GIMA

Geographical Information Management and Applications

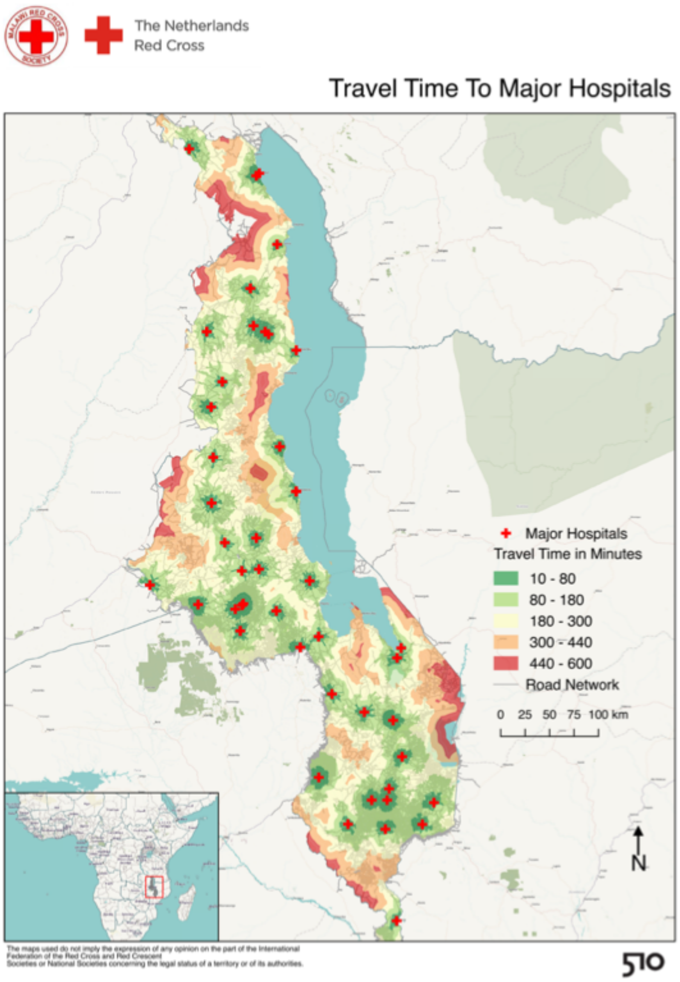
Remoteness as a proxy for social vulnerability in Malawian Traditional Authorities

An open data and open-source approach

|Master Thesis|

MSc Geographical Information Management and Applications | UU WUR UT TUD

Student	*J.G. Wilbrink*
Email	j.g.wilbrink@students.uu.nl
Supervisor	*dr.ir. R.L.G. (Rob) Lemmens*
Professor	prof. dr. M.J. (Menno-Jan) Kraak
Date	May 2017
Place	The Hague – Netherlands


The Netherlands Red Cross
Travel Time To Major Hospitals
Major Hospitals
Travel Time in Minutes
10 - 80
80 - 180
180 - 300
300 - 440
440 - 600
Road Network
0 25 50 75 100 km
N
The maps used do not imply the expression of any opinion on the part of the International Federation of the Red Cross and Red Crescent Societies or National Societies concerning the legal status of a territory or of its authorities.
510

Universiteit Utrecht

TUDelft

UNIVERSITY OF TWENTE.

WAGENINGEN UR

PREFACE

Six months of research and contributing to the humanitarian information sector has resulted in this thesis in the context of the Geographic Information Management and Applications Master's program. Throughout the period at the Netherlands Red Cross and 510 humanitarian data team I have gained knowledge and new skills to cope with open data and transform it into useful information. The humanitarian world of data fascinated me from the moment I started working with the first openly available excel files and produced the first results based on open-source geospatial analysis - it keeps surprising me the longer I work with it. I am amazed by the open repositories with not only open data, but open-source tools and plugins. I am pleased to be working in a time where skills, expertise, and created information are shared with such enthusiasm.

The use of open data and tools produced by people all over the world have motivated me to apply the data management, analysis, and visualisation skills obtained during my studies in this research for the benefit of real and on-going global problems.

Working within the motivated 510 humanitarian data team is an amazing experience. I want to thank all the team members for contributing with their knowledge and expertise to improve the humanitarian operations of the future. I want to thank Maarten van der Veen for providing me with the opportunities of going to Malawi to validate and calibrate the initial results and have a first-hand experience into the operations of the Netherlands Red Cross (NLRC) and the Malawi Red Cross Society (MRCS). I want to thank my supervisor, Rob Lemmens for supporting me along the way with all the remote and direct support. I want to express my gratitude to all the MRCS staff and volunteers for their help, confidence and most of all: kindness during the field-testing of the remoteness indicators. Finally, I want to express my sincere gratitude to my girlfriend, Rosa, my parents, Johan and Ada, and my dearest friends, for supporting me throughout my studies.

SUMMARY

Within the humanitarian aid sector there is a need to assess the social vulnerability to natural hazards within developing countries on a granular geo-spatial level. This is commonly done through the use of qualitative and paper-based field surveying methods, which are costly and time consuming. These field assessments are often project-based, linked to a certain area and consequently results in data filled with gaps. The current and most common methodology used within the humanitarian community, including the International Federation of the Red Cross and Red Crescent Societies (IFRC), is a so-called Vulnerability and Capacity Assessment (VCA). Through this methodology, insights into the vulnerability and coping capacity of a community with regard to natural hazards and climate change are gathered. In this way, humanitarian organisations are better informed on the susceptibility to natural hazards pre-disaster and the aid neediness of communities' post-disaster. More often than not, a VCA is not conducted or the outcomes on an affected community are not readily accessible by information managers, causing uninformed aid provisioning pre-disaster and in the immediate disaster response phase. The fact that there is no information readily available for these areas, creates a need for alternative evaluation methods based on secondary and open data sources. In this research, through the use of open data; OpenStreetMap (OSM), The Humanitarian Data Portal (HDX), Malawi Spatial Data Platform (MASDAP), and CEISIN/Facebook data, several remoteness indicators are created using Postgis, PgRouting, Osmosis, and other open-source geospatial analysis tools to generate a proxy for social vulnerability. The created remoteness indicators are field tested in Malawi using the digital surveying tools OpenDataKit (ODK) and OpenMapKit(OMK) to validate the initial outcomes of the geo-spatial analysis with the aim of calibrating the geospatial parameters to real-life values. The proxy is then created based on the remoteness indicators, which are trained through machine learning (ML) techniques to an existing social vulnerability index (SoVI) for Malawi. The results are visualized through an interactive map using a web application where the proxy result and the existing vulnerability index can be compared. To make this process generally applicable in other disaster-prone countries, the development of the proxy is done with the use of open data and open-source software and the results are interactively visualised through an online dashboard to make the process more transparent and replicable. The developed proxy predicts the SoVI scores for Malawian Traditional authorities with an accuracy of 64% and can be run on this granular level within hours, compared to months or even years using traditional vulnerability assessments. This potentially creates a fast and accurate assessment alternative for decision-makers who decide on the project areas of humanitarian organisations and it may provide a fast and evidence-based insight into vulnerability in the immediate response phase after a natural hazard.

Content

1 Introduction

"The key question is how to prevent known hazards from producing disasters" (Former UN Secretary-General Kofi Annan, 2003).

Lives and livelihoods are destroyed by natural and man-made disasters all around the world. The frequency and devastating impact of these disasters is increasing over the last two decades. The loss of lives, livelihoods, damages to property, infrastructure and assets has exposed the preparedness and response capacity of governments, businesses, international humanitarian assistance agencies, and communities (Sahay, Menon & Gupta, 2016). According to UNISDR (2013 p.5), *"Between 2002 and 2011, there were 4,130 disasters recorded, resulting from natural hazards around the world where 1,117,527 people perished and a minimum of US$ 1,195 billion was recorded in losses. In the year 2011 alone, 302 disasters claimed 19,782 lives; affected 206 million people and inflicted damages worth a minimum of estimated US$ 366 billion."* Sahey, Menon & Gupta (2016) identify the bottlenecks of humanitarian logistics that affected countries face during the immediate post-disaster phase. One of the major examples is *"the choice of the optimal way in which these relief materials can reach the disaster-affected communities".* Fast decisions with regard to identifying these communities are imperative to effective humanitarian logistics.

Natural hazards will not be preventable in the foreseeable future, alternatives to alleviate the consequences of natural hazards are highly sought after within the humanitarian aid community (UNISDR, 2013; IFRC, 2006). A report by the United Nations office of Disaster Risk Reduction (UNISDR 2002 p.392) identifies that there is a need to *"develop indicators for disaster risk reduction measures".* Within the Hyogo Framework for Action local risk assessments, maps, indicators, and vulnerability are stressed as key aspects in risk reductive measures (UNISDR, 2005). Research on disaster risk reduction (DRR) increasingly shows that it is often not the hazard that determines the disaster, but the vulnerability, exposure, and ability to anticipate, respond, and recover. Shifting from hazard response to identification, assessment, and ranking of vulnerabilities become more emphasized (Blaikie, Cannon, Davis, and Wisner, 2014).

The ability to measure vulnerability is increasingly seen as a means to effectively reduce risk of hazards and as a way to promote the culture of disaster resilience (Kasperson et al., 2005). In the light of the increasing frequency and intensity of disasters and environmental degradation, the measurement of vulnerability is instrumental in more accurate decision-making for disaster relief programs (Birkmann, 2006). Where research in the past has been solely focused on measuring the hazard characteristics (Lewis, 1999), there has been a paradigm shift towards taking into account the interaction of a vulnerable society, its environment, economy and infrastructure, with the potentially damaging physical event. Major disasters such as Hurricane Katrina and the Haiti earthquakes have shown how social processes underlying poverty and marginalization can lead to increased susceptibility to injury, death, dislocation, and difficulties in recovery (Tate, 2012). Within literature there is a call for the identification, assessment and ranking of several different vulnerabilities in the light of disaster prevention and relief policies (Cutter, 2003; Kasperson, 2006; Johnson et al., 2012).

In the humanitarian aid community, numerous initiatives are launched to accurately assess the vulnerability of communities in developing countries that are prone to natural hazards. The IFRC, the world's largest independent humanitarian agency, has applied vulnerability and capacity assessments (VCA's) in their work since 1996. (IFRC 1996, 1999, 2006a, 2006b, 2006c, 2006d, 2007). These VCA's often serve as a base- and end-line during Water, Sanitation, and Health (WASH), or Disaster Risk Reduction (DRR) projects carried out by the humanitarian organisation operating in the area. The VCA is conducted with the aim to reduce the impact of disasters, and public health emergencies whilst increasing the capacity to address the issues of vulnerability (IFRC, 2007). These assessments are focused on community involvement and are often conducted through a mostly qualitative and paper-based research on a community level. Within literature it has been identified that the links between researchers and humanitarian workers, and the application of vulnerability assessments can be further improved and strengthened: *'particularly in terms of the timeliness of such assessments in relation to the humanitarian agencies' requirements and ways in which assessments can minimize the burden on disaster-affected communities'* (Miller et al., 2010 p.5). In an interview with Jan Heeger, a Water, Sanitation, and Health (WASH) expert for the Netherlands Red Cross and Jeroen Bot, National Coordinator for Shelter and Care for the Netherlands Red Cross, several aspects in relation to information needs are discussed. *"Often, relevant indicators are missing due to the limited amount of information readily available; there is so much data, but with the limited amount of time and tools available it often comes down to experience when making decisions"* (Heeger, 2016). *"The first thing to do is look at the baseline data on vulnerability when waiting on a rapid assessment in country. Within this baseline data, often conducted through VCAs, many information gaps exist"* (Bot, 2016).

The VCA outcomes not only identify the areas on which a community can improve to become more resilient in the case of a natural disaster, it helps to identify the communities within an area or country which are likely to have higher aid neediness when a disaster occurs (IFRC, 2006). Humanitarian aid organizations are often the first actors of a larger response of national, international and transnational organizations that are deployed to take post-disaster measures. As stated above, humanitarian aid organizations assist in logistics and material deliveries to alleviate post disaster stress. The aim is to reduce human suffering and the loss of lives. The complex nature of a disaster response gets amplified by the lack of information or reliable field assessments (ACAPS, 2016). Aid response needs to be conducted in a timely manner, with limited amount of resources and workers; emergency managers often have to make decisions on assigning resources to affected areas based on assessments carried out by NGOs and other local institutions (Fiedrich et al., 2000; Heeger, 2016). In the initial response phase these assessments are either not available, contain information gaps, or are presented in paper-based qualitative reports and thus not useful for a fast and accurate overview of the vulnerability of the affected areas.

1.1 Problem Statement

The current way in which vulnerability assessments are conducted is labour-intensive in a way that it is too costly to be able to assess all communities' pre-disaster, and the available results are often paper-based and too qualitative to be of any assistance post-disaster in the rapid response phase. Furthermore, the vulnerability assessments conducted through different data collecting stakeholders; government, NGO's, and private sector are often project and context specific within Malawi (Jinon, 2017). This results in data- and subsequently information gaps. Emergency managers rely heavily on the information that is gathered by NGOs and local institutions that are active in the area to identify the areas that are most vulnerable, and thus most in need of aid. Within the humanitarian sector, alternatives to collect reliable and complete vulnerability data is limited. Humanitarian organizations therefore identified the need of a tool that generates reliable vulnerability information in the case of limited data availability (Heeger, 2016). Where the lack of reliable vulnerability information is a financial burden pre-disaster, it becomes a cause of unnecessary suffering and loss of lives post-disaster. A technical tool that can create granular, accurate and complete vulnerability data comprehensible for emergency managers and covering a complete country can contribute to current VCA data and potentially act as an alternative.

Increasingly, tools are being developed to assess the aid neediness of an area affected by natural disasters. The use of new technologies, with regard to crowd-sourced data collection have been proven effective with regard to humanitarian response and recovery. Increasingly, these unofficial data sources have a role in the disaster response phase and preparatory phases of humanitarian projects building resilience in hazard-prone countries. The Missing Maps project (Missing Maps, 2017) work hand in hand with NGOs like the Humanitarian OpenStreetMap Team (HOT) (HOT, 2017) to increase the amount and quality of geographic data of areas that have not been mapped properly as a means fill the gaps in geographic data of these areas. These initiatives distribute the collected data through OSM, and is therefore accessible to any organisation interested in using and enriching the data. However, while these technologies with the use of communities and citizens all over the globe, are often used to collect information on exposure, efforts directed towards assessing and monitoring hazards and vulnerability using these platforms remain limited (McCallin, 2016).

A tool that has recently gained traction is the Priority Index Model (PIM). It is used to identify communities most affected post-disasters (Benini, 2015; 510, 2016). A PIM visualizes the entities within an area that are in highest need of aid assistance, and is becoming an important tool for emergency managers. A PIM is composed in such a manner that a limited amount of data is necessary in the post-disaster phase to produce credible information. For a hurricane this can exist of wind speed and rain fall data, for an earthquake this can exist of a “shake map” and for floods this can be a flood-extent map and precipitation data (510, 2016). PIM’s rely heavily on quantified vulnerability data. As stated earlier, vulnerability data is often either not available, not usable or quantifiable for the purpose of a PIM. The development of a tool that can generate accurate, timely, and quantitative vulnerability data for the purpose of humanitarian response and logistics has not yet been established.

1.2 A proxy for vulnerability

The aim of this research is to reduce the workload and costs of the collection of these data and offer an alternative tool to fill information gaps on the vulnerability of hazard prone areas.

In this project, it is researched if ‘remoteness’ can act as a proxy for ‘social vulnerability’ of Traditional Authorities (TA’s), which are municipal areas in Malawi. Remoteness of these TA’s are identified through several ‘indicators’: (i) *distance*, such as to health facilities, water points, schools, general public and private facilities; (ii) *geographic properties,* such as ruggedness of the landscape, and (iii) *density Figures*, of the population, settlements, roads and other structures. Most indicators are created with the use of OpenStreetMap data (OSM, 2014), which is available for nearly any country and can therefore be applied to different countries when proven valid.

The final remoteness indicators are selected through the use of machine-learning (ML) techniques. An important source of secondary data for this research is an existing vulnerability tool for Malawi. It is a raster data set based on a multi-criteria analysis to indicate vulnerability of geographic areas in Malawi. The abovementioned remoteness indicators are selected based on their combined fit to this existing vulnerability index. The geographic granularity of the data is accumulated on TA level, which makes the data statistically comparable. The best set of indicators will then be chosen to create a proxy for social vulnerability of Malawian TA’s.

A major component of this research is focused on the amount and quality of the data used during the process. Due to the meagre data footprint of vulnerable communities (Cutter et al. 2008) it is essential to find a methodology that does not require a large amount of data and subsequently does so, using openly available data. To assess the possibilities of the development of credible proxy indicators based on openly available data, the set of remoteness indicators will be trained on the social vulnerability index score on a TA level. The social vulnerability index (SoVI) is adapted from the existing vulnerability index of the Regional Centre for Mapping of Resources for Development (RCMRD, 2015). The vulnerability index created by the RCMRD uses a large amount of indicators, mainly based on closed data, reliant on governmental agencies, are costly, and consequently labour intensive to obtain.

The SoVI will act as a dependent variable, referred to in this research as the Y-variable, in the statistical and machine learning (ML) analysis of the proxy indicators. In doing so, the most important remoteness indicators, these will be referred to as X-variables, are identified that can predict the social vulnerability of the areas; 'the remoteness proxy'. Based on this proxy, it will become possible to quickly assess the social vulnerability of other communities within Malawi that have not yet undergone a VCA and it may be applied to other countries where granularity of the available data is limited.

1.3 Challenges

During this research several challenges need to be overcome. In this section these challenges will shortly be introduced. To create an accurate proxy for vulnerability in Malawi several aspects need to be taken into account. As this research should be useful for humanitarian stakeholders, it is necessary to identify the type of vulnerability and remoteness that is relevant for decision-making for humanitarian intervention or in a crisis situation after a natural disaster. Subsequently, within the realm of open data, an existing source of accurate data needs to be identified to create a set of proxies and the dependent vulnerability index (SoVI). Therefore, the first challenge is identified: (i) The concepts of vulnerability, remoteness, and accessibility need to be explored carefully to fit within the humanitarian context of disaster risk reduction (DRR).

Whilst developing the proxy for vulnerability, the remoteness indicators need to be chosen carefully. The indicators need to be created in a relatively simple manner to make it applicable in different scenario's and regions around the world this led to the second challenge: (ii) The process of developing a proxy for SoVI is potentially useful in different disaster-prone countries around the world and should therefore be produced in a replicable manner.

The remoteness indicators in this study are partly based on the travel time over the road network to different points of interest for disaster prone areas; hospitals, secondary schools, primary schools, cities, trading centres, and water points. It is important to work within the Malawi context for this study, a field study is therefore necessary to identify the means of transportation, the used facilities, and the time that is necessary for households to reach the different points of interest. With these aspects in mind, the third challenge is identified: (iii) The indicators used to create the proxy need to be validated on their accuracy in the Malawi context and fed into the GIS analysis of this study.

As indicated above, the proxy should be easily reproducible for future use in different contexts. Therefore, it is important to find a fast and reproducible technique to correlate the remoteness indicators to the existing vulnerability index. Aspects herein again are the use of open-source software packages and scripting techniques. This leads to the fourth challenge: (iv) Accurate statistical methods should be explored to make the analysis convenient to reproduce.

As this analysis is aimed at providing an alternative to vulnerability data available to humanitarian decision-makers the presentation of the results should be useful. Through several interviews with information managers within the Red Cross movements, aspects as 'intuitive', 'map-based', 'transparent', and 'visual' have come forward (Heeger, 2016; Bot, 2016). This has led to the final challenge: (v) The presentation of the data for information managers, and humanitarian field operators is explored to turn the produced data into useful information.

The five challenges are briefly summarized in the following section.

i. *Defining social vulnerability and assigning a social vulnerability index for Malawi*

The concept of vulnerability is widely described through literature and its definition is not only context specific, but evolves per context. Within DRR the concept of vulnerability is often described as a component of disaster risk in combination with exposure and the hazard characteristics. For this research choice of concept for Malawi will be underpinned in the scientific context section.

ii. Developing the remoteness indicators in a rapid reproducible manner

The indicators for remoteness are chosen with several elements taken into consideration*. The data* used to create the indicators should be i) openly available ii) available for the entire research area iii) potentially scalable to other areas; with a lower or higher granularity. *The tools* used to create the indicators needed to be i) open-sourced ii) able to handle scripted work-flow, so the rapid reproducibility would be enhanced for future digital humanitarians or researchers.

iii. Validate and calibrate the remoteness parameters with a field study in Malawi

Several remoteness aspects used in this research needed to be validated within a field-study. For this field study the remoteness of several randomly selected Traditional Authorities and communities within the Thyolo district was evaluated. Using the outcomes of this field study, several remoteness indicators were calibrated to the measured values.

iv. Train the remoteness indicators on the selected social vulnerability index (SoVI)

The quantified SoVI for the TA’s in Malawi will act as the dependent (Y) variable which the remoteness indicators are trained on to accurately predict the SoVI, using machine learning (ML) techniques.

v. Visualize and present data in a transparent and interactive manner

The outcomes of the remoteness indicators and developed proxy is presented in a web-application where the difference between the predicted SoVI and the existing SoVi with both the original indicators and the remoteness indicators are presented.

2 Research Objectives

This chapter presents and defines the research objectives. The main objective (2.1) and subsequently the sub-objectives (2.2) are described. The objectives result is several research questions, that are shortly introduced in section 2.3.

2.1 Main objective

The main objective of this study is to develop a proxy for social vulnerability of communities in Malawi based on remoteness indicators. It is chosen to use the social vulnerability, simply because the vulnerability data should be usable in the case of different types of natural disasters, in different regions or countries over the globe. Within the concept of social vulnerability, the specific hazard component and the community's exposure is therefore not taken into account. The social component of vulnerability in relation to natural disasters is more commonly seen as imperative to informed humanitarian response and often shows to be complex and time consuming to collect. The proxy can act as a support on identifying the social vulnerability for communities that have not yet undergone extensive vulnerability assessments, in other words fill in the information gaps within the field of vulnerability data within Malawi. Furthermore, the use of proxies may show useful as a methodology in other hazard prone countries. This enables humanitarian aid decision-making to become more evidence-based in pre- and post- disaster management. Specifically, the problem of limited data availability is incorporated in the study. As many official data sources are commonly closed or limitedly available for humanitarian organisations, the aim is to identify accurate openly available alternatives to official data and data collected through field surveys (VCAs). Particularly the potential role of OpenStreetMap (OSM) data and open-source software for humanitarian purposes is researched through this study.

2.2 Sub-objectives

The main objective is structured through three subsequent sub-objectives. The first sub-objective is added to deal with the semantics and conceptualization of 'remoteness' and 'vulnerability'. Both terms are open to interpretation and are ambiguously described within literature and are differently used within humanitarian organisations.

The second sub-objective is aimed at developing the remoteness indicators through appropriate GIS techniques whilst creating an automated script/work flow to create the remoteness index and make it more usable and create transparency for policy makers, emergency managers and digital humanitarians. The automated workflow will be applied through the process of OSM data extraction, analysis, application, and visualisation. During this research, the available techniques to create such automated scripts are explored (e.g. PostGIS/PgRouting/Leaflet). The availability of open data based on OSM is important to make general assumptions about the remoteness proxy, however to implement the proxy to 'predict' the vulnerability of communities in other regions or countries, it is important to create an automated workflow in the form of a script that can be run with limited knowledge and research the minimum amount of data required to have a valuable result.

The third sub-objective is to distinguish the statistical relationships between the SoVI (dependent Y-variable) and the remoteness indicators (independent X-variables). This will be done through machine learning techniques using R-studio to script the algorithms, as it is deemed the most relevant methodology in this open-source and open data-driven study. The data set exists of the X-variables (remoteness indicators) and the Y-variable, (SoVI) on a municipal level (Traditional Authorities, TAs). With the use of machine learning, a part of the data set is used to 'train' the algorithm and prediction is done on the other part of the data set and subsequently compared if the predicted values match the existing SoVI scores for the individual municipalities. This is an essential step in the development of the proxy indicator. This part of the analysis will be conducted after the X-variables are developed and validated, and before the results are visualised through a web application.

2.3 Research Questions

Based on the problem statement and research objectives a main research question is formulated. Through the main research question it is aimed to identify communities in Malawi with a high social vulnerability to natural disasters through a remoteness proxy. The sub-questions are formulated to structure the research process.

'How can 'remoteness' contribute to the identification of vulnerability to natural disasters for communities in Malawi, using open data and open-source tools?'

Related to the first sub-objective the following research questions are formulated. It is important to conceptualize social vulnerability and remoteness in a valid manner. To correlate the index scores of remoteness to social vulnerability, both need to be conceptualized. Through an extensive literature review the initial indicators may be developed and subsequently, a practical analysis will lead to the final development of the index.

1. What indicators are used to create a social vulnerability index for natural hazards?
2. What data can be used to develop remoteness indicators as a proxy for social vulnerability?

Related to the second sub-objective; creating transparent, reproducible, and usable results, the sub-questions are formulated as follows:

3. What GIS-techniques and open-source tools are most relevant to extract remoteness indicators from OSM data?
4. How does scripted database management, analysis, and visualization contribute to the usability and reproducibility of the remoteness proxy?
5. How can the quality and completeness of open data for Malawi be validated to be sufficient for calculating a useful social vulnerability indicator?

Related to the final sub-objective, the relationship between the dependent and independent variables will be defined. The usability of such a proxy depends on the availability of data. Especially for remote areas in global south countries this shows to be challenging; remote and vulnerable communities often have a meagre data footprint (Pedraza-Martinez, 2013). With the use of open data, it often shows to be challenging to acquire the accurate data, in this case remoteness specific data, also the data availability may be problematic. Within the machine learning methodology, a part of the validation is done internally. For the validation of the remoteness indicators, thus the open data and OSM data, a 2,5-week field visit is conducted to Malawian communities located in different TA's in February 2017. The following sub-questions will be answered regarding these objectives:

6. What statistical models can be used to define the relationships between the SoVI (Y-variable) and the developed remoteness indicators (X-variables) for Malawian Traditional Authorities?
7. How accurate can the proxy, developed through the used models, predict SoVI?

3 Methodology

This chapter first provides an overview of the methodology and a visualisation of the conceptual methodology scheme. The following paragraphs are structured by the sub-objectives / research questions and will provide the methods used to result in answering the research questions.

3.1 Methodology Scheme

The research will be structured by the three sub-objectives presented in Chapter 2. The three sub-objectives and their elements are shown in the methodology scheme in *Figure 3.1****.*** The elements between the blue dashed lines are indicated separately; the analyses are completely done using scripting methods and are therefore easily repeatable. The scheme represents the letters A to G and the logo represents the scripting method used; command line, SQL, R, Java-script. G represents the GitHub where the code is shared. The complete script for every step can also be found in Appendix B. The three sub-objectives are indicated using the blue indicators with their coherent numbers. The chapters use the same indicators to make it more relatable to the scheme in *Figure 3.1.*

Sub-objective

1

The first sub-objective is aimed at an elaborate conceptualization of vulnerability and remoteness. For vulnerability, the RCMRD data, as well as other open data sources are investigated and consequently, the development of the social vulnerability index (SoVI) is done. Simultaneously, data on remoteness is gathered through the use of openly available data sources; HDX, the humanitarian data exchange (data.humdata.org), MASDAP, the Malawi Spatial Data Platform (masdap.mw), Open health sites data (Healthsites, 2017) and Madzi Alipo, water point data exchange (MadziAlipo, 2016) for the water point locations. Within this stage the remoteness indicators are p-coded; all the data is sorted by the p-code of the traditional authorities (TA), and ordered in a matrix database. This objective is related to research questions 1 and 2.

Sub-objective

2

The second objective runs parallel to the other objectives during the whole research. The presentation of the results and the role of visualizations in the form of a web application will be investigated in the light of usability and transparency for policy-makers and emergency managers and the applicability on future regions and countries. The scope of this research is primarily aimed at Malawian communities, more specifically on the Traditional Authority (TA) level. However, it is aimed to develop a general model for future applications in other hazard-prone areas across the global south. Different open data sources are used of which OSM forms the most important basis for the GIS analyses conducted during this study. The GIS analyses are validated through field testing in Malawi using the digital surveying tools OpenDataKit (ODK, 2016) and OpenMapKit (OMK, 2017). The field visit exists of ground measurements in the form of: distance, travel time, health site location, water point location, and semi-structured interviews with local authorities on community level. While validating the GIS analyses, data is collected simultaneously to contribute and enrich the existing OSM data set for Malawi. This objective is related to research questions 3, 4, and 5.

Sub-objective

3

Within this objective the dependent and independent variables are related using machine learning techniques. Primarily, the data quality is assessed, the data is checked for missing values, correctness of input or data that is not applicable. The used data is assessed and filtered on usability. Different data sets are created to investigate the minimum amount of data and parameters that are necessary to produce a scientifically sound result. By correlating the produced remoteness indicators to the existing social vulnerability index for Malawi, the proxy has a strong empirical underpinning. The Social Vulnerability Index (SoVI) will be based on the data behind the existing vulnerability tool for Malawi created by the Regional Centre for Mapping of Resources for Development (RCMRD, 2015). During the statistical analysis the data set is split into a training and test data set to make sure predictions are not over-fitted to the machine learning models. A so-called cross validation will take place as described in section 4.7. This objective is related to research question 6 and 7.

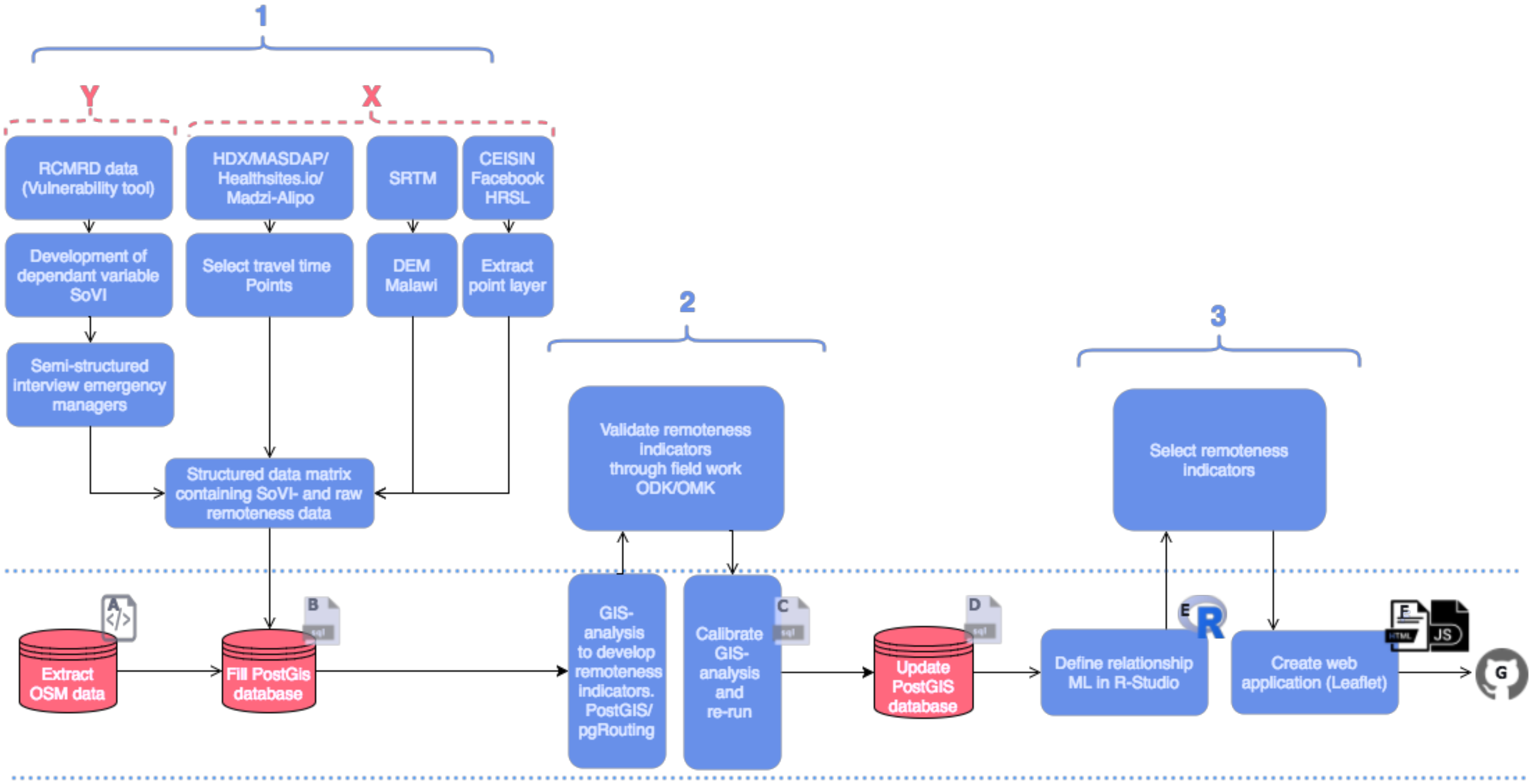

1
Y
X
RCMRD data (Vulnerability tool)
HDX/MASDAP/ Healthsites.io/ Madzi-Alipo
SRTM
CEISIN Facebook HRSL
Development of dependant variable SoVI
Select travel time Points
DEM Malawi
Extract point layer
Semi-structured interview emergency managers
Structured data matrix containing SoVI- and raw remoteness data
2
Validate remoteness indicators through field work ODK/OMK
3
Select remoteness indicators
A
B
C
D
E
Extract OSM data
Fill PostGis database
GIS-analysis to develop remoteness indicators. PostGIS/ pgRouting
Calibrate GIS-analysis and re-run
Update PostGIS database
Define relationship ML in R-Studio
Create web application (Leaflet)
sql
HTML
JS

Figure 3.1: Methodology Scheme following the 3 sub-objectives.
1) Searching for data on remoteness indicators in relation to vulnerability and developing the Y-variable SoVI
2) Developing and validating/calibrating remoteness indicators through open-source GIS tools
3) Select remoteness indicators for the final remoteness proxy through (open-source) machine learning tools and visualizing output online.
All steps indicated between blue dotted lines are scripted and can be repeated through an automated process. Scripts can be found on Github indicated in Appendix B

3.2 Social Vulnerability data assessment

RQ 1: What indicators are used to create a social vulnerability index for natural hazards?

For the initial phase of the research an extended literature study is necessary to identify the characteristics of vulnerability to natural hazards and the development of a proxy indicator based on remoteness, as well as the evaluation of existing vulnerability indices. One of the most accepted methods for humanitarian organisations at the moment to assess vulnerability is the Vulnerability and Capacity Assessment (VCA) method, developed by the International Federation of Red Cross and Red Crescent Societies (IFRC). This is an assessment method, designed to identify the vulnerability and capacity on a community level for developing countries as described in *Chapter 1*.

The primary work for this research is aimed at creating a suitable Y-variable to train the remoteness indicators on. The Y-variable should comply with the concept of vulnerability according to the consensus in current DRR literature. Complying to current consensus on the concept will contribute to the standardization of the concept and general understanding of vulnerability to natural hazards. Especially emergency managers within humanitarian organisations are increasingly interested in evidence-based and fast decision-making tools. To add to the process of creating such tools, an alternative to currently existing and validated vulnerability indices is researched in this study. This alternative is in the form of a proxy, based on remoteness indicators, as explained in chapter 2. To make the proxy as accurate as possible, the existing vulnerability index for Malawi is carefully chosen. As becomes clear in the research context Chapter 4, the SoVI is the most relevant conceptualisation choice for this study.

A data research on possible data sources for SoVI is performed. Within the NLRC and IFRC multiple VCA studies were performed for Malawi in the most recent years. Through the information channels of the Red Cross community, VCA data and meta data is evaluated. Alternative open data portals for Malawi and humanitarian data are explored to find a valid SoVI. During the research into the available VCAs conducted for Malawi, the relevance of this research became more evident; from within the NLRC it showed difficult to obtain VCA results. During the field visit to Malawi, after directly approaching the monitoring and evaluation manager of the Malawi Red Cross Society (MRCS) the first VCA report was obtained. The actual questionnaire data behind the VCA report could not be reproduced. As the scope of this research is aimed at using open data, an open vulnerability tool for Malawi is used to create the SoVI.

The selected index is subsequently discussed with several experienced emergency managers that have been active in the field of DRR. The concept of vulnerability in relation to hazards is discussed, as well as the possibility of a social vulnerability index to indicate pre- and post-disaster areas within a hazard-prone area that are most likely to be susceptible to disaster. The completeness of current data sets available to these information managers is often filled with information gaps.

3.3 Remoteness indicator data assessment

RQ 2: What data can be used to develop remoteness indicators as a proxy for social vulnerability?

Through this second research question the potential remoteness indicators are developed that are potentially useful for the social vulnerability proxy. Potentially, because the remoteness indicators can be excluded during the machine learning process. Based on the review on remoteness in *Section 3.3* several indicators for remoteness and techniques are selected to analyse the remoteness of areas within Malawi. Particularly, through the study of Maru et al. (2014) several indicators based on sanitation, health, and education are selected. The network analysis technique used in the Accessibility/Remoteness Index for Australia (ARIA, 2013) assessment of governance has shown to be effective in the identification of remote areas within Australia. Therefore, this technique is explored through the scripted geospatial analysis in this research.

An extensive search on available open data sources for the proxy development is conducted. Several data portals are used during this research: i) HDX, the humanitarian data exchange portal, where information useful for humanitarian organisations is published in a regulated manner, meta data is always available for this data portal, it is developed by UN-OCHA and official P-coded administrative boundaries are available. P-codes are used to identify administrative boundaries for all the countries globally in a standardised manner. ii) MASDAP, the Malawi Spatial Data Platform, openly accessible data maintained by the Malawi Department of Surveys and other open data authorities within Malawi. Iii) Healthsites.io, an open data platform and NGO, specifically initiated to make health site locations and their attributes around the world openly accessible through crowd-sourced data. IV) Madzi Alipo, a data platform developed specifically for water point location data and their attributes, including maintenance information. The portal is developed by Fisherman Rest, a local NGO for central Africa and partly relief on crowd-sourced data. V) CEISIN and Facebook developed High Resolution Settlement data (HRSL), where a pattern recognition algorithm detects human settlements based on high resolution Digital Globe satellite imagery. VI) SRTM (RCMRD, 2015) Digital Elevation Model (DEM) clipped to Malawi country admin boundary.

3.4 Remoteness indicator development

RQ 3: What GIS-techniques and open-source tools are most relevant to extract remoteness indicators from OSM data?

This third research question is aimed at making a structured data matrix with relevant X- and Y-variables based on TA-level administrative boundaries of Malawi. The X-variables will be developed using several GIS techniques. The techniques are chosen based on the speed in which remoteness indicators can be developed, the ease, costs, and the replicability for other countries or areas around the globe. It has therefore been chosen to explore scripting techniques and to use database management tools based on open-source tooling. As open-source tooling is often multi-platform, free, and modifiable to specific needs, it was deemed as highly relevant for the research purposes. Several different tools are explored and the most relevant for this research are selected. To modify the parameters used for the GIS-analyses, a field study is conducted, where random TA's within one district were visited with the help of 20 voluntary enumerators and the use of open-source surveying software OpenDataKit (ODK) and

OpenMapKit (OMK). All the buildings within a village were visited, and inhabitants were questioned through a survey to quantify the traveling time to major hospitals within the area. The village head, or chief, was asked permission to enter the village in a Malawi Red Cross Society capacity, and briefly interviewed on the traveling time to major hospitals and schools for his village to get a more general indication.

A short overview of the tooling and function is given for the purpose of a first indication of how different open-source tools are used, and where in the report these steps are described and visualised.

Table 3.1 Overview of Open-Source tooling used in this research and where in the report these are described

Open-Source Tool	Function	Report Section
Osmosis	OSM data extraction & data transformation	5.2.1
Imposm	OSM data extraction	5.2.1
OSM2PgRouting	Populate Postgis database with rouTable network	5.2.2 | 5.2.3
PgAdmin III	Postgis and PgRouting database management	5.2.2 | 5.2.3 | 5.3
Postgis	Geographic database management	5.2.2 | 5.3
PgRouting	GIS network techniques	5.3
QGIS	Raster operations & visualisation	5.2.2 | 5.2.3 | 5.2.4 | 5.3
R	Data cleaning, Multivariate linear regression, and Random forest regression analysis	5.5
Java-script dashboard	Results published in dashboard using leaflet library	5.5.4

3.5 Scripted work-flow

RQ 4: How does scripted database management, analysis, and visualisation contribute to the usability and reproducibility of the remoteness proxy?

As shown in the methodology scheme (*Figure 3.1*) the steps within the blue dashed lines are done through the use of scripting techniques. A) OSM data is extracted using different command-line tools. For the automated download of all the OSM data for Malawi, Imposm and Osmosis are used. For the extraction of rouTable network, based on the OSM road network of Malawi, OSM2PgRouting is used. B) To fill the Postgis database Osmosis is used to extract the data from the openly available PBF file of the OSM data for Malawi; the PBF file is extracted to an OSM (XML) file and subsequently loaded into the Postgis-enabled SQL-database C) All the GIS analyses are conducted using SQL queries. This database management scripting tool has shown to work fast, accurately, and precisely. Algorithms behind GIS analysis tools can be observed, and modified to specific needs and are therefore beneficial for the replicability. D) After calibrating the GIS algorithms to fit Malawi parameters, with regard to travel speeds over the road, the SQL-based analyses are re-run and a modified TA-based data matrix is exported. E) The remoteness indicators are trained on the existing SoVi using scripted machine learning techniques in R-studio. The script can easily be adapted to any remoteness matrix structured by administrative boundaries and is scalable to other countries. F) The output of the analyses and the final remoteness indicators is visualized in a leaflet web-application, created through java-scripting and HTML techniques. G) The code for all the steps is shared in a Github, an open online tool for sharing, maintaining and versioning of code. This makes it easily accessible for future scholars, working to improve the techniques and usable for information managers in need of a fast indication of social vulnerability of a particular country or area.

Within the GIS analyses, several SQL-based functions are utilized. The steps of the network analysis, isochrones creation based on travel times, settlement selection and assigning travel time scores, and weighted averaging based on a spatial join between the settlements and the TA admin boundaries Table, will be elaborated on in the analyses and results chapter. The validation of the initial remoteness indicators with the field-study will be subsequently explored in the analyses and results chapter.

3.6 Open data quality

RQ 5: How can the quality and completeness of open data for Malawi be validated to be sufficient for calculating a useful social vulnerability indicator?

The use of open data is ideal for the situation where data availability is limited, or closed by data producing agencies. The quality of the data needs to be assessed carefully to base accurate assumptions on the data. For the data that will be used that come from the open data portals in this research, the meta-data will be evaluated and sources of data will be inspected similarly. OSM data is used to create a routable network and create remoteness indicators based on travel times to a variety of amenities over the road. The analysis is validated through a field-study in Malawi. The outcomes of this study is with regard to travel times and location of amenities will act as a subsequent quality assessment.

3.7 Relationships SoVI - Remoteness

RQ 6: What statistical models can be used to define the relationships between the SoVI (Y) and the developed remoteness indicators (X) for Malawian Traditional Authorities?

Through this research question, the statistical relationships will be defined between the developed remoteness indicators and the pre-defined SoVI. A multivariate linear regression model and a random forest regression model are run on the data matrix using the open-source analysis software R-studio. The combination of best predicting, and relative importance of the remoteness indicators are explored. To answer this research question, the following steps are taken:

Data exploration

During this data exploration phase, the data matrix is made suitable for statistical analysis. I) The data matrix will be explored for missing values; possible reasons for missing values are defined, some will be left out of the data set, others will be filled in accordingly; either by using the minimum or maximum values of the column. Subsequently, all the missing values 'NA's' are omitted from the data set. II) The normal distribution of all the variables is explored and transformed e.g. logarithmically in the case of a skewed distribution. Finally, collinearity is explored for the X-variables by observing pairwise scatterplots and correlation coefficients. The solution to covariance is discussed in the analysis section of this research.

Multivariate Linear Regression

The multivariate linear regression model (LM) is chosen as a suitable model for this multivariate regression problem. Several steps will be taken to find the most fitting LM. An automated X-variable selection is conducted using the 'step' function in R-Studio. This function conducts an extensive search, whereby a large number of possible subsets of the X-variables are explored to find the optimal set of X-variables with the highest predictive ability for the Y-variable. This function is useful in variable selection due to the fact that it deselects the variables that are not significant, while preserving the variables that may become significant in the subsequent subset. These mistakes often occur whilst conducting manual forward selecting methods or backwards elimination techniques. After having made the subset selection of X-variables, the final LM is created using the adjusted R-squared as selection criteria; which is an indicator for the amount of variance that is explained in the Y-variable by the LM. It is referred to as adjusted, because the score is adjusted to the amount of observations taken into account in the model.

The LM is run multiple times. Different combinations of X-variables are taken into account. The combinations are based on the existing collinearity that exists within the X-variables and are therefore unfit to join simultaneously. The collinearity between X-variables are an indication of which remoteness indicators are best to use to predict social vulnerability; and duplication of GIS-based remoteness indicator extraction can be avoided in the future.

The final LM is validated using a set of assumptions; the linearity, heteroscedasticity, and normality of residuals, and no or limited multicollinearity.

Random Forest Regression

A random forest model (RF) runs multiple decision trees within a single model. The decision trees predict outputs by splitting the X-variables at relevant points. The classifier within a RF model makes use of a number of decision trees to increase the prediction rates. During each iteration of the RF model, corrections are made during the training and averaging is done on the multiple decision trees' output. In normal decision trees, every node is split taking the best split of all variables into account. For the RF model, the split is made based on the best fit in a subset of X-variables randomly gathered for each node. The RF is therefore referred to as machine learning (ML) regression.

If the RF is run in R-Studio, the number of trees, number of variables tried at each split, the mean of squared residuals (MSR), and the percentage of variance explained can be shown. RF, as any ML technique is often criticized for its 'black box' effect; not much further insight is given into the described relationships. The model is subsequently fitted with several runs, like with the LM and assessed by the selection criteria R-squared.

3.8 Proxy prediction

RQ 7: How accurate can the proxy, developed through the used models, predict SoVI?

Cross Validate

To be able to assess the predicting accuracy of the models, it is chosen to use cross-validation of the data matrix. The data matrix is split randomly into a train- and test-set. Based on the number of observations the percentages of the split data sets are chosen: in this case sixty percent will be used as the train set and forty percent for the test set. This is done, to check over-fitting of the prediction models to the training data. The test set contains the forty percent of SoVI data for which the model has not been trained, predictive accuracy is compared, and subsequently, assumptions on the model fit can be made.

The performance of the models is assessed by the R-Squared values, a scatterplot interpretation and the root of MSE. When the models perform less on the test set than on the train set, this is an indication of over-fitting. Over-fitting may be solved by the elimination of X-variables with the lowest significance. Geographic representation of predicted values may result in insights into the predictive accuracy of the models.

The outcomes will indicate the strength of the proxy based on the implemented remoteness indicators. The model will present a combination of remoteness indicators that can predict the social vulnerability on a municipal level for Malawi. The model may have the potential to be used on different granularity and within different geographic contexts.

4 Research context

This chapter will give an overview of the concepts highlighted in this research within the scientific context. An overview of humanitarian open data concepts, relevant research on hazard vulnerability, social vulnerability, vulnerability indices, remoteness, accessibility, and machine learning is given.

4.1 Humanitarian Open Data

Vetrò et al. (2016) describe open data as "open data are data that can be freely used, modified, and shared by anyone for any purpose". Within the contemporary data environment, open data and big data are becoming increasingly popular concepts. Major data collectors, like the public sector, are encouraged to openly disseminate their data. Open data initiatives have shown to increase cooperation and creativity (Hofmokl, 2010; Vetrò et al. 2016). Not only the public and expert data providers have increased their open data dissemination over the past years; the source for open data can often be found at communities and citizens all over the world (McCallum et al. 2016; MissingMaps, 2017). The most universal open data platform for geographic information is OpenStreetMap (OSM) and offers the possibility for data to achieve its greatest impact (McCallum et al. 2016). All these sources of open data contribute to the humanitarian data environment, also known as the humanitarian data ecosystem (Heimstädt, Saunderson & Heath 2014). The Humanitarian Data Exchange (HDX) developed by the United Nations office for Humanitarian Affairs (UN-OCHA) is an example of an initiative that takes advantage of the current data environment. HDX (2017) describes humanitarian data as follows: i) data on the context in which a humanitarian crisis is occurring ii) data on the people affected by the crisis and their needs and iii) data on the response by organisations and people seeking to help those who need assistance. Since vulnerability data is often limitedly available, these platforms form a potential for the use of preliminary open data sources. However, the open data available for development regions often come with their own limitations, like lacking metadata or standardization issues. These data should therefore be used with caution when performing analysis or when redistributing (Vetrò et al. 2016).

4.2 Hazard Vulnerability

Vulnerability can be described as *"the characteristics and circumstances of a community, system or asset that make it susceptible to the damaging effects of a hazard."* (UNISDR, 2009 p.30). This definition identifies vulnerability as a characteristic of the element of interest; a community, system, or asset which is independent of its exposure. However, the word vulnerability is often used more broadly to include the elements of exposure. Currently, a new version of the UNISDR definition is being discussed; draft documents include the following definition quoted in the Sendai framework adopted in Japan in 2015 p.10: *"Vulnerability is the conditions determined by physical, social, economic, and environmental factors or processes which increase the susceptibility of a community to the impact of hazards"*.

Vulnerability is caused by many factors; vulnerability is not a natural phenomenon, it is the human dimension of disasters and the entire range of physical, socio-cultural, institutional, political, economic, and environmental factors that shape lives and create their environment (Cardona et al., 2012). Vulnerability is usually associated with poverty; however, it can also arise when people are isolated, defenceless or facing stress (UNISDR, 2015).

Examples of physical vulnerability may exist of poor design or weak construction of the built environment, unstable conditions, close proximity to hazards, and fragile unprotected houses. Social and institutional vulnerability may include the lack of public awareness and information, lack of preparedness measure, low status in society, unequal gender relations, few decision-making possibilities, oppressive formal and informal institutional structures, and political, economic, and social hierarchies. Numerous social characteristics have been logically and empirically connected to human vulnerability. Insights into social vulnerability characteristics is beneficial for emergency preparedness, because social vulnerability plays a role in all the phases of disaster management; from mitigation, response, to recovery (Flanagan, 2011). Economic vulnerability can be described as the absence of productive assets, limited income earning opportunities, poor pay, single income revenues, no savings and insurance. The vulnerability of communities can furthermore increase by the disregard of environmental management (Wisner et al., 2012).

The parameters that influence vulnerability can be determined through a pre-defined framework. However, vulnerability differs greatly within a community, country or region, and over time. Thus there are many vulnerability parameters to be considered. Distinction can be made between hazard-independent vulnerability and hazard-dependent vulnerability; hazard-independent parameters should be considered in relation to any type of hazard and hazard-dependent parameters should be considered for a single natural hazard or for several hazards (Birkmann, 2007). Moreover, the parameters differ for individuals or households, administrative communities, countries, regions, or cultural communities. All these parameters can potentially determine vulnerability (Birkmann, 2007).

The concept of vulnerability related to hazards has widened over the years, *(Figure 4.1)* the understanding of the concept is important to be able to measure it in a similar manner for different regions and/or countries.

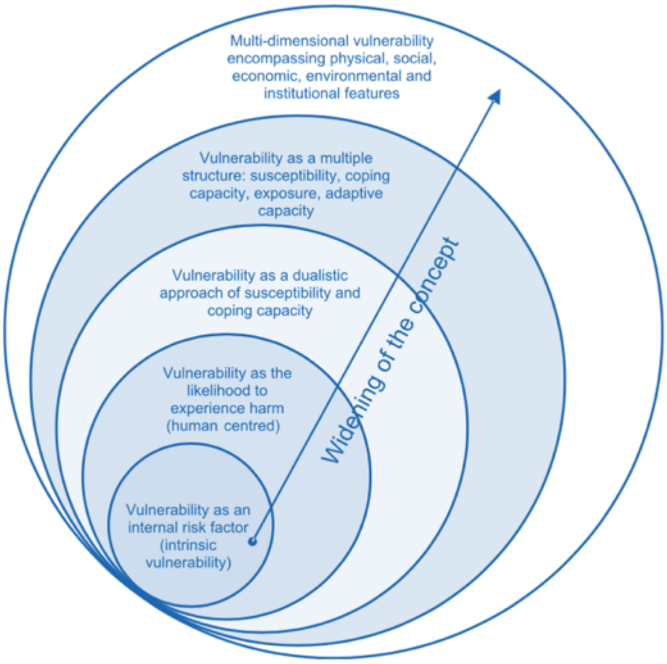

Multi-dimensional vulnerability encompassing physical, social, economic, environmental and institutional features
Vulnerability as a multiple structure: susceptibility, coping capacity, exposure, adaptive capacity
Vulnerability as a dualistic approach of susceptibility and coping capacity
Vulnerability as the likelihood to experience harm (human centred)
Vulnerability as an internal risk factor (intrinsic vulnerability)
Widening of the concept

Figure 4.1: Key spheres of the concept of vulnerability. Source: Birkmann, 2005

The ability to measure vulnerability has grown in popularity as a method to promote disaster resilience within communities (Kasperson et al., 2005). In the last decade, multiple NGO and GO initiatives have been initiated to assess the vulnerability of communities as a means to identify the risk of disasters (UNU-EHS, 2006; IFRC, 2006). Despite the differences in the conceptualization of vulnerability, several common elements can be found. One of the most important elements is the use of vulnerability assessments to identify hazard sites, which can act as the basis for pre-impact and mitigation planning (Brooks et al., 2005; Cutter et al., 2008; O'Brien et al., 2004). For this research, which is in line with the VCA outcomes of the IFRC and most recent conceptualization of vulnerability, the 'Disaster Risk Reduction' (DRR) framework of hazard-independent and social vulnerability is used in accordance with methodologies described by UNISDR (2015), and INFORM (De Groeve, Poljanšek, & Vernaccini, 2015). The framework integrates the systems sensitivity and the adaptive capacity to form hazard-independent vulnerability (Gallopin. 2006). The hazard characteristics, the exposure, otherwise described as 'biophysical vulnerability' and the hazard-independent vulnerability together, form the risk or potential for disaster. Vulnerability is conceptualized as the pre-event, inherent characteristics or qualities of social systems that create the potential for harm. It is a function of the sensitivity of the system (degree to which people and places can be harmed) and the adaptive capacity of the system (the capacity to adapt and respond to hazard threats) (Cutter, 2008) this form of vulnerability is often referred to as 'social vulnerability', as it is hazard independent and it does not take exposure into account (***Figure 4.2***). Although vulnerability is a dynamic process, for the purpose of measurement it will be viewed as a static phenomenon. Within this research, it is realized that it is important to set clear boundaries to what concepts, semantics and definitions are used within the research. Subsequently, it is realized that there is a wide variety of vulnerability frameworks described within different science disciplines, from social sciences to ecology. A critical review of Fekete (2012) on the quality of vulnerability indices shows the most important pitfalls within the development of vulnerability indices. The most prominent issue is how policy-makers, media, and public understand and interpret the result of the numbers, tables, and maps; therefore, the visualization and the path towards presenting the data is taken into account within this research.

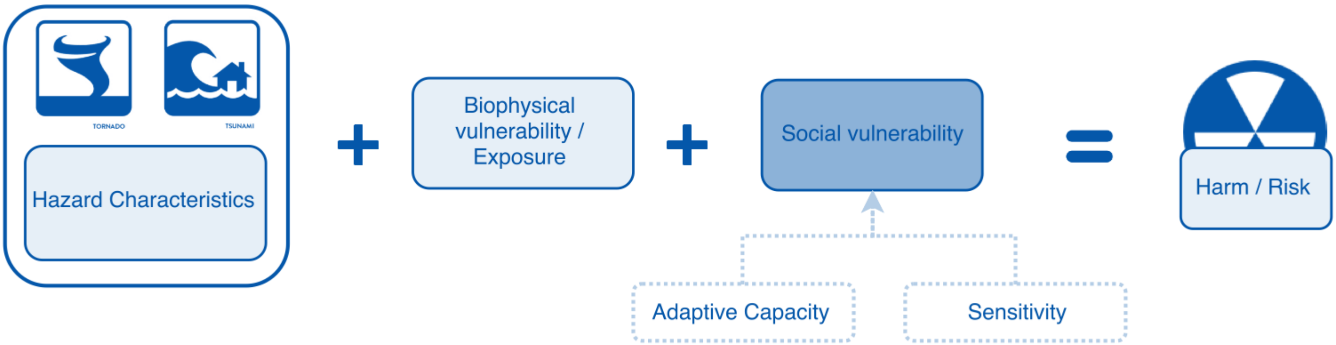


Figure 4.2: Identifying disaster risk within the DRR, UNISDR, and INFORM framework.
Risk = f (Hazard Characteristics, Exposure, (Social vulnerability= f(sensitivity, adaptive capacity))
adapted from Cutter, 2008.

4.2 Vulnerability indices

To understand the processes that cause vulnerability and the indicators that can be used to measure the difference in hazard susceptibility, the assessment of vulnerability are increasingly conducted. The key drivers of vulnerability are often identified through qualitative methods, which are instrumental for identifying the social constructs of vulnerability on a local level (Birkmann, 2006). The knowledge gained through these qualitative methods are then applied in quantitative methods to develop indicators used to compare the geographic and time trends of vulnerability. The composite of these indicators are commonly used to create vulnerability indices. The VCA tool, discussed in *Section 3.2,* is a widely used method within the humanitarian sector, more specifically within the IFRC. The VCA toolbox is evaluated as possible source of the SoVI for this research. As the results were difficult to obtain and very qualitative in nature, it was chosen to not take this source into consideration for the purposes of this research.

Several vulnerability indices have been created for different countries globally, including Malawi: the vulnerability & hazards tool for Malawi (RCMRD, 2015). However, as the conceptualization of vulnerability shows to be very diverse within current literature and disciplines, both in the theoretical underpinning and practical application of vulnerability. This should be taken into account when interpreting such indices (Adger, 2006; Green and Penning-Rowsell, 2007; Manuel-Navarette et al., 2007; McLaughlin and Dietz, 2008; Polsky et al., 2007). In *Figure 4.1* the key spheres of the concept of vulnerability are shown. The interpretation of a created vulnerability index or map is highly dependent on the choices made in presentation and visualization of the data. Subsequently, the meta data is often missing, which makes the interpretation of index based visualizations problematic (e.g. MASDAP, the open spatial data portal for Malawi).

An often mentioned approach to quantifying vulnerability is to define a set or composite of proxy indicators (Antwi-Agyei et al., 2012; Eakin and Luers, 2006). A common practice is to assess vulnerability by estimating indices or averages for the selected indicators (Gbetibouo, et al., 2010). Many aspects of data aggregation or the composition of indicators are standard challenges for researchers (Fekete, 2012). Indicators are however, often limited by the amount of available data. The USAID Food Emergency Warning System (FEWS, n.d.) is an example of a program which uses indices, calculated as averages or weighted averages of selected variables, to assess vulnerability to food insecurity in a variety of regions in Africa. In this case, vulnerability is a composite of data from different areas; 'crop risk', 'income risk', and 'coping strategies'. These indices are often hard to interpret and follow the logic and choices of the researchers developing the index (Fekete, 2012).

A way to overcome the conceptualization of different vulnerabilities within the DRR realm, social vulnerability indices are more frequently developed over the past decade as a quantitative measure of the social dimensions of natural hazards vulnerability. Social vulnerability is recognized as being critical to understanding natural hazards risks and in developing effective response capabilities (Blaike et al. 2014). This is in line with the developments regarding disaster understanding; It is more often believed to be the result of the social and demographic characteristics of communities, rather than solely from the interaction of physical and built environment systems (Tate, 2012).

The development of a Social Vulnerability Index (SoVI) by government organisations like the American Centre for Disease Control (CDC) and the Agency for Toxic Disease Registry (ATSDR) is described by Flanegan et al. (2011). Several socio-economic and demographic factors that affect the resilience of communities is taken into account. This SoVI development is based on the statement that studies have shown *"that in disaster events the socially vulnerable are more likely to be adversely affected, i.e. they are less likely to recover and more likely to die. Effectively addressing social vulnerability decreases both human suffering and the economic loss related to providing social services and public assistance after a* *disaster."* (Flanegan et al. 2011 p.12). The SoVI is created from 15 census variables and is specifically designed to identify areas in most need of assistance in emergency management.

The use of a social vulnerability index is believed to be applicable in the case of the vulnerability for Malawian communities to different types of hazards. Within Malawi a wide variety of natural hazards occur every year, the seasonal floods and droughts, but earth-quacks and crop pests as 'Army Worm' are becoming more common (Jinon, 2017). The remoteness proxy is developed to produce accurate social vulnerability scores for areas that have not been assessed, to fill in the vulnerability information gaps. The social vulnerability can be predicted for un-assessed areas by correlating the remoteness indicators to the existing data on the social vulnerability index.

The vulnerability index from the Regional Centre for Mapping of Resources for Development (RCMRD) is used to establish the SoVI for Malawi in this research. The vulnerability index of RCMRD is an averaged composite of sensitivity, adaptive capacity, and exposure indicators. To comply to the current DRR, UNISDR, and INFORM framework of the concept of social vulnerability where it is not hazard specific, the exposure component of the RCMRD index is not taken into account. Exposure is related to hazard specific components, and is therefore not useful for the purpose of the hazard-independent social vulnerability concept used in this research. The SoVI is composed of the average composite of sensitivity and adaptive capacity to form a score for social vulnerability for the Malawian Traditional Authorities.

4.3 Remoteness

Remoteness is used in several forms over a wide variety of fields. For remoteness several indices are currently in use in developed countries. For Australia the Accessibility and Remoteness Index for Australia (ARIA) is currently in use. ARIA assesses the level of government and private sector services available to regional and remote parts (Taylor & Sekkar et al., 2006). It uses a network analysis to calculate travel times from smaller towns to major towns within Australia. This index may show to be useful in the identification process of remoteness indicators. It shows the different indicators used to form a 'remoteness' score. The description of accessibility indices goes back to the late 1950s, where Hansen (1959) introduces the accessibility index from one location to a set of other locations. A widely accepted narrative is that many people in remote regions are chronically disadvantaged and therefore are among the most vulnerable to impacts of climate change; mainly in the form of natural hazards (Maru et al., 2014). In the study by Maru et al. (2014) the link between remoteness and vulnerability is researched through the use of three case studies, all within different regions, in a different stage of development. Different indicators are used to define remoteness of people and communities. Among the developed countries, Australia's ARIA index is referred to, which uses the density of the road network to identify the most remote areas. Remoteness within the Botswana context is often described using different indicators; it is found that these communities often live in small settlements that fall outside the coverage of basic services; communication, health, education, road transport, and retail and marketing facilities, which are described by Odysseos (2011). For the Brazilian context, indicators like access to clean water, health infrastructure, and other basic infrastructural services. Maru et al. (2014) argue that all the remote areas that were part of their studies were marginalised due to extreme climatic events, like floods and droughts. These areas have been neglected in regard to economic or environmental subsidies, and this reflects in the limited road density found in these remote areas. The link between remoteness and vulnerability is often described related to accessibility to different elements: agency, resources, and services. In the research, the disadvantages of remote areas are identified and related to vulnerability through indicators that fall within these three themes of remoteness. At this point in time, there is a limited amount of literature quantifying the empirical relationship between remoteness and vulnerability.

4.4 Accessibility

The creation of isochrones is often applied in accessibility studies (Li et al., 2011). The understanding of isochrones creation increases, when a short introduction to accessibility is given. Isochrone-maps are only used in a specific area of accessibility; this part of accessibility studies will be briefly touched upon.

The term 'accessibility' originates from the transportation planning literature, but is increasingly interdisciplinary. It is widely used within the scientific context of urban geography, regional science and geographic analysis, and spatial economics (Cascetta et al. 2016). The definition of accessibility within the scope of this research is coined by Bertolini et al. (2005 p.3) *"The amount and the diversity of places of activity that can be reached within a given travel time and/or cost"*. Often mentioned in accessibility literature, is that the places reached, can be replaced by the number of people or settlements reached. This fact is the inspiration of the methodology used within this research, where human settlement data is used within the GIS based isochrones creation. In this particular part of the GIS analysis human settlements, referring to household level data, is used to indicate the remoteness of a region or area within Malawi. This consequently means that remoteness data can be aggregated on any granular geographic level for a country.

A distinction can be made between passive and active accessibility. Passive accessibility is referred to by Cascetta et al. (2016) as the amount of effort needed to reach a certain activity. Active accessibility is aimed at the individual; the characteristics of an individual are taken into account, for example the ability of the people studied within the accessibility study. Isochrones are often regarded as passive accessibility. However, in this research the characteristics of individuals, with regard to traveling ability and speed are taken into account, see Section 5.4.

4.5 Machine learning for proxy development

In the case of social vulnerability data for communities, the set of variables to rely on is often limited. The development of a proxy needs to overcome this issue by providing an accurate alternative. Machine learning techniques are relevant in the sense that they explore the study and construction of algorithms that can learn from and make predictions on data. Machine learning techniques use data studying and algorithm construction to learn from and predict on data sets. Machine learning relies on predictive accuracy of the model, rather than on the static data modelling side of statistics, Breiman (2001) states that machine learning techniques are making progress due to this fact. There are two ways within the statistical modelling community to reach conclusions on data; "*One assumes that the data are generated by a given stochastic data model and then the assumptions for the model are tested. The other uses algorithmic models and treat the data mechanism as unknown*" (Breiman, 2001). The latter describes machine learning, the use of such algorithms overcomes the use of static programming by iterating data-driven predictions and decisions. In the case of a proxy indicator, the mechanism of the predictions is often not clear. For these complex prediction problems, the use of a pre-set data model is not sufficient (Breiman, 2001). Machine learning techniques overcome this by using the predictive accuracy of the data as a driver. Machine learning techniques contribute to the proxy development in the humanitarian data environment as it is not affected by as many model assumptions as linear regression techniques and results a higher predictive accuracy in a faster and more reproducible manner.

Within machine learning, there is a distinction between unsupervised and supervised learning techniques. During this study, supervised learning techniques will be used. In supervised learning, input data is referred to as training data, and has a known outcome; in this study the outcome is the social vulnerability index for Malawi. Within machine learning the data set is divided in learning data and test data; the results are known, hence: the test data and the predictions can be interpreted for its accuracy. For a multivariate regression problem as explored within this research, multiple algorithms are suitable; Linear Regression, Decision trees, Neural Network and Random Forest (Rohrer, 2016). Different considerations for choosing a certain algorithm are accuracy, training time, and linearity (Rohrer, 2016). As the remoteness indicators created through GIS analysis of open data are context specific and reliant on the available data sets, these may change when applying the algorithm to other countries. The linearity of the data set will therefore often change. The data set will always be relatively small and training time is not likely to become a constraint. A classic linear regression and a Random Forest regression model will be explored during this study.

The use of algorithmic modelling is a fast way to assess a set of data; it is possible to iterate multiple models with different parameter settings. This enables a relatively fast and simple assessment of the minimum amount of data necessary to accurately predict the social vulnerability index score. This may subsequently show to be useful when applied to different countries. When vulnerability data differs per country or region, semantically or otherwise, the model can learn from the new data, and consequently improve itself.

5 Analyses and Results

This chapter will provide a further elaboration on the presented methods in chapter 3. The GIS analysis and the machine learning analysis are described. All the research questions will be subsequently answered to accurately derive conclusions for the research.

5.1 Social vulnerability – Y-variable

The creation of the SoVI is explained

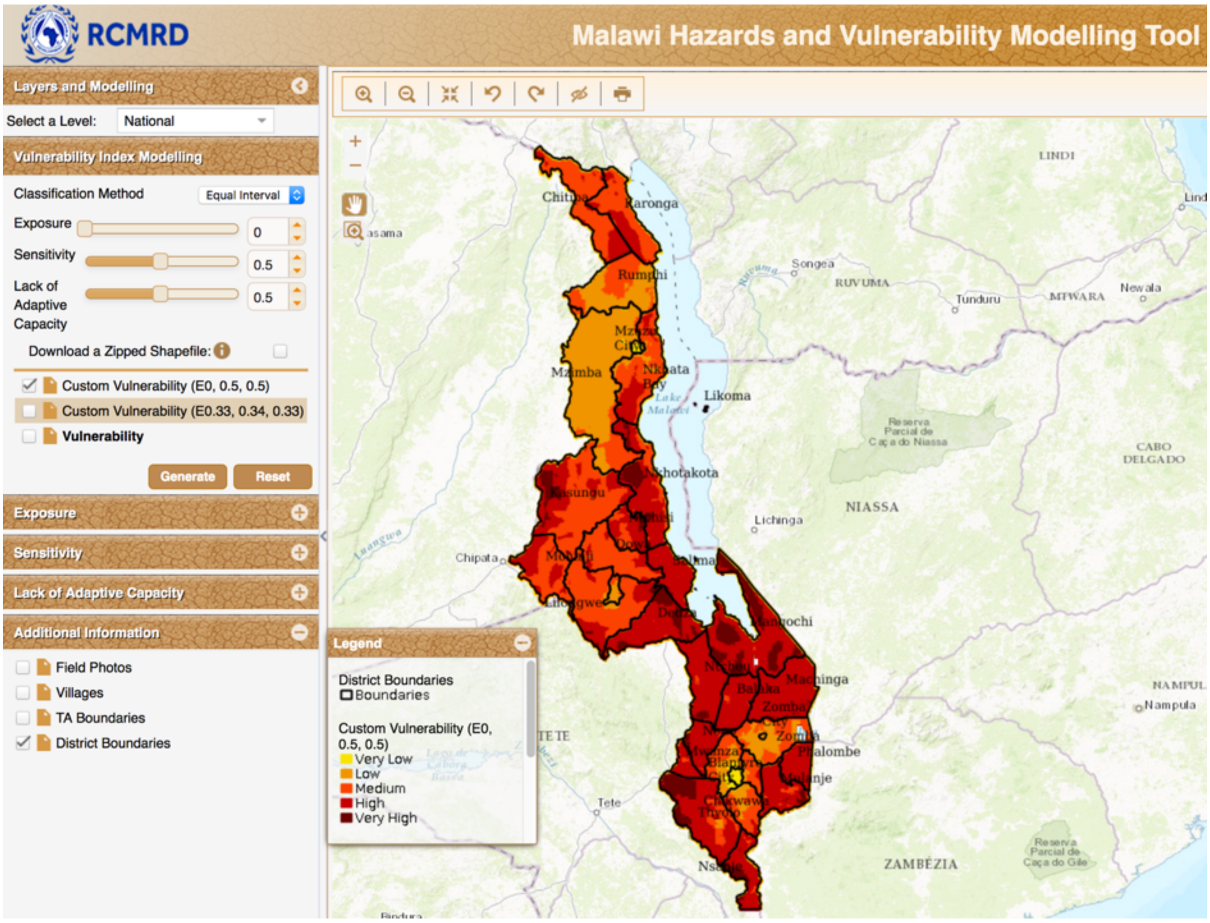
SoVI as the Y-variable

For the Y-variable, the RCMRD data behind the vulnerability tool for Malawi will be used. The indicators in the 'sensitivity' and 'adaptive capacity' categories for vulnerability in Malawi are depicted in ***Table 5.1.***

Table 5.1: RCMRD Data creating the social vulnerability index available at the RCMRD Geoportal (RCMRD, 2015)

Data set	Category	Type	Developed by	Source	Year
Soil organic carbon	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA).	ISRIC	1950 - 2005
Malawi National Poverty Levels	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA).	AfriPop	2010 - 2011
Malawi National Population Density	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	Landscan	2012
Malawi National Infant Mortality Rate	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	National Statistics Office (census)	2008
Malawi National Malaria Susceptibility	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	MAP	2010
Malawi National Female Headed Households	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	National Statistics Office (IHS3)	2011
Malawi National Building Material	Sensitivity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	National Statistics Office (IHS3)	2011
Malawi National Market Accessibility Time	Adaptive Capacity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	Joint Research Centre (JRC)	2008
Malawi National Irrigated Areas	Adaptive Capacity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	FAOSTAT	1990 - 2000
Malawi National Literacy Levels	Adaptive Capacity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	National Statistics Office (census)	2008
Malawi National Health Infrastracture Index	Adaptive Capacity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	MASDAP	2015
Malawi National Education Level of Mother	Adaptive Capacity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	National Statistics Office (IHS3)	2011
Malawi National Anthropogenic Biomes	Adaptive Capacity	Raster data set	RCMRD & Malawi Department of Disaster Management Affairs (DoDMA)	CEISIN	2001-2006

The vulnerability tool explicitly divides the indicators for vulnerability in three categories, as 'exposure' is taken into account for the Malawi vulnerability tool (***Figure 5.1).*** The tool is developed in the light of vulnerability to climate change and therefor, uses the 'global change' framework for vulnerability. In this specific case, this means that exposure is a third separate category and is filled with 'droughts', 'forest fires', 'precipitation', 'temperature trend', and 'flood frequency'. These indicators are all hazard-dependent variables and will be excluded from the Y-variable; 'SoVI'.



RCMRD Malawi Hazards and Vulnerability Modelling Tool

Layers and Modelling

Select a Level: National

Vulnerability Index Modelling

Classification Method: Equal Interval

Exposure: 0

Sensitivity: 0.5

Lack of Adaptive Capacity: 0.5

Download a Zipped Shapefile: □

☑ Custom Vulnerability (E0, 0.5, 0.5)

□ Custom Vulnerability (E0.33, 0.34, 0.33)

□ **Vulnerability**

Generate | Reset

Exposure

Sensitivity

Lack of Adaptive Capacity

Additional Information

□ Field Photos

□ Villages

□ TA Boundaries

☑ District Boundaries

Legend

District Boundaries

□ Boundaries

Custom Vulnerability (E0, 0.5, 0.5)

- 🟨 Very Low
- 🟧 Low
- 🟧 Medium
- 🟥 High
- 🟥 Very High

Figure 5.1. Web application of the RCMRD vulnerability tool where exposure is eliminated as an example of interactivity options the tool offers (Accessed on 25-12-2016).

All the data from RCMRD is presented in a continuous raster data set format with a resolution of 1 km. This makes the data suitable for the purpose of this study: the data is granular enough to use on a TA level. A matrix is developed following the P-codes of the TA admin boundaries. Each administrative area on the globe can be identified using a P-code, which becomes longer for lower administrative areas. An example for Malawi: the country 'Malawi' is identified by 'AFRMWI', the region 'Southern' as 'AFRMWI3', the district 'Balaka' as 'AFRMWI312', and the traditional authority 'Msamala' as 'AFRMWI31201'. A small section of the p-coded matrix is shown in *Table 5.4*. The data matrix is created in a .CSV file and can be added to an administrative TA-level Table so that it can be filled with the raster data within a GIS environment. Next to the raster data, the .CSV data format of the vulnerability tool is obtained directly from the lead researcher, Dennis Macharia, of the RCMRD vulnerability tool. A common mapping scale of 1km by 1km was used for the extraction of indicator values from the raw data files (rasters), a 1x1km centroid layer is used. The methodology of the vulnerability index was shared for the purpose of this research. In agreement with the RCMRD head researcher the vulnerability index based on the averaging of exposure, sensitivity, and adaptive capacity, was adapted to create the Social Vulnerability Index (SoVI). The averaging of the nominal values of sensitivity and adaptive capacity were used to create the adapted index.

5.2 Remoteness – X-variables data extraction

For the creation of relevant remoteness indicators, an extensive data search is conducted on openly available data which could potentially be of a good enough standard for analyses and assumptions. This chapter is related to sub-objective 1 as shown in the methodology scheme *(Figure 3.1)*

5.2.1 OpenStreetMap data

The basis of all the geospatial analyses conducted during this research is the road network for Malawi extracted from OSM. A general overview of the road network of Malawi is given in *Figure 5.2*. Furthermore, some amenities are extracted from the OSM data, such as city location, town location, health site location, water point data and schools. This is done to compare the OSM data with the other data sets obtained on the mentioned points of interest using other secondary open data sources. This is done to assess the data completeness and quality of the OSM data. As indicated in *Figure 5.2* some main road segments are missing when extracting the raw OSM data. The OSM road network is displayed as an overlay for the settlement data used in this analysis (HRSL) as described in *Section 5.2.3.*

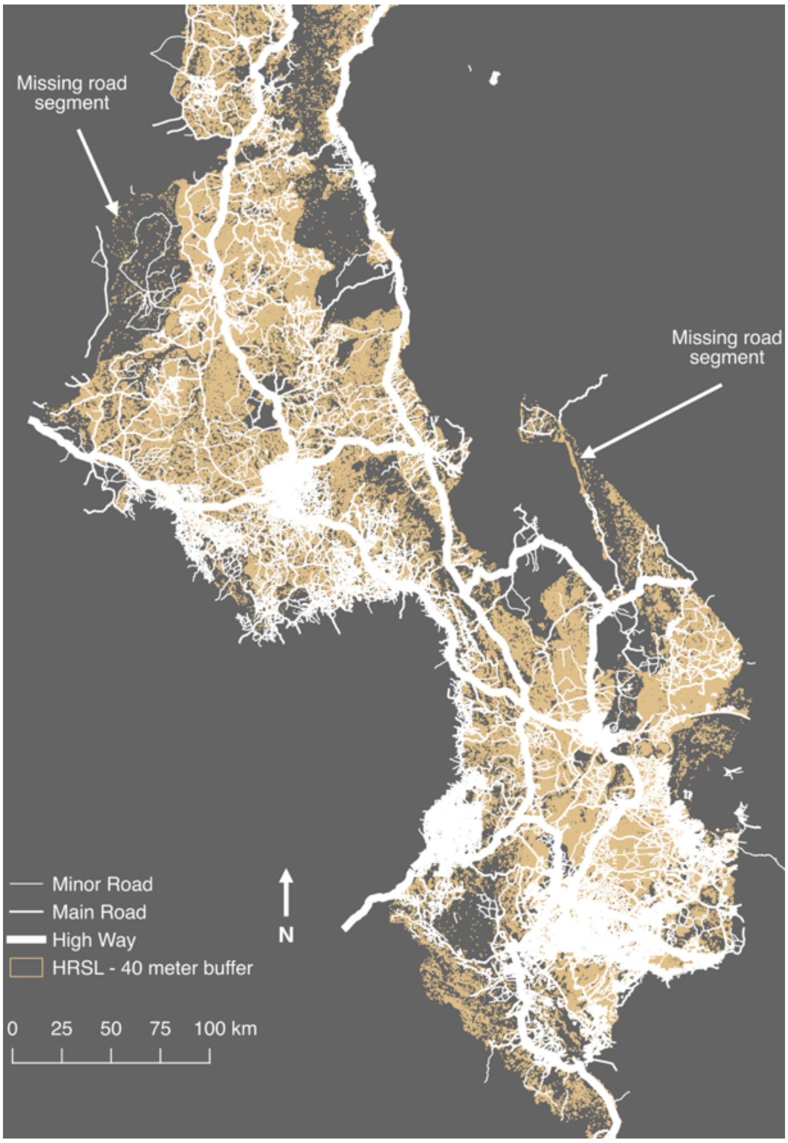


Figure 5.2 General overview of the OSM road network for Malawi (Latest abstract on 05-03-2017*)*

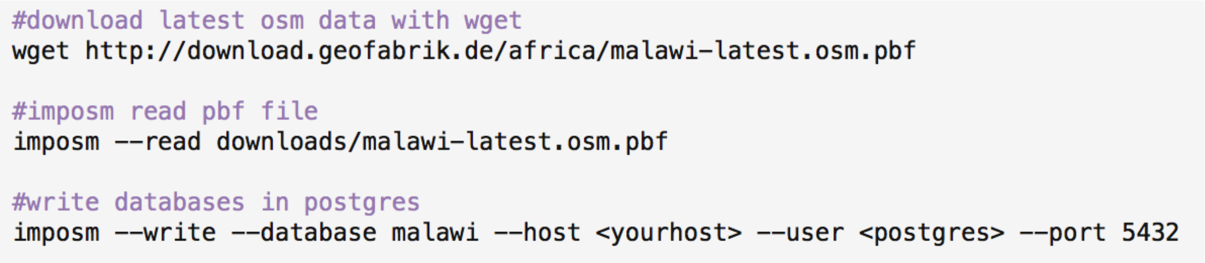
This openly available data will be one of the drivers for the creation of the remoteness indicators. Analysis on the data are executed using PostGIS and PgRouting to create relevant data sets of remoteness indicators. Each time an analysis is conducted, the latest 'nightly dump' of OSM is extracted from Geofabrik. Geofabrik is a web portal where OSM data for different geographic regions is prepared in a .pbf file format. For all the OSM data for Malawi, this is scripted into the analysis, and is therefore done automatically each time the script is run.

Data extraction and filling a database

As indicated in the methodology scheme in *Figure 3.1-A,* the extraction of the OSM data is done through several command line tools. The command-line tools used are discussed in the following section. To command these tools, scripts are used. This makes repeating the analysis more convenient. For the purpose of this study, it is chosen to extract OSM data through the use of these scripts to enhance the replicability. For the extract of a different country, solely the name of the country needs to be adapted and it will run automatically.

Imposm

The command line tools rely on the OSM data portal geofrabrik.de, where every night an extraction of the most recent OSM data is conducted. The command line operators are displayed in *Script 5.1*. Wget, a download tool for many repositories, is used to extract the .pbf file from the geofabrik data portal. The .pbf file is the packaged file format for .osm when extracted. The .osm file is a .xml file format. Imposm is used to read the .pbf file and write it into a preferred database. In this case the Postgis database is named 'malawi'.



```
#download latest osm data with wget
wget http://download.geofabrik.de/africa/malawi-latest.osm.pbf

#imposm read pbf file
imposm --read downloads/malawi-latest.osm.pbf

#write databases in postgres
imposm --write --database malawi --host <yourhost> --user <postgres> --port 5432
```

Script 5.1 Using WGET to download latest OSM data and Imposm to read and write to the Postgres database.

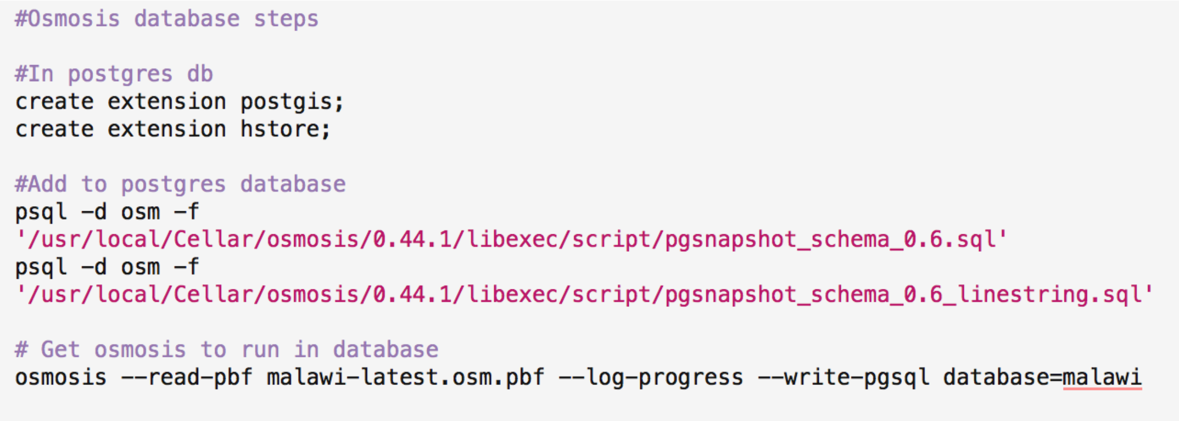
Imposm writes the major components of OSM data into a Postgres database. The database is enabled with a Postgis extension. PostGIS is a spatial extension for PostgreSQL, it adds support for geographic types and functions (PostGIS, 2016). This makes it possible to conduct Postgis queries and subsequently, GIS analyses on the data. The Imposm components are shown in *Table 5.1*.

Table 5.2: PostGIS database using Imposm for Malawi from extract: http://download.geofabrik.de/africa/malawi-latest.osm.pbf

PostGIS data set	Source
Osm_new_admin	*OSM*
Osm_new_aeroways	*OSM*
Osm_new_amenities	*OSM*
Osm_new_buildings	*OSM*
Osm_new_landusages	*OSM*
Osm_new_mainroads	*OSM*
Osm_new_minorroads	*OSM*
Osm_new_motorways	*OSM*
Osm_new_places	*OSM*
Osm_new_railways	*OSM*
Osm_new_roads	*OSM*
Osm_new_transport_areas	*OSM*
Osm_new_transport_points	*OSM*
Osm_new_waterways	*OSM*
OSM_NEW_WATERAREAS	*OSM*

Osmosis

Another command-line tool is Osmosis. The script used to fill a Postgres database is depicted in *Script 5.2.* The Postgres database needs to have the extension Postgis and Hstore loaded before this script will work and are therefore shown as the first steps in creating an Osmosis filled Postgres database. The Postgres database also needs some .sql functions loaded to function and is done in the second step. The .sql files come with the installation of Osmosis and the directory shown below is standard for Mac OS users. The final step is the actual command to run the latest OSM data into the Postgres database.



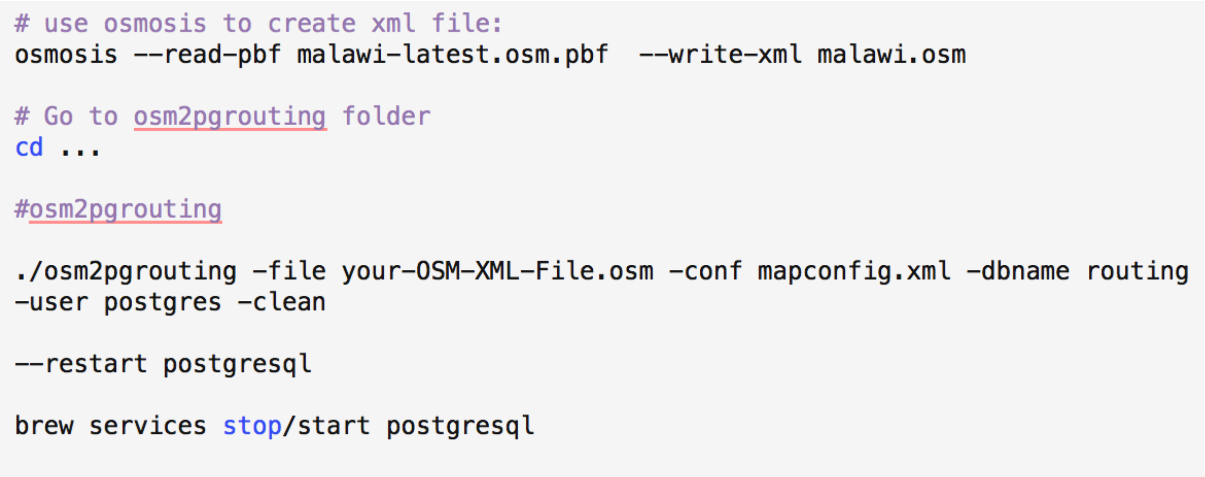
```
#Osmosis database steps

#In postgres db
create extension postgis;
create extension hstore;

#Add to postgres database
psql -d osm -f
'/usr/local/Cellar/osmosis/0.44.1/libexec/script/pgsnapshot_schema_0.6.sql'
psql -d osm -f
'/usr/local/Cellar/osmosis/0.44.1/libexec/script/pgsnapshot_schema_0.6_linestring.sql'

# Get osmosis to run in database
osmosis --read-pbf malawi-latest.osm.pbf --log-progress --write-pgsql database=malawi
```

Script 5.2 Using Osmosis to write to Postgres database

OSM2PgRouting

To write a 'rouTable' network topology Table to the Postgres database, the command-line tool OSM2PgRouting is used. The steps necessary are depicted in *Script 5.3.*

First, Osmosis is used to extract the .pbf file and write an .osm file. Osm2PgRouting is not able to read .pbf files. Secondly, to run the OSM2PgRouting command-line tool, it is necessary that it is in the same folder as where osm2PgRouting is located. Finally, the OSM2PgRouting function can be run to fill the Postgres database with a rouTable network Table compliant to the Postgis extension PgRouting.



```
# use osmosis to create xml file:
osmosis --read-pbf malawi-latest.osm.pbf  --write-xml malawi.osm

# Go to osm2pgrouting folder
cd ...

#osm2pgrouting

./osm2pgrouting -file your-OSM-XML-File.osm -conf mapconfig.xml -dbname routing
-user postgres -clean

--restart postgresql

brew services stop/start postgresql
```

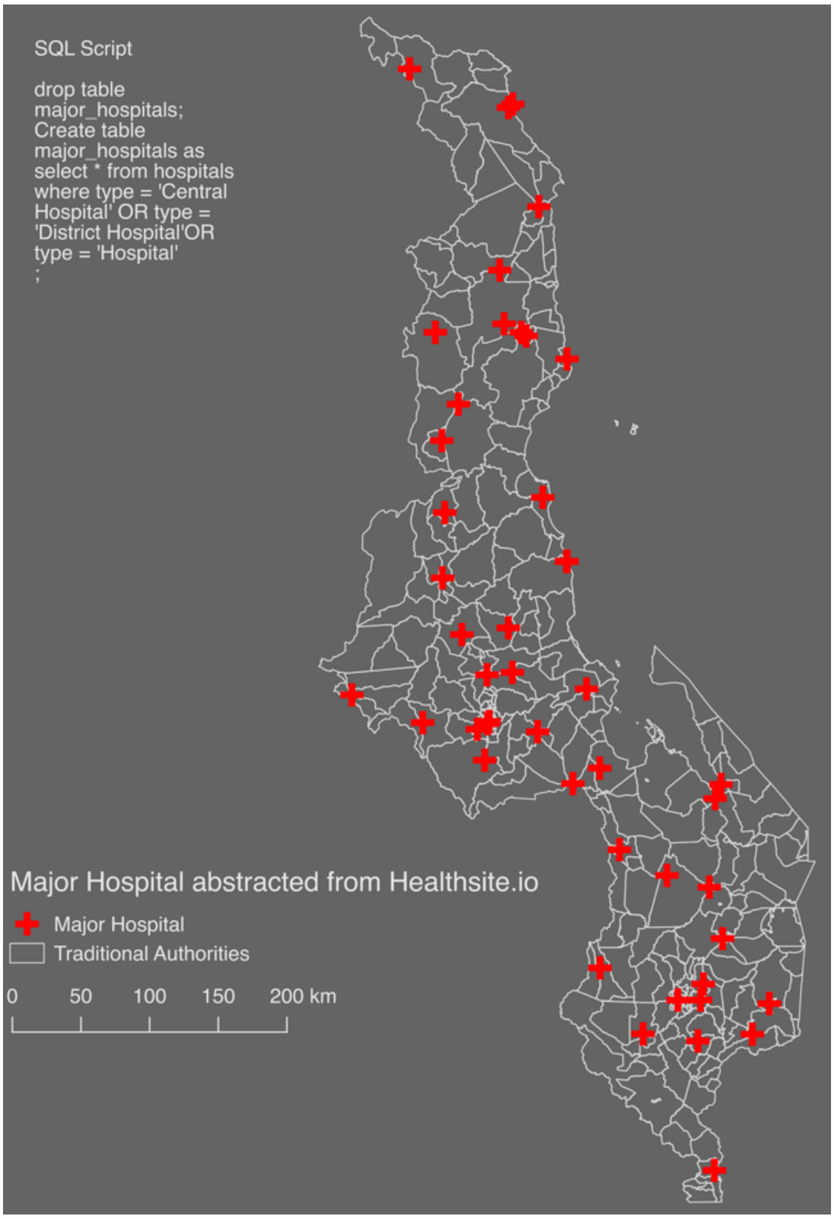
Script 5.3 Using OSM2PgRouting to write a rouTable network Table to a Postgres database

OSM2PgRouting fills the Postgis enabled database with two Tables: I) *ways,* a Table containing the network of roads with information on traveling speed, length, direction and is linked to the second Table through a coded indicator. Within the PgRouting terminology this Table is referred to as the 'edge Table'. II) *ways_vertices_pgr,* contains all the nodes or vertices on the edge Table. This table is referred to as the vertices Table. The attributes in the Table will be elaborated on in *Section 5.3*.

5.2.2 Points of Interest

The accuracy of the locational data of the points of interest from which isochrones will be calculated is essential. The points of interest chosen for this research are as follows:

I) Major Hospitals
II) Major Cities
III) Primary Schools
IV) Secondary Schools
V) Water points
VI) Trading Centres



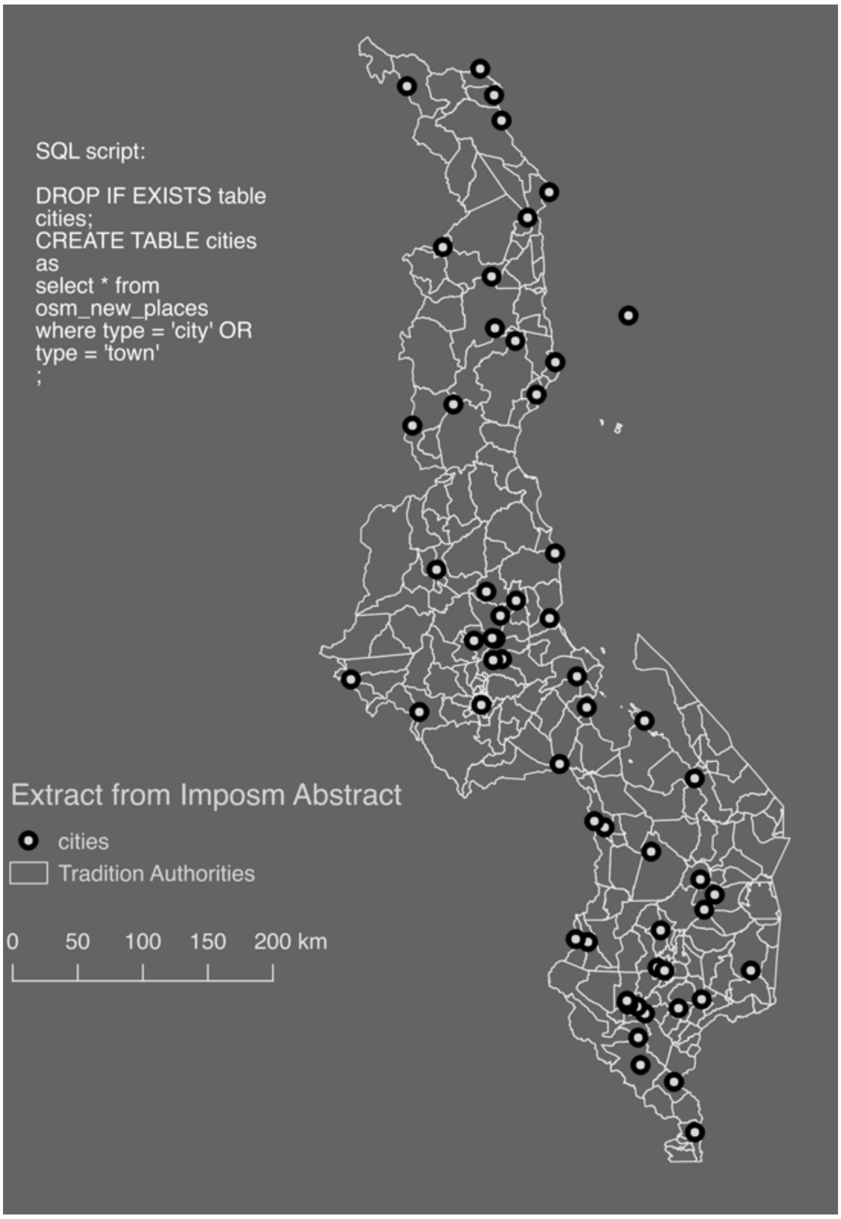
SQL Script

```
drop table
major_hospitals;
Create table
major_hospitals as
select * from hospitals
where type = 'Central
Hospital' OR type =
'District Hospital'OR
type = 'Hospital'
;
```

Figure 5.3 overview of major hospitals for Malawi,
extracted from Healthsites.io (Latest extract on 05-02-2017)

“The Sendai Framework for Disaster Risk Reduction 2015–2030 recognizes the strong connection between health and disasters and promotes the concept of health resilience throughout. Several of the seven global targets stated in the Sendai Framework are directly related to health in terms of reducing disaster mortality, the number of affected people, disaster damage to critical infrastructure, and disruption of basic services such as health facilities.” Maini et al. (2017 p.1)

(I) Major hospitals are extracted from Healthsites.io. The data set contains a large set of health sites for Malawi containing their function, working ours, number of beds, outpatient, inpatient and number of staff. The major hospitals were selected for Malawi, referred to as District Hospitals and Mission Hospitals within Malawi. These hospitals have the largest capacity, and are the go-to hospitals in case of a disaster. A spatial overview is given in *Figure 5.3*. The data set contains 46 major hospitals for Malawi. The Health sites data of HDX was explored and was similar to the healthsites.io data set. The HDX health sites data was created by UN-OCHA and stems from the year 2005. The SQL function within the overview shows how the hospitals were selected within the data set.

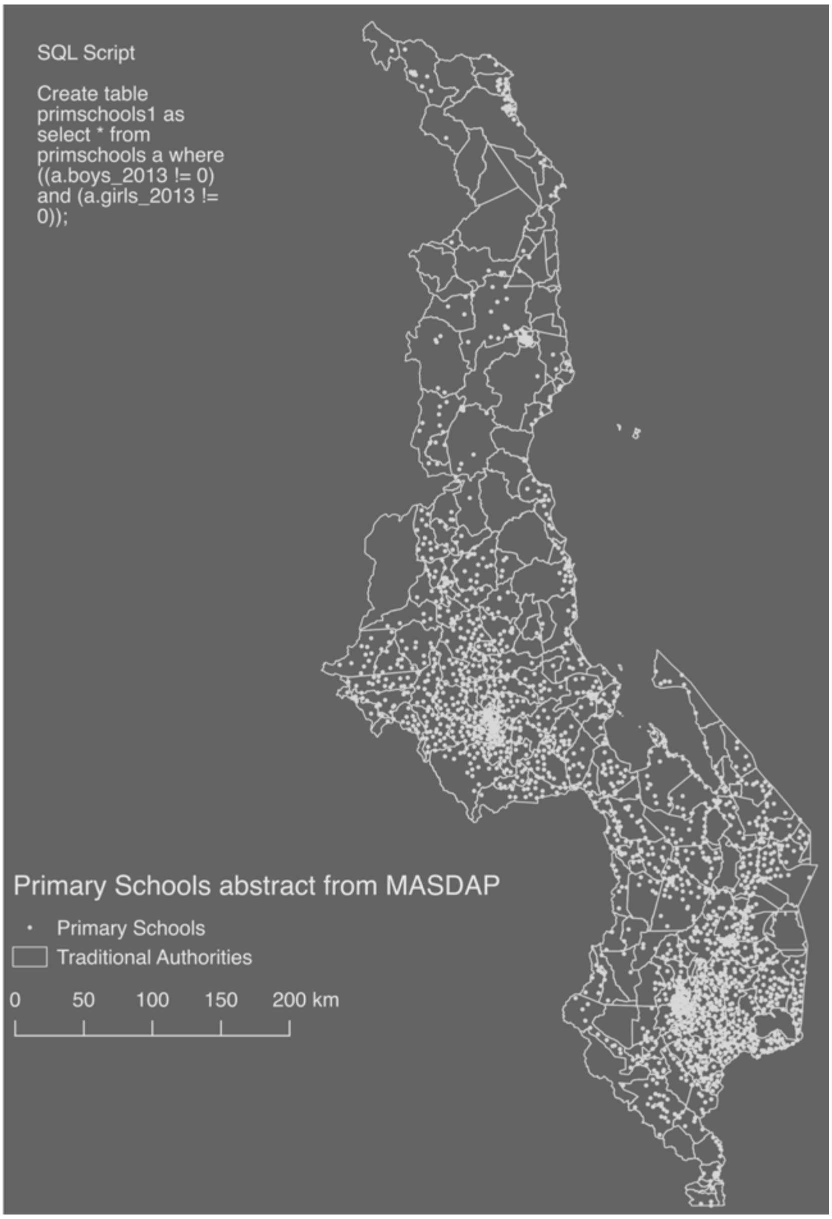


SQL script:

```
DROP IF EXISTS table
cities;
CREATE TABLE cities
as
select * from
osm_new_places
where type = 'city' OR
type = 'town'
;
```

Figure 5.4 Overview of all the major cities within Malawi.
Extract from Imposm abstract from OSM data. (Latest extract 05-02-2017)

(II) Major cities are extracted from the OSM data imported through Imposm. The Table 'osm_new_places' is used (Table 2) to abstract all the points with the attribute type as 'town' or 'city'. A spatial overview is given in ***Figure 5.4.*** The data set contain 58 major cities for Malawi. The SQL function within the spatial overview shows how the selection is made.



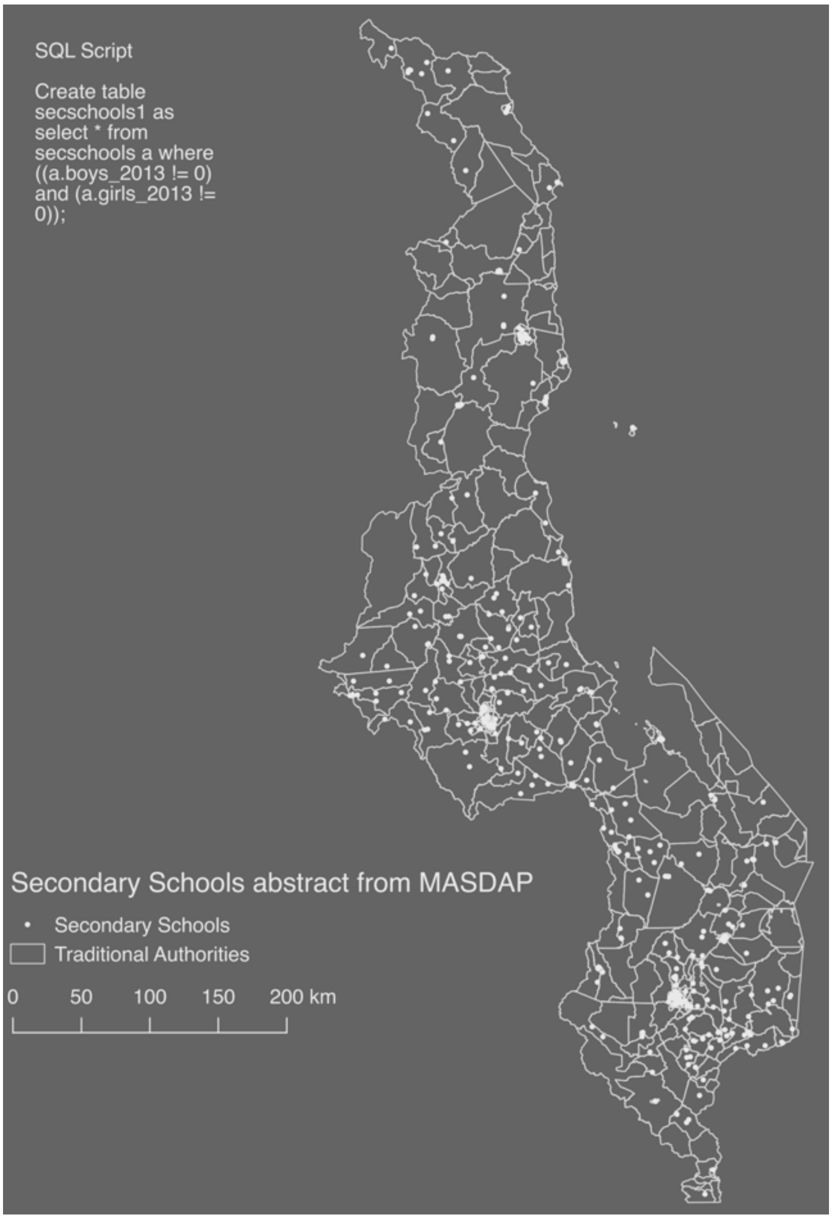
SQL Script

```
Create table
primschools1 as
select * from
primschools a where
((a.boys_2013 != 0)
and (a.girls_2013 !=
0));
```

Primary Schools abstract from MASDAP
• Primary Schools
□ Traditional Authorities
0
50
100
150
200 km

Figure 5.5 Overview of primary schools with attending children, adapted from MASDAP. (Latest extract on 05-02-2017)

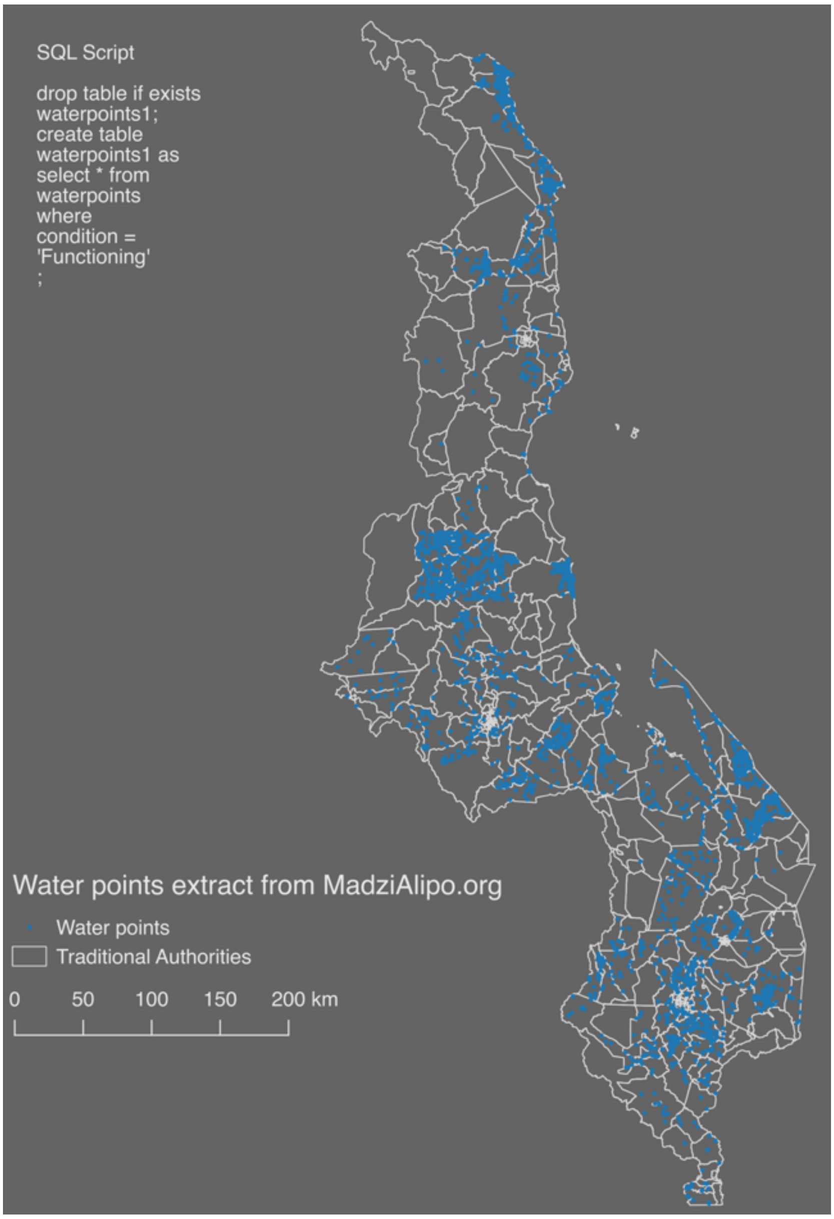
(III) Primary schools are extracted from MASDAP. The 'Primary schools' data set is created by the department of surveys and is the result of a survey in the year 2013. Through the SQL given, the schools where no children attend are left out of the analysis. A spatial overview is given in *Figure 5.5.* The data set contains 1891 primary schools for Malawi.



```
SQL Script

Create table
secschools1 as
select * from
secschools a where
((a.boys_2013 != 0)
and (a.girls_2013 !=
0));
```

Figure 5.6 Overview of secondary schools with attending children, adapted from MASDAP. (Latest extract on 05-02-2017)

(IV) Secondary schools are extracted from MASDAP. The 'secondary schools' data set is created by the Department of Surveys in Malawi in a survey conducted in 2013. Through the SQL shown in the spatial overview the secondary schools are eliminated based on school attending children. A spatial overview is given in *Figure 5.6.* The data set contains 1122 secondary schools for Malawi.



SQL Script

```
drop table if exists
waterpoints1;
create table
waterpoints1 as
select * from
waterpoints
where
condition =
'Functioning'
;
```

Figure 5.7 Overview of functioning water points within Malawi. Adapted from MadziAlipo.org (Latest extract on 08-02-2017)

(V) Water points are extracted from Madzi Alipo, a spatial data platform specifically for water point data. All the water points that are functioning for Malawi available in the data set are taken into account. The SQL script given in the spatial overview shows how the selection is done. A spatial overview is given in *Figure 5.7.* The data set contains 3805 water points for Malawi.

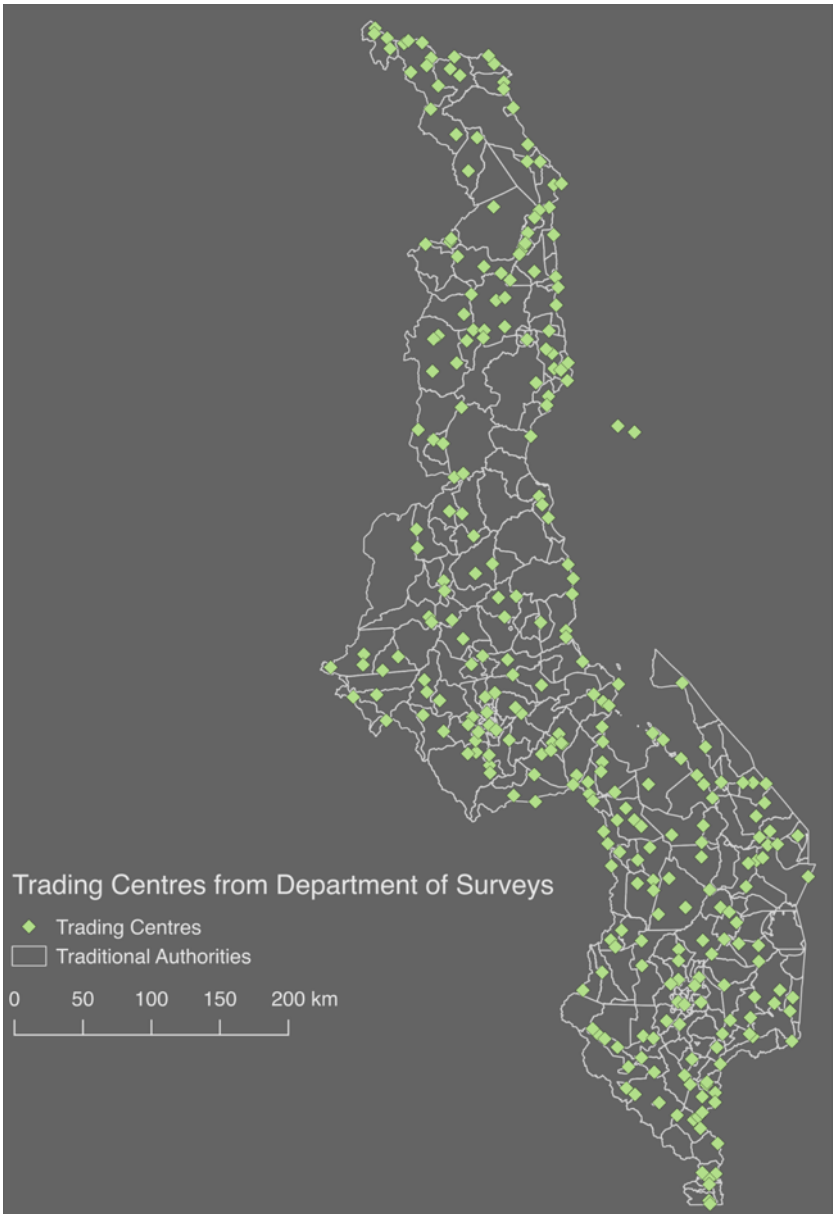

Trading Centres from Department of Surveys
◆ Trading Centres
□ Traditional Authorities
0
50
100
150
200 km

Figure 5.8 Overview of trading centres for Malawi. (Latest extract on *04-02-2017)*

(VI) Trading centres are extracted from the Department of Surveys in Malawi whilst visiting the organisation in Lilongwe, Malawi in February 2017. After the visit the data was made available by the organisation on the Malawi Spatial Data Platform (MASDAP) and is now openly available. The department was planning to upload the data to MASDAP. The data stems from the latest census data set from 2008. The data set contains 287 trading centres. A spatial overview is given in ***Figure 5.8.***

5.2.3 Human Settlement Layer

The High Resolution Settlement Layer (HRSL) is created by CEISIN and the Connectivity Lab at Facebook (CEISIN, 2017). It provides estimates of human population distribution at a resolution of approximately 30m for the year 2015. Population data is based on the most recent census data (2008) and high-resolution Digital Globe satellite imagery: at a 0.5-meter resolution. For both urban and rural areas, a detailed delineation of settlements is given (CEISIN, 2017). The number of points extracted from the HRSL count 3,047,846. According to the census data the population of Malawi is 17.2 million. This means that the average settlement point contains 5.7 people. The point extraction is done using the HRSL raster data and point extraction tools in QGIS. *Figure 5.9* shows the raster HRSL for the capital of Malawi.

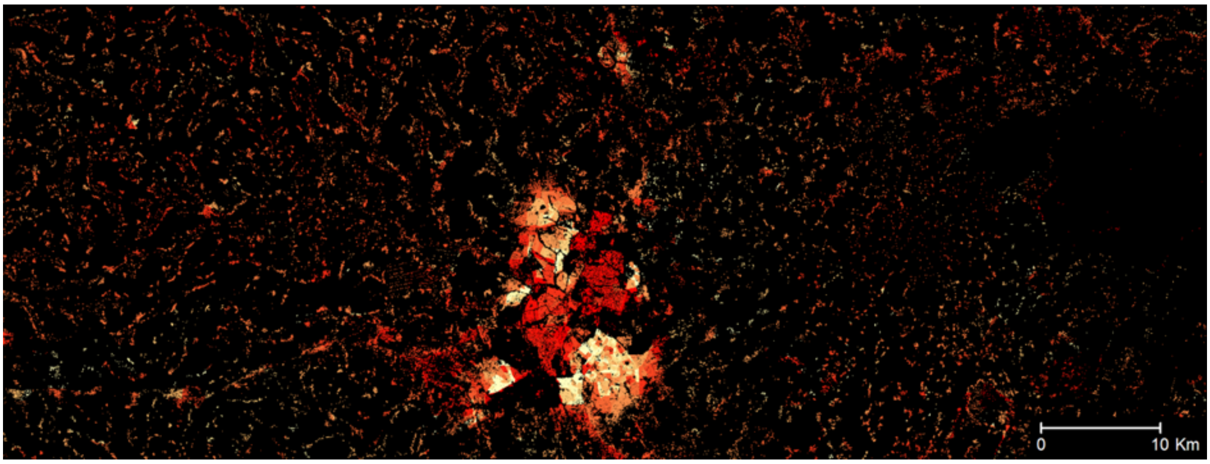


Figure 5.9 Overview of human settlements for Lilongwe, Malawi. Raster HRSL created by CEISIN/Facebook (2016)

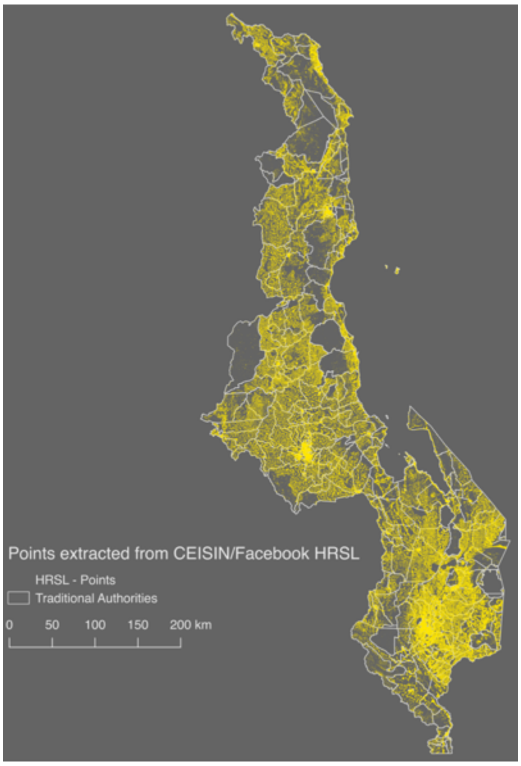


Figure 5.10 Overview of points extracted from CEISIN/Facebook HRSL (2015).

It is chosen to use the HRSL, rather than OSM buildings due to the completeness of the OSM data. For the OSM buildings data an overview is provided in *Figure 5.11*. The number of buildings in the OSM data set for Malawi is 869018; substantially less than the HRSL layer, taken into consideration that one point in the HRSL point layer may include multiple buildings, whereas the OSM building points are developed using the centroid of the polygon of the structure drawn in OSM.

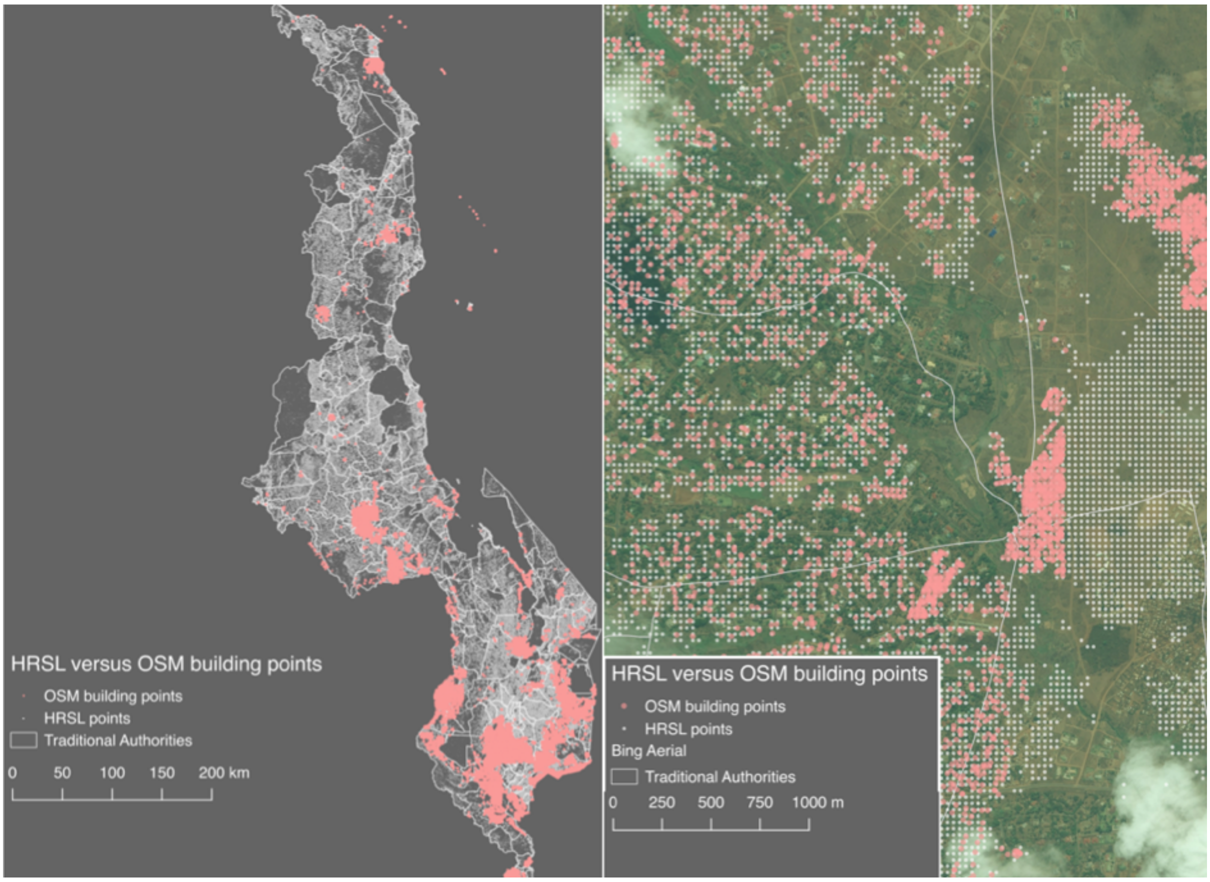


Figure 5.11 HRSL versus OSM building points. Left: overview of Malawi. Right: Lilongwe, Malawi. (OSM abstract from 08-03-2017)

The human settlement point layer is used as a continues input layer during the Postgis analysis. In Section 5.3 the use of the layer will be further elaborated on. The HRSL layer is highly accurate in the recognition of buildings and human settlements. A visual check using premium Digital Globe imagery shows that the majority of the structures are recognised through the Facebook algorithm *(e.g. Figure 5.11).*

5.2.4 Topographic roughness

The topographic feature used in this research is the topographic roughness (ruggedness) of the landscape as a measure to quantify terrain heterogeneity (Riley et al. 1999). The ruggedness function in QGIS is based on the Terrain Ruggedness Index (TRI) algorithm Riley et al. introduced in 1999. The TRI is the difference between the value of a raster cell and the mean of the 8-cell neighbourhood of the surrounding cells. The range of Riley et al. (1999) can be used to classify the results. The TRI is used as a remoteness indicator as terrain ruggedness can hinder accessibility, not only for transportation over the road, but also the reception of cellular network (Jinon, 2017).

As an input, the SRTM DEM is clipped to the administrative border of Malawi. The ruggedness layer output comes in a GeoTiff format (raster).

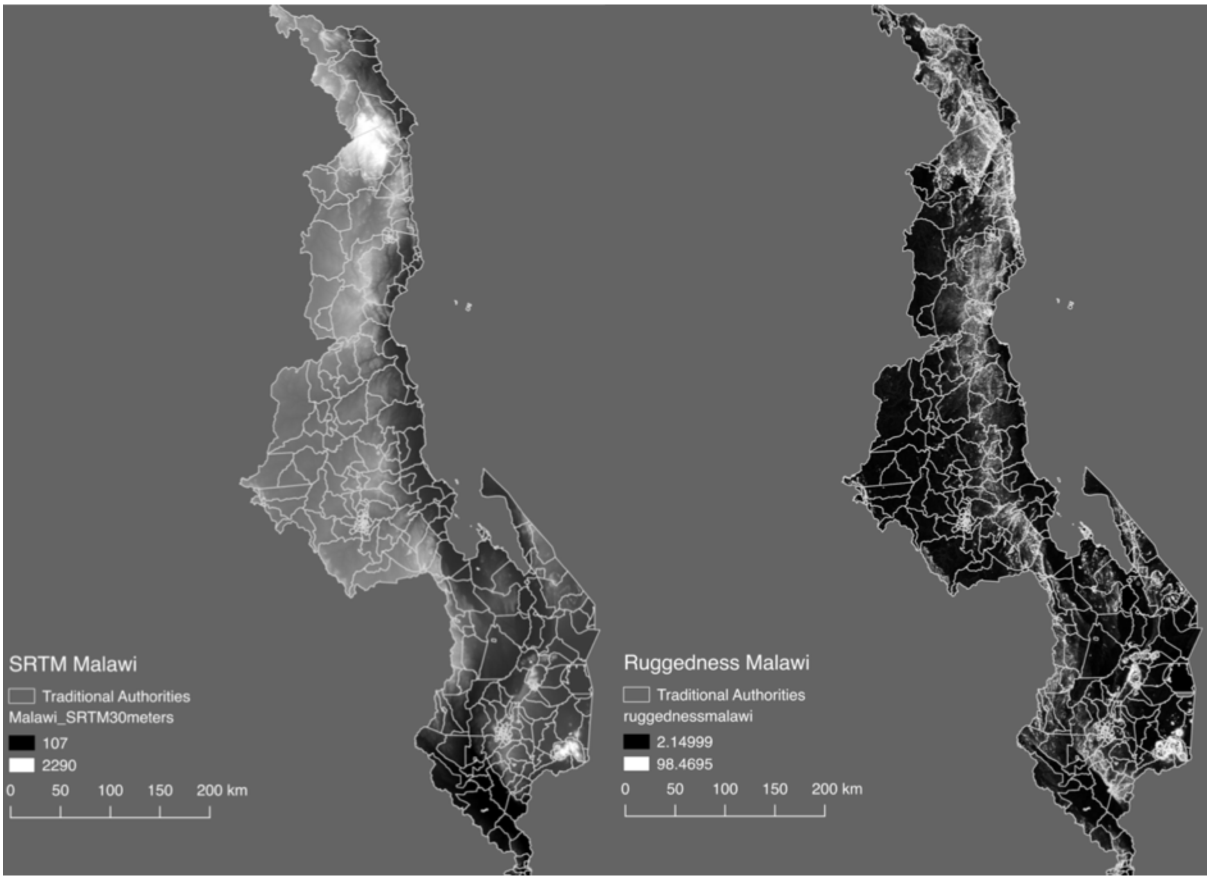


Figure 5.12 Left: SRTM Malawi Right: Ruggedness Index for Malawi

The mean value per TA is calculated for the ruggedness layer. The administrative boundaries for TA's Table is a polygon layer. The Ruggedness is a raster-based layer. For this calculation the 'zonal statistics' function within QGIS is used. This resulted in a mean score of ruggedness per TA. This is visualised in *Figure 5.13.*

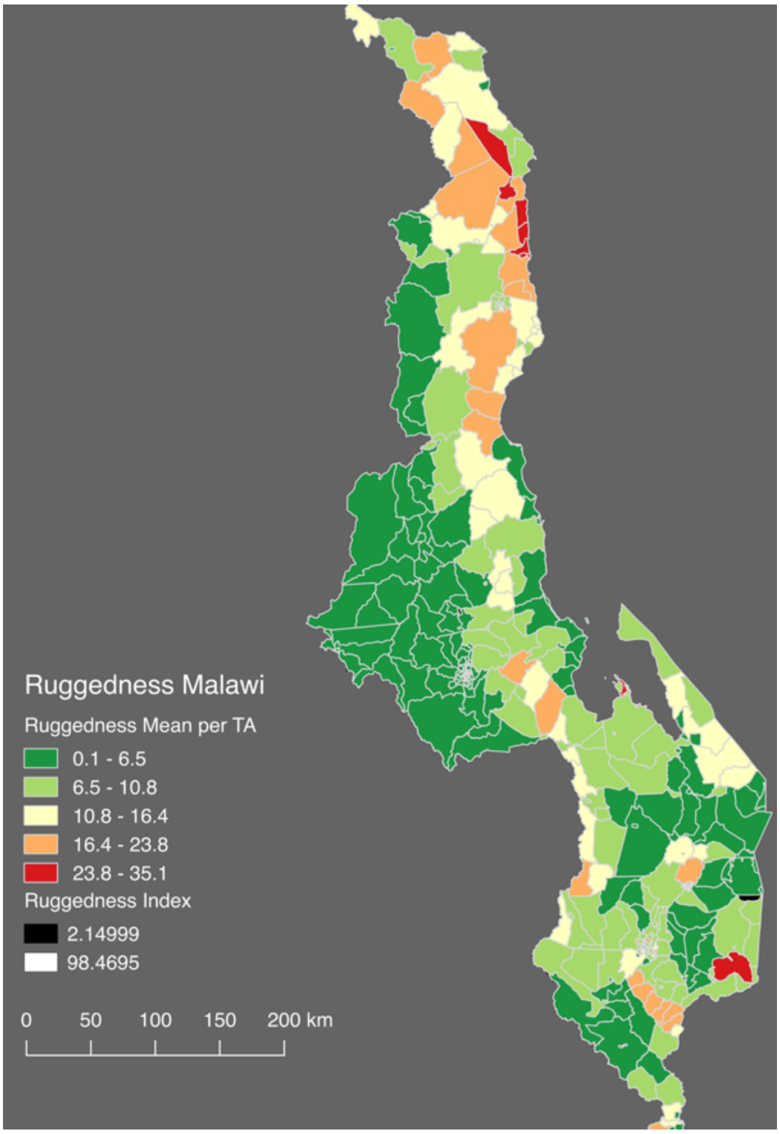


Figure 5.13 Mean Ruggedness Index score per TA.

5.3 Postgis and PgRouting analyses

An overview of the Postgis and PgRouting steps (related to sub-objective 2) is provided in the analysis scheme of *Figure 5.14*. Subsequently, the analysis steps A to F as depicted in the scheme are visualised for a single analysis step in *Figure 5.16-A to F*. As can be seen for step B, 30 Tables are created per iteration. So for the variable distance buffer creation for travel times from major hospitals, 30 variable distance buffers are conducted: for every 20 minutes' increase, another variable distance buffer is created. The same is done for step C to E. In step A, F and G, one Table is created.

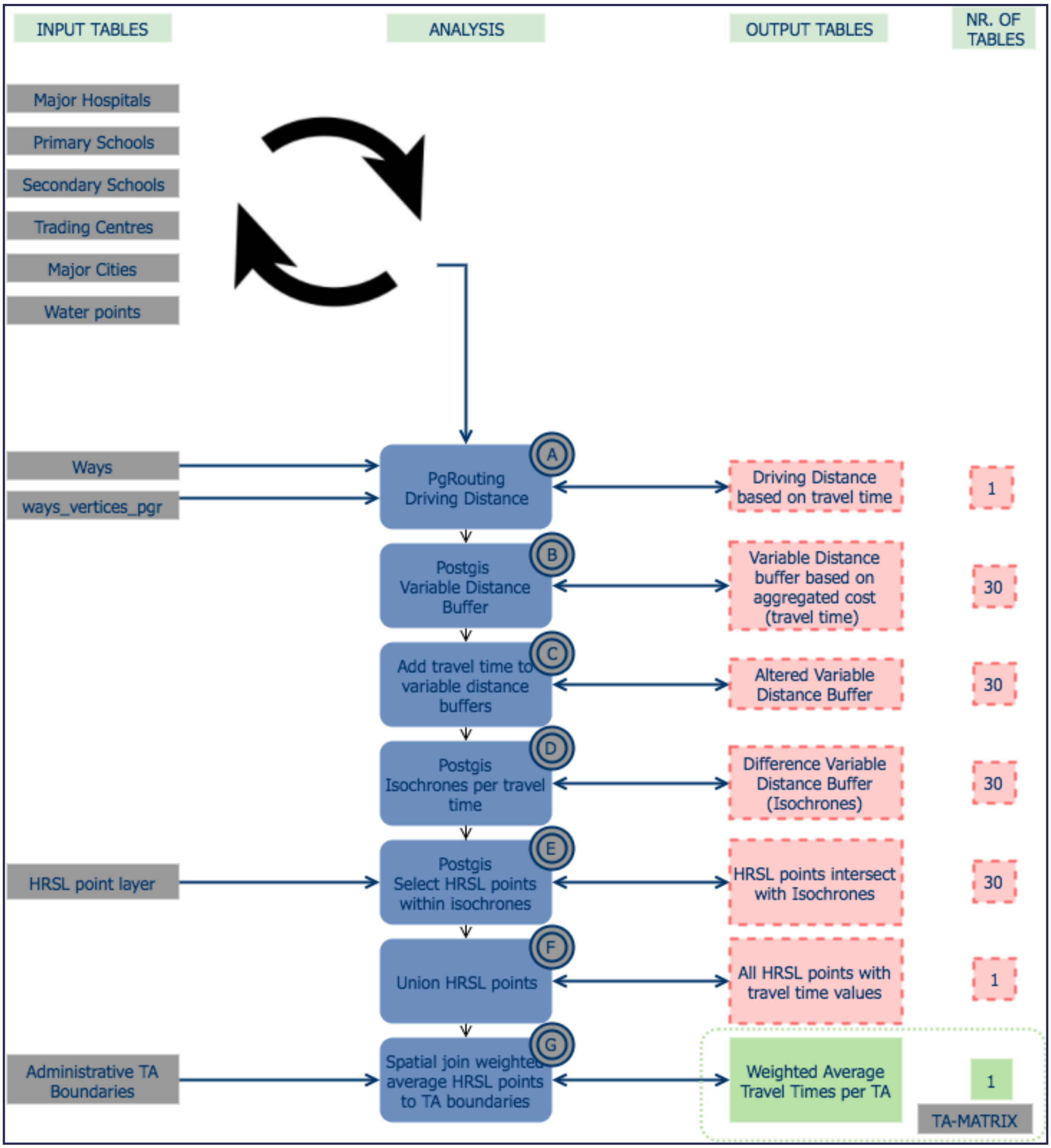

INPUT TABLES
ANALYSIS
OUTPUT TABLES
NR. OF TABLES
Major Hospitals
Primary Schools
Secondary Schools
Trading Centres
Major Cities
Water points
Ways
ways_vertices_pgr
HRSL point layer
Administrative TA Boundaries
A
PgRouting Driving Distance
B
Postgis Variable Distance Buffer
C
Add travel time to variable distance buffers
D
Postgis Isochrones per travel time
E
Postgis Select HRSL points within isochrones
F
Union HRSL points
G
Spatial join weighted average HRSL points to TA boundaries
Driving Distance based on travel time
Variable Distance buffer based on aggregated cost (travel time)
Altered Variable Distance Buffer
Difference Variable Distance Buffer (Isochrones)
HRSL points intersect with Isochrones
All HRSL points with travel time values
Weighted Average Travel Times per TA
1
30
30
30
30
1
1
TA-MATRIX

Figure 5.14 Analysis Scheme

'Input Tables' – All the used Tables with their moment of use

'Analysis' - Postgis and PgRouting steps A-G iterated for all the POI's

'Output Tables' – All the created temporary Tables and final result

'Nr. Of Tables' – Indicates the number of iterations per analysis step

Point tables

The analysis scheme in *Figure 5.14* should be read from left to right. Different geographic point Tables are shown top left, for different points of interest, for which the extraction is discussed in *Section 5.2.2* of this report. The circular arrows indicate an iteration; for each Table the same seven analysis steps (A to G depicted in blue) are taken. The Tables depicted mid-left of the analysis are input Tables that stay the same for every new iteration.

Temporary output tables

The intermediate Tables are depicted in red at the right side of the Table. Every analysis step results in an intermediate Table, which feeds into the subsequent analysis step. The green Table is the final step in the analysis and the results are written into the data matrix .csv ordered by TA.

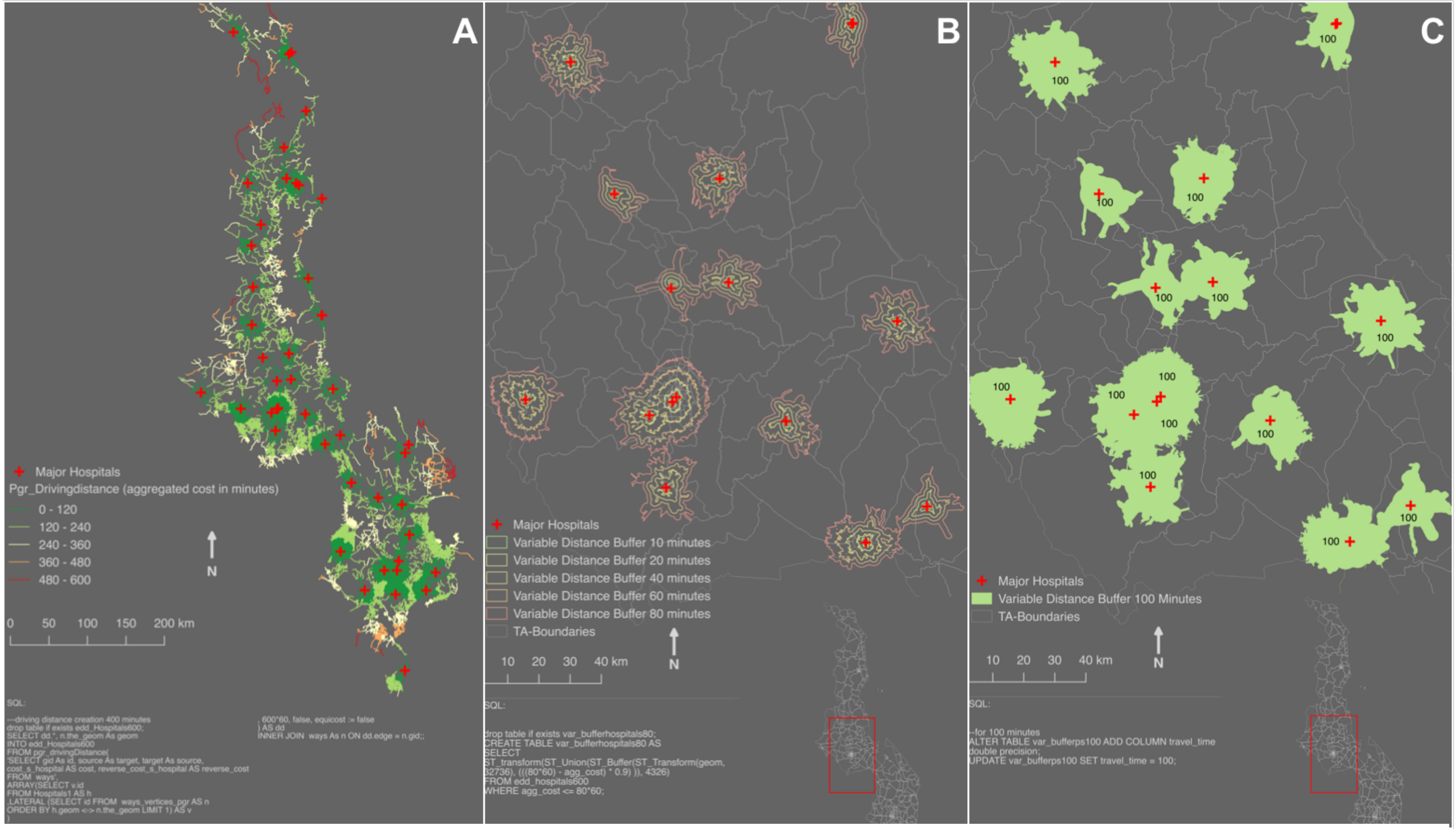
Iterations

The number of Tables is depicted on the far right. For every new iteration, the analysis steps themselves have iterations; for every analysis step where the number 30 is depicted, 30 iterations take place with increasing traveling time values; e.g. 20, 40, 60, to 600 minutes of traveling time from every point of interest in Malawi.

Settlement travel time

These PgRouting and Postgis analysis steps are taken to extract traveling times for all the human settlements within Malawi to the points of interest selected for the remoteness proxy development. To extract the traveling times, isochrones with a specific travel time value are created for every point of interest. For the major hospitals, this resulted in 46 x 30 isochrones; for every 20 minutes up until 600 minutes traveling time, one isochrone is created. As there are 46 major hospitals within the data set, 46 x 30 isochrones are calculated through the steps A to G depicted in the analysis scheme for hospitals alone. For the other points of interest, the same process is repeated.

The analysis steps A to G from ***Figure 5.14*** are spatially represented for the major hospitals in *Figures 5.15 – 5.18.* Thereafter the steps will be elaborately explained in *Sections 5.3.1 to 5.3.5.*


A
Major Hospitals
Pgr_Drivingdistance (aggregated cost in minutes)
0 - 120
120 - 240
240 - 360
360 - 480
480 - 600
N
0 50 100 150 200 km
B
Major Hospitals
Variable Distance Buffer 10 minutes
Variable Distance Buffer 20 minutes
Variable Distance Buffer 40 minutes
Variable Distance Buffer 60 minutes
Variable Distance Buffer 80 minutes
TA-Boundaries
10 20 30 40 km
N
C
100
Major Hospitals
Variable Distance Buffer 100 Minutes
TA-Boundaries
10 20 30 40 km
N

SQL:

```
---driving distance creation 400 minutes
drop table if exists edd_Hospitals600;
SELECT dd.*, n.the_geom As geom
INTO edd_Hospitals600
FROM pgr_drivingDistance(
'SELECT gid As id, source As target, target As source,
cost_s_hospital AS cost, reverse_cost_s_hospital AS reverse_cost
FROM  ways',
ARRAY(SELECT v.id
FROM Hospitals1 AS h
,LATERAL (SELECT id FROM  ways_vertices_pgr AS n
ORDER BY h.geom <-> n.the_geom LIMIT 1) AS v
)
, 600*60, false, equicost := false
) AS dd
INNER JOIN  ways As n ON dd.edge = n.gid;;
```

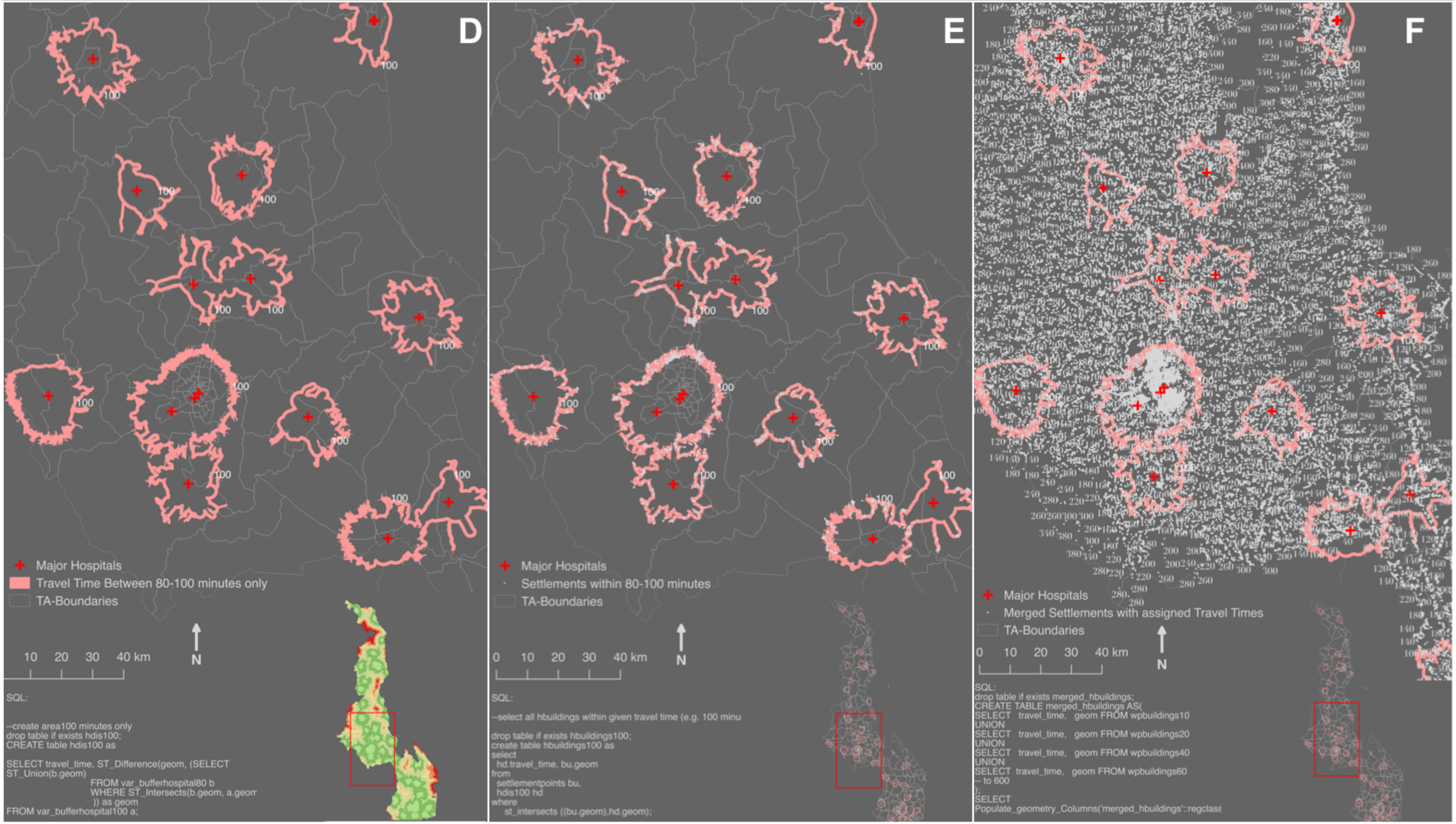
SQL:

```
drop table if exists var_bufferhospitals80;
CREATE TABLE var_bufferhospitals80 AS
SELECT
ST_transform(ST_Union(ST_Buffer(ST_Transform(geom,
32736), (((80*60) - agg_cost) * 0.9) )), 4326)
FROM edd_hospitals600
WHERE agg_cost <= 80*60;
```

SQL:

```
--for 100 minutes
ALTER TABLE var_bufferps100 ADD COLUMN travel_time
double precision;
UPDATE var_bufferps100 SET travel_time = 100;
```

Figure 5.15- A, B, C Postgis analysis steps visualised in spatial overview.
A) The Pgr_Drivingdistance output displayed using graduated colors to indicate the aggregated cost value in minutes of the road network
B) Multiple variable distance buffers, based on different aggregated costs in minutes, are displayed. The buffers are used to create service areas
C) The service areas are fitted with a value for the travel time they represent, again based on the aggregated cost in minutes


D
100
Major Hospitals
Travel Time Between 80-100 minutes only
TA-Boundaries
10 20 30 40 km
N
E
Settlements within 80-100 minutes
0 10 20 30 40 km
F
Merged Settlements with assigned Travel Times

```
SQL:

--create area100 minutes only
drop table if exists hdis100;
CREATE table hdis100 as

SELECT travel_time, ST_Difference(geom, (SELECT
ST_Union(b.geom)
                    FROM var_bufferhospital80 b
                    WHERE ST_Intersects(b.geom, a.geom
                     )) as geom
FROM var_bufferhospital100 a;
```

```
SQL:

--select all hbuildings within given travel time (e.g. 100 minu

drop table if exists hbuildings100;
create table hbuildings100 as
select
  hd.travel_time, bu.geom
from
  settlementpoints bu,
  hdis100 hd
where
    st_intersects ((bu.geom),hd.geom);
```

```
SQL:
drop table if exists merged_hbuildings;
CREATE TABLE merged_hbuildings AS(
SELECT   travel_time,   geom FROM wpbuildings10
UNION
SELECT   travel_time,   geom FROM wpbuildings20
UNION
SELECT   travel_time,   geom FROM wpbuildings40
UNION
SELECT  travel_time,   geom FROM wpbuildings60
-- to 600
);
SELECT
Populate_geometry_Columns('merged_hbuildings'::regclass
```

Figure 5.15- D, E, F Postgis analysis steps visualised in spatial overview.

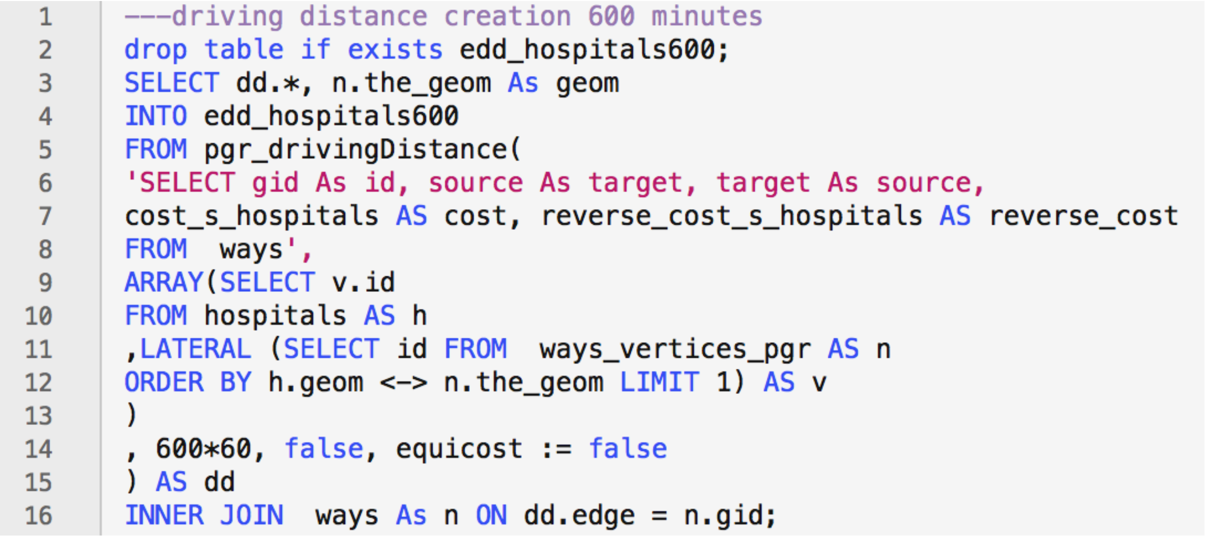
D) The service areas are 'subtracted' based on their aggregated costs in minutes to create specific sections of travel time (e.g. between 80-100 minutes). Overview show the joined isochrones from 20 to 600 minutes (green to red).

E) The settlement points are selected based on their location within the specific travel time areas, and subsequently assigned the value of that time in minutes.

F) The settlement points with their assigned values are joined in a single Table to be aggregated on TA-level in the final step of the analysis *(G – Figure 5.18)*

5.3.1 (A) PgRouting 'DrivingDistance'

Step A in the analysis scheme of *Figures 5.14 & 5.15* indicate the PgRouting function: pgr_drivingdistance. The function produces one Table per iteration, because the maximum traveling distance possible for Malawi is used: 600 minutes. The function for the driving distance from hospitals is depicted in *Function 5.1.*



```
1   ---driving distance creation 600 minutes
2   drop table if exists edd_hospitals600;
3   SELECT dd.*, n.the_geom As geom
4   INTO edd_hospitals600
5   FROM pgr_drivingDistance(
6   'SELECT gid As id, source As target, target As source,
7   cost_s_hospitals AS cost, reverse_cost_s_hospitals AS reverse_cost
8   FROM  ways',
9   ARRAY(SELECT v.id
10  FROM hospitals AS h
11  ,LATERAL (SELECT id FROM  ways_vertices_pgr AS n
12  ORDER BY h.geom <-> n.the_geom LIMIT 1) AS v
13  )
14  , 600*60, false, equicost := false
15  ) AS dd
16  INNER JOIN  ways As n ON dd.edge = n.gid;
```

Function 5.1 pgr_drivingDistance for hospitals

SQL function explained

Row 2

Within the function a Table is created for the driving distance results: edd_hospitals600. The Table has the indicator 'edd' because the driving distances are joined ***row 16:(***INNER JOIN) on the 'ways' (edge Table). The 'Ways' Table is indicated as the first Table in the analysis scheme that remains the same for every iteration of points of interest.

Row 3

The geometry of the node Table is selected. The node Table is called 'ways_vertices_pgr' and is shown in the analysis scheme under 'input Tables' as the second Table that remains the same for every iteration of the points of interest.

Row 5 - 8

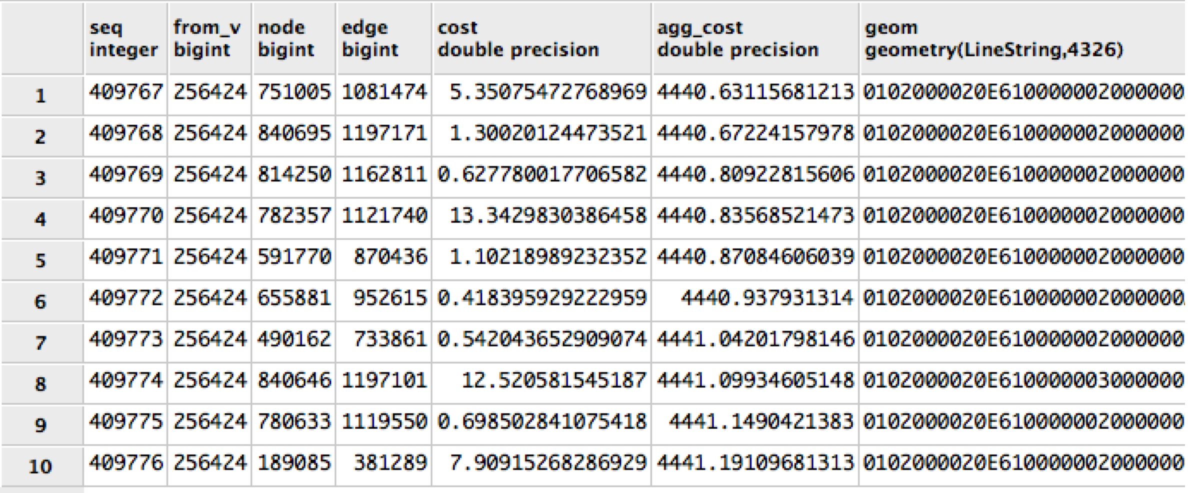
The PgRouting function pgr_drivingDistance is called upon and the attributes are indicated. The source and target nodes from the edge Table are switched, due to the fact that the traveling time to the hospital is preferred, rather than from the hospital to the settlements and some network sections may be directed. Subsequently the cost_s columns are indicated. These columns are the cost of every edge part of the ways Table calculated through the length in meters and possible traveling speed in meters per second. The costs of the edge Table are therefore in seconds. For the points of interest, a different traveling speeds is used; e.g. cost_s_hospital, cost_s_secschools, cost_s_primschools for the average traveling speed for hospitals and secondary schools. These traveling speeds are based on the calibration measurements conducted during the field study in Malawi.

Row 9 - 12

The points of interest table, in this case the hospital point data, is selected as an array. A lateral selection of the node table (ways_vertices_pgr) is conducted to select the nodes on the network that fall within the traveling cost (in seconds). The amount of cost that will be taken into account is indicated in row 14: 600 minutes (600 x 60 seconds). This means that the nodes on the network reachable within 600 minutes will be taken into account. Finally, the aggregated cost in seconds is joined to the edge Table (ways).

Visualising the aggregated cost in minutes for hospitals on the edge table results in the output depicted in ***Figure** 5.15-A.* The maximum travel time for hospitals is 600 minutes for Malawi. A small section of the output Table created by the Pgr_drivingDistance function is depicted in *Table 5.3.*

Table 5.3: A selection of the Table created by the pgr_drivingDistance function (edd_hospitals600)



	seq integer	from_v bigint	node bigint	edge bigint	cost double precision	agg_cost double precision	geom geometry(LineString,4326)
1	409767	256424	751005	1081474	5.35075472768969	4440.63115681213	0102000020E610000002000000
2	409768	256424	840695	1197171	1.30020124473521	4440.67224157978	0102000020E610000002000000
3	409769	256424	814250	1162811	0.627780017706582	4440.80922815606	0102000020E610000002000000
4	409770	256424	782357	1121740	13.3429830386458	4440.83568521473	0102000020E610000002000000
5	409771	256424	591770	870436	1.10218989232352	4440.87084606039	0102000020E610000002000000
6	409772	256424	655881	952615	0.418395929222959	4440.937931314	0102000020E610000002000000
7	409773	256424	490162	733861	0.542043652909074	4441.04201798146	0102000020E610000002000000
8	409774	256424	840646	1197101	12.520581545187	4441.09934605148	0102000020E610000003000000
9	409775	256424	780633	1119550	0.698502841075418	4441.1490421383	0102000020E610000002000000
10	409776	256424	189085	381289	7.90915268286929	4441.19109681313	0102000020E610000002000000

An important constraint of the pgr_drivingDistance function is the node selection on the network. The accuracy of the analysis depends on the node density of the network. The problem is depicted in *Figure 5.16,* the actual driving time over the network is indicated on the network. However, if there is no node available, the node closest to the point of interest is returned. In the network created through the OSM2PgRouting command-line tool, the node density for some sections of the network graph was limited.

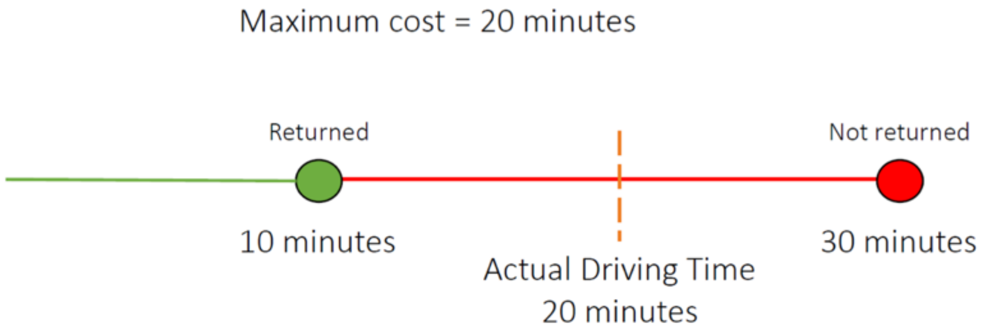


Figure 5.16 Limitation of the node selection of the network through the pgr_drivingDistance function (Adapted from Obe & Hsu, 2017).

To fix certain limitations within the network Table, PgRouting offers a number of 'network-fixing' functions:

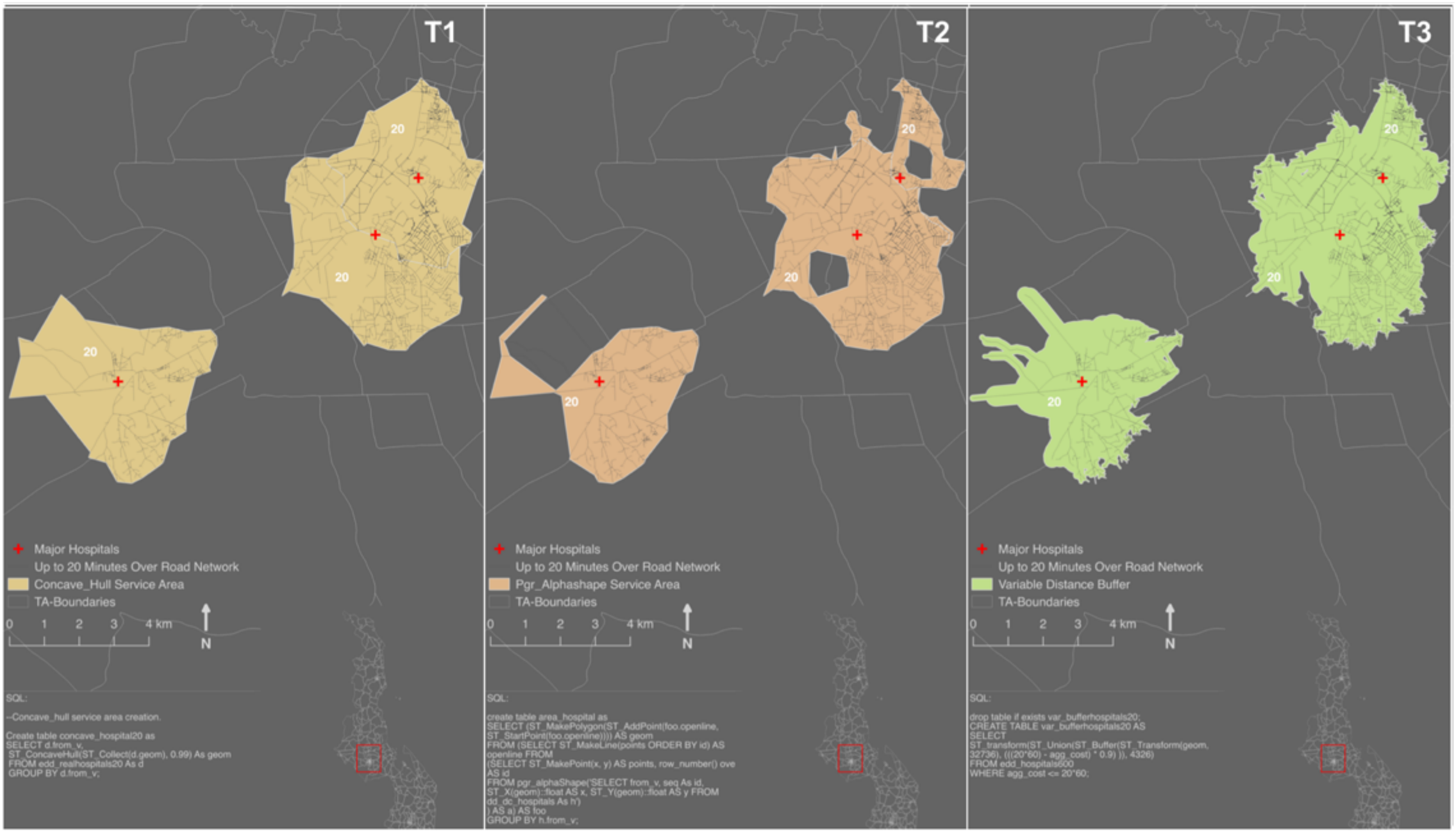
- Pgr_analyzeGraph - After the network is built, this function was run to fix gaps between nodes and fix dead-ends.
- Pgr_analyzeOneway - Analyses directionality, calls out edges that violate directionality rules.
- Pgr_nodeNetwork - Ensures that all edges connect at nodes. Fix disconnected edges and edges that cross each other by introducing additional edges as necessary.

After these steps are taken, the ability to analyse the network was improved slightly. The number of nodes is increased, the node density however, is not a parameter that can be adjusted in the network fixing functions.

5.3.2 (B) Variable distance buffer

This paragraph is related to step B in *Figure 5.15* and will describe the steps taken and the process towards the selection of the techniques used.

Step B is the second step toward the isochrones creation. Several techniques (T) are investigated to arrive at service areas based on the aggregated costs joined to a network table. *Figure 5.17* depicts the results of the three techniques used a concave hull function (T1), the PgRouting function Alphashape (T2), and the variable distance buffer (T3). All the techniques are based on the aggregated cost derived through the Pgr_Drivingdistance function, in minutes over the network. The Concave Hull and the Alphashape technique were finally disregarded due to under or over selection of settlements without a strong enough theoretical underpinning. As can be seen in *Figure 5.17 – T1* the Concave Hull over selects the service area; the farthest points of the network are taken, in this case 20 minutes, and joined with lines not taking the gaps within the road network, or the areas not reachable by the network into account. In *Figure 5.17 – T2* the Pgr_Alphashape function is displayed. This service area creation technique shows more potential, as it takes the road network into account, and areas appear within the service areas that are not served by the hospital within 20 minutes. However, if walking speed is taken into account for areas not served by the road network, this technique shows an under-selection of service area; this effect gets increased with higher travel times and if more points of interest are taken into account that lay close to each other. This was the case with water points, which had 3805 points distributed over the whole of Malawi, rather than 46 for the hospitals. Finally, *Figure 5.17 – T3* shows the result of the variable distance buffer which was used for this analysis.



```
--Concave_hull service area creation.

Create table concave_hospital20 as
SELECT d.from_v,
 ST_ConcaveHull(ST_Collect(d.geom), 0.99) As geom
 FROM edd_realhospitals20 As d
 GROUP BY d.from_v;
```

```
create table area_hospital as
SELECT (ST_MakePolygon(ST_AddPoint(foo.openline,
ST_StartPoint(foo.openline)))) AS geom
FROM (SELECT ST_MakeLine(points ORDER BY id) AS
openline FROM
(SELECT ST_MakePoint(x, y) AS points, row_number() ove
AS id
FROM pgr_alphaShape('SELECT from_v, seq As id,
ST_X(geom)::float AS x, ST_Y(geom)::float AS y FROM
dd_dc_hospitals As h')
) AS a) AS foo
GROUP BY h.from_v;
```

```
drop table if exists var_bufferhospitals20;
CREATE TABLE var_bufferhospitals20 AS
SELECT
ST_transform(ST_Union(ST_Buffer(ST_Transform(geom,
32736), (((20*60) - agg_cost) * 0.9) )), 4326)
FROM edd_hospitals600
WHERE agg_cost <= 20*60;
```

Figure 5.17 Area creation techniques (T): St_ConcaveHull (T1), Pgr_Alphashape (T2), and Variable distance buffer (T3).

The variable distance buffer function used for the analysis for 20 minutes is depicted above in *Figure 5.17 – T3*, to increase the understanding of the function, the variable distance function for 100 minutes is depicted in the following section in ***function* *5.18***.

```
1  drop table if exists var_bufferhospital100;
2  CREATE TABLE var_bufferhospital100 AS
3  SELECT  ST_transform(ST_Union(ST_Buffer(ST_Transform(geom, 32736),
4   (((100*60) - agg_cost) * 0.9) )), 4326)
5  FROM edd_hospital600
6  WHERE agg_cost <= 100*60;
```

Function 5.1 Variable distance buffer example for the travel time to majors hospitals of 100 minutes.

In *Function 5.1* the SQL is given for the variable distance buffer used throughout the GIS analysis where the service areas are created for every point of interest (*see Section 5.2.2 for the points of interest used in this research to calculate travel times or Figure 5.14 under 'input tables'*).

SQL function explained
Rows 1-2
Through rows 1 and 2 the variable distance table name and creation statement is given. The 'drop table if exists' statement is used for the case the Table needs to be re-run, the existing Table within the database is deleted (dropped) and can be re-created again. Postgis functions do not allow to overwrite

existing tables. This statement is useful in the case many iterations need to be run within a larger SQL query.

Rows 3-4

Through rows 3 and 4 the input table is transformed to the local Spatial Reference Identification (SRID) for Malawi. UTM-zone 36s with the code 32736 is used. Postgis functions work with SRID codes for data Tables with a geometry column implemented. Subsequently, the parameters of the buffer is indicated; this is where the normal Postgis buffer is used to create a variable distance buffer. The buffer characteristics are indicated with a function, rather than a fixed distance. The function ((100*60) – agg_cost) * 0.9)) is used to indicate the distance the buffer should take into account. In this case 100 minutes is the maximum travel time taken into account for the service area; hence the 100 *60 (as aggregated cost is in seconds, it is multiplied by 60). The aggregated cost is subtracted from the maximum travel time (in seconds) and multiplied by the speed people are able to walk straight from the road (in meters per second). The outcome is a value in meters. So for every point on the network within 100 minutes traveling to a hospital, a different size buffer is created based on the time left *((maximum in minutes) – (aggregated cost in minutes of that point on the network))* when arriving at that specific point on the road network.

Rows 5-6

As stated above, the parameters of the variable distance buffer rely on the aggregated costs in minutes. The aggregated cost Table of the network is created in the Pgr_Drivingdistance function and is depicted in *Figure 5.16 – A.* The 'FROM' statement refers to a Table in the Postgis database and the 'WHERE' statement indicated the aggregated costs that need to be taken into account.

5.3.3 (C&D) Specified Travel Time

This paragraph is related to steps C and D of Figure 5.15. The adding of the actual travel times and the creation of the polygon areas that describe the specific travel time only.

In step C the travel time value is added to the created variable distance buffers of the area that is reachable within the time specified in the variable distance buffer. This means that the score between 80 and 100 minutes receives the value 100 minutes. The SQL is given in the spatial visualisation to analysis step in C.

Step D describes the difference of the polygons, to create the travel time 'only' polygons. This is done through the SQL given in step D. The Postgis St_difference function is used. This is done for all the different travel time-specified polygons created through the variable distance buffer. This results in the output visualized in step D of ***Figure 5.15.*** The map indicated as overview of Malawi for step D is an output of all the specific travel time areas joined in one Table to give a visual overview of the service areas of the hospitals for the whole of Malawi from green (low travel times) to red (high travel times). Directly noticeable are the red areas. Inhabitants of these areas have to travel the longest to reach a major hospital.

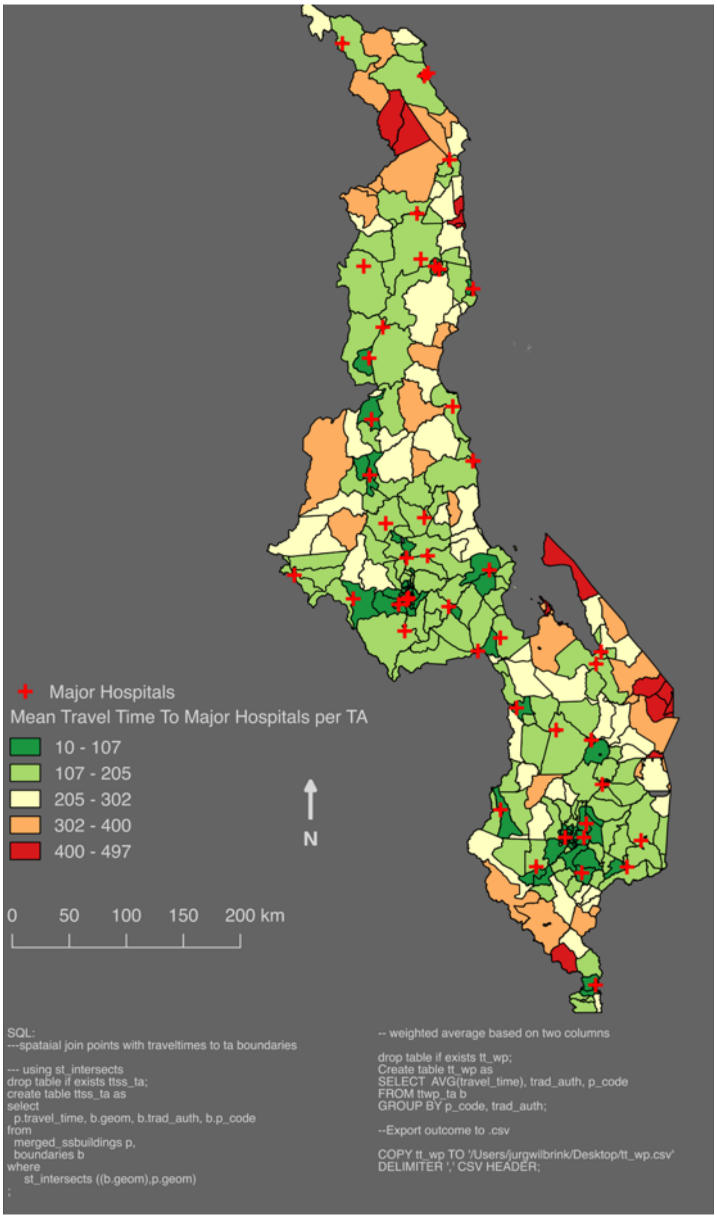
5.3.4 (E&F) HRSL Point Selection

This paragraph is related to steps E and F of *Figure 5.15*. Step E is where the HRSL is added to the analysis, also see *Figure 5.14* under 'input Tables' and notice that the Table gets added for the analysis in step E. The HRSL point Table is create through the steps indicated in *Section 5.2.3.*

Step E is done to assign a travel time to all the available settlement points extracted from the HRSL layer created by CEISIN and Facebook. This is shown in the SQL displayed in step E of *Figure 5.15*; rows 3-4 under 'select'. The travel time of the polygons is selected, and the geometry of the settlement point file. Both are combined and assigned to a new geometry Table.

In step F all the settlement points are joined within one larger table. This is done to have all the points with the assigned travel times within one Table to perform the final analysis step performed in step G, following the analysis scheme *Figure 5.14 – step G*. The output of this step is presented in ***Figure 5.18.***

5.3.5 (G) Data Matrix on TA-level



```
SQL:
---spataial join points with traveltimes to ta boundaries

--- using st_intersects
drop table if exists ttss_ta;
create table ttss_ta as
select
  p.travel_time, b.geom, b.trad_auth, b.p_code
from
  merged_ssbuildings p,
  boundaries b
where
    st_intersects ((b.geom),p.geom)
;
```

```
-- weighted average based on two columns

drop table if exists tt_wp;
Create table tt_wp as
SELECT  AVG(travel_time), trad_auth, p_code
FROM ttwp_ta b
GROUP BY p_code, trad_auth;

--Export outcome to .csv

COPY tt_wp TO '/Users/jurgwilbrink/Desktop/tt_wp.csv'
DELIMITER ',' CSV HEADER;
```

Figure 5.18 Step G. Example for travel times to major hospitals per traditional authority with SQL script of the analysis added to the bottom of the Figure.

Figure 5.18 shows the final output of one iteration (for hospitals) of the GIS travel time analysis. The Postgis and PgRouting analysis resulted in 6 similar outputs for all the points of interest. The SQL query is given for this final step in the GIS analysis. The first section indicates the spatial join between the settlement points with the assigned travel times and the administrative TA-boundaries. Subsequently, the weighted average of the points is taken per TA. The created remoteness indicators are filled into the data matrix on TA-level. A section of the data matrix with all the variables included is shown in *Table 5.4.*

The final data matrix is filled with the following variables:

- PCODE: The P-code corresponding to the Traditional Authority. This code is used to join different Tables on within the Postgis database. The use of the P-code results in faster, but more importantly, more accurate outputs. The latter part is especially important for Malawian TA's, as some TA's within Malawi are named the same, or are a variation of the same name.
- **TA:** The Tradional Authorities, filled with the names of all the Traditional Authorities within Malawi. Malawi counts 367 TA's, the section shown in Table 5.4 counts 19 of these TA's to give an indication of what the data matrix looks like.
- **Nr_Settlements:** The number of settlements within a specific TA. The number of points within a TA are counted and the data matrix is filled with this information.
- Settlements_km2: The settlement density is calculated as the number of settlements per square kilometre. This is taken into account, as the real number of settlements may also rely on the TA area size.
- **TT_Cities:** The travel time to major cities in minutes, as calculated through the steps A to G as indicated in *Figures 5.14 – 5.18.*
- TT_Hospitals: The travel time to hospitals in minutes, as calculated through the steps A to G as indicated in *Figures 5.14 – 5.18.*
- TT_Secschools: The travel time to secondary schools in minutes, as calculated through the steps A to G as indicated in *Figures 5.14 – 5.18.*
- TT_Tradcentres: The travel time to trading centres in minutes, as calculated through the steps A to G as indicated in *Figures 5.14 – 5.18.*
- TT_Primschools: The travel time to primary schools in minutes, as calculated through the steps A to G as indicated in *Figures 5.14 – 5.18.*
- TT_Waterpoints: The travel time to water points in minutes, as calculated through the steps A to G as indicated in *Figures 5.14 – 5.18.*
- Mean_Ruggedn_Index: The mean ruggedness index, as calculated through the steps indicated in chapter 5.2.4.
- **Rd_Length_km:** The road length in kilometre per TA.
- RoadDens_km2: The road density, as calculated through the road length per square kilometre.
- **Settlement_Size:** The settlement size, calculated by dividing the population per TA (census, 2011) by the number of settlements.
- **SoVI:** The Social Vulnerability Index per TA, created through the steps indicated in *Section 5.1* and acts as the Y-variable.

Table 5.4 Section of the data matrix with all the developed remoteness indicators (X-variables) and the SoVI (Y-variable) included. All the travel times (TT) are depicted in minutes.

PCode	TA	Nr_Settlements	SettlmDens_km2	TT_Cities	TT_Hospitals	TT_Secschools	TT_Tradcentres	TT_Primschools	TT_Waterpoints	Mean_Ruggedn_Index	Rd_Length_Km	RoadDens_km2	Settlement_Size	SoVI
AFRMWI30501	Senior TA Kapeni	24968	117.1385829	81	128	36	45	36	38	5.720946415	310.181179	1.455230045	2.893143223	40.23366094
AFRMWI30502	TA Lundu	8155	61.08297132	90	251	78	74	56	92	10.12500456	25.11451162	0.188113917	5.686327407	48.63071139
AFRMWI30503	TA Chigaru	10228	40.02192644	106	289	62	94	58	88	4.596246552	138.4812359	0.54187386	3.894798592	49.77092684
AFRMWI30504	TA Kunthembwe	8971	19.43312703	200	216	160	132	92	100	9.412356288	76.42695887	0.165557329	3.7643518	50.41944752
AFRMWI30721	Luchenza Township	2075	210.2254746	17	129	16	17	31	33	5.252761107	28.71486725	2.909203179	5.251084337	53.94120216
AFRMWI20501	TA Maganga	5649	31.77616472	105	124	102	44	31	33	6.210514195	135.4522069	0.761931605	8.408213843	62.98035608
AFRMWI20502	TA Karonga	11640	26.0418677	103	89	71	109	63	84	6.698703887	194.1371339	0.434337934	4.772594502	65.84830902
AFRMWI20503	TA Pemba	2971	30.06679119	103	125	100	65	53	55	5.170817495	52.87993657	0.535149785	6.584315045	64.84423683
AFRMWI20504	STA Kambwiri	6296	23.68591313	92	102	96	106	61	113	7.530537644	103.7844328	0.390442989	4.306543837	68.31010014
AFRMWI20505	TA Ndindi	7275	33.7015209	75	190	69	80	83	180	5.585963301	74.92607253	0.34709589	5.532646048	65.3061334
AFRMWI30805	TA Nkanda	20201	54.87190002	128	109	48	48	36	75	4.303052349	1875.127196	5.093400923	4.647888718	64.77400264
AFRMWI30806	TA Juma	14705	43.36975918	179	165	74	92	55	68	3.683461381	213.2316576	0.628888517	5.314450867	64.73746549
AFRMWI30807	Mulanje Mountain Reserve	4634	9.79721665	153	143	121	124	113	176	35.07413116	130.0392602	0.274929393		65.8733187
AFRMWI21054	Area 24	3041	946.3847682	58	41	22	25	17	20	4.349999116	35.81800998	11.14686586	7.4672805	45.64847851

5.4 Field validation – Speed Parameter Calibration

In the methodology scheme (*Figure 3.2*) as a part of sub-objective 2, the validation and calibration of the remoteness indicators is taken into account. This chapter will elaborate on the conducted steps during this field validation.

During the field visit within Malawi several steps were undertaken to calibrate the parameters used to conduct the travel time analysis steps in Postgis and PgRouting as described in ***Section 5.3***. Random villages in the working area of the MRCS, within the mapped area through the Missing Maps project (Missing Maps, 2016) and logistically feasible to reach within a week, were selected. Fifteen Malawi Red Cross volunteers and three staff members were identified to support the field calibration steps. During the field calibration through digital surveying methods new data was collected simultaneously. The collected data consists of the travel method and time to major hospitals, travel method and time to primary and secondary schools, building materials, state of the building, water and sanitation points, and the accuracy of school-, trading centres-, cities- and hospital locations. Next to the data collected in the digital survey, digital imagery of the roads to- and around the villages were collected with the use of cameras on the vehicles. The imagery is collected as part of a crowd-sourced effort to create openly available street-view imagery for the world (Mapillary, 2017). All the data was collected to enrich the existing OSM data, add to the MRCS database of their working areas and to calibrate the GIS analyses conducted in this research.

Thyolo district is an important work area of the MRCS. The district was identified two months before the field study as the location of the field-study. Through the Missing Maps projects hosted by the Netherlands Red Cross, the major working areas of MRCS within the district were mapped as complete as possible; this entails the mapping of roads and buildings that can be found on satellite imagery. The mapping of these areas were conducted through crowd-sourced mapping events, often referred to as ‘Mapathons’, as described by Missing Maps (2017). These mapping events were driven by the anticipation of the field calibration of the travel time parameters and in general to create more accurate maps for humanitarian organizations working within the area.

An overview of the OSM road data, locations of major hospitals, primary schools, secondary schools, and the location of the visited villages for Thyolo district is given in ***Figure 5.19***. The five villages were identified, based on the random selection of locations in the northern region of Thyolo district. The main reason for this selection was the logistic feasibility to visit the villages within a week, while transporting a total of twenty voluntary enumerators (data collectors). Furthermore, within the northern region of Thyolo the MRCS is most active and the OSM data completeness is relatively high compared to the rest of the district. This can be seen in road density in the north, compared to the south depicted in ***Figure 5.19.*** The village names were not known before hand, and were identified through part of the conducted survey. A list of all the survey questions as well as the form design steps for ODK and OMK surveys are displayed in Appendix C. During the field-surveys the locations of the schools were validated by visiting the school and acquiring the location to be able to compare the school locations to the available data set. For the major hospitals, the map of the locations was shown to the health experts within MRCS to confirm the location, names, and capacities of the hospitals.

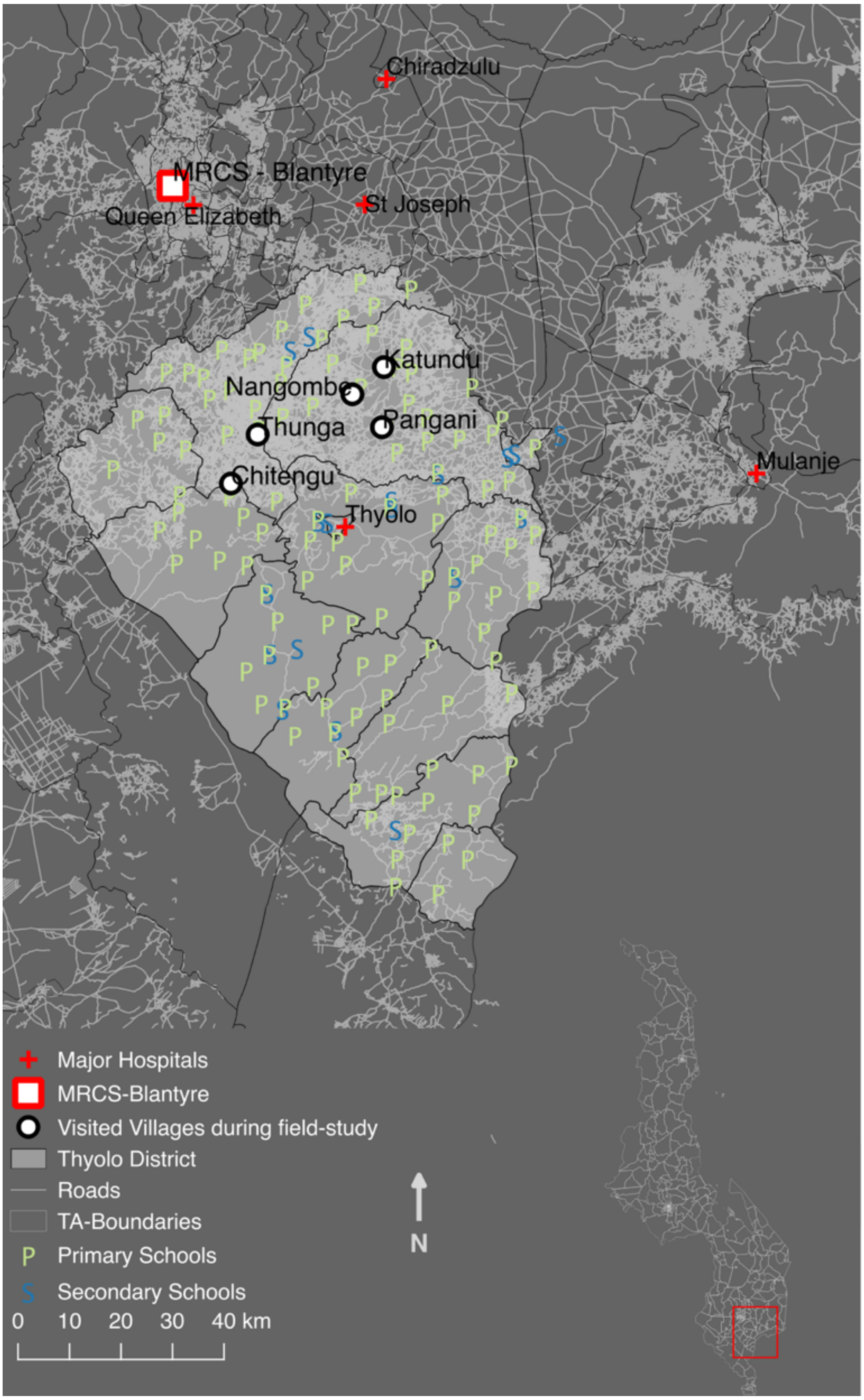


Figure 5.19 Overview of villages visited during the field study. The MRCS-Blantyre served as the base from where the daily field trips were organised. The Major hospitals, primary schools, secondary schools, and roads are indicated in this spatial overview.

OpenDataKit and OpenMapkit surveys

For the surveying in the villages to collect the relevant data, digital open-source tools were used: OpenDataKit (ODK) and OpenMapKit (OMK). ODK is the digital tool where survey questions are incorporated and OMK is built upon ODK to add the OSM-based spatial attributes to the survey. Both applications are (android) smartphone or tablet based tools. Spatial attributes are referred to as 'tags', and spatial points are referred to as 'nodes' within the OSM ontology. The MRCS staff and volunteers received a one-day training on operating the digital tooling and the relevancy of the survey and survey questions were explained. The following five days were spent in the field, visiting the villages and conducting the survey questions at every household within the villages. Whilst conducting the survey, spatial attributes were added to the existing OSM data layer. *Figure 5.20* gives a spatial overview of the changes and additions made in the OSM tags and nodes through the conducted surveys. The open-source software Java OpenStreetMap editor(JOSM) was used to upload the collected data to OSM.

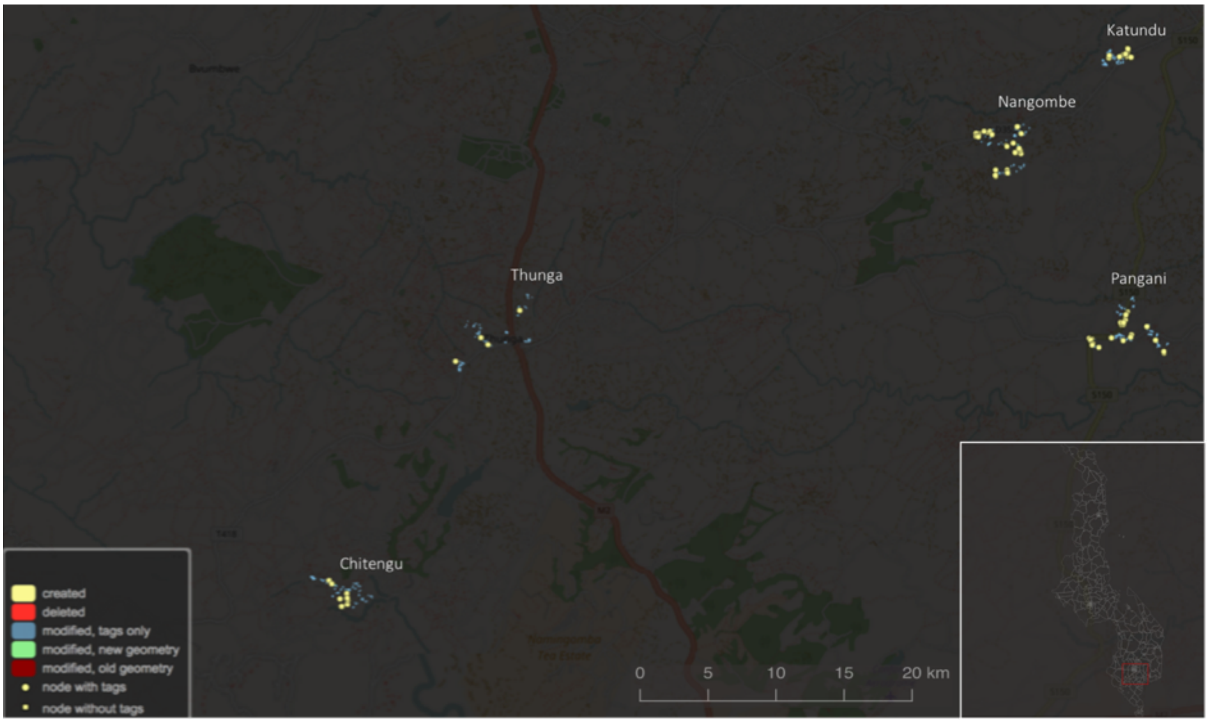


Figure 5.20 Overview of OSM data added and edited as a result of the field validation within Thyolo district in Malawi.

ODK – OMK outcomes

The outcomes of the surveys are used to calibrate the GIS analyses conducted to determine travel times from all the points of interest mentioned in *Section 5.2.2.* The output of all the field survey questions relevant to calibrate these travel times can be found in Appendix C. A short write-up on the enriching of the OSM data and the use of cameras on the cars during the field-work is elaborated on in Appendix D.

The main results identified through the weighted averaging of the travel times for hospitals are depicted in *Table 5.4.* The same is done for primary schools and secondary schools to calibrate the travel speeds for these points of interest.

Table 5.5 Output for different steps in the calibration of traveling speed for major hospitals ordered by village and traveling method.

Village	Distance to hospital (km)	Travel method	Count	Total count	percentage of total	percentage	Travel time in minutes	Avg. Travel time	Avg. Travel time total	Avg. speed km/h	Avg. village speed weighted	Avg. total Speed weighted	Avg. total speed weighted m/s
Thunga	15	publictransport	40	51	0.226666667	0.784313725	56	69.4	94.48	16.07142857	15.8888959	10.5973725	2.943714583
Thunga	15	byfoot	3	51	0.226666667	0.058823529	130	69.4	94.48	6.923076923	15.8888959	10.5973725	2.943714583
Thunga	15	car	6	51	0.226666667	0.117647059	41	69.4	94.48	21.95121951	15.8888959	10.5973725	2.943714583
Thunga	15	bike	2	51	0.226666667	0.039215686	120	69.4	94.48	7.5	15.8888959	10.5973725	2.943714583
Thunga	15	motorbike	0	51	0.226666667	0	0	69.4	94.48	0	15.8888959	10.5973725	2.943714583
Pangani	12.5	publictransport	25	74	0.328888889	0.337837838	60	107.2	94.48	12.5	8.644639108	10.5973725	2.943714583
Pangani	12.5	byfoot	22	74	0.328888889	0.297297297	137	107.2	94.48	5.474452555	8.644639108	10.5973725	2.943714583
Pangani	12.5	car	9	74	0.328888889	0.121621622	71	107.2	94.48	10.56338028	8.644639108	10.5973725	2.943714583
Pangani	12.5	bike	16	74	0.328888889	0.216216216	118	107.2	94.48	6.355932203	8.644639108	10.5973725	2.943714583
Pangani	12.5	motorbike	2	74	0.328888889	0.027027027	150	107.2	94.48	5	8.644639108	10.5973725	2.943714583
Nangombo	17.5	publictransport	19	35	0.155555556	0.542857143	128	94.2	94.48	8.203125	8.478481923	10.5973725	2.943714583
Nangombo	17.5	byfoot	4	35	0.155555556	0.114285714	107	94.2	94.48	9.813084112	8.478481923	10.5973725	2.943714583
Nangombo	17.5	car	0	35	0.155555556	0		94.2	94.48	0	8.478481923	10.5973725	2.943714583
Nangombo	17.5	bike	9	35	0.155555556	0.257142857	133	94.2	94.48	7.894736842	8.478481923	10.5973725	2.943714583
Nangombo	17.5	motorbike	3	35	0.155555556	0.085714286	103	94.2	94.48	10.19417476	8.478481923	10.5973725	2.943714583
Katundu	19	publictransport	12	41	0.182222222	0.292682927	108	107	94.48	10.55555556	10.19208369	10.5973725	2.943714583
Katundu	19	byfoot	13	41	0.182222222	0.317073171	138	107	94.48	8.260869565	10.19208369	10.5973725	2.943714583
Katundu	19	car	5	41	0.182222222	0.12195122	60	107	94.48	19	10.19208369	10.5973725	2.943714583
Katundu	19	bike	10	41	0.182222222	0.243902439	189	107	94.48	6.031746032	10.19208369	10.5973725	2.943714583
Katundu	19	motorbike	1	41	0.182222222	0.024390244	40	107	94.48	28.5	10.19208369	10.5973725	2.943714583
Chitengu	16	publictransport	16	24	0.106666667	0.666666667	97	94.6	94.48	9.896907216	9.156230427	10.5973725	2.943714583
Chitengu	16	byfoot	3	24	0.106666667	0.125	140	94.6	94.48	6.857142857	9.156230427	10.5973725	2.943714583
Chitengu	16	car	0	24	0.106666667	0		94.6	94.48	0	9.156230427	10.5973725	2.943714583
Chitengu	16	bike	3	24	0.106666667	0.125	116	94.6	94.48	8.275862069	9.156230427	10.5973725	2.943714583
Chitengu	16	motorbike	2	24	0.106666667	0.083333333	120	94.6	94.48	8	9.156230427	10.5973725	2.943714583

The tables contain all the outputs of steps taken to identify a weighted average in traveling speed of point of interests per village. *Table 5.4* contains the output for major hospitals found through the field survey of every village. A total of 225 households indicated to have travelled to a major hospital from their current home. The distance to the major hospitals, primary schools, and secondary schools from the designated villages are measured through the use of the shortest possible route through the PgRouting Dijkstra algorithm. The algorithm is designed to find the shortest path between two nodes on a network, in this case the OSM road network for Malawi, see *Figure 5.21*. The majority of the respondents indicated that public transport is the main method of transportation to a major hospital. Only in Katundu, transportation by foot was slightly higher; which was the village with the longest travelling distance to the major hospital of choice. Thunga has the highest percentage traveling by public transport to major hospitals: 78.4 %, Katundu the lowest percentage: 29.2%. Personal motorized transportation methods (motorbike and car) are the least represented traveling methods for all the villages. In the villages that do have households indicating to travel by personal car to the major hospital, the percentage is around 12%. The percentage traveling by foot to the major hospital differs per village; Thunga shows the least households traveling by foot (5.8%), and Katundu the highest percentage (31.7%). Pangani also shows a relative high percentage traveling by foot: 29.7%. Pangani and Nangombe have the lowest weighted average in traveling speed to hospitals: 8.6 and 8.5 km/h. Thunga and Katundu show the highest weighted average traveling speeds: 15.9 and 10.2 km/h. Both resulting from the relative high percentage of personal car transportation with high traveling speeds. The weighted average speed of all the villages (total weighted average speed) is 10.6 km/h. As the cost of the network used in PgRouting (explained in *Section 5.3.1*) is in meters, the cost_s parameter (traveling speed calculated per road segment based on the road length in meters) for major hospitals is calibrated to 2.94 m/s for the GIS analysis.

The same data is collected for primary and secondary schools, which went through a similar process. The cost_s parameter for primary schools is set to 1.38 m/s and the cost_s parameter for secondary schools is set to 3.04 m/s.

Before the field validation, the GIS analyses were conducted with the parameter set to the OSM ‘maxspeed’ tag, based on the maximum allowed driving speed for a particular road. Travel time outputs based on this tag were approximately under-estimating the travel times for all points of interest with a factor nine. This resulted in major over-estimation of service areas for all the points of interest. Most of the speed tags in OSM road network are set to 50 km/h, as there is no other relevant information added. In the context of Malawi, this tag was not deemed useful. Therefore, the calibrated traveling speeds were applied for the entire network.

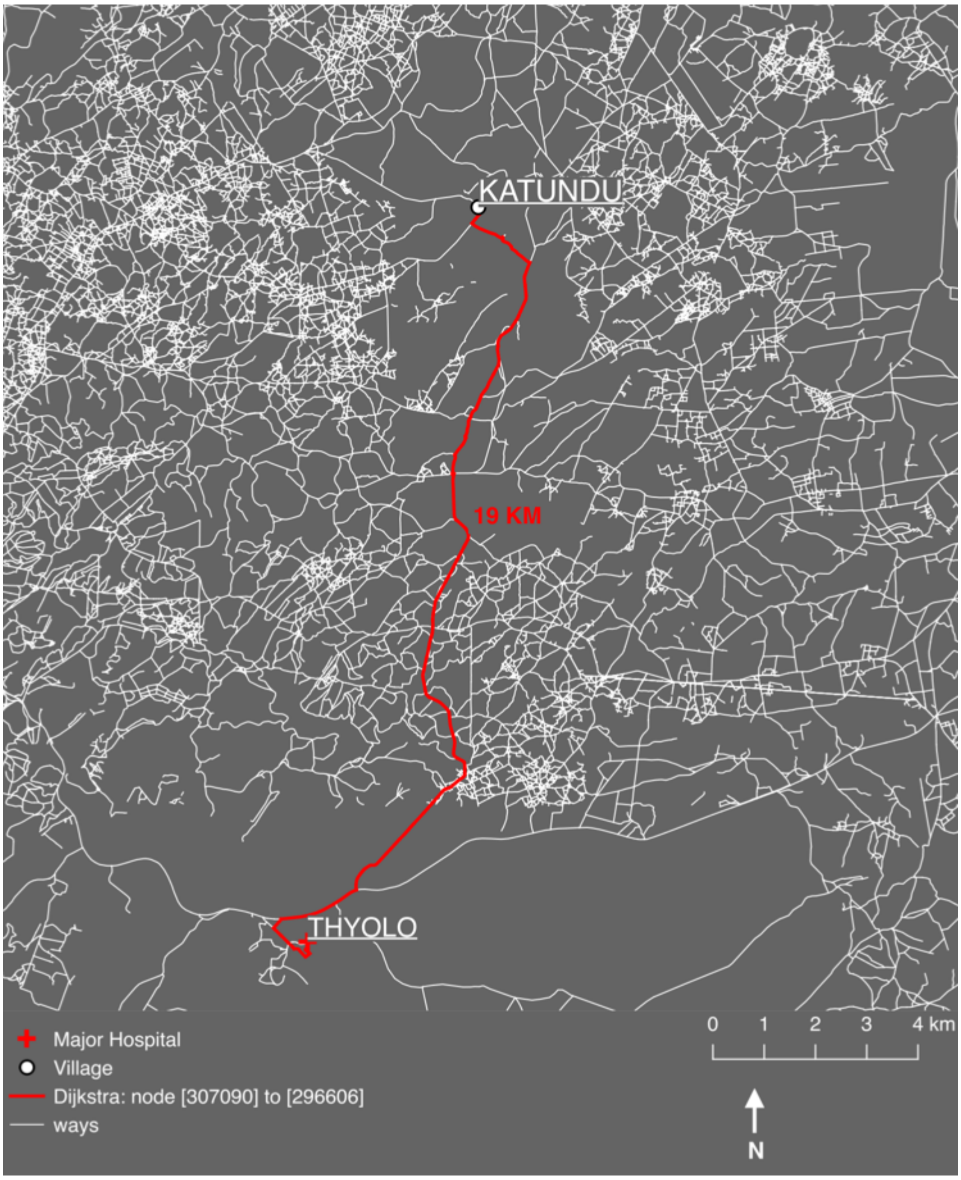


Figure 5.21 Example of Dijkstra algorithm calculating the shortest path from one of the visited village to the visited major hospital

5.5 LM and RF Outcomes

The third sub-objective is concerned with finding statistical relationships between the remoteness indicators and the social vulnerability index to accurately select and weigh the remoteness variables to develop the final proxy. The created remoteness indicator data (X-variables) and the SoVI (Y-variable) is structured in a final data matrix. Through machine learning techniques the X-variables will be correlated to the Y-variable. The data matrix is split into a training data set which contains a fixed 60% of the data matrix and a test data set which contains a fixed 40% of the data matrix. For this research the composite indices that predict vulnerability need to be carefully chosen; the choices made in weights and aggregations made by analysts who create indices are often based on weak rationale (Fakete, 2010). This research will extract the relevant indicators and their weights, as in the creation of an index, a data-driven methodology is often described as preferable (Benini, 2015). The Netherlands Red Cross has successfully applied machine learning techniques in the development of a priority index model for typhoon Haiyan in the Philippines (510, 2016). In Machine learning, algorithms are used that learn and make predictions on data. These types of algorithms avoid using strict program instructions by making data-driven predictions and decisions. It is designed to overcome the lack of argumentation of weighing factors, and different types of statistical analysis methods can be compared in a relative fast manner.

Predictive accuracy

Machine learning's advantage over traditional statistical methodology is that it uses the predictive accuracy of models; a part of the observations or data is used for 'testing' and another part is used for 'training' the accuracy of the model's predictions. Different techniques can be quickly compared to choose the model with the highest predictive accuracy. Furthermore, if new relevant data sets are found, which is a common phenomenon within the humanitarian data environment, machine learning algorithms can be re-run with relative ease. An example is the water point data set for Malawi, the functioning water pumps will vary greatly within a short time span. The new data set can be run within machine learning algorithms instantly, without much preparation. The vulnerability within a country, district, or municipality may change rapidly when access to a hospital or water point may be blocked. New and most current assessments may be run with relative ease.

Data use

In the case where limited amount of data is available, all the data can be used in the analysis. Irrelevant data is delineated by the algorithm, and will thus eliminate pre-processing choices in a timely and unbiased manner. With a classic multivariate linear regression model this is often the case, due to multicollinearity, normal distribution, and linearity requirements and assumptions, the limited amount of data can often be a problem in this regard.

Model use

In this research, supervised machine learning will be used, as the result or label is known (social vulnerability index adapted from the RCMRD index). Many supervised learning methods are available; Decision Trees, Naïve Bayes Classifications, Neural Nets, Logistic Regression, Support Vector Machines, and Random Forest (Caruana & Niculescu-Mizil, 2006). Due to the scope and timeframe of this research, the multivariate linear regression is compare to a random forest regression model. The limitations and benefits of each model will be touched upon.

5.5.1 Data exploration

All the steps described in this section of the research can be found in the R-script (*Appendix B*); all the analysis steps are described there, shortly explained within the hashtag per section. The data matrix is checked for missing values ('NA's). Different choices can be made to deal with missing values; the values are actually missing, or there is no data or output for a certain column. Sometimes it is possible to accurately 'fill' the missing data values, based on certain assumptions, estimations or by comparing data obtained from other (independent) sources. In the data matrix, a very limited amount of data was missing. For some TA's the road density and length had a missing value, and for these data, due to missing roads within OSM, it was chosen to fill the data with the minimum amount of the data matrix for that specific indicator. For the lakes and national parks within Malawi there was no data, and it can be assumed that there are indeed no human settlements within lakes or wild parks and these entries were therefore removed from the data set.

The data matrix is filled with data with different scaling and units. To be able to make a composite score after the weighing of the indicators it is chosen to normalize the data according to the SoVI, from zero to hundred. A 'rescale' function is created in R, in which the following function is used:

```
rescale <- function(x) (x-min(x))/(max(x) - min(x)) * 100
```

For all the variables this transformation is done during the pre-analysis. The rescaling of the data does not affect the structure of the data and can be done without affecting the analysis (Breiman, 2001). In the interpretation of the analysis this needs to be taken into account.

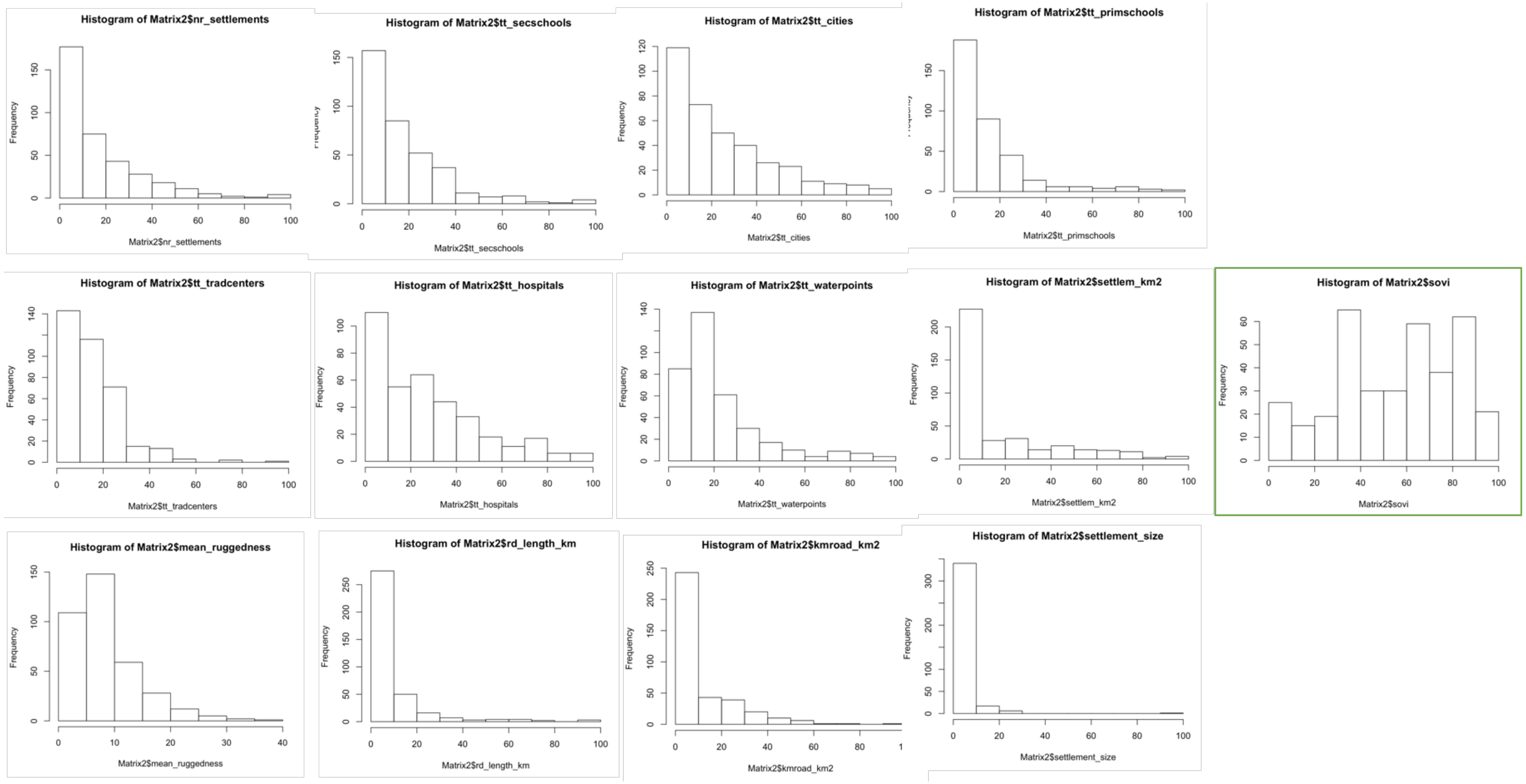


Figure 5.22-A Frequency distribution histogram of original variables

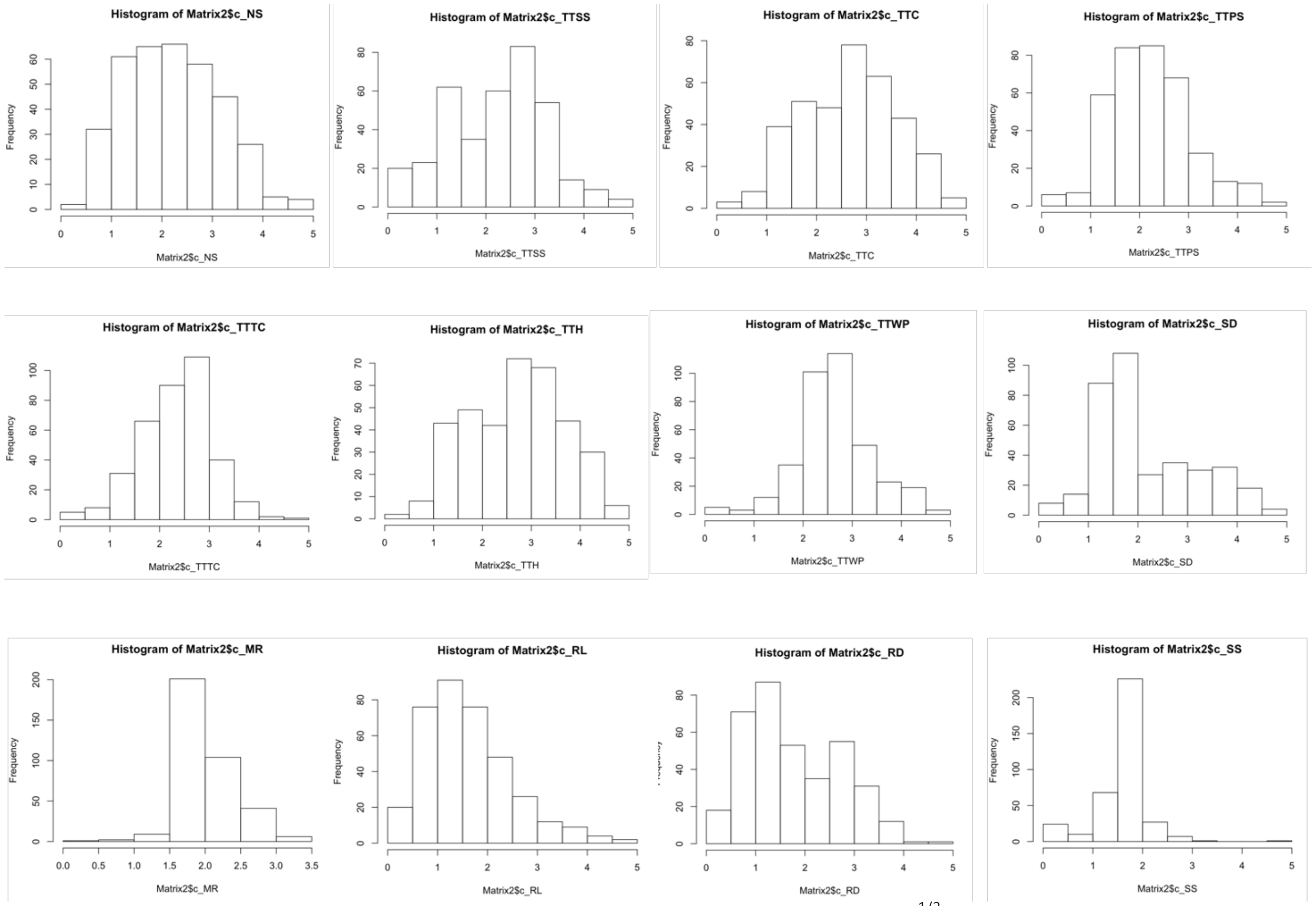


Figure 5.22-B Frequency distribution histograms of cube root ($x^{1/3}$) X-variables

After evaluating the frequency distributions of the variables (*Figure 5.22-A),* it was chosen to transform the data to better fit the normal distribution assumption for the variables. SoVI sufficiently meets the assumption of normality. The X-variables are all right-skewed and were primarily Log-transformed, which improved the distribution, but normality could still not be assumed sufficiently. Finally, it was chosen to use de cube root ($x^{1/3}$) transformation for all the X-variables. The outputs of the cube root transformations for the X-variables are displayed in *Figure 5.22-B*.

All cube root transformed X-variables can be assumed to be normally distributed following the frequency histograms from *Figure 5.22-B.* It can be argued that the frequency distribution of the cube root road length indicator (c_RL) is still slightly right skewed. However, it has been chosen to use the cube root, as it was the transformation that best fitted the distribution to normality.

For several variables outliers could be detected. For road length, road density, and settlement size. For the settlement size variable, there is no reason to assume incorrect data. For the road length and road density variable, both depend on the completeness of OSM data. As indicated in *Figure 5.19*, the road density was used as a measure for OSM completeness and can be clearly seen when the road network is spatially visualised. In the case of the road length variable the outputs of the boxplots are depicted in *Figure 5.23*. In boxplot B the normal road length variable is depicted in a boxplot, with 33 outliers. In boxplot C the log transformed road length variable is depicted, with 15 outliers. In boxplot A, the cube root transformed road length variable is depicted, with 9 outliers. The cube root transformed variable fits the assumed normality best, and shows the least amount of outliers, and thus seems the most appropriate. The outliers are all scattered across the high spectrum of the distribution. These outliers are TA's that are either larger, have a higher road density, or have a higher level of OSM completeness, and are thus essential to take into account within the analysis.

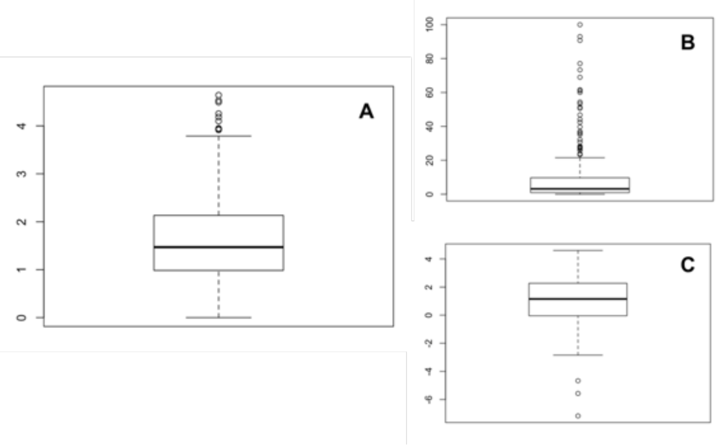


Figure 5.23 example boxplot for A) cube root transformed road length B) normal road length C) Log transformed road length variable with outliers.

Train and test data

The following step that was conducted is the splitting of the data set into a test and a train data set, this is done before further exploration of the data so that the train and test set can be explored further individually. The train and test data set are created to have a separate data set to test the predictive accuracy on. This will be explained further in *Sections 5.5.3* and *5.5.4,* where the LM and the RF model outputs are explained. The splitting is done based on a 60% training data set and a 40% test data set, which are numbers commonly used within machine learning techniques with comparable data set sizes (Rohrer, 2016). The train and test data set are randomly split, but kept the same for the rest of the process. The train and test set are used for both the LM- and RF-model. Within R-studio, the *set.seed* function is used to create the random data sets, and maintain it for the rest of the analysis. Both the linear regression and the random forest model use the same train and test set. This is done, so results can be compared with each other, avoiding any change to be related in different train and test sets.

Multicollinearity

The final step of the data exploration is to check any dependency among the X-variables, the correlation between the X-variables: multicollinearity. A correlation plot is created for the entire data matrix. The output is displayed in *Figure 5.24*. The gradient of blue indicates the measure of positive correlation between indicators where darker values are higher, the gradient of red is the measure of negative correlation between indicators where darker values are higher (negative correlation). It can be seen that the SoVI's strongest correlation is between travel times to hospitals (TTH), secondary schools(TTSS), cities (TTS), settlement density(SD), and road density(RD). Furthermore, there exists multicollinearity between the travel times for hospitals, cities, trading centres, primary schools, secondary schools, road density, and settlement density as can be expected. The travel times are based on GIS analysis based on the road network and the settlement layer of Malawi. More than 50% of the indicators show multicollinearity, as mentioned earlier, this can be expected with developing indicators with limited data sources based on one common denominator: 'remoteness'. The multicollinearity is addressed in the proceeding steps for the linear regression model. The Random Forest model does not operate under the assumption that the X-variables are not correlated, and addressing of the multicollinearity problem is therefore not necessary for the Random Forest model.

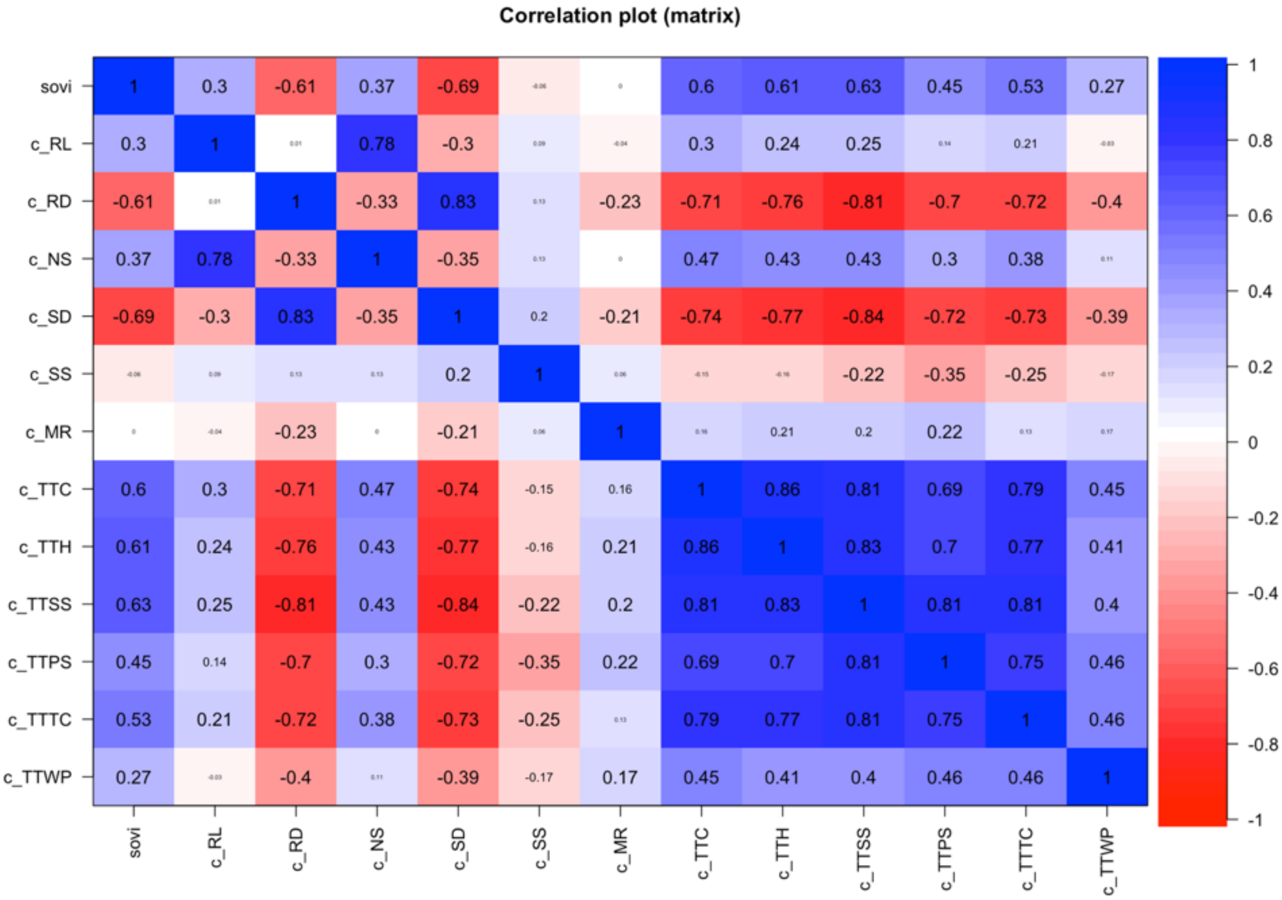


Figure 5.24 Correlation plot for the entire matrix, where the gradient of blue indicates positive correlation, and red negative. RL: Road Length – RD: Road Density – NS: Number of settlements – SD; Settlement Density – SS: Settlement Size – MR: Mean Ruggedness – TTC: Travel Time to Cities – TTH: to Hospitals – TTSS: to Secondary Schools – TTPS: to Primary Schools – TTTC: to Trading Centres – TTWP: to Water Points.

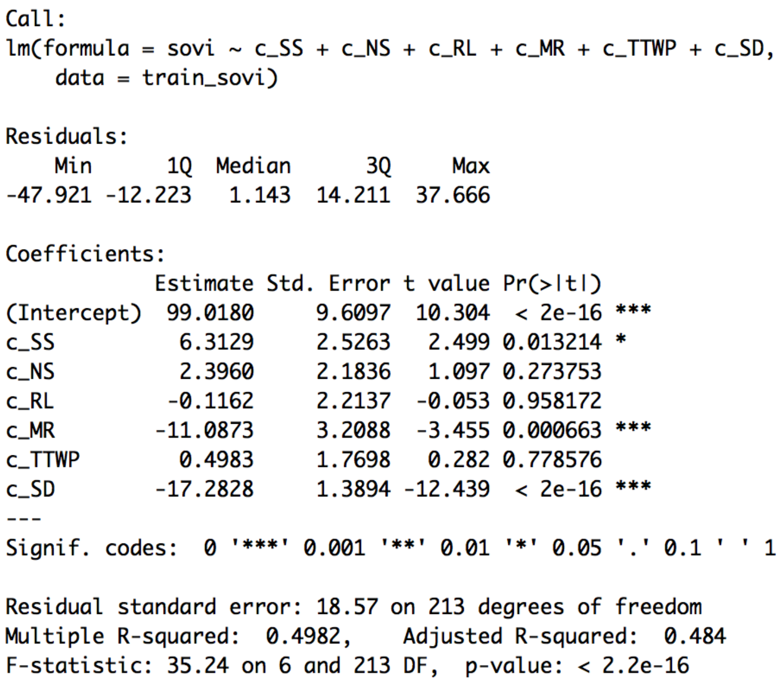
5.5.2 Linear Regression Model Training

The multivariate linear logistical regression model is run several times to select the best predicting set of variables. The selection of the X-variables is iterated, based on the multicollinearity between the variables. Different combinations of X-variables are run, and the significance of the variables and the variance explained by the models (adjusted R-squared) are compared. The final LM performance estimates need to be unbiased, therefore the model is tested on data that was not taken into account for the tuning of the model, including the variable selection. For this reason, only the training data is taken into account in the analysis (60% of the total data set).

With all the variables removed where multicollinearity exists, except for one, the highest R^2_{adj} score is sought after. For every iteration one single variable is used that falls within the list of variables where multicollinearity is deemed to be too high (>55%):

- Travel Time to Primary Schools (R^2_{adj} of LM with variable: 0.2754)
- Travel Time to Trading Centres (R^2_{adj} of LM with variable: 0.3178)
- Travel Time to Cities (R^2_{adj} of LM with variable: 0.3765)
- Travel Time to Hospitals (R^2_{adj} of LM with variable: 0.4127)
- Travel Time to Secondary Schools (R^2_{adj} of LM with variable: 0.4284)
- Road Density (Adjusted R-Squared of LM with variable: 0.5082)
- Settlement Density (Adjusted R-Squared of LM with variable: 0.5083)

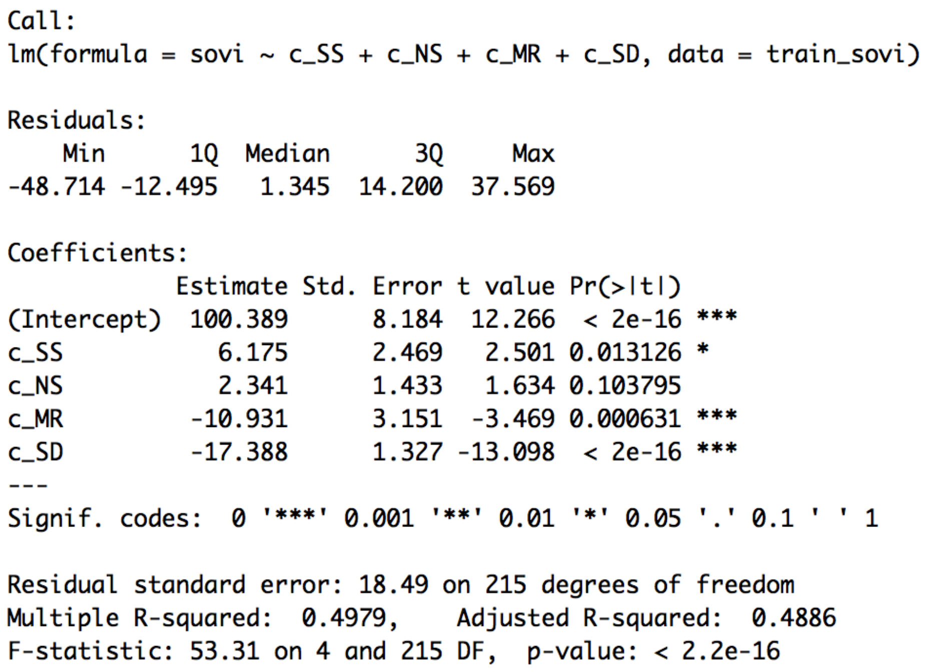
Leaving out Travel Time to hospitals, cities, trading centres, and primary schools, road density, and taking settlement density into account, the model prediction is highest for the entire data set: an adjusted R^2_{adj} of 0.484. The output is shown in *Figure 5.25.*



```
Call:
lm(formula = sovi ~ c_SS + c_NS + c_RL + c_MR + c_TTWP + c_SD,
    data = train_sovi)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.921 -12.223   1.143  14.211  37.666 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  99.0180     9.6097  10.304  < 2e-16 ***
c_SS          6.3129     2.5263   2.499 0.013214 *  
c_NS          2.3960     2.1836   1.097 0.273753    
c_RL         -0.1162     2.2137  -0.053 0.958172    
c_MR        -11.0873     3.2088  -3.455 0.000663 ***
c_TTWP        0.4983     1.7698   0.282 0.778576    
c_SD        -17.2828     1.3894 -12.439  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.57 on 213 degrees of freedom
Multiple R-squared:  0.4982,	Adjusted R-squared:  0.484 
F-statistic: 35.24 on 6 and 213 DF,  p-value: < 2.2e-16
```

Figure 5.25 Output of multivariate linear regression, selecting single multicollinearity variable with highest adjusted R-squared for entire data set (Settlement Density)

Subsequently the ‘step’ function is used in R-Studio. This function provides an automated selection of the combination of variables that results in the highest R^2_{adj} score. The output is shown in *Figure 5.26.* The adjusted R-Squared increases to 0.4886, which is limited. The step function removes the travel time to water points. All the remaining variables have significant p-values.



```
Call:
lm(formula = sovi ~ c_SS + c_NS + c_MR + c_SD, data = train_sovi)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.714 -12.495   1.345  14.200  37.569 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  100.389      8.184  12.266  < 2e-16 ***
c_SS           6.175      2.469   2.501 0.013126 *  
c_NS           2.341      1.433   1.634 0.103795    
c_MR         -10.931      3.151  -3.469 0.000631 ***
c_SD         -17.388      1.327 -13.098  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.49 on 215 degrees of freedom
Multiple R-squared:  0.4979,	Adjusted R-squared:  0.4886 
F-statistic: 53.31 on 4 and 215 DF,  p-value: < 2.2e-16
```

Figure 5.26 Output LM after using ‘step’ function for automated selection of variable combinations with the highest possible adjusted R^2_{adj} outcome.

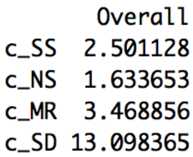
The R^2_{adj} value indicates that 48.86% of the variance in the Y-variable can be explained by the sub-set of X-variables. The normal R^2 is 0.4979. This model can be written in the following equation:

$$SoVI = 100.39 + 6.18\, X_1 + 2.34\, X_2 + -10.93\, X_3 + -17.39\, X_4$$

SoVI	*Social Vulnerability Index*
X_1	*(Settlement Size)*$^{1/3}$
X_2	*(Number of Settlements)*$^{1/3}$
X_3	*(Mean Ruggedness)*$^{1/3}$
X_4	*(Settlement Density)*$^{1/3}$

The coefficient assigned to each X-variable represents the mean change in the Y-variable for one unit of change in the X-variable. The negative or positive values can be interpreted as follows: according to the coefficients, and increased settlement size results in the increase of social vulnerability. The increased number of settlements also results in the increase of social vulnerability. For both mean ruggedness and settlement density, the increased results in the decrease in social vulnerability. The increased settlement density is expected to have a decreased SoVI score as a result; cities show lower social vulnerability. Less mean ruggedness results in higher SoVI for Malawi, this is somewhat unexpected. The accessibility to a community through rugged terrain, for physical transport as well as radio, phone, and internet reception will be less than smooth terrain. For the Malawi context, however the lower and flat regions have a higher SoVI score; these areas are more often struck by floods and droughts, which will lead to a higher social vulnerability within these communities in Malawi.

The relative importance of the variables is determined through the t-statistic value of the individual model variable *(Figure 5.27)*. The t-statistic is a measure that indicates the difference, relative to the variation in the sample data.



```
        Overall
c_SS   2.501128
c_NS   1.633653
c_MR   3.468856
c_SD  13.098365
```

Figure 5.27: absolute t-statistic predictors of the LM

The outputs indicate that the Settlement Density and the Mean Ruggedness are the most important indicators. This shows that for the model, the settlement density and the mean ruggedness of an area are most relevant in predicting social vulnerability in Malawi from the created remoteness indicators.

For the multivariate linear regression model to be valid, the following assumptions need to be met: linearity of residuals, heteroscedasticity of residuals, and normality of residuals. In *Figure 5.28* the residuals are plot against the fitted predicted values.

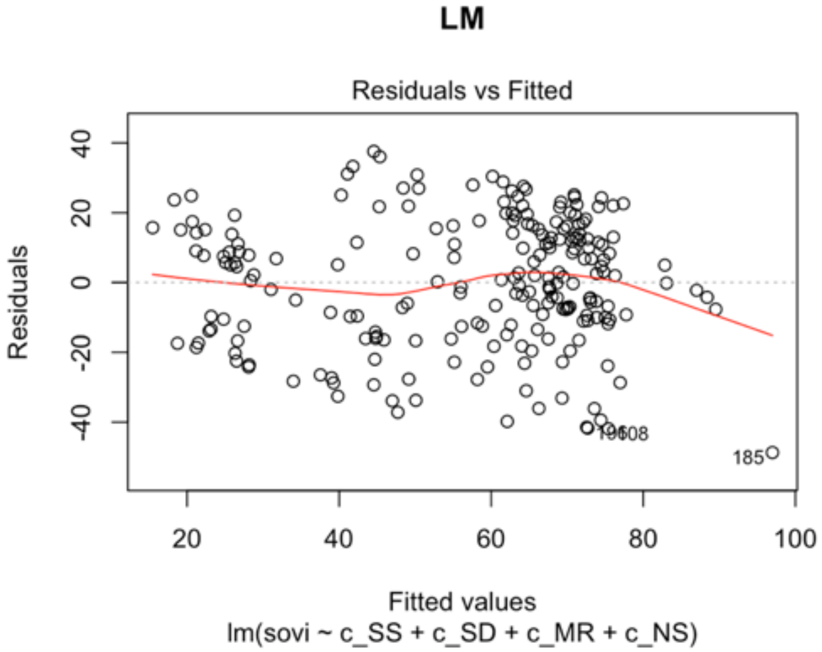


Figure 5.28 Residuals plotted to fitted values of LM

For the assumption of linearity, the residuals are spread randomly around the 0 line. It can be assumed that linearity of the relationship is reasonable. The residuals are similar for all the values, although the high values rely on a limited amount of observations, heteroscedasticity is however sufficiently met. In regard to normality, the model tends to over-estimate in the high values (SoVI >80), and relies on limited observations.

5.5.3 Random Forest Model Training

As described earlier, the RF model relies on less assumptions that the multi variate linear regression model. As with any other model, it does rely on the assumption that the sampling is representative for reality. The algorithm behind the RF model can handle multicollinearity and non-linearity. The RF is run for the cube root transformed as well as the normal variables. The results of the best fit model are presented below.

The model with the highest variance explained and R^2-score is the one using the cube root transformed X-variables to predict the SoVI. Twelve X-variables are used in this model and are displayed in *Figure 5.29-A.* For every time the model is run, some randomness in the fitted values exists. The highest R^2 score after 50 iterations is 0.6477, which means that **64.77%** of the variance in SoVI is explained by the RF model. The RF algorithm runs through 500 decision trees, trying eight trees at each split. After +/- 90 trees the error rate stabilizes *(Figure 5.29-B).* No information regarding significance of variables or reason of influence on the Y-variable is given by the model. The relative importance of the X-variables can be interpreted by the variable importance plot in *Figure 5.29-A.* The x-axis indicates the percentage in increase of Mean Squared Error (MSE) on a scale from 0 to 100%, in the case the concerning X-variable is permutated.

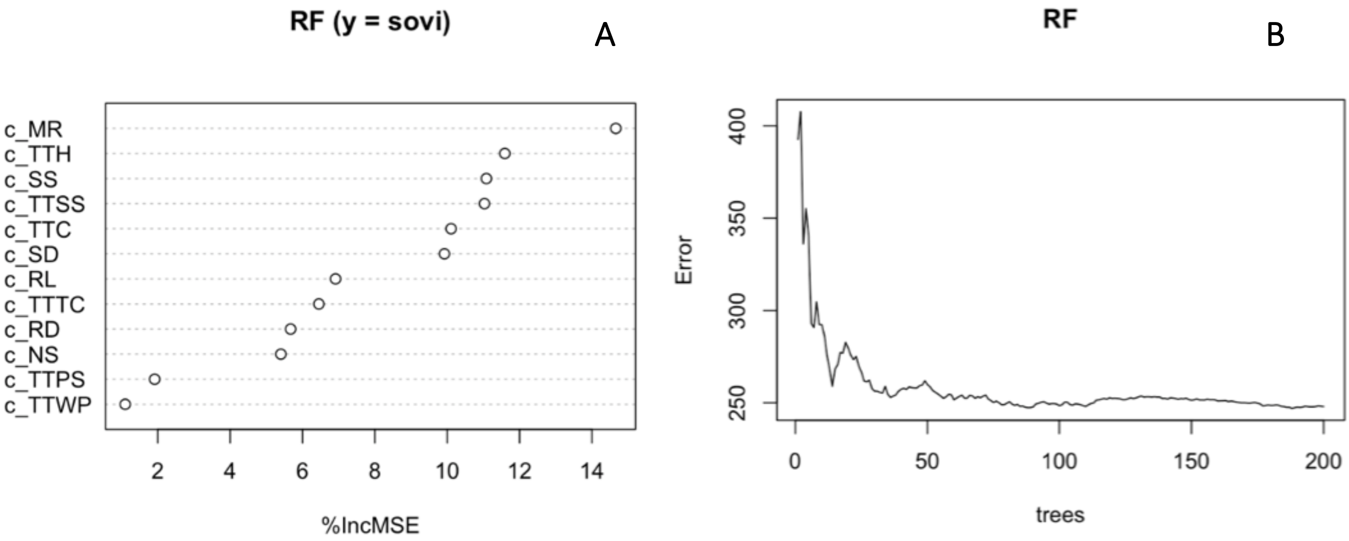


Figure 5.29 A) Variable importance plot of RF B) Error vs number of decision trees

The mean ruggedness (MR) is the most important X-variable in relation to predicting the SoVI using the RF algorithm, followed by the travel time to hospital (TTH), settlement size (SS), travel time to secondary schools (TTSS), and travel time to cities (TTC). The settlement density (SD) on the sixth place, shows to be less important comparing the RF with the LM model. It is noticeable that the travel time to hospitals and the travel time to secondary schools are important in the RF model. The travel time to water points and primary schools are the least important X-variables. Road length, travel time to trading centres, road density, and number of settlement have an average importance from 5 - 7%.

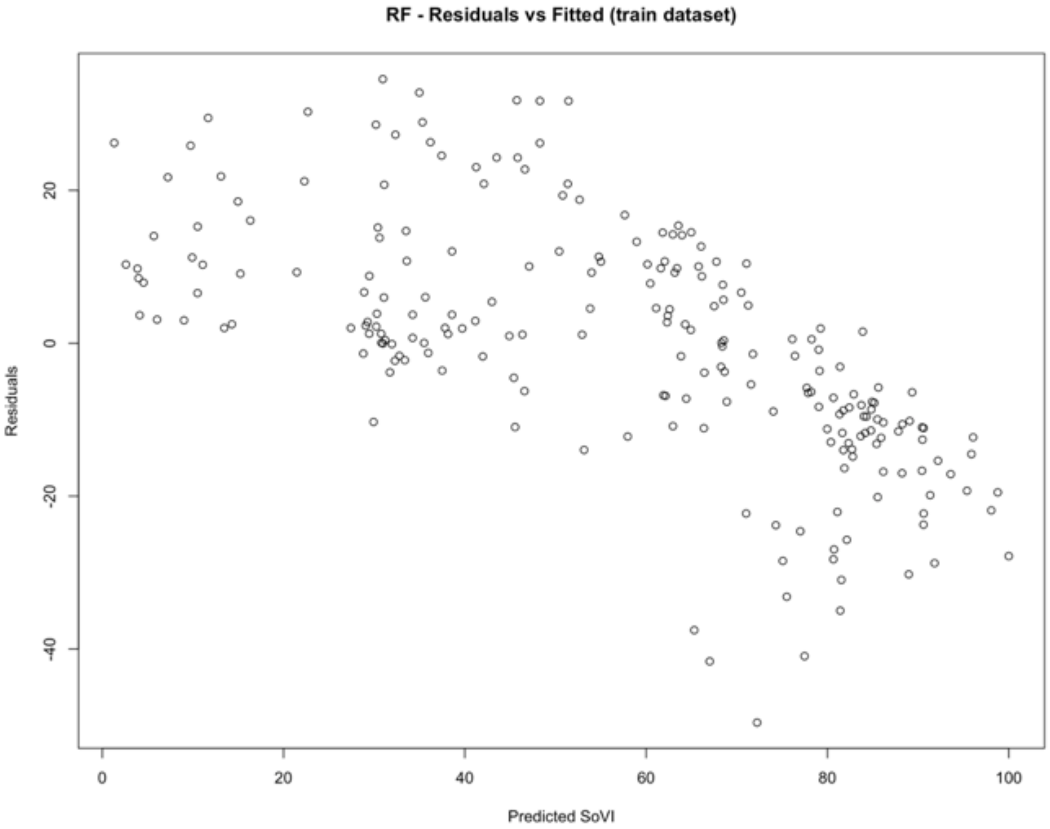


Figure 5.30 Absolute residuals to predicted values of RF

The residuals of the prediction (predicted SoVI – measured SoVI) is plotted against the predicted values using the train data set. This results in a non-linear pattern ***(Figure 5.30).*** The social vulnerability in areas with high social vulnerability is under-estimated and the low values are over-estimated. This means that the RF 'flattens' reality, resulting is smaller differences in SoVI in the different TA's. The model fit is less for low and high values.

5.5.4 Model validation

The LM and RF best fitting models have been defined. The model is validated through cross validating the data model to the test data set created as described in *Section 5.5.1* train and test data paragraph. The test data consists of 40% of the remaining data which was not used in calibrating the parameters of both models; the so called out-of-sample validation.

Multivariate linear regression model

The R^2_{adj} of the LM when applied on the test data is 0.53. This indicates that 53% of the variance in SoVI variable of the test data set is explained by the LM. The LM has a root mean squared error (RMSE) of 18.07. The RMSE is calculated through the following formula: RMSE = sqrt(mean((pred – obs)^2). The RMSE of the training data is 18.3. The RMSE has decreased with 1.14%, which is acceptable and does not indicate any deficiency or overfitting in the model. The average prediction of the LM model deviates 18.07 SoVI from the original (measured) SoVI. *Figure 5.31* depicts the residuals versus the predicted (fitted values) and the measured versus predicted values of the SoVI for the test data set. In the prediction versus measured score, the overall predictive accuracy is the same across the entire range of possible scores. The mid-section seems to be less accurate, as the scores between 40 and 70 seem to deviate more than in the lower and higher scores.

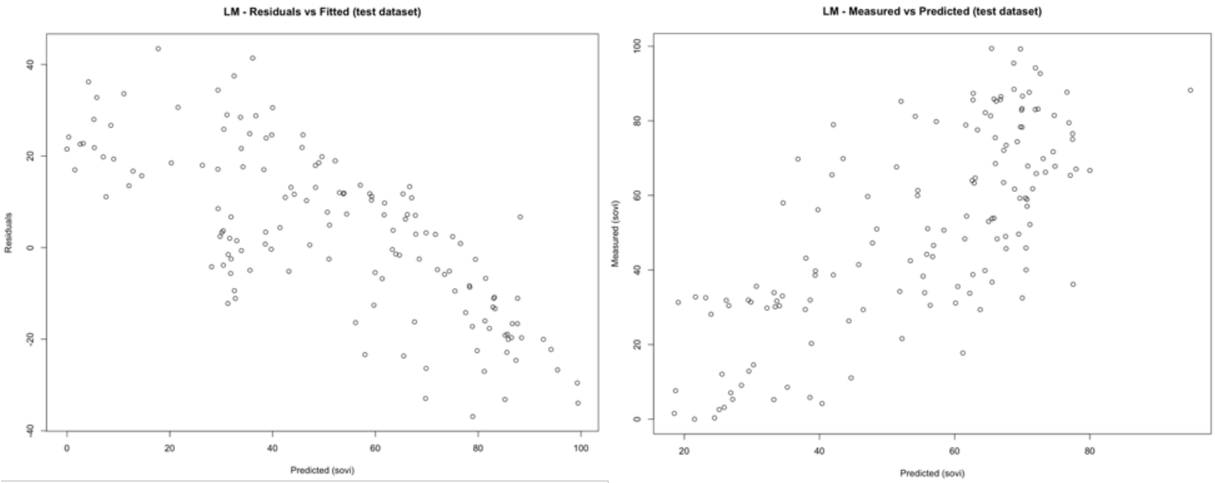


Figure 5.31 Predictive accuracy of LM on test data set ($R^2_{adj:}$ 0.53, RMSE: 18.07)

The measured SoVI according to the existing social vulnerability index (RCMRD, 2015), and the predicted SoVI using the LM are displayed in *Figure 5.32*. The ability to identify the most vulnerable TA's within Malawi with the help of a proxy is important. The identification of the most vulnerable TA's contributes to the ability to prioritise in decision-making within humanitarian interventions. The TA's with the highest SoVI score are mostly identified correctly with the LM. Again, the visualisation shows that the high and low scores in SoVI tend to be more accurate than the mid-section scores. The values between 30 and 70 tend to show more variation. The LM result based on the test data set is applied to the entire data set and contribute to the spatial representation displayed in *Figure 5.32.* Overall the pattern shows many similarities; the major cities show lower SoVI values, and the more rural areas show higher SoVI values.

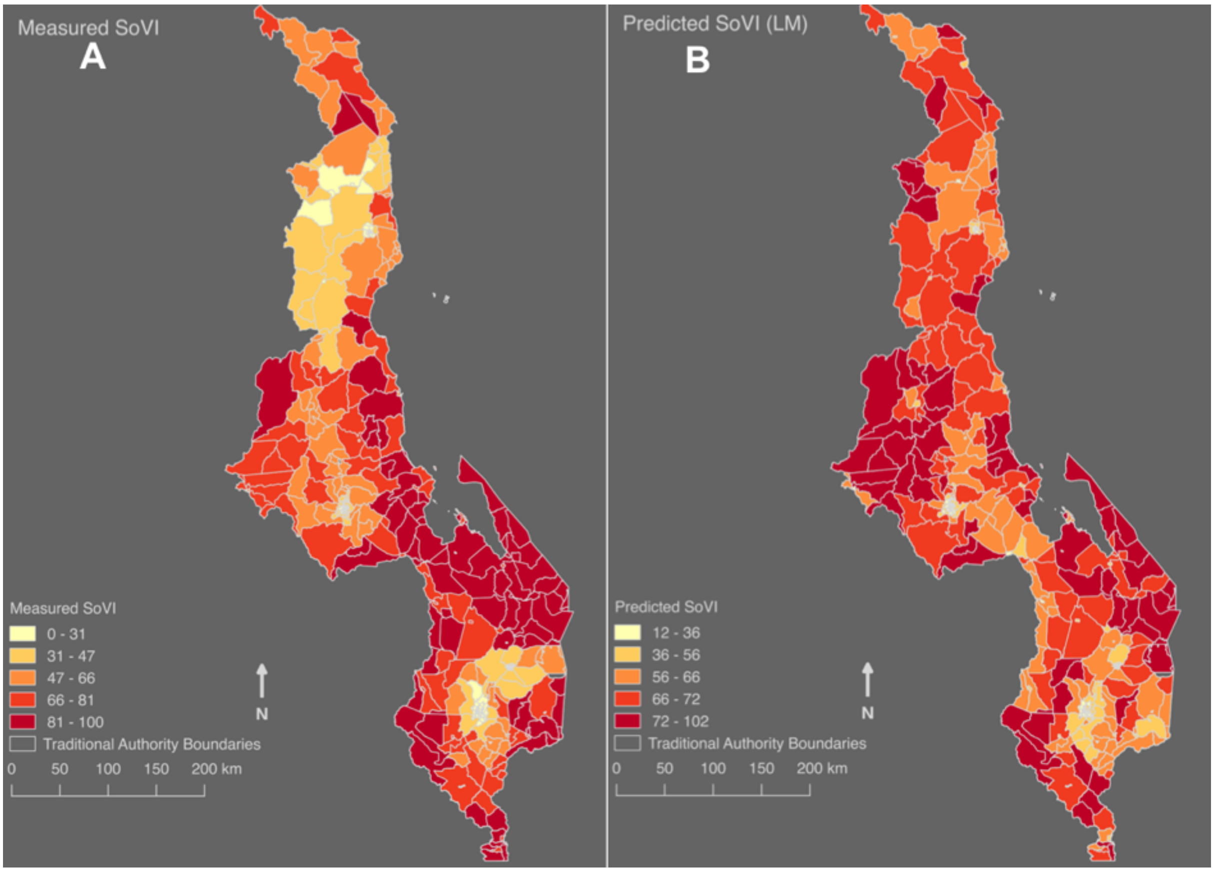


Figure 5.32 (A) Measured SoVI values (B) Predicted SoVI values using LM

Random Forest Model

The RF model is applied to the test data set. This results in a R^2 of 0.64 and a RMSE of 16.4. On average the prediction of SoVI deviates with a value of 16.4. The decrease of 4.2% compared to the RF model run on the train data set indicates that there is no model deficiency or overfitting. The measured versus the predicted values are presented in *Figure 5.33*. The distribution shows a similar pattern as the LM prediction. The low values tend to overestimate, and the higher values tend to under-estimate the measured SoVI values.

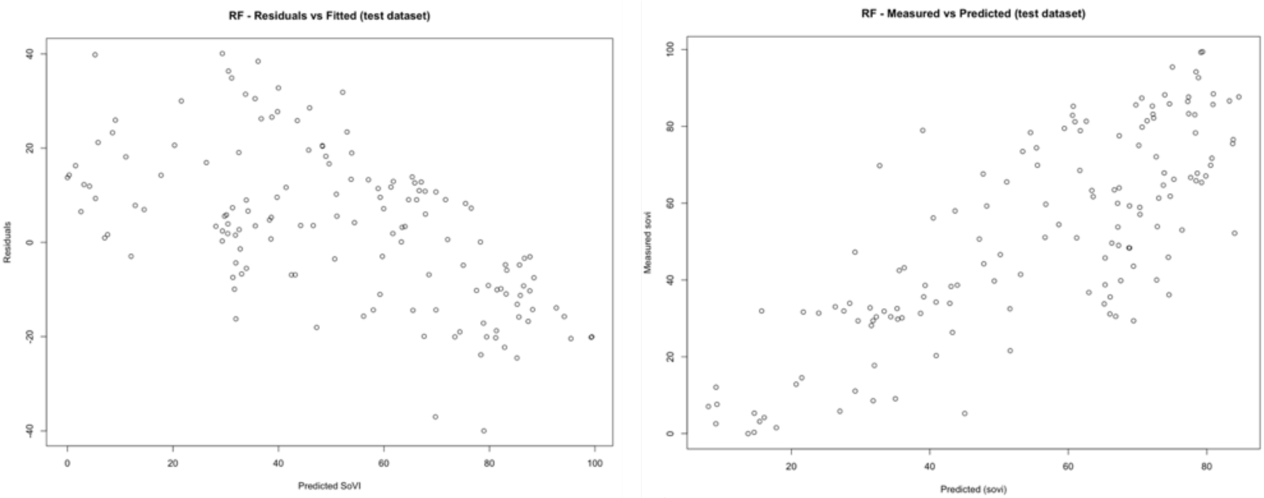


Figure 5.33 Predictive accuracy of RF R^2 of 0.615 and RMSE of 16.4

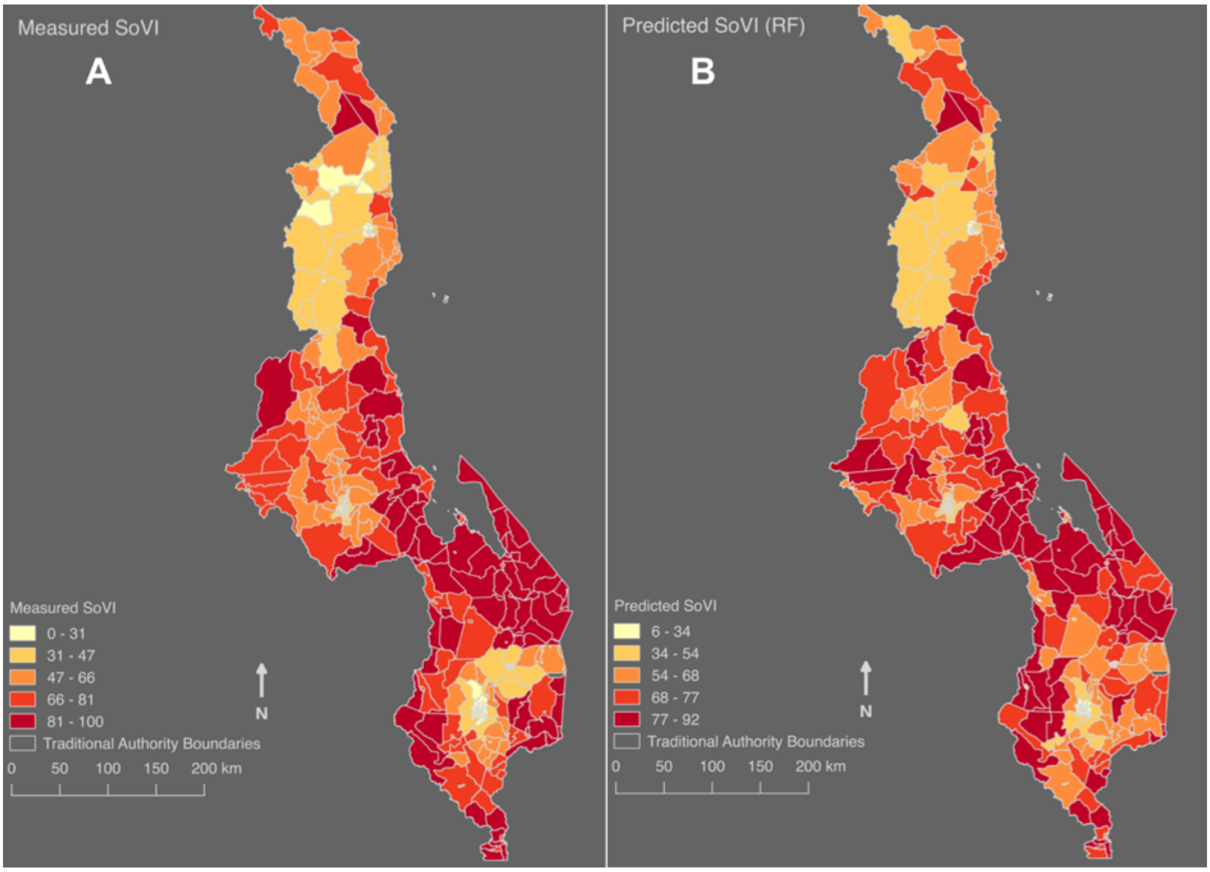
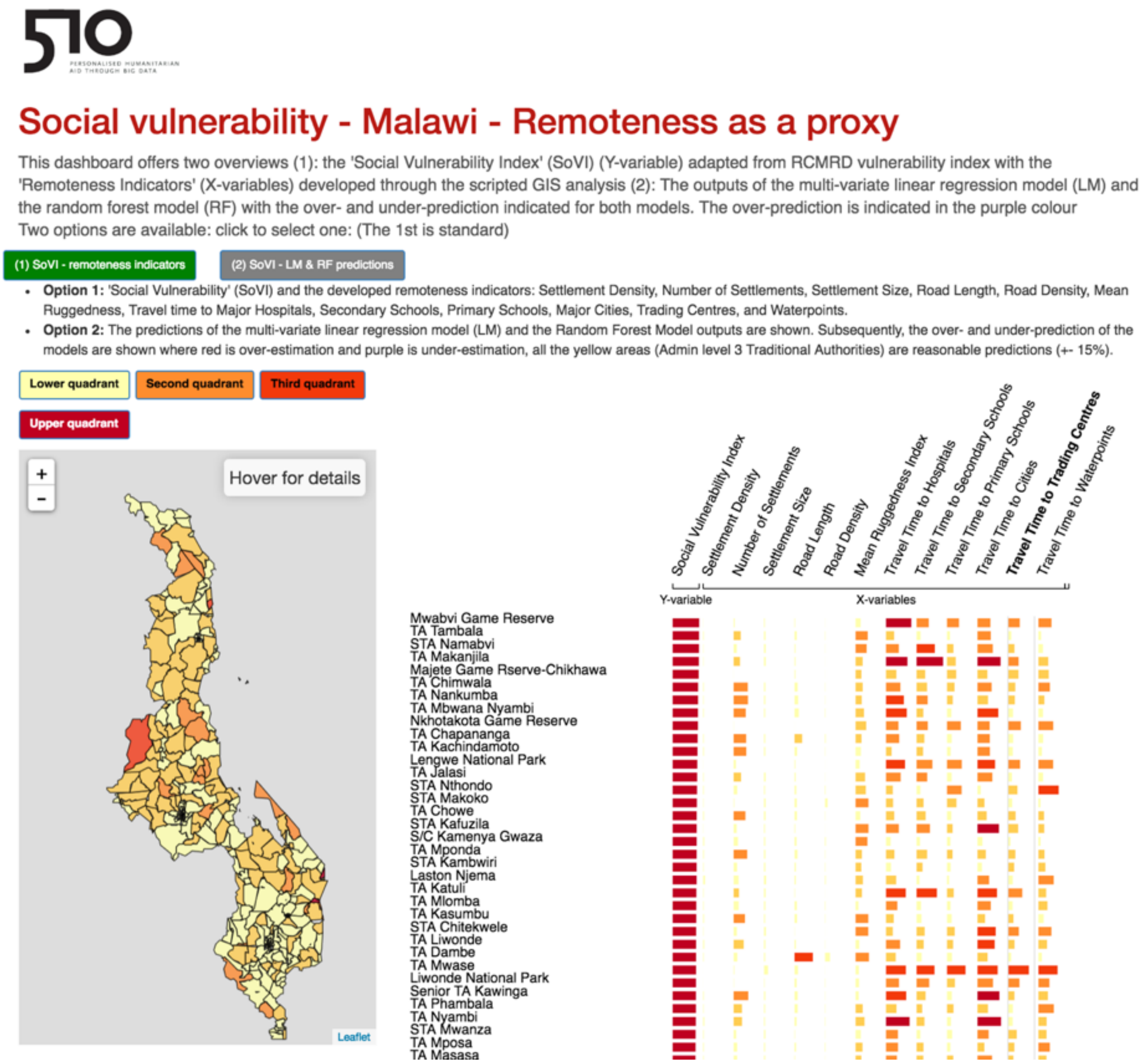


Figure 5.34 (A) Measured SoVI values (B) Preidcted SoVI values using RF

Figure 5.34 displays the difference between the measured SoVI and the predicted SoVI using the RF prediction for the complete area of Malawi. More high and low SoVI values are predicted correctly compared to the LM outcomes. Especially in the northern region the lower values are predicted more accurate than with the LM prediction (*Figure 5.32*). The high values in the mid-eastern region are subsequently more accurate in the RF prediction.

5.5.4 Online Dashboard

The results of the data matrix, the remoteness indicators, the SoVI, and the prediction of the LM and RF models are presented in an interactive online-dashboard. The dashboard is created using Java-script and HTML coding. The geographic representation is created using a Leaflet extension. On the presented map, the TA’s can be hovered over, and details on the selection will become visible for that TA. The overview of all the data is presented in a Table on the right side of the dashboard. The Table can be sorted per parameter when selected.



Social vulnerability - Malawi - Remoteness as a proxy

This dashboard offers two overviews (1): the 'Social Vulnerability Index' (SoVI) (Y-variable) adapted from RCMRD vulnerability index with the 'Remoteness Indicators' (X-variables) developed through the scripted GIS analysis (2): The outputs of the multi-variate linear regression model (LM) and the random forest model (RF) with the over- and under-prediction indicated for both models. The over-prediction is indicated in the purple colour
Two options are available: click to select one: (The 1st is standard)

- **Option 1:** 'Social Vulnerability' (SoVI) and the developed remoteness indicators: Settlement Density, Number of Settlements, Settlement Size, Road Length, Road Density, Mean Ruggedness, Travel time to Major Hospitals, Secondary Schools, Primary Schools, Major Cities, Trading Centres, and Waterpoints.
- **Option 2:** The predictions of the multi-variate linear regression model (LM) and the Random Forest Model outputs are shown. Subsequently, the over- and under-prediction of the models are shown where red is over-estimation and purple is under-estimation, all the yellow areas (Admin level 3 Traditional Authorities) are reasonable predictions (+- 15%).

Figure 5.31 Screenshot of online dashboard indicating results of SoVI remoteness proxy.

The data matrix is transformed to a json format, and traditional authority boundaries to geojson to adhere to the dashboard settings used. The code for the web application (java script and html) and the steps on how to publish the dashboard can be found in *Appendix B.*

The link to the online dashboard:
https://jwilbrink.github.io/Remoteness-proxy/

6 Discussion

This chapter provides an overview of the most important findings during this research. Subsequently, an overview of the shortcomings, recommendations, and possible future research is given.

6.1 Research overview

Humanitarian organisations lack the access to complete, granular, and up-to-date geo-spatial information on the vulnerability of communities within their operation areas. These organisations currently gather information from the government on a less granular level, and collect granular and recent data through their Monitoring and Evaluation (M&E) departments by conducting field assessments (e.g. VCAs). These assessments are often project-based and therefore aimed at a certain geo-spatial area. This makes it difficult to fulfil the aspects of granularity, timeliness, and completeness. To be able to make more evidence-based decisions, the humanitarian sector is in need of a tool that can assess vulnerability in a granular, timely, and complete manner. The tool should use openly available data, and analysis should be conducted with the use of open-source tooling; this will enhance the replicability of such a tool in other geo-spatial areas. Malawi is selected to create remoteness indicators using open-source GIS tools (Qgis, Postgis, PgRouting, JOSM, and Leaflet). The remoteness indicators are calibrated through a field-study in Malawi, and the points of interest used to create the indicators are validated for their completeness and accuracy. The calibrated remoteness indicators are trained on an existing SoVI for Malawi to create the final proxy for social vulnerability in Malawi. The indicators are trained using a multivariate linear regression technique, and a Random Forest regression technique. Based on the overall usability and predictive accuracy, a conclusion is drawn as to which model is the most relevant. The results are visualised geo-spatially in an online dashboard in an interactive manner where the different results may be compared with the aim to present the findings transparently.

6.2 Main findings, limitations, and recommendations

RQ 1: What indicators are used to create a social vulnerability index for natural hazards?

The primary objective, related to research question 1, was to find a useful and fitting definition of vulnerability to natural hazards. Vulnerability with regard to natural hazards are described in various ways and many frameworks exist to define this type of vulnerability. Recent DRR literature and leading organisations within this field, define vulnerability to natural hazards in general as ‘social vulnerability’ when the type of hazard is not taken into account. Often, the risk or potential for harm is measured based on the components; hazard characteristics, exposure, and social vulnerability, often referred to as ‘vulnerability’, see *Figure 4.2*. When an assessment is done for a communities’ vulnerability to a specific type of natural hazard, e.g. a flood, the exposure and hazard characteristics become important. The exposure, often referred to as ‘physical vulnerability’ is related to the built structures within communities. The exposure is only important to take into account when the type of hazard is known. For example, a wooden house may be less resistant to a flood or a hurricane, but is less likely to be affected in case of an earthquake. The hazard characteristics are often defined as external, and are therefore not taken into account when defining vulnerability to natural hazards.

The next step in this research was to find a way to quantify social vulnerability for Malawi on a granular level. The most described quantification of social vulnerability is through the use of a vulnerability index. After elaborate research on these indices, to describe social vulnerability the elements within sensitivity and adaptive capacity of communities are commonly used. Then the search for existing data on social vulnerability commenced. The primary goal was to use the data generated by the Netherlands Red Cross and its Partner National Societies (PNS) for Malawi using the VCA methodology. After elaborate research on the VCA and the limited data availability with regard to these VCA's, the need for a tool that could generate vulnerability data became more evident. After a search for an alternative secondary open data source, the RCMRD vulnerability index for Malawi was decided on. This index was created with the aim to define vulnerability specific to the Malawian context, including exposure and hazard characteristic indicators. In collaboration with the lead researcher on the RCMRD project, the indicators were adapted to create the SoVI used as the baseline (Y-variable) within this study. Some critical points on the use of this index would be that it is created using a specific conceptualization of vulnerability. It could be stated that adapting the index will lead to unwanted results. However, the index could be adapted in a structured manner, as the elements of sensitivity and adaptive capacity fall within the conceptualisation of 'social vulnerability' in current DRR literature. The second critical remark that can be made for the index is that it is created in 2015, using data sources that go back to the year 2000. It could be questionable if the retention period of the data behind the indicators used in the RCMRD index are still within their limits. The aim of this research is however, to find a methodology that can potentially predict the vulnerability of a granular geo-spatial area within a country using remoteness indicators, the index should be suitable to make assumptions about the different remoteness indicators and if they can potentially predict social vulnerability. Furthermore, the available vulnerability data within Malawi on a granular level is very limited. Therefore, it was chosen to use the existing index to train the remoteness indicators to.

RQ 2: What data can be used to develop remoteness indicators as a proxy for social vulnerability?

Related to research question 2, the definition of remoteness is studied and the remoteness indicators created for this study are inspired by the indicators used in several existing remoteness indices. The remoteness indicators used in these indices are based on western countries as well as global south countries. For the choice of remoteness indicators for the purpose of this study a combination of the literature on remoteness, accessibility, and the possibilities within OSM, and other open data sets for Malawi were considered. For the final remoteness indicators, the travel times for communities to certain points of interest were created; the travel times to major hospitals, secondary schools, primary schools, water points, trading centres, and major cities/towns. These 'points of interest' are described in ***Section 5.2.2.*** The choices for these points of interest lay within the assumption that access to education (schools), health care (major hospitals), water (water points), commercial activities (trading centres), and government (major cities/towns) play an important role in the social vulnerability of communities. The road length and density is often mentioned within literature (Maru et al. 2014) as an indicator of government involvement or funding within an area. The settlement size could be described as a characteristic of locality; it could potentially be used as a distinction between urban and rural areas. The ruggedness, or topographic roughness of the terrain could potentially mean that access to a community is challenging. In this regard the physical accessibility, as well as cell phone, radio or even internet signal reception could be limited by the roughness of the terrain a community is located in. Many other remoteness indicators can be taken into account in the prediction of social vulnerability, but were not within the scope of this research.

RQ 3: What GIS-techniques and open-source tools are most relevant to extract remoteness indicators from OSM data?

In the creation of the remoteness indicators different open data sets are used. The basis for the GIS analyses related to the travel times are based on the road network created with the use of OSM data. The data is extracted from OSM using open-source command line tools. The extraction is done with several tools; Imposm, Osmosis, and OSM2PgRouting. Imposm is a tool to extract all the data for a certain geo-spatial extent and structures the data is different classes, see *Table 5.2*. Osmosis is a general OSM extraction tool which does not provide any pre-classification options. It can also function as a file format converter to feed into other OSM extraction tools or GIS software. OSM2PgRouting is specifically designed to extract all the road (line) attributes from OSM for a certain geo-spatial extent and create a routable network with a useful topology and nodes. These open-sourced tools are used for the analysis. During this study, no other command-line based tools were discovered that could handle the same amount of data with the functionalities described above. The scripts to extract the OSM data are command-line based, and need to run separately from the other scripts in this analysis. This can be difficult to maintain when other researchers want to replicate the extraction and analysis without any scripting knowledge.

All the open data sets that are used in this research are stored in an SQL database. To conduct any GIS analysis on this set in a scripted manner, a Postgis and PgRouting extension are added. The OSM road network showed some issues. The completeness of the minor roads varied greatly within Malawi, the core network with main roads was relatively complete, with some exceptions as indicated in *Figure 5.2.* To make the OSM-based road network suitable for routing, some fixing of the data was necessary. The functionalities with PgRouting are not sufficient in that regard, some gaps between the network were too large to accurately snap to the network without making the network unreliable. This indicates that the completeness of the OSM data for Malawi is somewhat limited. Further improvements of the OSM data is highly recommended to improve the roads for routing abilities and consequently the quality of the network analyses that may be conducted using the data. For the visualisation for the purpose of this report, QGIS is used. The visualisation for the purpose of decision-makers and transparency, the online dashboard is created using a Leaflet plugin

RQ 4: How does scripted database management, analysis, and visualisation contribute to the usability and reproducibility of the remoteness proxy?

As the OSM extraction tools are command-line based, the extraction can be scripted and any re-runs for updated OSM data can be done instantly. The OSM database is continuously updated, and new data is added constantly. To be able to use the latest data as an input, this needs to be repeated every so often. This subsequently enhances the replicability of the analysis as by using the script, the exact same parameters for data extraction will be used every time. However, some scripting knowledge is necessary at this time, as all the scripts within this analysis are not combined in a pipe-line or other similar tool. All the GIS analyses are conducted using SQL scripts, and can therefore be used with relative ease for other geo-spatial areas. Another benefit of GIS analysis using SQL scripts is that there is a complete insight in how the analysis is conducted, with what parameters, and what techniques are used. Within a guided user interface, like QGIS, it is often unclear to the user what techniques are used behind certain tools. Again, some SQL knowledge is necessary to edit the scripts to another area. The visualisations for this report are done within QGIS. In this regard, a template can be used to result in similar outputs. However, sharing of these templates between different operating systems showed some issues. The statistical analyses are conducted through an R-script in R. Again, the script gives insight into the methods and techniques used and can be re-run on a different set of indicators for a different geo-spatial extent. To run the script, some knowledge of R is necessary. For the final visualisation of the results, an online dashboard is created using java-script and html. These files can easily be tweaked and shared. The online dashboard can easily be published on Github. The application in another geo-spatial area can therefore, from data, through analyses, and visualisation be adapted with relative ease. For the applicability in the future, it is acknowledged that knowledge of command-line, SQL, R, Java-script, and HTML is necessary. This is a limitation, if a non-expert should be able to run the analyses. Follow-up development and research should be aimed at creating a pipe-line in which the separate scripts can be run simultaneously. This will increase the usability for non-experts, but limit the transparency related to the used methods and techniques for the technical experts.

.

RQ 5: How can the quality and completeness of open data for Malawi be validated to be sufficient for calculating a useful social vulnerability indicator?

Primarily, it is researched what open data sources specific to vulnerability and remoteness were available. The data sources were evaluated on the recentness, available meta-data, and on their retention period. On the open data platform HDX, all necessary meta-data is shared with the data sets. For MASDAP, this is the case but to a lesser extent. MadziAlipo does the sharing of their water point data sets with all the meta-data attached. The recentness of the water points, in the form of last visited, is given. For the OSM data, the road network was of importance for the analyses. A gap analysis was conducted whilst creating a routable network using PgRouting. Some portions of the road network contained large gaps, and snapping the roads to the closest nodes resulted in unwanted results. Furthermore, the 'Maxspeed' tag was used for the initial travel speeds for community members to the points of interest. After calibrating the travel speeds through the conducted field surveys using ODK and OMK, it was concluded that the 'maxspeed' tag overestimated the travel speed for community members to the points of interest with an average factor of nine. During the field-surveys the locations of the schools, hospitals, and nearest trading centre and town or city were validated. The location data on these points of interest were judged as accurate. A limitation is the geo-spatial extent used within Malawi to calibrate the remoteness indicators; more accurate results could be yielded through a validation in different Malawian districts. For the training of the remoteness indicators on the existing SoVI, the open-source software R is used. The final matrix with the SoVI and all the remoteness indicators were exported from the SQL database in a comma separated file format (CSV), which is a format often used in open data that can handle large amounts of data whilst staying compact. The structure of the file is sensitive, and should be respected, as with limited documentation it is difficult to reproduce the file. The matrix is read by R and checked for missing values before conducting the statistical analyses. The only missing values fell within the road density and road length values for 3 TAs. The locations of the roads and houses within OSM are evaluated using the latest openly available satellite imagery. These data are found to be accurate, the limitations for Malawi OSM data is the completeness. During the multivariate linear regression and the random forest regression models, the data matrix is split into a train data set of 60% of the TAs and a test data set of 40% of the TAs. This is done to assess if the models are not over-fitted to the data. The predictive accuracy is tested on the test data set to evaluate if this is the case. It subsequently acts as an internal validation of the data.

RQ 6: What statistical models can be used to define the relationships between the SoVI (Y) and the developed remoteness indicators (X) for Malawian Traditional Authorities?

For the prediction of SoVI using the remoteness indicators several supervised machine learning models could potentially be relevant. Many supervised learning methods are available; Decision Trees, Naïve Bayes Classifications, Neural Nets, Logistic Regression, Support Vector Machines, and Random Forest (Caruana & Niculescu-Mizil, 2006). Due to the timeframe and scope of this research a multivariate linear regression model and a random forest regression model are compared.

Both models LM, and RF are fitted to a train data set of 60% of the data within the matrix containing 220 of the 367 TAs. The LM uses 4 remoteness indicators as predictor variables for SoVI: Settlement Size, Number of Settlements, Mean Ruggedness, and Settlement Density. The Settlement Size and the Mean Ruggedness are the most important in the LM. The RF contains all the 12 predictor remoteness indicators. The Mean Ruggedness, Travel Time to Hospitals, Settlement Size, and Travel Time to Secondary Schools are most important in the RF. In both models, the Mean Ruggedness variable show to be among the most important variables. The RF can take all the remoteness indicators into account as it does not operate under the assumption that no multicollinearity exists among the predictor variables. The LM does operate under that assumption and thus can take less variables into account. With the scarce data availability, the LM thus provides some limitations. Within the remoteness indicators, around 50% of the variables show multicollinearity, especially the travel times to the points of interest. This can be expected when creating indicators with a common denominator: remoteness. The travel time indicators added to the analysis as a composite value did not yield a better result, and would subsequently result is loss of insight as to what travel time indicator would be of most relevance in the case when the data is not available for a certain point of interest in another geo-spatial extent.

It is difficult to make assumptions based on these models in other geo-spatial extents. Some of the remoteness indicators may be specific to the Malawi context. It is therefore necessary to test the models to other cases before implementing the model as such.

RQ 7: How accurate can the proxy, developed through the used models, predict SoVI?

To give an accurate answer to research question 7, the models were validated using the remaining 40% of the data set within the matrix. The data within the 40% falls outside of the sample where the models are fitted to. The LM predicts 53% of the SoVI using the test data set. It has an R^2 of 0.53 and a RMSE of 18.07. The RF predicts 64% of the SoVI using the test data set. It has a R^2 of 0.64 and a RMSE of 16.4. Both models show no indication of over-fitting to the train data set. The RF model 'flattens' the scores of the original SoVI somewhat; it underestimates the highest values and overestimates the lowest values. The LM flattens the mid-sections scores. This could indicate that some remoteness indicators are missing to predict these lowest and highest values for the RF model. The LM however, provides more insight into the relationships between the remoteness indicators and in what manner they predict the SoVI compared to the RF model. According to the coefficients in the LM, an increased settlement size results in the increase of social vulnerability. The increased number of settlements also results in the increase of social vulnerability. For both mean ruggedness and settlement density, the increased values result in the decrease in social vulnerability. The settlement density and the mean ruggedness are the most important indictors according to the t-statistic for the LM. The settlement density is higher in cities, which often show lower SoVI scores. For the Malawi context, less ruggedness leads to higher social vulnerability. This can be explained by the fact that lower, and thus flatter terrain for Malawi is more often struck by droughts and floods which leads to loss of crops and unsustainable livelihoods. This is a cause for higher social vulnerability. The most important remoteness indicators for the RF are the mean ruggedness, travel time to hospitals, settlement size, and travel time to secondary schools. The limitation for the RF model is that no insight into the relationship of correlation is given. The importance is determined using the percentage of increased mean square error (%incMSE). The importance of the travel time to hospitals can be explained by the access to healthcare. One could presume that further traveling times to hospitals will lead to increase in SoVI scores due to access to healthcare. For secondary schools, multiple reasons can be given; for Malawi it can indicate the access to secondary education, which is often limited. Furthermore, the secondary schools often function as a shelter during emergencies. It can be assumed that more distance to shelter locations can result in higher social vulnerability in the case of a natural hazard in Malawi. The mean ruggedness has most likely the same mechanism as the LM; but it could have a different relationship in other contexts. It could also be argued that a higher topographic roughness can lead to an increased social vulnerability in relation to accessibility, physical as well as radio, phone, and internet connectivity. For the case of Malawi, the flatter regions show a higher SoVI score. The settlement size is also an important predictor value in the RF. I can be expected that settlement sizes have a relationship with increased SoVI scores. For the LM, an increased settlement size means an increase in SoVI scores. This could be explained by the fact that areas with higher settlement densities (e.g. cities) have smaller settlement sizes. The settlement sizes in more rural areas are often larger in average. Rural areas are often more remote, and show higher SoVI scores. As stated before, the remoteness indicator set could be extended for future research in this regard. Extended research should focus on e.g. the difference between urban and rural TAs; it could be added to the statistical analysis as a dummy variable whereby the model could consequently improve. Within the boundaries of this research, an accurate data set to categorise the TAs within these classes was not found. Another aspect that could be beneficial for the predictions would be adding a predictor variable based on the number of settlements located on increased slope levels, as the mean ruggedness plays an important role as a stand-alone variable. The last issue that could be brought forward is the data set the indicators are trained on; it could be stated that the dataset is not an accurate representation of the SoVI, as the index is created using data that could potentially be outdated, and on choices made by the research team that developed the index for Malawi.

6.3 Future Research

This research is aimed at exploring the potential use of remoteness indicators for the accurate prediction of social vulnerability through the use of open data and open-source tooling. Due to the character of this research, being explorative, some follow-up research possibilities will be touched upon in this section.

Primarily, the research was aimed at the case for Malawian TAs. Different geo-spatial levels within Malawi could be used as an input to test the models on. A short exploration with a more granular boundary level for Malawi has been conducted, outside of the scope of this research; the Enumeration Area (EA) boundary. As this data set was not openly available at the time of this research it has been chosen not to include the findings. For future research this could potentially lead to even more granular results.

Furthermore, other countries (geo-spatial extents) could be explored to study how the models perform within a different context. As this study was performed on a relatively small country in a sub-Saharan African context, follow-up research could be on a larger country, or within a South-American or Asian context.

As stated earlier, the testing of different remoteness indicators was not exhausted. Different types of remoteness indicators could be assessed on their added value with regard to predictive accuracy of the LM and RF models. These could e.g. be the distinction between urban and rural areas, and the combination of settlements and the slope data. Furthermore, different machine learning techniques could be explored for their usability and accuracy within the Malawian context and beyond that.

As for the field calibration. Research into the use of enumeration tools could be expanded on. For OMK, it could be researched if there is a possibility to incorporate more possibilities in regard to OSM tagging; directly incorporate a water point, sanitation point, or road data to OSM through an OMK survey.

The completeness of the OSM data resulted in some challenges with regard to the network analysis and the accurateness of the travel times. Future research could be aimed at the possibility of different tags that would allow traveling speeds of e.g. pedestrians or public transport. Possible examples of further tagging within OSM would be ruggedness of the road, types of transportation, and traffic figures. For OSM data for Malawi, in general village names are missing, large segments of the road network, types of road, and locations of schools, water points, and hospitals.

As far as the scripting of the data collection, analyses, and visualisation goes, explorations could be made on creating a pipe-line to combine the separate scripts into one functioning script. This could increase the usability for non-technical (end-users). Furthermore, the visualisation of the dashboard could be validated by presenting the outputs to different information managers and decision-makers in the humanitarian sector. Any feedback would improve the visualisation and adhere better to their information needs.

Finally, a different vulnerability index may be used to fit the models to. If accurate and improved vulnerability data can be found for Malawi, the models can be tested on the data for their predictive accuracy. It would strengthen the validation of the models.

7 Conclusion

Main Research Question:

'How can 'remoteness' contribute to the identification of vulnerability to natural disasters for communities in Malawi, using open data and open-source tools?'

This chapter is aimed at giving an overview of the main conclusions that can be drawn from this research and in doing so, answering the main research question.

The main objective of this research was to identify if a measure of remoteness could identify whether Malawian communities are more or less vulnerable to natural hazards and where these communities are located. It is explored if this identification would be feasible through the use of open data and open-source tooling. Based on the conducted research and the discussion, this chapter attempts to answer the main research question.

For Malawi, the vulnerability data that is openly available was mainly provided through one source, the RCMRD vulnerability tool. As this data is openly available and has the rights to be shared and used for research purposes it is highly usable for the purpose of this study. It could be adapted to the current consensus on the concept of vulnerability to natural hazards in general: 'social vulnerability'. The data was deemed to be slightly out-dated. Mainly due to the lack of data in general, this data set was deemed most appropriate as the basis for the independent Y-variable in this research: SoVI.

Remoteness in Malawi could be identified through the use of open-source GIS analyses of travel times for communities to the most relevant points of interest within Malawi; major hospitals, schools, cities, trading centres, and water points. The locations of these points could be acquired through different open data platforms. The most important geo-spatial platform used for the analyses were OSM, HDX, MASDAP, CIESIN/Facebook, and MadziAlipo. The CEISIN/Facebook HRSL provided the basis for all the settlement location data within Malawi and provided a method to create an overview of the completeness of OSM settlement data for Malawi. OSM provided the basis for the road network within Malawi, the OSM road data is not complete and needed several fixing techniques to be able to perform the network analyses on. MASDAP provided the locations of most of the points of interest. HDX provided official boundary files for the different administrative boundaries within Malawi. The most relevant and up-to-date water point data was extracted from MadziAlipo, a Malawi-based geo-platform for water points. Different density analyses were conducted to add to the remoteness indicators. Finally, the openly available SRTM DEM for Malawi provided the basis for the mean ruggedness scores for all the TAs in Malawi.

The most relevant open-source tools for the GIS analyses for this research are QGIS, Postgis, and PgRounting. All the density and travel time analyses are scripted in SQL and stored in an SQL database. This resulted in fast, stable, and accurate outcomes providing a transparent overview of the methods and techniques for future use. The most relevant open-source tool for the statistical analyses for this research is R. It provides the possibility to script the analyses, contributing to the transparency and consequently replicability of the used methods and techniques. The visualisation of the results is subsequently done through an online dashboard using java-script and html scripting. The dashboard provides the results in an interactive manner, potentially improving the transparency of the results. It is recognized that due to the scripts being separate, the usability for a non-technical user will be more limited that for a more advanced, technical user.

The statistical methods in R consisted of the exploration of two machine learning models: the LM and the RF. Through the LM 53% of the SoVI scores for the TAs were predicted correctly, versus 64% for the RF. The RF relies on less assumptions than the LM, and is more flexible and timely in handling changing data sets than the LM. Overall, the RF is deemed to be the most useful model within the changing humanitarian data environment in predicting social vulnerability for Malawian communities using remoteness indicators. It does however, show less insight into the actual relationships between the predictor variables and dependent variable compared to the LM. For both models, the mean ruggedness came out as important predictor data sets. The ruggedness of the landscape can have several relations with a community being social vulnerable; accessibility by road, access to radio, phone, or internet access due to blockage of the landscape. For Malawi however, the areas with that are less rugged show higher SoVI scores; this can be due to the fact that these areas are more often struck by droughts and floods in Malawi which leads to unsustainable livelihoods, and thus higher SoVI scores. For the RF, the remoteness indicators travel time to hospitals (TTH) settlement size (SS), and travel time to secondary schools (TTSS) are the most important predictor indicators. For the hospitals this could be explained by the access to healthcare. The travel time to secondary schools could be explained by access to secondary education, but another aspect is that secondary schools often function as shelters in the aftermath of a natural hazard. The travel time to a shelter in the case of a natural hazard could affect the SoVI scores of communities. The settlement size has is positively correlated in the LM model, and is therefore assumed to have the same mechanism within the RF. Cities often show smaller settlement sizes, due to the increased density. In average, the bigger settlements are located in more rural, and remote areas. These areas often show higher SoVI scores.

The quality of the open data providers for Malawi are reasonable. The locations of the points of interest provided through MASDAP were reasonably accurate. OSM data caused the most challenges, with regard to completeness, and therefore accuracy of the possible network analyses conducted for Malawi using the road network. The HRSL dataset provided by CEISIN/Facebook showed to be accurate, the data was compiled in an overlay with OSM data and openly available satellite imagery, and showed overall similarities. The data was more complete than the OSM settlement data and the HRSL is therefore deemed as the most relevant open data set for the location of Malawian communities. The water point data provided by MadziAlipo is the best option with regard to openly available water point data as it contains meta data on visitation dates and state of the water points. As this data set is large and subject to fast-paced change, it is a data set that is difficult to validate.

To summarise, it is possible to accurately predict the vulnerability of Malawian communities to natural hazards using open data and open-source tooling for 64% of the TAs using a random forest regression model. There are some limitations in the RF, due to the limited insight into the relationship between predictor and dependent variables. A LM does show this insight, but is able to this on limited data due to model assumptions. There are limitations in the open data sets, as to completeness, retention period, and recentness. With the use of OSM, it is possible to create accurate remoteness indicators, however, a field calibration is necessary to comply to realistic values. The OSM data should be updated further, for its completeness and with improved tagging to enrich the data for humanitarian purposes and to further improve the analysis-possibilities of this data set.

8 References

510 (2016). A Priority Index For Humanitarian Aid After A Typhoon, Available at: https://510.global/philippines-typhoon-haima-priority-index/ [Accessed Nov, 2016]

ARIA (2013). Remoteness structure, Available at: http://www.abs.gov.au/ [Accessed Dec, 2016]

Assessment Capabilities Project (ACAPS). (2016). *Review of Information Needs*, Available at: https://www.acaps.org/library/assessment#resource-562 [Accessed Feb, 2017].

Adger, W. N. (2006). Vulnerability. *Global environmental change*, *16*(3), 268-281.

Antwi-Agyei, P., Fraser, E. D., Dougill, A. J., Stringer, L. C., & Simelton, E. (2012). Mapping the vulnerability of crop production to drought in Ghana using rainfall, yield and socioeconomic data. *Applied Geography*, *32*(2), 324-334.

Benini, A. (2015). The use of Data Envelopment Analysis to calculate priority scores in needs assessments. *Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP), London, UK*.

Bertolini, L., Le Clercq, F., & Kapoen, L. (2005). Sustainable accessibility: a conceptual framework to integrate transport and land use plan-making. Two test-applications in the Netherlands and a reflection on the way forward. *Transport policy*, *12*(3), 207-220.

Birkmann, J. (2006). Measuring vulnerability to promote disaster-resilient societies: Conceptual frameworks and definitions. *Measuring vulnerability to natural hazards: Towards disaster resilient societies*, *1*, 9-54.

Birkmann, J. (2007). Risk and vulnerability indicators at different scales: applicability, usefulness and policy implications. *Environmental Hazards*, *7*(1), 20-31.

Blaikie, P., Cannon, T., Davis, I., & Wisner, B. (2014). *At risk: natural hazards, people's vulnerability and disasters*. Routledge.

Bot, Jeroen. Personal communication. National Coordinator Shelter and Care, The Netherlands Red Cross. Interviewed by: Wilbrink, Jurg. (22 December 2016).

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231.

Cardona, O. D., van Aalst, M. K., Birkmann, J., Fordham, M., McGregor, G., & Mechler, R. (2012). Determinants of risk: exposure and vulnerability.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). ACM.

Cascetta, E., Cartenì, A., & Montanino, M. (2016). A behavioral model of accessibility based on the number of available opportunities. *Journal of Transport Geography*, *51*, 45-58.

Centre For International Earth Science Information Network (CEISIN) (2017) Available at: https://ciesin.columbia.edu/data/hrsl/ [Accessed Jan, 2017].

Maini, R., Clarke, L., Blanchard, K., & Murray, V. (2017). The Sendai Framework for Disaster Risk Reduction and Its Indicators—Where Does Health Fit in?.*International Journal of Disaster Risk Science*, 1-6.

Cutter, S. L., Barnes, L., Berry, M., Burton, C., Evans, E., Tate, E., & Webb, J. (2008). A place-based model for understanding community resilience to natural disasters. *Global environmental change*, *18*(4), 598-606.

Damm, M. (2010) Mapping Social-Ecological Vulnerability to Flooding. United Nations University, Institute for Environment and Human Security (UNU-EHS).

Eakin, H., & Luers, A. L. (2006). Assessing the vulnerability of social-environmental systems. *Annual Review of Environment and Resources*, *31*(1), 365.

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

Fekete, A. (2010). *Assessment of Social Vulnerability River Floods in Germany*. K. Brach (Ed.). United Nations University, Institute for Environment and Human Security (UNU-EHS).

Fekete, A. (2012). Spatial disaster vulnerability and risk assessments: challenges in their quality and acceptance. *Natural hazards*, *61*(3), 1161-1178.
Famine Early Warning System Network (FEWS) (n.d.). Available at: http://www.fews.net/fewspub.html [Accessed Dec, 2016]

Flanagan, B.E., Gregory, E.W., Hallisey, E.J., Heitgerd, J.L., and Lewis, B. (2011). A social vulnerability index for disaster management. Journal of Homeland Security and Emergency Management: Vol. 8: Iss. 1, Article 3.

Füssel, H. M. (2007). Vulnerability: a generally applicable conceptual framework for climate change research. *Global environmental change*, *17*(2), 155-167.

Gallopín, G. C. (2006). Linkages between vulnerability, resilience, and adaptive capacity. *Global environmental change*, *16*(3), 293-303.

Gbetibouo, G. A., Ringler, C., & Hassan, R. (2010, August). Vulnerability of the South African farming sector to climate change and variability: An indicator approach. In *Natural Resources Forum* (Vol. 34, No. 3, pp. 175-187). Blackwell Publishing Ltd.

De Groeve, T., Poljanšek, K., & Vernaccini, L. (2015). "Index For Risk Management – INFORM." Retrieved from INFORM website: http://www.inform-index.org/InDepth [Accessed Nov, 2016].

Healthsites (2017). About healthsites.io, Available at: https://healthsites.io/#about [Accessed Feb, 2017]

Heeger, Jan. Personal communication. Water and Sanitation Specialist, The Netherlands Red Cross. Interviewed by: Wilbrink, Jurg. (20 December 2016).

Heimstädt, M., Saunderson, F., & Heath, T. (2014). Conceptualizing Open Data ecosystems: A timeline analysis of Open Data development in the UK. In *CeDEM14: Conference for E-Democracy an Open Government* (p. 245). MV-Verlag.

Hofmokl, J. (2010) The Internet commons: towards an eclectic theoretical framework. *International Journal of the Commons*, 4(1), pp.226–250.

Humanitarian Data Exchange (HDX) (2017). About the Humanitarian Data Exchange, Available at: https://data.humdata.org/faq [Accessed Oct, 2017]

Humanitarian OpenStreetMap Team (HOT) (2017). About, Available at:
https://www.hotosm.org/about?pk_campaign=cpc-Branding-NoUsa [Accessed Jan, 2017]

International Federation of Red Cross and Red Crescent Societies (IFRC). (1996). *Vulnerability and capacity assessment toolbox.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (1999). *Vulnerability and capacity assessment: an international federation guide.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2004). *World disasters report 2001: focus on community resilience.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2006*a). What is VCA? A guide to vulnerability and capacity assessment.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2006*b). How to do a VCA: A practical step-by-step guide for Red Cross Red Crescent staff and volunteers.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2006*c). VCA toolbox and tool reference sheets.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2006*d). Vulnerability and capacity assessment: lessons learned.* IFRC, Geneva, Switzerland.

International Federation of Red Cross and Red Crescent Societies (IFRC). (2007)*. VCA toolbox and tool reference sheets.* IFRC, Geneva, Switzerland.

Jinon, Rene. Personal communication. Disaster Risk Reduction Delegate Malawi, The Netherlands Red Cross. Interviewed by: Wilbrink, Jurg (11 February 2016).

Johnson, D. P., Stanforth, A., Lulla, V., & Luber, G. (2012). Developing an applied extreme heat vulnerability index utilizing socioeconomic and environmental data. *Applied Geography*, *35*(1), 23-31.

Kasperson, J. X., & Kasperson, R. E. (2005). *The social contours of risk: publics, risk communication and the social amplification of risk* (Vol. 1). Earthscan.

Lewis, J. (1999). *Development in disaster-prone places: studies of vulnerability*. ITDG Publishing.

Li, Q., Zhang, T., Wang, H., & Zeng, Z. (2011). Dynamic accessibility mapping using floating car data: a network-constrained density estimation approach. *Journal of Transport Geography*, *19*(3), 379-393.

Luers, A. L., Lobell, D. B., Sklar, L. S., Addams, C. L., & Matson, P. A. (2003). A method for quantifying vulnerability, applied to the agricultural system of the Yaqui Valley, Mexico. *Global Environmental Change*, *13*(4), 255-267.

MadziAlipo (2016) Background, Available at: http://www.madzialipoapp.org/background [Accessed Dec, 2016]

Maru, Y. T., Smith, M. S., Sparrow, A., Pinho, P. F., & Dube, O. P. (2014). A linked vulnerability and resilience framework for adaptation pathways in remote disadvantaged communities. *Global Environmental Change*, *28*, 337-350.

McLaughlin, P., & Dietz, T. (2008). Structure, agency and environment: Toward an integrated perspective on vulnerability. *Global Environmental Change*, *18*(1), 99-111.

McCallum, Ian, Wei Liu, Linda See, Reinhard Mechler, Adriana Keating, Stefan Hochrainer-Stigler, Junko Mochizuki et al. (2016). "Technologies to Support Community Flood Disaster Risk Reduction." *International Journal of Disaster Risk Science* 7, no. 2: 198-204.

Miller, F., Osbahr, H., Boyd, E., Thomalla, F., Bharwani, S., Ziervogel, G., ... & Hinkel, J. (2010). Resilience and vulnerability: complementary or conflicting concepts?. *Ecology and Society*, *15*(3).

Missing Maps (2017). About, Available at: http://www.missingmaps.org/about/ [Accessed Dec, 2016]

Obe, R.O & Hsu L.S. (2017). PgRouting, a practical guide. Locate Press LLC

Odysseos, L. (2011). Governing dissent in the Central Kalahari game reserve: 'development', governmentality, and subjectification amongst Botswana's bushmen. *Globalizations*, *8*(4), 439-455.

OpenDataKit (ODK). (2016). About. Available at: https://opendatakit.org/about/ [Accessed Okt, 2016]

OpenMapKit (OMK). (2016). How it Works. Available at: http://www.openmapkit.org/index.html [Accessed Nov, 2016]

OpenStreetMap (OSM). (2014). Main. Available at: https://wiki.openstreetmap.org/wiki/Main_Page [Accessed Sep, 2016].

Pedraza-Martinez, A.J. (2013). On the Use of Information in Humanitarian Operations. In *Decision Aid Models for Disaster Management and Emergencies*. pp. 1–16. Available at: http://link.springer.com/10.2991/978-94-91216-74-9_6 [Accessed Nov, 2016].

Peduzzi, P., Dao, H., Herold, C., & Mouton, F. (2009). Assessing global exposure and vulnerability towards natural hazards: the Disaster Risk Index. *Natural Hazards and Earth System Sciences*, *9*(4), 1149-1159.

PgRouting (n.d.). Pgrouting documentation, Available at: http://pgrouting.org/documentation.html [Accessed Dec, 2016].

Polsky, C., Neff, R., & Yarnal, B. (2007). Building comparable global change vulnerability assessments: The vulnerability scoping diagram. *Global Environmental Change*, *17*(3), 472-485.

Regional Centre for Mapping Resources of Resources for Development (RCMRD). (2015). Malawi SRTM 30 Meters, available at: http://geoportal.rcmrd.org/layers/servir%3Amalawi_srtm30meters [accessed Nov, 2016]

Regional Centre for Mapping Resources of Resources for Development (RCMRD). (2015). Malawi Data Layers. Available at: http://servirportal.rcmrd.org/layers/?limit=100&offset=0&title__icontains=malawi&extent=-375.46875,-88.94504236931593,637.03125,82.1183836069127 [Accessed Nov, 2016].

Riley, S. J. (1999). Index that quantifies topographic heterogeneity. *intermountain Journal of sciences*, *5*(1-4), 23-27.

Rohrer, B. (2016). How to choose machine learning algorithms | Microsoft Azure. Available at: https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/ [Accessed Dec, 2016].

Sahay, B. S., Menon, N. V. C., & Gupta, S. (2016). Humanitarian logistics and disaster management: the role of different stakeholders. In *Managing Humanitarian Logistics* (pp. 3-21). Springer India.

Simelton, E., Fraser, E. D., Termansen, M., Forster, P. M., & Dougill, A. J. (2009). Typologies of crop-drought vulnerability: an empirical analysis of the socio-economic factors that influence the sensitivity and resilience to drought of three major food crops in China (1961–2001). *Environmental Science & Policy*, *12*(4), 438-452.

Tate, E. (2012). Social vulnerability indices: a comparative assessment using uncertainty and sensitivity analysis. *Natural Hazards*, *63*(2), 325-347.

Taylor, M. A., Sekhar, S. V., & D'Este, G. M. (2006). Application of accessibility based methods for vulnerability analysis of strategic road networks. *Networks and Spatial Economics*, *6*(3-4), 267-291.

United Nations Internarial Strategy for Disaster Reduction (UNISDR) (2002). Living with Risk: A Global Review of Disaster Reduction Initiatives, Geneva: UN Publications.

United Nations Internarial Strategy for Disaster Reduction (UNISDR) (2005). Hyogo framework for action 2005-2015: building the resilience of nations and communities to disasters, 2004 version, Geneva: UN Publications

United Nations Internarial Strategy for Disaster Reduction (UNISDR) (2009). UNISDR *Terminology on Disaster Risk Reduction.* Available at: https://www.unisdr.org/we/inform/terminology [Accessed Dec, 2016].

United Nations Internarial Strategy for Disaster Reduction (UNISDR) (2015). Sendai Framework for Disaster Risk Reduction 2015 – 2030. Available at: http://www.unisdr.org/files/43291_sendaiframeworkfordrren.pdf [Accessed Dec, 2016].

Wisner, B., & Gaillard, J. C. I. Kelman (2012) 'Framing Disaster: Theories and stories seeking to understand hazards, vulnerability and risk'. *The Routledge Handbook of Hazards and Disaster Risk Reduction*, 18-33.

Vetrò, A. et al. (2016) Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2), pp.325–337.

Appendix A - NLRC Terms of reference: field visit Malawi

Terms of reference – National Society review on data preparedness

Background

The Netherlands Red Cross has a long term partnership with Malawi Red Cross. In 2016 the NLRC has started a crowdfunding campaign for a project in Malawi to map vulnerable communities. Two graduates have started preliminary work on the project, to use data to identify vulnerable communities and to help the national society and other stakeholders to identify highly impacted communities after a flood.

The Netherlands Red Cross has also had the opportunity to apply for funding with the World Bank on a data project for the new Sustainable development goals (SDGs). The objective of this proposal is to create a national data collaborative in Malawi through which organizations can share data that are used to make humanitarian and development programs more efficient, and which can be used to monitor and report on the SDGs in Malawi.

This TOR has the objective to support the Netherlands Red Cross country representative to review the state of data preparedness in Malawi. It is a preliminary assessment that already provides the National Society with important insights on what data is available in Malawi, and it paves the way for more extensive work on this topic if the World Bank projects are selected for funding.

Additionally, through this TOR we want to take the opportunity to ground truth the work done by the researchers, by combining the mission with field validation exercises. For their vulnerability analysis to be accurate, the data quality and recentness of the data must satisfy certain standards. By participating in the data preparedness review they can make assumptions about the future usability of their models, and interact with Malawi Red Cross staff to learn how their models can be used in development of programs and humanitarian operations.

Purpose:

In order to assess needs, plan operations, implement activities and report on results during disaster responses, quality and timely data and information is essential.

Activities

The initial data preparedness review:

- A dialogue with decision makers in the national society to identify what kind of data they need to be faster and more effective during a disaster.
- An internal assessment of what data and systems are available in the National Society
- A dialogue with a few key partners to identify what data other organizations have, and which of these can be shared openly/in a trusted network. Build a trusted network for data sharing and support:
  - Visiting Ministry headquarters Lilongwe
  - Visit Ministry of Agriculture

  - Visit Department of Climate Change and Meteorological Services
  - Visit National Statistics Offices
  - Visit Regional Centre For Mapping Resource For Development
  - Visit Humanitarian Aid Organizations
  - Meetings with various district commissioners
  - Meetings with disaster preparedness officers (government)
- An introduction to movement wide data collection tools and methodologies (mobile data collection through RAMP, Missing Maps and OpenStreetMap, basic information management). Providing training to the National Society on these tools where requested.

The following activities are identified for a second stage data preparedness assessment:

- Develop a strategy for how to collect missing data, by the National Society, or through partners.
- Data collection of key data that is identified as missing using the best fit tools.
- Involving decision makers in how data-driven decisions can be made, and to increase the understanding of its value. Focus on establishing processes and preparing operations managers and decision makers to use data, maps, information in assessment, planning. Integrating the use of data in standard operating procedures where possible.
- Establish connection with existing initiatives that can support the NS in the future (i.e. SIMS, IFRC Go Network)

The additional activities of ground-truthing vulnerability models:

- Visit some of the identified most vulnerable communities to validate the model, by answering the following questions:
  - How can a remoteness index be used to find the most vulnerable people in Malawi?
  - How does travel time in reality compare to what an analysis of OpenStreetMap shows?
  - Validate the number of hospitals, sanitation points, houses, building materials.
- Discuss with Red Cross operations managers in the headquarters and in the field how they define vulnerability, how a vulnerability map can be used, and how priority index model/map be used after a disaster;
- How can a certain model or map be included into the work of the Malawi Red Cross?

Duration and timing

The duration of this project in Malawi would be 2-3 weeks and take place in January / February 2017.

Expenses

The team can cover their own expenses for travel and accommodation in Malawi, as well per diems for staff or volunteers travelling with them.

The Team

One senior team member will be selected to manage the project. The project manager will be working closely with the NLRC country delegation and Malawi Red Cross focal points.

The researchers are Thomas Plaatsman and Jurg Wilbrink. Both are graduating at the Red Cross Netherlands. Thomas Plaatsman is finalizing his masters in econometrics/big data analytics and is developing a priority index model for disaster response to flooding in Malawi.
Jurg Wilbrink is finalizing his masters in Geographic Information Management and Applications and is developing a remoteness index as a proxy indicator for vulnerability on the community level in Malawi. The data focus is on the use of open data (provided by governments, NGOs, OpenStreetMap).

Appendix B - Scripts

This appendix presents all the different scripts used in this research. The scripts consist of command-line based scripts, SQL scripts, R-script, and JavaScript.

1. OSM extract scripts

These scripts are command-line based.

```
#download latest osm data with wget using Geofabrik nightly dump.
wget http://download.geofabrik.de/africa/malawi-latest.osm.pbf

#imposm read pbf file
imposm --read downloads/malawi-latest.osm.pbf

#write databases in postgres
imposm --write --database malawi --host 192.168.220.254 --user postgres --port 5432

OSM data to PgRouting ready


#Osmosis steps

(after initiating postgres db)

--In postgres db:
create extension postgis;
create extension hstore;

--Add to postgres database
psql -d osm -f '/usr/local/Cellar/osmosis/0.44.1/libexec/script/pgsnapshot_schema_0.6.sql'
psql -d osm -f '/usr/local/Cellar/osmosis/0.44.1/libexec/script/pgsnapshot_schema_0.6_linestring.sql'


-- get osmosis to run in database:
osmosis --read-pbf malawi-latest.osm.pbf --log-progress --write-pgsql database=osm

-- use osmosis to create xml file:
osmosis --read-pbf malawi-latest.osm.pbf  --write-xml malawi.osm


#osm2PgRouting

./osm2PgRouting -file your-OSM-XML-File.osm -conf mapconfig.xml -dbname routing -user postgres -clean

--restart postgresql

brew services start postgresql
```

2. SQL scripts

These scripts are SQL-based

1. Pgr_DrivingDistance example (PgRouting function)

```
---driving distance creation 600 minutes
drop Table if exists edd_secschool600;
SELECT dd.*, n.the_geom As geom
INTO edd_secschool600
FROM pgr_drivingDistance(
'SELECT gid As id, source As target, target As source,
cost_s_secschool AS cost, reverse_cost_s_secschool AS reverse_cost
FROM  ways',
ARRAY(SELECT v.id
FROM secschools AS h
,LATERAL (SELECT id FROM  ways_vertices_pgr AS n
ORDER BY h.geom <-> n.the_geom LIMIT 1) AS v
)
, 600*60, false, equicost := false
) AS dd
INNER JOIN  ways As n ON dd.edge = n.gid;;
```

2. Variable distance buffer example

```
drop Table if exists var_bufferps500;
CREATE TABLE var_bufferps500 AS
SELECT  ST_transform(ST_Union(ST_Buffer(ST_Transform(geom, 32736), (((500*60) - agg_cost) * 0.9) )), 4326)
FROM edd_cities500
WHERE agg_cost <= 500*60;
```

3. Add travel time example

```
--for 80 minutes
ALTER TABLE var_bufferps80 ADD COLUMN travel_time double precision;
UPDATE var_bufferps80 SET travel_time = 80;
```

4. Create specific travel time only example

```
--create area20 minutes only (delete area20 from area40)
drop Table if exists psdis20;
CREATE Table psdis20 as

SELECT travel_time, ST_Difference(geom, (SELECT ST_Union(b.geom)
                              FROM var_bufferps10 b
                              WHERE ST_Intersects(b.geom, a.geom)
                               )) as geom
FROM var_bufferps20 a;
```

5. Select settlement points within specific travel time example

```
--select all cbuildings within given travel time (e.g. 20 minutes)
drop Table if exists cbuildings20;
create Table cbuildings20 as
select


  hd.travel_time,


  bu.geom
from
  fbpoints1 bu,
  cdis20 hd
where
    st_intersects ((bu.geom),hd.geom)
;
```

6. Union all settlement with travel time added example

```
drop Table if exists merged_wpbuildings;
CREATE TABLE merged_wpbuildings AS(
SELECT   travel_time,   geom FROM wpbuildings10
UNION
SELECT   travel_time,   geom FROM wpbuildings20
UNION
SELECT   travel_time,   geom FROM wpbuildings40
UNION
SELECT  travel_time,   geom FROM wpbuildings60
UNION
SELECT   travel_time,   geom FROM wpbuildings80
UNION
SELECT  travel_time,   geom FROM wpbuildings100
UNION
SELECT  travel_time,   geom FROM wpbuildings120
UNION
SELECT   travel_time,   geom FROM wpbuildings140
UNION
SELECT   travel_time,   geom FROM wpbuildings160
UNION
SELECT  travel_time,   geom FROM wpbuildings180
UNION
SELECT   travel_time,   geom FROM wpbuildings200
UNION
SELECT   travel_time,   geom FROM wpbuildings220
UNION
SELECT   travel_time,   geom FROM wpbuildings240
UNION
SELECT   travel_time,   geom FROM wpbuildings260
UNION
SELECT   travel_time,   geom FROM wpbuildings280
UNION
SELECT   travel_time,   geom FROM wpbuildings300
UNION
SELECT   travel_time,   geom FROM wpbuildings320
UNION
SELECT   travel_time,   geom FROM wpbuildings340
UNION
SELECT   travel_time,   geom FROM wpbuildings360
UNION
SELECT   travel_time,   geom FROM wpbuildings380
UNION
SELECT   travel_time,   geom FROM wpbuildings400
UNION
SELECT   travel_time,   geom FROM wpbuildings420
UNION
SELECT   travel_time,   geom FROM wpbuildings440
UNION
SELECT   travel_time,   geom FROM wpbuildings460
UNION
SELECT   travel_time,   geom FROM wpbuildings480
UNION
SELECT   travel_time,   geom FROM wpbuildings500
UNION
SELECT   travel_time,   geom FROM wpbuildings520
UNION
SELECT   travel_time,   geom FROM wpbuildings540
UNION
SELECT   travel_time,   geom FROM wpbuildings560
UNION
SELECT   travel_time,   geom FROM wpbuildings580
UNION
SELECT   travel_time,   geom FROM wpbuildings600
UNION
);
SELECT Populate_geometry_Columns('merged_wpbuildings'::regclass);
```

7. Spatial join to TA-boundaries example

```
---spataial join points with traveltimes to ta boundaries
--- using st_within
drop Table if exists ttwp_ta;
create Table ttwp_ta as

select b.geom, b.trad_auth, b.p_code, b.district, p.travel_time
 from merged_wpbuildings p, boundaries1 b where st_within (p.geom, b.geom);

--- using st_intersects
drop Table if exists ttss_ta;
create Table ttss_ta as
select
  p.travel_time, b.geom, b.trad_auth, b.p_code
from
  merged_ssbuildings p,
  boundaries b
where
    st_intersects ((b.geom),p.geom)
;

--Weighted average based on two columns
drop Table if exists tt_wp;
Create Table tt_wp as
SELECT  AVG(travel_time), trad_auth, p_code
FROM ttwp_ta b
GROUP BY p_code, trad_auth;

COPY tt_wp TO '/Users/jurgwilbrink/Desktop/tt_wp.csv' DELIMITER ',' CSV HEADER;
```

8. Density scripts

```
drop Table if exists settlement_density;
create Table settlement_density as

select b.geom, b.trad_auth, b.p_code,COUNT(p.geom) as nr_settlements, ST_AREA(ST_Transform(b.geom, 32736)) as
ta_area, COUNT(p.geom) / (ST_AREA(ST_Transform(b.geom, 32736)) /1000000) as settlment_per_km2
 from fbpoints p, boundaries1 b where st_within (p.geom, b.geom)
 Group BY b.geom, b.trad_auth, b.p_code;

drop Table if exists roadlength_ta;
create Table roadlength_ta as

select b.geom, b.trad_auth, b.p_code,SUM(p.length_m) as rd_length, ST_AREA(ST_Transform(b.geom, 32736)) as ta_area,
(SUM(p.length_m) /1000) / (ST_AREA(ST_Transform(b.geom, 32736)) /1000000) as km_per_km2
 from ways p, boundaries1 b where st_within (p.the_geom, b.geom)
 Group BY b.geom, b.trad_auth, b.p_code;
```

3. R-Script

```
---
title: "Code multivariate linear regression and random forest model for predicting social vulnerability in Malawi"
author: "Jurg Wilbrink"
date: "<14-06-2017>"
RStudio version Version 1.0.136 – © 2009-2016 RStudio, Inc.
---

 # install all required packages

install.packages("rmarkdown")

require('randomForest')
require('partykit')

library(caret)
library(readxl)
library(C50)
library(leaps)
library(psych)
library(randomForest)
library(leaps)
library(readxl)

 #----------------------------------------READ DATA MATRIX----------------------------------------
# Read data set

Matrix <- read_excel("~/Desktop/MATRIX.xls")"containing all the X-variables and Y-variable in one matrix"



 #----------------------------------------DATA EXPLORATION----------------------------------------
# Check for NA -- use minimum value

sum(is.na(Matrix$tt_cities))
Matrix$tt_cities[is.na(Matrix$tt_cities)]=min(Matrix$tt_cities[!is.na(Matrix$tt_cities)]);

sum(is.na(Matrix$tt_waterpoints))
Matrix$tt_waterpoints[is.na(Matrix$tt_waterpoints)]=min(Matrix$tt_waterpoints[!is.na(Matrix$tt_waterpoints)]);

sum(is.na(Matrix$tt_hospitals))
Matrix$tt_hospitals[is.na(Matrix$tt_hospitals)]=min(Matrix$tt_hospitals[!is.na(Matrix$tt_hospitals)]);

sum(is.na(Matrix$tt_secschools))
Matrix$tt_secschools[is.na(Matrix$tt_secschools)]=min(Matrix$tt_secschools[!is.na(Matrix$tt_secschools)]);

sum(is.na(Matrix$tt_primschools))
Matrix$tt_primschools[is.na(Matrix$tt_primschools)]=min(Matrix$tt_primschools[!is.na(Matrix$tt_primschools)]);

sum(is.na(Matrix$tt_tradcenters))
Matrix$tt_tradcenters[is.na(Matrix$tt_tradcenters)]=min(Matrix$tt_tradcenters[!is.na(Matrix$tt_tradcenters)]);

sum(is.na(Matrix$settlement_size))
Matrix$settlement_size[is.na(Matrix$settlement_size)]=min(Matrix$settlement_size[!is.na(Matrix$settlement_size)]);

sum(is.na(Matrix$settlem_km2))
Matrix$settlem_km2[is.na(Matrix$settlem_km2)]=min(Matrix$settlem_km2[!is.na(Matrix$settlem_km2)]);

sum(is.na(Matrix$nr_settlements))
Matrix$nr_settlements[is.na(Matrix$nr_settlements)]=min(Matrix$nr_settlements[!is.na(Matrix$nr_settlements)]);

sum(is.na(Matrix$settlement_size))
Matrix$settlement_size[is.na(Matrix$settlement_size)]=min(Matrix$settlement_size[!is.na(Matrix$settlement_size)]);

sum(is.na(Matrix$rd_length_km))
Matrix$rd_length_km[is.na(Matrix$rd_length_km)]=min(Matrix$rd_length_km[!is.na(Matrix$rd_length_km)]);
```

```
sum(is.na(Matrix$kmroad_km2))
Matrix$kmroad_km2[is.na(Matrix$kmroad_km2)]=min(Matrix$kmroad_km2[!is.na(Matrix$kmroad_km2)]);



# delete all na's that are left
Matrix1 <- na.omit(Matrix)

 # Create rescale function - data set from 0-100
rescale <- function(x) (x-min(x))/(max(x) - min(x)) * 100

# rescale all the indicators

Matrix2 <- Matrix1

Matrix2$sovi <- rescale(Matrix2$sovi)
Matrix22$mean_ruggedness   <- rescale(Matrix2$mean_ruggedness)
Matrix2$tt_hospitals   <- rescale(Matrix2$tt_hospitals)
Matrix2$settlement_size   <- rescale(Matrix2$settlement_size)
Matrix2$settlem_km2   <- rescale(Matrix2$settlem_km2)
Matrix2$tt_secschools   <- rescale(Matrix2$tt_secschools)
Matrix2$tt_cities   <- rescale(Matrix2$tt_cities)
Matrix2$rd_length_km   <- rescale(Matrix2$rd_length_km)
Matrix2$tt_tradcenters   <- rescale(Matrix2$tt_tradcenters)
Matrix2$nr_settlements   <- rescale(Matrix2$nr_settlements)
Matrix2$kmroad_km2   <- rescale(Matrix2$kmroad_km2)
Matrix2$tt_waterpoints   <- rescale(Matrix2$tt_waterpoints)
Matrix2$tt_primschools   <- rescale(Matrix2$tt_primschools)


 # Data exploration
hist(Matrix2$sovi)
hist(Matrix2$nr_settlements)
hist(Matrix2$settlem_km2)
hist(Matrix2$settlement_size)
hist(Matrix2$mean_ruggedness)
hist(Matrix2$rd_length_km)
hist(Matrix2$kmroad_km2)
hist(Matrix2$tt_cities)
hist(Matrix2$tt_hospitals)
hist(Matrix2$tt_secschools)
hist(Matrix2$tt_tradcenters)
hist(Matrix2$tt_primschools)
hist(Matrix2$tt_waterpoints)

hist(Matrix2$c_NS)
hist(Matrix2$c_SD)
hist(Matrix2$c_SS)
hist(Matrix2$c_MR)
hist(Matrix2$c_RL)
hist(Matrix2$c_RD)
hist(Matrix2$c_TTC)
hist(Matrix2$c_TTH)
hist(Matrix2$c_TTSS)
hist(Matrix2$c_TTTC)
hist(Matrix2$c_TTPS)
hist(Matrix2$c_TTWP)

#Boxplot
summary(boxplot(Matrix2$rd_length_km))
summary(boxplot(Matrix2$c_RL))
summary(boxplot((log(Matrix2$rd_length_km))))

summary(boxplot(Matrix2$c_TTH))
summary(boxplot(Matrix2$c_TTPS))

#check for skewedness
#log of x variables:
hist(log(Matrix2$sovi))
hist(log(Matrix2$nr_settlements))
hist(log(Matrix2$settlem_km2))
hist(log(Matrix2$settlement_size))
hist(log(Matrix2$mean_ruggedness))
hist(log(Matrix2$rd_length_km))
hist(Matrix2$rd_length_km)
hist(log(Matrix2$kmroad_km2))
hist(log(Matrix2$tt_cities))
hist(log(Matrix2$tt_hospitals))
hist(log(Matrix2$tt_secschools))
hist(log(Matrix2$tt_primschools))
hist(log(Matrix2$tt_tradcenters))
hist(log(Matrix2$tt_waterpoints))

hist((Matrix2$tt_waterpoints)^(1/3))
hist((Matrix2$tt_cities)^(1/3))
hist((Matrix2$tt_hospitals)^(1/3))
hist((Matrix2$tt_secschools)^(1/3))
hist((Matrix2$tt_primschools)^(1/3))
hist((Matrix2$tt_tradcenters)^(1/3))

hist((Matrix2$mean_ruggedness)^(1/3))
hist((Matrix2$rd_length_km)^(1/3))
hist((Matrix2$kmroad_km2)^(1/3))
hist((Matrix2$nr_settlements)^(1/3))
hist((Matrix2$settlem_km2)^(1/3))
hist((Matrix2$settlement_size)^(1/3))

hist(Matrix2$c_RL)
hist(Matrix2$c_)
# Cube root variables
Matrix2$c_RL = ((Matrix2$rd_length_km)^(1/3))
Matrix2$c_RD = ((Matrix2$kmroad_km2)^(1/3))
Matrix2$c_NS = ((Matrix2$nr_settlements)^(1/3))
Matrix2$c_SD = ((Matrix2$settlem_km2)^(1/3))
Matrix2$c_SS = ((Matrix2$settlement_size)^(1/3))
Matrix2$c_MR = ((Matrix2$mean_ruggedness)^(1/3))
Matrix2$c_TTC = ((Matrix2$tt_cities)^(1/3))
Matrix2$c_TTH = ((Matrix2$tt_hospitals)^(1/3))
Matrix2$c_TTSS = ((Matrix2$tt_secschools)^(1/3))
Matrix2$c_TTPS = ((Matrix2$tt_primschools)^(1/3))
Matrix2$c_TTTC = ((Matrix2$tt_tradcenters)^(1/3))
Matrix2$c_TTWP = ((Matrix2$tt_waterpoints)^(1/3))


 #---------------------------------------TRAIN AND TEST SET---------------------------------------
# split up test and training data (hdf data set)
set.seed(123)
train_ind=(runif(nrow(Matrix2))<=0.60)
train_sovi <- Matrix2[train_ind, ]
test_sovi <- Matrix2[!train_ind, ]



 # check for multicollinearity
pairs.panels(Matrix2[, c(15:27)],cor=TRUE,lm=TRUE,hist.col="cyan",method="pearson",scale=FALSE,pch = 20, cex = 1)
cor.plot(Matrix2[, c(15:27)],colors=TRUE,main="Correlation plot (matrix)")

pairs.panels(Matrix2[, c(15:27)],cor=TRUE,lm=TRUE,hist.col="cyan",method="pearson",scale=FALSE,pch = 20, cex = 1)

# On train data set
pairs.panels(train_sovi[, c(15:27)],cor=TRUE,lm=TRUE,hist.col="cyan",method="pearson",scale=FALSE,pch = 20, cex = 1)
cor.plot(train_sovi[, c(15:27)],colors=TRUE,main="Correlation plot (abs)")



 #---------------------------------------MULTIVARIATE LOGISTICAL REGRESSION---------------------------------------



# manual selection of X-variables where no covariance exists
```

```
Model1= lm(sovi ~ c_SS + c_NS  + c_RL + c_MR + c_TTWP
          + c_SD
         , data=train_sovi)

summary(Model1)

# automated selection using 'step' function
stepmodel1 <- step(Model1)
summary(stepmodel1)

plot(stepmodel1,scale="adjr2",main="y = sovi")

# train the linear regression model with the selected variables (by automated selection)
lm1=lm(sovi ~ c_SS+ c_SD +c_MR + c_NS, data=train_sovi)
summary(lm1)

# relative importance of the predictor variables
varImp(lm1, scale = FALSE)
varImp(stepmodel1, scale = FALSE)

# checking performance of the model on the training data set
plot(lm1,main="LM")


# assign the prediction to the matrix
train_sovi$pred_lm1=lm1$fitted.values

# calculate root mean squared error and rquared on training data
postResample(train_sovi$pred_lm1,train_sovi$sovi)

#plot measured to predicted
plot(train_sovi$pred_lm1,train_sovi$sovi,
    main="lm1 - Measured vs Predicted (training data set)",xlab="Predicted ",ylab="Measured ")

# run the LM model on the test data set
pred_lm1=predict(lm1,test_sovi)

# assign the prediction to the matrix
test_sovi$pred_lm1=pred_lm1



# calculate the RMSE and rsquared
postResample(test_sovi$pred_lm1,test_sovi$sovi)

# plot the residuals
test_sovi$res_lm1=((test_sovi$pred_lm1)-(test_sovi$sovi))


plot(test_sovi$sovi,test_sovi$res_lm1,main="lm1 - Residuals vs Fitted (test data set)",xlab="Predicted (sovi)",ylab="Residuals")

# plot measured to predicted
plot(test_sovi$pred_lm1,test_sovi$sovi,main="lm1 - Measured vs Predicted (test data set)",xlab="Predicted (sovi)",ylab="Measured (sovi)")

# apply the LM model on the complete study area
pred_lm1=predict(lm1,Matrix2)
Matrix2$pred_lm1=pred_lm1

plot(Matrix2$sovi, Matrix2$pred_lm1)



 #------------------------------------------------RANDOM FOREST--------------------------------------------


#rf1 model predicting sovi
RF=randomForest(sovi ~ c_SS + c_NS + c_MR  + c_TTTC + c_TTH + c_TTSS + c_TTPS + c_TTC
              + c_TTWP + c_RL+ c_RD + c_SD
              ,data=train_sovi,mtry=8,importance=TRUE,ntree=200)
print(RF)
plot(RF)
```

```
# RANDOM FOREST with normal variables
rf2=randomForest(sovi ~ nr_settlements + settlem_km2 + settlement_size
            + kmroad_km2 + rd_length_km + tt_cities + tt_hospitals +
             tt_secschools + tt_tradcenters + tt_primschools
            + tt_waterpoints + mean_ruggedness, data=train_sovi,mtry=8,importance=TRUE,ntree=200);

print(rf2)
plot(rf2)

# checking performance of the model on trainig data (only RF2)
summary(rf1)
summary(rf2)
#calculate RMSE and rsquared on training data
train_sovi$predRF=RF$predicted
postResample(train_sovi$predRF,train_sovi$sovi)


# check relative importance of predictors
importance(RF,type=1)
varImpPlot(RF,type=1,main="RF (y = sovi)")


# plot the residuals train
train_sovi$resRF=((train_sovi$predRF)-(train_sovi$sovi))
plot(train_sovi$sovi,train_sovi$resRF,main="RF1 - Residuals vs Fitted (train data set)",xlab="Predicted SoVI",ylab="Residuals")


# add predicted values and retransformation of them to Table (and plot against eachother)
train_sovi$predRF1=(train_sovi$predRF1)
plot(train_sovi$predRF1,train_sovi$sovi,main="Measured VS predicted (training data)")


# run the RF model on the test data set
predRF1=predict(rf1,test_sovi)

# assign the prediction to the matrix
test_sovi$predRF1=predRF1


# calculate the RMSE and rsquared
postResample(test_sovi$predRF1,test_sovi$sovi)
plot(test_sovi$predRF1,test_sovi$sovi, main="RF1 - Measured vs Predicted (test data set)",xlab="Predicted (sovi)",ylab="Measured sovi")

# plot the residuals
test_sovi$resRF1=((test_sovi$predRF1)-(test_sovi$sovi))
plot(test_sovi$sovi,test_sovi$resRF1,main="RF1 - Residuals vs Fitted (test data set)",xlab="Predicted SoVI",ylab="Residuals")

# plot measured to predicted
plot(test_cdh$predRF2_cr,test_cdh$cuberoot_cdh,main="RF2 - Measured vs Predicted (test data set)",xlab="Predicted
(CDH)^1/3",ylab="Measured (CDH)^1/3")

plot(test_sovi$predRF2,test_sovi$sovi,main="RF2 - Measured vs Predicted (test data set)",xlab="Predicted sovi",ylab="Measured sovi")


# apply the RF model on the complete study area (SA)
predRF1=predict(rf1,Matrix2)
Matrix2$predRF1=predRF1

# save dataframes with prediction to local drive
write.csv(Matrix2, file = "/Users/jurgwilbrink/Desktop/Matrix2.csv")


#END
```

4. ‘Java-script’ script online dashboard

```
function generateDashboard(data,geom){

  var map = new lg.map('#map').geojson(geom).nameAttr('TRAD_AUTH').joinAttr('P_CODE').zoom(6.6).center([-13,34]);

  var sovi = new lg.column("#sovi").label("Social Vulnerability Index").axisLabels(false);

  var SD = new lg.column("#settlem_km2").label("Settlement Density").axisLabels(true);

  var SS = new lg.column("#settlement_size").label("Settlement Size").axisLabels(false);

  var NS = new lg.column("#nr_settlements").label("Number of Settlements").axisLabels(false);

          var TTC = new lg.column("#tt_cities").label("Travel Time to Cities").axisLabels(false);

  var TTH = new lg.column("#tt_hospitals").label("Travel Time to Hospitals").axisLabels(false);

  var TTSS = new lg.column("#tt_secschools").label("Travel Time to Secondary Schools").axisLabels(false);

          //The below entry was missing the .axisLabels(false) part at the end.
          var TTTC = new lg.column("#tt_tradcenters").label("Travel Time to Trading Centres").axisLabels(false);

          var TTPS = new lg.column("#tt_primschools").label("Travel Time to Primary Schools").axisLabels(false);

  var TTWP = new lg.column("#tt_waterpoints").label("Travel Time to Waterpoints").axisLabels(false);

          //Below line mentioned "Povert level" instead of "Poverty level":
  var MR = new lg.column("#mean_ruggedness").label("Mean Ruggedness Index").axisLabels(false);

          var RL = new lg.column("#rd_length_km").label("Road Length").axisLabels(false);

  var RD = new lg.column("#kmroad_km2").label("Road Density").axisLabels(false)

  var PLM = new lg.column("#pred_lm1").label("Predicted LM").axisLabels(false);

  var PRF = new lg.column("#predRF1").label("Predicted RF").axisLabels(false);

          var DLM = new lg.column("#dif_lm2").label("Over and under prediction LM").axisLabels(false)
                    .colorAccessor(function(d){ if (d>0.15) {return 0;}  else if (d>=-0.15) {return 2;}  else if (d<-0.15) {return
4;}})
                    .colors(['#d7191c','#fdae61','#ffffbf','#DA70D6','#8B008B']);

          var DRF = new lg.column("#dif_rf2").label("Over and under prediction RF").axisLabels(false)
                    .colorAccessor(function(d){ if (d>0.15) {return 0;}  else if (d>=-0.15) {return 2;}  else if (d<-0.15) {return
4;}})
                    .colors(['#d7191c','#fdae61','#ffffbf','#DA70D6','#8B008B']);

  lg.colors(["#ffffb2","#fecc5c","#fd8d3c","#f03b20","#bd0026"]);

          var group1 = 1;
          var group2 = 12;
          var group3 = 0;
          var group4 = 0;

          var name1 = 'Y-variable';
          var name2 = 'X-variables';

  var grid1 = new lg.grid('#grid1')
    .data(data)
    .width($('#grid1').width())
    .height(5000)
```

```
                  .width(750)
    .nameAttr('#trad_auth')
    .joinAttr('#p_code')
    .hWhiteSpace(4)
    .vWhiteSpace(4)
                  //Adjust the below parameters to re-size the Table within the maximum assigned height and width of
the grid.
                  //The parameters for left and right are adjusted, so that the full names of the entries in the 1st column
do not overflow into the 2nd column.
                  //The text "Independent (X) variables" should appear correctly now.
    .margins({top: 250, right: 80, bottom: 30, left: 270})
     .columns([sovi,SD,NS,SS,RL,RD,MR,TTH,TTSS,TTPS,TTC, TTTC,TTWP])
                  ;



  $('#run1').on('click',function(){
  var group1 = 1;
         var group2 = 7;
         var group3 = 7;
         var group4 = 0;

    lg._gridRegister = [];
                  $('#run2').css({'background-color': 'grey' });
                  $('#run1').css({'background-color': 'green' }); //was: '#BF002D'
    $('#map-container').html('<div id="map"></div>');
    $('#grid1').html('');
    grid1 = new lg.grid('#grid1')
      .data(data)
      .width($('#grid1').width())
      .height(5000)
                           .width(750)
      .nameAttr('#trad_auth')
      .joinAttr('#p_code')
      .hWhiteSpace(4)
      .vWhiteSpace(4)
                           //Adjust the below parameters to re-size the Table within the maximum assigned height and
width of the grid.
                           //The parameters for left and right are adjusted, so that the full names of the entries in the 1st
column do not overflow into the 2nd column.
                           //The text "Independent (X) variables" should appear correctly now.
      .margins({top: 250, right: 120, bottom: 30, left: 270})
      .columns([sovi,PLM,PRF,DLM,DRF])
      ;


         $('#run2').on('click',function(){

    lg._gridRegister = [];
                  $('#run1').css({'background-color': 'grey' });
                  $('#run2').css({'background-color': 'green' }); //was: '#BF002D'
    $('#map-container').html('<div id="map"></div>');
    $('#grid1').html('');
    grid1 = new lg.grid('#grid1')
      .data(data)
      .width($('#grid1').width())
      .height(5000)
                           .width(750)
      .nameAttr('#trad_auth')
      .joinAttr('#p_code')
      .hWhiteSpace(4)
      .vWhiteSpace(4)
```

```
                           //Adjust the below parameters to re-size the Table within the maximum assigned height and
width of the grid.
                           //The parameters for left and right are adjusted, so that the full names of the entries in the 1st
column do not overflow into the 2nd column.
                           //The text "Independent (X) variables" should appear correctly now.
      .margins({top: 250, right: 80, bottom: 30, left: 270})
      .columns([sovi,SD,NS,SS,RL,RD,MR,TTH,TTSS,TTPS,TTC, TTTC,TTWP])
      ;

                  lg.init();
    initlayout(data,sovi,'#sovi');
    $("#map").width($("#map").width());
  });



                  lg.init();

    $("#map").width($("#map").width());
  });



         lg.init();
  initlayout(data,sovi,'#sovi');
  $("#map").width($("#map").width());

  function initlayout(data,sort_indicator1,sort_indicator2){

    //sort Table and color map by priority after loading dashboard
    var newdata = [];
    data.forEach(function(d){
      newdata.push({'key':d['#p_code'],'value':d[sort_indicator2]});
    });
    map.colorMap(newdata,sort_indicator1);
    grid1._update(data,grid1.columns(),sort_indicator1,'#trad_auth');



                  ////////////////////////////////////////
                  //Create the category lines above the grid
                  ////////////////////////////////////////

                  var g = d3.select('#grid1').select('svg').select('g').append('g');

                  //Add the number of variables per group
                  var offset_hor = 0;
                  var offset_vert = -30;

                  //horizontal line 1
                  g.append('line').attr("x1", 0+offset_hor)
                                                      .attr("y1", offset_vert)
                                                      .attr("x2",
(lg._gridRegister[0]._properties.boxWidth)*group1+(lg._gridRegister[0]._hWhiteSpace)*(group1-1)+offset_hor)
                                                      .attr("y2", offset_vert)
                                                      .attr("stroke-width", 1)
                                                      .attr("stroke", "black");

                  //horizontal line 2
                  g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*group1+offset_hor)
                                                      .attr("y1", offset_vert)
```

```
                                    .attr("x2",
(lg._gridRegister[0]._properties.boxWidth)*(group1+group2)+(lg._gridRegister[0]._hWhiteSpace)*(group1+group2-
1)+offset_hor)
                                    .attr("y2", offset_vert)
                                    .attr("stroke-width", 1)
                                    .attr("stroke", "black");

            //horizontal line 3
            g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2)+offset_hor)
                                    .attr("y1", offset_vert)
                                    .attr("x2",
(lg._gridRegister[0]._properties.boxWidth)*(group1+group2+group3)+(lg._gridRegister[0]._hWhiteSpace)*(group1+group2+
group3-1)+offset_hor)
                                    .attr("y2", offset_vert)
                                    .attr("stroke-width", 1)
                                    .attr("stroke", "black");

/*          //horizontal line 4
            g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3)+offset_hor)
                                    .attr("y1", offset_vert)
                                    .attr("x2",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3+group4))
                                    .attr("y2", offset_vert)
                                    .attr("stroke-width", 1)
                                    .attr("stroke", "black"); */

            //vertical line 1.1
            g.append('line').attr("x1", 0+offset_hor)
                                    .attr("y1", offset_vert)
                                    .attr("x2", 0+offset_hor)
                                    .attr("y2", (offset_vert-5))
                                    .attr("stroke-width", 1)
                                    .attr("stroke", "black");

            //vertical line 1.2
            g.append('line').attr("x1",
lg._gridRegister[0]._properties.boxWidth*(group1)+(lg._gridRegister[0]._hWhiteSpace)*(group1-1)+offset_hor)
                                    .attr("y1", offset_vert)
                                    .attr("x2",
lg._gridRegister[0]._properties.boxWidth*(group1)+(lg._gridRegister[0]._hWhiteSpace)*(group1-1)+offset_hor)
                                    .attr("y2", (offset_vert-5))
                                    .attr("stroke-width", 1)
                                    .attr("stroke", "black");

            //vertical line 2.1
            g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1)+offset_hor)
                                    .attr("y1", offset_vert)
                                    .attr("x2",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1)+offset_hor)
                                    .attr("y2", (offset_vert-5))
                                    .attr("stroke-width", 1)
                                    .attr("stroke", "black");

            //vertical line 2.2
            g.append('line').attr("x1",
lg._gridRegister[0]._properties.boxWidth*(group1+group2)+(lg._gridRegister[0]._hWhiteSpace)*(group1+group2-
1)+offset_hor)
                                    .attr("y1", offset_vert)
```

```
                                                        .attr("x2",
lg._gridRegister[0]._properties.boxWidth*(group1+group2)+(lg._gridRegister[0]._hWhiteSpace)*(group1+group2-
1)+offset_hor)
                                                        .attr("y2", (offset_vert-5))
                                                        .attr("stroke-width", 1)
                                                        .attr("stroke", "black");

                            //vertical line 3.1
                            g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2)+offset_hor)
                                                        .attr("y1", offset_vert)
                                                        .attr("x2",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2)+offset_hor)
                                                        .attr("y2", (offset_vert-5))
                                                        .attr("stroke-width", 1)
                                                        .attr("stroke", "black");

                            //vertical line 3.2
                            g.append('line').attr("x1",
lg._gridRegister[0]._properties.boxWidth*(group1+group2+group3)+(lg._gridRegister[0]._hWhiteSpace)*(group1+group2+gr
oup3-1)+offset_hor)
                                                        .attr("y1", offset_vert)
                                                        .attr("x2",
lg._gridRegister[0]._properties.boxWidth*(group1+group2+group3)+(lg._gridRegister[0]._hWhiteSpace)*(group1+group2+gr
oup3-1)+offset_hor)
                                                        .attr("y2", (offset_vert-5))
                                                        .attr("stroke-width", 1)
                                                        .attr("stroke", "black");

                            /* //vertical line 4.1
                            g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3)+offset_hor)
                                                        .attr("y1", offset_vert)
                                                        .attr("x2",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3)+offset_hor)
                                                        .attr("y2", (offset_vert-5))
                                                        .attr("stroke-width", 1)
                                                        .attr("stroke", "black");

                            //vertical line 4.2
                            g.append('line').attr("x1",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3+group4))
                                                        .attr("y1", offset_vert)
                                                        .attr("x2",
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3+group4))
                                                        .attr("y2", (offset_vert-5))
                                                        .attr("stroke-width", 1)
                                                        .attr("stroke", "black"); */

                            //horizontal text 1
                            g.append('text').attr('x', lg._gridRegister[0]._properties.boxWidth*(group1/2)+offset_hor)
                                                        .attr('y', (offset_vert+15))
                                                        .text(name1)
                                                        .style("text-anchor", "middle")
                                                        .attr("font-size",12);

                            //horizontal text 2
                            g.append('text').attr('x',
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2/2)+offset_hor)
                                                        .attr('y', (offset_vert+15))
                                                        .text(name2)
                                                        .style("text-anchor", "middle")
                                                        .attr("font-size",12);
```

```
                    /*//horizontal text 3
                    g.append('text').attr('x',
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3/2)+offset_hor)
                                                            .attr('y', (offset_vert+15))
                                                            .text(name3a)
                                                            .style("text-anchor", "middle")
                                                            .attr("font-size",12);

/*                  //horizontal text 4
                    g.append('text').attr('x',
(lg._gridRegister[0]._properties.boxWidth+lg._gridRegister[0]._hWhiteSpace)*(group1+group2+group3+group4/2)+offset_h
or)
                                                            .attr('y', (offset_vert+15))
                                                            .text('Demographics')
                                                            .style("text-anchor", "middle")
                                                            .attr("font-size",12);      */
          }
}

function hxlProxyToJSON(input,headers){
  var output = [];
  var keys=[]
  input.forEach(function(e,i){
    if(i==0){
      e.forEach(function(e2,i2){
         console.log(e2);
        var parts = e2.split('+');
        var key = parts[0]
        if(parts.length>1){
          var atts = parts.splice(1,parts.length);
          atts.sort();
          atts.forEach(function(att){
            key +='+'+att
          });
        }
        keys.push(key);
      });
    } else {
      var row = {};
      e.forEach(function(e2,i2){
        row[keys[i2]] = e2;
      });
      output.push(row);
    }
  });
  return output;
}

function stickydiv(){
  var window_top = $(window).scrollTop();
  var div_top = $('#sticky-anchor').offset().top;
  if (window_top > div_top){
    $('#map-container').addClass('sticky');
  }
  else{
    $('#map-container').removeClass('sticky');
  }
};

$(window).scroll(function(){
  stickydiv();
});
```

```
//load data

var dataCall = $.ajax({
    type: 'GET',
    url: 'data/data.json', //https://proxy.hxlstandard.org/data.json?merge-keys01=%23adm2%2Bcode&strip-
headers=on&filter01=merge&merge-
url01=https%3A//docs.google.com/spreadsheets/d/1klRixK82iRk1JnDOpAqKrry4VQiFcTGrfFZWr9ih-
Z8/pub%3Fgid%3D777123392%26single%3Dtrue%26output%3Dcsv&url=https%3A//docs.google.com/spreadsheets/d/1Olxh
Q_ejRKNvohbnfJ7yJPKD6U6pXcPPfsFnwBbP2nc/pub%3Fgid%3D0%26single%3Dtrue%26output%3Dcsv&filter02=select&sele
ct-query02-01=%23indicator%2Bcategory%21%3D1&merge-
tags01=%23affected%2Bdeaths%2C%23affected%2Bmissing%2C%23affected%2Bwounded%2C%23affected%2Binshelter%2
C%23affected%2Bbuildings%2Bdestroyed%2C%23affected%2Bbuildings%2Bpartially%2C%23affected%2Bschools',
    dataType: 'json',
});

//load geometry

var geomCall = $.ajax({
    type: 'GET',
    url: 'data/geom.json',
    dataType: 'json',
});

//when both ready construct dashboard

$.when(dataCall, geomCall).then(function(dataArgs,geomArgs){
    geom = topojson.feature(geomArgs[0],geomArgs[0].objects.geom);
                console.log(dataArgs);
    overview = hxlProxyToJSON(dataArgs[0],false);
    generateDashboard(overview,geom);
});
```

5. HTML script online dashboard

```
<html>
  <head>
    <title>Social vulnerability predicted through remoteness indicators for Malawi</title>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <link rel="stylesheet" href="css/bootstrap.min.css">
    <link rel="stylesheet" type="text/css" href="css/leaflet.css"/>
    <link rel="stylesheet" type="text/css" href="css/site.css"/>
    <link rel="stylesheet" type="text/css" href="css/leaflet-grid-graph.css"/>
    <script src="js/jquery.js"></script>
    <script src="js/d3.min.js"></script>
    <script src="js/crossfilter.v1.min.js"></script>
    <script src="js/leaflet.js"></script>
    <script src="js/topojson.v1.min.js"></script>
    <script src="js/leaflet-grid-graph.js"></script>
    <script src="js/papaparse.min.js"></script>
  </head>
  <body class="container">
    <div class="row">
      <div class="col-md-12">
                           <!--Removed the below line to show the 510 logo on a separate line. Display behaviour in IE 11 and Chrome is now the same. -->
                                    <!--<div style="display:flex">-->
                                            <img src="images/510.PNG" style="height: 80px; top: 50px">
                                            <div>
                                                     <!--Full title is now shown on one line: -->
                                                     <h1>Social vulnerability - Malawi - Remoteness as a proxy</h1>
                                            </div>
                                    <!-- </div> -->
        <p class="titledesc">
                                    This dashboard offers two overviews (1): the 'Social Vulnerability Index' (SoVI) (Y-variable) adapted from RCMRD vulnerability index (http://tools.rcmrd.org/vulnerabilitytool/) with the 'Remoteness Indicators' (X-variables) developed through scripted GIS analyses (2): The outputs of the multi-variate linear regression model (LM) and the random forest model (RF) used to predict the social vulnerability for Malawi using the remoteness indicators. The over- and under-prediction are indicated for both models.                       </br>
                                    Two options are available: click to select one: (The 1st is standard)
                                    </div>
                                    <!-- Colour settings of buttons have been changed to meet user-experience requirements. Green indicates: selected button. -->
                                    <button style="background-color: green; margin-right=5px;" class="btn btn-primary btn-sm" id="run2">(1) SoVI - remoteness indicators</button>
                                    <button style="background-color: grey" class="btn btn-primary btn-sm" id="run1">(2) SoVI - LM & RF predictions </button>
                                    <ul>
                                            <li><b>Option 1:</b> 'Social Vulnerability' (SoVI) and the developed remoteness indicators: Settlement Density, Number of Settlements, Settlement Size, Road Length, Road Density, Mean Ruggedness, Travel time to Major Hospitals, Secondary Schools, Primary Schools, Major Cities, Trading Centres, and Waterpoints.  </li>
                                            <li><b>Option 2:</b> The predictions of the multi-variate linear regression model (LM) and the Random Forest Model outputs are shown. Subsequently, the over- and under-prediction of the models are shown where red is over-estimation and purple is under-estimation, all the yellow areas (Admin level 3 Traditional Authorities) are reasonable predictions (+- 15%). </li>
                                    </ul>
                                    </div>
                           </div>
    </div>
    <div class="row">
      <div class="col-md-4">
                                    <button style="margin-bottom:10px;font-weight:bold;margin-right: 1px;background-color: #ffffb2;cursor:default; color:black" class="btn btn-primary btn-sm">Lower quadrant</button>
                                    <button style="margin-bottom:10px;font-weight:bold;margin-right: 1px;background-color: #fd8d3c;cursor:default; color:black" class="btn btn-primary btn-sm">Second quadrant</button>
                                    <button style="margin-bottom:10px;font-weight:bold;margin-right: 1px;background-color: #f03b20;cursor:default; color:black" class="btn btn-primary btn-sm">Third quadrant</button>
                                    <button style="margin-bottom:10px;font-weight:bold;margin-right: 1px;background-color: #bd0026;cursor:default" class="btn btn-primary btn-sm">Upper quadrant</button>
                                    <br>
        <div id="sticky-anchor"></div>
        <div id="map-container" style="margin-bottom:20px;"><div id="map"></div></div>
      </div>
      <div class="col-md-8">
                             <div id="grid1"></div>
      </div>
    </div>
    </div>
                  Dashboard made by Jurg Wilbrink: jwilbrink(at)RedCross(dot)nl
    <script src="js/site.js"></script>
  </body>
</html>
```

6. Commit to Github and publish online

Commit to github and publish online: https://jwilbrink.github.io/Remoteness-proxy/

Uploading to Github:

a) After preparing the data the dashboard is uploaded through a public Github account.
b) A new repository is created: ‘Remoteness-proxy’
c) The data is uploaded to Github through the Git command-line tool:
    i) Open Git Bash commandline window and go to dashboard folder (cd /c/github/folder/)
    ii) Git init
    iii) Git add -A
    iv) Git commit –m “first commit”
    v) Git remote add origin https://github.com/JWilbrink/Remoteness-proxy.git
    vi) Git push origin master

Appendix C - Field survey ODK/OMK

An ODK survey may be designed in a Microsoft Excel format to work on the tablets used in the field survey. For this particular study, an ODK format is chosen with the link to OMK implemented in the form; this will link the questionnaire answers to the building location using the OSM data as a base layer. The four tabs within the excel format are presented below. The first tab is displayed in *Table 1* and represents the tab where the survey questions are created. The survey questions were translated with the help of a MRCS volunteer in the local language: Chichewa. *Table 2* represents the second tab necessary for the ODK survey to function properly on the Tablets. It contains the possible answer choices that can be made within the survey. *Table 3* represents the third tab in the ODK form, it contains the possible choices in the OMK part of the survey.

Table 1: Tab 1 of ODK form in Excel. Representing the questions of the field survey

note	Consent	Hello, my name is……… And I am working under the Malawi Red Cross Society Integrated Disaster Reduction Project in Thyolo as a enumerator. We are carrying out an evaluation to learn about your practices regarding hospital visits and school use. The interview takes up to 5 minutes. You have been randomly selected to participate in this study. The information collected on your personal behaviour and life will be handled with strict confidence and the information will not be openly shared. Participation in this study is voluntary and we hope you will participate in the study since your views are important.
integer	deviceid	Device ID
select_one interviewer	interviewer	Name of Interviewer
select_one yes_no	hospital	Have you ever visited the district hospital of Thyolo or any other district hospital? --- Munayamba mwapita ku chipatala cha boma chachikulu?
select_multiple hospital	what_hospital	What district hospital have you visited? -- Munapita chipatala chanji? Kapena mukhoza kupita ku chipatala chanji mutadwala?
select_multiple travel	travelmethod_hospital	How do you travel to the hospital? -- Mumayenda bwanji kuti mukafike ku chipatala?
integer	traveltime_hospital	How many minutes does it take to reach the hospital? -- Mumatenga nthawi yaitali bwanji kuti mukafike ku Chipatalako?
select_multiple service	servicehospital	What types of services does the hospital supply? -- Ndichithandizo chanji chimene chimaperekedwa ku chipatalako?
select_one yes_no	treatment	Did you receive prescribed treatment? -- Mumapatsidwa chithandizo chimene mwalemberedwa ndi adotolo?
select_one yes_no	school	Do your children go to school? -- Kodi ana anu amapita ku sukulu?
select_multiple school	what_school	What school do you use? --Amaphunzira kuti?
select_multiple travel	travelmethod_school	How do your children go to school? -- Amapita bwanji popita ku sukulu?
integer	traveltime_school	How many minutes does it take the children to reach the school? -- Amatenga nthawi yaitali bwanji kuti akafike ku sukulu?
select_multiple school	what_school1	What school do you use?(2) (use this answer if children use different schools) -- Amaphunzira kuti?(2)
select_multiple travel	travelmethod_school1	How do your children go to school?(2) -- Amapita bwanji popita ku sukulu?
integer	traveltime_school1	How many minutes does it take the children to reach the school?(2) -- Amatenga nthawi yaitali bwanji kuti akafike ku sukulu?
osm building_tags	osm_building	Building

geopoint	geopoint_building	This will record the GPS location of the building
image	building_photo	Take a picture of the building from the street
end	form_completed	

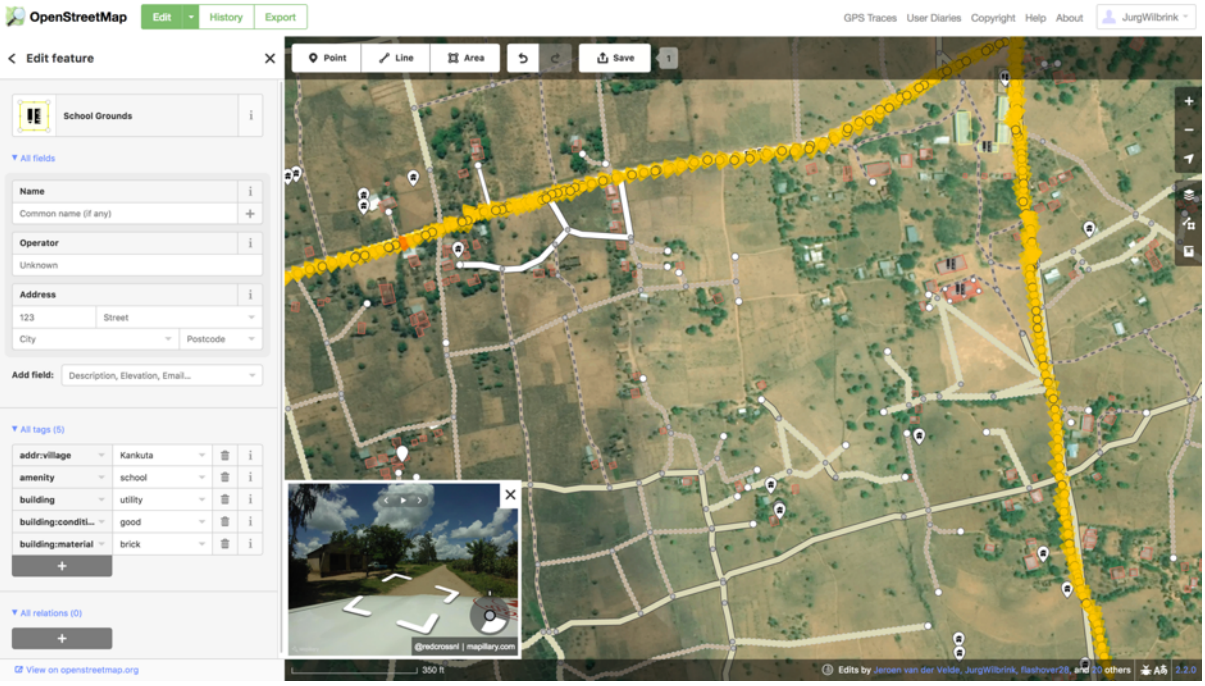
Table 2: Tab 2 of ODK form in Excel. Representing all the possible answer choices in the form. This is a specific example for Thunga Village.

hospital	stjoseph	St. Joseph Hospital
hospital	makwasa	Makwasa Hospital
hospital	thyolo	Thyolo District Hospital
hospital	bvumbwe	Bvumbwe Hospital
hospital	blantyre	Blantyre District Hospital
travel	car	Car
travel	motorbike	Motorbike
travel	bike	Bicycle
travel	publictransport	Public transport
travel	byfoot	On foot
travel	combination	Combination
service	Pharmacy	Pharmacy
service	operation	Operation
service	Malaria	Malaria
service	HIV	HIV/AIDS
service	blood	Blood transfusion
service	Maternity	Maternity
school	1	Mulambala School
school	2	Thunga School
school	3	Grengarry School
school	4	Chimkwende School
school	5	Mphedzu
school	6	Nkaombe School
school	7	Bvumbwe School
school	8	Namaona School
school	9	Kankhomba School
school	10	Chamasowa School
school	11	Chikapa School
school	12	Naming'omba School
school	13	Kalintulo School
school	14	Chimwavi Primary School
school	15	Mpeni School
school	16	Thunga CDSS
school	17	Naming'omba CDSS
school	18	Victory Christian Pvt

Table 3: Tab 3 of the ODK survey form in Excel. Containing the OMK choices of the field survey.

list name	name	label
building_tags	building	Building
building_tags	addr:village	Village
building_tags	amenity	Building Type
building_tags	building:material	Building Material
building_tags	building:condition	Building Condition
building	residential	Residential
building	commercial	Commercial
building	industraial	Industrial
building	utility	Utility
building	mixed	Mixed
building	hospital	Hospital
building	civic	Community Center
building	construction	Under Construction
amenity	school	School
amenity	church	Church
amenity	office	Office Building
amenity	fuel	Fuel Station
amenity	bank	Bank
amenity	restaurant	Restaurant
amenity	community_center	Community Center
amenity	clinic	Clinic
building:material	plaster	plaster
building:material	brick	brick
building:material	burnt_brick	burnt brick
building:material	tin	tin
building:material	cement_block	cement block
building:material	glass	glass
building:material	bamboo_sheet	bamboo sheet
building:material	wood	wood
building:condition	poor	poor
building:condition	average	average
building:condition	good	good

Appendix D - OSM contribution



School Grounds

Name: Common name (if any)

Operator: Unknown

Address: 123, Street, City, Postcode

Add field: Description, Elevation, Email...

All tags (5)

addr:village	Kankuta
amenity	school
building	utility
building:conditi...	good
building:material	brick

Edits by Jeroen van der Velde, JurgWilbrink, flashover28, and 20 others

Figure D1 Example of Katundu/Kankuta village output in OSM editor with Mapillary overlay.

Two Mapillary cameras were mounted to the front of the two cars during the field visits. The cameras are slightly pointed outwards and set to take a picture every 2 seconds or every 15 meters. The cameras record a GPS point for every photo taken. The results are filtered for privacy issues and uploaded to OSM. The yellow line and dots in *Figure D1* indicates the road photographed during the field visit. When a yellow dot is selected, the corresponding picture is displayed in the bottom left corner. The other items in the figure indicate the edits that were conducted during the visit of Katudu and Kankuta village. The school is selected in the upper east side of the map, the village name, amenity, building material, and building conditions are uploaded to OSM. To go through the street view imagery collected during the field survey visit: https://www.mapillary.com/app/user/redcrossnl [accessed March, 2017].