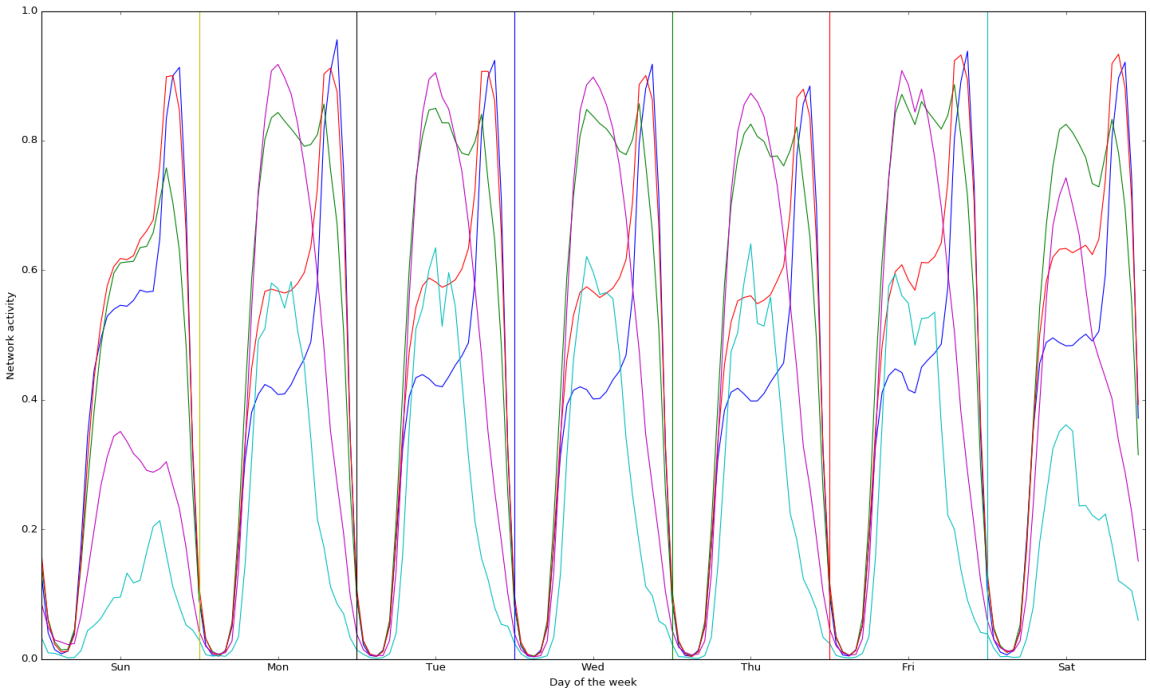
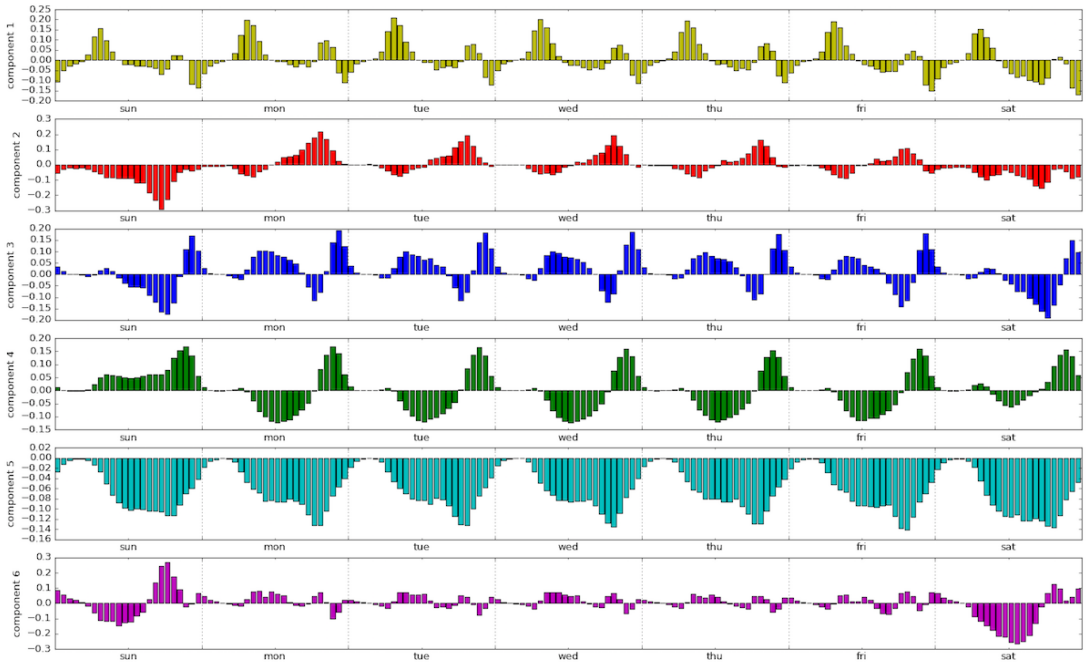


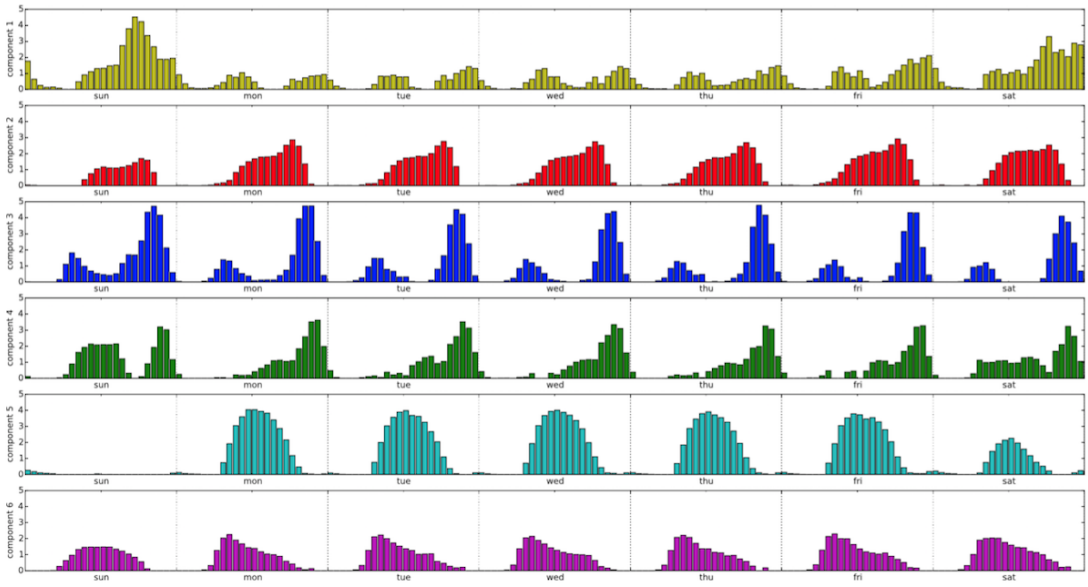
Figure 3.2: Signature 2 of Centroids for each of the clusters identified through K-means.

the dataset was used for verification of labels assigned to the identified clusters. Labels are primarily for the benefit of practitioners and to provide a qualitative understanding whether the differences in behaviour corresponds to differences in land use as it has traditionally been measured.

Similar to existing research [226, 232, 315] K-means clustering was used for land use classification following the creation of latent feature spaces in the previous section. K-means is an iterate spatial proximity-based clustering technique, which refines initial cluster centroid estimates into final centroids in an attempt to minimise the sums of squares of distances from each data point to the nearest cluster centroid. Unlike existing studies, clustering was performed on the latent features rather than the original signatures. The choice of n remains an arbitrary one dependent on the task in hand. As such, the number of clusters n was varied from 2 to 15 clusters. Silhouette scores were considered and an n of 5 was chosen. The signatures of the centroids for each of the resulting clusters are illustrated in Figure 3.2.



(a) Principal components



(b) Latent features

Figure 3.3: Six trends extracted via PCA (a) and NMF (b) from the CDR data. Each describes a different underlying population behaviour, which form the building blocks for our activity based land use clustering approach. With the exception of component 5, the latent features generated through NMF are far more intuitive to interpret than the principal components of PCA.

3.4 Results and Discussion

Manual analysis of the latent features and principal components extracted in §3.3.3 allowed for the generation of land use labels for ascription to the distinct groups identified through clustering. Empirical evidence suggests that the different behaviour components illustrated in Figure 3.3 distinguish between usage patterns that correspond to more traditional concepts of land-use. As aforementioned, labels were generated for the benefit of practitioners and to improve qualitative understanding. These are informative of population behaviours, with common weekly trends being revealed. Component 2, for example, reveals general underpinning network activity patterns (and is very similar to the average weekly time series for all towers) with a gradual increase from 7 am until 10 am, plateauing out before an early evening spike in network events. Component 3, however, reflects a predominant residential activity pattern, with population leaving an area in the middle of the day, and returning after work. In contrast, Component 5 indicated workplace behaviour, with high daytime activity, zero nighttime events, and a significant weekend drop. Any particular area may be composed of a combination of land uses (for example, half residential and half industrial), and so may express each of these building block behaviours to a different amount. These components provide a vocabulary through which those combinations can be discussed, without need for imagery or demographic data. Due to the presence of negative loadings, the principal components extracted using PCA (Figure 3.3a were found to be far harder to interpret than the latent features extracted using NMF (Figure 3.3b). The components are not expected to represent land use behaviour but rather serve as indicators to understand what cluster summaries, generated through K-means, potentially represent.

Cluster 1 - Affluent-Commercial: consistent activity throughout the day (activity spaces that bring people in due to tourism, job opportunities, amenities, etc.).

Cluster 2 - Slum: characteristic of a poor demographic with lower daytime activity, low morning activity and significant peak in the early evening (perhaps due to lack of mobility).

Cluster 3 - Residential-Commuting: this profile expresses a far higher expression of component 3 (the residential activity pattern) than other behaviours, suggesting a commuting pattern.

Cluster 4 - Industrial: high expression of component 5 (non-residential). Some commuting, but a highly significant lack of mobility activity in the mornings, evenings and weekend.

Cluster 5 - Formal-Night-Active: average activity over the day, but with significant spikes in the evening and night.

The map in Figure 3.4 shows a plot of the spatial distribution of these clusters for the centre region of Dar es Salaam. In order to protect the commercial interests of the network operator that provided the data, BTS catchment areas were interpolated with an overlaid $1km \times 1km$ grid-cell representation. Purple grid cells represent 'Affluent-Commercial' regions, green areas are identified as informal areas or 'Slums', yellow areas represent those classed as 'Formal-Night-Active', red areas are identified as 'Industrial' with blue areas classed as 'Residential-Commuting'.

3.4.1 Study Limitations

Representativeness and sub-sample bias CDR data is restricted to subscribers of the network operator providing the data, which represents only a sub-sample of the entire population in a country. It is not guaranteed, that the subset of mobile phone subscribers used as part of this study evenly represents population samples across all of the different areas.

High diversity in land use Many areas of Dar es Salaam are highly diverse areas with different land uses in very close proximity, as seen in Figure 3.5. Ahas *et al.* (2015) summarised this diversity in a study of urban activity in Harbin, China; Tallin, Estonia; Paris, France:

> "Cities not only contain monofunctional areas such as suburban 'sleeping'

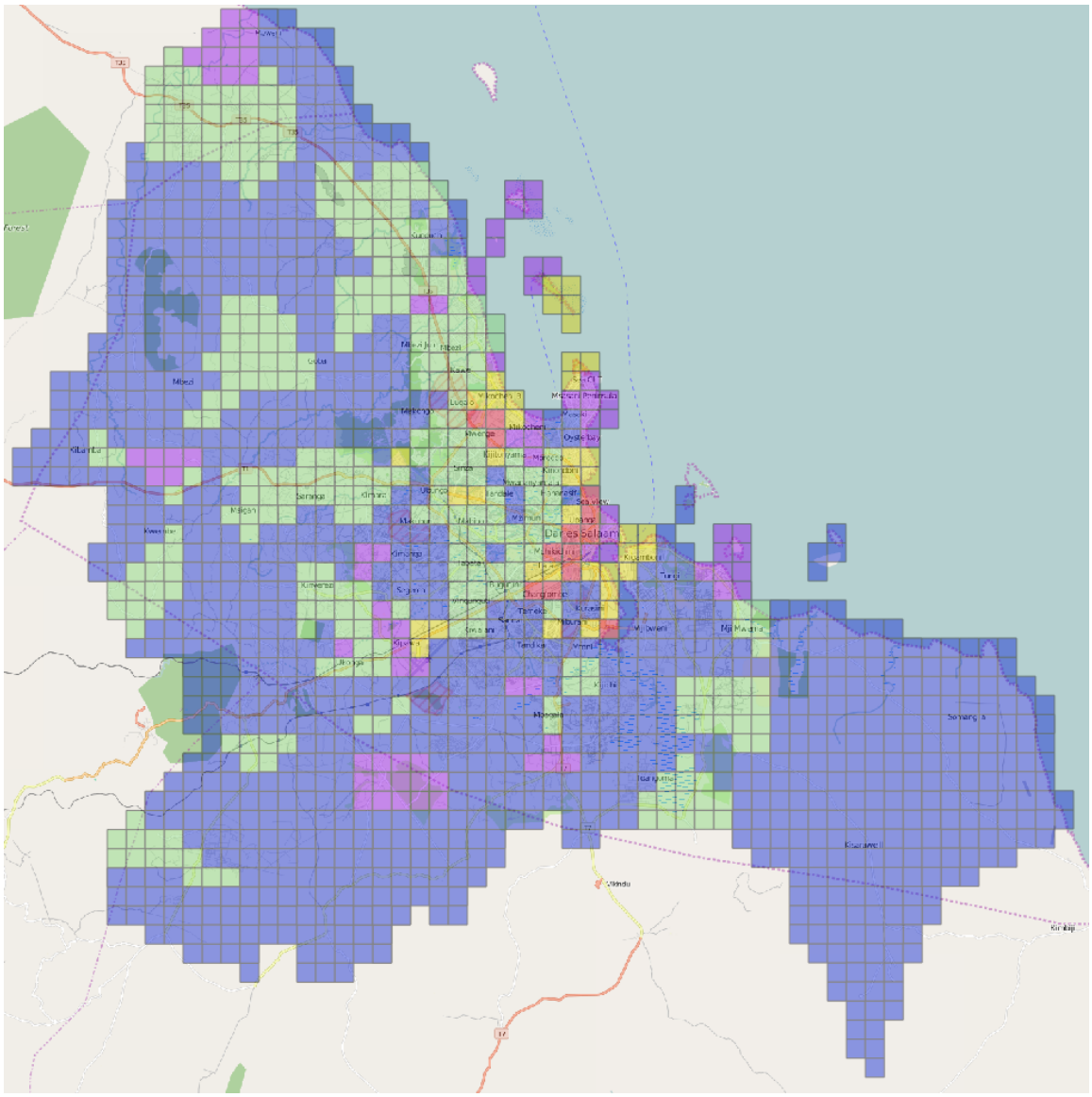


Figure 3.4: Spatial distribution of activity-based land use areas across the metropolitan area of Dar es Salaam region.

Figure 3.5: Sample area within Dar. Residential, Commercial and Industrial activity is condensed in a small area of high diversity

> districts and industrial work areas but also include many multifunctional areas as well with various activities entwined in the same buildings and/or districts. For example, some people may work at home while conversely many household activities are taken care of at work during the day." [7, p.2021]

Non-uniform BTS density The decreasing density of BTS outside urban areas reduces the accuracy of the classification algorithm, as different usage patterns will be averaged over a larger area. As population density reduces in the peri-urban outskirts of the metropolitan area of Dar es Salaam, so does the density of BTS and associated

coverage area size. This reduction in density and increase in area size further exacerbates the issue of high diversity in land use discussed above.

Verification of cluster label interpretation Labels denoting different land use interpretations were derived from the interpretation of latent components shown in Figure 3.3 in respect of traditional land use categorisations, and in the absence of accurate and up to date ground reference data. This method provides no guarantee, however, that clusters can be interpreted with respect to traditional land use characteristics.

3.5 Chapter Summary

The analysis leveraged activity signatures derived from 450.2 million call and SMS event logs to generate insight into area-level activity-based land use across the Dar es Salaam metropolitan area. Constructed signatures were feature scaled for comparison with NMF used for extraction of matrices of latent factors representing underlying land use classes. Unsupervised clustering was subsequently used to cluster BTS with similar signatures to those represented by the latent factors.

This study demonstrated that NMF and k-means clustering could be used to identify interpretable activity-based land use classifications from mass CDR datasets. However, limitations remain, largely due to a need for more high-quality ground reference data for validation and the inherent nature of MND data collection that results in sparse sampling frequencies (§7.2.1). At the same time, this analysis will provide useful information in metropolitan areas such as Dar es Salaam, where the majority of the city is made up of informal housing, complicating both data collection and governance, and land management and land use planning approaches.

The following chapter will use CDR and MFS data to generate insights into SEL, which cannot be captured through activity signatures alone, and factors that may contribute to RSS as endogenous factors for subsequent land use and socio-economic – transport

interaction analysis in chapter 6.

Chapter 4

Tracking Urban Socio-Economic Levels

4.1 Chapter Introduction

The previous chapter used CDR data to examine activity-based land use to understand the diversity of the built environment of Dar es Salaam. This chapter is concerned with the extraction features from the socio-economic dimension described in §2.3.2 from CDRand MFS data for subsequent area-level SEL classification. The following research objective guides this chapter:

Research Objective 2: the investigation of small area SEL classification using CDR and MFS data through supervised machine learning, and subsequent analysis of features used for classification to understand the main determinants behind classification results.

The SEL of an area can be seen as a proxy indicator for poverty within the area. Tackling poverty requires an intricate understanding of poverty in the area itself and by extension, accurate poverty measurements to measure both progress and success of interventions. In line with the 'Statistical Tragedy' affecting emerging economies around the world, re-

search by Dang *et al.* (2019) [83] has shown that poorer countries with an increased need for poverty reduction are also facing greater barriers in measuring poverty.

Traditionally, the most reliable way to estimate SEL and poverty has been through censuses or household surveys. This pathway has proven both impractically expensive and time-consuming in the majority of fast-evolving urban spaces in the Global South. These surveys are often only conducted every few years and by the time results are available, they are often already outdated. In addition to those general shortcomings, statistical institutes in emerging economies often lack in technical and logistical capacity further prohibiting the conduction of high-quality and large scale surveys [83]. As a result, (survey) data collection is rarely a high priority among emerging economies, further exacerbating the 'Statistical Tragedy' [93].

A solution, however, exists in the proliferation of cheap mobile phones and the new sources of data generated through them, such as CDR and anonymized MFS transaction logs used as part of this thesis research. MFS is a viable alternative to traditional banking services in emerging economies, particularly utilized by citizens unable to obtain a bank account, MFS transaction logs provide a mostly unexplored yet potentially more powerful data set held by the same MNO [115]. Appendix B examines the nature and rise in popularity and market penetration of MFS as an enabler of large-scale economic data collection in emerging economies.

As a first step, this chapter provides a review of the literature on MFS research, mobile money in Africa and approaches to mapping poverty and urban SEL. Second, the chapter uses a sample of 329,530 mobile phone subscribers resident in Dar es Salaam, Tanzania, (7.6m MFS records, 450.2m call and SMS event logs) to demonstrate the improvements that can be made in SEL classification compared to traditional data sources. A custom ground reference survey conducted over a 2 month period between late 2015 and early 2016

as part of the EPSRC *Neodemographics* project[1] and in collaboration with the *Ramani Huria* program [84] was used as a poverty measure [115].

4.2 Literature Review

4.2.1 Mobile Financial Services

Existing research on MFS has mostly been limited to interview and survey-based work that broadly falls under three areas:

Benefits and impact of MFS

There are several studies on the general reasons for the adoption and success of MFS [118, 274, 289]. Minto-Coy and McNaughton (2016) [244] review reasons for the success of MFS service in some countries with a more comparative review on why MFS services are successful in emerging but not developed economies undertaken by Chaix and Torre (2015) [71]. Others investigated more specific factors on MFS uptake and success, such as the influence of word of mouth and social networks in rural areas [189, 235].

The reduced time commitment, costs and paperwork for initiating transfers compared to traditional banking services [250] have been found to have a positive impact on female empowerment in particular due to "simpler registration process and less burdensome documentation requirements" - transcending issues of financial illiteracy [180, 190].

Part of the success of MFS and its inherent benefits was increasing the availability of financial data, allowing providers to make more informed decisions and by extensions provide a wider range of micro-credits and loans [22]. The digital nature of MFS accounts make these services the second most popular savings instrument after traditional hiding places [245] and help build resilience by allowing for budgeting and saving of assets for emergencies through the M-Pawa service. M-Pawa, a service offering mobile credit and

[1] "Neo-demographics: Opening Developing World Markets by Using Personal Data and Collaboration", EPSRC Reference: EP/L021080/1, 2015

savings, was launched jointly by Vodacom and CBA in late 2014 and attracted nearly 5 million customers within the first two years since its introduction in Tanzania [151].

Inclusive development and rural usage

Second, research has investigated the impact of MFS services on rural and agricultural areas in terms of inclusive development [23, 286]. Kim *et al.* (2018) [196] provide a comprehensive review on mobile-based financial inclusion and "contribute to the evidence-based policy-making and practices on inclusive MFS by contouring the broad landscape of current knowledge" at the intersection of MFS, financial inclusion and development.

Transporting money more securely is particularly vital in remote regions, where households commonly rely on remittances from family members working in urban centres, which have traditionally been transported through informal services such as paying bus drivers or through other risky means. It has been found, that the digital storage identified in the previous section, and the ability to transfer money between account holders, allows for the safer transmission of remittances [179, 180, 191].
There has also been work looking at MFS adoption among dedicated user groups such as the work by Mpogole *et al.* (2016) [250] on students and businesses owners in Iringa, Tanzania; Dzogbenuku (2013) [101] on students in Ghana; Mbogo (2010) [236] on micro-businesses in Kenya; and Kikulwe *et al.* (2014) [195] on smallholder farmers in rural Kenya [2].

Safety and Security

Finally, previous research has examined the safety and security of MFS services [155]. Compared to traditional hiding places, which are susceptible to theft or loss, MFS allows significantly safer storage and savings options [180, 257]. The digital nature of MFS accounts make these services the second most popular savings instrument after traditional hiding places [245] and help build resilience by allowing for budgeting and saving of assets

[2] A wider review on review on mobile banking can be found in [308]

for emergencies [10].

At the same time, however, insufficient trust, privacy concerns and technical literacy and sophistication for both sender and receiver have been identified as non-price related barriers to increased service adoption [107, 192, 229]. Trust in the agent that is the interface for a lot of the customer base and essentially the backbone of the MFS system is particularly key [235, 302]. Distrust can stem from a lack of liquidity management of vendors [100] or professionally and qualification level of agents [22, 235]. Having agents within one's social network has been found to improve trust significantly [235].

Regulation of the MFS ecosystem can have a tremendous impact on the wide adoption and by extension success of MFS in a country [81, 235]. Sanz and De Lima (2013) [302] investigate the level of restrictions that allow an environment to thrive while preventing malicious activity such as money laundering and terrorism financing [20, 81].

4.2.2 Poverty Mapping

The majority of previous work on poverty estimation using non-conventional data sources has been conducted using satellite imagery. Satellite imagery has numerous advantages over more conventional manual sampling approaches, including global coverage, a high revisiting capability and relative ease of access. It allows for the collection of a range of derived data such as Night Time Light (NTL), vegetation cover [265], and base data for GIS driven analysis of an area's proximity to the physical location of POI such as schools and hospitals, and infrastructure (e.g. density, proximity to paved roads) [373]. NTL or Light Based Growth Indicator, in particular, have received a lot of attention, showing a good correlation with a country's GDP [105, 110] and have been proposed as a supplement to national accounting in data-poor countries [162]. NTL has shown a good correlation with electrification and economic growth data for 5000 sub-districts in Indonesia between 1992 and 2008 [266].

Recent work in the US [206], Bangladesh [313] and China, however, suggests that as urban areas are nearing saturation levels of NTL and consequently the value of NTL as an indicator is beginning to decrease. The analysis of satellite imagery is both costly and static, contrasting sharply with dynamic and digitally logged behavioural data streams such as CDR. While satellite imagery allows us to observe and understand the role of natural resources and certain aspects of the spatial dimensions discussed in §2.3.2, it cannot provide insight networks such as socio-economic ties, cultural interactions as well as micro- and macro behaviour that is intrinsic to understanding poverty.

4.2.3 Poverty Mapping Using Mobile Network Data

Aside from Gutierrez *et al.* (2013) [153, 154], who used airtime top-up behaviour logged in CDR records as an indicator or a users wealth under the assumption, that poorer people are likely to have frequent top-ups in small amounts compared to infrequent large amount top-ups [3], and Eagle *et al.* (2010) [103], who used average normalised entropy of the social ties in a neighbourhood to find that they correlate strongly with the level of socio-economic deprivation, a number of previous studies have utilized aggregate CDR data for poverty prediction. An annotated bibliography is provided in Table 4.2.3, summarising details and results of the key research papers on poverty mapping using CDR data discussed in more detail below.

[3] Results, which were not validated against established wealth indicators in Côte d'Ivoire

Table 4.1: Overview of research on poverty prediction and socio-economy analysis using Mobile Network Data

Reference	Data Source	Model (# Features)	Sample Size	Time Period	Findings	Spatial Resolution	Poverty Measure	Region
Soto *et al.* (2011) [317]	CDR	SVM and Random Forest (279)	500k users	6 months	r=0.8	920 BTS	Region-level SEL	Urban area in a Latin American city
Smith-Clarke *et al.* (2014) [313]	CDR	Linear regression (10)	5M (Set1), and 928k users	5 months and 6 weeks		255 sub-prefectures/ 176 areas at the next administrative level down	IMF and Asset-based index	Côte d'Ivoire and undisclosed region
Blumenstock *et al.* (2015) [47]	CDR and phone survey	Supervised learning, Linear regression (5088)	1.5M users (CDR) and 856 (survey)	9 months	Survey respondents (r=0.68), DHS households (r=0.917), DHS cluster (r=0.79)	2148 Cells, 30 districts, 492 DHS Clusters	DHS composite wealth index	Rwanda
Pokhriyal and Dong (2015) [279]	CDR	Linear regression	9.54m users (Set1), 146k users (Set3)	12 months	r=0.82	14 regions and 123 *arrondissements*	OPHI MPI	Senegal

Letouze (2016) [215]	CDR	Linear Regression	5 million (Set1) and 50k users (Set3)	5 months	r=0.3 (head-count), r=0.15 (inten-sity)	1214 BTS	MPI (Poverty head-count)	Côte d'Ivoire
Pokhriyal and Jacques (2017) [280]	CDR and environ-mental data	Gaussian Process Regression	9.54m users (Set1), en-vironment data	12 months (CDR), 1960-2014 (Environ-ment)	r=0.91	552 communes, $100m^{-1}km$ (Environment)	MPI	Senegal
Engelmann *et al.* (2018) [115]	CDR, MFS, Survey	Random Foret	7.6m MFS records for 147k users, 450.2m CDR for 329,530 users	4 months	65.9% - 73.4%	517 BTS	Survey	Dar es Salaam, Tanzania

CDR Data Studies

One of the first was Soto *et al.* (2011) [317], who used SVM and Random Forest predictors and CDR data for 500k users over six months between February and July 2010 to predict region-level derived SEL within an urban area in a major Latin American city. They extracted 279 features including 69 behavioural variables (e.g. total number of calls and SMS), 192 social network features (e.g. in-degree and out-degree) and 18 mobility variables (e.g. total travel distance and the number of visited BTS) from CDR from 920BTS for 500k users over six months. In addition to CDR, they made use of 1200 Geographical Regions between $1km^2$ and $4km^2$ in size and labelled as one of three socio-economic levels as ground reference. Each BTS was assigned a weighted SEL variable for classification. Their model showed 80% accuracy in the classification of SEL using 38 features selected for the final classification task.

Focusing on Rwanda, Blumenstock *et al.* (2015) [47] combined a geographically stratified random sample of 856 phone surveys with CDR data for 1 year for 1.5 million users. They constructed a composite wealth index from the first principal component of several survey responses related to wealth and DHS data from 2007 and 2010. The DHS does not include income or consumption data, but rather information on asset ownership, health and education that can be used for the construction of a Multidimensional Poverty Index (MPI) [14, 15]. Following a feature engineering process to generate 5088features from CDR data and feature selection, they generated a generalisable classification model and found a cross-validated correlation coefficient of $r = 068$ for the 856 users with survey responses. Using the survey-CDR model to predict wealth for the remaining users in the CDR data set and linear correlation against DHS responses for 12,972 households, they found strong correlations for households owning at least one mobile phone (r=0.917), all surveyed households (r=0.916) and dis-aggregated clusters akin to a village (r=0.79).

Orange D4D Challenge Côte d'Ivoire

Smith-Clarke *et al.* (2014) [313] used CDR data from the Orange D4D challenge and poverty rate estimates IMF for 2008 at 11 subnational region level for Côte d'Ivoire, and CDR data and an asset-based index derived from census data for an unspecified region. The D4D data used was Set 1 (BTS-to-BTS traffic on an hourly basis) for 5 million customers from December 1st 2011 until April 28th 2012 while the undisclosed region set covered 928k customers and 40 million CDR for six weeks in early 2012. Features included extracted activity (call volume and duration), gravity residual (the difference between observed and expected interregional flow), network advantage (entropy measure capturing call diversity across areas, and a measure of degree distribution) and introversion (volume comparison of inter- and intra-traffic) measures [313]. Poverty level estimation was undertaken using a linear ordinary least squares regression model for 225 sub-prefectures in Côte d'Ivoire and 176 areas at the next administrative level down in the unspecified region. They found a negative correlation between activity and poverty that is hypothesised to become less of an effective proxy as the telecommunications market matures, similar to the weakening of NTL as a proxy [206, 313].

Using Set1 (BTS-to-BTS traffic on hourly basis) used by [313] and Set3, which covers individual trajectories for 50k users over 12 months, Letouze (2016) [215] combined CDR data with DHS survey data. Each of the 1214 BTS in the datasets was assigned per capita call variables and self-constructed MPI headcount data from DHS clusters based on the coverage of the Voronoi shapes of the BTS location. Population counts for normalization were calculated from 2010 2.5 arc-minutes resolution raster data from the Center for International Earth Science Information Network, Columbia University and administrative population estimates at the level of 225 sub-prefectures provided by UN OCHA. Using linear regression, he found a strong significance of the coefficients, with weak correlations with the MPI headcount variable (r=0.3) and intensity of multidimensional poverty as the average share of deprivations experienced by people classified as living in multidimensional poverty (r=0.15).

Orange/Sonatel D4D Challenge Senegal

Using 12 months' worth of data from Senegal provided by Orange as part of the D4D challenge, Pokhriyal and Dong (2015) [279] created poverty maps for 14 regions and 123 *arrondissements* in Senegal using two different approaches. The first is using hourly BTS-to-BTS traffic for 1666 BTS (Set1 data) similar to those for Côte d'Ivoire used by [215, 313] to generate virtual connectivity maps at the BTS, *arrondissement* and regional level for linear correlation of graph-theoretic measures (e.g. centrality) and direct features (activity, eigenvector and page rank centrality, page rank, gravity residual and introversion) for each area with an MPI. They found a strong negative correlation between the metrics and the headcount ratio of poverty, a marked negative correlation with the incidence of the poor, and also a region's MPI. The second was using hourly location recordings on the *arrondissement* level for 146,352 Orange customers (Set3 data) plus "a monthly set of 33 behavioural indicators which capture calling/texting patterns (14), mobility patterns (6), and social behaviour (13) of each user" [279]. Conducting linear regression between the median region level for each indicator and the MPI, they found that 11 of those indicators were shown to have a Pearson's R of greater than r=0.9.

Extending on previous research [279], Pokhriyal and Jacques (2017) [280] used the CDR datasets for Senegal provided by Orange in combination with environmental data (i.e. food security, economic activity and facility access) to conduct Gaussian Process Regression. CDR was provided by Orange for 9.54m users and 11 billion interactions for 1666 BTS over 12 months, while environment data were collected from 1960-2014 in a resolution of 100m^{-}1km. This was combined with 2013 census data for 1.4 million individuals at the household level and a MPI for 2013 at the regional level (14 regions). Testing correlation on the commune level (552), they found a strong correlation with Pearson's R of r=0.91, similar to their previous findings [279].

Research gaps

Two gaps arise in the prior literature:

First, a gap in conducting poverty prediction in small geographic areas such as wards or Lower Super Output Area (LSOA) in part due to a lack in socio-economic data at such a fine granularity. The spatial resolution of previous research highlighted in Table 4.2.3 and discussed in more detail above varies from BTS-level [47, 115, 215, 317], DHS cluster[4] or commune level [47, 280], sub-prefectures or *arrondissements* level [279, 313], and regional or district level [47, 279]. Other studies including Engelmann *et al.* (2018) were carried out on the BTS level, Soto *et al.* (2011) [317] used CDR data for an undisclosed region with undisclosed sources for their area level SEL used for validation. Blumenstock *et al.* [47] combined CDR data with a composite wealth-index based on phone surveys with individuals contained within the same dataset, resulting a validation index that cannot be easily replicated. Finally Letouze (2016)[215], who found very weak results for CDR measures (r=0.3 for MPI headcount and r=0.15 for intensity) compared to the classification accuracy for 65.9% to 73.4% achieved as part of this analysis.

Second, a reliance on communication rather than mobile financial service data. While the research on MFS has looked in detail at the impact and barriers to adoption of MFS services, such analysis has been limited to ethnographic studies rather than the analysis of raw MFS data streams themselves. To date, no previous studies exist that use features extracted from MFS transaction logs, a data stream that one would expect to provide significant insight into a fine-grained socio-economic analysis with the exception of Engelmann *et al.* 2018 [115] that is based on the research conducted as part of this thesis. As the results of this research §4.4 show, financial data can bring a significant improvement to classification results compared to models using only CDR derived features.

[4] a geographic area designed to be comparable to a village

BTS Level	DHS cluster/ commune level	sub-prefecture/ *arrondissements*	region / district
Soto *et al.* (2011) [317]	Blumenstock *et al.* (2015) [47]	Smith-clarke *et al.* (2014) [313]	Blumenstock *et al.* (2015) [47]
Blumenstock *et al.* (2015) [47]	Pokhriyal and Jacques (2017) [280]	Pokhriyal and Dong (2015) [279]	Pokhriyal and Dong (2015) [279]
Letouze (2016)[215]			
Engelmann *et al.* (2018) [115]			

Table 4.2: Spatial granularity of analysis in MFS papers.

4.3 Research Approach

Similar to the empirical analysis of activity-based land use conducted in the previous chapter 3, the analysis focused on BTS for the metropolitan area of Dar es Salaam. A core dataset of 517 BTS cells from Dar es Salaam were identified, with each cell being labelled with a SEL based on a ground reference survey discussed below. Two testing scenarios were formed, one considering prediction for all BTS cells within Dar es Salaam that were assigned a SEL as part of the survey (n=517), and the second considering only cells labelled as being predominantly residential (n=384). Three models were constructed for each scenario (producing six scenarios in total) to allow direct statistical comparison of the effectiveness of MFS versus CDR features. Competing models were generated from either MFS derived features, CDR derived features or combined feature sets. The performance of each model in predicting SEL was tested via a strict cross-validation methodology with features further analysed through variable importance measures and Partial Dependency plots to gauge their explanatory value for inclusion as input features for the socio-economic dimension for subsequent land use and socio-economic – transport interaction in chapter 6.

4.3.1 Data Description

To address the research gaps highlighted in §4.2.3, this study leverages three disparate data sets: mass CDR and MFS datasets for the generation of regularity, diversity, activity and spatial features, and ground reference survey data for validation of classification performance. The following subsections will provide an overview of the three data sets used as part of the study. The spatial granularity for the features for subsequent analysis was set at the BTS and surrounding Voronoi polygon resulting in an average study area size of approximately $6.23km^2$.

CDR data

Residential location was calculated from CDR logs for SMS and call events for the whole of Tanzania through the calculation of the mode BTS favoured by users between 10 pm and 6 am. Only those users whose BTS was located within the Tanzanian municipalities of Kinondoni, Ilala and Temeka, which are classed as the Dar es Salaam metropolitan area, were included within the study. The approach used for home BTS detection is similar to related approaches by Calabrese *et al.* (2011) [60], Isaacman et al. (2011) [177] and Mamei and Ferrari (2013) [231] who chose 9pm to 6am for 'home-events' and 11am - 4pm for 'work events'. An adapted home/work detection accuracy index based on prior work by Berlingerio *et al.* (2013) [40] was further used to overcome issues of the frequent visitation ranking technique misclassifying other frequently visited locations as home or work respectively. A BTS was confirmed as a home location if the repetitiveness of the home cell was greater than 0.5 and a user-generated network events on at least 60 days during the study period. The repetitiveness of a home cell can be defined as

$$BTS_{home} = \frac{\beta_{home}}{\beta_{events}}$$

With BTS_{home} as the confirmed home BTS, β_{home} as the number of nights when the BTS is the most frequently used for home-events, and β_{events} as the number of nights when network events are generated.

The final CDR dataset used as part of this study covers a total of approximately 476 million call and SMS events for 329k mobile phone subscribers resident within the Dar es Salaam region of Tanzania over 122 days from August 1st to December 1st, 2014. Due to both individual and commercial privacy, the anonymised data used as part of this study is not publicly available and was provided through a partnership with a Tanzanian MNO with high market penetration in the case study area of Dar Es Salaam. Features pertaining to residents of a BTS were aggregated to the respective BTS level as the smallest granularity as highlighted in Table 4.2, ensuring both strict privacy and for analysis purposes. By undertaking a Voronoi tessellation of each BTS, a set of irregularly shaped cells was generated at an average size of $6.23km^2$. The subset of areas classed as predominantly urban was $2.77km^2$ on average.

MFS data

As discussed in §2.2.2 MFS 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 ones mobile phone account to bank accounts and vice versa" [250, p.4]. Similar to CDR being recorded whenever an active or passive network event takes place as discussed in §2.2.1, a MFS transaction record was generated whenever a transaction took place. The MFS features, whose extraction is described in more detail below, were associated with a 'home' BTS extracted from the CDR data set described in the previous section. While an antenna identifier was included in the data set as a spatial reference, it was truncated to the point where it was not possible to assign a MFS record to a BTS and by extension a rough geographical area. Only users who made use of both call or SMS and MFS services could, therefore, be included within the analysis of the MFS records. Ultimately, the analysis is based on 47.6m MFS records of approximately 147k customers classed as residing with the Dar es Salaam metropolitan area for the same 122 day period as covered by the CDR data.

Ground Reference Data

In the tradition of the 'Statistical Tragedy' [93], accurate and fine-grained ground reference data necessary for supervised machine learning is extremely hard to find in East Africa. Additionally, most of the available ground reference data sets are collected within existing administrative zoning boundaries which seldom correspond to the shape of capture areas of BTS. In order to overcome this limitation a custom ground reference survey that captured SEL was conducted over a 2 month period between late 2015 and early 2016 as part of the EPSRC *Neodemographics* project[5] and in collaboration with the *Ramani Huria* program [84]. In addition to SEL, the survey captured a range of attributes from the spatial, socio-economic and mobility dimensions including the predominant land use, whether an area is residential, and 'social mobility' for 517 areas across Tanzania's largest city, Dar es Salaam. The labelled areas are contiguous, irregular and cover nearly the whole of the metropolitan area of Dar es Salaam. Survey coverage areas have been derived via a Voronoi tessellation of WGS84 locations of each BTS in order to address the mismatch between BTS capture areas expressed as Voronoi cells and administrative boundaries within official statistics. Surveying was undertaken by local inhabitants from *Ramani Huria*, with the aim of each area being surveyed at least twice for confirmation. Income labels were chosen to be recorded at the most common recording interval. Ultimately, areas associated with six BTS out of 565 were inadvertently missed out during surveying. The overall SEL variable was chosen as the focus of our study, with each area being labelled as either 'very poor' (n=5), 'poor' (n=137), 'average' (n=274), 'wealthy' (n=81) or 'very wealthy' (n=42) within the study. This covariate served as the key poverty measure for the modelling and evaluation process.

4.3.2 User Selection

Several users were excluded from the analysis from the onset. The majority of exclusions from the MFS dataset were due to the nature of spatial constraints discussed above.

[5] "Neo-demographics: Opening Developing World Markets by Using Personal Data and Collaboration", EPSRC Reference: EP/L021080/1, 2015

Additional exclusions were undertaken to prevent the occurrence of a high-frequency bias during the aggregation from individual user features to BTS level features. This was done following a multi-step process.

1. Exclusion of MFS users who were not contained in both the CDR and the MFS datasets due to the aforementioned spatial considerations. The use of CDR services was a necessary requirement to be able to add a geographical component to the MFS data. While the MFS data did contain a spatial reference to the sector antenna servicing the transaction, it was truncated and could not be assigned to a BTS with enough confidence for use in this study. Features generated through the MFS data are therefore assigned to the BTS identified as the preferred home BTS through CDR data;

2. Exclusion of users not classified as subscribers as those are very likely to introduce a high-frequency bias due to high-frequency usage patterns compared to the rest of the area population;

3. Pair-wise deletion of users without both incoming and outgoing transactions in the same month. This step was chosen to exclude users who are simply using MFS services as a savings mechanism, rather than following an incoming and outgoing pattern that is more akin to general financial usage;

4. Users with an average of more than 40 incoming or 100 outgoing MFS events per month were excluded from the analysis, as these most likely represented unlicensed businesses or informal street traders operating as regular subscribers.

4.3.3 Input Feature Engineering

To populate the input feature space of each of the three models, 17 features were extracted from the CDR data similar to those in prior work by Pokhriyal and Jacques (2017) [280] to form model 1. Going beyond their work, 22 features were derived from MFS data to form model 2, with a combination of those features representing model 3. Features were first extracted at the individual level before being aggregated to the BTS level as the

lowest reliable level of spatial granularity available with the CDR dataset used as part of this research. As discussed in §4.3.1, the home BTS corresponds to the most frequently recorded BTS with timestamps between 10pm and 6am within a user's CDR data and was used for aggregation of features from individual users. Such aggregation provides the analysis with an additional layer of privacy provision (recent estimates indicate over 4.3 million residents in Dar es Salaam, so the average number of people living in each of the analysis cells is >8000). The derived-features were selected as they capture a diverse range of human behaviour. This ranges from very basic usage of mobile devices to understanding regularity of interactions including responsiveness (expressed through inter event time for example); diversity, which highlights the social network a user engages with such as percent pareto; as well as activity to give insight into their initiative to start interactions (percent initiated) or financial responsibility (e.g. percentage defaulted); and spatial features as a proxy for movement.

While features in Model 1 and Model 2 are broadly comparable with each other, a number of features unique to both CDR and MFS data set were generated. Examples for Model 1 (CDR) are features such as *ratio of text and calls* and response delay in SMS conversations; with features such as *average MFS in/out*, *percentage defaulted* and *percentage balance checks* for Model 2 (MFS). A closer examination of feature breakdown is supplied below, with an overview of all features used provided in Table 4.3.3.

Table 4.3: List of features generated from CDR and MFS data for area-level socio-economic prediction

Feature (total no. of features)	**Data source**	**Description**
Basic Use		
The number of interactions (2)	CDR, MFS	The total number of incoming and outgoing SMS events and MFS transactions for a user.
Number of users (2)	CDR, MFS	The total number of users within the area
Average transaction size (1)	MFS	Size of an average MFS transaction across both incoming or outgoing transactions.
Average MFS in/out (2)	MFS	Average monthly inflow and outflow over the study period.
Total MFS in/out (2)	MFS	Total inflow and outflow over the study period.
Active Days (2)	CDR, MFS	The number of days during which the user was active.
The ratio of text and call interactions (1)	CDR	The ratio of text and call interactions.
Spending uptake (1)	MFS	The total spend in an area divided by the MFS uptake (the number of MFS users divided by the number of CDR users in an area).
The ratio of incoming and outgoing transactions (1)	MFS	The ratio of incoming and outgoing MFS transactions.
Regularity		
Inter event time (4)	CDR, MFS	The inter-event time between two records of the user. This feature is calculated as mean and SD for MFS, mean for calls and SD for calls and SMS.
Monthly events (1)	MFS	Average number of transactions per month.
Diversity		

Balance of contacts (2)	CDR, MFS	The balance of interactions per contact. This feature is calculated-each for text and MFS. For every contact, the balance is the number of outgoing interactions divided by the total number of interactions (in + out).
Interactions per contact (2)	CDR, MFS	The number of interactions a user had with each of his or her contacts via call or MFS
Percentage Pareto interactions (2)	CDR, MFS	The percentage of user's contacts that account for 80 of his or her call or MFS interactions
The entropy of contacts (2)	CDR, MFS	The entropy of the user's contacts for calls or MFS
Activity		
Response delay (2)	CDR	The response delay of the user within a conversation (in seconds). Calculated for text (SD and mean of the response delay).
Percentage initiated (3)	CDR, MFS	The percentage of network events initiated by the user for calls, call and SMS, or MFS.
Percentage defaulted (1)	MFS	The percentage of transactions that failed due to insufficient account funds.
Percentage balance checks (1)	MFS	The percentage of transactions representing balance checks.
Spatial		
Number of BTS (3)	CDR, MFS	The number of unique cells or BTS visited.
Frequent BTS (1)	CDR	The number of BTS that accounts for 80% of locations where the user was.
Entropy of BTS (1)	CDR	The entropy of visited BTS.

CDR derived features:

Features for Model 1 (CDR) were derived from transactions generated by 329,530 users over the study period and classed as living in the Dar es Salaam metropolitan area. These include a range of features identified as important for poverty prediction as identified by Pokhriyal and Jacques (2017) [280]. The features can be broadly classified as falling under five different domains:

Basic usage: active days for both call and SMS, the ratio of call and text interactions and the number of SMS interactions;

Regularity: the mean inter-event time for calls and standard deviation of inter-event time for calls and text;

Diversity: the mean balance of contacts for SMS, the percentage of Pareto interactions for calls, the mean call interactions per contact and the Entropy of contacts;

Activity: the mean, and standard deviation in response delay for texts, the per cent initiated interactions for calls and the per cent of initiated conversations for call and text;

Spatial: the total number of visited BTS, the most frequent BTS, the entropy of BTS and the radius of gyration.

MFS derived features:

Features for Model 2 (MFS) were derived from transactions covering 147k users with both incoming and outgoing MFS transactions occurring in every month of the sampling period, and who were classed as living in Dar es Salaam. The 'residency' was determined by cross-referencing records with the anonymized *callingpartynumber* id contained in both CDR and MFS datasets and using the BTS identified as the most likely 'home' location. As discussed in §4.3.2, only transactions by those users deemed to be regular subscribers rather than commercial users were taken into account to calculate those features. With some error codes changing over the year, only transactions without error codes or those with code '200' or 'error000' indicating success were included. A full list of error codes

is included in appendix B. Similarly, transaction amounts equal to or under 50Tanzanian Shilling (TZS) were excluded from the analysis, as they mostly referred to balance inquiries or pin changes, which involved minuscule service charges, that introduced significant noise into the MFS feature calculation. The final data set used for feature engineering was comprised of n=10,011,674 MFS records. Where applicable, features were calculated as the average (across users) of averages (individual users' transactions). The list below comprises features that were not contained in Model 1 (CDR) but rather unique to being derived from MFS data:

Basic usage: the average transaction size across both incoming or outgoing transactions, the average MFS in/out representing the amount of money which was received or spent during an average month, the total amount received and spent over the study period, spending uptake as the total spending in an area divided by the MFS uptake (calculated as the number of MFS users divided by the number of CDR users within the same area), and the ratio of incoming and outgoing MFS transactions.

Regularity: the average number of MFS transactions for a given month within the study period.

Activity: number of defaulted transactions identified as having failed due to error codes indicating an insufficient account balance to complete a particular transaction as highlighted in appendix B, and balance checks identified as transactions with an event type for 'balance inquiries'.

Purchase categories:

Each MFS record contained a *servicetype* and *subtype* with a coded reference to the type of transaction carried out (*servicetype*), and the vendor involved in the particular transaction (*subtype*). Nineteen distinct categories for the categorisation of service and product purchases were generated using the MNO bill-paying categories advertised through their website. The majority of codes contained in the *subtype* column were related to a vendor and subsequently categorised based on Google searches of the code and local knowledge.

While the *subtype* allowed for the categorisation of most vendors, not all *subtypes* could be reliably identified, leading to the grouping of a large proportion of purchases as unknown.

Ultimately, category indicators were not used in the generation of the poverty classification models due to their sparse nature. While there was a consideration to generate additional scenarios by differentiating between all MFS users (n=523592 including those without bill payments) and only MFS users that use the service for bill payments (n=122179 with bill payments, n=122211 with bills and SMILE, and n=231938 with Xtreme Purchase), this approach resulted in the exclusion of too large number a number of users to be viable.

4.3.4 Experimental Method

In order to investigate the utility of features derived from MFS versus CDR transaction logs for remote SEL classification, a core prediction task was formulated. The prediction task underpinning the scenarios was formulated as a ternary classification problem in order to overcome the uneven distribution of class memberships in the raw survey data introduced in §4.3.1. Output feature labels 'very poor' and 'poor' were merged, as were labels 'wealthy' and 'very wealthy'. This resulted in a relatively balanced data set with 202/143 areas labelled as poor, 153/102 as average and 163/140 as wealthy classes for all/residential scenarios respectively.

For each scenario, Model 1 was trained using an input feature set drawn from CDR data, with features fastidiously engineered to correspond directly to those used in the most recent literature (see §4.3.3 for more details of this process). Model 2 was trained using an MFS derived feature set. Many features in Model 2 echo those in Model 1 (to provide a fair comparison), despite being seeded by transaction logs denoting MFS rather than call or SMS events. A final predictor, Model 3, was trained using a combination of all features used in the previous models. A number of different classification methods were tested including logistic regression, decision tree, random forest and nearest neighbour alongside a basic dummy classifier. The selection of methods was guided by prior research discussed

earlier (see Table 4.2.3).

Logistic regression

Logistic regression modelling is an effective technique for the analysis of linear continuous domains. Easy to implement regularisation methods exist for non-binary, discrete categorical domains [18, 171]. The resulting model can be easily updated for future scenarios. It is an attractive technique to use because it is (1) better than probit by providing easier interpretation (2) logistic regression accuracy is largely unaffected by the distribution of the predictors [18], and (3) easy implementation while providing quick and robust results.

Unlike the majority of previous studies on SEL using CDR data discussed earlier in §4.2.3, logistic regression is used purely as a benchmark against which to compare the more advanced classification techniques against as part of this analysis. Traditional linear per-feature correlation analysis was omitted as a core analysis model, given that the primary focus is concerned with the analysis of the utility of features from the socio-economic dimension, and all features were based on pre-collected and re-purposed data with negligible acquisition cost.

Decision tree

Decision Trees are relatively simply predictive models, which have nonetheless proven to be highly effective in both regression and classification tasks. They are an increasingly popular choice for solving classification tasks because of their ease of use and interpretability, and their ability to accommodate predictors measured at different measurement levels (including nominal variables) [325]. Such models take a very different approach to classify problems from SVM and Artificial Neural Networks, forming a set of classification decision rules based on the given attributes of the data. Iteratively, a set of decision boundaries are formed that create binary splits within the data set, minimising the entropy of each class either side of the boundary, and hence gradually cutting the feature space into areas favouring one of the k-classes.

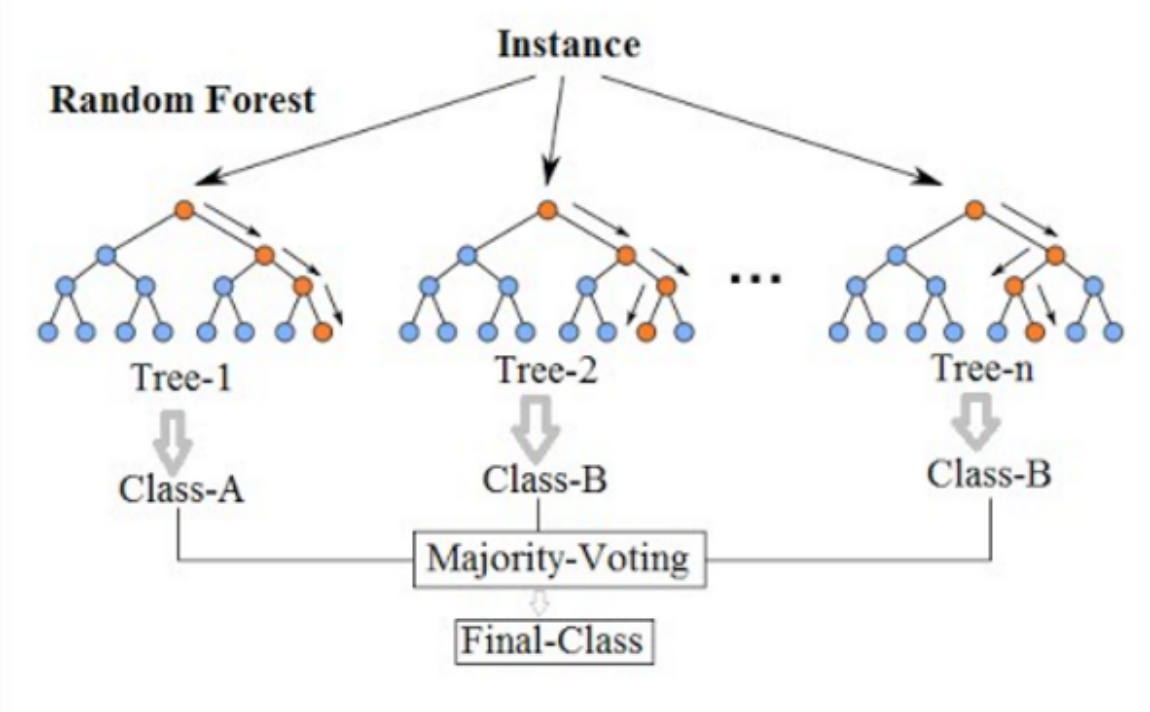


Figure 4.1: High-level overview of the Random Forest prediction model

Model	CDR feature set	MFS feature set	Combined feature set	Avg accuracy
Linear regression	51.663	59.067	59.545	56.797
Decision tree	52.751	59.828	54.222	53.774
Random Forest	57.320	58.754	56.983	56.924
Nearest Neighbour	52.202	59.256	51.653	55.320
Dummy classifier	35.343	37.987	38.037	37.448

Table 4.4: Accuracy Results for different classification techniques without grid search for all three feature sets

Random Forests

Random Forest is an ensemble classifier based on the ideas of bootstrap sampling and random feature selection [54, 317]. With ensemble learning an input dataset is split into a bootstrap sample as a learning set and a training set respectively. The key objective of ensemble learning techniques is the reduction of bias and variance. During the training of the the three (an example of which can be seen in Figure 4.1) each node and its split is calculated using a predetermined number m of randomly selected features with m ¡¡ M with M as the dimensions of the feature space.

4.3.5 Evaluation Setup

Initial exploratory analysis over a range of model classes carried out in the previous step (Table 4.4) showed Random Forest to be performing highly effectively for the given scenarios. Given the ability of Random Forest to directly model multi-class classification problems, handle non-linear relationships, and their association with well understood and tractable variable importance measures [301], the subsequent evaluation focused on this class of model.

To evaluate the comparative utility of the CDR versus MFS features in real-world applications, two predictive tasks are considered (1) predicting the economic status of *only residential* regions populations and (2) predicting the economic status of *all* regions populations. The first represents the more focused task of interest, while the second represents a common real-world use case with an absence of knowledge to the residential/non-residential status of areas under study. Areas were excluded from the first prediction task based on the collected ground-reference survey (§4.3.1). In each case, a combined model with all features (CDR + MFS) and two models per data source (one only containing CDR features, one only containing MFS features) was evaluated. Output labels were generated based on this ground-reference data set, with each region labelled as either 'poor', 'average' or 'wealthy'.

For each model, the data was split via stratified random sampling into a training (66%) and test set (33%) and the parameters for the Random Forest (number of trees, maximum depth and the minimum samples per split) selected via a grid search underpinned by stratified five-fold cross-validation using only the training set and the model finally trained with the best parameters. The performance of each model was then tested on the held-out test set. This was repeated 30 times arriving at 30 performance scores for each of the three models across the six scenarios.

4.3.6 Evaluation Criteria

The performance of each model was measured via the precision, recall and the F1 score (harmonic mean of the precision and recall) for each of the 30 runs per model. The average results overall runs are shown in Table 4.5, along with classification accuracy. Box-plots showing the per-class distribution of the overall F1 scores per model are shown in Figure 4.7a for all areas and Figure 4.7b for residential areas only.

4.3.7 Analysis of Variable Importance

Of equal, if not greater importance than overall model accuracy, is the ability to break apart the models to investigate the importance of the individual CDR and MFS features. To achieve this, the final step in the investigation was a three-stage variable importance analysis.

Firstly, and due to a large number of features proposed in the literature (and included in this work), the candidate set was filtered to exclude those features shown to have minimal impact on model performance using the *Boruta algorithm* [208, 324]. In contrast to other popular feature selection approaches such as sklearn's *SelectFromModel*, *Boruta* compares features to a permutated (i.e. randomized) version of the feature. *Boruta* only considers variables as important if they provide a statistically significant increase in prediction strength compared to the permutated version. The level required to achieve statistical significance was set low ($p = 0.1$) to minimize the probability of inadvertently discarding relevant features.

Secondly, surviving features were used to fit optimized Random Forest models for the whole dataset, selecting optimized parameters via a grid search. The out-of-bag [6] OOB F1 score was then checked to ensure the reduction in features did not cause a significant decrease in overall generalized performance.

[6] Out-of-bag (OOB) refers to the use of only the trees in the forest for which the sample being predicted was not part of during training. In this way, the resulting performance measure can be considered to represent the generalized error [182].

This then allowed a full permutation importance analysis to be undertaken. Permutation importance was used due to both its interpretability (illustrating the mean decrease in performance of omitting each feature) as well as its ability to attribute variance in the case of non-linear interactions.

With important features thus identified, their behaviour within the model and across sub-populations of the data was further examined using partial dependence plots (§4.4.3).

4.4 Discussion

4.4.1 Model Performance Results

The prediction task underpinning the models was formulated as a ternary classification problem in order to overcome the uneven distribution of class memberships in the raw survey data. Output feature labels 'very poor' and 'poor' were merged, as were labels 'wealthy' and 'very wealthy', resulting in a relatively balanced dataset (with 202/143 areas labelled as poor, 153/102 as average and 163/140 as wealthy classes for all/residential scenarios respectively). Table 4.5 reports all results for the experiments, detailing the F1, precision and recall scores averaged over each of the 30 experimental runs, accompanied by an overarching classification accuracy score. Per-class F1 scores were calculated using sklearn.metrics.precision_recall_fscore_support [272]. Figure 4.2 illustrate the confusion matrices for all three feature sets across the residential and all-area scenarios. As can be seen, in all scenarios, Model 2 containing MFS features strongly outperform Model 1 that uses CDR features alone. The final column, indicating per-class F1 scores for each model, illustrates that this improvement is not due to one class being favoured, but occurs across the board – and in particular improves prediction of average, middle-income areas.

Of note is the improvement in F1 scores of Model 1 against Model 2, with an increase in accuracy of $\sim$ 9.7% for all areas, and $\sim$ 9.2% for residential areas. In both cases,

Table 4.5: Accuracy Results for Random Forest prediction using different feature sets over 30 randomly seeded experimental runs

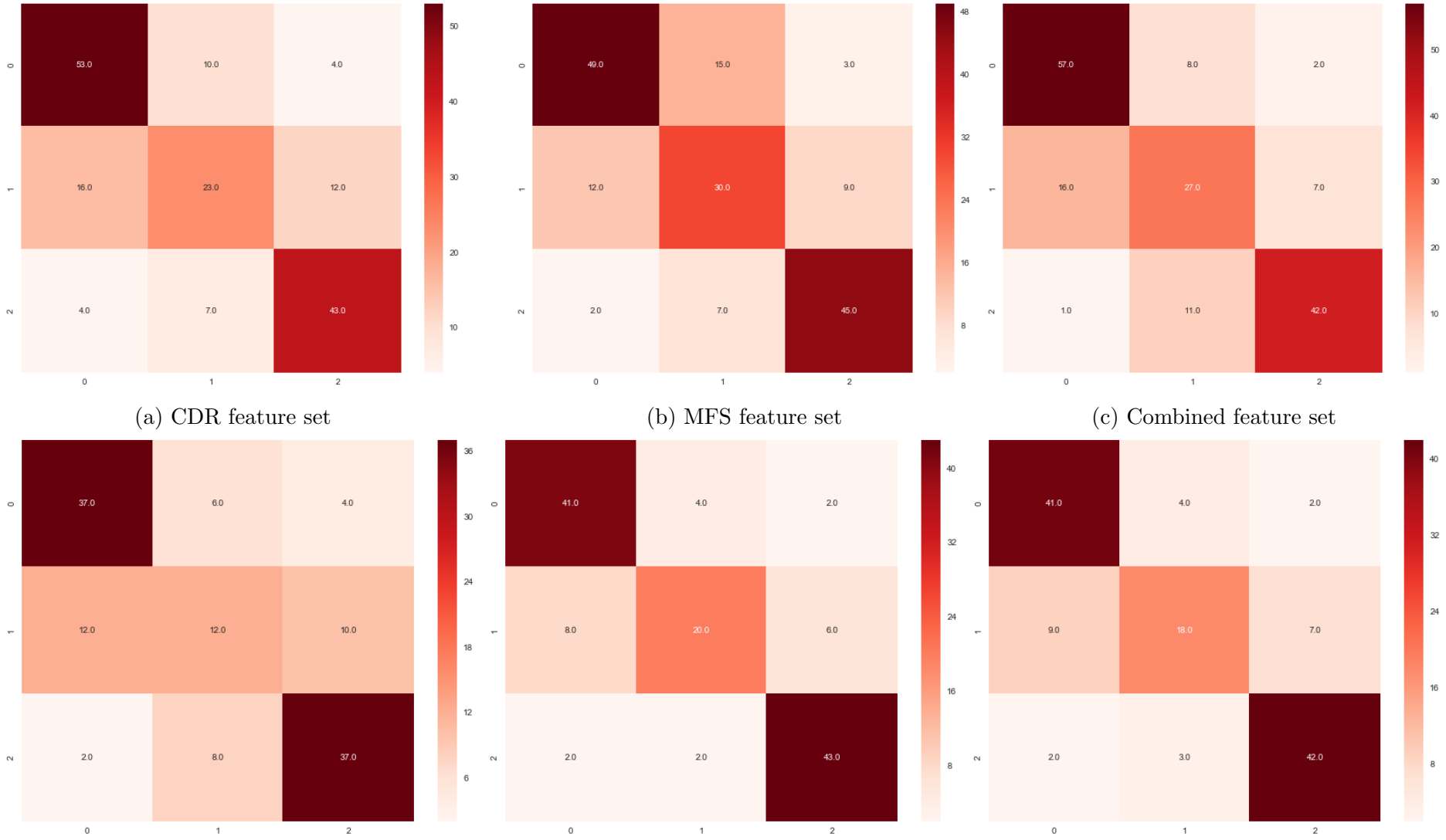
Model Feature Set	**Model Performance**	**Classification Accuracy**	**Per-class F1 score**
CDR features	f1 score: 0.63 precision: 0.64 recall: 0.64	65.9%	Poor: 0.74 Average: 0.43 Wealthy: 0.74
MFS features	f1 score: 0.7 precision: 0.71 recall: 0.7	71.3%	Poor: 0.77 Average: 0.54 Wealthy: 0.8
Combined (CDR + MFS)	f1 score: 0.71 precision: 0.72 recall: 0.71	72.3%	Poor: 0.78 Average: 0.55 Wealthy: 0.8
CDR features residential	f1 score: 0.63 precision: 0.64 recall: 0.64	67.2%	Poor: 0.74 Average: 0.38 Wealthy: 0.77
MFS features residential	f1 score: 0.71 precision: 0.71 recall: 0.71	73.4%	Poor: 0.79 Average: 0.51 Wealthy: 0.82
Combined residential	f1 score: 0.7 precision: 0.71 recall: 0.71	73.4%	Poor: 0.79 Average: 0.49 Wealthy: 0.82

while we can observe an increase in the performance of models containing MFS features, there are only marginal differences in using combined data sets. These results indicate that the MFS features appear to be a better indicator of underlying SEL compared to baseline CDR features. Further, it provides evidence that MFS features are subsuming the information contained in CDR features when it comes to SEL classification, and CDR is likely capturing only a subset of the variance covered by the MFS feature set.

Models trained on residential-only data clearly function more effectively than those also attempting to predict the level of poverty in non-residential areas (improving by 1.4%, 2.1% and 1.1% for Models 1-3 respectively), highlighting the difficulty in categorising non-residential areas with an SEL label.

The scope of improvements made by models leveraging MFS features is illustrated in Figures 4.7a and 4.7b, which show box plots of per-class F1 scores for all three feature sets in all six scenarios. We can observe in Figure 4.7a that F1 scores for middle-income areas (0.51) are on average $\sim$ 39% lower than scores for poor (0.76) and $\sim$ 42% for wealthy (0.78) areas. A similar, albeit even more pronounced trend, can be observed in the residential scenario with a low F1 score for average (0.46) and a difference in $\sim$ 50% to poor (0.77) and $\sim$ 54% to wealthy (0.8) areas. Predicting average areas is tricky, with most model features only supporting a binary poor/not-poor decision boundary.

Nevertheless, it is for middle-income areas that MFS features appear to make the most gains over CDR, with a large gap between Model 1 (0.43) and Model 2's performance (0.54) across all areas. The box plots in Figure 4.7a clearly show that the variance of CDR features in middle-income areas is extremely wide, rendering them ineffective in delineating 'average' areas from any other.



(a) CDR feature set

(b) MFS feature set

(c) Combined feature set

(d) CDR feature set - residential

(e) MFS feature set - residential

(f) Combined feature set - residential

Figure 4.2: Subfigures (a) to (c) illustrate represent the confusion matrices for all three feature sets for all BTS in Dar es Salaam that were used in this study. Subfigures (d) to (f) represent confusion matrices for all feature sets for areas classed as residential.

4.4.2 Variable Importance

Given the advantage that experiments indicate MFS features have over CDR, consideration turned to why this is occurring. To this end, permutation importance was investigated as per the methodology described in §4.3.7. For each model 1 feature within the test set (OOB samples) at a time is permuted (simulating holding out that feature without requiring a full model training) and the performance (F1 score) compared to the performance obtained with the full model. Note that the use of OOB samples provides the variable importance of the generalized predictor, rather than fitted (potentially overfit) model. Results for the combined feature set model are shown in Figure 4.7c, with the x-axis being interpreted as the mean increase in F1 score that the inclusion of the variable provides (assuming all other features are kept in the model). Six out of the seven most predictive features are generated via MFS data (in fact the only useful CDR feature relates to *number of users*, which is likely to reflect the population density in an area and can be drawn from other sources). The top 4 MFS features overall are: *average MFS in*; *active MFS users*; *percentage balance checks* and *average MFS spend*. The average income received by residents across an area clearly explains the most variance, contributing a 0.18 increase in F1 score. This is followed by the number of MFS users in second at 0.11; the percentage balance checks at 0.03; and the average MFS out at 0.01.

4.4.3 Understanding MFS Feature Effects

To break down the feature effects further Partial Dependency Plot (PDP) were used to visualize the increase/decrease of the probability of predicting a given output class (i.e. wealthy) when a factor is varied while all others are kept fixed. PDP are highly effective in showing us the model's sensitivity to the feature in question and how its predictions will respond as the variable's value changes. In each diagram, the heavily weighted line shows the mean change in the probability of a data-point being assigned to a particular class, as the variable increases. Considered in conjunction with the permutation importance scores, the PDP highlight the nature of the relationship between select MFS variables and SEL [126, 138].

As one might expect, the model identifies that as the amount of income of a BTS increases the likelihood of that area being affluent also increases (as denoted by the large areas under the curve for Figure 4.3a and c). However, the feature provides minimal information to the model for categorization for 'average' areas. There is, in fact, a relatively wide variance of *avg money in* in 'average' areas (which can be both middle-income areas or zones with combined informal/residential housing). Thus, we are left with a feature which provides binary classification - Figure 1a and 1c being mirror images of each other. There is a plateau to the informativeness of this effect, however, and as a predictor, its partitioning effectiveness peaks at $\sim 35,000$ TZS (which thus might well define a heuristic boundary between poor/wealthy users).

The number of residents who use mobile financial services in an area is negatively correlated with affluence, as illustrated in Figure 3c, where higher uptake of MFS increases the probability of a wealthy classification. This may initially seem counter-intuitive, but it is the 'unbanked' that have the highest propensity to need an alternative to traditional financial services. As wealthier users are more likely to make use of traditional 'brick and mortar' banking institutions than those 'unbanked' users on the lower end of the income scale. At the same time, however, the convenience of payment services such as M-Pesa for subscriptions may still have some of those users utilise MFS to pay for a range of services from TV subscriptions and airline tickets to school fees and utility bills for electricity and water. While the poor remain disengaged from credit card usage, mobile phone usage is nigh ubiquitous, even in slum areas, resulting in this effect.

The most important feature to the model in classifying middle-income towers is average MFS spend. It is this feature, which is likely improving the most over CDR data in the assessment of such areas. Figure 4.4 shows that while the model broadly associates an average MFS spend of $\sim 25,000$ TZS as the cut off line for poorer areas, it is only at $\sim 29,000$ TZS where the likelihood of assignment as a rich area begins to jump. In

between, the likelihood of a middle-income area being assigned increases. However, the feature is not decisive, and the true situation remains blurry to the model with two broad trends visible in Figure 4.4b. Half of the blue lines form a hump in the middle of the graph (reflecting a strict middle-income sub-population perhaps) dropping once spend crosses over a certain threshold, while the other half remains high (and thus will not be distinguished from residents of a wealthy area). Such fuzziness is likely due to the different types of 'average' areas that occur in reality - true middle-income areas, and those with distinct mixtures of wealthy and less affluent communities.

A further useful feature identified by the model is the average of balance checks users make to their mobile money accounts. As the number of balance checks increases, so does the likelihood of an area correctly classed as poor (although this is not a monotonic relationship). This is likely to be reflecting the fact that those living closer to the breadline need to assess their exact financial situation far more regularly than those who are more affluent.

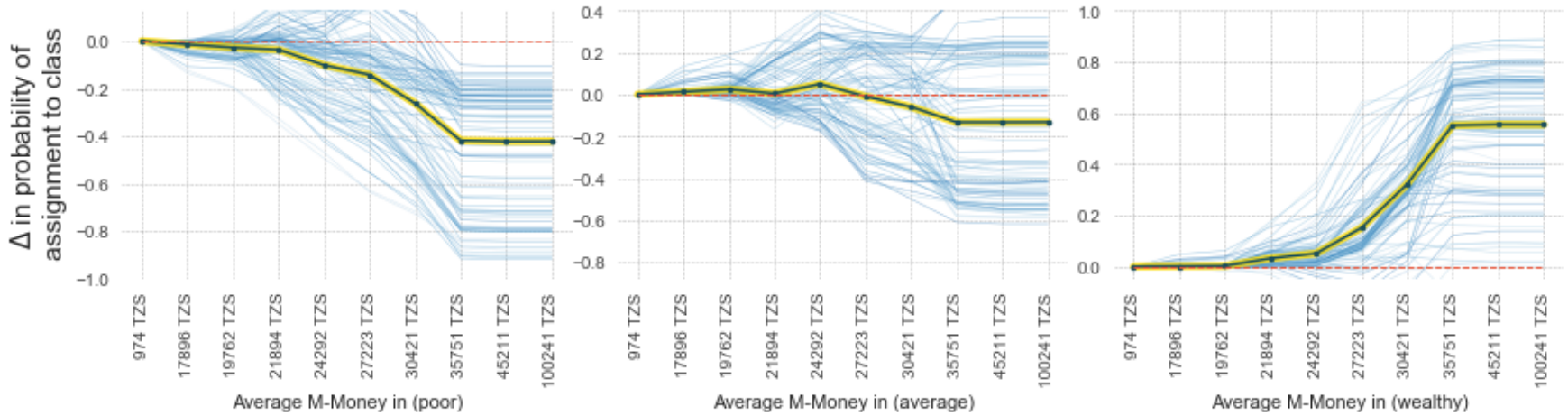


Figure 4.3: Partial Dependence Graphs highlighting the likelihood of an area being classed as poor, average or wealthy based on Average TZS received via MFS by residents of a given area.

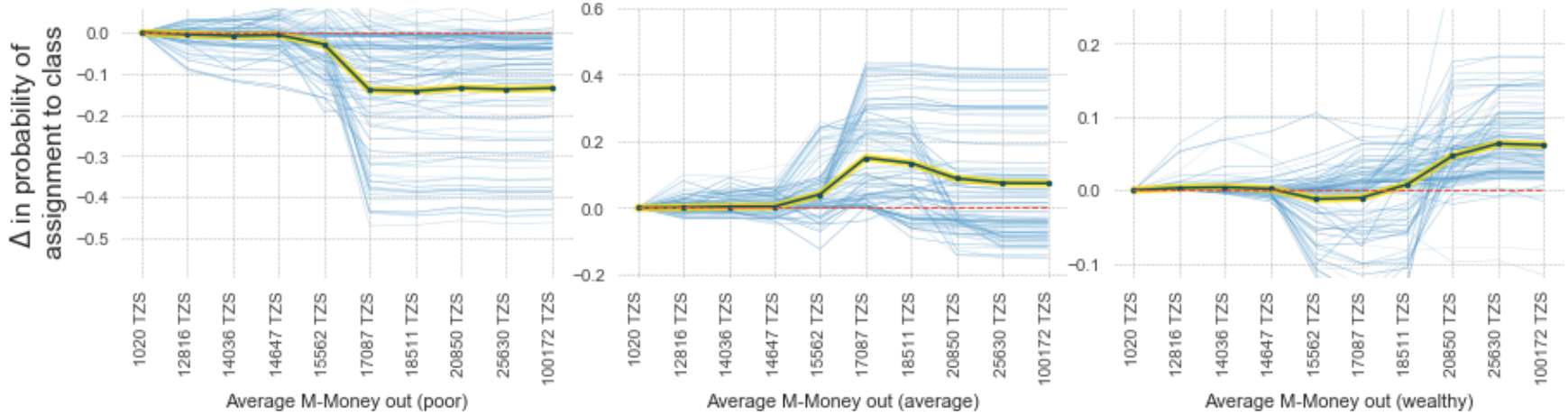


Figure 4.4: Partial Dependence Graphs highlighting the likelihood of an area being classed as poor, average or wealthy based on average TZS spending via MFS by residents of a given area.

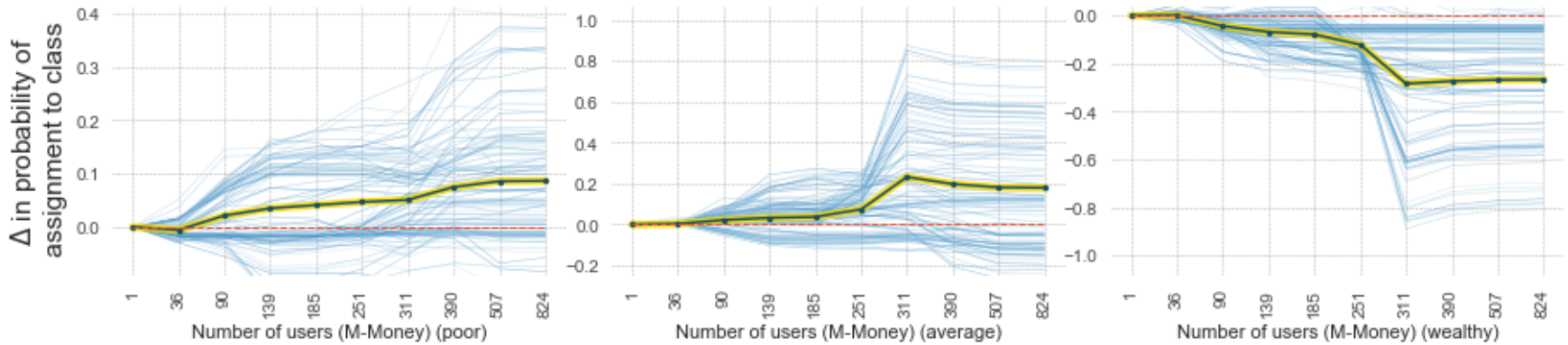


Figure 4.5: Partial Dependence Graphs highlighting the likelihood of an area being classed as poor, average or wealthy based on the number of MFS users within a given area.

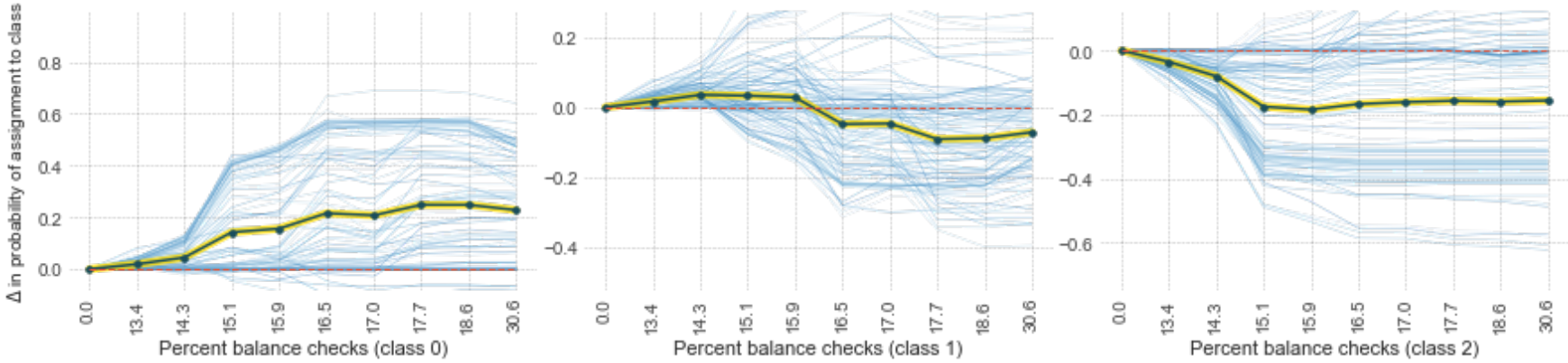
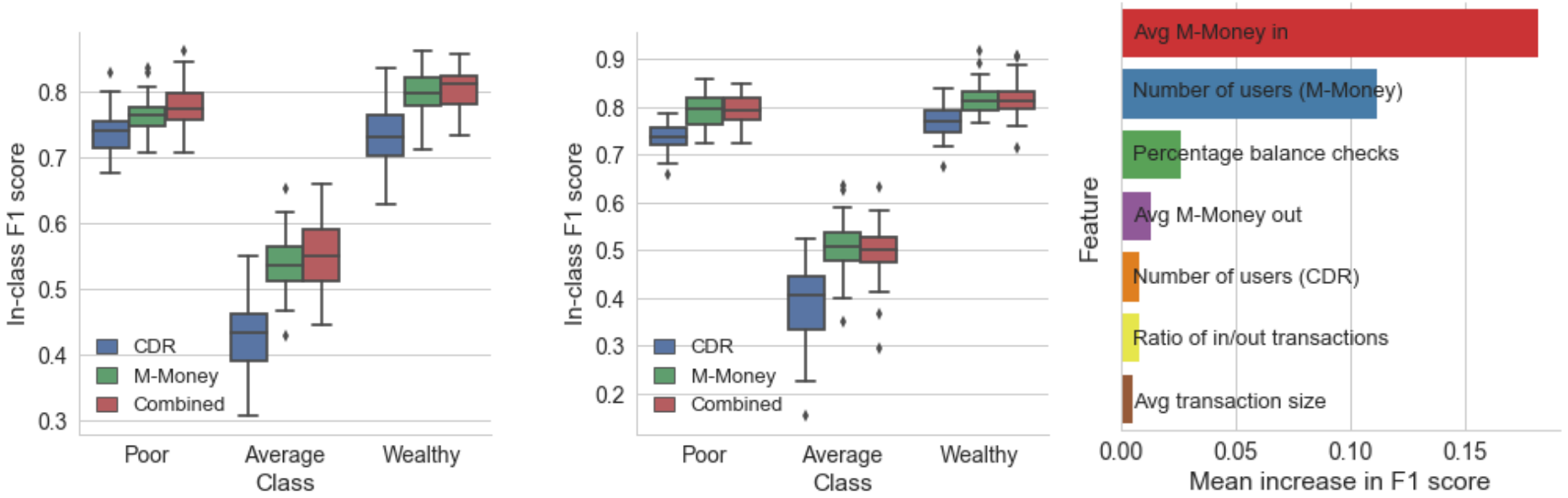


Figure 4.6: Partial Dependence Graphs highlighting the likelihood of an area being classed as poor, average or wealthy based on the number of balance checks among users within a given area.



(a) Per-class F1 scores across all BTS areas in the dataset

(b) Per-class F1 scores across all Residential areas in the dataset

(c) Variable importance for combined (best performing) residential model

Figure 4.7: Subfigures (a) and (b) illustrate the significant improvements made across all 3 classes (*poor*, *average*, *wealthy*) using M-Money rather than CDR features. Variable importances in (c) are for a combined CDR/M-Money model.

4.4.4 Study Limitations

User selection The study includes only MFS data for users also generating CDR records through call and SMS events due to the lack of geographic indicators included in the MFS data as geographic indicators were not present in MFS logs (although this is a symptom of our sample, rather than the raw data). Furthermore, user types such as subscriber or agent were provided by the MNO and accuracy was assumed. On the ground analysis, however, has indicated a large informal economy of street traders operate under the auspices of being regular subscribers. This led to the development of a stringent multi-step user selection process, which may have led to the exclusion of several legitimate high-frequency users from the analysis.

There are also potential issues around selection bias arising from mobile phone ownership among different groups of society. In urban spaces in Tanzania mobile phone penetration has been reported to be close to 92% resulting in a negligible uptake bias. Furthermore, the MNO providing the study's datasets has a 28% per cent market share, with over 70% of overall customers reporting that they had not used another network, alleviating issues of multiple-SIM ownership that are commonplace in emerging economies [245].

Accuracy of training data Supervised machine learning techniques require accurate data for training and test. As part of this study, a custom ground reference survey collected by volunteers within Dar es Salaam over a two month period was used. While there is often a gap in collection periods and granularity between the socio-economic data used as a poverty measure for validation and mobile phone data used to derive proxies, the survey used as part of this research was collected around the same time as the mobile phone data underpinning this research. Here, the sampled areas conformed to the same BTS derived Voronoi tessellation, assignment of these areas into clear SEL was often difficult for those collecting the data in (1.) the rural and peri-urban outskirts of Dar es Salaam, where large coverage areas of individual BTS are encompassing a mixture of affluence levels, and (2.) the dense urban center, where slums and formal residential areas

can sometimes be found in very close proximity (as shown in a sample satellite image in Figure 4.8). This issue of uncertainty over the accuracy of the ground reference data was further exacerbated as there was no clear indication of what distinguish a poor from an 'average or a wealthy from very wealthy for example, leading to a potential mislabelling of areas.

Data accessibility Many features require detailed knowledge of individuals' mobile phone usage, raising privacy concerns, and subsequently limiting the opportunity to obtain data from MNOs [154, 313]. While initial guidelines for best-practice engagement with third-party organizations to analyze CDR data have been developed by the GSMA [149] (in the wake of the recent Ebola crisis), overarching policies are yet to be established in regions where research can be of the most benefit.

Heterogeneity in usage Initial research by Jack and Suri (2011) [180] has shown, that MFS usage is equally widespread across both banked and unbanked populations. They do however, acknowledge that initiatives such as Equity Bank and Family Bank are increasing the availability to traditional 'brick and mortar' bank accounts for the previously 'unbanked' which could affect this balance in a similar way as light saturation has weakened the explanatory value of NTL as a proxy for poverty.

4.5 Chapter Summary

The analysis leveraged a range of regularity, diversity, activity and spatial features derived from 450.2m call and SMS event logs and 7.6m MFS records to generate insight into area-level socio-economic status. Using ground reference, CDR and MFS data, features were identified and quantified to accurately classify small-scale area SEL across the Dar es Salaam metropolitan area. The approach compared baseline metrics extracted from CDR data in line with previous approaches on harnessing CDR data for area-level poverty assessment discussed in Table 4.2.3 with features extracted from MFS data, combinations of those feature sets, and scenarios incorporating a differentiation in underlying land use.

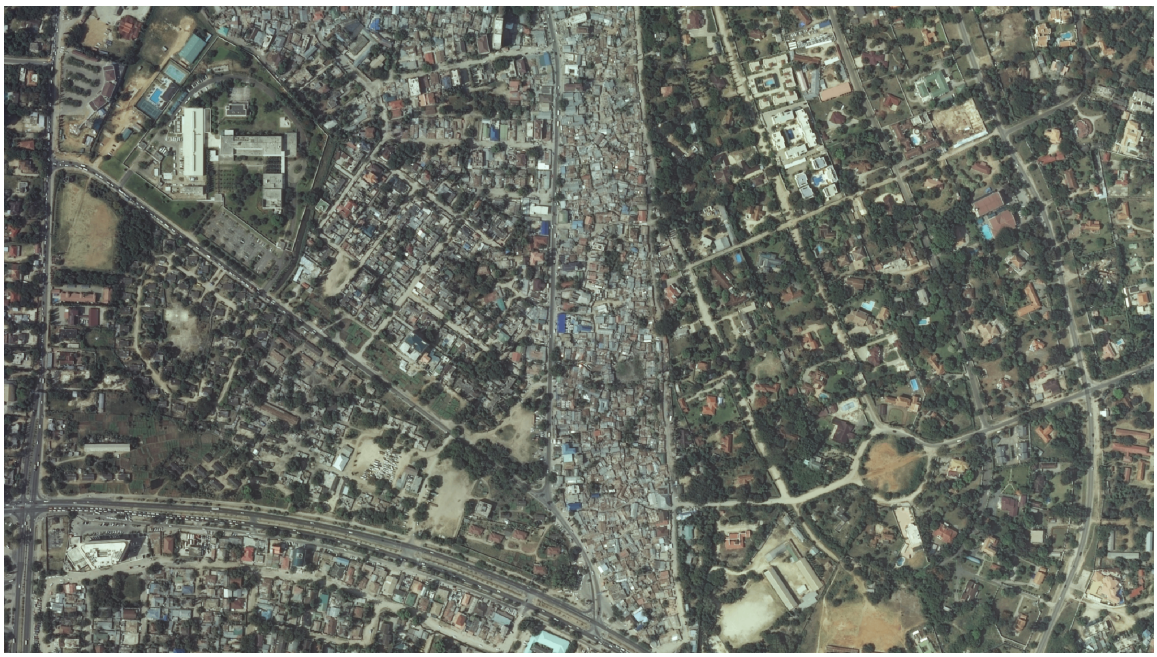


Figure 4.8: Sample area within Dar. A slum occurs in the centre with affluent housing left and right of the corridor.

This study demonstrated that mass MFS datasets could provide sizable improvements in SEL classification accuracy over recent CDR approaches. However, limitations remain, largely due to a need for more high-quality ground reference data for training. At the same time, this analysis will provide useful information to empower policymakers and local municipalities to identify areas requiring interventions and revitalization programs and assess the progress and success of those interventions, as according to the UNFPA:

> "any indicative estimates would provide in certain situations where none are currently available; even if they carried with them a significant level of uncertainty such estimates would still represent a large improvement in many cases" [313, p.519].

Comparing MFS data directly to CDR as used in prior work, the results show that MFS provides an increase in SEL classification accuracy (average F1 score) from 65.9% (0.63) to 71.3% (0.7) at a fine-grained spatial level. Notably, the combined use of MFS and CDR data only increased prediction accuracy (expressed through the average F1 score) from 71.3% (0.7) to 72.3% (0.71), providing evidence that MFS is informationally subsuming

CDR data. The following chapter will use CDR data to generate insights into transport demand and travel distance as endogenous factors for subsequent land use and socio-Economic – transport interaction analysis in chapter 6.

Chapter 5

Understanding Urban Mobility Patterns

5.1 Chapter Introduction

The previous chapter used CDR and MFS data to extract features from the socio-economic dimension described in §2.3.2 for subsequent area-level SEL classification and use in the land use and socio-economic – transport interaction analysis to be undertaken in Chapter 6. This chapter is using CDR data to extract some of the mobility variables described in §2.3.3. Specifically, trip distance and trip frequency measures for further analysis in Chapter 6 through the creation of OD matrices, one of the fundamental tools in transport planning, and synthetic daily activity plans. The following research objective guides this chapter:

Research Objective 3: exploration of synthetic daily activity plans based on the previously-evidenced assumption that the majority of human movement is predictable, and the generation of transient OD matrices to understand travel and mobility patterns for Dar es Salaam.

OD matrices represent the sum of trips undertaken in a specific time window between

given geographical areas and are used to inform transport forecasting models as a proxy of mobility demand [26]. Each trip is made up of two stops, an origin and a destination for the trip. A stop can simultaneously be a destination for one trip and an origin for another. Transient OD matrices are generated from stops and trips detected through a transient-based approach, which, unlike other approaches, identifies stops solely based on temporal parameters. While other approaches might only discover 'major' stops of home and work as one trip as a user returns home, a transient-based approach might identify the user stopping at a supermarket on the way, thereby identifying an additional trip (i.e. work - supermarket, supermarket - home). Traditionally, mobility insights have been derived from manual surveys, Road Side Interview (RSI), panel surveys or road traffic counts. These data collection approaches are slow, expensive, infrequent while only providing a micro-level snapshot of a particular period in time and space [147, 163, 164]. Rapid urban expansion make them particularly unsuited in emerging economies, where mobile phones are now nearly ubiquitous. CDR data similar to that used as part of this thesis research has shown huge potential for the analysis of mobility behaviour as will be discussed in the next section. An alternative to traditional surveying and sensor infrastructure generated data discussed in §5.2.1, they provide the opportunity to monitor human behaviour at scale, as the mobile phones have become nigh ubiquitous with human movement. This research goes beyond existing research reviewed in §5.2.2 by comparing three different approaches (transient-based, stop-based and frequency-clustering based) to OD matrix generation.

As a first step, background on mobility demand and transport forecasting is provided. Second, relevant literature on the analysis of different aspects of mobility behaviour using traditional data sources and CDR data is reviewed. Third, the chapter uses a sample of 329,530 mobile phone subscribers resident in Dar es Salaam, Tanzania, (450.2m call and SMS event logs) to derive travel demand in the form of OD matrices using both a synthetic plan and a more traditional transit-based trip extraction approach. In addition to CDR data provided by a MNO in Tanzania, the research builds on work conducted as

part of an industrial placement with the Transport Systems Catapult in 2017.

5.1.1 Mobility Demand

Increased mobility demand through urban expansion, increasing commuting distances and times caused by urban sprawl, and congestion are serious problems impacting the quality of life in cities around the globe [193]. While mass transit provides an opportunity to alleviate some of these issues, most of the worlds mass transit systems are ineptly designed if they exist at all due to a lack of understanding off local mobility demand and patterns [31, 298]. An estimated 35% of the worlds 100 largest cities do not have complete transit route maps, with the number increasing to 92% in the 100 largest lower-middle-income cities [204].

Tanzania, like many other emerging economies, is experiencing rapid urban expansion while suffering from poor mass transit provision, poor infrastructure and high levels of congestion [165, 281]. This urban expansion puts increasing pressure on often already limited road infrastructure leading to a range of issues such as severe traffic congestion, parking difficulties and increased traffic accidents. High levels of congestion and poor infrastructure conditions, in particular, have been recognised as major challenges within Dar es Salaam.

Besides being an inconvenience, endemic traffic congestion leads to huge economic costs, which account for 411 billion TZS ($USD195m) annually in Dar es Salaam alone [165]. The building of the Mfugale overpass at the junction of the Mandela Expressway and Julius Nyerere Road has significantly contributed to a reduction in traffic congestion in the ring road area on the way from Julius Nyere airport into the city centre along Julius Nyere road.

In emerging economies, informal bus services make up most of the mass transit provision. Which, while filling a void left by formal operators, give rise to their own set of challenges

[202, 207]. Up to 50% of bottleneck traffic is caused by small informal mass transit vehicles with low fuel efficiency that could instead be served by larger vehicles which are too expensive for small operators [40]. Low safety standards have contributed to African countries having the highest road fatality ratings globally [156].

5.1.2 Transport Forecasting

Transport forecasting models use OD matrices to help estimate future mobility demand and form the basis to determine the need for additional road capacity, changes to mass transit services and infrastructure, and land use policy and patterns by forecasting the impact of urban and transit interventions [64, 269]. Forecasting models "provide an average picture of the current state (average working day of a current year describing mainly the number of trips between zones, used modes and routes, link flows, travel times, and congestion problems), and a prediction of future states and expected effects of interventions" [363, p.68]. The models can be broadly categorised as trip-based, activity-based and system-based.

Trip-based models were first developed and implemented in the Detroit Metropolitan Area Traffic Study *RR-040A* in 1956 and Chicago Area Transportation Study *CATS* in the late 1950s and early 1960s on the back of major investments in road infrastructure and rapidly increasing car ownership and usage. They were predicated on a need for long-term aggregate level demand prediction of trips across modes and routes [269]. They are fundamentally aggregate - focusing on mobility trends between zones rather than individual mobility behaviour [292]. Trip-based analysis models are based on a four-step forecasting process:

1. **trip generation**
2. **trip distribution**
3. **mode choice**

4. **route assignment**

The first three represent the demand model and the latter, a network or infrastructure assignment model [263]. In the trip generation step, the frequency of origins or destinations of trips in each TMZ or other specified analysis area is determined by trip purpose as a function of either or a combination of land use, household demographics, and other socio-economic factors. TMZ have traditionally been designed manually based on basic principles of (1) cluster social, land use and economic characteristics; (2) frame zone boundaries around natural and man-made boundaries such as rivers and rails; and (3) selecting manageable traffic zone sizes; outlined by Ortuzar and Willumsen (2011) [269]. Geographical information, population data, land use characteristics, and socio-economic characteristics as used to divide the study area into TMZ [98, p.279]. During the trip distribution step origins and destinations are aggregated into *a priori* OD matrices. During the mode choice step the *a priori* OD matrices from step 2 are split by transport modes such as car, activity-based (e.g. walking, cycling, etc.) and rail. Finally, the route or traffic assignment step is used to allocate trips to potential routes between and origin and destination of a trip along the existing road and rail infrastructure.

Activity-based models were first developed as part of the model improvement program in the early 1990s based on theoretical frameworks proposed by Hgerstraand (1970) [157] and Chapin (1974) [120]. They are the current state of the art approach for travel forecasting as they bring with them a shift from aggregate statistics and relationships of trip-based models to dis-aggregate models and micro-scale simulations. In activity-based models, travel is considered in the broader context of activity scheduling in time and space, and as an intrinsic component to satisfy the economic, physiological and social needs of an individual to participate in spatially distributed activities [52, 64, 158, 278, 292].

In contrast to trip-based models, activity-based models consider travel as dis-aggregate. The focus is on the daily trajectories, the chronological chain of users trips, rather than aggregate summaries of individual trips between two points in space. They expand on

trip-based models by considering the activity-motifs (i.e. the reason for conducting a trip in the first place). Activities are defined to reflect basic personal and household needs and are usually categorised as basic motifs such as home, work, education, recreation, shopping, etc. Individuals are presumed to follow weekly/daily activity schedules and to optimise trips to perform all activities with the required frequency taking into account time constraints and available transportation and activity location infrastructure which is shared with other individuals [131, 359].

System-based models, per-driver or per-agent models are based on initially derived daily activity plans comprising activity locations, activity start time, duration and end time, and mode and routes of trips connecting activities. Those plans are then fed into microscopic time-dynamic traffic simulations to prescribe synthetic daily plans based on system constraints given by the transport network and its attributes [28]. These synthetic plans and individuals or agents can then be used for testing of future transportation, land use and smart city concepts. The most common simulation environments are SimMobility [3], SimAGENT [146], and MATSim [170], which combine mode, time, destination and activity scheduling processes into a single consistent framework.

5.2 Literature Review

5.2.1 Traditional Data Sources

Data for transport forecasting models have traditionally been collected through manual surveys, fixed sensor infrastructure or mobile sensor tracking. Each of those is discussed in more detail below.

Manual sampling

Manual surveys and traffic counting are the most widely used sources of data for trip generation. They are undertaken when the use of automated methods is not feasible due to cost and effort implications or to gather data which cannot be obtained effectively through

automated methods such as occupancy rate, travel mode classification or pedestrians.

Traffic counts are generally collected over a 'representative' one-day period when travel flows are maximal without accounting for reasons for travelling or tracking the same individuals to assess long-term changes [269].

RSI are similar to traffic counts in that they involve a small sample size over a time period of a few hours at any given location [58]

Panel surveys are conducted at regular intervals, effectively tracing the same individuals over time. They provide opportunities to analyse (changes in) and model travel behaviour, while often leading to completion fatigue, affecting both adherence and accuracy [269].

Manual surveys are costly to undertake and quickly outdated in addition to being affected by sampling biases and reporting errors [147, 163, 164]. The presence of new travel trends is frequently only discovered after the release of new results [35].

Point detection methods

Point detection methods involve the use of spatially-fixed sensors for traffic data recording. They help overcome the issues of data collection at sparse temporal intervals and human observation bias inherent to manual data collection discussed in the previous section. Prime data sources are point-based sensors such as loop detectors and overhead detectors, which allow for the recording of occupancy rate, volumes and speed for particular road sections.

Inductance loops are an expensive but reliable technology for measuring traffic along roads. Installation and maintenance require the closure of lanes.

Video image detection was introduced to deal with some of the shortcomings affiliated with inductance loops. Cameras are mounted above ground but do not work as well in low-light and poor weather conditions.

Microwave radar technology was mainly developed to overcome the shortcomings posed by inductance loop and video image detection. While not being affected by low-light and bad weather, it reports traffic counts with less accuracy than other point detection methods.

While recording basic metrics, point-based methods do not allow for the generation of time-space trajectories as they are unable to track specific cars as they are moving through the sensor network. The only point-based data source allowing for tracking of individual vehicles are automatic number plate recognition systems [129]. They are not widely used, however, due to high installation costs and privacy and data protection issues. Bacon *et al.* (2008) [24] examined the feasibility of wide-scale deployment of static sensor infrastructure for traffic monitoring and found it unfeasible due to the high cost involved. Ehmke *et al. 2010* [108] have further found that data from static sensor networks need to be supplemented with additional area-wide traffic data collection due to the large gaps in static sensor infrastructure provision. Equipping cars with sensors and networking capabilities were found to be a much better solution at the time.

Vehicle based detection methods

GPS technology has been the core focus of automated data collection methods for more than 20 years. A lot of work has been undertaken using GPS traces from mobile phones, busses and taxis for mobility analysis in particular, as it is easier and more cost-effective to deploy than static sensing infrastructure [24, 259, 365].

GPS data has been applied to a wide range of methodological components inherent to transportation forecasting as discussed in §5.1.2 including: Trip generation [21, 323, 364] and activity-motif detection through matching of trip end locations with land use data [364], Route assignment [167, 184, 216, 287], Travel mode detection [296, 337, 358] and the calculation of expansion factors [55, 365].

While GPS traces generally boast high levels of spatial and temporal accuracy and there-

fore Spatio-temporal insight about individuals movements, the attainable sample size and observation periods of GPS-assisted surveys are still limited [24, 37]. This is partly due to the high privacy concerns that come with the recording of GPS data, as it is possible to track individuals at all times fully [283, 284]. Another issue is the buy-in nature of most GPS based studies, as data for GPS travel diaries are generally recorded through specific mobile phone applications, which require user buy-in on a large scale to generate meaningful levels of data. Opportunities exist in passive data collection through Automated vehicle location sensors, which are increasingly installed in regular vehicles for security purposes [200, 254, 299].

5.2.2 CDR Derived Stop Extraction and Trip Generation

Section 5.1.1 introduced the concepts of mobility demand and shortcomings of its analysis in areas facing rapid urban expansion. CDR data allows us to observe combined trip generation and distribution directly, and to a certain extent also route choice for a subset of the population that are captured by the data. Further, several studies have discussed two core and one ancillary method to estimate mobility demand using CDR data. Transient-based, Stay-based, and Frequency-based clustering. As discussed in the chapter introduction, both transient-based and frequency-based approaches were chosen as applicable tools to derive mobility demand for Dar es Salaam. This is due to the more stringent spatial requirements akin to individual handset triangulations afforded by stay-based approaches, as opposed to transient-based and frequency-based clustering, which can be undertaken using BTS-level CDR data. Each of the three approaches is used to identify trips.

> A "trip(u,o,d,t) is characterised by a user ID u, origin location o, destination location d and starting time t" [60, p.38].

Each origin location o and destination location d is defined as a stop. In the transient-based and frequency-based clustering approach stops correspond with BTS associated with one or multiple network events while they are often referred to as virtual locations

rather than stops in the stay-based approach. Each stop can be a destination for one trip and an origin for another at the same time. A trip consists of a pair of non-identical consecutive pairs of stops (an origin and a destination). A user has n-1 trips per day, with n being the number of detected stops for that day. The chronological order of stops is a user's daily trajectory.

OD matrices are generated through the aggregation of trips into predefined regions and temporal windows. The OD matrix count is the flow, the sum of trips between matching o and destination location d pairs during the specified period. As multiple approaches for trip extraction exist, it is important to understand how they are related to understanding their subtle differences in real-world application utility. Accordingly, the following subsections explain the vocabulary, definitions and methods of transient-based, stay-based and frequency-based clustering.

Transient-based approach

Transient-based approaches are sometimes also referred to as temporal-based or cell-flow clustering approaches. Rather than simply selecting consecutive network events with non-identical BTS to identify stops equaling to origin locations o and destination locations d, this approach uses temporal filters to identify meaningful locations.

As discussed in more depth in §2.2.1, CDR data does not record individual's locations, but rather a non-precise proxy via the location of the BTS delivering the service. Sector antenna range is dependent on factors relating to the location, height, and the technology involved. Limitations in range are generally accounted for by MNO through intentional overlaps in service area to reduce the risk of 'holes' in signal coverage. Handsets will automatically seek the BTS with the strongest, and therefore generally nearest, signal, which can potentially result in false displacement where a user may appear to be moving while remaining stationary in the physical realm [109]. As a result and due to the uncertainty over a subscribers proximity to any specific BTS, location information cannot be

assured, and devices may appear to be moving rapidly between surrounding BTS at times.

Within the transient-based approach, temporal filters are used to filter out noise that may simply be an artefact of those rapid 'location jumps' between surrounding BTS. Transient-based approaches resemble trip-based approaches that identify segments of travel based on limited data that lose much of its value in areas with a lower spatial resolution (i.e. low BTS density) and high road network density. As filtering within the transient-based approach is limited to the temporal scale, the approach is used to capture intermittent (transient) points in addition to overall origin and destination points of a users daily trajectory.

One of the first to apply a transient-based approach was [305], who used floating phone data on LAU sequences from several months of phone activity of T-Mobile (Germany) customers to attempt to close an initially perceived gap in the monitoring of long-distance trips above 20km. They used three different temporal rules: a 60min-rule (remaining in a location area longer than is necessary to traverse the area potentially has interzonal trip), an extended 60min-rule (two or less LAU, more than 60 mins between the first log in and the last logout), and a jumpiness rule (LAU/unique LAU >2 or 3+ unique LAU).

Among the first studies to use CDR for transient OD estimation was undertaken by Wang *et al.* (2012) [347] using data from 360k Bay area users and 892 towers as well as 680k users in 750 census tracts in Boston over three weeks. They applied a frequent-sequence mining algorithm to identify frequent mobility trips for individuals. A trip was extracted for BTS displacement in subsequent records of a user taking place within 10 minutes and 1 hour. The 10-minute re-sampling rate was used to alleviate the effects of localisation errors and event-driven location measurements on individual trip determination similar to Calabrese *et al.* (2011) [58]. False displacement effect can occur during peak times when users are sometimes transferred to towers that are further away in order to balance the load on the cell tower infrastructure. The upper bound of 1 hour was used to mitigate

the effects of false displacement.

Using temporal variation of association rules, Fritz-Martinez *et al.* (2012) [125] infer OD matrices from aggregated CDR for calls, SMS and MMS for a 2-month period from October to November 2009 for 3.5m unique handsets in Madrid. The rules were first introduced by Agrawal *et al.* (1993) [5] and used to identify home-work and work-home commuting trips. They performed *a priori* OD matrix validation on home-work commuting time windows using National Statistical Institute (NSI) mobility matrices for the state of Madrid.

Bahoken and Raimond (2013) [26] used CDR data for calls and SMS for 10 million users over six weeks in France. They extracted all trips, which involved a displacement within the space of 24 hours, and from 6 am-10 pm and 4 pm-8 pm. The time windows were chosen to estimate travel demand over a day, as well as to identify demand for home-work, and work-home commute. Unlike the more general definition of flows as the sum of all trips in a specified time for a particular user by Calabrese *et al.* (2011) [60], they calculate flow using the first and last two points of a users trajectory for the given temporal window.

Following a similar approach to Wang *et al.* (2012) [347], Iqbal *et al.* (2014) *et al.* [176] use traffic count, and CDR data for 2.87million users in Dhaka, Bangladesh over a month to estimate node-to-node transient OD matrices. 10 minutes were chosen as a minimum time in order to reduce the false displacement effect. 1 hour was chosen as a maximum time for a trip to ensure that only meaningful trips are inferred. CDR data is used to generate *a priori* OD matrices before using traffic counts generated from video vehicle detection to estimate final *posteriori* OD matrices for four time periods: 7am-9am, 9am-12pm, 3pm-5pm, 5pm-7pm. The optimisation based approach was chosen to minimise the difference between observed and simulated traffic.

Larijani *et al.* (2015) [209] build on prior work by [26] by examining active and passive

network event logs (which include active CDR, and passive handover and LAU) for 1.4 million phone users in Paris over the course of a single day, with a particular focus on morning commutes (6 am-10 am) and afternoon commutes (2pm-9pm). While having access to 57 million logs, they identified nearly 50% of network events as being LAU.

Stay-based approach

Stay-based or time-distance clustering approaches are very similar to transient-based approaches, in that they are used to identify the trip start and end locations from consecutive CDR records. In addition to temporal filters, spatial filters are used to further reduce noise in the identification of meaningful locations where a user has stopped at the beginning or end of a trip.

> The maximum diameter "controls the spatial resolution at which stays are identified. Higher values for D result in lower spatial resolution with a larger upper bound for the area applicable for a stay" [228, p.788]

This follows the same approach as with GPS data, where boundaries are defined in line with positioning errors and minimum dwell times [160]. In contrast to the transient-based approach, the more strict spatial requirements lead to this approach, only capturing travel between significant stay areas. Those stay areas can encompass one or more BTS within the spatial parameter used for filtering, with the centroid often selected as the stop or virtual location. This allows the summarising of consecutive network events that are close enough into a single stop - activity location identification by filtering out passing-by points/trips obtained from flows between stay locations. Identifying stay locations where people conduct activities can help overcome those limitations by reducing noise in the data [379].

As one of the first to apply a transient based approach, Calabrese *et al.* (2011) [60] examined CDR for 1 million users in the Boston Metropolitan area containing 829 million CDR logs including BTS and a triangulated position for each network event provided by the US-based company *AirSage* collected over a period of 4 months. In addition to a

temporal low-pass filter of 10 minutes to alleviate the false-displacement effect similar to transient-based approaches discussed above, they applied a spatial filter of 1km to cluster locations and further reduce the false displacement effect. The data provider *AirSage* identified a 1km uncertainty radius to be optimal in reducing the false displacement effect during a pre-processing step common with GPS trace analyses. They chose the centroid of all points within a cluster as the virtual location. Agglomerative clustering was used to consolidate stars into a single semantic location regardless of the temporal sequence of the CDR records. Each non-identical consecutive pair of virtual locations corresponds to a trip. While agglomerative clustering to identify virtual locations did not consider the temporal sequence, the daily trajectory is determined by the chronological order of virtual locations.

Jiang *et al* (2013) [185] used a grid-clustering approach in their analysis of approximately 835m CDR for 1 million users over two months in Boston in 2010 for a more coarse-grained analysis of travel motifs, preferential return and explorative characteristics. As part of the grid-clustering approach, the entire study region was split into stay regions with a roaming distance of 300m, corresponding with the maximum distance between any two BTS collecting location data in the study area and as a reasonable walking distance for activity detection. First, spatially close points are clustered by Euclidean distance between consecutive CDR and compared to the spatial threshold set at 300m. Second, clusters are considered stays when the time between the first and last CDR within the cluster is greater than 10 minutes. Trips were identified if there was a change in stay region and a minimum separation of 10 minutes between the first and the last record.

Comparing both transient-based and stay-based approaches, Maldeniya *et al.* (2015) [228] used 1.4 billion CDR records from 10million users from different network operators in Sri Lanka to test approach efficacy to derive *a priori* OD matrices. The average number of network events per user per day was sparse, with around 25 events per user on average. Ten minutes between displacement records was chosen as a low-pass filter to alleviate

false displacement effect, where calling parties appear to be moving due to connecting to different towers while remaining stationary. An upper-bound of 1 hour was chosen to ensure meaningful trips were detected. 1km grids were chosen as a trade-off between spatial granularity and noise reduction due to localisation errors (particularly in high-BTS density areas) similar to the $300m \times 300m$ grid approach chosen by Jiang *et al.* (2013) [185].

Alexander *et al.* (2015) [12] use a stay-based approach to estimate OD's for four periods (AM, midday, PM, night) and purpose (home-based work, home-based other, non-home-based) from triangulated data for 2 million users over 2 months in Boston, USA. Their results indicate a strong correlation between trips inferred from CDR data and ground truth data from Census Transportation Planning Products (average of 0.5) and NHTS (over 0.95) similar to the MODLE project (§5.3.1.

Using the same Boston, USA dataset as [60, 185] and additional triangulated location management data for 1 million customers for two weeks in 2012 Vienna, Austria, Widhalm *et al.* (2012) [359] combined a low-pass filter with an incremental clustering algorithm for robust stop extraction and daily trajectory identification. In order to account for a potential low-frequency bias, they chose to exclude users with less than six observations per day similar to [306]. The chosen approach considers the geometry of trajectories and travel speed to adjust clusters to identify activity locations incrementally. In a follow-on step, they combined the derived OD flows with external land use data to model location choice behaviour for inclusion of *a posteriori* OD matrices in activity-based transport forecasting models as discussed in §5.1.2 [197, 242, 271].

More recently, Wismans (2018) [363] used pre-aggregated data for one month for about a third of dutch mobile phone subscribers provided by the data provider *Mezuro* for the Netherlands to inform stay-based OD matrices for November 2014. Ground truth data from the annual dutch NHTS *OViN* survey and the Rotterdam transport model is used

for validation. Rather than using a low-pass filter of 10minutes as related to previous research, the period for a trip is set as the mid-time between two consecutive stays of over 30 mins in different areas. Location data was aggregated at the village or district level, splitting the Netherlands into 1259 study zones. Scaling was undertaken using a combination of trip-length distribution between *OViN*, and CDR derived OD counts based on the assumption that both are inherently biased to some degree (§7.6.3 resulting in an increase of trip counts between 8km and 13km, and a reduction of trip counts for those between 14km and 40km long.

Frequency-based clustering approach

As location information is not directly observed but rather only collected when a user actively engages with a mobile device through the creation of a network event, CDR data contains a large amount of hidden movement. A user could use his or her mobile phone at one point for example and then be in a completely different part of a city 6 hours later without us having an idea how the person got there or what happened in the meantime. The issue of hidden movement is especially pronounced with users with a low number of irregular network events [63]. It can have similar effects to the oscillating ping-pong effect caused by signal strength challenges that are not unique to sector antenna in the underlying GSM network [211].

The frequency-based clustering approach indirectly addresses this phenomenon by seeking to extract frequent (sequential) virtual location visit patterns. Rather than applying temporal and/or spatial filters to consecutive CDR records, BTS are instead ranked by visitation frequency to derive frequently visited points of interest such as home and work and to extract synthetic daily activity plans representing a common trajectory for each mobile phone subscriber. This approach is based on the key assumption that people have regular patterns of mobility the physical locations of POI and geographies such as home and work [6, 34, 142, 314].

Bayir *et al.* (2010) [33] used the MIT Media Lab Reality Mining project dataset that tracked BTS locations of a 100 students equipped with Nokia 6600 feature phones over nine months for path construction. They observed the time spend at each BTS, transition time between adjacent BTS and observed and hidden destinations with duration's above-set thresholds. Weight-based hierarchical graph clustering is used to group adjacent BTS clusters and prevent the ping-pong effect before an AprioriAll [5] is used to identify frequently visited stay areas. Resulting individual mobile phone profiles for each student are specified by day of the week and specified periods of 6 am-12 pm, 12 pm-6 pm, 6 pm-12 am, and 12 am-6 am. Similar to existing research by Gonzalez *et al.* (2008) [142], human trajectories were found to have a high level of regularity with an average of 85% of observed stay time was at and between top-ranked BTS locations. Observed mobility profiles are heavily biased toward specific locations; however, due to the limited collection area, which mostly focussed on the MIT campus in Boston, MA.

Isaacman *et al.* (2011) [177], used a cluster-leader algorithm to rank the frequency of BTS visits. In order to assign semantic meaning to the identified points, a logistic regression model trained on a ground truth sample was used. The metrics included the number of contact days during the study period, the period between first and last contact, and inferred home and work 'visits'. Despite the ground truth, a confusion between significant and pass-by points remained.

Using Data from the D4D challenge datasets described in §2.1.1, Berlingerio *et al.* (2013) [40] examined CDR for 50k users over a period of 2 weeks. They identified frequent travel patterns that are outside the realm of existing public transit infrastructure in order to identify service provision gaps. First, they determined travel and activity patterns by deriving frequent virtual locations and patterns of mobility between those. Second, activity patterns and OD flows were aggregated to design new transit services. Finally, they evaluated how well different categories of users were served by public transport services in the area.

Schneider *et al.* (2013) [306] adopted 'motifs' from complex network theory [243] with the assumptions that a) the daily trajectory starts and ends at the identified home BTS based on the intrinsic need for sleep, and b) each stop is visited at least once. Similar to previous research by Bayir *et al.* [33], time spend at a stop and transition time between similar activities or motifs is incorporated in the development of a model (perturbation-based) to reproduce the frequency of a motif occurring. BTS with more than three oscillations in a day were merged. The most frequently visited BTS in a 30minute segment on weekdays was selected as a stay. Any day with less than eight observations was discarded so as not to bias motif generation. A home BTS was identified as the most frequently visited between midnight and 6 am.

5.2.3 From *a Priori* to *a Posteriori* OD matrices

The stop extraction and generation approaches described in the previous section are used to generate *a priori* OD matrices representing travel patterns for a limited sub-sample of the population. Scaling is necessary to generate more representative *a posteriori* OD matrices from the sub-sampled matrices. The following section introduces scaling approaches, and ways to validate OD matrices.

Scaling of *a priori* OD matrices

The scaling of *a priori* to *a posteriori* OD matrices requires the use of external data to calculate appropriate expansion factors. The most commonly used expansion factor in related research has been based on the ratio of the census populated and identified BTS-based [12, 186, 347] or Location area [376] based home users within a census area. Calabrese *et al.* (2011, 2013) [58, 59] for example compared population density calculated from CDR data with census level population data and found that biases are exacerbated in areas with low cell phone penetration. Census-derived expansion factors have a high risk of introducing significant biases due to the (temporal) gaps in census collection,

making them particularly unsuitable for application in fast-changing urban environments in emerging economies. Even in a western context outside Africa's 'Statistical Tragedy', however, they can be of limited utility as became apparent in a study of the Bristol region as part of the Mobility on Demand Laboratory Environment (MODLE) project, where developments under study only emerged following the last census collection.

A second approach exists in the derivation of expansion factors from traffic volume counts collected manually or through road-side detectors at major node points [128]. From traffic counts, expansion factors can be calculated either iteratively [225], through correlation analysis [176] or as an optimisation problem [176, 335] using simulations.

Alternatively, Cai *et al.* 2017 [57] have used mobile phone ownership and MNO penetration rates to generate expansion factors. They calculated the expansion factor as

$$OD_{people} = \frac{OD_{mobile}}{\alpha_1 \alpha_2 \alpha_3 \alpha_4}$$

With OD_{people} as the scaled passenger volume between each OD pair, α_1 as the number of mobile phones per person = 1.077, α_2 as the penetration rate of mobile phone usage among all residents = $min(\frac{\text{number of mobile phone users}}{\text{permanent residents}}), 1$, α_3 as the market penetration of the MNO, and α_4 as the probability of a mobile phone being detected = 0.84.

Validation

Following the generation of scaled *posteriori* OD matrices, related studies have sought to validate their matrices using travel survey data such as *OViN* or NHTS. Early approaches sought to use a combination of GPS data, GIS data, and individual and household demographic data for validation [48, 74, 77, 140]. Calabrese *et al.* (2011) [60] and Frias-Martinez *et al.* (2012) [125] used external NSI OD matrices for validation to compare time intervals for home-work commuting.

Using NHTS data, Gong *et al.* (2015) [141] compared trip purpose distribution of NHTS

trips and semantically labelled CDR derived trips. In an alternative approach, [363], compared CDR derived *a priori* OD matrices with *OViN*, the dutch NHTS and used it to improve traditional models of transport forecasting, finding a particular under representation in short trips (§7.3.1) similar to validation results in the MODLE project discussed in §5.3.1.

Mapping of *posteriori* OD data

CDR derived OD matrices commonly represent transport demand on either the BTS level or between Voronoi cell areas depending on the underlying datas spatial granularity and trip extraction approach. In related research, these OD matrices have been spatially interpolated to pre-determined zone structure such as grids, census tracts of traffic analysis zones [12, 58, 76, 228, 374]. Interpolation to specified geographical representations has been undertaken to either align CDR derived OD matrices with existing TMZ for integration into implemented transport forecasting models and/or to protect individual privacy and commercial interest for visualisation of derived mobility trends.

Using triangulated device data, Jiang *et al.* (2013) [185], used a grid-clustering approach with $300m \times 300m$ grids as part of the stop extraction and trip generation process prior to the aggregation of trips to OD matrices. Maldeniya *et al.* (2015) [228] instead use spatial interpolation of generated trips as part of the *a priori* OD matrix generation to distribute trips among pairs of existing TMZ in their comparative study using CDR data from Sri Lanka.

While helping to preserve individual privacy, which is a particular concern with triangulated device data used in stop-based approaches [60, 185], and commercial interests such as the BTS location, this approach has been shown to lead to noisy and unbalanced OD matrices when aggregating/apportioning to small spatial areas [76, 374]. As Alexander *et al.* (2015) [12] have shown, this concern can be alleviated somewhat when aggregating to larger zone sizes. An alternative to spatial interpolation of CDR-derived OD matrices to

pre-determined zone structures is the mapping of CDR derived BTS Voronoi polygons to the nearest road network node or metro station [176, 209].

5.3 Research Approach

Following the empirical analysis of activity-based land use conducted in chapter 3 and socio-economic levels in chapter 4, the analysis focussed on BTS for the Tanzanian port city of Dar es Salaam. A core dataset of 565 BTS located in the metropolitan area of Dar es Salaam were identified. Two testing scenarios were formed, one considering a transient-based approach for OD matrix estimation, and the second considering a frequency-based clustering approach. In addition to OD matrices as a proxy for the inbound and outbound trip attraction of an area, travel distance metrics are derived from the synthetic daily activity plans created through the frequency-based clustering approach as dependent variables in the land use and socio-economic – transport interaction analysis in Chapter 6.

Based on a fixed definition of origins, destinations and trips, as well as understandings of the transient-based and frequency-based clustering approach outlined in §5.2.2, the following high-level process was undertaken to construct OD matrices and synthetic daily activity plans from CDR data:

1. **Data Cleansing:** were raw CDR was converted into CDR time series. As part of this step, BTS in very close spatial proximity where the location of a user cannot be distinguished were merged. Data was further filtered to comply with the set temporal and spatial thresholds for this study and low and very high-frequency users excluded to prevent the introduction of frequency-biases.

2. **Stop Extraction:** Both transient-based and stop-based approaches were applied to extract BTS represent potential *origins* and *destinations* for trips at the individual level.

3. **Trip generation:** Trips, which do not meet predetermined thresholds are removed at this stage.

4. ***A priori OD matrix generation*:** the trips generated as part of the previous step are filtered depending on the chosen application before being aggregated into an initial OD matrix representation. Temporal (i.e. weekday/weekend and by time of day) is undertaken as part of this step.

5. **Scaling:** raw OD matrices are scaled using traffic count data to more appropriately scale to the number of trips detected to the full population from the sub-sample represented in the *a priori* OD Matrices.

6. ***Posteriori OD matrix generation*:** scaled *a priori* OD matrices are interpolated to the Ward Level

This process was similar to the data cleansing and OD generation process used as part of the MODLE project discussed in more detail in §5.3.1. Here, data processing included four key stages.

Data cleansing CDR and location management data was selected for neutral days (i.e. weekdays that are not public holidays) during the study period. Event data was converted into dwells and journeys. Internet of Things devices, tablets and users on business contracts were removed from the analysis.

Stop extraction Identification of points of interests from frequent visits and categorisation of POI into the home, work and other. Using each users home location, an expansion factor calculated as the number of homes in an area divided by the census population was calculated.

Trip generation and semantic annotation Journeys were categorised by purpose according to origin and destination POI category. Identification of journey mode for road and rail trips. Using these journeys, only trips that penetrated the study area cordon were selected. Those trips were assigned to one of the four-time of day sections according to their start time.

***Posteriori* OD matrix generation** Journeys were aggregated by their respective origin and destination, trip purpose, time, period and travel mode to form OD matrices.

5.3.1 Data Description

This study uses mass CDR datasets for the generation of OD matrices and daily activity-plans as proxies for the creation of mobility variables for further analysis in Chapter 6. The spatial granularity for the features for subsequent analysis was set at the BTS and surrounding Voronoi cell level resulting in an average study area size of approximately $6.23km^2$. The analysis of mobility behaviour in Dar es Salaam is based on CDR data collected for approximately 2 million users in Tanzania during the year 2014 stored in a Postgres [9.13] database. In the present case, only CDR for SMS and calls recorded between August to December for users that connected to at least one BTS located within the Tanzanian municipalities of Kinondoni, Ilala and Temeka, which are classed as the Dar es Salaam metropolitan area and shown in Figure 1.1, were included within the study.

The final CDR dataset used as part of this study covers a total of 433.601.508 call and SMS events for 415.341 mobile phone subscribers resident within the Dar es Salaam region of Tanzania over a period of 122 days from August 1st to December 1st, 2014 [1]. Similar to prior empirical analysis in Chapter 3 and Chapter 4 mobile data usage was excluded from the analysis as there was no access to mobile data usage and SMS for the same period of the year. A period with access to SMS transaction records was subsequently chosen over mobile data, as only a small subset of users has access to mobile data capabilities.

MODLE

In addition to the CDR data provided by a Tanzanian MNO used for the empirical analysis of mobility behaviour in Dar es Salaam, CDR derived OD matrices are used for a comparative assessment. Those OD matrices have been generated by a UK based MNO

[1] Due to both individual and commercial privacy, the anonymized data used as part of this study is not publicly available, and was provided through a partnership with a major Tanzanian MNO

using both active and passive events generated on weekdays between February and March 2016 using a multi-step process for the MODLE project. The Transport Systems Catapult worked on an agent-based model to visualise demand in the peri-urban area to one side of a major UK-based city. The MODLE project aims to demonstrate how activity-based models can inform the operation of a mobility service provider and in which way new data sets can be exploited and used in modelling tool to have a deeper understanding of demand patterns and trip chains in urban areas.

The MODLE OD matrices were aggregated to UK census geography levels for use with activity-based transport forecasting models. In addition to origin and destination, they contained information on trip purpose, periods defined to be consistent with traditionally transport modelling periods [2] and inferred travel mode.

5.3.2 Data Cleansing

Only CDR records for SMS and call events with timestamps between August 1st and December 1st were selected. Although MFS logs were available for the entire study period, they were excluded from the analysis as the cell identifiers corresponding with BTS at the centre of the cell were truncated, and BTS could therefore not be accurately identified in most cases. Furthermore, mobile data usage was excluded from the analysis as there was no access to data for mobile data usage and SMS for an overlapping period of the year. A period with access to SMS data was subsequently chosen over mobile data, as only a small subset of users has access to phones with mobile data capabilities. This lack of access would have resulted in the exclusion of a large proportion of users from the analysis. Furthermore, users that did not connect to at least one BTS located within the Tanzanian municipalities of Kinondoni, Ilala and Temeka, which are classed as the Dar es Salaam metropolitan area, were excluded from the dataset.

Records were further checked for integrity and missing attributes. This included the iden-

[2] AM peak (07:00 10:00), Interpeak (10:00 16:00), PM peak (16:00 19:00), Off-peak (19:00 07:00)

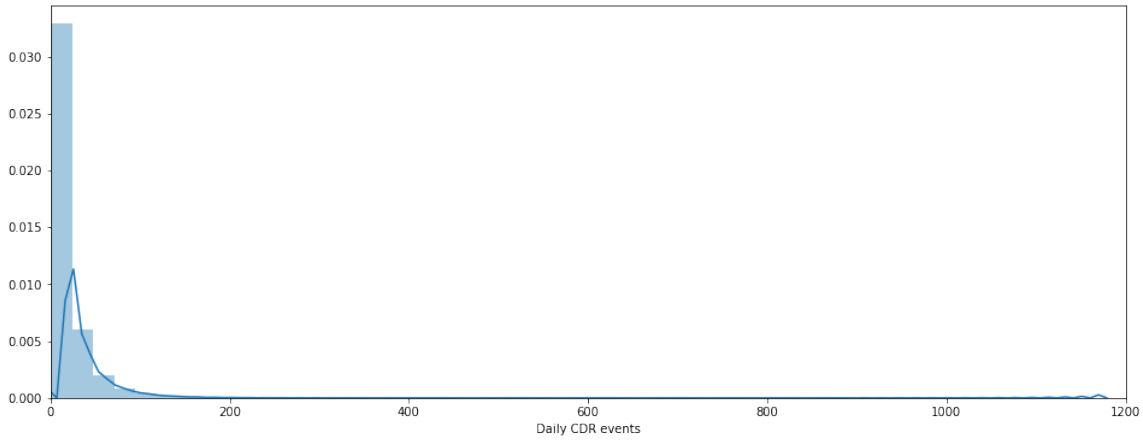


Figure 5.1: Average number of CDR events incorporating calls and SMS considering days when a mobile phone subscriber generated events.

tification of CDR logs with missing BTS information, which in the present case affected approximately 0.56% of records. Those were removed from the analysis, as they could not be reliably linked to a BTS as with MFS data discussed earlier.

As part of the integrity analysis, the spatial proximity of BTS was investigated through the generation of the Cartesian product of all BTS in the Dar es Salaam metropolitan area and the spheroid distance between the geographical locations. 14 BTS were found to be within 50 meters of each other with 8 out of the 14 being within a proximity of fewer than 10 metres to each other. Those BTS were subsequently merged to prevent the location of a user becoming indistinguishable and to reduce noise within the data set.

As a final step, individual calling activity was examined in order to identify low-frequency and high-frequency users, which could introduce potential biases into the identified mobility behaviour. As showing in Figure 5.1, the majority of mobile phone subscribers only produce a small number of CDR events, while the opposite extreme includes businesses and agents that were most likely classed as subscribers erroneously, or constitute informal businesses. Due to the sparsity in the event and by extension location recordings of low-frequency users, consecutive events may be separated by significant time intervals, with high amounts of hidden movement [63]. In order to identify and subsequently remove these users from the analysis, a three-step process was used:

1. all network events per user were summarised per weekday.

2. the average weekday activity per user was calculated over the course of a month.

3. five n-tiles were calculated to classify the population into different activity groups. Earlier work by Wang *et al.* (2012) [347] also categorised users as belonging to one of five categories, but considered fixed numbers of network events per month [3] rather than a dynamic assignment. A similar approach also exists in the blanket exclusion of all users with less than six network events per day [306, 359]. While this approach accounts for a potential low-frequency bias, it neither addresses a potential high-frequency bias nor does it account for differences in usage on weekdays/weekends (§7.4.3).

5.3.3 Stop Extraction and Trip Generation

Two categories of parameters are used to extract stops from the cleansed CDR time series generated through the previous step: temporal and frequency-based parameters. As part of the frequency-based clustering approach and filtering, BTS associated with multiple network events are extracted.

Temporal filter Stops for the transient-based approach were extracted using a temporal low-pass and upper-bound filter similar to related research [60, 176, 228]. The filters were applied to the inter-event time between two subsequent records. Records which fell outside either of the two filters were discarded from subsequent trip generation. The low-pass filter was set to an inter-event time of 10 minutes in order to alleviate some of the false displacement effect, where users appear to be moving as they are connecting to different BTS while remaining stationary (§7.3.2). The upper bound filter was set to 60 minutes similar to previous approaches [176, 228] to mitigate the effect of hidden motion caused by the sparsity of records detected in the statistical analysis of average user activity levels. Sequential network events pairs with inter-event times that fit with the

[3] below 10, 10-500, 500-1000, 1000-2000, over 2000

temporal filters were selected as stops with the BTS servicing the initial network event selected as a stop location.

Usage frequency Stops for the frequency-based clustering, were identified through filtering on regular BTS usage rather than the filtering of inter-event times. All network events for weekdays over the entire study period were aggregated into 30-minute bins with a count assigned to each BTS, resulting in a potential 48 observations (24 hours / 30 minutes = 48 slots) for each BTS. Three parameter scenarios were formed for weekday activity, one considering no filtering at all, one considering a minimum of 4 events within the 30-minute bin, and one considering a minimum of 4 events that account for more than 60% of observations within the 30-minute bin. Using the parameters the most commonly used BTS in each 30-minute bin was selected as a stopping point.

Trip generation Trips are generated through the chronological order of stops identified using parameters discussed in the previous section. Each stop can be a destination for one trip and an origin for another at the same time. A trip consists of a pair of non-identical consecutive pairs of stops (an origin and a destination). In the case of usage frequency-based clustering, the identified daily trajectory was regarded as a users synthetic daily activity plan for common weekday activity.

5.3.4 *A Priori* OD Matrix Generation

In order to generate the *a priori* OD matrices, the trips generated as part of the previous step were then aggregated for different periods with the matrix count as the flow, the sum of all trips in a specified time frame, between matching origin and destination pairs. Choosing appropriate periods requires an understanding of intrinsic rhythms that govern urban spaces. Periods for flow analysis are defined according to different goals, including peak hours, slack periods or work-day flow analysis. Using general perceptions of mobility in Paris, Bahoken and Raimond (2013) [26] chose time windows of 6 am-10 pm and 4 pm-8 pm as well as a full 24 hour period for their OD matrices. Building on this research, Larijani *et al.* (2015) [209] use 6am-10am and 2pm-9pm to detect morning commutes and

afternoon flow. Sinha *et al.* (2014) [312] chose to only focus on the morning AM peak period between 9.30 am to 11 am in their analysis of mobility demand in Mumbai, India. Diao *et al.* (2015) [95] chose 8 time segments to capture intra-day mobility variations in Boston, USA: early morning (3-6am), morning peak-hour (6am-9am), morning-work (9am-12pm), noon (12pm-2pm), afternoon work (2pm-5pm), afternoon peak hour (5pm-8pm), night (8pm-12am), and midnight (12am-3am).

Four different periods were chosen in the generation of transient-based *a priori* OD matrices similar to those in related research by Bahoken and Raimond (2013) [26]. Those included 6 am to 10 am to capture morning commutes from home to work, 10 am to 4 pm to capture general movement while the majority of the population is at work, 4 pm to 8 pm to capture the commute home from work, 8pm to 6am to capture general movement while the majority of the population is expected to rest at home. While other approaches as discussed above have been tested, one only captured morning AM periods without providing windows for the remainder of the day [312], while another broke the day into too many time segments [95] and would have exacerbated issues of data sparsity in some areas. Weekends were excluded from the analysis, as the activity patterns differ from working activities taking place during the week. This heterogeneity phenomenon (§7.4.3) has been both discovered in existing literature [63, 113, 114] and in the analysis of activity signatures to identify activity-based land use in Chapter 3 and specifically Figure 3.3b.

5.3.5 Scaling and Verification

One of the issues with undertaking research in developing nations is the fallacy of the 'Statistical Tragedy', a common sparseness of accurate and timely base data. In the present case, neither traffic counts, nor GPS traces or information from travel surveys were available to confirm the findings of the current study. Instead, the sparseness of OD matrices was used as a verification criterion. Analysing the sparseness of BTS in an OD matrix was first undertaken by Ton and Hensher (2002) [334] and later De Dios and Willumsen (2011) [269]. They searched for BTS with 0 values between regions as

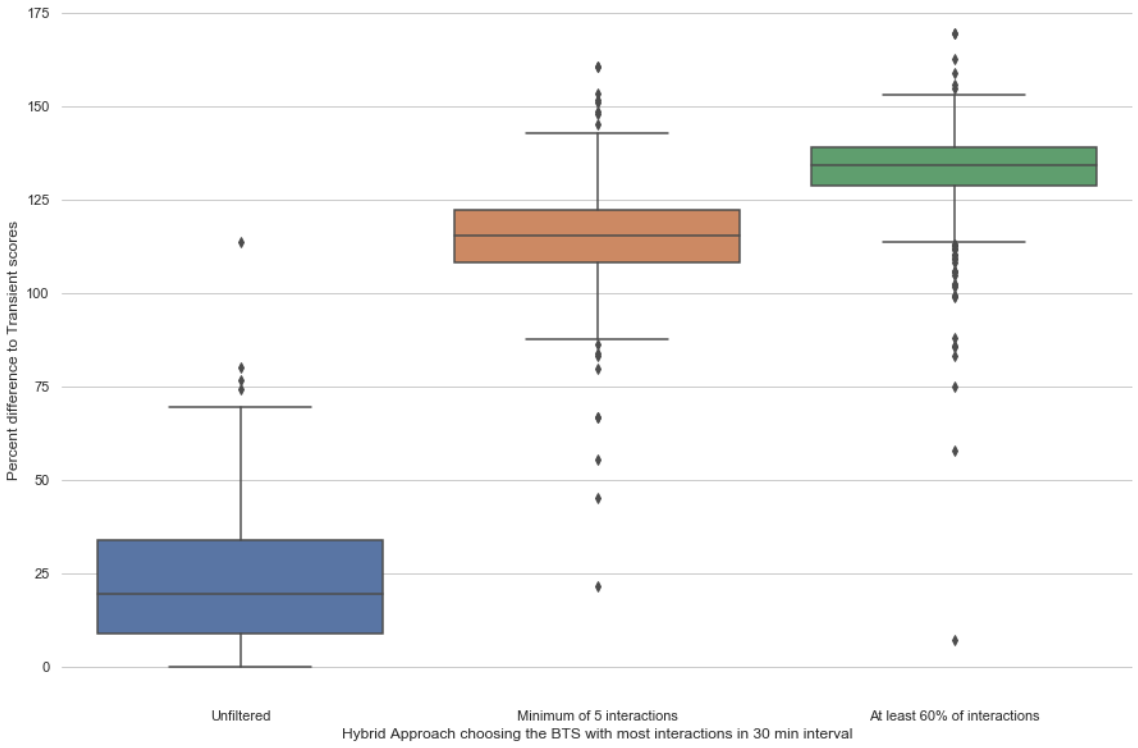
indicators that those regions are not generating or attracting trips. 52 BTS were found to not attract any incoming or outgoing trips during the study period following the application of temporal filters during the stop extraction step. This cannot always be directly attributed to the use of temporal filters but rather a result of changes in BTS operations and the proximity of BTS.

5.4 Results and Discussion

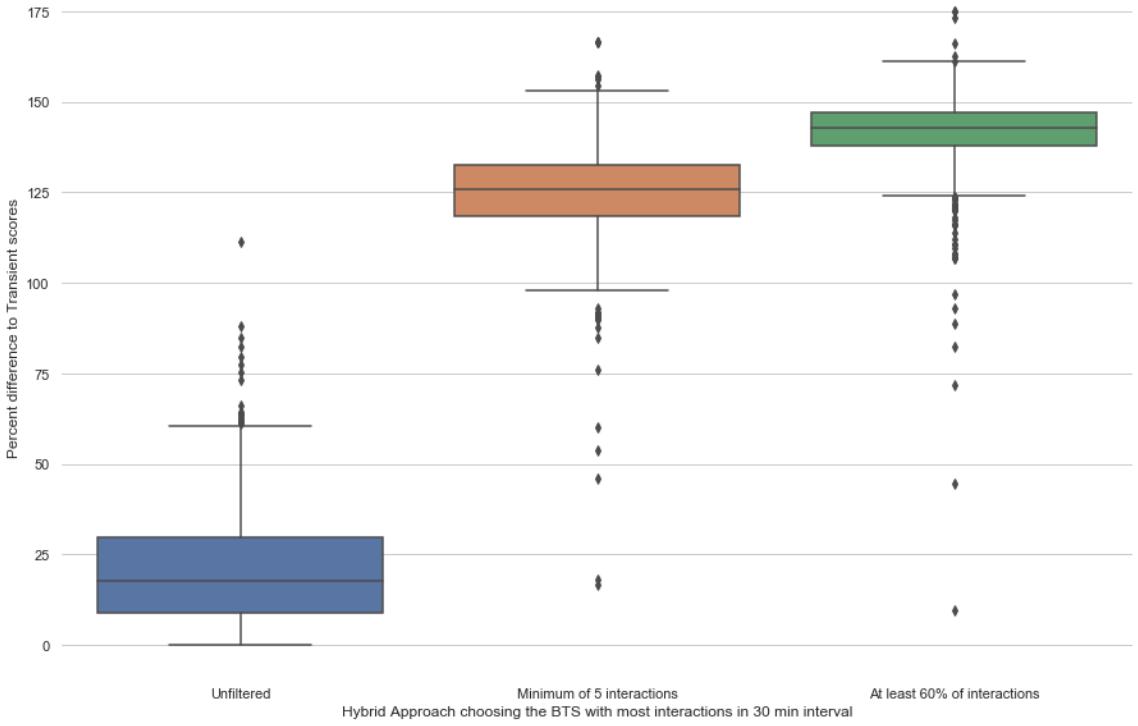
5.4.1 Origin-Destination Matrices

Descriptive statistics for the transient-based, and frequency-based clustering approach to trip extraction for different parameter scenarios is reported in Table 5.1. Figures 5.2a and 5.2b illustrate the percent difference of detected trips between the transient-based approach and three different filter scenarios for the frequency-based clustering OD matrix generation approach.

The transient-based approach was run using a temporal low-pass filter of 10 minutes and an upper-bound filter of 60 minutes between sequential CDR pairs with those within the parameters and showing BTS displacement selected as trip pairs. A total of 2,546,677 inbound and 3,002,348 outbound trips were detected with a mean number of 4564 and 5471 trips respectively. The result is a percent difference of 16.246 % in the mean number of detected inbound and outbound trips across the whole day. While the percentage difference across most of the day is negligible with 0.420 % in the morning, 0.229 % around noon and 0.005% in the evening, the difference in detected inbound and outbound trips spikes to 38.107% for the night-time period. This may be in part due to a known negative morning bias as mobile phone subscribers tend to use their phone more toward later parts of the day leading to fewer network events and by extension trips being detected in the earlier portions of the day [11]. Additionally, this approach relies on the identification of pairs from individual network events, which are more susceptible to differences in individual usage patterns with previous research finding strong effects of heterogeneous usage



(a) Inbound



(b) Outbound

Figure 5.2: Subfigures (a) and (b) illustrate the percent difference of detected trips between the transient-based approach and three different testing scenarios for the frequency-based clustering OD matrix generation approach.

patterns and individual-level sampling on network event generation and by extension trip rate identification [11, 29, 63, 113, 114, 220]. As discussed in §5.3.4 a set of four time periods, 6 am to 10 am, 10 am to 4 pm, 4 pm to 8 pm and 8pm to 6am was chosen to lend insight into morning commutes, workdays, evening commutes and night time patterns respectively. Those time periods are useful for practitioners looking to design effective interventions such as road pricing to address challenges brought about by increasing mobility demand [269]. A shortening of any of the periods to below 4 hours would have exacerbated issues of sparsity (see §5.4.3 in the data used as part of this research. Additionally, it may have introduced additional biases related to heterogeneity in call rates and particularly negative morning bias.

The usage of a frequency-based clustering approach helps address some of these issues. The approaches pattern-based nature helps uncover 'hidden movement' caused by sparse temporal frequency and heterogeneity in usage. Here, the frequency-based clustering approach was run using three different parameter scenarios: unfiltered; a minimum of 5 network events at a given BTS within a 30-minute interval; and a minimum of 5 network events accounting for more than 60 % of observations at a given BTS within a 30-minute interval. The percent difference between the mean number of inbound and outbound trips was significantly lower than for the whole day of the transient-based approach at 0.045% (unfiltered), 0.061% (minimum of 5), and 0.003% (minimum of 5 and 60%) respectively. The percent difference between the total number of detected trips for inbound and outbound journeys between both the inbound and unfiltered frequency-based clustering approach is 10.584% for inbound and 5.820% for outbound trips, respectively. The discrepancy is most likely caused by the significant difference in detected inbound and outbound trips for the night-time period for the transient-based approach. Overall, the percentage difference for the different parameter settings indicates the frequency-based clustering approach to be less susceptible to individual-level sampling and heterogeneity biases. Figures 5.2a and 5.2b highlight the percentage difference in detected trip numbers between the three parameter settings for the frequency-based clustering approach,

compared to the transient-based approach.

Table 5.1: Descriptive statistics for transient-based and frequency-based clustering trip extraction scenarios

Approach	Time of day	Parameter	Direction	Total	Mean	S.D.	Min	Max	% [4]
Transient	Morning	Lower &	In	403188	721.267	421.875	1	2537	0.420
		Upper bound	Out	403019	720.964	427.766	3	2587	
	Noon	Lower &	In	729950	1308.154	769.851	13	4561	0.229
		Upper bound	Out	729587	1305.165	768.202	1	4586	
	Evening	Lower &	In	444568	795.292	465.831	2	2872	0.005
		Upper bound	Out	444590	795.331	466.277	1	2885	
	Night	Lower &	In	968981	1733.419	1078.709	1	6481	38.107
		Upper bound	Out	1425152	2549.467	1606.950	3	10224	
	All day	Lower &	In	2546677	4563.937	2599.912	25	15814	16.246
		Upper bound	Out	3002348	5370.927	3120.560	24	19703	
Frequency-based	All day	None	In	2831268	5064.880	3046.040	27	17831	0.045
			Out	2832536	5067.148	3050.035	28	17883	
		5 minimum	In	686138	1227.438	709.007	9	4153	0.061
			Out	686553	1228.181	717.918	11	4209	
		5 minimum & 60%	In	509530	911.503	530.957	4	3176	0.003
			Out	509515	911.476	544.194	5	3271	

[4]Percent difference between mean inbound and outbound trips

	Mean	S.D.	Min	Max
Trajectory distance	23599.521	8368.864	3346.101	79297.873
Trip Distance	4249.839	1492.331	411.593	11913.608

Table 5.2: Descriptive statistics for trip and trajectory travel distances inferred through frequency-based clustering

5.4.2 Travel Distance

The frequency-based clustering approach with filtering of a minimum of 5 network events was used to calculate both the trip and the total trajectory distance travelled by mobile phone subscribers resident within Dar es Salaam. Each distance was calculated as the euclidean distance between identified BTS pairs. Table 5.2 shows the descriptive statistics for both distance metrics in meters. The average number of each trip across the whole population sample is 4250m between BTS with the average trajectory, the sum of all trips for a user in a given day, showing an average of 23600m across the population sample. Distances travelled vary widely across the population as is evidenced by the high standard deviation of 8368.864 and 1492.331 for trajectories and trips respectively.

5.4.3 Study Limitations

Individual-level sampling Due to the sampling nature of MND data (§2.2.1) a users position is only recorded when they engage in a network event (§7.2.1). As a result, low-frequency users generating only a small number of network events exhibit a potentially very large amount of hidden movement. High-frequency users generating a disproportionate amount of network events, on the other hand, can lead to the introduction of non-representative mobility behaviour into the final results (§7.2.2). As can be seen in Figure 5.1 briefly discussed above, the CDR data used as part of this research has a very high number of low-frequency users, with many only generating one or two network events per day. The usage of a frequency-based clustering approach helps address the sampling issue, as it both fills gaps left by hidden movement through the identification of common patterns based on the assumption that the majority of human movement is predictable [6, 34, 142, 314] as well as with high-frequency users, as OD matrices can be based on

common routings rather than observation counts.

BTS-level accuracy In contrast to GPS, MND data does not record individual's locations, but rather a non-precise proxy via the physical location of the BTS servicing the mobile phone subscriber at a given time. Using signal strength data from a sector antenna, which was not available as part of the data used for this research, triangulation of a mobile phone subscribers location is theoretically possible (§7.3.1). The lack of this data affects both the identification of trip rates and distances, and daily trajectory distances, as BTS displacement and by extension, movement may not be detected in the larger peri-urban areas with a lower BTS density (§7.3.3). Additionally, false displacement caused by fluctuations in signal strength can lead to BTS displacement in consecutive network events while the handset associated with these events remains stationary in the physical realm (§7.3.2).

Population-level bias The aforementioned heterogeneity in network event generation rates can lead to the generation of vastly different network events registered across the time period (§7.4.3). Due to its pattern-based nature, the frequency-clustering based approach to drip detection can help overcome this effect. Nonetheless, issues of representativeness and sub-sample demographic bias inevitably remain as neither the MNO providing the data for this study nor mobile devices itself have a market penetration of 100% (§7.4.1).

5.5 Chapter Summary

The analysis leveraged CDR data for 433.601.508 call and SMS events for 415.341 mobile phone subscribers resident within the metropolitan area of Dar es Salaam over a period of 122 days from August 1st to December 1st, 2014. Both a transient-based and a frequency-based clustering approach with three different parameter sets were used to extract trips for the generation of OD matrices and distance metrics as mobility variables for further analysis in the next chapter.

This study demonstrated that the frequency-based clustering approach is less susceptible to differences in individual usage patterns including heterogeneity in call rates and individual low- and high-frequency biases, which affected trip rates identified through transient-based trip extraction. However, limitations remain, largely due to BTS-level accuracy and population-level biases that result in a potentially large number of unidentified trips for mobile phone subscribers within the data set, and a full exclusion of patterns of those not generating data with the MNO in the first place. At the same time, this analysis will provide useful information in fast-changing metropolitan areas such as Dar es Salaam, where the use of traditional methods of data collection, including manual sampling and sensor-based approaches are prohibitive.

The following chapter will use variables from the spatial dimension 3, socio-economic dimension, and mobility metrics generated within this chapter to analyse the land use and socio-economic – transport interaction for the metropolitan area of Dar es Salaam, Tanzania.

Chapter 6

Land Use and Socio-economic – Transport Interaction

6.1 Chapter Introduction

The previous chapters 3-5 used CDR, MFS and ground reference data to generate insight into land use characteristics, SEL and mobility trends within the metropolitan area of Dar es Salaam, Tanzania. This chapter is using factors initially generated at the BTS and Voronoi-polygon level for the ward-level analysis of the land use and socio-economic – Transport Interaction. The following research objective guides this chapter:

Research Objective 4: analysis of the alternative land use – transport interaction accounting for socio-economic characteristics for Dar es Salaam using variables identified from CDR and MFS data through Research Objective 1-3.

The importance of incorporating land use within the analysis of transport patterns have been recognised as early as the 1960s with initial work by Forrester (1969) [123] in his *Theory of Urban Interaction*. The initial spatial land use model was an aspatial model to study the interaction between population, employment and housing. Land use – transport interaction models have evolved since then, as research has shown that the built environ-

ment alone cannot account for individual differences in mobility behaviour [25, 197, 320]. Instead, transport interaction is now recognised as being influenced through a range of variables in addition to the built environment, commonly described through the spatial dimension expressed through the 3D's [201] – Density, Diversity, Design – and more recently 5D's [119] – expanding 3D's to include Distance to transit, Destination accessibility. This 'traditional' relationship has been further expanded to also consider the socio-economic dimension or personality dimension discussed in Chapter 2 to propose an 'alternative' relationship as shown in Figure 1.4. Owing to the absence of up to date, fine-grained and reliable data, often incompatible levels of aggregation among different datasets, and choice in analysis approach, however, the different dimension remain analysed in isolation through 'traditional' relationships in most cases. There has been minimal work using data generated through non-traditional data sources such as LSS data, aerial imagery or MND in the analysis of either LUTI relationships. As discussed in §2.3.1, there is also a distinct scarcity in empirical studies LUTI studies in emerging economies. Those few studies focused on India [8], Iran [117], China [218, 258], Thailand [343] with Ghana as the only African country on the list [281].

This chapter will be addressing these shortcomings by analysing the influence of independent variables from the spatial and socio-economic dimension on dependent mobility variables generated from non-traditional MND in Dar es Salaam in an 'alternative' relationship using Structural Equation Modelling (SEM) as it allows for the analysis of factors in combination to account for indirect interaction. It will first provide a review of literature using SEM for the analysis of land use – Transport Interaction before detailing the research approach building on factors generated from CDR and MFS data.

6.2 Literature Review

Traditionally, features used in the analysis of land use – transport interaction have been generated from official statistical sources and analysed in de-facto isolation using more traditional multiple linear regression approaches. While the literature review in §2.3.1 has

focused on research on the interaction in emerging economies, the section below highlights LUTI research carried out using SEM, which allow for the parameterisation of independent relationships to account for indirect interaction effects in the interconnected relationships explaining mobility behaviour through multiple dimensions.

6.2.1 SEM Using the Spatial and Socio-economic Dimension

The different dimensions discussed in Chapter 2 have traditionally been analysed using multiple linear regression. Structural Equation Models, on the other hand, allow for the parameterization of independent relationships to explain mobility trends and RSS [25, 85]. Among SEM are three different types: SEM with a measurement model and a structural model known as SEM with latent variables, a structural model without any measurement models (SEM with observed variables), or a measurement model alone (confirmatory factor analysis). Latent variables are created to measure the joint effect of all characteristics of a dimension. The mechanics behind those models are introduced more in-depth in §6.3.3. An overview of SEM papers that investigated the relationship between factors of the spatial and socio-economic dimension and its influence on mobility trends can also be found in Table 6.2.1.

Bagley and Mokhatarian (2002) [25] constructed a SEM with observed variables to analyse determinants for RSS and travel demand in an attempt to overcome the shortcomings of traditional LUTI analysis using regression analysis. They used data from site surveys, mail-out surveys and travel diaries conducted in 5 neighbourhoods in the San Francisco area representing a range of design, diversity and density metrics in 1993 akin to previous work by Kitamura *et al.* (1997) [198]. The observed variables were mostly from the socio-economic dimensions, while also accounting for attitudes and lifestyles from the personality dimension. Dependent variables included residential location, attitudes to explain RSS and mode choice, travel demand across modes and commuting distance for work trips. Variables from the personality dimension were found to have the most significant impact on RSS. Once the personality and socio-economic dimension has been

accounted for, land use has negligible explanatory value for mobility behaviour. Their work supports the argument that the interaction between land use and transport is driven by correlation of each of the dimensional variables rather than one direct causality as traditionally assumed.

Simma and Axhausen (2003) [311] used travel survey and transport model-derived data for 1992 for regions in Upper Austria to construct a SEM with latent variables of spatial structure and personal and household characteristics. They investigated car-ownership (as an endogenous, rather than is commonly used as an exogenous feature of the socio-economic dimension), and mode split into walking, public transit and car trips. Similar to other non-SEM based work [320, 205] they found gender and work status to be the most important socio-economic characteristics. While the number of facilities was the most important variable within the spatial structure construct, the spatial dimension itself contributed to little explanatory value compared to socio-economic variables.

Using municipal level Flemish Regional Travel Survey data for 2000-2001 for 5696 respondents, Van Acker *et al.* (2007) [340] used Confirmatory Factor Analysis (CFA) to construct latent variables for use in an SEM with latent variables to determine the causal influence on travel distance (in km), travel time (in minutes) and the number of trips. As the travel survey included only a limited amount of land use variables such as distance to public transit facilities and residential environment, a synthetic feature based on municipal categories was created as a proxy for differences in the spatial dimension concepts of density, diversity and design. Factors from the socio-economic dimension, on the other hand, were split across two latent constructs of social status (of the individual) and household responsibility. Travel behaviour was found to be predominantly influenced by the respondents social status: a high social status was associated with more complex travel behaviour. Travel behaviour was affected, especially indirectly, by the individuals role within the household. The effect of land use was limited. Furthermore, indirect effects remained important to understand the complexity of travel behaviour.

Focusing on the Pudget Sound region, USA, Silva and Goulias (2009) [85] used transportation panel survey data from 2000 for 1025 respondents and land use data for TMZ and $750m \times 750m$ grids to construct and SEM with observed variables and compare their results with findings from Lisbon, Portugal. Observed variables included socio-economic features, residence and workplace land use patterns, and commuting distance. The latter in log form also being part of the endogenous travel behaviours alongside mobility decisions classed as either short-term or long-term by the authors. They found that both diversity and density have a significant impact on the attractiveness of activity spaces and peoples willingness to travel. Activity space is a concept originating in the research field of time geography and reflects attempts to understand the actual and possible undertaking of an activity provided through given land use. People with different socio-economic characteristics and income levels tend to work and live in areas of different density and diversity. Those living or living and working in the city centre are more likely to travel by public transit and less likely to own a car, while those living in less dense peri-urban areas are willing to travel longer distances. While the models account for RSS based on socio-economic characteristics, they also show, that the spatial dimension has a significant impact on long-term decision making of individuals even when accounting for RSS. Minor differences between the two cities can be attributed to the level of implementation of a public transit system in the two cities.

Focusing on TMZ in Taipei, Taiwan, using trip distribution model and City Bureau of Transport for 2000, Lin and Yang (2009) [219] use the maximum likelihood function within *LISREL* to construct a SEM with latent variables to analyse trip generation and private mode split. While data on vehicle-miles travelled was available, it was disregarded as an exogenous variable due to the level of aggregation used. The authors created latent variables on diversity; design with associated constructs of transit service and private-mode facility; and density based on the concept of the 3Ds by Cervero and Kockelman (1997) [69]. The socio-economic condition was captured through the social-condition and

economic condition. Density was found to increase trip generation and reduce private mode split, whereas diversity showed an opposite effect. Pedestrian-friendly design led to a reduction in the private-mode facility. In contrast to US-based studies but consistent with studies on Asian countries, they found that diversity increased private-mode split.

Eboli *et al.* (2012) [106] used census data from 2001 and GIS-based measures of land use and accessibility to analyse census parcels for Cosenza, Italy, and surrounding areas to construct a SEM with latent variables to analyse the number of internal and external trips. They constructed two latent variables to measure the joint effect of both the socio-economic dimension through socio-demographic and economic characteristics constructs and spatial dimension through constructs of density-based land use and design based accessibility. The construct 'economic characteristics' was found to have the most significant influence on both internal and external trips.

Analysing home-work commuting in Shirza, Iran, Etminani-Ghasrodashti and Ardeshiri (2016) [117] used survey data for 22 residential areas from 2014 to construct their SEM. They use a range of land use attributes, individual characteristics, job characteristics and household characteristics to investigate their influence on the modal split for work-based and non-work trips. They simplified traditional RSS assumptions by ignoring the personality dimension and assuming that RSS occurs per socio-economic characteristics. While diversity and density findings were in line with previous studies, design features were found to be ambiguous in explaining the modal choice.

Focusing on the Netherlands, Puello *et al.* (2017) [285] used panel survey data for 2013-2015 for 5042 households and 11,322 individual respondents to build a hybrid choice model. They used individual level and household level features from the socio-economic dimension, accessibility and urbanity from the spatial dimension and survey characteristics to analyse both trip rates per group of respondents and attrition and completeness of panel surveys. All of the variables were found to be significant in estimating mobil-

ity while the explanatory power of attrition and completeness varies significantly across waves, due to issues of adherence.

Table 6.1: Overview of research on the relationship between the spatial and socio-economic dimension using SEM

	Data source, sample size, study area	Analysis approach	Spatial dimension	Socio-economic dimension	Mobility variables
Bagley and Mokhtarian (2002)[25]	1993 site surveys, mail-out surveys and travel diaries for 5 neighbourhoods in the San Francisco area	SEM with observed variables		age, female, household size, number of under 16yo, square root number of vehicles, years in the bay area	residential location (traditional, suburban), attitudes (pro high-density, pro-driving, pro-transit), travel demand (log vehicle miles, log transit miles, log walk/bike miles), commute distance (miles) to job
Simma and Axhausen (2003) [311]	1992 Upper Austrian travel survey and transport model	SEM with latent variables	**Spatial structure** (distance to district capital, share of farms, share of working women, share of commuters, size of shop base/ number of work places, accessibility measure, pt-supply/car-supply)	**Personal and household** (employed, male, number of infants, number of pupils, number of reachable facilities)	car-ownership, walking trips, public transit trips, car trips

Van Acker *et al.* (2007) [340]	2000-2001 Flemish Regional Travel Survey on municipal level for 5696 respondents	CFA and SEM with latent variables	**Land use** (Categorization by size [e.g. suburban, large city], residential percentage, distance to public transit)	**Social status** (number of cars, education, household income, job status, full/part time employment), **Household responsibility** (Age, number of hh members, gender frequency, marital status)	Travel distance (km), travel time (minutes), number of trips)
Silva and Goulias (2009) [85]	2000 Pudget Sound panel survey with 1025 respondents, land use and TMZ data for Seattle, USA	SEM with observed variables	**Residence and workplace land use patterns** (global population density, built floor space density, density of arterial intersections, distance to regional center, land use entropy to measure diversity balance, bus supply)	**Socio-economic** (age, gender, income (low, medium, high), household size, average age, household with two members, household with teenagers), **commuting distance**	Log commute distance, **short-term decisions** (time spent between first and last trips, number of trips (non-motorized, transit, car)) **long-term decisions** (Transit pass, No. of cars)

Lin and Yang (2009) [219]	2000 City bureau of Transport data, Trip distribution model for 173TMZ in Taipei	Maximum Likelihood and SEM with latent variables	**diversity** (type-mix, housing-job, housing-retail, retail-job, land use entropy), **design** (road density, grid network, sidewalk density), **density** (residential, building, employment) **transit service** (bus stop density, distance to metro station, bus route density, transit accessibility) **private-mode facility** (accessibility (car, motorcycle), parking space density)	**social-condition** (alimentation ratio, student ratio, female ratio) **economic condition** (household income, car ownership, motorcycle ownership)	
Eboli *et al.* (2012) [106]	2001 census data and GIS-based measures of land use and accessibility for census parcels in Cosenza, Italy	SEM with latent variables	**Land use** (Houses surface, Residential environment), **Accessibility** (Distance to bus stop, distance to road junction, Attractiveness)	**Socio-demographic characteristics** (density, household members, workforce population, gender, marital status, education), **Economic characteristics** (Employment rate, Resident unemployed, resident non-worker, sector employees (primary, secondary, tertiary)	Number of internal and external trips

Etminani-Ghasrodashti and Ardeshiri (2016) [117]	2014 survey data for 22 residential areas in Shirza, Iran	SEM	**Built environment attributes** (Residential density, Job density, Entropy index, Street density, Internal Connectivity, Distance (bus stop, intersection, sub-center, central business district))	**Individual Characteristics** (Sex, Age, Income, Max desired walking distance), **Job characteristics** (No. of jobs (part, full time), work commute time), **Household characteristics** (18yos (under, above), workers, car ownership)	Number of car, transit, non-motorized trips
Puello *et al.* (2017) [285]	2013-2015 panel survey with 5042 households and 11322 individual respondents	Hybrid Choice Model, discrete choice model	**Spatial level** (Accessibility, urbanity)	**Individual level** (age, gender, employment, head, education, license, life event, e-shopping, preferences), **household level** (household size, gross income, children, household type, number of cars), **survey characteristics** (travel day, diary day, stayer, year, month)	Trip rates per group of respondents, attrition and completeness of panel surveys

6.3 Research Approach

The empirical analysis of land use and socio-economic – transport interaction for the metropolitan area of Dar es Salaam is undertaken on the ward level using factors generated through the analysis of MND in Chapters 3-5 and additional data obtained from a ground reference survey (§4.3.1) carried out over a 2 month period between late 2015 and 2016, and building counts of Dar es Salaam obtained from the World Bank. The analysis followed a three-step process:

1. **Spatial interpolation**
2. **Ward-unique feature engineering**
3. **SEM model construction**

6.3.1 Spatial Interpolation

Factors from the spatial dimension, socio-economic dimension and mobility were generated at the BTS level in chapters 3-5. Voronoi tesselation was used to generate Voronoi polygons as coverage area proxies for each BTS. Those polygons, however, do not match with pre-existing administrative boundaries.

In the analysis of activity-based land use in Chapter 3, grid interpolation was used for visualisation of land use maps in an effort to protect individual privacy (§7.6.2) and commercial interests (§7.6.4) of the MNO that provided the data. Here, Voronoi polygons are apportioned with existing administrative ward boundaries to allow for the calculation of ward-level statistics on density and mixture from the spatial dimension (§2.3.2), and statistics on gender and income that have been recognised as influential in explaining the socio-economic dimension in §2.3.2.

Dar es Salaam consists of 90 wards with 26 in Ilala, 34 in Kinondoni and 30 in Temeke. Voronoi polygons generated around BTS classed as being located within the municipalities of Ilala, Kinondoni and Temeke, which make up the Dar Es Salaam region, intersect

with 89 of the 90 wards located within the region. No BTS within the ward of Pemba Mnzani in Temeke municipality were included in the initial MND analysis carried out in Chapters 3-5.

Similar to Mao *et al.* (2017) [233], simple area weighting was used for apportioning of BTS voronoi polygons with official ward boundaries for Dar es Salaam. In simple area weighting the value C of a ward polygon N_i is based on the proportion of the area of intersections with other BTS Voronoi polygons V_j. Using equation 6.1, the value of a ward C_{N_i} is calculated based on the proportion of the wards area A_{N_i} shared with that of voronoi polygons A_{v_j} and their CDR and MFS derived attributes C_{V_j}. The attribute value limitations of this approach are discussed in §7.5.2.

$$C_{N_i} = \sum_{V_j} C_{V_j} \frac{A_{N_i} \cap V_j}{A_{N_i}} \tag{6.1}$$

There is a maximum number of 1276 intersections between coverage area of BTS-level features expressed through the theoretical maximum of 600 Voronoi-polygons for BTS within the wider metropolitan area of Dar es Salaam, and the 90 official ward boundaries for each of the generated factors as shown in Table 6.2. There are 1084 intersections for 88 wards with data for the factors available on the Voronoi level listed above. The discrepancy between theoretical and observed intersections, wards and BTS stems from different approaches to data filtering employed in the respective analyses. During the interpolation process, the ward of Kimbiji in Temeke, directly adjacent to Pema Mnazi (see Figure 6.1), was identified as lacking data for several metrics. It is located on the coastal area outskirts to the south of the central part of Dar es Salaam. In addition, the peri-urban wards of Chanika ($58.77km^2$) and Msongola ($65.02km^2$) within the Ilala municipality were shown to have a low-level of Voronoi area intersection coverage of 29.98% and 54.07% respectively. Figure 6.1 shows the relative location of Chanika, Msongola, Kimbiji and Pema Mnazi within the Dar es Salaam metropolitan area.

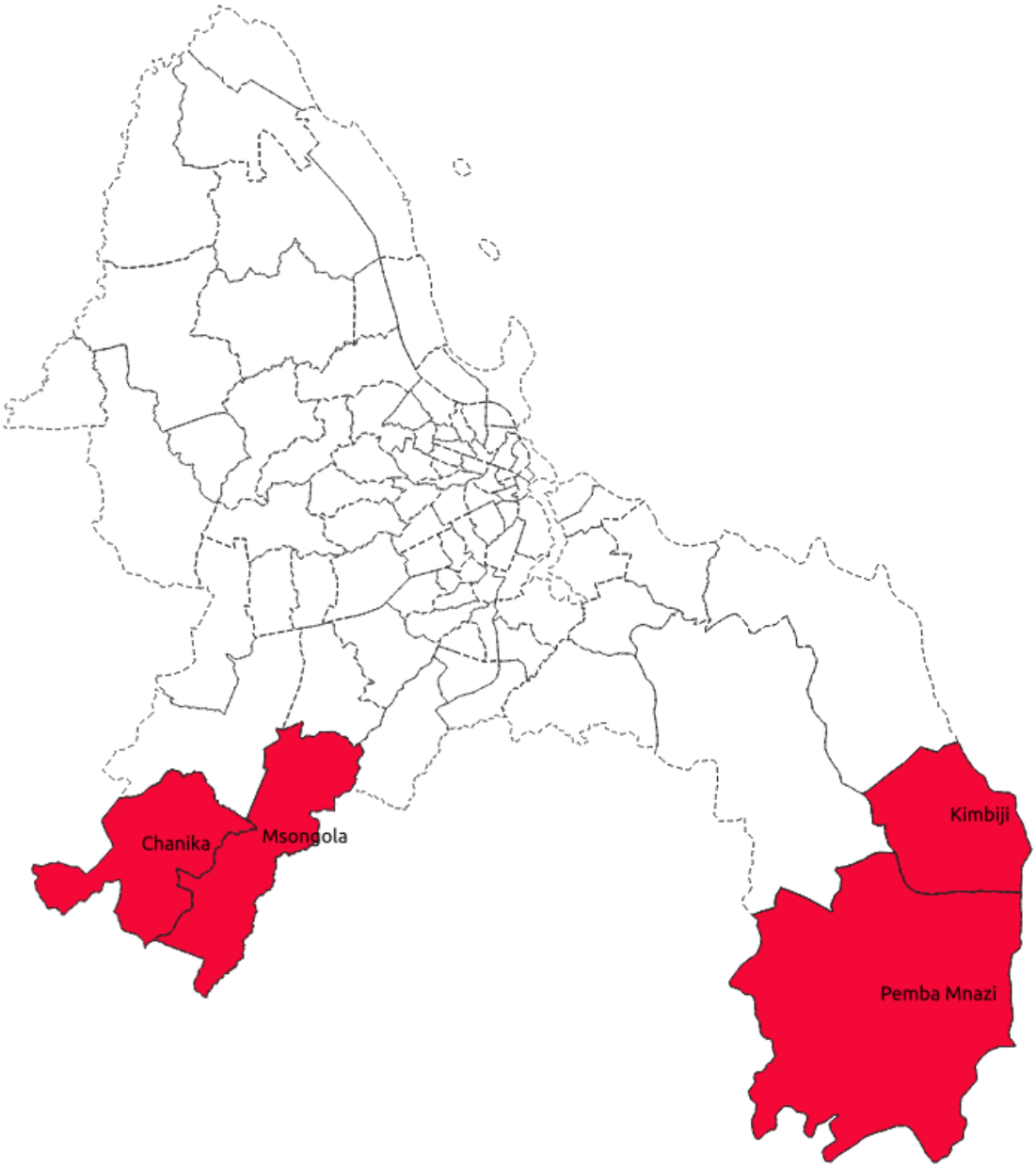


Figure 6.1: Wards of Chanika and Msongola in Ilala, and Kimbiji and Pema Mnazi in Temeke

Factor	Dimension	Source	Intersections	Unique Wards
Land use cluster	Spatial - Diversity	Chapter 3	1168	89
Percentage residential	Spatial - Diversity	§4.3.1	1271	81
CDR event count	Spatial - Density	Chapter 4	1196	89
Building Count	Spatial - Density	World Bank	1271	89
CDR users	Socio-Economic	Chapter 4	1197	89
MFS users	Socio-Economic	Chapter 4	1172	89
Income	Socio-Economic	Chapter 4	1166	89
SEL level	Socio-Economic	Chapter 4	1172	88
OD counts inbound	Mobility	Chapter 5	1165	89
OD counts outbound	Mobility	Chapter 5	1165	89
Avg trip distance (Weekday)	Mobility	Chapter 5	1170	89
Avg trajectory distance (Weekday)	Mobility	Chapter 5	1170	89
Frequent BTS	Mobility	Chapter 4	1169	89

Table 6.2: Factors generated on the Voronoi level

6.3.2 Ward-Unique Feature Engineering

In addition to the interpolated factors initially generated on the Voronoi-level shown in Table 6.2 as part of Chapters 3-5, additional ward-level statistics were included in the SEM model. The Voronoi-level factors were chosen as commonly used factors from the spatial, socio-economic and mobility dimension respectively as discussed earlier in Sections 2.3.2 and 2.3.3.

Network Event density of a BTS is used as a proxy for density, one of the three spatial dimensions to discuss the built environment discussed in §2.3.2. Similar to existing research by Mao *et al.* (2016) [234], the network event density of a BTS is calculated as

$$p_1 = \frac{v_i}{a_i}$$

in which v_i is the total number of network events made from or to a BTS i, and a_i is the area size of the Voronoi polygon of BTS i. BTS within the top quintile (20%) are considered high-density areas with the bottom quintile representing low-density areas and other quintiles representing medium-density areas.

Land use mixture is based on a degree of 'entropy', a measure originally defined for Thermodynamics based on the Greek word *Entropein* meaning transformation and change. In information theory it describes the average uncertainty within a variables possible outcomes. The use of entropy calculations in urban and regional models dates back to Wilson (1969) [361] for the estimation of traffic distribution. It was first applied to the calculation of balance between different land use classes by Cervero in 1989 [67] with other well-known calculations of land use mix as an entropy index conducted by Frank and Pivo (1994) [124] using geospatial analysis, and Kockelman (1997) and Cervero and Kockelman(1997) [69, 201], who introduced standardisation through the log calculation of the number of land use clusters $log(n)$. Based on the Kockelman, and Cervero and Kockelman approach, each ward is ranked from 0 (single-use) to 1 (high-mixture) with the land use mixture defined as:

$$Entropy = -\sum_{n}^{i=1} \frac{P_n \times \ln(P_n)}{\ln(n)}$$

in which n is the overall number of distinct land use clusters identified through the activity-based land use analysis in Chapter 3 (n=5), and P_n is the proportion of land use n's coverage of the total ward polygon area.

While other approaches such as the Herfindahl index exist, a meta-analysis of different LUTI papers by Ewing and Cervero (2010) [70] found the entropy-based land use mixture index to be the most appropriate metric for assessing the Diversity aspect of the spatial dimension (§2.3.2).

Table 6.3.2 provides the descriptive statistics of ward level variables from the spatial and socio-economic dimension used as independent, and mobility factors used as depended variables for the subsequent analysis through a SEM described in the next section. Figure 6.2 highlights the pearson correlation coefficients for the ward level variables. The trip frequency variables are perfectly correlated as was expected, since entering an area will ultimately require someone to leave the same area again. Trip distance and residential

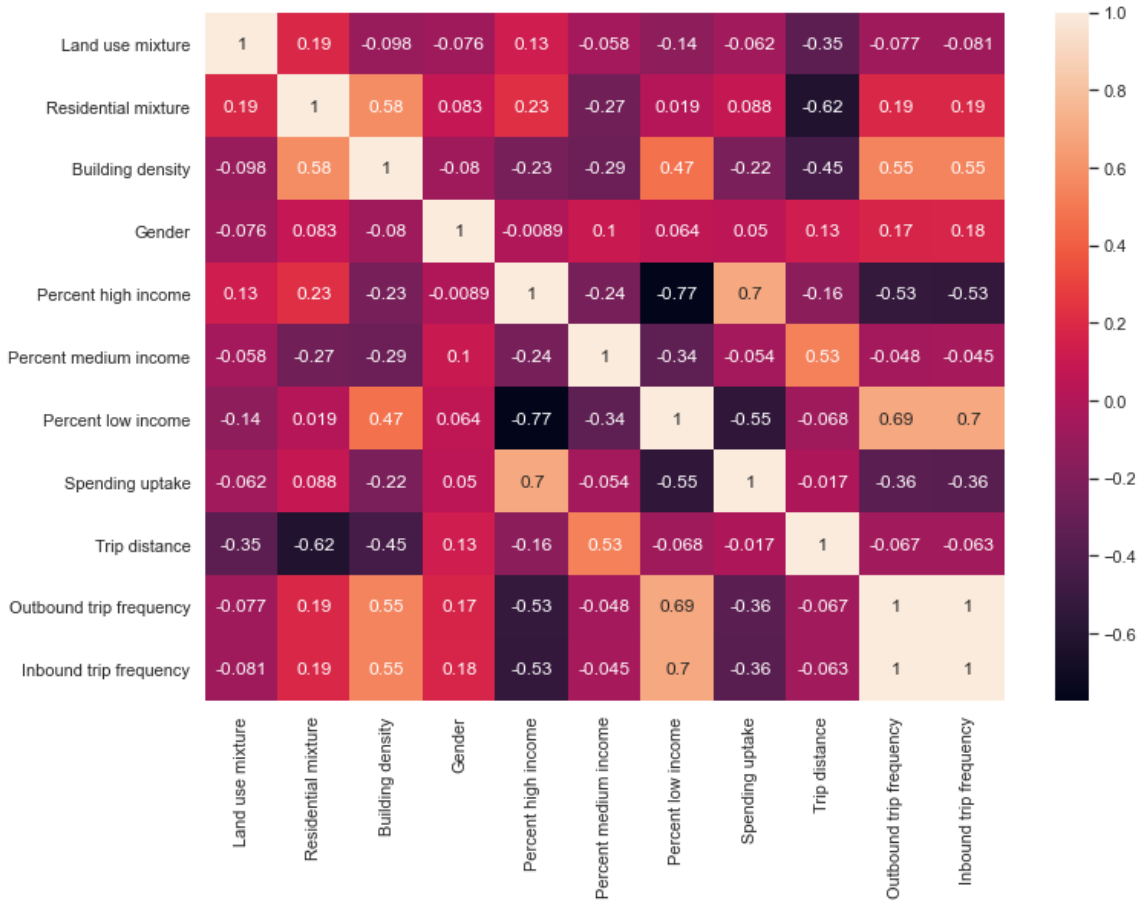


Figure 6.2: Pearson correlation coefficients for ward-level variables used for the analysis of and use and socio-economic – transport Interaction

mixture are strongly negatively correlated indicating that high levels of residential mixture is associated iwth shorter trips, while the correlation between overall land use mixture and trip distance is weaker. There is a strong positive correlation between spending uptake and high income as more disposable income is available in those areas. There is also a strong negative correlation between percent low income and percent high income, as poorer areas tend to have less wealthier residents and vice versa. There are also strong positive correlations between the trip frequency and percent low income in an area, indicating high numbers of trips in and out of the areas. Maps visualising the different features are furthermore included in Appendix C.

Table 6.3: Descriptive ward-level statistics of dependent and independent variables for the analysis of land use and socio-economic – transport Interaction

Factor	Description	Mean	S.D.	Min.	Max.
Spatial dimension	(independent) §2.3.2 and Chapter 3				
Network event density	Density of generated network events, continuous	274203.834	311048.0345	2739.0133	1343664.765
Land use mixture	Entropy Index, continuous	0.478	0.2	0	0.948
Residential mixture	Proportion of ward classed as residential, continuous	63.442	32.549	0	99.993
Building density	Building count relative to area size, continuous	339.4999	331.662	7.534	1466.452
Socio-Economic Dimension	(independent) §2.3.2 and Chapter 4				
Gender	1 = Male, 0 = Female, binary	0.501	1.63	0.431	0.546
High income	Proportion of area classed as Wealthy, continuous	36.203	26.679	0.966	95.577
Medium income	Proportion of area classed as Average, continuous	27.369	22.872	0.007	99.453
Low income	Proportion of area classed as Poor, continuous	49.153	31.801	0.013	99.993
Spending uptake	Income relative to the difference between MFS users and mobile phone subscribers, continuous	1623257.292	740673.185	658395.717	4451629.988
Mobility Factors	(dependent) §2.3.3 and Chapter 5				
Trip distance	Individual trajectory distance, continuous	21707.215	7585.104	10455.925	49947.1
Outbound trip frequency	Frequency of trips starting in the area, OD count	4774.802	1915.095	633	9455
Inbound trip frequency	Frequency of trips ending in the area, OD count	4778.221	1925.898	620	9456

6.3.3 Structural Equation Modelling

SEM is an umbrella term for statistical methods for the investigation of relationships between variables through the description of variance/co-variance structures of given data sets.

> "Structural Equation Modelling (SEM) can be defined as a diverse set of tools and approaches for describing and estimating causal relationships between variables, whether they be observable or latent" [135, p.1].

Simple SEM models were initially developed by Sewall Green Wright (1921) [367] in the form of Path analysis to analyse observed variables. Modern SEM date back to the 1960s and 70s and the development of factory analysis that allows for the analysis of latent variables, CFA [188] and multivariate regression models within a single analysis.

SEM has been used in a range of fields from psychology, educational research and political science to marketing research. In this Chapter, an SEM is constructed and tested alongside multiple linear regression to investigate the 'alternative' relationship between land use, socio-economic and mobility factors. SEM have advantages over more traditional multiple regression analysis [139, 251, 294] including:

- the ability to build a more complex model to better account for direct and indirect interaction effects such as the impact of socio-economic levels on land use density within an area and subsequently on mobility
- the ability to include measurement models to describe latent variables with multiple observed factors directly within the SEM to express constructs such as density as an aspect of land use
- the correction of residual measurement errors across observed variables

Variables

SEM are used to analyse structural relationship between independent (exogenous) and dependent (endogenous) variables. Variables can be either:

Observed/manifest variables (indicators) which can be directly measured or observed and can be both endogenous and exogenous.

Latent variables (constructs) which cannot be directly observed and are defined in terms of underlying observed variables, which are assumed to represent the latent variable.

An SEM can have any number of independent and dependent variables with each latent variable defined through its own measurement model. Each variable has either a measurement or a residual error assigned to it:

Measurement errors are associated with observed variables and reflect the adequacy in measuring the related underlying variables within the latent construct(s).

Residual errors are associated with the prediction of endogenous from exogenous variables.

Exogenous variables can have a mutual effect on the endogenous variable (Figure 6.3a) or one exogenous variable can influence the effect of another exogenous variable on the endogenous variable (Figure 6.3b,c) [139]. Each variable can have one of three effects on another:

1. **Direct effect** going directly from one variable to another (e.g. latent land use to latent mobility, latent socio-economics to latent mobility and latent socio-economics to latent land use)

2. **Indirect effect**, when two variables are influenced by one or more intervening variables, "such as the relationship between socio-economic status and [mobility] through land use".

3. **Total effect**, which is determined by the combined direct and indirect effects of an exogenous on an endogenous variable.

The arrows between latent and their manifest variables do not correspond to direct effects. Instead, they contribute to the understanding of the structure of the latent construct itself.

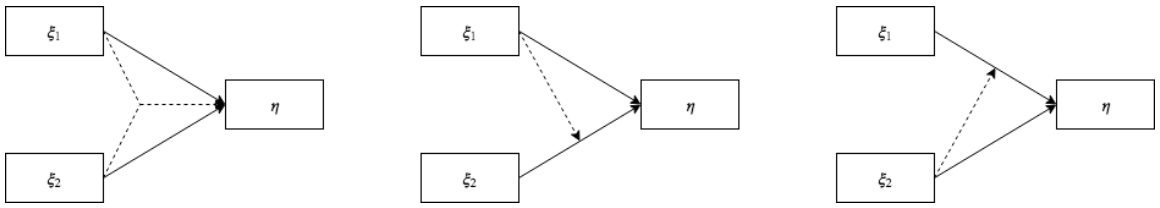


Figure 6.3: Interaction effects of endogenous and exogenous variables with full arrows as main, and dotted arrows as indirect interaction effects in which $\xi_1 \times \xi_2$

Structure

Unlike the single equation model of regression analysis, SEM can accommodate up to three different sets of equations simultaneously.

1. **Measurement models** define the composition of latent variables through their observed variables. It indicates how observed variables make up latent variables (i.e. $independent_1$ and $independent_2$ may be indicators of the latent *socio-economic status*)

   (a) A measurement (sub)model for the dependent variables

   $$y = \lambda_y \eta + \epsilon$$

   (b) A measurement (sub)model for the independent variables

   $$x = \lambda_x \xi + \delta$$

   in which x and δ are vectors of observed, manifest exogenous variables and errors, respectively; λ_x is a structural coefficient matrix capturing the effects of latent exogenous on observed variables; y and ϵ are vectors of observed, manifest endogenous variables and errors, respectively; λ_y is a structural coefficient matrix capturing the effects of latent endogenous on observed variables [49, 106].

   SEM's can be made up only of a measurement model (CFA). CFA is commonly used as either: **exploratory** to determine the number of and which latent vari-

ables/constructs are needed to explain the relationship between the observed variables, or **confirmatory** to model the known relationships among observed variables.

2. A **Structural model** defines how observed and latent variables are linked to each other for testing of the hypothesized causal dependencies. A basic structural model is expressed through the following equation [49]:

$$\eta = \beta\eta + \gamma\xi + \zeta$$

in which η is a vector of the latent endogenous variables, ξ is a vector of latent exogenous variables, and ζ is a vector of random variables [106]. Both β (for endogenous) and γ (for exogenous) represent structural coefficients or parameters of the model.

In a SEM all of the equations are estimated simultaneously and make up the 'SEM with latent variables'. A 'SEM with observed variables' consists of a structural model without any measurement model (The measurement model is not needed if all independent and dependent variables are observed variables). Ordinary regression is a special SEM consisting of one observed dependent variable and multiple observed independent variables.

Implementation

A two-step approach [19, 294] was used to generate latent variables (land use, socio-economics, mobility) from observed variables similar to [340] and subsequently analyse their relationship using the Python *Semopy* package. The package was chosen as it is well documented, freely available and developed for use with Python unlike many other popular SEM packages such as *lavaan* and *sem* that were developed for use within R [135].

First, measurement models are designed and set-up to corroborate what is known based on prior research described in Chapter 2. CFA is used to confirm whether the underlying observed variables appropriately measure the latent constructs or variables 'land use',

'socio-economics', and 'mobility'. Latent variables are considered appropriately designed [340] if they fulfil the following properties:

1. **Unidimensionality:** each observed variable is related to only a single latent variable;

2. **Convergence validity:** there is a degree of confirmation between two observed variables of the same latent variable;

3. **Reliability:** can be evaluated by the composite reliability and the variance extracted;

4. **Discriminant validity:** is ascertained when two constructs are not correlated.

Three absolute fit indices (X^2, GFI, RMSEA) and one incremental fit index (CFI) were chosen to determine model fit as four commonly used measures of fit [168]:

- Chi2 (X^2) as a traditional measure of absolute model fit and measurement of the discrepancy between the input sample and fitted co-variances matrices

- Goodness of fit index (GFI) as an alternative to X^2 for calculation of the difference in variance accounted for through estimated population variance

- Root mean square error of approximation (RMSEA) to identify model fit to the population covariance matrix with unknown yet optimized parameter estimates

- Confirmatory Fit Index (CFI) which compares the X^2 of the model to that of the null model, while accounting for small sample sizes making it a particularly suitable measure of model fit in the present case due to the comparatively low sample size of 86 wards compared to the more common sample sizes of 200+ for SEM with latent variables.

Table 6.4 shows the improvements in the goodness of fit for three different sets of measurement models. The 'Original model' considered three measurement models, 'Socio-economics', 'Land use' and 'Travel patterns' (see Figure 6.4a) based on the alternative

Indices	X^2 test		GFI		RMSEA		CFI	
	MLW	ULS	MLW	ULS	MLW	ULS	MLW	ULS
Benchmark	p>0.05		>0.9		<0.03		p>0.95	
Original model	521.860	510.181	0.663	0.656	0.326	0.322	0.683	0.677
Interim model	534.58	400.560	0.655	0.730	0.341	0.291	0.673	0.752
Adjusted model	562.607	277.774	0.669	0.822	0.368	0.247	0.684	0.845

Table 6.4: Goodness of fit in model development

relationship proposed by Stead [320] (see Figure 1.4). The 'Interim' and 'Adjusted' model instead consider land use as two separate latent variables density and diversity (see Figure 6.4b). While the interim model still considered 'trajectory distance' to be a part of the endogenous latent variable 'Travel patterns' this was removed in the 'Adjusted' model. Here, measurement models express the observed (see Table 6.3.2) variables through four latent constructs:

1. The latent exogenous variable 'Socio-economics' is explained by five observed variables, which are percent low income, percent medium income, percent high income, spending uptake and percent female.

2. The latent exogenous variable 'Density' is explained by network event density and building density

3. The latent exogenous variable 'Diversity' is explained land use mixture and percent residential.

4. The latent endogenous variable 'Travel patterns' is explained by two observed variables inbound trips and outbound trips.

Second, a SEM with latent variables is designed and implemented to test the hypothesised explanatory relationships and effects. The original model is akin to the alternative relationship identified by Stead (2001) [320] shown in Figure 1.4. The interim model extends this by considering density and diversity as separate latent constructs in line the D's identified by Cervero and Kockelman (1997) [69] while considering 'trajectory distance' and inbound/outbound trip frequency as part of a combined latent construct 'Travel Patterns'.

Based on the goodness of fit results from the previous step, a third 'adjusted' model was designed (see Figure 6.4b). The adjusted model has three exogenous latent variables ('Socio-economics', 'Density' and 'Diversity'), an endogenous latent variable ('Travel patterns') and considers ('trajectory distance') outside of the latent construct 'Travel patterns'.

6.4 Results and Discussion

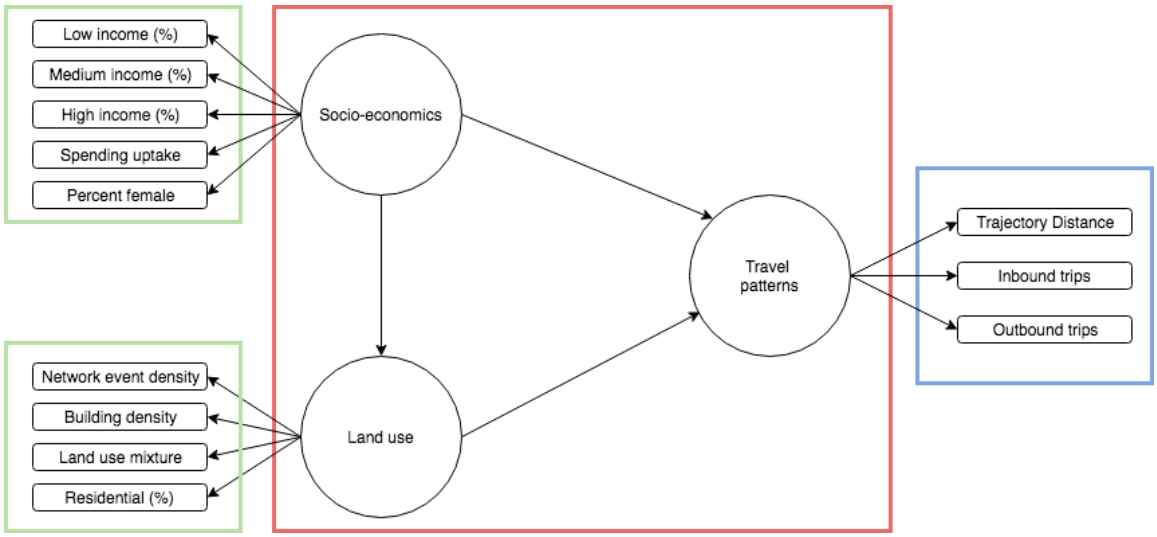
Both the more conventionally used multiple linear regression and a SEM with latent variables were used to analyse the LU(S)TI relationship in the Tanzanian port city of Dar es Salaam.

6.4.1 Multiple Linear Regression

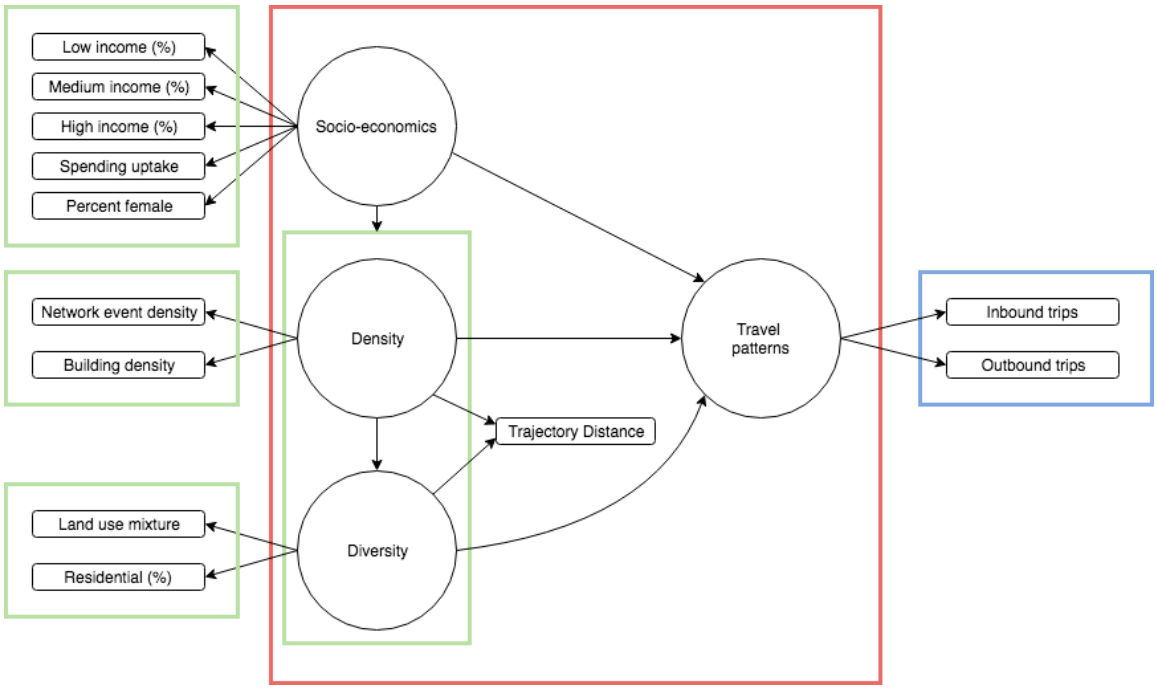
The data set was standardised before the application of multiple linear regression to test the relationship between the different dependent variables for each set of independent variables. The results for the regression analysis with each trajectory distance, inbound trips, and outbound trips are shown in Table 6.4.1.

The results confirm earlier findings from India [8], Taipei [219] and elsewhere [69, 119, 178, 198], that both density, expressed through network event density, and land use mixture are both significantly negatively correlated to the trajectory distance travelled. While the trajectory distance travelled has a significant positive correlation to the percentage in low, medium and high SEL in an area, there seems to be no strong differentiation between the three and the distance travelled.

Inversely, however, both network event density and land use mixture are positively correlated with the number of inbound and outbound trips, indicating that increasing density and mixture lead to higher trip attraction. Similar to the trajectory distance, percent low, medium and high-income exhibit a significant positive correlation to both inbound



(a) Original SEM model



(b) Adjusted SEM model

Figure 6.4: Subfigures (a) and (b) illustrate the different SEM with latent variables with the measurement sub-models for the exogenous constructs (green), sub-models for the endogenous constructs (blue) and the structural model (red).

and outbound trip rates to a higher degree than for distance.

Table 6.5: Results of multiple linear regression for each of the three dependent variables and the set of variables from the socio-economic and spatial dimension post-standardisation.

Category	Variable	trajectory distance		Inbound trips		Outbound trips	
		Coeff	t-Score	Coeff	t-Score	Coeff	t-Score
Socio-economics	Percent low income	**1.0091*** **	4.993	**1.6209*** **	6.837	**1.6226*** **	6.787
	Percent medium income	**0.8683*** **	6.801	**0.8358*** **	5.580	**0.8362*** **	5.537
	Percent high income	**1.0502*** **	5.113	**1.0215*** **	4.239	**1.0261*** **	4.223
	Spending uptake	-0.1022	-1.276	-0.0863	-0.918	-0.0832	-0.878
	Percent female	-0.0491	-0.842	0.0485	0.708	0.0476	0.689
Land use	Network event density	**-0.3362*** **	-3.795	0.1215	1.169	0.1240	1.183
	Building density	-0.0486	-0.493	**0.2242* **	1.939	**0.2245* **	1.925
	Land use mixture	**-0.2731*** **	-4.653	0.1043	1.515	0.1084	1.561
	Percent residential	**-0.3619*** **	-4.422	-0.0687	-0.715	-0.0693	-0.716
Summary Statistics	Number of observations	86		86		86	
	Adj. R-squared	0.764		0.676		0.670	
	Log likelihood	-55.050		-68.781		-69.499	

Notes: significance is expressed: *** denotes significance at p <0.01, **denotes significance at p <0.05 and *denotes significance at p <0.1

6.4.2 Structural Equation Model

Table 6.4.2 shows the results of the adjusted SEM model. The first and second column contains the model variables; the third column contains the values of the coefficients of the model (here named as 'Regression Weights'); the fourth to sixth columns contain the value of the standard error (S.E.) of each coefficient, the Z-Score and the probability level (P) that the estimated coefficient is significantly different from zero, respectively.

Percent low income (ξ_1) has the biggest impact on the latent variable socio-economics (η_1) followed by a significant negative impact of the percent high income. Percent medium-income contribution to the overall latent construct socio-economics is almost negligible, which may stem from the fuzziness of the label caused by the high level of diversity within Dar es Salaam identified during the analysis of SEL in Chapter 4.
Network event density (ξ_6) has the biggest impact on land use density (η_2) as it can be seen as a better proxy for identifying higher level of activity than the sheer building density inferred from satellite imagery.
The land use mixture (ξ_8) identified through analysis of activity signatures in Chapter 3 appears to have the biggest impact on the diversity of an area compared to the percent residential (ξ_9) of a ward area.

The latent endogenous variable travel patterns (η_4) seems to be nigh equally affected by the number of inbound and outbound trips due to relative similarity in scores. The latent exogenous variable with the biggest positive effect on Travel patterns is socio-economics. This finding confirms that the built environment alone cannot account for differences in mobility behaviour and that once the socio-economic (and personality) dimension become accounted for, the explanatory value of land use for mobility behaviour becomes nigh negligible [197, 205, 320]. Considering the latent endogenous construct of 'travel patterns' made up of inbound and outbound trip frequencies, density carries a significantly smaller weight than socio-economics, while the explanatory value of diversity is negligible. While density was found to increase trip generation similar to earlier research in Taipei by Lin

and Yang (2009) [219], they found diversity to show an opposite effect on trip generation. Both density and diversity have a significant negative effect on the trajectory distance travelled, however. This confirms findings of the multiple linear analysis discussed above, highlighting that high density and land use mixture increases the likelihood of activity locations being in close proximity, therefore reducing trip distances.

In addition to the direct effects and regression weights discussed above and highlighted in Table 6.4.2, it is necessary to discuss the indirect effects caused by the interrelationships among the latent constructs. Those can have a combined total effect that may be different from the direct effects above and can have different signs, leading to different conclusions. Table 6.7 shows that, in terms of total effect, socio-economics remains the major latent exogenous variable to influence travel patterns (η_4), while Diversity now has a negative effect on the travel trip attraction confirming earlier findings by Lin and Yang (2009) [219] that were not confirmed without accounting for the socio-economic dimension.

Table 6.6: Results of the Structural equation Model with latent variables.

			St. Regression Weights	SE	Z-Score	P-Value
Latent endogenous variable	<–	*Latent exogenous variable*				
Travel patterns (η_4)	<–	Socio-economics (η_1)	0.796	0.101	7.886	0.000
Travel patterns (η_4)	<–	Density (η_2)	0.182	0.082	2.230	0.026
Travel patterns (η_4)	<–	Diversity (η_3)	0.036	0.096	0.377	0.706
Latent exogenous variable	<–	*Latent exogenous variable*				
Socio-economics (η_1)	<–	Density (η_2)	-0.023	0.102	-0.227	0.820
Socio-economics (η_1)	<–	Diversity (η_3)	-0.123	0.124	-0.992	0.321
Diversity (η_3)	<–	Density (η_2)	-0.083	0.106	-0.789	0.430
Observed endogenous variable	<–	*Latent exogenous variable*				
Trajectory distance (ξ_{12})	<–	Density (η_2)	-0.756	0.062	-12.288	0.000
Trajectory distance (ξ_{12})	<–	Diversity (η_3)	-0.502	0.213	-2.356	0.018
Observed exogenous variable	<–	*Latent exogenous variable*				
Percent low income (ξ_1)	<–	Socio-economics (η_1)	1.000	-	-	
Percent medium income (ξ_2)	<–	Socio-economics (η_1)	-0.0760	0.126	-0.604	0.546
Percent high income (ξ_3)	<–	Socio-economics (η_1)	-0.970	0.089	-10.866	0.000
Spending uptake (ξ_4)	<–	Socio-economics (η_1)	-0.752	0.104	-7.218	0.000
Percent female (ξ_5)	<–	Socio-economics (η_1)	0.074	0.126	0.586	0.558

Network event density (ξ_6)	<–	Density (η_2)	1.000	-	-	
Building density (ξ_7)	<–	Density (η_2)	0.761	0.079	9.650	0.000
Land use mixture (ξ_8)	<–	Diversity (η_3)	1.000	-	-	-
Percent residential (ξ_9)	<–	Diversity (η_3)	0.217	0.142	1.525	0.127
Observed endogenous variable	<–	*Latent endogenous variable*				
Inbound trips (ξ_{10})	<–	Travel patterns (η_4)	1.000	-	-	-
Outbound trips (ξ_{11})	<–	Travel patterns (η_4)	1.001	0.003	376.208	0.000

Notes: significance is expressed: *** denotes significance at p <0.01, **denotes significance at p <0.05 and *denotes significance at p <0.1

Latent exogenous variable	Direct	Indirect	Total effect
Socio-economics	0.796	-	0.796
Density	0.182	-0.023 × -0.083	0.184
Diversity	0.036	-0.123	-0.087

Table 6.7: Standardised total effects on latent variable Travel Patterns

6.4.3 Study Limitations

Missing data Due to different data gaps, and outlier detection procedures applied during the analysis of MND in Chapters 3-5, BTS and by extension Voronoi-polygon level metrics were not generated for all areas of Dar es Salaam across all the different features listed in Table 6.2. Tandale, one of the most notorious slums in Dar es Salaam, for example, was classed as 86% poor without any information for the remaining 14% as the particular BTS accounting for the gap was excluded as part of the analysis of SEL due to a gap in the ground-reference data used (§4.3.1). Figure 6.5 highlights the completeness of ward-level metrics (Table 6.3.2) across the different wards in the metropolitan area of Dar es Salaam included in the above analysis of land use and socio-economic – transport interaction.

Objective-Subjective divide Prior research has recognised the explanatory power of objective factors from the personality dimension in explaining mobility behaviour (§2.3.2. Unfortunately, however, data on attitudes and lifestyles were not available for inclusion within this thesis research and was therefore not considered.

Sample size One of the main advantages of MND over traditional data sources is the scale at which it can be collected. Due to limited availability of external data such as building counts and other ground reference (§4.3.1) data used as part of this research, variables were only generated for 86 wards within the Dar es Salaam metropolitan area resulting in a comparatively small sample size of <200, which is often regarded as the minimum for the use of an SEM with latent variables. Despite this shortcoming, findings appear to confirm earlier research.

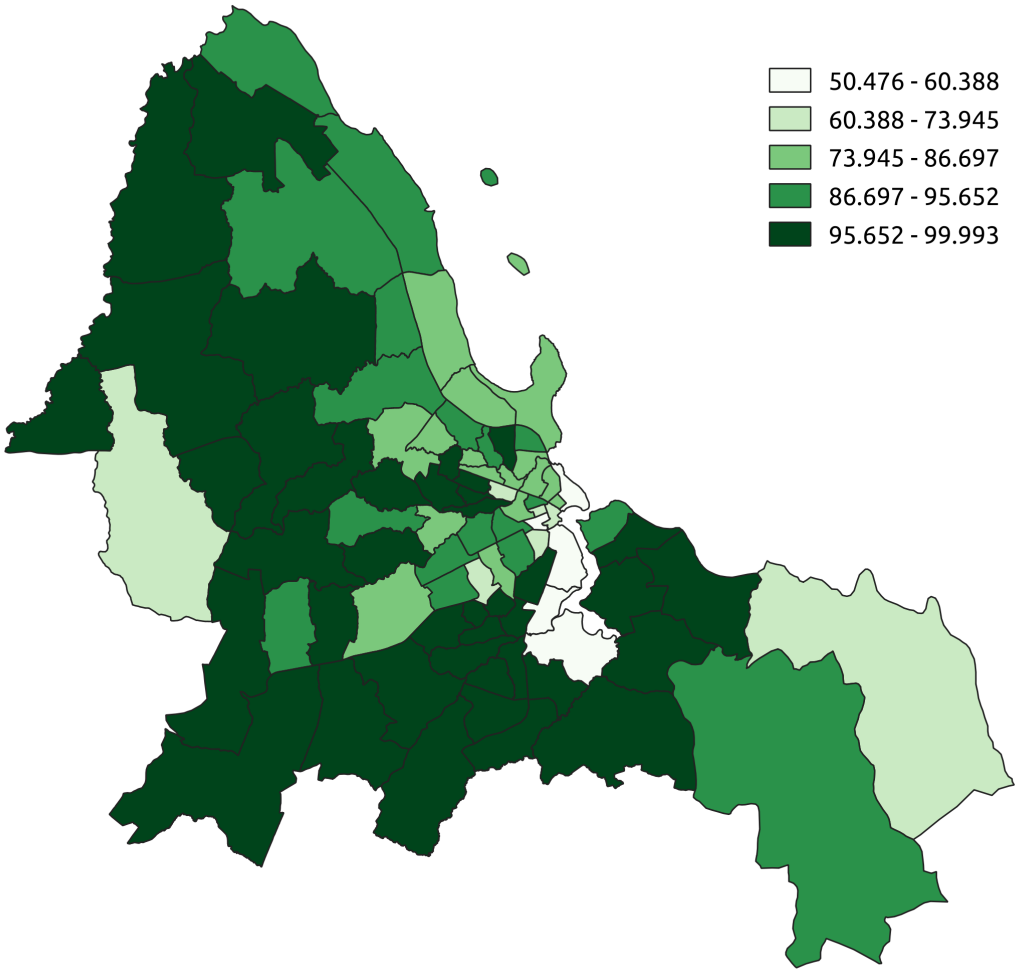


Figure 6.5: Percentage coverage overlap of BTS-level voronoi polygons with official Ward boundary polygons for the metropolitan area of Dar es Salaam

6.5 Chapter Summary

The analysis leveraged variables from the spatial dimension, socio-economic dimension, and mobility mobility metrics generated at the Voronoi-polygon level in Chapters 3 to 5 to analyse the land use and socio-economic – transport interaction for the metropolitan area of Dar es Salaam. Variables were apportioned with existing ward boundaries to allow for the calculation of ward-level statistics on density, land use mixture, gender and SEL for 86 wards within the metropolitan area of Dar es Salaam.

This study confirmed earlier findings that once the socio-economic dimension is accounted for, the explanatory value of land use on mobility behaviour becomes nigh negligible [197, 205, 320]. Similarly, and in line with findings by earlier research [8, 69, 119, 178, 198, 219] density and diversity were found to have a significant negative effect on the trajectory distance travelled indicating, that high density and land use mixture increases the likelihood of activity locations being in close proximity, therefore reducing trip distances.

In addition to the direct effects and regression weights, indirect interaction effects were found to have an influence on trip frequency generation. While diversity was initially showed a positive influence on trip attraction, it was found to have a negative effect on the frequency of inbound and outbound trips captured through the latent endogenous variable travel patterns, confirming earlier findings by Lin and Yang (2009) [219] that were not confirmed without accounting for the socio-economic dimension.

The following chapter will explore limitations arising from the usage of CDR and MFS data, and potential solutions to address those.

Chapter 7

Limitations and Opportunities of Mobile Network Data

7.1 Chapter Introduction

This chapter explores the limitations arising from the usage of temporally and spatially sparse CDR data identified during the analysis of land use in Chapter 3, socio-economics in Chapter 4, and mobility trends in Chapter 5 highlighted in Figure 7.1. The following research objective guides this chapter:

Research Objective 5: identification of shortcomings of both CDR and MFS data, and potential solutions to address those.

Discussing each of the limitations in turn a brief description, and approaches to overcome the limitations is presented. The challenges are broadly classed according to the salient issue they affect or where they occur. Each of the limitations affects a different part of the data collection, analysis and usage life cycle. Some of these shortcomings can be addressed during the pre-processing stage, while others require a trade-off between granularity and volume during subsequent analysis and usage.

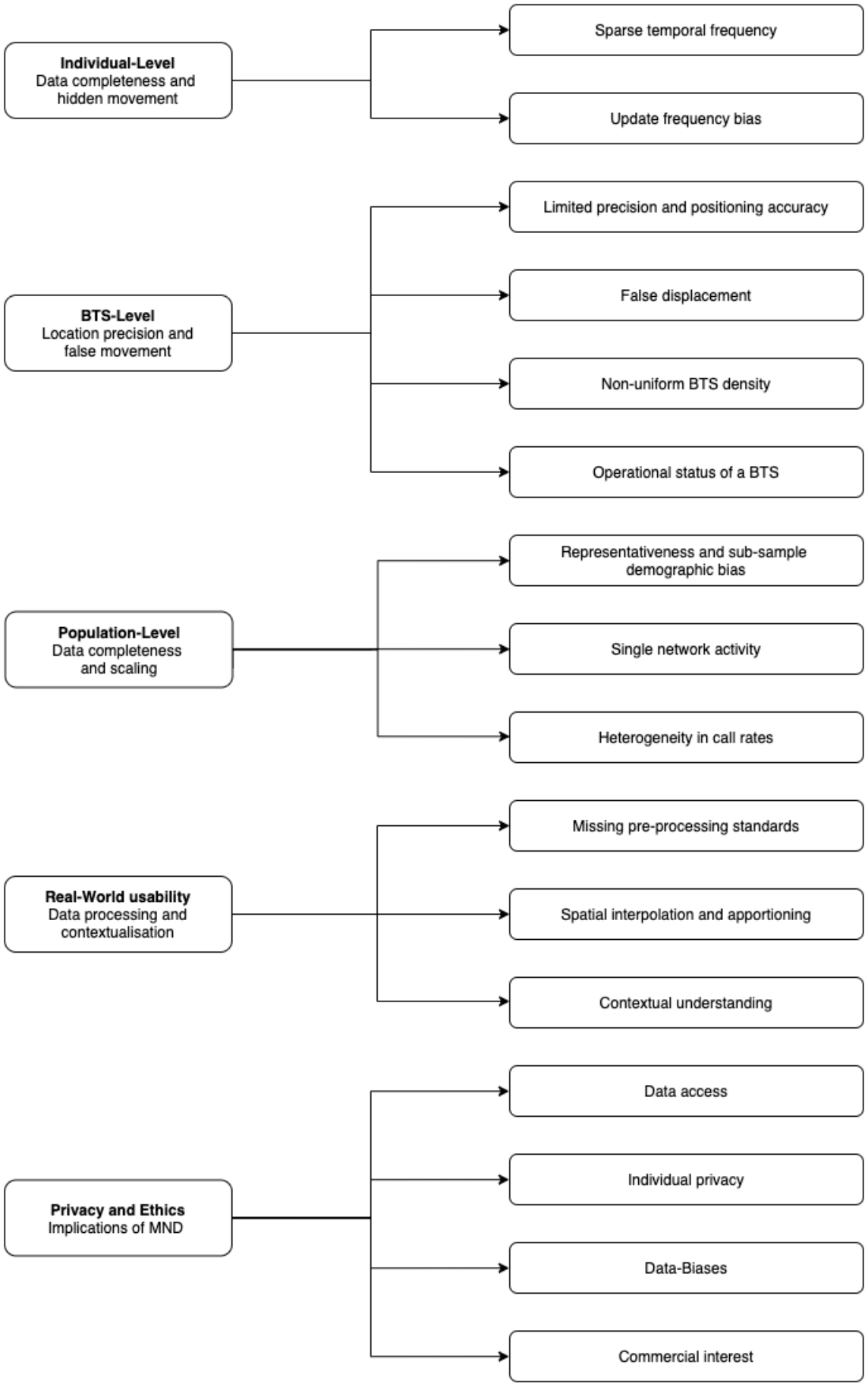


Figure 7.1: Overview of limitations of MND across the Individual-level, BTS-level, Population-level, Usability, and Privacy and Ethics

7.2 Individual-Level Data Completeness and Hidden Movement

7.2.1 Sparse Temporal Frequency

Challenge: In contrast to GPS traces with customizable sampling frequencies, MND data is only recorded when a network event takes place as discussed in §2.2.1. GPS traces are generally recorded with a high-enough frequency to be treated as continuous trajectories [63]. The average inter-event time, the duration between two consecutive network events for a user such as sending an SMS before placing a phone call, for CDR data, on the other hand, can range from a few seconds or minutes to multiple hours depending on a user's usage patterns. Related research has found average inter-event times of 8.2 hours for 100.000 users over six months counting just active network events [142]. The lowest durations were found to be 260 minutes on average with a median 84 minutes once CDR was used in conjunction with location management data [58].

Given the long inter-event times within CDR datasets, it can be assumed that most users exhibit some level of hidden movement. A user could use his or her mobile phone at one point for example and then be in a completely different part of a city 6 hours later without the dataset capturing how a user got there or what happened in the meantime. While every user is likely to have some form of 'hidden movement' and contribute to a resulting 'low-frequency bias', this is significantly exacerbated with users displaying low numbers of irregular network events [63]. The low-frequency bias is likely to significantly affect the representativeness of predictions unless findings are adjusted accordingly.

Discussion and solution: Human mobility is highly predictable; however, as humans tend to follow similar patterns such as commuting from home to work over time [6, 34, 187, 314]. This assumption can be used to reduce the amount of hidden movement and the associated low-frequency bias. The extraction of frequent mobility patterns from longitudinal time series can be used to aggregate location information over time to fill

gaps within a user's mobility patterns. The frequency-clustering based approach to OD matrix and synthetic activity plan creation used in Chapter 5 is based on this principle.

Alternatively, the inter-event times can be lowered through the use of passive location management data such as LAU or handovers discussed in §2.2.1 and CDR's on mobile data usage. There is an increasing shift in usage patterns away from SMS and phone calls toward social network services such as Facebook Messenger, Telegram and WhatsApp, which rely on mobile data usage. This results in a decrease in the number of CDR's associated with SMS and phone calls being generated, further exacerbating the sparse temporal frequency problem. While mobile data usage increases the number of logged network events significantly [61], smartphone penetration varies quite strongly between different demographic groups and between emerging economies and more economically developed countries. This over-representation can lead cast severe doubts over the representativeness of any predictions made using CDR data as the sole data source [63], an issue which is discussed in more detail below in §7.4.1.

A low-frequency bias present in CDR data can also be addressed through the removal of all low-frequency users from the analysis. The most straightforward approach would be to identify users with distinct gaps in their usage, which could indicate that they have moved network for intermittent periods, and remove them from the analysis. Demissie *et al.* (2016) [91] for example, chose to exclude users with less than 75% of days containing network events during the study period, and users with fever than 1000 network events per week. A similar approach also exists in the blanket exclusion of all users with fever than six network events per day [306, 359]. While this approach can generate more representative insights, it is potentially excluding a vast majority of users from any subsequent analysis. Using the same approaches as in the generation of transient and frequency-based OD matrices in Chapter 5, for example, over 80% of users would have had to have been excluded from the analysis. Nevertheless, normalising or even removing outliers with abnormal usage patterns seems to be the most effective strategy for addressing a low-

frequency bias.

7.2.2 Update Frequency Bias

Challenge: As CDR data is only created when network events are taking place, users with a high network activity are over-proportionally represented within the data set and may, therefore, appear significantly more mobile, leading to a high-frequency bias [79, 80, 291]. Similar to the observation-expectancy bias inherent in manual surveying, a high-frequency bias present in CDR data can significantly affect the representativeness of any predictions based on CDR data alone without further adjustments [63]. When analysing transport demand, Wang *et al.* (2010) [346] for example, found it necessary to adjust generated OD matrices to more accurately reflect transport demand. This issue is exacerbated in the Global South, where costs can lead to multiple users using a single device in exchange for a small fee, generating a high number of network events while generating limited movement [46, 183].

Discussion and solution: Similar to the exclusion of low-frequency users discussed in the previous section, various approaches have sought to address this issue through user exclusion [91, 346]. Early work by Wang *et al.* (2012) [347] categorised users as belonging to one of five categories, based on fixed numbers of network events per month [1]. Those mobility profiles were further developed by Berlingerio *et al.* (2013) [40] into low-high intensity for network activity and low/high intensity for time travelled as user classes based on the aggregation of individual behaviour within CDR data. Jiang *et al.* (2016) [187] instead categorised active users, those with more than 50 stays and more than ten homestays over six weeks, into commuters and non-commuters over the age of 16.
As part of the transient-based OD matrix analysis carried out in Chapter 5, users were split across 5 activity levels through a dynamic rather than a fixed assignment as carried out by Wang (2012) [347].

[1] below 10, 10-500, 500-1000, 1000-2000, over 2000

7.3 BTS-Level Location Precision Issues and False Movement

7.3.1 Limited Precision and Positioning Accuracy

Challenge: CDR data does not record an individual's locations, but rather a non-precise proxy via the location of the BTS delivering the service represented as a Voronoi shape as a coverage area proxy. Sector antenna reception range is dependent on factors relating to the location, height, and the technology involved. A device will seek the sector antenna providing the strongest, and therefore usually nearest, signal [109]. Connection or handover to the nearest BTS is therefore not guaranteed. As a result, the location of a handset cannot be assured. Despite this uncertainty, it can be expected that BTS activity is roughly indicative of activity in the surrounding coverage area commonly estimated as a Voronoi polygon. Combined with a sparse sampling frequency (§7.2.1), the non-precise, areal nature of Voronoi locations significantly affects the detection of short trips. It also affects the estimation of geographical and built environment factors such as distance to the nearest public transit point, which is recognised a key factor in the analysis of accessibility and design within the Spatial dimension (§2.3.2).

Discussion and solution: Partly in order to overcome this issue, and partly due to concerns over individual privacy and commercial interests discussed in §7.6.2, several studies [59, 60, 185] have used triangulated data. In those studies, data was provided by AirSage, a US-based private-sector company using its proprietary Wireless Signal Extraction technology to anonymise, aggregate and analyse mobile phone signal data from multiple MNO's for real-time traffic speed and travel time prediction. Similar companies include IntelliOne in the US, ITIS holding in the UK, Delcan in Canada and CellInt in Israel [145]. Using procedures initially developed for GPS traces, Steenbruggen *et al.* (2013) [321] have triangulated handset locations using signal strength data. As signal strength data was not available in the datasets used as part of this research, handsets could not be triangulated resulting in an aggregation of activity at the BTS level as well as their

surrounding Voronoi cells instead.

Beyond collecting BTS information when a user is engaged in a network event, operators need to keep track of a user's location to direct incoming calls to the appropriate BTS in the network. Redirecting is done through location management processes involving the recording of passive time-based and movement events described in more detail in §2.2.1. This location management can provide more reliable insights into a user's journey. Depending on the operator's choice, location management can be either undertaken as:

Never-update: location information is not collected but rather all sector antenna are 'pinged' to identify the active antenna to direct an incoming call to.

Always-update: the handset is informing the network whenever it is moving into a new cell. While there is no paging cost, the network would most likely get quickly overwhelmed by the frequent updates.

Location-area-update: this approach is a combination of the previous two, as cells are grouped into location areas. The MNO is informed when a user moves between location areas, and the sector antenna in the last recorded location area for a handset are pinged to identify the currently active sector antenna for the direction of an incoming call.

Outside of emerging economies in the United States and the European Union network operators are mandated by law to keep track of handsets to provide emergency services with location approximations in emergency scenarios. Under enhanced 112 in the EU [293] and enhanced 911 in the US [318], operators have to locate users within a 50 meter radius in 67% of cases and 150 meter in 95% of cases. Considering that the imposed accuracies are not currently achievable through BTS-only positioning, new techniques for handset tracking were developed. Operators can use network-centric cellphone positioning, which uses existing network capabilities to triangulate handsets based on time, angle and distance measurements generated from the signal strength of sector antenna and handover information captured through MSC's [293, 318] or device-centric cellphone positioning

using measurements and calculations within the handset itself to effectively triangulate handsets.

7.3.2 False Displacement

Challenge: During periods of peak activity, users may be transferred to nearby BTS in order to balance the load on the mobile network. As the location is based on the BTS information contained in a CDR log, this effect leads to the misidentification of a user's location. This false displacement occurs without the knowledge of the user and follows a similar pattern as handovers related to physical movement, as shown in Figure 2.1. This issue is exacerbated when multiple unrelated network events take place in a short space of time in areas with a high BTS and, by extension, cell density.

Signal jumps can also be caused by similar signal strength of different BTS being registered by the mobile device [175]. As antennas providing the highest signal strength in a given area vary under different circumstances such as time of day, built environment changes, surface reflection, etc. the link is mostly stochastic rather than deterministic [181]. Limitations in the signal range are accounted for by MNO through the intentional overlap of service coverage to minimise the risk of gaps in coverage, thus further exacerbating the likelihood of false displacements taking place [109].

Discussion and solution: Different approaches have been developed to identify false displacement caused by peak BTS redistribution: speed, patterns, and hybrid approaches.

Speed-based Methods require a high-temporal frequency within the dataset and can be used to identify sequences where switch speed are above a predetermined threshold. Identified logs are removed from the data with the main challenge in the identification of an appropriate threshold [175]. Horn [169] for example chose a filter of 250km/h to smooth travel speeds and times in their comparison of a recursive naive filter, recursive look-ahead filter (both effectively low-pass filter) and Kalman filter. One of the most effective ways to tackle the false displacement effect is through

the usage of a lower-end temporal filter as part of the data analysis. A temporal filter can be used to discard consecutive records with an 'unrealistic' displacements within 10 minutes of each other [347, 349]. Previous studies [60, 176, 185, 347] have routinely chosen a low-pass filter of 10 minutes in the detection of stops for transient-based mobility analysis. While a temporal filter can alleviate the detection of false displacement somewhat, it can also lead to an increase in the amount of hidden movement and reduction of short journeys detected. Similarly, it can prevent the detection of movement occurring when a user is travelling on the boundary between two BTS coverage areas or location areas.

Pattern based methods involve the identification of oscillations in chronological recordings. Lee and Hou (2006) [211] identified an oscillation during initial research with Wi-Fi networks when three consecutive switches between a pair of locations are observed and referred to it as the ping-pong effect. In those cases, the BTS with the longest duration was selected as the stay BTS. Bayir *et al.* (2010) [33] recognised a misidentification risk of oscillations for user frequently travelling between 2 locations. They chose to generate multiple mobility paths per day with trips estimated using a frequent-sequence mining algorithm. Other studies have instead used a spatial filter either ignoring cells that are Voronoi neighbours and only considering those above a certain distance [152] or grouping nearby events into a stop [60]. While this approach does not require the application of a temporal low pass filter, it can still lead to a non-detection of activity unless movement takes place in densely packed urban areas where BTS are generally close to one another. Another solution to addressing peak BTS distribution can be the exclusion of surrounding towers as part of the stop detection in areas of high BTS density.

Wang (2014) [348] combined speed and pattern-based methods by first detecting subsequences using a pattern based approach before determining switching speeds and updating BTS pairs above a specified speed threshold.

7.3.3 Non-Uniform BTS Density

Challenge: Due to a lower population density and associated lower levels of activity, network operators tend to operate fewer BTS in rural areas. This lack in demand can result in BTS being located many kilometres apart in peri-urban and rural areas. Densely populated urban areas, on the other hand, often have a median distance of just a few hundred meters between BTS's [59, 73]. The Colombo District in Sri Lanka, for example, is both the most urbanised region in the country and accounts for a significantly higher BTS density than any other district in the country [228]. The varying density in BTS coverage impacts the accuracy of descriptive, predictive and prescriptive insight generated outside of densely covered urban areas. Short and intra-area trips become harder to detect as coverage areas increase [253].

Discussion and solution: Grid-based approaches have been suggested as a possible way to mitigate the issues of appearing less mobile in rural areas and account for varying the spatial intensity of BTS distribution [222, 360]. The study area is split into rectangular grid cells with identical dimensions based on the level of analysis performed, for example, 5km square cells for district-level mobility flows or more high-resolution 1km cells for road network traffic analysis. BTS situated within a cell are assigned to the centre of the cell and grid cells not containing any BTS are excluded from the analysis. One major limitation is the rising generation of artifacts for events that are close to the border of the cells resulting from the quantisation of space. This is exacerbated through the false-displacement effect, when an individual calling from within a particular cell may be connected to a BTS in a neighbouring cell.

> "Such artifacts can be accounted for by performing the analysis based on all possible grids at the same resolution by shifting the grid on the area of interest based on a suitable shifting distance and considering the average value of the measures being used." [228, p.792]

This approach also helps mitigate the impact of a single location being served by multiple nearby BTS at different points in time by assigning average locations to neighbouring

BTS based on the considered spatial resolution.

7.3.4 Operational Status of a BTS

Challenge: Operational status of BTS infrastructure can change over time due to malfunctions and/or the installation of additional BTS's to meet demand. What may appear as a data integrity issue at first, can be directly related to real-world events such as power outages or disasters affecting the structural integrity of a BTS. In the case of Kashmir, for example, BTS had been deliberately shut off as they were used by militants for targeting of specific groups [297]. At other periods, BTS may be put out of operation temporarily by network operators to cut down on operation costs during periods of low network activity. Besides limiting communication capabilities for users in the coverage areas, BTS blackouts may affect predictions and results gained from longitudinal analysis using CDR data. Such alterations directly affect the precision of location, and by extension, the insights generated from CDR data analysis. In a study on population displacement following Port au Prince, Haiti earthquake in 2010 Bengtsson *et al.* (2011) [39] found an impact on geographic positioning precision due to BTS malfunctions.

Discussion and solution: Detecting power down-time and assessing the impact on the reliability of predictions can be difficult depending on the level of aggregation present in CDR data available. Checking for towers with very low levels of activity in a process called sparsity analysis can help identify BTS that are subject to frequent changes in operation. Analysing the sparseness of BTS in an OD matrix was undertaken by [269]. They searched for towers with 0 values between regions as indicators that those regions are not generating or attracting trips. Another approach could be to visualise analysis results in the form of time-sliced Voronoi. Visualisation can be undertaken through grid-interpolation to protect commercial interests, similar to the approach described in §7.3.3 on BTS density.

7.4 Population-Level Data Completeness and Scaling

7.4.1 Representativeness and Sub-Sample Demographic Bias

Challenge: Representativeness is largely neglected in big data research, and some form of demographic bias will always be inherent in any study on human behaviour [221, 378]. CDR data is restricted to subscribers of the network operator providing the data, which represents only a sub-sample of the entire population in a country. This is caused by competing operators and less than complete market penetration of the MNO providing the data across the entire population with and without handsets. Each MNO's CDR data is likely to be further biased towards different population groups due to operator marketing and pricing strategies.

Even with access to large amounts of CDR data from all network operators in a country and the assumption that mobile phone penetration is close to 100%, findings will not necessarily be representative. There will always be certain groups of people such as those without access to mobile phones that are often from a lower socio-economic background and/or within certain age group demographics. There are, however, also those users with anomalous patterns that were removed as part of the pre-processing of data so as not to skew results, that are not informing the final insights.

> Zhao *et al.* (2016) [378] argue, that "the representativeness of individual trajectories derived from CDRs are strongly influenced by peoples habit of using mobile phones at certain locations and time in a day. For example, CDR trajectories of a traveling salesperson who talks to his/her customers on a mobile phone regularly may well depict his/her daily movements, whereas CDR trajectories of a person who uses his/her phone occasionally should not be used to understand his/her mobility pattern in space and time. As a result, it is important to investigate to what extent we can trust the mobility indicators derived from CDR trajectories and the conclusions drawn from these indicators. It should be noted that using CDRs collected over a long period of

> time as a workaround cannot address this issue as people who rarely engage in phone communication remain underrepresented."

Due to these sub-sample biases, care must be taken not to further bias results towards the specific demographic of users subscribing to the network providers services. Consequently, all results must be considered as intermediary *a priori* results offering relative, rather than absolute insight until scaling occurs.

Discussion and solution: Most studies have used some form of expansion factor to account for sub-sample biases, which has shown some initial success [91, 354]. For population-statistics-derived expansion factors, the actual population of a polygon is divided by the number of users who have been classified as a zone's residents [78]. The approach, however, requires population statistics which generally come from census surveys [12, 58, 78, 335]. Expansion factors derived from census information may be grossly inaccurate in certain areas, however. Generally, censuses are only carried out every couple of years with the gap increasing rapidly in emerging economies around the world due to the general 'Statistical Tragedy' affecting Africa in particular. Furthermore, new growth location areas may be underrepresented in census information due to the relative infrequency of data collection as was the case in the work on MODLE project with the Transport Systems Catapult §5.3.1. Low numbers of CDR users in certain areas can further contribute to inflated expansion factors [78].

Alternative data sources to census surveys for the calculation of expansion factors for mobility studies include travel surveys such as RSI or panel surveys [38, 56, 59, 186, 238, 333, 335, 347, 357], combinations of census and travel surveys [38, 187], probe vehicle traces [347], traffic flows from video recordings at main intersections [176], national travel survey and odometer readings from safety inspections of private vehicles [59] and pre-existing OD's from Census Transportation Planning Products and Rio transportation plan [78]. While existing information on transport demand and other mobility trends can be used to validate the results, questions over the accuracy of the external information itself remain.

An alternative approach for mobility studies with CDR data is the extrapolation of expansion factors through modelling and micro simulations. OD matrices have been shown to work well with gravity models as they do not require knowledge of model parameter values beyond a high adjusted R-squared [60, 82, 333]. Flow models, on the other hand, have been shown to perform poorly with spatially and temporarily sparse CDR's for OD estimation and network assignment [152]. Another approach has been the usage of the microsimulation platform MITSIMLab for calibration by Iqbal *et al.* (2014) [176]. The scaling factors were based on cellular penetration (different provider and non-user), mobile phone non-usage (not used at every transient point) and vehicle usage (users may not use cars for every trip).

7.4.2 Single Network Activity

Challenge: It is not uncommon for mobile users in emerging economies to have access to multiple SIM cards with each operator subsequently only receiving a sample of their overall network events, potentially biased towards certain usage situations [9]. Switching network operators briefly can offset some of the issues associated with limited network coverage when travelling. Another reason for multi-SIM use includes the anticipation of network downtime, which is a particular concern in rural areas within emerging economies. Furthermore, intra-network calls, calls between two users of a particular network operator, are often associated with lower usage costs than inter-network calls. Some users might, therefore, opt to use different SIM cards for calls to different user groups [92, 97].
Single Network Activity bias can be broken down into three main challenges:

1. The systematic loss of network events due to usage behaviour;

2. The temporary loss (or gain) of network events due to the CDR operator's network (or their competitors) becoming temporarily unavailable; and

3. The systematic, per individual, loss of events in a given region as they change providers to maintain/improve service.

Discussion and solution: Although further work on bias correction methods is necessary, ground-truth data has previously been used to estimate the impact of differential ownership on inferred human mobility models in Kenya and found CDR generated insight to be promising [354].

7.4.3 Heterogeneity in Usage

Challenge: Individual level insights are both positively and negatively affected by biases in call rates and general usage [220].

- Positive biases occur in waiting areas such as bus or ferry terminals, where users are stopping to wait for their transit services while potentially using their phones at higher rates than normal to 'kill time' which can lead to an overestimation of stop-frequencies.

- Negative landline biases may occur in locations such as home or work where landlines may be used in favour of mobile connections, which can lead to a significant reduction of detected traffic. As users remain stationary for prolonged periods, there can be a shift from using mobile phones to using landline connections. With reduced mobile phone usage also comes an increased low-frequency bias (§7.2.1).

- Negative morning biases also occur as individuals tend to use their phone more during the later portions of the day, leading to fewer events and therefore fewer AM trips being detected [11]. Similar variations in seasonality also occur between weekdays and weekends that generally show different levels of activity [29].

Discussion and solution: Negative biases related to landline usage are less of an issue in emerging economies due to the high ratio of mobile phone vs landline connections. A recent report by the International Telecommunication Union has shown that in 2012 mobile phone connections were on average 6× higher than landline connections in Asia-Pacific, Latin America and the Middle East, compared to 3× in the European Union and 2.3× in Northern America. The most substantial gap was found in Sub-Saharan Africa

where mobile connections were almost 50× higher than landline connections as can be seen in 7.2. Ground truth information can be used to identify areas with heterogeneous movement patterns such as ferry ports, airport terminals and bus terminals, where users remain stationary for prolonged periods of times while using their phones more often than usual. One can assume that these stationary high-activity patterns are higher in users with access to mobile data services than those without. A more detailed classification of scaling filter can potentially be used to overcome heterogeneity issues [176].

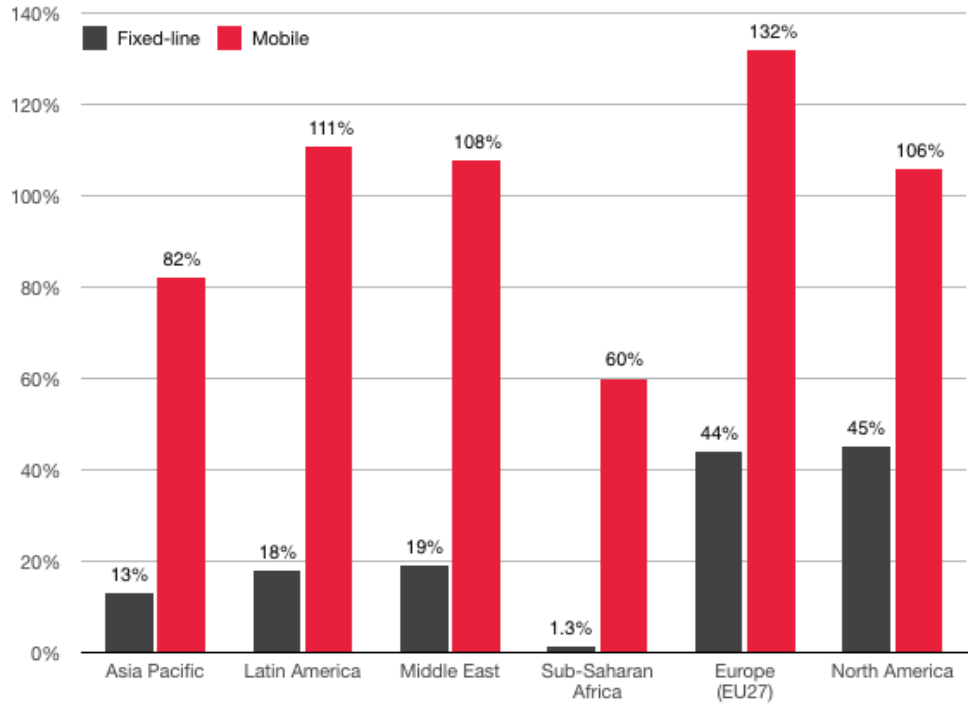


Figure 7.2: Disparity between land-line and mobile connections based on registration statistics from 2012 [148]

7.5 Real-World Usability, Data Processing and Contextualisation

7.5.1 Missing Pre-Processing Standards

Challenge: There also remains much work left to do on developing standardised guidelines for the acquisition, conservation and usage of personal data. No pre-processing

standards for CDR data currently exist due to the large variety in both CDR granularity available to third parties due to considerations on privacy (§7.6.2) and commercial interest (§7.6.4). Automatic procedures for the processing of large amounts of data in real-time, and more flexible models, are however required for inclusion of this MND into already implemented transport forecasting models as those discussed in §5.1.2 [212, 331].

Discussion and solution: Initial work in the new field of Human Data Interaction is considering how infrastructure, models and interfaces are built to enable users of these to understand and engage with data processing [249]. Common ways of storing and sharing microdata should be defined. At least for geospatial data, the ESRI shapefile is generally recognised as a standard data storage format with a white paper by ESRI providing specifications for sharing of shapefiles across software solutions and platforms [116]

7.5.2 Spatial Interpolation and Apportioning

Challenge: As discussed in §5.1.2, inferred mobility insights are used for Transport Forecasting across different TMZ in a city. TMZ have traditionally been designed manually based on basic principles of (1) cluster social, land use and economic characteristics; (2) frame zone boundaries around natural and human-made boundaries such as rivers and rails; and (3) selecting manageable traffic zone sizes outlined by Ortuzar and Willumsen (2011) [269]. Voronoi cells used to visualise approximate coverage areas for CDR-based studies rarely overlap with pre-existing TMZ areas, instead of requiring apportioning or spatial aggregation.

Discussion and solution: One possible approach involves assigning BTS to adjacent TMZ as was undertaken by Jiang *et al.* (2015) [186] for sub 200 meter distances in Singapore. An alternative approach involves designing a conversion matrix for zoning areas based on building polygons from *BDTopo* database for filtered area weighting [50] and the normalisation of observations [181]. The interpolation of discreetly bordered areal units through apportioning or aggregation invokes the modifiable areal unit problem defined by Openshaw [268]. The Modifiable areal unit problem, however, raises the question of what

scale, and zonation or aggregation are appropriate for the analysis task at hand [75].

7.5.3 Contextual Understanding

Challenge: Traditional data collection approaches collect information on individual travel reasons, mode choice, socio-economic and demographic characteristics, which are not explicitly included in CDR data. Also, the understanding of individual behavioural phenomena requires an understanding of external contexts that can influence user behaviour. The analysis of commuting patterns through transport forecasting models, for example, involves an understanding of modal split and route options for network assignment. When commuting from the south-eastern outskirts into the centre of Dar es Salaam, for example, the two most common routes are a land route past the airport via Miburani or taking a ferry via Kigamboni. Each route is affected by external factors such as the anticipation of prolonged traffic jams or the impact of floods on the road infrastructure and ferry timetables at different times. Traffic jams, in particular, affect commuting times significantly and influence the time of the day at which people travel to and from work. In contrast to RSI and panel surveys, CDR data lack directly recorded movement metadata such as trip purpose, mode and vehicle occupancy.

Beyond the understanding of individual contexts, more comprehensive understandings are necessary to design effective interventions. In the case of the Bus Rapid Transit system in Dar es Salaam, for example, one of the depots was built in a high flood-prone area, leading to frequent flooding of the depot. While additional walls to prevent flooding of the depot were installed, these have instead pushed the water into nearby areas leading to flooding of multiple-occupied tenements.

Discussion and solution: When generating descriptions of transport demand, it is necessary to distinguish between activity-based travel such as running and cycling and vehicle-based travel. While distinguishing between transport modes is possible using traditional road-side surveys or to a certain degree using sensor technologies such as smart

cards for use with public transport or road-side cameras, the use of those technologies is prohibitive in emerging economies and rapidly changing urban spaces due to scale factors outlined in the previous section. CDR data does not currently allow us to detect whether a user is travelling using an activity-based or a vehicle-based mode.

Some work has been undertaken on transport mode recognition and human mobility modelling using different smartphone sensors [277, 296, 380]. Caceres *et al.* (2008) [253] proposed a method to estimate the transport mode from *a priori* flows using criteria such as travel speed, travel time and traffic information. An alternative approach by Larijani *et al.* (2015) [209] sought to identify subway travel, assuming that dedicated BTS serves subway tunnels. As part of the MODLE project discussed in §5.3.1, rail trips were differentiated from automobile trips via the analysis of LAU patterns. Rail trips were shown to have distinctively clustered LAU patterns compared to scattered patterns for car journeys as passengers on a fast-moving train were travelling across boundaries quasi simultaneously. Rail trips may have been misattributed as road trips in the case of short trips between LSOA areas. Some trips may have been misclassified as walking trips and subsequently been removed from the data during the mode detection step undertaken by the data provider. Aeroplane journeys were identified as making use of BTS in two distant LAU's at high travel speeds and without distinctive LAU handover patterns.

These projects have however been undertaken in the developed world, where there are fewer and easier distinguishable transport modes, i.e. train, tram, bus, metro [161]. Vehicle classification in developing countries has so far only been attempted by Mallikarjuna *et al.* (2009) [230] using video image processing from roadside cameras and Garg *et al.* (2014) [133] using smartphone sensors for recognition. Both have limitations as camera infrastructure may be limited, and smartphones are not as widely used as feature phones in the developing world as discussed in §5.2.1.

Route detection from non-continuous data such as spatially and temporally sparse CDR

data is nearly impossible due to the absence of intermediary stop points. Fillekes *et al.* (2014) [122] tried applying map-matching techniques originally developed for use with GPS trajectories for route assignment, but found those to perform poorly with CDR data due to the temporal sparsity (§7.2) and spatial granularity (§7.3.1) of CDR data. This issue is further exacerbated through the absence of accurate base-maps [99, 173]. Using location management data to address both issues of temporal sparsity (§7.2.1) and spatial granularity (§7.3.1), Gundlegard *et al.* (2016) [152] have shown, that ad hoc route detection is possible for trips that cross multiple LAU boundaries only. Human mobility was found to be highly predictable, and several studies have argued, that near-random individual trajectories are rare as we follow similar patterns over time, allowing for longitudinal identification of common routes such as home-work commutes [6, 34, 142, 314].

In addition to modal split and route options for network assignments of trip-based models, activity-based models use activity-motifs, which correspond with daily activities such as work, home, shopping, etc. There have been attempts to extract and characterise where individuals stay and pass by in urban locations and subsequently infer types of activities (i.e. activity-motifs) they engage in [185]. One option to identify motifs is through the identification of points of interest in different parts of a study area and use these to guide the identification of new activity area classes beyond home and work, such as going-out, shopping or administration. Schneider *et al.* (2013) [306] for example identified 17 unique network motifs with simple rules present in everyday activity. They were able to account for 90% of the study population in different countries: each person was found to have a motif persisting for several months similar distributions found phone and survey data [185].

7.6 Privacy and Ethical Implications of MND

7.6.1 Data Access

Challenge: Negotiating access to CDR data is difficult. There are no clear legal and regulatory frameworks surrounding the sharing of CDR information. There are also privacy and security implications involved in sharing data in the first place. Replicability of CDR based research, and (incremental) improvements continue to be a challenge due to the limited access of third parties to mobile network data, including CDR and location management datasets. Raw mobile network data are inherently proprietary, private data sets, and it is not possible to make them fall under the purview of Open Data policies.

Discussion and solution: A clear plan for what research will be undertaken with the CDR data and an understanding of what magnitude of data is required are useful for showing confidence and building trust with network operators as part of access negotiations. In the absence of clear ethical and regulatory frameworks, the Groupe Spcial Mobile (GSMA) [2] developed guidelines for sharing MND with third parties following the deadly Ebola epidemic in Western Africa [149, 112]. Those were first steps toward guiding the third party and academic work with CDR; however, national legislation/government agencies in many countries still fail to regulate the usage of MND formally - frameworks instead exist for other sensitive data sources including NHTS, census surveys and data on taxation [86].

Due to the private nature of personal information encoded in CDR repurposing should only be undertaking using pre-anonymised data in line with GSMA emergent guidelines [149] on the use of CDR data. This is a vital and initial, but not yet sufficient, approach to addressing privacy concerns. Specifically, additional steps need to be undertaken to prevent the re-identification of mobile phone subscribers identities through individual-specific patterns within the data and external information [145].

[2] Most mobile phone operators around the world are members of the GSMA, the mobile phone industry body.

7.6.2 Individual Privacy

Challenge: The sharing and analysis of CDR raise serious individual privacy concerns as mobile devices have become more of an integral part of our daily lives [43, 86]. As discussed in earlier sections, human movement is highly predictable as we tend to visit a limited number of places many times periodically, and many others just once [50]. The identification of a few frequently visited points and approximate visit times can theoretically be enough to identify individuals reliably [87, 375]. Additional concerns include data use without consent, the use of identifiable data, data security and lack of transparency. They are accompanied by technical obstacles, institutional fragmentation and the cost associated with anonymisation and access control. Linking CDR datasets with demographic and/or socio-economic data can raise additional privacy concerns. There are potential challenges with overseas research due to differences in privacy protection laws.

Discussion and solution: The invasive nature of CDR data calls for considerations over what data is actually required for a given analysis task. Out of considerations for user's privacy, all CDR should be anonymised by the network operators in line with GSMA guidelines. There are several technical challenges associated with anonymising large data sets that have been addressed to some extent, such as developing aggregation algorithms that still maintain identity [255, 326] and finding the right level of granularity in data abstraction without losing core characteristics [288]. Previous work, however, has shown, that users can be re-identified even after the anonymisation of data sets [264]. While being possible, the process is difficult and time-consuming [66]. The success of re-identification also highly depends on the spatial granularity of the data. In contrast to GPS data with a longitude and latitude, the geographical information contained with a CDR record is linked to a BTS, which serviced the network event the CDR was created for, and may not even be the closest one in the proximity of the user.

One of the most effective ways to avoid placing individual user's privacy at risk is to focus on the behaviour of masses rather than individuals through data aggregation. Analysing hundreds, thousands or even millions of user's data simultaneously can be an effective

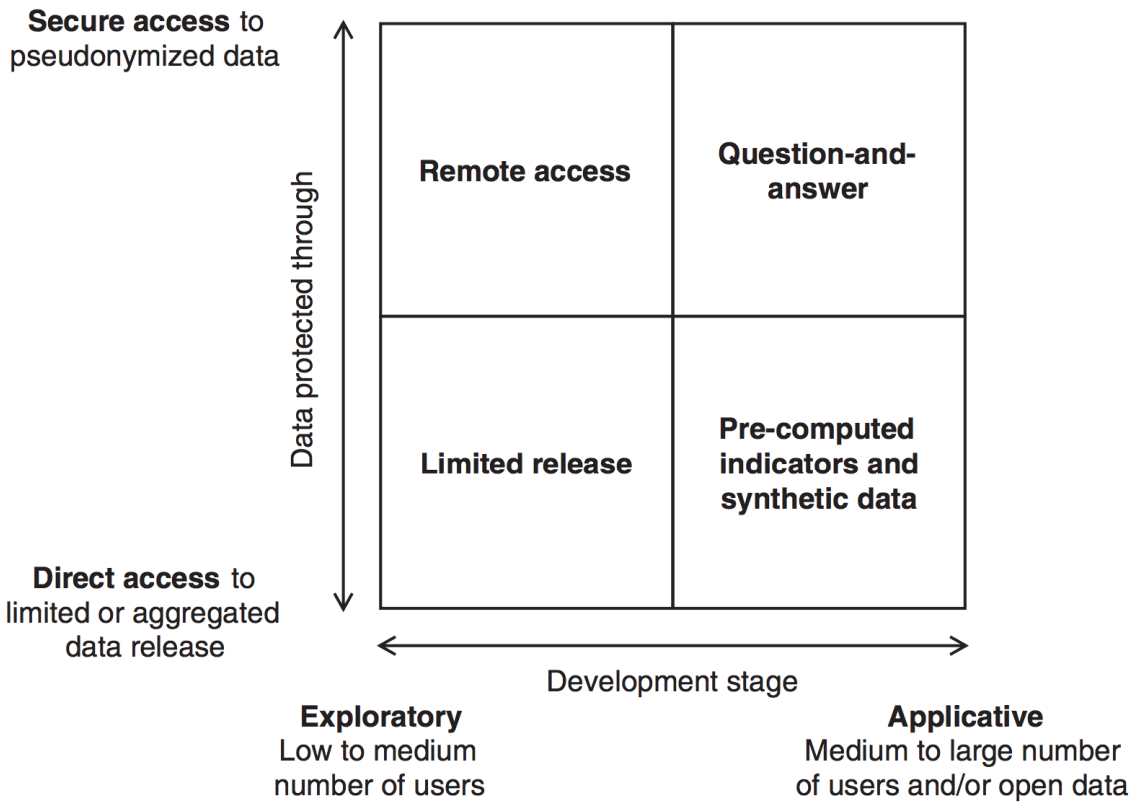


Figure 7.3: Matrix of four models providing a balance between utility and privacy for the privacy-conscientious use of MND proposed by De Montjoye *et al.* (2018) [86]

strategy for ensuring 'privacy in numbers'. While the principle of 'privacy in numbers' is inherent in the aggregation process, the difficulty lies in finding the right balance between retaining detail necessary for generating insights and abstracting to safeguard privacy [112].

Montjoye *et al.* (2018) [86] have proposed four different models, shown in Figure 7.3, for the privacy-conscientious use of MND. All four models assume the use of pre-aggregated rather than raw-data and provide a balance between intended use case and associated privacy risk. This involves regulating who has access to the data (i.e. third parties or only the operator), where it will be stored (i.e. on the operators servers or whether it can be stored off-site) and how processing of the data will be carried out (e.g. whether the operator will be processing queries themselves or if it can be done by third parties remotely).

7.6.3 Data-Biases

Challenge: Beyond individual privacy considerations (§7.6.2), there are also ethical concerns as certain findings can lead to (unintended) structural discrimination in the design of policy interventions. Poorly designed transport interventions for new mass transit routes for example can lead to the isolation of certain groups or areas through access to limited routes or modes of transport as input data may show limited demand stemming from or trip attraction to those areas [267]. This often stems from but is not limited to the aforementioned sub-sample bias (see §7.4.1), which can be further exacerbated in the analysis, as low data areas can be subsumed by others or erroneously excluded:

1. The efficacy of a model is limited by the input data used for training of a model. Learning from inherently biased data leads to biased insights [270], a issue that is only exacerbated with the common lack of accurate and timely ground truth data symptomatic of Africa's 'Statistical Tragedy'.

2. Defining ground reference, such as what constitutes poor, average and wealthy in the analysis of SEL in chapter 4 is often biased by cultural and social assumptions, leading to a trade-off between prediction accuracy and fairness in the choice and design of machine learning models. As discussed in Section 4.4.4, there was no clear indication of what distinguished a 'poor' from an 'average' or a 'wealthy' from a 'very wealthy' area for example in the input data used as part of this thesis, therefore labels were assumed to be 'fair' within the analysis.

3. Sample-size disparity with low-representation groups (§7.2) can further influence the results of algorithms [159].

4. Confounding variable removal such as ZIP codes does not prevent biases from occurring in a lot of cases. Existing research has shown that algorithms are capable of probabilistically infer previously excluded variables as discussed in the previous section on individual privacy [264, 270].

Discussion and solution: Overcoming bias requires an understanding of the data and a combination of technical and design considerations in the selection of the analysis approach. Design considerations of input feature selection and definition of fairness, equitability and ethical standards can ultimately be a vicious circle considering point two discussed in the challenge. Technical considerations such as similarity metrics, auditing behaviour or statistical independence checks can result in an enforced trade-off between prediction accuracy and fairness. An understanding of the (input) data, algorithmic literacy and transparency in terms of informed scepticism over the results of analytical tasks appear to be the most promising avenue for tackling biases [270].

7.6.4 Commercial Interest

Challenge: Beyond individual privacy, there are concerns for the network operators themselves, including damage of public perception in case of data leaks, and commercial interests such as the location of BTS. If it is known which operator's data has been utilised, the analyses can reveal insights into localised network penetration rates.

Discussion and solution: A common process for visualising coverage around a BTS is through Voronoi cells. Voronoi cells are geographic areas, which include all points that are closer to the BTS within the corresponding area than to any other BTS in the area. While visualisations of Voronoi cells do not always include the points used to draw the polygons, they can be easily reverse-engineered. As the points corresponding to the geographical location of BTS cannot be given and are regarded as a commercial secret, estimations have to be given. A common approach is to perform a grid interpolation, a geographic process where grid cells are attributed to the underlying cell, in our case Voronoi cells, occupying the largest part of the respective grid cell, in a process similar to that used to address non-uniform BTS density (§7.3.3). As neither the sizes of the respective Voronoi cells nor the distances between the different BTS can be revealed, an effective approach involves assigning confidence scores to grid cells based on the building counts for a respective cell. Steenbruggen *et al.* (2015) [322]

> "find that previous research follows one of three approaches best-serving polygons, rasterisation, and Voronoi tessellation. Best-serving polygons and rasterisation are representations of the areas covered by specific BTS, and are estimated by the mobile service provider on the basis of tower capabilities and nearby urban form. Voronoi tessellations are a simplified representation of service zones, creating regions associated with cell towers based on nearest proximity alone."

As BTS locations were provided in the current study without indications of the best-service zone and signal strength, Voronoi tesselation is the most suitable approach to the present dataset.

7.7 Chapter Summary

This chapter explored the shortcomings of potentially temporally and spatially sparse CDR and MFS data identified through the analysis of activity-based land use in Chapter 3, socio-economics in Chapter 4, and mobility trends in Chapter 5. Those ranged from individual-level challenges of data completeness and hidden movement caused by the nature of network event generation, to BTS-level challenges of location precision and false movement caused by differences in BTS density, fluctuations in sector antenna range, and missing signalling data preventing the triangulation of handsets, population-level challenges of data completeness of scaling caused by missing ground-reference data and less than complete market penetration of the MNO providing the data across the entire population with and without handsets as well as differences in heterogeneity of usage, to challenges related to real-world usability of data processing and contextualisation, privacy and ethics concerns. Discussing each of the limitations, in turn, a brief description and approaches to overcome those limitations were presented. The proceeding chapter will conclude this thesis.

Chapter 8

Summary and Reflections

8.1 Chapter Introduction

This thesis examined the utility of Call Detail Record (CDR) and Mobile Financial Services (MFS) data for generating insights into urban land use, socio-economic levels and mobility trends; how those can be used to analyse the alternative Land Use – Transport Interaction (LUTI) relationship in an emerging economy context; and identified shortcomings of both CDR and MFS data, and potential solutions to address those. This chapter concludes the thesis, examining the research questions, aims and objectives, how they were achieved, and sets a future research agenda.

8.2 Meeting the Research Question, Aims and Objectives

Chapter 1 introduced the context of rural-urban migration, the 'Statistical tragedy' and barriers to effective data collection and governance, which necessitate the introduction of new approaches to data collection and analysis, in an emerging economy context using the metropolitan area of Dar es Salaam, Tanzania, as a pertinent example of a fast-growing city in the Global South. Further, it set the research question, aims and objectives guiding this thesis. The following subsections examine how the overarching research question, the

aims and objectives were met in the context of this thesis.

8.2.1 Mobile Network Data Analysis

Chapters 3-5 used CDR and MFS data to examine its utility as a novel data source to help overcome shortcomings of more conventional methods of data collection such as manual sampling and physical sensor infrastructure to help address Africa's 'Statistical Tragedy'. Each chapter, in turn, addressed a particular research objective set in the introduction in Chapter 1.

> **Research Objective 1:** to examine whether differences in activity-based land use and density can be distinguished from behavioural patterns contained within CDR data.

Within chapter 3, Non-negative matrix factorization (NMF) was used as a dimension reduction technique to identify key activity-based land use patterns expressing underlying activity spaces such as commuting, working and residential usage. Subsequently, unsupervised k-means clustering was used to categorize areas of the metropolitan area of Dar es Salaam according to the previously identified activity-based land use patterns. The analysis highlighted the potential to use mass CDR datasets for the dynamic analysis of easily interpretable activity-based land use classifications. Furthermore, it was found to be possible to calculate network event density as a proxy for residential density from the observed network events contained within the dataset.

> **Research Objective 2:** the investigation of small area SEL classification using CDR and MFS data through Supervised machine learning, and subsequent analysis of features used for classification to understand the main determinants behind classification results.

The analysis in chapter 4 demonstrated that previously unused MFS datasets could provide sizable improvements in SEL classification accuracy over existing CDR approaches. Comparing MFS data directly to CDR as used in prior work, the results show that MFS

provides an increase in SEL classification accuracy (average F1 score) from 65.9% (0.63) to 71.3% (0.7) at a fine-grained spatial level below most official administrative boundaries as those of wards used within the analysis of the land use and socio-economic – transport interaction in chapter 6. Notably, the combined use of MFS and CDR data only increased prediction accuracy from 71.3% (0.7) to 72.3% (0.71), providing evidence that MFS is informationally subsuming CDR data.

> **Research Objective 3:** exploration of synthetic daily activity plans based on the previously-evidenced assumption that the majority of human movement is predictable, and the generation of transient OD matrices to understand travel and mobility patterns for Dar es Salaam.

Finally, CDR data was used to analyse mobility patterns in the form of trip frequencies, expressed as Origin-Destination (OD) matrices, and trajectory distances in chapter 5. A comparison of a transient-based and frequency-clustering based analysis approach demonstrated that frequency-based clustering approaches, which rely on assumptions of human movement being highly predictable, were found to be less susceptible to differences in individual usage patterns. Those differences included heterogeneity in call rates and individual low- and high-frequency biases, which affected trip rates identified through transient-based trip extraction.

8.2.2 Land Use and Socio-Economic – Transport Interaction

> **Research Objective 4:** analysis of the alternative land use – transport interaction accounting for socio-economic characteristics for Dar es Salaam using variables identified from CDR and MFS data through Research Objective 1-3.

Chapter 6 used variables generated through the analysis of both CDR and MND data in the preceding chapters 3-5 to analyse the land use and socio-economic – transport interaction for the metropolitan area of Dar es Salaam. The analysis confirmed earlier findings that once the socio-economic dimension is accounted for, the explanatory value of land use on mobility behaviour becomes nigh negligible [197, 205, 320]. Similarly in line

with earlier findings, density and diversity were found to have a significant negative effect on the trajectory distance travelled indicating, that high density and land use mixture are associated with a higher likelihood of activity locations being in proximity, therefore reducing trip distances [8, 69, 119, 178, 198, 219].

In addition to the direct effects and regression weights, indirect interaction effects were found to influence trip frequency generation. While diversity has initially shown a positive influence on trip attraction, it was found to have a negative effect on the frequency of inbound and outbound trips captured through the latent endogenous variable travel patterns once indirect effects of the socio-economic dimension, which were identified through the use of Structural Equation Modelling (SEM)'s with latent variables, were taken into account.

8.2.3 Limitations and Opportunities

> **Research Objective 5:** identification of shortcomings of both CDR and MFS data, and potential solutions to address those.

Chapter 7 explored the shortcomings of non-continuously collected CDR and MFS data identified through the analysis of land use in Chapter 3, socio-economics in Chapter 4, and mobility trends in Chapter 5. Challenges ranged across different levels and aspects from the nature of data generation to analysis and real world implications:

- Individual-level challenges of data completeness and hidden movement caused by the nature of network event generation in terms of relative sparsity as events are only generated whenever a network event such as an incoming SMS or placing a phone call occurs.

- BTS-level challenges of location precision and false movement caused by differences in BTS density and activity, fluctuations in sector antenna range, and missing signalling data preventing the triangulation of handsets.

- Population-level challenges of data completeness of scaling caused by missing ground-reference data and less than complete market penetration of the MNO providing the data across the entire population with and without handsets as well as differences in usage of devices across the day and week

- Real-world usability challenges of data processing, contextualisation and spatial interpolation and apportioning with existing analysis areas such as wards or traffic zones.

- Privacy and ethics related challenges ranging from data access due to individual privacy and commercial sensitivity to data biases in the analysis results.

Discussing each of the limitations, in turn, a brief description and approaches to overcome those limitations were presented.

8.3 Suggested Further Research

This thesis has highlighted Africa's 'Statistical tragedy' and barriers to effective data collection and governance, which necessitate the introduction of new approaches over conventional methods of data collection. The research has joined a growing body of research on the use of automatically generated CDR data and illuminated both CDR's and MFS data's potential as a pertinent data source for the generation of insights into land use, socio-economic and mobility patterns, and their interaction at scale and speed to suit the needs of policymakers in fast-changing urban environments. The analysis of MFS data, in particular, is still in its infancy, and there are countless avenues to both continue to research and to explore new potential research directions, offering exciting potential.

Chapter 4 briefly discussed purchase categories, which are included in the MFS dataset provided by the Tanzanian MNO that forms the basis of this thesis research. Despite the exclusion in this piece of research due to their sparse nature, there is the potential to investigate more nuanced usages of MFS uptake among different user groups within Dar

es Salaam and potentially help improve the Socio-economic Level (SEL) carried out.

Chapter 6 analysed the relationship between land use, socio-economic characteristics and mobility patterns in the metropolitan area of Dar es Salaam. It breaks new ground in the Global South by both considering these interactions using SEM in an East African context, and by using variables generated through CDR and MFS transaction logs as proxies of human behaviour as the main input. One direction would be to both scale the analysis to other urban spaces within Tanzania and to consider more nuanced variables from network analysis that have seen study elsewhere.

The longitudinal nature of MND makes it a pertinent candidate for assessing change over time, from changes in the sprawl of informal areas and socio-economic levels to changes in mobility patterns caused by the extension of the cities Bus Rapid Transit Network to changes caused by annual flooding. Further research should consider to what extent rather than if historical data can be used to assess changes within urban areas in emerging economies, and to what extent those changes can be predicted.
The use of CDR and MFS data and machine learning techniques in the generation of insights and (official) statistics by policymakers would be a new direction. It could help bridge challenges to effective data collection and governance in Tanzania and elsewhere.

8.4 Final Conclusions

This thesis concludes that despite the many shortcomings of CDR and MFS data, it provides a significant opportunity to overcome current gaps in official statistics by providing timely and at-scale insight into activity-based land use, socio-economic and mobility patterns. Provided agreements with MNO‘s are in place, MND data can empower policymakers and local municipalities to identify areas requiring interventions and revitalization programs and assess the progress and success of those interventions.

While results obtained through the analysis of CDR and MFS data may not necessarily

be completely accurate in the absolute sense, they can provide a significant improvement relative to the insight currently available as was summarized by the UNFPA:

> "any indicative estimates would provide in certain situations where none are currently available; even if they carried with them a significant level of uncertainty such estimates would still represent a large improvement in many cases" [313, p.519].

Bibliography

[1] Aburas, M. M., Abdullah, S. H., Ramli, M. F., and Ash'aari, Z. H. Measuring land cover change in Seremban, Malaysia using ndvi index. *Procedia Environmental Sciences 30* (2015), 238–243.

[2] ADBG. Tracking Africa's progress in figures. Tech. rep., ADBG, 2014.

[3] Adnan, M., Pereira, F. C., Azevedo, C. M. L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., and Ben-Akiva, M. Simmobility: a multiscale integrated agent-based simulation platform. In *Transportation Research Board 95th Annual Meeting* (2016).

[4] Aggarwal, C. C. On the effects of dimensionality reduction on high dimensional similarity search. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (2001).

[5] Agrawal, R., Imielinski, T., and Swami, A. N. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (1993).

[6] Ahas, R., Aasa, A., Silm, S., and Tiru, M. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research, Part C 18* (2010), 45–54.

[7] Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M., and Zook, M. Everyday spacetime geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and

Tallinn. *International Journal of Geographical Information Science 29* (2015), 2017–2039.

[8] Ahmad, S., and de Oliveira, J. A. P. Determinants of urban mobility in India: Lessons for promoting sustainable and inclusive urban transportation in developing countries. *Transport Policy 50* (2016), 106–114.

[9] Ahonen, T. T., and Moore, A. A mobile phone for every living person in Western Europe: penetration hits 100%. In *Communities dominate brands*. 2013.

[10] Aker, J. C., Boumnijel, R., McClelland, A., and Tierney, N. Zap it to me: the short-term impact of a mobile cash transfer program. Tech. rep., Centre for Global Development, 2011.

[11] Aledavood, T., Lpez, E., Roberts, S. G. B., Reed-Tsochas, F., Moro, E., Dunbar, R. I. M., and Saramki, J. Daily rhythms in mobile telephone communication. *PLoS ONE 10* (2015).

[12] Alexander, L., Jiang, S., Murga, M., and Gonzalez, M. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies 58* (2015), 240–250.

[13] Alexandre, C., Mas, I., and Radcliffe, D. Regulating new banking models that can bring financial services to all. *Challenge Magazine 54* (2010), 116–134.

[14] Alkire, S., Foster, J. E., Seth, S., Santos, M. E., Roche, J., and Ballon, P. Chapter 5 the Alkire-Foster counting methodology. In *Multidimensional Poverty Measurement and Analysis*. Oxford University Press, Oxford, UK, 2015.

[15] Alkire, S., and Santos, M. E. Measuring acute poverty in the developing world: Robustness and scope of the multidimensional poverty index. Tech. rep., OPHI, 2013.

[16] Allan, C., Jaffe, A. B., and Sin, I. Diffusion of green technology: A survey. *International Review of Environmental and Resource Economics 7* (2014), 1–33.

[17] Anderson, J. A land use and land cover classification system for use with remote sensor data. Tech. rep., US Government Printing Office, 1976.

[18] Anderson, J. Handbook of statistics. In *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds., vol. 2. Elsevier, 1982, ch. Logistic discrimination.

[19] Anderson, J. C., and Gerbing, D. W. Structural equation modelling in practice: A review and recommended two-step approaches. *Psychological Bulletin 103* (1988), 411–423.

[20] Anong, S. T., and Kunovskaya, I. Mfinance and consumer redress for the unbanked in South Africa. *International Journal of Consumer Studies 37* (2013).

[21] Ashbrook, D., and Starner, T. Using gps to learn significant locations and predict movement across multiple users. *Pers. Ubiquitous Comput. 7* (2003), 275–286.

[22] Asongu, S., Anyanwu, J. C., and Tchamyou, V. S. Technology-driven information sharing and conditional financial development in Africa. Tech. rep., African Governance and Development Institute, 2017.

[23] Asongu, S., and Odhiambo, N. M. Mobile banking usage, quality of growth, inequality and poverty in developing countries. Tech. rep., African Governance and Development Institute, 2017.

[24] Bacon, J., Bejan, A. I., Beresford, A. R., Evans, D., Gibbens, R. J., and Moody, K. Using real-time road traffic data to evaluate congestion. In *Dependable and Historic Computing*. Springer, Berlin, Heidelberg, 2008.

[25] Bagley, M. N., and Mokhtarian, P. L. The impact of residential neighborhood type on travel behavior: a structural equations modeling approach. *The Annals of Regional Science 36* (2002), 279–297.

[26] Bahoken, F., R. A. Designing origin-destination flow matrices from individual mobile phone paths: The effect of spatiotemporal filtering on flow measurement. In *Proceedings of the ICC 2013 International Cartographic Conference* (2013).

[27] Baker, J. Climate change, disaster risk, and the urban poor. Tech. rep., World Bank, 2011.

[28] Balmer, M., Axhausen, K. W., and Nagel, K. Agent-based demand-modeling framework for large-scale microsimulations. *Transportation Research Record: Journal of the Transportation Research Board 1985* (2006).

[29] Barlacchi, G., Nadai, M. D., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., and Lepri, B. A multi-source dataset of urban life in the city of milan and the province of Trentino. *Scientific Data 2* (2015).

[30] Barnes, S. J., and Corbitt, B. Mobile banking: Concept and potential. *International Journal of Mobile Communications 1* (2003).

[31] Batty, M. *The new science of cities*. MIT Press, Cambridge, MA, 2013.

[32] Batty, M., Hudson-Smith, A., Milton, R., and Crooks, A. Map mashups, web 2.0 and the gis revolution. *Annals of GIS 16* (2010), 1–13.

[33] Bayir, M., Demirbas, M., and Eagle, N. Mobility profiler: a framework for discovering mobility profiles of cell phone users. *Pervasive Mobile Computing 6* (2010).

[34] Becker, R., Caceres, R., Hanson, K., Loh, J., Urbanek, S., Varshavsky, A., and Volinsky, C. Route classification using cellular handoff patterns. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp'11)* (2011).

[35] Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., and Volinsky, C. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing 10* (2011).

[36] Beige, S., and Axhausen, K. W. Interdependencies between turning points in life and long-term mobility decisions. *Transportation 39* (2012), 107–128.

[37] Bejan, A., Gibbens, R., Evans, D., Beresford, A., Bacon, J., and Friday, A. Statistical modelling and analysis of sparse bus probe data in urban areas. In *13th International IEEE Conference on Intelligent Transportation Systems* (Madeira Island, Portugal, 2010), pp. 1256–1263.

[38] Bekhor, S., Cohen, Y., and Solomon, C. Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. *Journal of Advanced Transportation 47* (2013), 435–446.

[39] Bengtsson, L., Liu, X., Thorson, A., Garfield, R., and von Schreeb, J. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Medicine 8* (2011), e1001083.

[40] Berlingerio, M., Calabrese, F., DiLorenzo, G., Nair, R., Pirelli, F., and Sbodio, M. Allaboard: a system for exploring urban mobility and optimising public transport using cellphone data. In *Machine Learning and Knowledge Discovery in Databases*, H. Blockel, K. Kersting, S. Nijssen, and F. Zelezny, Eds., vol. 8190 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, 2013, pp. 663–666.

[41] Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis 52* (2007), 155–173.

[42] Beshouri, C., and Gravrak, J. Capturing the promise of mobile banking in emerging markets. *McKinsey Quarterly* (2010).

[43] Biermann, K. Was vorratsdaten ueber uns verraten. *Die Zeit* (2011).

[44] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[45] Blondel, V., de Cordes, N., Decuyper, A., Deville, P., Raguenez, J., and Smoreda, Z., Eds. *Mobile Phone Data for Development* (2013).

[46] Blondel, V. D., Esch, M., Chan, C., Clrot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., and Ziemlicki, C. Data for development: the d4d challenge on mobile phone data. *arXiv.org* (2013).

[47] Blumenstock, J., Cadamuro, G., and On, R. Predicting poverty and wealth from mobile phone metadata. *Science 350* (2015).

[48] Bohte, W., and Maat, K. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C 17* (2009), 1430–1441.

[49] Bollen, K. A. Structural equations with latent variables. In *Structural Equations with Latent Variables*. Wiley-Interscience, 1989.

[50] Bonnel, P., Smoreda, Z., Hombourger, E., and Olteanu-Raimond, A.-M. Potential of 'passive' mobile phone dataset to construct origin-destination matrix. In *4th Symposium of the European Association for Research in Transportation (HEART)* (2015).

[51] Boulos, M., Resch, B., Crowley, D., Breslin, J., Sohn, G., Burtner, R., Pike, W., Jezierski, E., and Kuo-Yu, S. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International Journal of Health Geographics 10* (2011).

[52] Bradley, M., Bowman, J. L., and Griesenbeck, B. Sacsim: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling 3* (2010).

[53] Brdar, S., Gavric, K., Culibrk, D., and Crnojevic, V. Unveiling spatial epidemiology of HIV with mobile phone data. *Scientific Reports* (2016).

[54] Breiman, L. Random forests. *Machine Learning 45* (2001), 5–32.

[55] Bricka, S., and Bhat, C. Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Rec. 1972* (2006), 9–20.

[56] Caceres, N., Wideberg, J. P., and Benitez, F. G. Deriving origin-destination data from a mobile phone network. *IET Intelligent Transport Systems 1* (2007), 15–26.

[57] Cai, Z., Wang, D., and Chen, X. M. A novel trip coverage index for transit accessibility assessment using mobile phone data. *Journal of Advanced Transportation* (2017), 1–14.

[58] Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., and Ratti, C. A system for real-time monitoring of urban mobility using cell phones: A case study in Rome. *IEEE Trans. Intelligent Transportation Systems 12* (2011), 141–151.

[59] Calabrese, F., Diao, M., Lorenzo, G., Ferreira Jr, J., and Ratti, C. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transportation Research Part C 26* (2013).

[60] Calabrese, F., Lorenzo, G. D., Liu, L., and Ratti, C. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing 10* (2011), 36–44. 2011b.

[61] Calabrese, F., Pereira, F., DiLorenzo, G., Liu, K., and Ratti, C. The geography of taste: analyzing cell-phone mobility and social events. In *Proceedings of the Int. Conf. of Pervasive Computing* (2010), pp. 22–37.

[62] Campbell, J. International health officials and Tanzania clash over potential ebola case, 10 2019.

[63] Candia, J., Gonzalez, M. C., Wang, P., Schoenharl, T., Mdaey, G., and Barabasi, A. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical 41* (2008), 1–11.

[64] Castiglione, J., Bradley, M., and Glieve, J. Activity-based travel demand models: A primer. Tech. Rep. S2-C46-RR-1, Transportation Research Board, Washington, D.C., 2015.

[65] Castri, S. D., and Gidvani, L. Enabling mobile money policies in tanzania. a test and learn' approach ton enabling market-led digital financial services. Tech. rep., GSMA, 2014.

[66] Cavoukian, A., and Castro, D. Big data and innovation, setting the record straigh: De-identification does work. Tech. rep., Information and Privacy Commissioner Ontario, Canada, 2014.

[67] Cervero, R. *America's Suburban Centers: The Land Use–Transportation Link*. Unwin Hyman, Boston, MA., 1989.

[68] Cervero, R. Mixed land-uses and commuting. evidence from the American housing survey. *Transp. Res. A 30* (1996), 361377.

[69] Cervero, R., and Kockelman, K. Travel demand and the 3ds: Density, diversity, and design. *Transportation Research Part D: Transport Environment 2* (1997), 199–219.

[70] Cervero, R., and Murakami, J. Effects of built environments on vehicle miles traveled: Evidence from 370 US urbanized areas. *Environment Planning A 42* (2010).

[71] Chaix, L., and Torre, D. The dual role of mobile payment in developing countries. Tech. Rep. WP 2015-01, GREDEG, 2015.

[72] Che-Mponda, A. H. To run a city government: The case of Dar es Salaam, Tanzania. *African Study Monographs* (1986), 71–81.

[73] Chen, C., Bian, L., and J., M. From sightings to activity locations: how well can we guess the locations visited from mobile phone sightings. *Transportation Research Part C: Emerging Technologies 46* (2014).

[74] Chen, C., Gong, H., Lawson, C., and Bialostozky, E. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. *Transportation Research Part A 44* (2010), 830–840.

[75] Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M. The promises of big data and small data for travel behaviour (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies 68* (2016).

[76] Chung, E., and Kuwahara, M. Mapping personal trip od from probe data. *International Journal of ITS Research 5* (2007).

[77] Chung, E., and Shalaby, A. A trip reconstruction tool for GPS–based personal travel surveys. *Transportation Planning Technologies 28* (2005).

[78] Colak, S., Alvim, L. A. B., Mehdiratta, S., and Gonzalez, M. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transporation Research Record: Journal of the Transportation Research Board* (2015).

[79] Couronne, T., Smoreda, Z., and Olteanu, A.-M. Chatty mobiles: Individual mobility and communication patterns. In *NetMob 2011* (2011).

[80] Couronne, T., Smoreda, Z., and Olteanu, A.-M. Urban mobility: Velocity and uncertainty in mobile phone data. In *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust* (2011), pp. 1425–1430.

[81] Cousins, K. C., and Varshney, U. The regulatory issues affecting mobile financial systems: Promises, challenges, and a research agenda. *Communications of the Association for Information Systems 34* (2014).

[82] Csaji, B., Browet, A., Traag, V., Delvenne, J., Huens, E., Dooren, P., Smoreda, Z., and Blondel, V. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications 392* (2013), 1459–1473.

[83] Dang, H.-A., Jolliffe, D., and Carletto, C. Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments. *Journal of Economic Surveys 33* (2019), 757–797.

[84] Dar Ramani Huria. `http://ramanihuria.org/`.

[85] de Abreu e Silva, J., and Goulias, K. G. Structural equations model of land use patterns, location choice, and travel behavior: Seattle, Washington, compared with Lisbon, Portugal. *Transportation Research Record 2135* (2009), 106–113.

[86] de Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., de Cordes, N., Deletaille, S., Eng-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., Krings, G., Letouz, E., Luengo-Oroz, M., Oliver, N., Rocher, L., Rutherford, A., Smoreda, Z., Steele, J., Wetter, E., Pentland, A. S., and Bengtsson, L. Comment: On the privacy conscientious use of mobile phone data. *Nature: Scientific Data 5:180286* (2018).

[87] de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., and Blondel, V. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports 3* (2013).

[88] de Montjoye, Y.-A., Rocher, L., and Pentland, A. S. Bandicoot: a Python toolbox for mobile phone metadata. *Journal of Machine Learning Research 17* (2016), 1–5.

[89] de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. D4d-senegal: The second mobile phone data for development challenge. *arXiv.org* (2014).

[90] Demirguc-Kunt, A., Klapper, L., Singer, D., and Oudheusden, P. V. The global findex database 2014: measuring financial inclusion around the world. Tech. rep., The World Bank, Washington, D.C., 2015.

[91] Demisse, M. G., Antunes, F., Bento, C., Phithakkitnukoon, S., and Sukhivbul, T. Inferring origin-destination flows using mobile phone data: A case study of Senegal. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016 13th International Conference on* (2016).

[92] Desai, A. M., and Forrest, E. E-marketing in developed and developing countries: Emerging practices. In *E-Marketing in Developed and Developing Countries: Emerging Practices*, H. El-Gohary and R. Eid, Eds. Business Science Reference, Hershey, PA, USA, 2013, ch. Mobile Marketing: The Imminent Predominance of the Smartphone, pp. 97–115.

[93] Devarajan, S. Africa's statistical tragedy. *The Review of Income and Wealth 59* (2013).

[94] di. Lorenzo, G., Sbodio, M., Calabrese, F., Berlingerio, M., Pinelli, F., and Nair, R. AllAboard: Visual exploration of cellphone mobility data to optimise public transport. *IEEE Trans. Vis. Comput. Graph. 22* (2014).

[95] Diao, M., Zhu, Y., Ferreira Jr, J., and Ratti, C. Inferring individual daily activities from mobile phone traces: A boston example. IRES Working Paper Series IRES2015-012, IRES, 2015.

[96] Dieleman, F. M., Dijst, M., and Burghouwt, G. Urban form and travel behaviour: micro-level household attributes and residential context. *Urban Studies 39* (2002), 507527.

[97] Ding, K. The disintegration of production: Firm strategy and industrial development in China. In *The Disintegration of Production: Firm Strategy and Industrial Development in China*, M. Watanabe, Ed. Edward Alger Publishing Limited, Cheltenham, UK, 2014, ch. The specialized market system: the market exploration of small businesses, pp. 149–176.

[98] Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., and Zhou, X. Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C 58* (2015), 278–291.

[99] Doyle, J., Hung, P., Kelly, D., McLoone, S., and Farrell, R. Utilising mobile phone billing records for travel mode discovery. In *ISSC* (Trinity College Dublin, Dublin, 2011).

[100] Duncombe, R., and Boateng, R. Mobile phones and financial services in developing countries: A review of concepts, methods, issues, evidence and future research directions. *Third World Quarterly 30* (2009).

[101] Dzogbenuku, R. K. Banking innovation in Ghana: Insight of students' adoption and diffusion. *Journal of Internet Banking and Commerce 18* (2013), 1–20.

[102] Eagle, L., and Bruin, A. D. Advertising restrictions: protection of the young and vulnerable? *International Journal of Advertising and Marketing to Children 2* (2001), 259–271.

[103] Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science 328* (2010).

[104] Eagle, N., and Pentland, A. S. Reality mining: sensing complex social systems. *Journal Personal and Ubiquitous Computing 10* (2006), 255–268.

[105] Ebener, S., Murray, C., Tandon, A., and Elvidge, C. C. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. *International Journal of Health Geographics 4* (2005).

[106] Eboli, L., Forciniti, C., and Mazzulla, G. Exploring land use and transport interaction through structural equation modelling. *Procedia - Social and Behavioral Sciences 54* (2012), 107–116.

[107] Economides, N., and Jeziorski, P. Mobile money in Tanzania. *Marketing Science 36* (2017).

[108] Ehmke, J., Meisel, S., and Mattfeld, D. C. Floating car data based analysis of urban travel times for the provision of traffic quality. In *Traffic Data Collection and its Standardization*, J. Barcelo and M. Kuwahara, Eds., vol. 144 of *International Series in operations Research & Management Science*. Springer Science+Business Media, New York, NY, 2010, pp. 129–149.

[109] electronics notes. Gsm handover, 2017.

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

[111] Elwood, S., and Leszczynski, A. Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum 42* (2011), 6–15.

[112] Engelmann, G. Ethics and privacy implications of using CDR data for social good, 2016.

[113] Engelmann, G., Goulding, J., and Golightly, D. Estimating activity-based land-use through unsupervised learning from mobile phone event series in emerging economies. In *GISRUK* (2017).

[114] Engelmann, G., Goulding, J., Smith, G., and Golightly, D. Estimating population behaviour to describe activity-based land-use in emerging economies using mobile phone event series. In *NetMob* (2017), pp. 146–148.

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

[116] ESRI. Esri shapefile technical description. Tech. rep., ESRI, 1998.

[117] Etminani-Ghasrodashti, R., and Ardeshiri, M. The impacts of built environment on home-based work and non-work trips: an empirical study from Iran. *Transportation Research Part A: Policy and Practice 85* (2016), 196–207.

[118] Evans, D. S., and Pirchio, A. An empirical examination of why mobile money schemes ignite in some developing countries but flounder in most. *Review of Network Economics 3* (2015), 397–451.

[119] Ewing, R., and Cervero, R. Travel and the built environment. a meta-analysis. *Journal of the American Planning Association 76* (2010), 265–294.

[120] F. Stuart Chapin, J. *Human Activity Patterns in the City.* Wiley-Interscience, 1974.

[121] Fengler, W. Big data and development: the second half of the chess board", 2013.

[122] Fillekes, M. *Reconstructing Trajectories from Sparse Call Detail Records.* PhD thesis, University of Tartu, Tartu, Estonia, 2014.

[123] Forrester, J. *Urban dynamics.* MIT Press, Cambridge, MA, 1969.

[124] Frank, L. D., and Pivo, G. Impacts of mixed use and density on utilization of three modes of travel: Single-occupant vehicles, transit and walking. *Issues in land use and transportation planning, models, and applications* (1994), 44–52.

[125] Friaz-Martinez, V., Soguero, C., and Frias-Martinez, E. Estimation of urban commuting patterns using cellphone network data. In *ACM SIGKDD International Workshop on Urban Computing* (Beijing, PRC, 2012).

[126] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[127] Friedman, T. L. *The World Is Flat*. Penguin, London, 2006.

[128] Friedrich, M., Immisch, K., Jehlicka, P., Otterstttter, T., and Schlaich, J. Generating OD matrices from mobile phone trajectories. In *Transportation Research Board 89th Annual Meeting* (2010).

[129] Friedrich, M., Jehlicka, P., and Schlaich, J. Automatic number plate recognition for the observance of travel behavior. In *CD-OM, Proceedings of the 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability* (2008).

[130] Fras-Martnez, V., Soto, V., Hohwald, H., and Fras-Martnez, E. Characterizing urban landscapes using geolocated tweets. In *SOCIALCOM-PASSAT '12 Proc. 2012 ASE/IEEE Int. Conf. on Social Computing and 2012 ASE/IEEE Int. Conf. on Privacy, Security, Risk and Trust* (2012), pp. 239–248.

[131] Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., and Smoreda, Z. A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Transactions on Mobile Computing 16* (2017), 2682–2696.

[132] Garca-Palomares, J., Salas-Olmedo, M., Moya-Gmez, B., Condeo-Melhorado, A., and Gutirrez, J. City dynamics through twitter: Relationships between land use and spatiotemporal demographics. *Cities 72* (2018), 310–319.

[133] Garg, S., Singh, P., Ramanathan, P., and Sen, R. Vividhavahana: Smartphone based vehicle classification and its applications in developing region. In *Pro-*

ceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (2014).

[134] Gautam, B. P., and Shrestha, D. Document clustering through non-negative matrix factorization: a case of hadoop for computational time reduction of large scale documents. In *International MultiConference of Engineers and Computer Scientists IMECS* (2010), vol. 1, pp. 1–10.

[135] Georgy, M., and Anna, I. semopy: A Python package for structural equation modeling. *arXiv.org* (2019).

[136] Gilman, L. The impact of mobile money interoperability in Tanzania. Tech. rep., GSMA, September 2016.

[137] Giugale, M. Fix Africa's statistics, 2012.

[138] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics 24*, 1 (2015), 44–65.

[139] Golob, T. F. Structural Equation Modeling For Travel Behavior Research. University of California Transportation Center, Working Papers qt2pn5j58n, University of California Transportation Center, Dec. 2011.

[140] Gong, H., Chen, C., Bialostozky, E., and C, L. A GPS/GIS method for travel mode detection in new york city. *Computers, Environment and Urban Systems 36* (2011).

[141] Gong, L., Liu, X., Wu, L., and Y, L. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science 43* (2015).

[142] Gonzalez, M., Hidalgo, C., and Barabasi, A. L. Understanding individual human mobility patterns. *Nature 453* (2008), 779–782.

[143] Goodchild, M. Citizens as sensors: The world of volunteered geography. *GeoJournal 69* (2007).

[144] Goodchild, M., and Glennon, J. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth 3* (2010), 231–241.

[145] Goulding, J., Engelmann, G., Iliffe, M., and Smith, G. Best practices and methodology for od matrix creation from cdr data. Tech. rep., N/Lab, 2014.

[146] Goulias, K. G., Bhat, C. R., Pendyala, R. M., Chen, Y., Paleti, R., Konduri, K. C., Lei, T., Yoon, S. Y., Huang, G., and Hu, H.-H. Simulator of activities, greenhouse emissions, networks, and travel (simagent) in Southern California. In *Transportation Research Board 91st Annual Meeting* (2012).

[147] Groves, R. Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Quart. 70* (2006), 646–675.

[148] GSMA. Gauging the relationship between fixed and mobile penetration. Tech. rep., GSMA, 2014.

[149] GSMA. GSMA guidelines on the protection of privacy in the use of mobile phone data for responding to the ebola outbreak, 2014.

[150] GSMA. State of the industry: Mobile financial services for the unbanked. Tech. rep., GSMA, 2014.

[151] GSMA. The impact of mobile money interoperability in Tanzania. Tech. rep., GSMA, 2016.

[152] Gundlegard, D., Rydergren, C., Breyer, N., and Rajna, B. Travel demand estimation and network assignment based on cellular network data. *Computer Communications* (2016).

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

[154] Gutierrez, T., Krings, G., and Blondel, V. D. Indicators of wealth, economic diversity and segregation in Côte d'Ivoire using mobile phone datasets. In *Netmob* (2013).

[155] Gwahula, R. Risks and barriers associated with mobile money transactions in tanzania. *Business Management and Strategy 7* (2016).

[156] Gwilliam, K. Africa's transport infrastructure. Tech. rep., World Bank, 2011.

[157] H"agerstraand, T. What about people in regional science? *Pap. Reg. Sci. 24* (1970), 7–24.

[158] Hao, J., Hatzopoulou, M., and Miller, E. Integrating an activity-based travel demand model with dynamic traffic assignment and emission models. *Transp. Res. Record J. Transp. Res. Board 2176* (2010).

[159] Hardt, M. How big data is unfair, 2014.

[160] Hariharan, R., and Toyama, K. Project Lachesis: Parsing and modeling location histories. In *Proc. 3rd Int. Conf. Geographic Inf. Science* (2004).

[161] Hemminki, S., Nurmi, P., and Tarkoma, S. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems* (2013).

[162] Henderson, J. V., Storeygard, A., and Weil, D. N. Measuring economic growth from outer space. Tech. Rep. No. 15199, The National Bureau of Economic Research, 2009.

[163] Herrera, J., Amin, S., Bayen, A., Madanat, S., Zhang, M., Nie, Y., Qian, Z., Lou, Y., Yin, Y., and Li, M. Dynamic estimation of OD matrices

for freeways and arterials. Tech. rep., Institute for Transportation Studies, UC Berkeley, 2007.

[164] Herrera, J. C., Work, D., Herring, R., Ban, X., Jacobson, Q., and Bayen, A. Evaluation of traffic data obtained via gps-enabled mobile phones: the mobile century field experiment. *Transportation Research Part C: Emerging Technologies. 18* (2010), 568–583.

[165] Hily, D. N. How dar traffic jams cost 411bn annually, 2013.

[166] Hinson, R. E. Banking the poor: The role of mobiles. *Journal of Financial Services Marketing 15* (2011).

[167] Hood, J., Sall, E., and Charlton, B. A gps-based bicycle route choice model for san francisco. *Calif. Transp. Lett. 3* (2011), 63–75.

[168] Hooper, D., Coughlan, J., and Mullen, M. R. Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods 6* (2008), 53–60.

[169] Horn, C., Klampfl, S., Cik, M., and Reiter, T. Detecting outliers in cell phone data: Correcting trajectories to improve traffic modeling. In *2014 TRB Annual Meeting Compendium of Papers* (2014).

[170] Horni, A., Nagel, K., and Axhausen, K. W. *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press, 2016.

[171] Hosmer, D. W., and Lemeshow, S. Introduction to the logistic regression model. In *Applied Logistic Regression*, second edition ed. 2000, p. 130.

[172] Hu, S., and Wang, L. Automated urban land-use classification with remote sensing. *International Journal of Remote Sensing* (2013), 790–803.

[173] Iliffe, M. *The praxis of community mapping in developing countries*. PhD thesis, University of Nottingham, 2017.

[174] International Monetary Fund. Regional economic outlook: Sub–Saharan Africa. Tech. rep., International Monetary Fund, 2016.

[175] Iovan, C., Olteanu-Raimond, A., Couronne, T., and Smoreda, Z. Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In *Geographic Information Science at the Heart of Europe. Lecture Notes in Geoinformation and Cartography*, D. Vandenbroucke, B. Bucher, and J. Compvoets, Eds. Springer, Cham, 2013, pp. 247–265.

[176] Iqbal, M., Choudhury, C., Wang, P., and Gonzalez, M. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies 40* (2014).

[177] Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. Identifying important places in people's lives from cellular network data. In *9th International Conference, Pervasive 2011, San Francisco, USA* (2011).

[178] Jabareen, Y. R. Sustainable urban forms: Their typologies, models, and concepts. *Journal of Planning Education and Research 26* (2006), 38–52.

[179] Jack, W., Ray, A., and Suri, T. Transaction networks: Evidence from mobile money in Kenya. *American Economic Review 103* (2013).

[180] Jack, W., and Suri, T. Mobile money: The economics of M-Pesa. Tech. rep., National Bureau of Economic Research, 2011.

[181] Jacobs-Crisioni, C., Rietveld, P., Koomen, E., and Tranos, E. Evaluating the impact of land-use density and mix on spatiotemporal urban activity patterns: an exploratory study using mobile phone data. *Environment and Planning A 46* (2014), 27692785.

[182] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*. Springer, 2013.

[183] James, J., and Versteeg, M. Mobile phones in Africa: how much do we really know? *Soc Indic Res 84* (2007), 117–126.

[184] Jan, O., Horowitz, A. J., and Peng, Z. Using global positioning system data to understand variations in path choice. *Transp. Res. Rec. 1725* (2000), 37–44.

[185] Jiang, B., Fiore, G., Yang, Y., Ferreira, J., Frazzoli, E., and Gonzalez, M. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp* (Chicago, IL, 2013).

[186] Jiang, S., Jr, J. F., and Gonzalez, M. C. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* (2015), 1–12.

[187] Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., and Gonzlez, M. C. The timeGeo modeling framework for urban mobility without travel surveys. *PNAS 113* (2016), E5370–E5378.

[188] Joereskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika 34* (1969), 183–202.

[189] Johnson, S. Competing visions of financial inclusion in Kenya: The rift revealed by mobile money transfer. *Canadian Journal of Development Studies 37* (2016).

[190] Johnson, S., and Arnold, S. Inclusive financial markets: Is transformation under way in Kenya? *Development Policy Review 30* (2012).

[191] Jones, J. H. M., Williams, M. J., and Joshi, M. P. Domestic migration and remittances in India: Rajasthani tribal migrants working in Gujarat. *Enterprise Development and Microfinance 25* (2014).

[192] Kadui, E., Bojovi, P., and galj, A. Consumer adoptionrisk factor of mobile banking services. *World Academy of Science, Engineering and Technology 80* (2011).

[193] Kandt, J. Heterogeneous links between urban form and mobility: A comparison of Sao Paulo, Istanbul and Mumbai. *The Journal of Transport and Land Use 11* (2018), 721–745.

[194] Kanobe, F., Alexanderand, P. M., and Bwalya, K. J. Policies, regulations and procedures and their effects on mobile money systems in Uganda. *The Electronic Journal of Information Systems in Developing Countries 83* (2017).

[195] Kikulwe, E. M., Fischer, E., and Qaim, M. Mobile money, smallholder farmers, and household welfare in Kenya. *PLoS One 9* (2014).

[196] Kim, M., Zoo, H., Lee, H., and Kang, J. Mobile financial services, financial inclusion, and development: A systematic review of academic literature. *The Electronic Journal of Information Systems in Developing Countries 84* (2018).

[197] Kitamura, R., Chen, C., and Pendyala, R. Generation of synthetic activity-travel patterns. *Transportation Research Record 1607* (1997).

[198] Kitamura, R., Mokhtarian, P. L., and Laidet, L. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation 24* (1997), 125–158.

[199] Klinger, T., and Lanzendorf, M. Moving between mobility cultures: what affects the travel behavior of new residents? *Transportation 43* (2016), 243–271.

[200] Knoop, L., and Eames, T. Public transport technology in the United Kingdom: Annual survey 2008. Tech. rep., RTIG Ltd, 2009.

[201] Kockelman, K. Travel behavior as function of accessibility, land use mixing, and land use balance: Evidence from san francisco bay area. *Transportation Research Record: Journal of the Transportation Research Board 1607* (1997), 116125.

[202] Kouakou, E. Public transport in Sub-Saharan Africa: Major trends and case studies. Tech. rep., UITP: International Association of Public Transport, 2010.

[203] KPMG. Monetizing mobile: How banks are preserving their place in the payment value chain, 2012, March 5.

[204] Krambeck, H. The general transit feed specification (GTFS) and implications for international development. Transforming Transportation, 2015.

[205] Krizek, K. J. Residential relocation and changes in urban travel. does neighborhood-scale urban form matter? *J. Am. Plan. Assoc 69* (2003), 265–281.

[206] Kulkarni, R., Haynes, K. E., Stough, R. R., and Riggle, J. D. Light based growth indicator (LBGI): exploratory analysis of developing a proxy for local economic growth based on night lights. *Regional Science Policy & Practice 3* (2011), 101–113.

[207] Kumar, A., and Barrett, F. Stuck in traffic: Urban transport in Africa. Tech. rep., World Bank, Washington, D.C., 2008.

[208] Kursa, M. B., and Rudnicki, W. R. Feature selection with the Boruta package. *Journal of Statistical Software 36* (2010).

[209] Larijani, A. N., Oletanu-Raimond, A.-M., Perret, J., Bredif, M., and Ziemlicki, C. Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. *Transportation Research Procedia 6* (2015), 64–78.

[210] Lee, D. D., and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature 401* (1999), 788–791.

[211] Lee, J., and Hou, J. Modeling steady-state and transient behaviours of user mobility: formulation, analysis, and application. In *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing* (2006).

[212] Lee, R. J., Seneer, I. N., and III, J. A. M. An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. *Transportation Letters 8* (2016).

[213] Lenormand, M., Picornell, M., Cant-Ros, O. G., Louail, T., Herranz, R., Barthelemy, M., Fras-Martnez, E., Miguel, M. S., and Ramasco, J. J. Comparing and modelling land use organization in cities. *Royal Society Open Science* (2015).

[214] Letouze, E., Vinck, P., and Kammourieh, L. The law, politics and ethics of cell phone data analytics. Tech. rep., Data-Pop Alliance, 2015.

[215] Letouze, E. F. *Applications and Implications of Big Data for Demo-Economic Analysis: The Case of Call-Detail Records*. PhD thesis, UC Berkeley, 2016.

[216] Li, H., Guensler, R., and Ogle, J. Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data. *Transp. Res. Rec. 1926* (2005), 162–170.

[217] Lima, A., Pejovic, V., Rossi, L., Musolesi, M., and Gonzalez, M. Progmosis: Evaluating risky individual behavior during epidemics using mobile network data. *arXiv.org* (2015).

[218] Lin, D., Allan, A., and Cui, J. Sub-centres, socio-economic characteristics and commuting: a case study and its implications. *International Journal of Urban Sciences 21* (2017), 147–171.

[219] Lin, J.-J., and Yang, A.-T. Structural analysis of how urban form impacts travel demand: Evidence from Taipei. *UrbanStudies 46* (2009), 1951–1967.

[220] Liu, Y., Kang, C., and Wang, F. Towards big data-driven human mobility patterns and models. *Geomat. Inform. Sci. Wuhan University 39* (2014).

[221] Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., and Shi, L. Social sensing: a new approach to understanding our socio-economic environments. *Annals of of the Association of American Geographers 105* (2015).

[222] Louail, T., Lenormand, M., Cantu-Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., and Bathelemy, M. From mobile phone data to the spatial structure of cities. *Scientific Reports 4* (2015).

[223] Lucas, K., and Porter, G. Mobilities and livelihoods in urban development contexts: Introduction. *Journal of Transport Geography* (2016), 129131.

[224] Lyons, A., and Scherpf, E. Moving from unbanked to banked : evidence from the money start program. *Financial Services Review 13* (2004), 215–231.

[225] Ma, J., Yuan, F., Joshi, C., Li, H., and Bauer, T. A new framework for development of time-varying O-D matrices based on cellular phone data. In *4th TRB Innovations in Travel Modelin (ITM) Conference* (Tampa, FL, 2012).

[226] Madhawa, K., Lokanathan, S., Maldeniya, D., and Samarajiva, R. Using mobile network big data for land use classification. Tech. rep., LIRNEasia, 2015.

[227] Majeed, R. Disseminating the power of information: Kenya open data initiative, 2011-2012. Tech. rep., Innovations for Successful Societies, 2012.

[228] Maldeniya, D., Lokanathan, S., and Kuramage, A. Origin-destination matrix estimation for Sri Lanka using mobile network big data. In *Proceedings of the 13th Internationational Conference on Social Implications of Computers in Developing Countries* (2015).

[229] Mallat, N. Exploring consumer adoption of mobile payments-a qualitative study. *The Journal of Strategic Information Systems 16* (2007).

[230] Mallikarjuna, C., Phanindra, A., and Ramachandra, K. Traffic data collection under mixed traffic conditions using video image processing. *Journal of Transportation Engineering 135* (2009), 174182.

[231] Mamei, M., and Ferrari, L. Daily commuting in Ivory Coast: Development opportunities. In *D4D Challenge @ 3rd Conf. on the Analysis of Mobile Phone datasets*. NetMob, Milan, Italy, 2013.

[232] Manley, E., and Dennett, A. New forms of data for understanding urban activity in developing countries. *Appl. Spatial Analysis 12* (2018), 45–70.

[233] Mao, H., Ahn, Y.-Y., Bhaduri, B., and Thakur, G. Improving land use inference by factorizing mobile phone call activity matrix. *Journal of land use science 12* (2017), 138–153.

[234] Mao, H., Thakur, G., and Bhaduri, B. Exploiting mobile phone data for multi-category land use classification in Africa. In *UrbanGIS 16* (2016).

[235] Maurer, B., Nelms, T. C., and Rea, S. C. bridges to cash': Channelling agency in mobile money. *Journal of the Royal Anthropological Institute 19* (2013).

[236] Mbogo, M. The impact of mobile payments on the success and growth of microbusiness: The case of M–Pesa in Kenya. *Journal of Language, Technology & Entrepreneurship in Africa 2* (2010), 182–203.

[237] Melamed, C. Development data: how accurate are the figures?, 2014, January.

[238] Mellegard, E., Moritz, S., and Zahoor, M. Origin/destination-estimation using cellular network data. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (2011), pp. 891–896.

[239] Merritt, C. Mobile money transfer services: The next phase in the evolution of person-to-person payments. *Journal of Payments Strategy & Systems 5* (2011), 143–160.

[240] Mesev, V. The use of census data in urban image classification. *Photogrammetric Engineering and Remote Sensing 64* (1998), 431436.

[241] Meurs, H., and Haijer, R. Spatial structure and mobility. *Transportation Research D 6* (2001), 429–446.

[242] Miller, E., Hunt, J., and Abraham, J.E. Salvini, P. Microsimulating urban systems. *Computers, Environment and Urban Systems 53* (2004).

[243] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs: Simple building blocks of complex networks. *Science 298* (2002).

[244] Minto-Coy, I., and McNaughton, M. Barriers to entrepreneurship and innovation: An institutional analysis of mobile banking in Jamaica and Kenya. *Social and Economic Studies 65* (2016).

[245] Mirzoyants, A. Mobile money in tanzania: Use, barriers, and opportunities. the financial inclusion tracker survey project. Tech. rep., InterMedia, 2013.

[246] Mishra, V., and Bisht, S. Mobile banking in a developing economy: a customer-centric model for policy formulation. *Telecommunications Policy 37* (2013).

[247] Morawczynski, O., and Pickens, M. Poor people using mobile financial services: Observations on customer usage and impact from M-Pesa. Tech. rep., World Bank Group, Washington, D.C., 2009.

[248] Moro, E., de Montjoye, A. Y., Blondel, V., and Pentland, A. S. *Data for Development Challenge Senegal*. 2015.

[249] Mortier, R., Haddadi, H., Henderson, T., McAuley, D., and Crowcroft, J. Human-Data Interaction: The Human Face of the Data-Driven Society. *Social Science Research Network* (Oct. 2014), 1–14.

[250] Mpogole, H., Tweve, Y., Mwakatobe, N., and Serijo Mlasu, D. S. Towards non-cash payments in Tanzania: the role of mobile phone money services. In *IST-Africa 2016 Conference* (2016).

[251] Muijs, D. One step beyond: Introduction to multilevel modelling and structural equation modelling. In *Doing Quantitative Research in Education with SPSS*. Sage, 2011, ch. 12, pp. 225–240.

[252] Must, B., and Ludewig, K. Mobile money: cell phone banking in developing countries. *Policy Matters Journal* (2010), 27–33.

[253] N. Caceres, J. P. Wideberg, F. G. B. Review of traffic data estimations extracted from cellular networks. *IET Intelligent Transport Systems 2* (2008), 179–192.

[254] Naranjo, J. E., Jimnez, F. J. S., and Zato, J. G. Comparison between floating car data and infrastructure sensors for traffic speed estimation. In *IEEE ITSC2010 Workshop on Emergent Cooperative Technologies in Intelligent Transportation Systems* (Madeira Island, Portugal, 2010).

[255] Narayanan, A., and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy* (2008), pp. 111–125.

[256] National Bureau of Statistics. United republic of tanzania national accounts statistics. Tech. rep., National Bureau of Statistics, June 2019.

[257] NetHope, and MEDA. Tanzania mobile money assessment and case study: Examining cash payment streams and their electronic alternatives among usaid implementing partners. Tech. rep., USAID, 2013.

[258] Ni, L., Wang, X. C., and Chen, X. M. A spatial econometric model for travel flow analysis and real world applications with massive mobile phone data. *Transportation Research Part C 86* (2018), 510–526.

[259] Nitsche, P., W. P. B. S. B. . N. M. P. Supporting large-scale travel surveys with smartphones–a practical approach. *Transp. Res. C. 43* (2013), 212–221.

[260] Noulas, A., Mascolo, C., and Frias-Martinez, E. Exploiting foursquare and cellular data to infer user activity in urban environments. In *2013 ieee 14th international conference on mobile data management* (2013), pp. 167–176.

[261] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011).

[262] O2. Telefnica, the Open Data Institute and the MIT set data challenge for campus party 2013.

[263] O'Flaherty, C. A. *Highways Vol. 1 Highways and Traffic. 2nd Edition.* Edward Arnold Publishers Ltd., London, 1974.

[264] Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review 57* (2010), 1701–1777.

[265] Okwi, P. O., Ndeng'e, G., Kristjanson, P., Arunga, M., Notenbaert, A., Omolo, A., Henninger, N., Benson, T., Kariuki, P., and Owuor, J. Spatial determinants of poverty in rural Kenya. *PNAS 104* (2007).

[266] Olivia, S., Gibson, J. K., Brabyn, L. K., and Stichbury, G. A. Monitoring economic activity in indonesia using night light detected from space. In *The 12th Indonesian Regional Science Association Conference* (Makassar, Indonesia, 2014).

[267] Olmos, L. E., Tadeo, M. S., Vlachogiannis, D., Alhasoun, F., Alegre, X. E., Ochoa, C., Targa, F., and Gonzlez, M. C. A data science framework for planning the growth of bicycle infrastructures. *Transportation Research Part C: Emerging Technologies 115* (2020).

[268] Openshaw, S. *The Modifiable Areal Unit Problem.* Geo Books, Norwich, UK, 1984.

[269] Ortuzar, J. D. D., and Willumsen, L. G. *Modelling Transport*, 4th ed. John Wiley & Sons Ltd., 2011.

[270] Osoba, O., and IV, W. W. An intelligence in our image. Tech. rep., RAND Corporation, 2017.

[271] Pas, E. State of the art and research opportunities in travel demand: another perspective. *Transportation Research Part A* (1985).

[272] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.,

Vanderplas, J., Passos, A., and Cournapeau, D. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[273] Pei, T., Sobolevsky, S., Ratti, C., Shaw, S., and Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science* (2014).

[274] Peruta, M. D. Adoption of mobile money and financial inclusion: a macroeconomic approach through cluster analysis. *Economics of Innovation and Technology 27* (2018).

[275] Pew Research Center. Cell phones in Africa: Communication lifeline, 2015.

[276] Phithakkitnukoon, S., Horanont, T., Lorenzo, G., Shibasaki, R., and Ratti, C. Activity-aware map: Identifying human daily activity pattern using mobile phone data. Tech. rep., MIT-Senseable City Lab, Boston, MA, 2010.

[277] Piggott, D. *Inferring transportation mode using smartphone sensor data.* PhD thesis, Fitzwilliam College, 2011.

[278] Pinjari, A. R., and Bhat, C. R. Activity-based travel demand analysis. *A Handbook Transp. Econ. 1* (2011).

[279] Pokhriyal, N., and Dong, W. Virtual network and poverty analysis in Senegal. In *NetMob Book of Abstracts* (MIT Media Labs, Boston, 2015).

[280] Pokhriyal, N., and Jacques, D. C. Combining disparate data sources for improved poverty prediction and mapping. *PNAS 114* (2017).

[281] Poku-Boansi, M., and Cobbinah, P. B. Land use and urban travel in Kumasi, Ghana. *GeoJournal 83* (2018), 563–581.

[282] Pronello, C., and Camusso, C. Travellers' profiles definition using statistical multivariate analysis of attitudinal variables. *J. Transp. Geography 19* (2011), 1294–1308.

[283] Pu, W., Lin, J., and Lon, L. Urban travel time estimation using real time bus tracking data. Tech. rep., Transport Chicago, Chicago, IL, 2008.

[284] Pu, W., Lin, J., and Lon, L. Real-time estimation of urban street segment travel time using buses as speed probe. *Transportation Research Record 2129* (2009), 81–89.

[285] Puello, L. L. P., Olde-Kalter, M.-J., and Geurs, K. T. Measurement of non-random attrition effects on mobility rates using trip diaries data. *Transportation Research, Part A 106* (2017), 51–64.

[286] Qiang, C., Kuek, S., Dymond, A., and Esselaar, S. Mobile applications for agricultural and rural development. Tech. rep., ICT Sector Unit, World Bank, 2011.

[287] Quddus, M.A., O.-W. Z. L. N. R. A general map matching algorithm for transport telematics applications. *GPS Solutions 7* (2003), 157–167.

[288] Qui, Q., Ward, M. O., Rundensteiner, E., and Yang, J. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics 12* (2006), 709–716.

[289] Rahman, M. A., Qi, X., and Islam, M. T. Banking access for the poor: Adoption and strategies in rural areas of Bangladesh. *Journal of Economic and Financial Studies 4* (2016), 1–10.

[290] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. Searching and mining trillions of time series subsequences under dynamic time warping. In *KDD International Conference on Knowledge Discovery & Data Mining* (2012).

[291] Ranjan, G., Zang, H., Zhang, Z., and Bolot, J. Are call detail records biased for sampling human mobility. *Mobile Comput Commun Rev 16* (2012).

[292] Rasouli, S., and Timmermans, H. Activity-based models of travel demand: Promises, progress and prospects. *International Journal of Urban Sciences 18* (2014).

[293] Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. Mobile landscapes: Using location data from cellphones for urban analysis. *Environment and Planning B: Planning and Design 33* (2006), 727–748.

[294] Raykov, T., and Marcoulides, G. A. *A First Course in Structural Equation Modeling*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[295] Reades, J., Calabrese, F., and Ratti, C. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design 36* (2009).

[296] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. Using mobile phones to determine transportation modes. *ACM Trans. Sens. Netw. 6* (2010).

[297] Reuters. Attack force locals to shut down dozens of cell phone towers in Indian Kashmir, 2016.

[298] Rodrigue, J. P., Comtois, C., and Slack, B. *The Geography of Transport Systems*. Routledge Taylor and Francis Group, London and New York, 2006.

[299] Rose, G. Mobile phones as traffic probes: Practices, prospects and issues. *Transport Reviews 26* (2008), 275–291.

[300] Rosen, J. W. This Tanzanian city may soon be one of the world's most populous. is it ready?, 4 2019.

[301] Rudnicki, W. R., Wrzesień, M., and Paja, W. All relevant feature selection methods and applications. In *Feature Selection for Data and Pattern Recognition*, U. Stańczyk and L. C. Jain, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pp. 11–28.

[302] Sanz, F. P., and Lima, P. D. The uptake of mobile financial services in the Middle East and North Africa region. *Enterprise Development and Microfinance 24* (2013).

[303] Scheiner, J., and Holz-Rau, C. Travel mode choice affected by objective or subjective determinants? *Transportation, 34* (2007), 487–511.

[304] Scheiner, J., and Holz-Rau, C. Gendered travel mode choice: a focus on car deficient households. *J. Transp. Geography 24* (2012), 250–261.

[305] Schlaich, J., Otterstatter, T., and Friedrich, M. Generating trajectories from traces. In *89th Annual Meeting of the Transportation Research Board* (Washington, DC, 2010).

[306] Schneider, C. M., Belik, V., Couronn, T., Smoreda, Z., and Gonzlez, M. C. Unravelling daily human mobility motifs. *J. R. Soc. Interface 10* (2013).

[307] Schwegmann, C. Open data in developing countries. Tech. Rep. 2013/02, European Public Sector Information Platform, 2013.

[308] Shaikh, A. A., and Karjaluoto, H. Mobile banking adoption: A literature review. *Telematics and Informatics 32* (2015), 129–142.

[309] Shalita, S. World Bank statement on amendments to Tanzania's 2015 Statistics Act, 2018.

[310] Sihvonen, M. Ubiquitous financial services for developing countries. *The Electronic Journal of Information Systems in Developing Countries 28* (2006), 1–11.

[311] Simma, A., and Axhausen, K. W. Interactions between travel behaviour, accessibility and personal characteristics: The case of the Upper Austria region. *European Journal of Transport and Infrastructure Research 3* (2003), 179–197.

[312] Sinha, A., Saini, T., and Srikanth, S. V. Distributed computing approach to optimize road traffic simulation. In *2014 International Conference on Parallel, Distributed and Grid Computing* (2014), pp. 360–364.

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

[314] Song, C., Qu, Z., Blumn, N., and Barabasi, A. L. Limits of predictability in human mobility. *Science 327* (2010), 1018–1021.

[315] Soto, V., and Frias-Martinez, E. Automated land use identification using cell-phone records. In *HotPlanet '11 Proceedings of the 3rd ACM International Workshop on MobiArch* (2011), ACM, pp. 17–22.

[316] Soto, V., and Frias-Martinez, E. Robust land use characterization of urban landscapes using cell phone data. In *Proceedings of the 1st workshop on pervasive urban applications, in conjunction with the 9th international conference on pervasive computing* (San Francisco, CA, 2011).

[317] Soto, V., Frias-Martinez, V., Virseda, J., and Frias-Martinez, E. Prediction of socioeconomic levels using cell phone records. In *19th Int. Conference on User Modeling, Adaption and Personalization* (Gerona, Spain, 2011).

[318] Spinney, J. E. Mobile positioning and LBS applications. *Geography 88* (2003), 256–265.

[319] Stammers, N., and Eschle, C. Socials movements and global activism. In *Global activism global media*. Pluto Press, London, UK, 2005, pp. 50–67.

[320] Stead, D. Relationship between land use, socioeconomic factors and travel patterns in Britain. *Environment and Planning B 24* (2001), 499–529.

[321] Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., and Scholten, H. Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities. *GeoJournal 78* (2013), 223–243.

[322] Steenbruggen, J., Tranos, E., and Nijkamp, P. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy 39* (2015), 335346.

[323] Stopher, P. R., Jiang, Q., and FitzGerald, C. Processing GPS data from travel surveys. In *2nd International Colloqium on the Behavioural Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications* (Toronto, Canada, 2005).

[324] Stoppiglia, H., Dreyfus, G., Dubois, R., and Oussar, Y. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research 3* (2003), 1399–1414.

[325] Stork, D. G., Hart, P. E., and Duda, R. O. *Pattern Classification*. Wiley, New York, MA, 2001.

[326] Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10* (2002), 557–570.

[327] Tatem, A. J., Qiu, Y., Smith, D. L., Sabot, O., Ali, A. S., and Moonen, B. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium Falciparum rates among Zanzibar residents. *Malaria Journal 8*, 1 (2009), 287–12.

[328] TCRA. Quarterly communications statistics report: January-March 2016 quarter. Tech. rep., TCRA, 2016.

[329] Telecom Italia. Enter the big data challenge 2015, 2015.

[330] Terroso-Saenz, F., and Munoz, A. Land use discovery based on volunteer geographic information classification. *Expert Systems with Applications 140* (2020).

[331] Texas A&M Transportation Institute. Emerging data collection techniques for travel demand modeling: A literature review. Tech. rep., Texas A&M, 2014.

[332] Tizzoni, M., Bajardi, P., Decuyper, A., King, G. K. K., Schneider, C. M., Blondel, V., Smoreda, Z., Gonzlez, M. C., and Colizza, V. On the use of human mobility proxies for modeling epidemics. *PLOS Computational Biology 10* (2014).

[333] Tolouei, R., Psarras, S., and Prince, R. Origin-destination matrix development: Conventional methods versus mobile phone data. *Transport Research Procedia 26* (2017), 39–52.

[334] Ton, T. T., and Hensher, D. A. A spatial and statistical approach for imputing origin-destination matrices from household travel survey data: A Sydney case study. Tech. rep., Institute of Transportation Studies, Sydney, Australia, 2002.

[335] Toole, J., Colak, S., Sturt, B., Alexander, L., Evsukoff, A., and Gonzalez, M. The path most traveled: travel demand estimation using big data resources. *Transportation Research Part C 58* (2015).

[336] Toole, J., Ulm, M., Gonzalez, M. C., and Bauer, D. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (2012).

[337] Tsui, S., and Shalaby, A. S. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transp. Res. Rec. 1972* (2006), 483–498.

[338] UN General Assembly. Transforming our world: the 2030 agenda for sustainable development. adopted by the General Assembly in its Seventieth session.

[339] United Nations. Spatial planning: Key instruments for development and effective governance with special reference to countries in transition. Tech. rep., United Nations Economic Commission for Europe, 2008.

[340] van Acker, V., Witlox, F., and Wee, B. V. The effects of the land use system on travel behavior: A structural equation modeling approach. *Transportation Planning and Technology 30* (2007), 331–353.

[341] Vanhoof, M., Combes, S., and de Bellefon, M.-P. Mining mobile phone data to detect urban areas. In *SIS 2017. Statistics and Data Science: new challenges, new generations* (2017), Proceedings of the Conference of the Italian Statistical Society, Firenze University Press, pp. 1005–1012.

[342] Vasileva, R. Interviews, 2019.

[343] Vichiensan, V., Miyamoto, K., and Rujopakarn, W. An empirical study of land use/transport interaction in Bangkok with operational model application. In *Proceedings of the 7th International Conference of Eastern Asia Society for Transportation Studies* (2007).

[344] Vogel, N., Theisen, C., Leidig, J. P., Scripps, J., Graham, D. H., and Wolffe, G. Mining Mobile Datasets to Enable the Fine-grained Stochastic Simulation of Ebola Diffusion. *Procedia - Procedia Computer Science 51* (2015), 765–774.

[345] Vos, J. D., Derudder, B., Acker, V. V., and Witlox, F. Reducing car use: changing attitudes or relocating? the influence of residential dissonance on travel behavior. *J. Transp. Geography 22* (2012), 1–9.

[346] Wang, H., Calabrese, F., Lorenzo, G. D., and Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th International IEEE Conference on Intelligent Transportation Systems* (2010), pp. 318–323.

[347] Wang, H., Hunter, T., Bayen, A., Schechtner, K., and Gonzalez, M. C. Understanding road usage patterns in urban areas. *Scientific Reports 2* (2012), 1–6.

[348] Wang, M. *Understanding Activity Location Choice with Mobile Phone Data.* PhD thesis, University of Washington, Seattle, WA, 2014.

[349] Wang, M., Schrock, S. D., Broek, N. V., and Mulinazzi, T. Estimating dynamic origin-destination data and travel demand using cell phone network data. *Int. J. Intell. Transp. Syst. Res. 11* (2013), 76–86.

[350] Wartmann, F., Acheson, E., and Purves, R. Describing and comparing landscapes using tags, texts, and free lists: An interdisciplinary approach. *International Journal of Geographical Information Science 32* (2018), 1572–1592.

[351] Wee, B. V. Land use and transport: research and policy challenges. *J. Transp. Geography 10* (2002), 259–271.

[352] Wee, V. B. Evaluating the impact of land use on travel behaviour: The environment versus accessibility. *Journal of Transport Geography 19* (2011), 1530–1533.

[353] Wegener, M., and Fuerst, F. Land-use transport interaction: State of the art. Tech. rep., Institut f'ur Raumplanung, Dortmund, DE, 1999.

[354] Wesolowksi, A., Eagle, N., Noor, A. M., Snow, R. W., and Buckee, C. O. The impact of biases in mobile phone ownership on estimates of human mobility. *Interface (Journal of the Royal Society) 10* (2013).

[355] Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. Quantifying the impact of human mobility on malaria. *Science 338* (2012), 267–270.

[356] Wesolowski, A. P., and Eagle, N. Parameterizing the dynamics of slums. In *Artificial Intelligence for Development - Papers from the AAAI Spring Symposium* (2010), pp. 103–108.

[357] White, J., and Wells, I. Extracting origin destination information from mobile phone data. *Road Transportation and Control 486* (2002), 30–34.

[358] Widhalm, P., Nitsche, N., and Brandie, N. Transport mode detection with realistic smartphone sensor data. In *21st International Conference on Pattern Recognition (ICPR)* (2012), pp. 573–576.

[359] Widhalm, P., Yang, Y., Ulm, M., Athavale, S., and Gonzalez, M. Discovering urban activity patterns in cell phone data. *Transportation* (2015), 1–27.

[360] Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., and Dobra, A. Measures of human mobility using mobile phone records enhanced with GIS data. *PloS ONE 10* (2015).

[361] Wilson, A. G. The use of entropy maximising modes in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy 3* (1969), 108–126.

[362] Wirtz, S., Jakobs, E. M., and Ziefle, M. Age-specific usability issues of software interfaces. In *IEA 2009 - 17th World Congress on Ergonomics* (2009).

[363] Wismans, L. J. J., Friso, K., Rijsdijk, J., de Graaf, S. W., and Keij, J. Improving a priori demand estimates transport models using mobile phone data: A Rotterdam-region case. *Journal of Urban Technology 25* (2018), 1466–1853.

[364] Wolf, J., G. R. B. W. Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. *Transp. Res. Rec. 1768* (2001), 125–134.

[365] Wolf, J., Loechl, M., Thompson, T., and Arce, C. Trip rate analysis in GPS-enhanced personal travel surveys. *Transp. Surv. Qual. Innov. 28* (2003), 483–498.

[366] World Bank. Financial inclusion.

[367] Wright, S. Correlation and causation. *Agricultural Research 20* (1921), 557–558.

[368] Wu, S., Qiu, X., Usery, E. L., and Wang, L. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. *Annals of the Association of American Geographers 99* (2009), 7698.

[369] Xing, Z., Lin, W., Qiao, Y., Zhang, H., Sun, K., and Yang, J. Inferring land use type in urban area with mobile big data. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)* (2018), pp. 1835–1840.

[370] Yang, X., and Lo, C. P. Using a time series of satellite imagery to detect land use and land cover changes in the atlanta, georgia metropolitan area. *International Journal of Remote Sensing 23* (2002), 1775–1798.

[371] Yin, J., Yin, Z., Zhong, H., Xu, S., Hu, X., Wang, J., and Wu, J. Monitoring urban expansion and land use/land cover changes of Shanghai Metropolitan area during the transitional economy (19792009) in china. *Environmental Monitoring and Assessment 177* (2011), 609621.

[372] Yuan, F., Sawaya, K. E., Loeffelholz, B. C., and Bauer, M. E. Land cover classification and change analysis of twin cities (minnesota) metropolitan area by multitemporal landsat remote sensing. *Remote Sensing of Environment 98* (2005), 317–328.

[373] Yuan, J., Wang, D., and Li, R. Remote sensing image segmentation by combining spectral and texture features. *IEEE Transactions on Geoscience and Remote Sensing 52* (2014), 1624.

[374] Zandelbergen, P. Accuracy of iPhone locations: a comparison of assisted GPS, wifi and cellular positioning. *Transactions in GIS 13* (2009).

[375] Zang, H., and Bloot, J. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile Computing and Networking* (2011), pp. 145–156.

[376] Zhang, Y., Qin, X., Dong, S., and Ran, B. Daily o-d matrix estimation using cellular probe data. In *Transportation Research Board 89th Annual Meeting* (2010).

[377] Zhang, Y., Wu, W., He, Q., and Li, C. Public transport use among the urban and rural elderly in China effects of personal, attitudinal, household, social-environment and built-environment factors. *Journal of Transport and Land Use 11* (2018), 701–719.

[378] Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., and Yin, L. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science 30* (2016), 1738–1762.

[379] Zheng, Y. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol 6* (2015), 29–41.

[380] Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W. Understanding mobility based on GPS data. In *Proceedings of UbiComp'08* (Seoul, South Korea, 2008).

[381] Ziefle, M., and Jakobs, E. New challenges in human computer interaction: Strategic directions and interdisciplinary trends. In *4th International Conference on Competitive Manufacturing Technologies (COMA)* (2010), pp. 389–398.

Appendix A

Mobile Financial Services in Africa

A.0.1 Mobile Financial Services and the 'Unbanked'

More than 2 billion people are struggling to access traditional 'brick and mortar' banking services with the highest numbers found in emerging economies in the Global South [90]. The intrinsic lack of access forces them to rely on generally unsafe, inconvenient and costly informal financial services or cash. At the same time, traditional banking institutions are struggling to make their business model work for those at the lower end of the socio-economic spectrum. A plight that is only exacerbated in rural areas due to sparse residential density. Over the past decade, Mobile Financial Services (MFS) have emerged as a major alternative to traditional banking services.

MFS is an umbrella term for several services offered by Financial Service providers:

Mobile banking is a channel to interact with a bank via mobile devices [30] and has gained major traction with the introduction of the smartphone and mobile apps.

Mobile payment is the payment for goods and services using mobile devices either remotely or at the point of sale [203]

Mobile money transfer or **cross-operator transaction** are Person to Person or *P2P* transfers within and depending on inter-operability agreements across carriers within a country [81, 151]

Mobile international remittance services are closely linked to mobile money trans-

> fers and are often used by migrant workers for remittance payments to family members. There has been a shift from traditional providers such as Western Union or MoneyGram to MNO services due to lower cost of services and a wider availability of mobile devices [239]

Those MFS offered by MNO have been identified as a novel and effective avenue for progressing toward financial inclusion, which has been recognised by the World Bank as a key enabler for economic development [366], by providing accessible financial services to the previously 'Unbanked' [166, 194, 246, 310].

> "The main goal of the service is financial inclusion for the financially excluded. Mobile money is based on the offer of simple financial services for customers. It provides access to electronic accounts where customers can deposit cash up to a certain ceiling and from which they can withdraw cash and manage their electronic money. Access and subscription to these accounts and associated services are enabled usually by ownership of a national identity card. Opening, crediting and managing accounts is free (only money transfers are taxed). Mobile money services allow subscribers to send or receive money to/from subscribers using the same service, or banked customers (domestic transfers and/or international remittances) and/or allow bill payments. Mobile money services users can rely on a growing network of service provider employees and retail commercial partners, which allow them to deposit and withdraw cash." [274, p.155]

The offered services range from simple deposits, withdrawals and transfers to interoperable transfers between the MFS services of different MNO, bill payments, insurance and saving accounts [150, 151, 246]. MFS offerings overcome two key issues: the cost of traditional banking services and spatial proximity to banking infrastructure [247]. The use of agents (such as small shops, street vendors, bank branches) enables 'banking beyond branches' [13]. Agents can step in to tackle the lack of distribution network [224] and close the proximity gap to traditional brick and mortar branches [42] by administering

cash deposits and withdrawals for MFS customers in the field [235].

MFS are offered by MNO across Africa, Asia and Latin America [252].

> "The numbers of mobile money users and countries with access to mobile money services are growing constantly. Diffusion of mobile money within countries was in an introductory stage up to 2012, after which it entered the growth stage. The diffusion of mobile money services across countries has continued to increase, but at a slower pace and can be said to be entering a saturation phase." [16]

While the availability of MFS offerings has increased drastically, a large number of services have already had to shut down as they were not economically viable and failed to offset setup, infrastructure and maintenance costs MNO incurred [118]. The growth and failure of MFS services have received significant attention within the literature, as will be discussed in more detail in §4.2.1. The most successful services can be found in sub-Saharan Africa, where they were originally developed.

A.0.2 Mobile Financial Services in Africa

M-PESA [1] was first introduced by Safaricom in Kenya in March 2007 with similar systems soon established in Sudan, Ghana, Tanzania and elsewhere. When surveying M-Pesa usage in Kenya, Jack and Suri (2011) [180] found a sharp increase in account ownership from 43% in 2008 to 70% in 2009 as services became more widely known and customers developed trust into the services.

Mobile Money in Tanzania

The uptake in MFS services has been extremely high across East African countries. In Tanzania, the MFS user base grew to 5.5 million users within the first four years from the introduction of M-Pesa by Vodacom in Tanzania in 2008 and Tigo Pesa in 2010. By

[1] Pesa is the Kishawhili word for 'money' - M[obile] Money

2013 MFS agents and outlets significantly outnumbered traditional offerings in Tanzania with 17,541 MFS agents compared to only 1,117 ATM's and 504 bank branches [65, 257]. This disparity reflects a wider trend within emerging markets, where at least nine other countries have been found to have more MFS accounts than bank accounts [150]. It was estimated, that over 50% of subscribers use MFS with over 35% of households having at least one MFS user compared to only 2% having active bank accounts [107, 245, 257]. The volume of MFS transactions nearly doubled between 2013 and 2015 as nearly one-third of active MFS accounts in East Africa were registered with Tanzanian MNO by 2015 [136, 174]. The TCRA estimated there to be 16.5 million MFS accounts in March 2016 [328].

Appendix B

MFS Error Codes

The list below contains all error codes contained within the MFS dataset described in §2.2.2. Those transaction that do not have an error code attached to them are assumed to be successful transactions.

- Success
  - 200 success
  - error000
- Maximum balance exceeded
  - 60028 Unable to complete transaction as transaction amount is more than the maximum txn value for the recipient
  - 60030 Unable to complete transaction as the payee account would go above the maximum balance
- Maximum number of transactions
  - 00008 Max percentage transfer reach limit
  - 60011 Unable to complete transaction as maximum number of transactions per day reached
  - 60021 Unable to complete transaction as maximum number of transactions per day for payee reached

- Amount below minimum transaction size
  - 00409 The transaction amount is less than the minimum value defined for this service
  - 60017 Unable to complete transaction as transaction amount is less than the minimum txn value for sender
  - 60027 Unable to complete transaction as amount is less than the minimum limit
  - elec016 Amount too low. Request amount exceeds the minimum amount for this meter
  - error015 requested amount is lower than the minimum amount set per transaction
- Amount above transaction limit
  - 00031 Request amount more then allowed in the network
  - 60024 Unable to complete transaction as maximum transaction value per day reached
  - 60026 Unable to complete transaction as maximum transaction value per month reached
  - 60017 Unable to complete transaction as amount is more than the maximum limit
  - 00410 Unable to complete transaction as amount is more than the maximum limit
  - 60014 Unable to complete transaction as maximum transaction value per day reached
  - 60016 Unable to complete transaction as maximum transaction value per day reached
  - error012 Amount is out of the range set for purchased in a single transaction

- Insufficient balance
  - 60019 Unable to complete transaction as account would go below minimum balance
  - error013 insufficient balance to complete transaction
  - elec001 The specified customer has been blocked by Credit Control
- 00042 Requested amount not in multiple of allowed value

Appendix C

Ward-Level Feature Maps

The following maps for ward-level statistics on socio-economics, land use density, land use diversity and mobility patterns were generated using CDR and MFS data as well as from external ground reference data as discussed in more detail in Chapter 6.

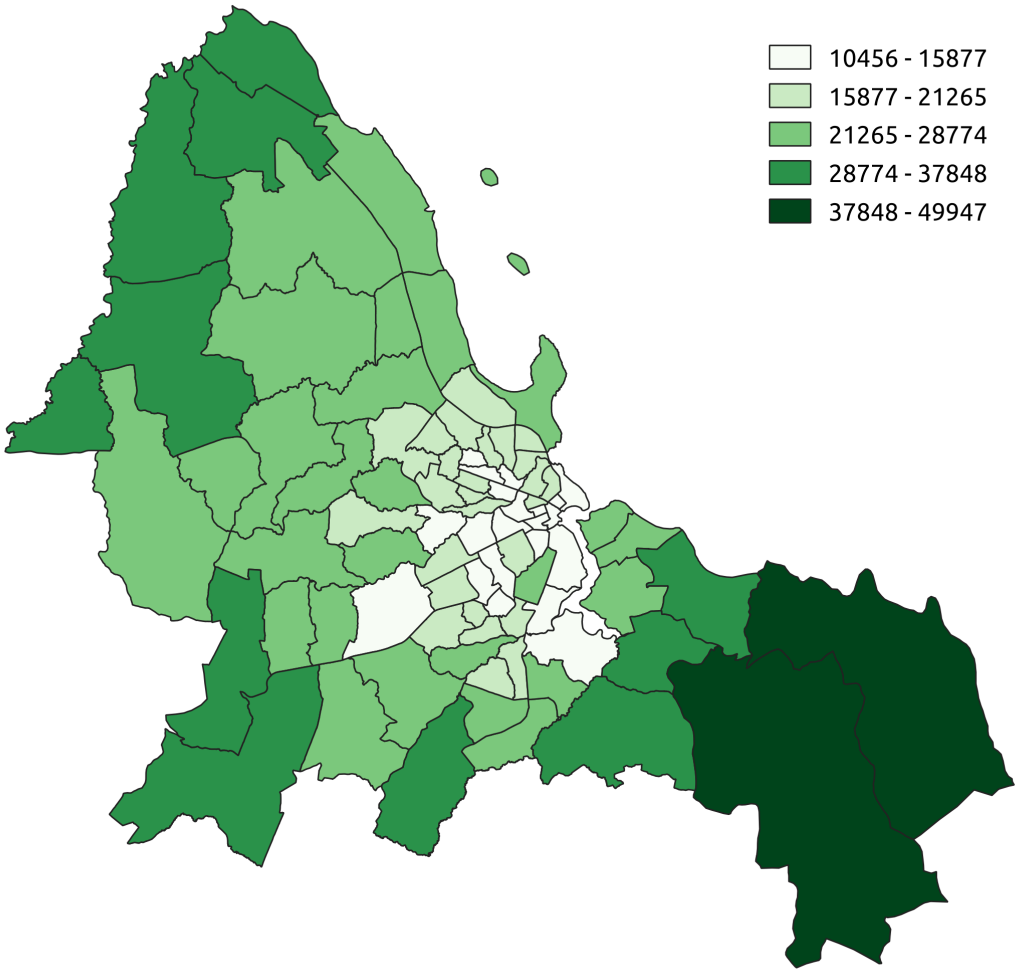


Figure C.1: Average trajectory distance across wards within the metropolitan area of Dar es Salaam

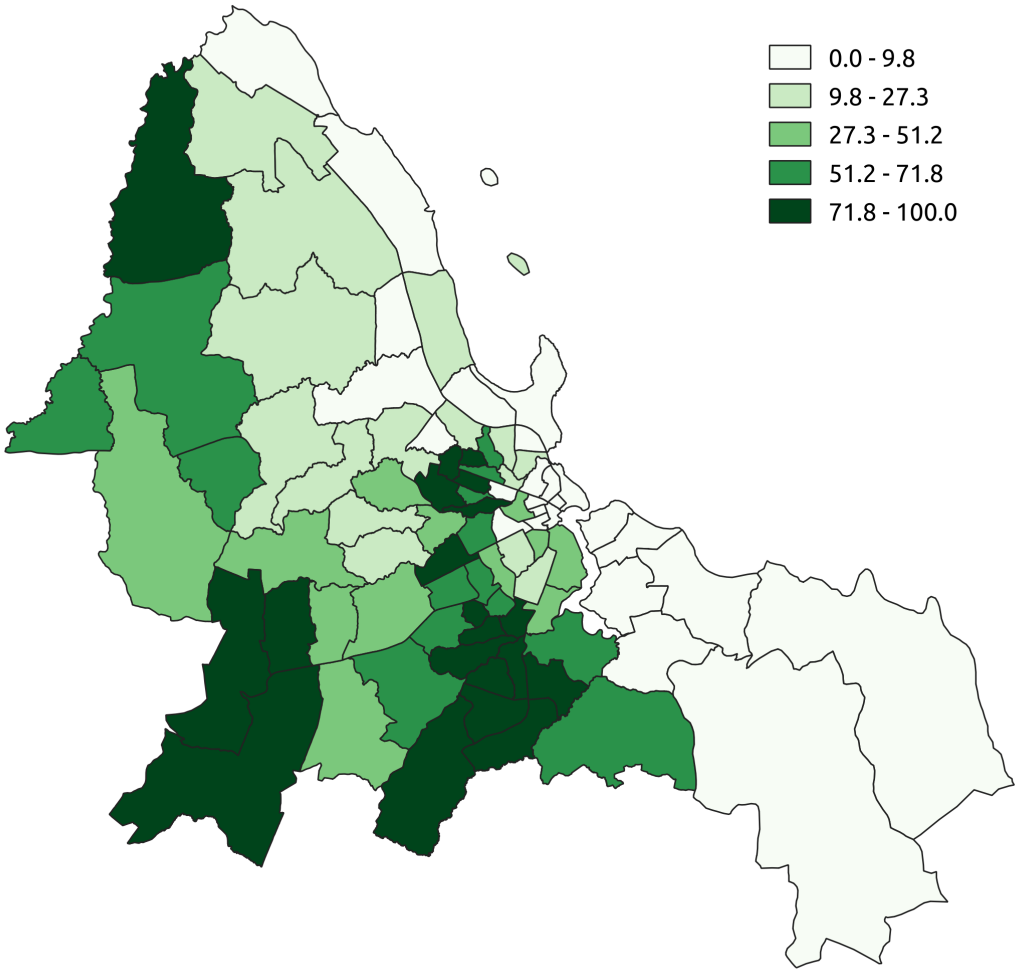


Figure C.2: Percent low-income across wards within the metropolitan area of Dar es Salaam

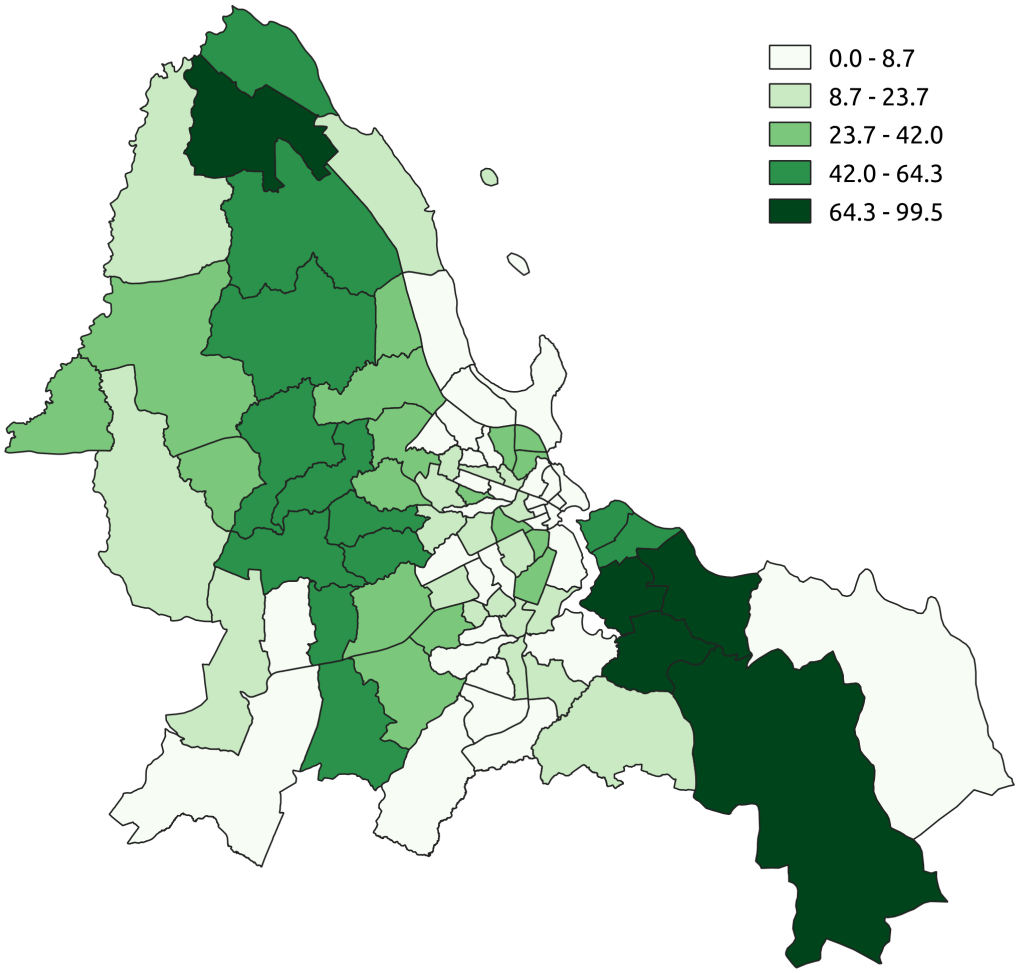


Figure C.3: Percent medium-income across wards within the metropolitan area of Dar es Salaam

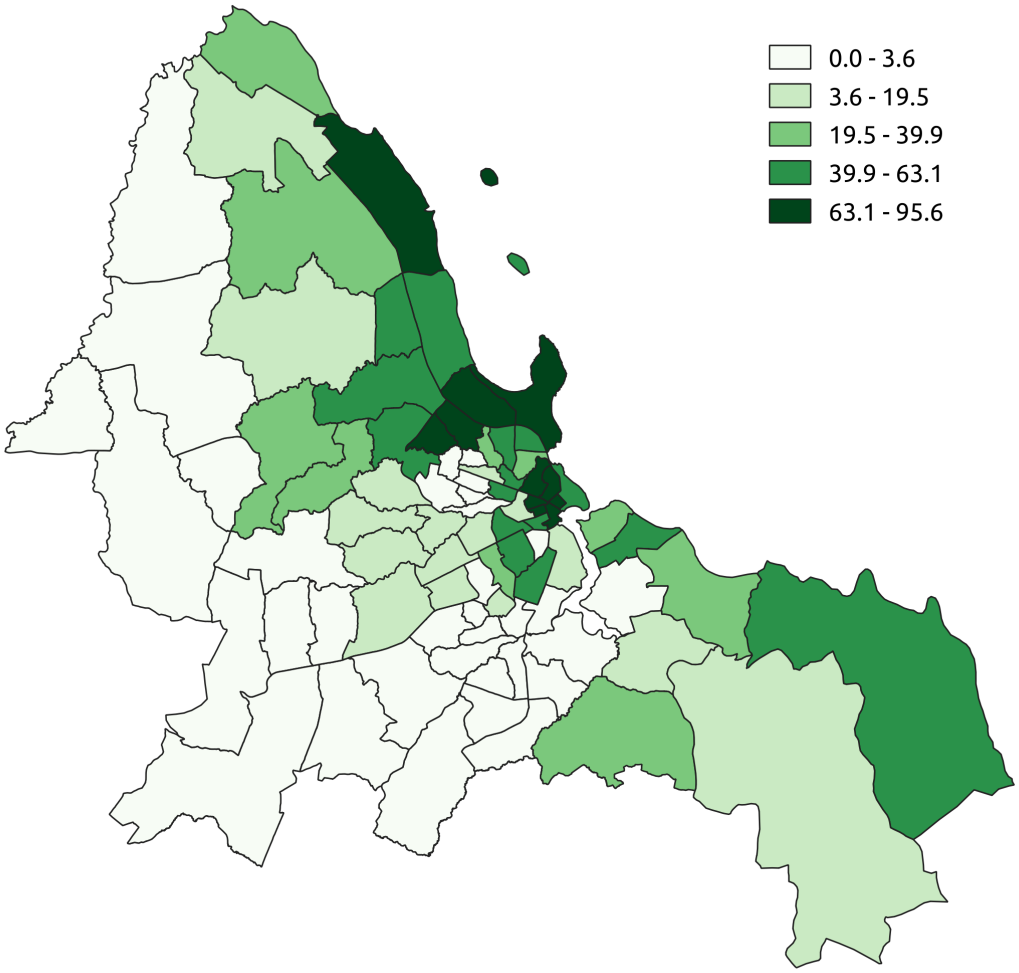


Figure C.4: Percent high-income across wards within the metropolitan area of Dar es Salaam

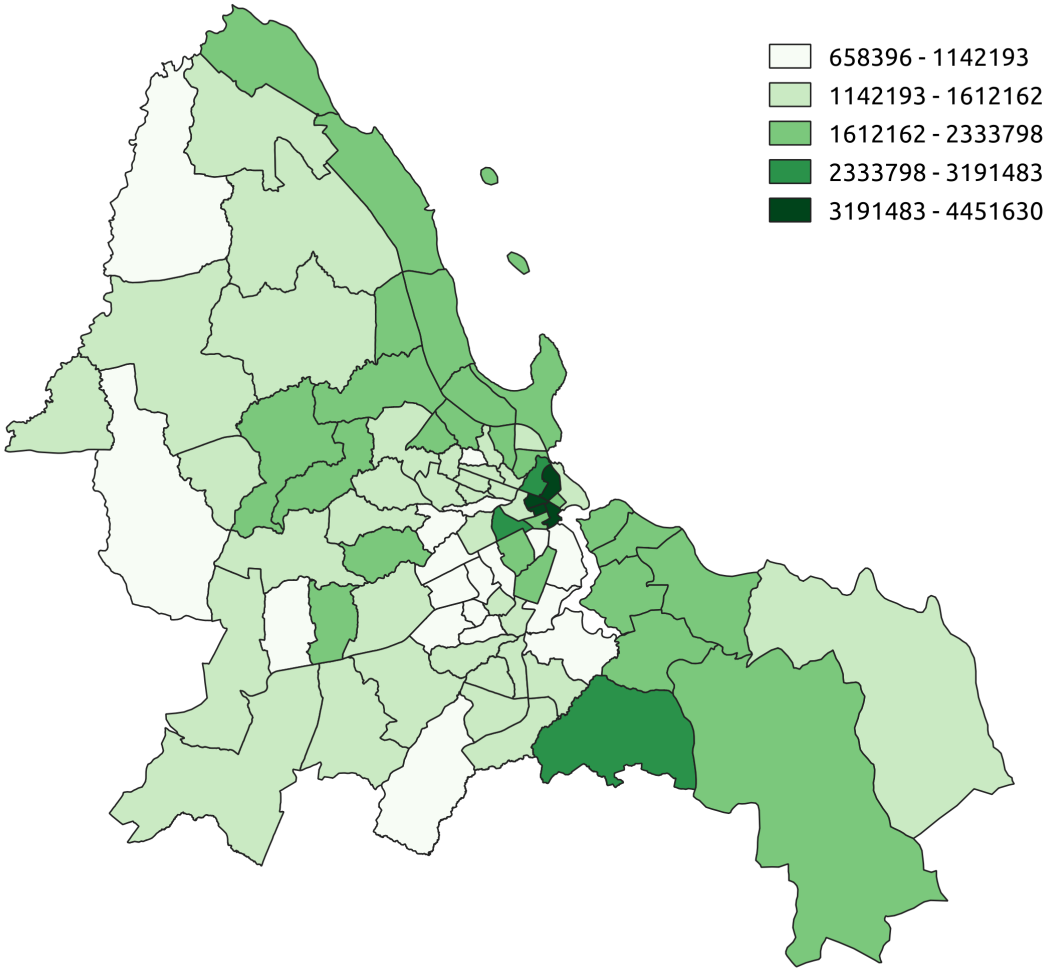


Figure C.5: Spending uptake across wards within the metropolitan area of Dar es Salaam

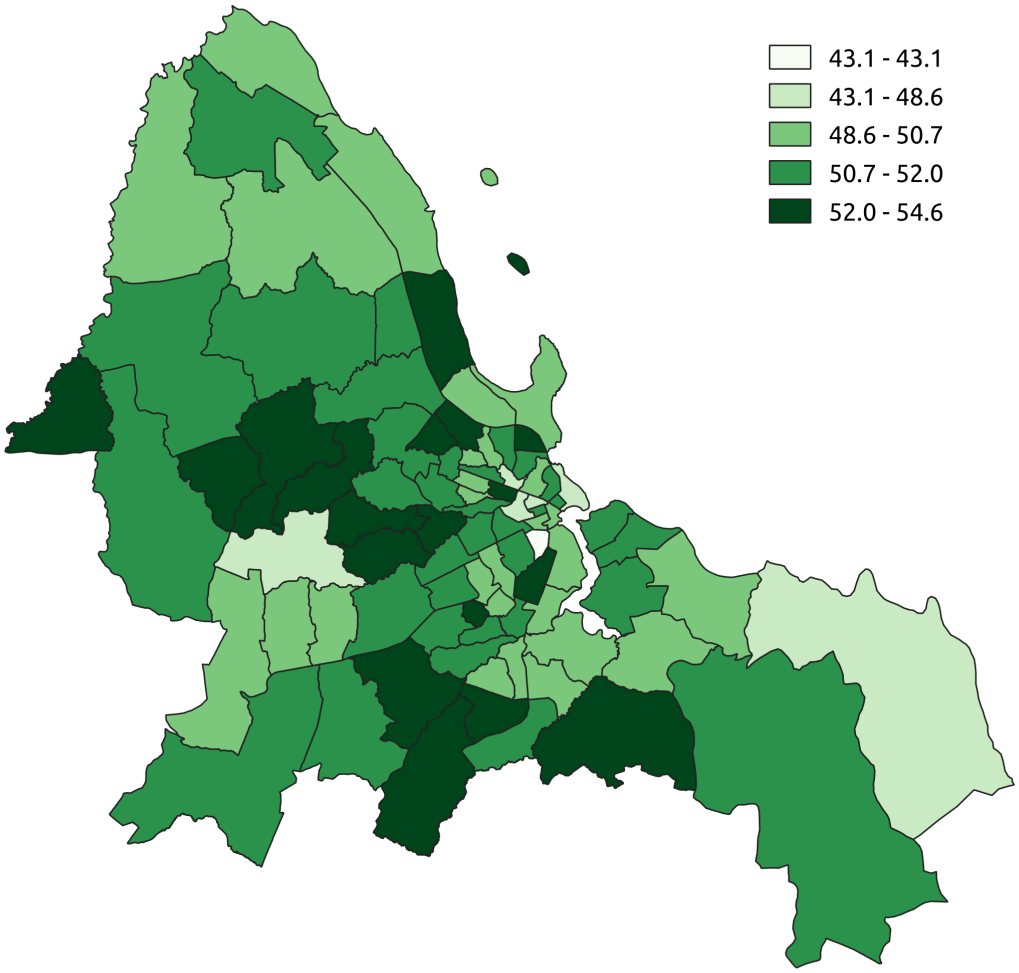


Figure C.6: Gender split across wards within the metropolitan area of Dar es Salaam as percentage female

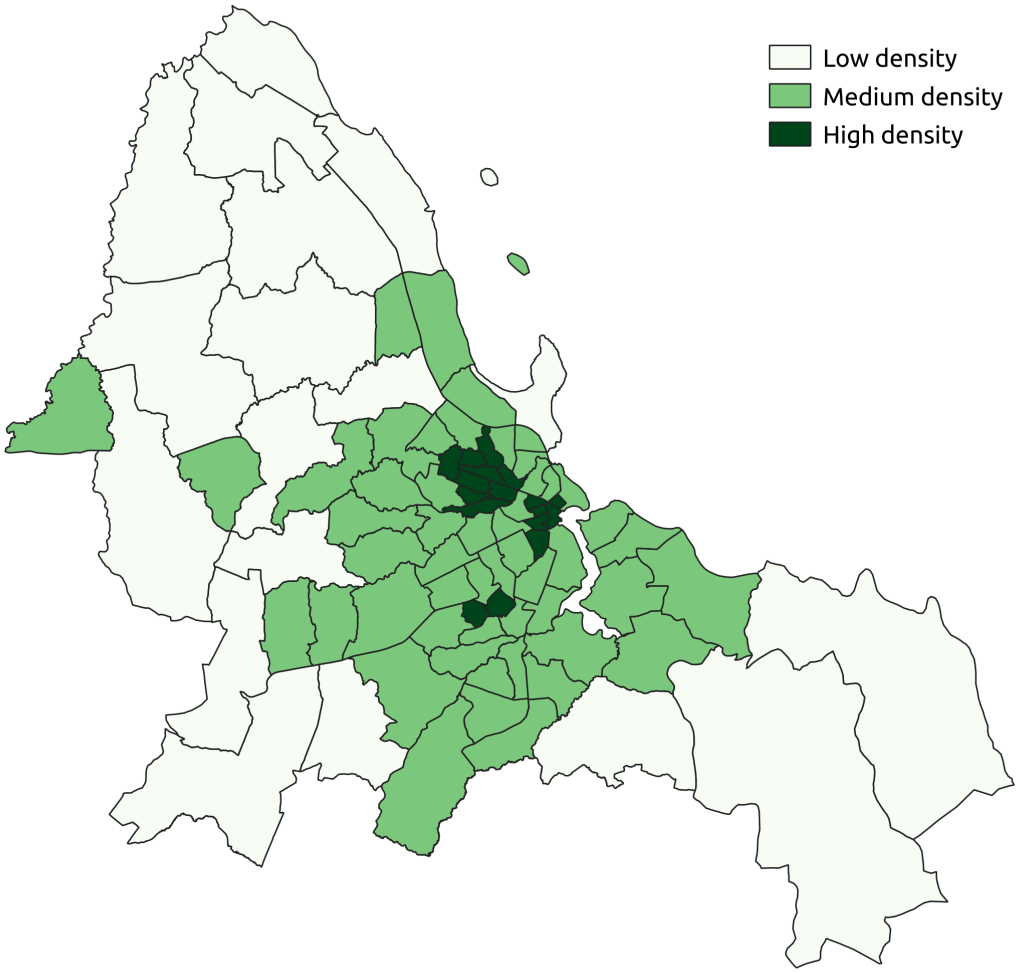


Figure C.7: Network event density across wards within the metropolitan area of Dar es Salaam

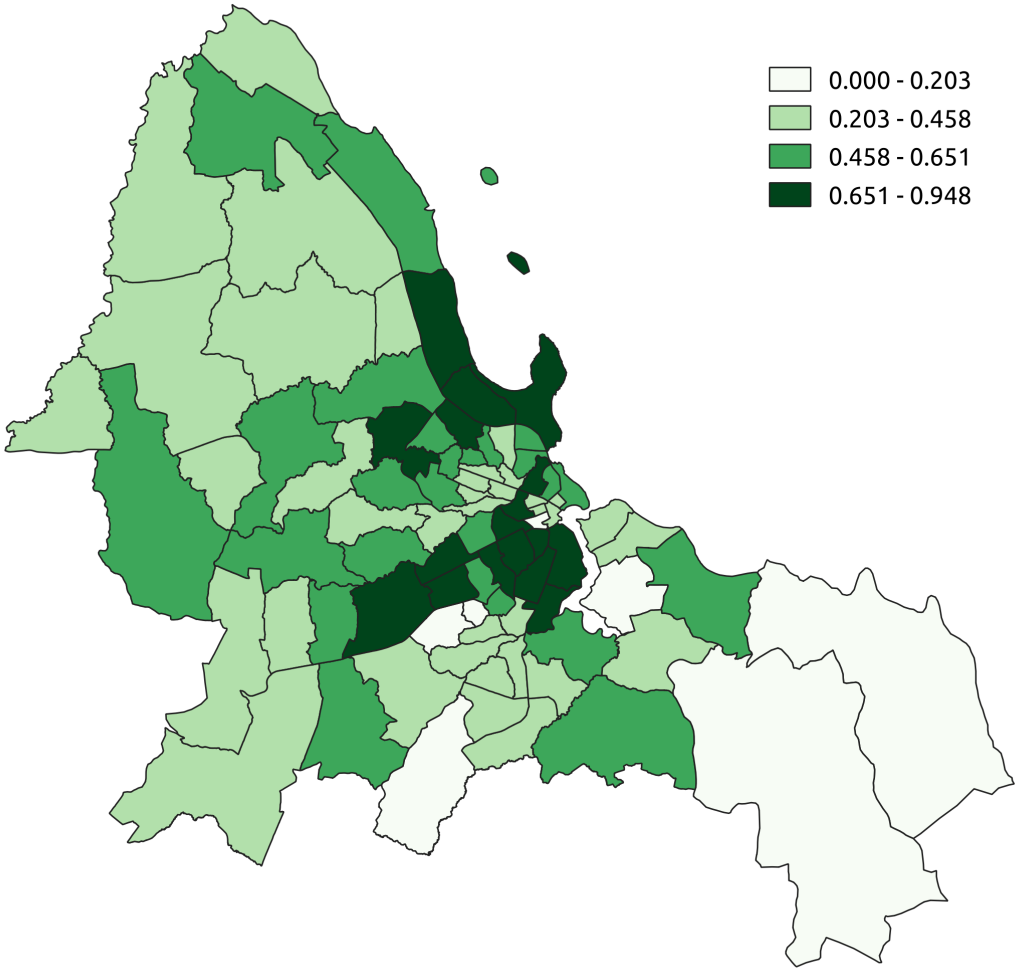


Figure C.8: Land use mixture across wards within the metropolitan area of Dar es Salaam

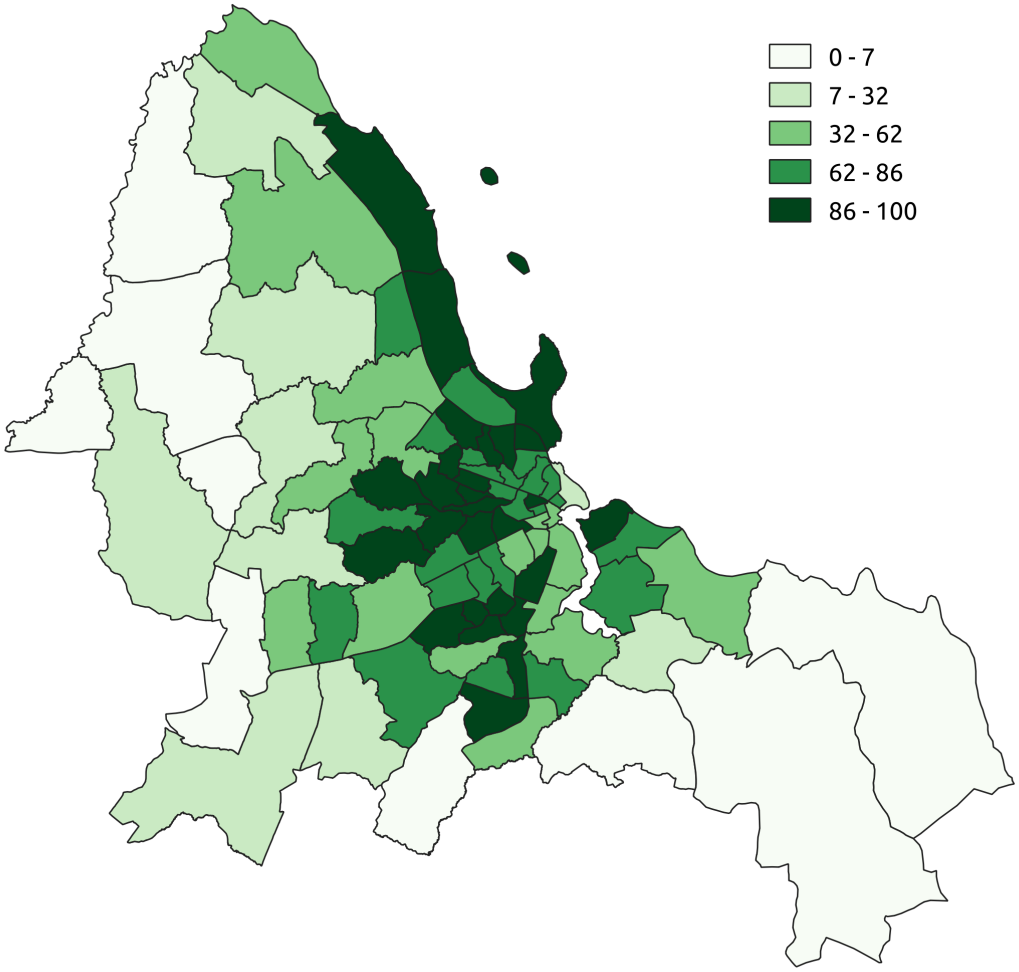


Figure C.9: Percent residential across wards within the metropolitan area of Dar es Salaam

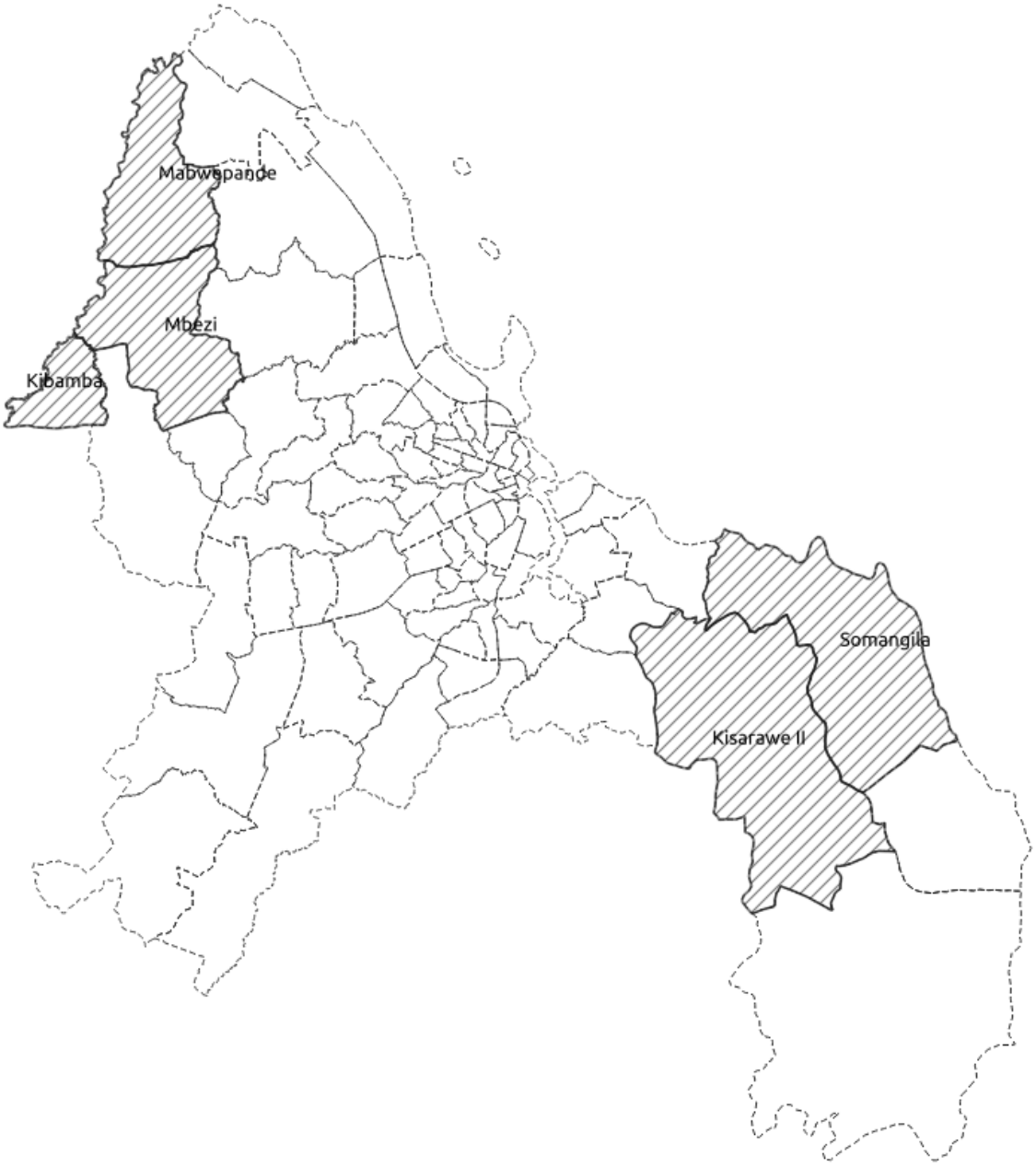


Figure C.10: Wards of Kibamba, Mabwepande, Somangila, Kisarawe II and Mbezi as the only wards within the metropolitan area of Dar es Salaam classed as entirely non-residential through their Voronoi intersections

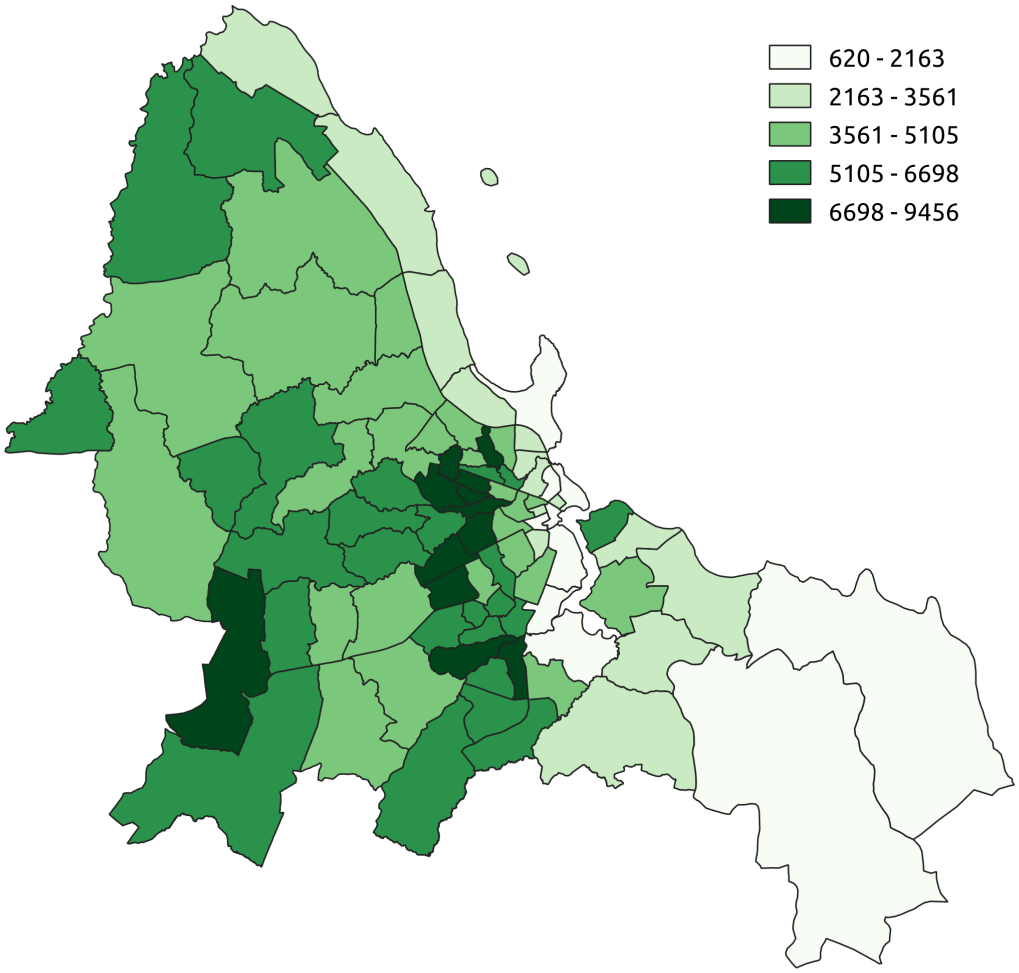


Figure C.11: Number of inbound trips across wards within the metropolitan area of Dar es Salaam