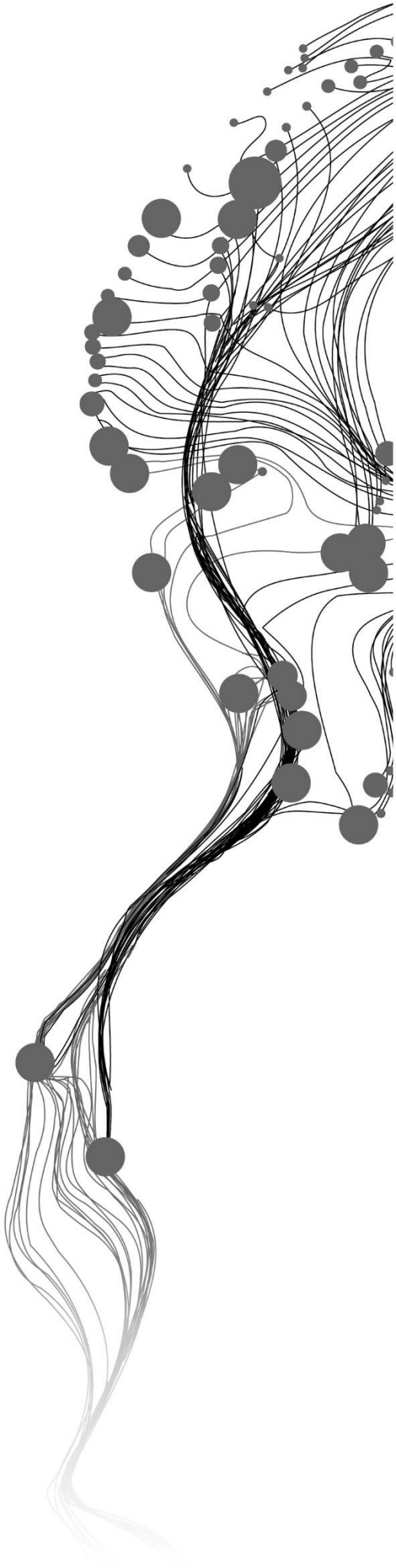
Combining Earth Observation and Social Media Data for Informal Settlements Identification in Dar es Salaam, Tanzania

WENFEI DONG
February, 2018

SUPERVISORS:
Dr. M. Kuffer
Dr. R. V. Sliuzas



Combining Earth Observation and Social Media Data for Informal Settlements Identification in Dar es Salaam, Tanzania

WENFEI DONG
Enschede, The Netherlands, February, 2018

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Urban Planning and Management

SUPERVISORS:
Dr. M. Kuffer
Dr. R. V. Sliuzas

THESIS ASSESSMENT BOARD:
Prof. Dr. K. Pfeffer (Chair)
Dr. H. Taubenböck (External Examiner, German Aerospace Center)

ABSTRACT

Identifying informal settlements in urban areas is important for policy making of informal settlement upgrading. Conventionally, informal settlements are detected by field surveys but this method is time-consuming and costly, especially in large areas. Remote Sensing (RS) has become a popular approach to map informal settlements in urban areas. However, RS-based approaches only can reflect morphological disparities between informal and formal residential areas. While, social media data as a novel data source is capable to reflect socio-economic disparities.

The objective of this study is to combine a Machine learning (ML) approach and social media data to identify informal settlements in Dar es Salaam (Tanzania). We employed the Support Vector Machine (SVM) algorithm to classify informal settlements and formal settlements from VHR (Pleiades-1A) images and HR (Sentinel-2A) images, and used twitter data to detect digital neighbourhoods (hot spots) and digital deserts (cold spots). The results showed VHR image obtained a higher accuracy (82.5%) than HR image (72.5%). Then, we compared the classification result of VHR image with cold and hot spots derived from twitter data. About 56.5% of cold spots matching with classified informal settlements and 19.7% of hot spots matching with classified informal settlements. It means twitter is not limited in informal settlements. Analysing the areas classified as informal settlements but recognized as digital neighbourhoods were explored whether this is due to misclassification. We found that only 28% of these areas are informal settlements, while 72% of these areas were misclassified as informal areas or mixed residential areas. Visual image interpretation and contextual information show, the informal settlements which were recognized as digital neighbourhoods are located in better-off informal areas with some infrastructures (like schools, markets, bars and health center).

In conclusion, only 2.4% of classified informal settlements are digital neighbourhoods, while 97.6% of classified informal settlements were regarded as digital deserts. Although the informal settlements were very diverse in Dar es Salaam, the combination of EO data and twitter data allows to analyse the socio-economic disparities in informal settlements.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to:

The ITC Directorate for giving me the opportunity to complete the MSc study through the UTS-ITC Excellent Scholarship programme;
Dr. Monika Kuffer and Dr. Richard Sliuzas, my supervisors, who gave their professional supervision, patience and guidance throughout my thesis study and also provided me the needed data. Dr. Karin Pfeffer, the Chair, for the advice and suggestions to improve the thesis during the proposal defence and mid-term presentation.
Dr. Hannes Taubenböck, for the permission to use the twitter data, which is very important to support my study.
European Space Agency (ESA) for facilitating the access to the Pleiades image.
Gina Leonita and Mustak Sheikh for the helpful discussion about Machine Learning and useful comments on thesis writing.

I also extend my thanks to my lover, Kuan Chai. I am lucky to meet him in ITC. Thanks for his continued encouragement and support during my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SSA	Sub-Saharan African
RS	Remote Sensing
EO	Earth Observation
VHR	Very High Resolution
HR	High Resolution
OBIA	Object Based Image Analysis
ML	Machine Learning
SVM	Support Vector Machine
RF	Random Forest
GSO	Generic Slum Ontology
RBF	Radial Basis Function
GLCM	Grey-Level Co-occurrence Matrix
NDVI	Normalized Difference Vegetation Index
FS	Feature Selection
DR	Dimensionality Reduction
SMFS	Supervised Multiview Feature Selection
LQ	Location-based Social Network Population Location Quotient
GPS	Global Positioning System
OSM	Open Street Map
HSIC	Hilbert-Schmidt Independence Criterion
SFS	Sequential Forward Selection
LISA	Local Indicators of Spatial Associati

1. INTRODUCTION

The existence of informal settlements directly reflects urban poverty in cities, which cannot be ignored considering the extent of such settlement in many Sub-Saharan African (SSA) cities. Substantial physical improvements are challenging but also urgently required. This research focus on the combination of remote sensing-based approach and social media data for informal settlements identification. This section introduces the background and significance of this research. After reviewing relevant literatures on combining remote sensing and social media data, the research problem is introduced. Next, based on the identified research gap, research objectives and corresponding questions are identified. Moreover, this chapter also includes hypotheses of this study and provides the conceptual framework.

1.1. Background and Justification

In past decades, the pressure of the rapid urbanization process has accelerated the growth of cities at worldwide scale. Due to the large number of immigrants aggregating in cities, informal settlements (also referred to as slums) are growing as migrants often do not find a better alternative place to reside. This is not only caused by the high urbanization pressure, but also by local land markets that cannot provide adequate land and affordable housing to low-income groups (Kuffer, Pfeffer, Sliuzas, & Baud, 2016). Informal settlements are undesirable residential areas which have been constructed illegally on land either administrated by the government or land managed privately (Ishtiyaq & Kumar, 2011). This phenomenon is particularly arrestive in developing countries, about 43% of the total population in developing countries live in informal settlements but only 6% of the people in developed countries (Ishtiyaq & Kumar, 2011). Especially in SSA, approximate 55% of the population live in informal settlements (United Nations, 2015). Asia and Pacific, the home of half of the urban population in the world, has 28% of the population living in informal settlements (UN-Habitat, 2016b). Therefore, cities in developing countries need to cope with development dynamics of informal settlements.

Informal settlements are often not an appropriate place for long-term living. People may face socioeconomic vulnerability like poverty, unemployment, underemployment because of the shortage of educational opportunities (Engstrom et al., 2015), the living environment is quite different compared with citizens living in formal areas. Thus as the most visible manifestation of urban poverty, informal settlements represent the socioeconomic divide in the city (Klotz, Wurm, Zhu, & Taubenböck, 2017). The increasing number of people aggregating in informal settlements will accelerate socioeconomic disparities. Strategies need to be developed to alleviate these socioeconomic disparities.

Slum upgrading is an appropriate tool to supply adequate housing and infrastructures for low-income citizens (UN-Habitat, 2016a). It reduces social inequalities, improves urban safety, improve access to basic services and triggers local economic development (UN-Habitat, 2015). To achieve slum upgrading, the basic task is to differentiate informal settlements from formal built-up areas. According to UN-Habitat (2003), informal settlements are characterized by lack of basic services, substandard housing, high density, hazard location, insecure tenure, poverty and small building size. These criteria help to detect informal settlements but require the use of census, socioeconomic or spatial data. However, these data are rarely up-to-date or complete. Even if researchers are willing to collect these data, it is an undoubtedly expensive and time consuming exercise because the need of a large number of people to be interviewed or performing a visual inspection of potential areas (Engstrom et al., 2015).

With the development of technology, Remote Sensing (RS) systems and image analysis techniques have been widely used in civil and commercial earth observation (EO). EO instruments contribute to the mapping, characterization, spatial feature extraction, and monitoring the evolution of large-scale urban landscapes (Taubenböck & Kraff, 2014). RS data is easy to capture and manifest the spatial and temporal information from an area. In principle, it allows for fast mapping of large areas at high frequencies (Duque, Patino, Ruiz, & Pardo-Pascual, 2015). Informal settlements are characterized by small building size, irregular and narrow road network, lack of vegetation and open space, and dense building construction (Graesser et al., 2012). Very-high resolution (VHR) images with resolution from 1 to 4 m (Murray, Lucieer, & Williams, 2010) allow the extraction of these important features and have been widely used in recent studies to identify informal settlements (e.g., Graesser et al., 2012; Hofmann, 2001; Gruebner et al., 2014; Naorem, Kuffer, Verplanke, & Kohli, 2016). Moreover, high resolution (HR) images with 5-10 m spatial resolution (Chehata, Orny, Boukir, Guyon, & Wigneron, 2014) also have been used in urban mapping (Pesaresi et al., 2013; Iannelli et al., 2014). For example, Wurm, Weigand, Schmitt, Geiß, & Taubenböck, (2017) utilized Sentinel-2A (10 m resolution) for informal settlements extraction due to such images are freely available.

Informal settlement identification can be implemented by various approaches. The most popular method is Object-Based Image Analysis (OBIA), followed by visual interpretation, texture or morphology analysis and machine learning (Duque, Patiño, & Betancourt, 2016). Machine Learning (ML) algorithm performs the highest accuracies and are capable to implement the computation with large indicator sets (Kuffer, Pfeffer, & Sliuzas, 2016). Furthermore, the combination of ML and textural, spectral and structural features has been applied in informal settlements extraction (Kuffer, Pfeffer, Sliuzas, et al., 2016). For example, (Duque et al., 2016) extracted spectral, textural and structural features from VHR images as input data to classify slums and non-slums in urban areas and evaluated the classification capacity of three machine learning algorithms (Logistic Regression, Support Vector Machine and Random Forest), and found Support Vector Machine (SVM) with radial basis kernel had the best performance. Wurm, Weigand, Schmitt, Geiß, & Taubenböck, (2017) employed textural and morphological image features mapping informal settlements by Random Forest (RF) classifier to assess pixel-based and patch-based accuracy. Engstrom et al. (2015) utilized spatial structural and contextual features for mapping informal settlements. All these literatures focus on the use of VHR image, image features and ML method to detect informal settlements.

Thus, RS is a popular approach to map informal settlements in urban area, nonetheless, RS-based informal settlements identification only can reveal the disparities of building types from the view of morphology (Wurm & Taubenböck, 2018). In recent years, a large number of spatio-temporal geographic footprints were produced by social network and information sharing applications (e.g., Facebook, Twitter, and Flickr). This novel data source can reflect citizens' social behaviour and can be related to aspects of the physical surrounding (Li, Goodchild, & Xu, 2013). Social media data capture the ambient geospatial information, allow to tag tweets with users' current location. Previous studies have shown the feasibility of combination between satellite data and social media data. For example, Cervone, Schnebele, Waters, Moccaldi, & Sicignano, (2017) utilized social media data collected from Twitter integrating them with remote sensing image for transportation assessment. Frias-Martinez, Soto, Hohwald, & Frias-Martinez (2012) focused on the use of geolocated tweets to characterize urban land use type. Hu et al. (2015) used geotagged photos to extract the urban regions of interest. However, few studies concentrate on the application of social media data for informal settlements mapping. Only Klotz et al., (2017) shown that social media data can characterize slums in Mumbai, India. No study has yet focused on the use of social media data for characterizing informal settlements in African cities. Thus, this research will study the

utility of social media data, to explore its potential for the identification of informal settlements in Dar es Salaam, Tanzania.

1.2. Research Problem

RS is an advanced tool to observe the earth and its environment by means of various sensors. Although the quantity of RS data available increased in the last years, gaps cannot be avoided because of particular limitations of instruments, carrier platforms and atmospheric interference (Cervone et al., 2017). VHR and HR satellite data provide adequate details for informal settlements identification. VHR images help researchers to identify buildings, vegetation, narrow roads and other objects (Veljanovski, Kanjir, Pehani, Oštir, & Kovačič, 2012). However, VHR images are costly data and the processing procedure is time consuming and complex. Especially in large-area applications of informal settlements mapping, one VHR image scene only covers a small part of the entire city. Thus several VHR images are needed to cover an entire city (Wurm et al., 2017). However, HR images are popular because it provides an increasing about of image details with the increasing in spatial resolution. Nonetheless, VHR and HR images allow only the extraction of features of informal settlements in form of their physical appearance (Kohli, Sliuzas, Kerle, & Stein, 2012), representing the morphological disparities of the physical environment, but lack access of socioeconomic data, which can reflect socioeconomic disparities effectively.

Lately, a novel data source - social media data can fill this gap. In this research, data collected from twitter can provide temporal and spatial information and possibly assist in informal settlements extraction. Twitter has a function that it can record the geographic position of the users at each time when the user creates a tweet (Frias-Martinez et al., 2012). These geolocated tweets can be regarded as a spatio-temporal footprints of the creators (Li et al., 2013). In this way, the area where people are more or less likely to use twitter can be shown in hot and cold spots, respectively. To some extent, it can reflect socioeconomic disparities in the city assuming that people living in informal settlements might use much less such services. A similar research has been done by Klotz et al., (2017), slums derived from remote sensing imagery had a strong accordance with the clusters assembled by cold spots, in Mumbai. This research will use both VHR and HR images to identify informal settlements by SVM classifier and aims to explore whether the same methodology is transferable to an African city - Dar es Salaam, for the mapping of informal settlements.

1.3. Research Objectives and Questions

1.3.1. Main Objective

The main objective of this research is to analyse the combination of SVM algorithm using VHR and HR images and social media data for mapping informal settlements in Dar es Salaam.

1.3.2. Specific Objectives

1. To select the most significant image features to classify informal settlements comparing VHR and HR images.
2. To apply SVM using VHR image and HR image to extract informal settlements.
3. To analyze the utility of Twitter data to depict the spatial distribution of digital deserts.
4. To analyze the relationship between informal settlements extracted from the best classification result and digital deserts (cold spots) derived by twitter data.

1.3.3. Specific Objectives and Questions

1. To select the most significant image feature to classify informal settlements comparing VHR and HR images.
   - What are the local characteristics of informal settlements using the conceptualization of the generic slum ontology (GSO, introduce the concept in section 2.2)?
   - What image features describe the local ontology for informal settlements mapping?
   - Which image features are the most significant for the local informal settlements identification?
2. To apply SVM using VHR image and HR image to extract informal settlements.
   - What spatial unit allows the best feature aggregation (e.g. regular grids, building blocks extracted from street network)?
   - What are the accuracies mapping informal settlements with both images (VHR and HR)?
3. To analyze the utility of twitter data to depict the spatial distribution of digital deserts.
   - How to aggregate geolocated tweets and disaggregate census data to spatial unit?
   - Where are the digital neighborhoods (hot spots) and digital deserts (cold spots) of twitter data within the city?
4. To analyze the relationship between informal settlements extracted from the best classification result and digital deserts (cold spots) derived by twitter data.
   - To what extent are the cold spots matching with informal settlements extracted from the best classification result?
   - Is the methodology used in Mumbai to map informal settlements by social media data transferable to Dar es Salaam?

1.4. Hypotheses

- Informal settlements in Dar es Salaam could be characterized by different image features in VHR image and HR image.
- The informal settlements identified from satellite images match with the digital desert detected by twitter data.

1.5. Conceptual Framework

The aim of this study is to identify informal settlements in Dar es Salaam based on an earth observation approach and social media data analysis. As mentioned in section 1.1, owing to satellite images only capture physical characteristics, remote sensing methods can measure morphological disparities between informal settlements and other urban areas. The limitation of this method is the lack of socioeconomic information which can reflect socioeconomic disparity in the city. However, as a new data source, social media data has the potential to fill this gap. As it is shown in figure 1.1, GSO can describe morphological characteristics of informal settlements based on three levels (object, settlement and environs level), image features extracted from satellite images can reflect the morphological characteristics of informal settlements. Machine learning as a fast mapping approach identifying informal settlements form images. Social media data is capable to reflect socio-economic disparities. Digital deserts could be derived from social media data and they are related to informal settlements because normally informal areas have less twitter users. Finally, the exploration of the relationship between informal settlements and digital desert will reveal the transferability of the methodology proposed by Klotz et al., (2017) to Dar es Salaam.

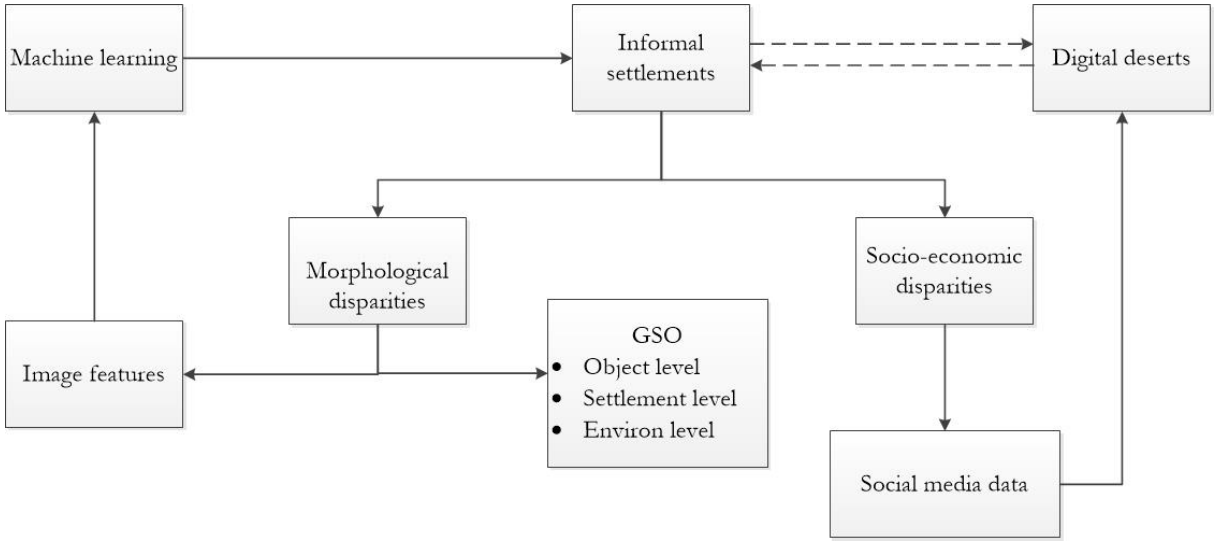


Figure 1.1 Conceptual framework.

2. LITERATURE REVIEW

2.1. Informal Settlements

There are diverse terms used in different parts of the world to describe the same urban phenomenon, i.e., informal settlements, slums, squatters, squatter settlements, favelas, poblaciones, shacks, barrios bajos and bidonvilles (UN-Habitat, 2015). Informal settlements are defined as residential areas where the dwellers lack tenure security of dwellings, lack access to basic services, insufficient living area, and the housing may not comply with building regulations(UN-Habitat, 2015). In principle, informal settlements are living places for all income groups including rich and poor, while slums are commonly described as settlements which are located near the hazardous urban land, aggregated by many dilapidated housing and characterized by poverty (UN-Habitat, 2015).

2.2. Informal Settlements Ontology

Before conducting image feature selection and classification, the most important step is the understanding of what characteristics make informal settlements differ from formal areas. Kohli et al., (2012) developed the generic slum ontology (GSO), providing a systematic conceptualization of indicators, which have the potential for slum identification. This framework contains three levels: object level, settlement level and slum environment, each level has two relative indicators. The relationship between each level and the associated indicators are shown in figure 2.1.

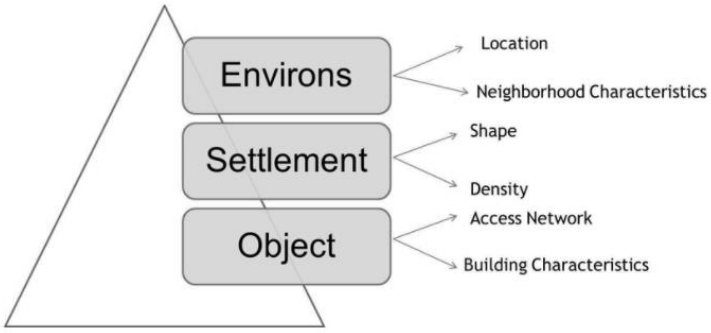


Figure 2.1 GSO in three spatial levels with relative indicators (Kohli et al., 2012).

Object level is the basis in this framework, it represents the components of the settlements, such as buildings and roads (Kohli et al., 2012). Building characteristics differentiate formal and informal areas based on roof type, footprint, shape as well as orientation in the satellite image (Kohli et al., 2012). Various roof materials of informal settlements can be found in different region. For instance, SSA mainly use tin or iron, Asian has mud, plastic or straw thatch etc.(Kohli et al., 2012). In informal settlements, footprints of dwellings are in general smaller than formal areas, shape and orientation do not comply with building regulations (Kohli et al., 2012). Moreover, informal settlements usually comprise irregular road networks, which contains different street types, surfaces and widths (Kohli et al., 2012).

Settlement level is a description of the overall form in a settlement (Kohli et al., 2012). Generally, informal settlements were constructed based on the shape of features such as roads, railways or drainage channels (Kohli et al., 2012). They may be located near to planned areas or main roads. Some experts used terms such as elongated or linear to describe the shape of informal settlements, others preferred irregular. Thus, these indicators can be used to seek potential locations of informal settlements (Kohli et al., 2012). In

addition, building density can also be an indicator at settlement level, due to high densities commonly found in informal settlements caused by the absence of open spaces and vegetation (Kohli et al., 2012).

Environ level includes location and neighbourhood characteristics as indicators (Kohli et al., 2012). Commonly, informal settlements are located at hazardous place, such as flooding areas, marshy areas, near railways or steep slopes (Kohli et al., 2012). Neighbourhoods which can offer unskilled or low-skilled job opportunities attracts the growth of informal settlements, therefore, can be used as potential indicator (Kohli et al., 2012).

Thus, the GSO provides a clear approach which combines all the characteristics used for informal settlements identification. As a good start in classification, it is a crucial step to understand the characteristics of informal settlements before using machine learning algorithm to classify images or using OBIA approach to detect objects from the image.

2.3. Support Vector Machine

Support Vector Machine is one of the most powerful machine learning algorithms in RS image classification (Mountrakis, Im, & Ogole, 2011). This method contributes to find an optimal solution to classify images using a relatively small sample dataset. SVM differs from other machine learning algorithms that it can transform a non-linear problem to a linear problem (Wang, Wan, Ye, & Lai, 2017).

As a supervised classification approach, SVM aims at seeking a hyperplane which can optimally split classes (e.g. two classes) (Kavzoglu & Colkesen, 2009). Training sets are used to optimize the hyperplane, and test sets are used to verify its generalization ability (Kavzoglu & Colkesen, 2009). As it shown in figure 2.2 (1), there are various hyperplanes which can separate two classes, but only one hyperplane is the best (figure 2.2 (2)) which keeps the maximum margin between two classes (Kavzoglu & Colkesen, 2009). The position of this best hyperplane would be equidistant, on the average between the border pixels of each class (Richards & Jia, 2006), therefore, the border pixels which are referred to as support vectors are the main focus in SVM.

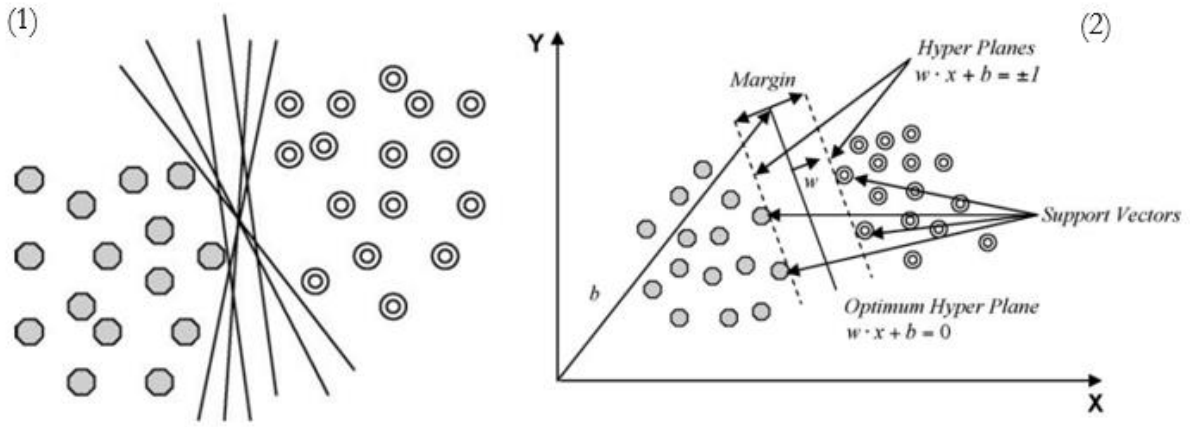


Figure 2.2 Various hyperplanes for linearly separable data (1), support vectors and the optimal hyperplane with maximum margin (2) (Kavzoglu & Colkesen, 2009).

The classical task of SVM is binary (two-class) classification. Assume we have a training set that contains samples from two different classes, the class in which each sample is assigned to belong is labelled y_i

$$y_i \in \{-1, 1\} \quad (1)$$

it represents that y_i is equal to 1 for one class, equal to -1 for another class (Zuo & Carranza, 2011). The hyperplane is defined as

$$w \cdot x_i + b = 0 \quad (2)$$

in formula 2, x is any point on the hyperplane, w represents the orientation of hyperplane in the space, and b is the bias represents the distance between the hyperplane and origin (Kavzoglu & Colkesen, 2009). If these two classes are linear separable, thus, there would be a d-dimensional hyperplane in feature space to split the training data into disparate classes (Zuo & Carranza, 2011). The hyperplane conforms to following equations:

$$w \cdot x_i + b \geq +1 \quad \text{for } y_i = +1 \quad (3)$$

$$w \cdot x_i + b \leq -1 \quad \text{for } y_i = -1 \quad (4)$$

formula 3 and 4 are equivalent to:

$$y_i(wx_i + b) \geq 1, \ i = 1,2, \dots, n \quad (5)$$

Support vectors are located on two hyperplanes which are parallel with the best hyperplane, defined by $w \cdot x_i + b = \pm 1$. The hyperplane which satisfies formula 5 means the classes are linearly separable. Hence, the margin between these two hyperplanes can be represented by $2/||w||$, the optimal hyperplane could be determined by minimizing $||w||2$, like formula 6:

$$\min[\tfrac{1}{2}||w||^2] \quad (6)$$

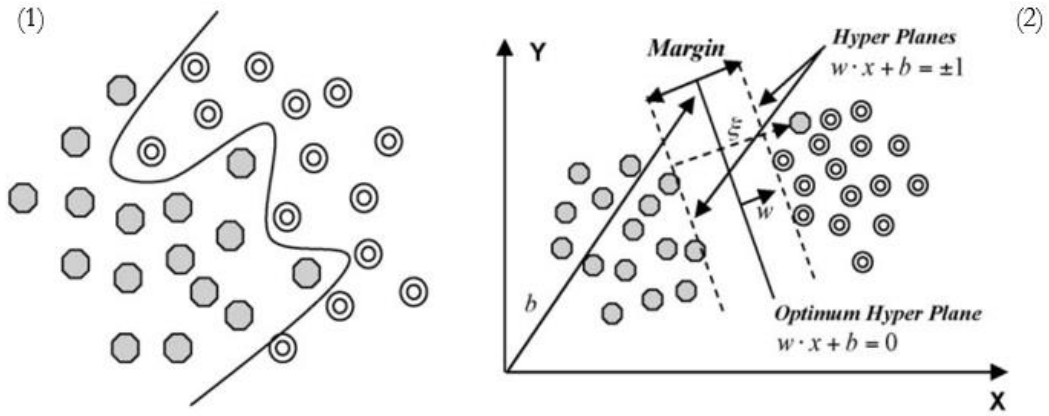


Figure 2.3 Nonlinear data set separation (1), slack variables in nonlinear data set separation (2) (Kavzoglu & Colkesen, 2009).

However, some classification cases are not linear separable (shown in figure 2.3 (1)), applying a linear classifier may misclassify some of the samples. Therefore, a nonlinear decision surface can be developed for such cases, the optimization problem changes to:

$$\min[\tfrac{1}{2}\left||w|\right|^2 + C\sum_{i=1}^{r}\xi_i] \quad (7)$$

Parameter C is the cost of penalty associated with margin maximization and error minimization and ξ is slack variables (shown in figure 2.3 (2)) that indicate the distance from the misclassified points to the best hyperplane (Kavzoglu & Colkesen, 2009).

Kernel methods are conducted to build a non-linear decision boundary in a high-dimensional feature space (Zuo & Carranza, 2011). SVM contains four types of kernels: Linear kernel, Polynomial kernel, Radial basis function (RBF) and Sigmoid kernel. Among all the kernels, RBF is commonly used in general cases for classification of remote sensing imagery since it has the capacity to classify multi-dimensional data, few parameters than polynomial kernel (Lin, Lee, Chen, & Tseng, 2008). Overall, RBF has a better performance than other kernels (Lin et al., 2008; Duque et al., 2016; Kavzoglu & Colkesen, 2009; Wang et al., 2017). According to Amami, Ayed, & Ellouze, (2015), RBF kernel is defined as

$$K_{RBF}(x_i, x_j) = \exp\left(-\sigma\left|\left|x_i - x_j\right|\right|^2\right), \sigma > 0 \quad (8)$$

In this formula, $||x_i - x_j||$ in the exponent could be expressed in different manners, such as Euclidean distance, least squares errors, inner product, etc (Chen, Ouyang, & Chang, 2012).There are two parameters σ and C in SVM with a RBF kernel. C is the cost of the penalty, the choice of C influences the classification outcome (Lin et al., 2008). Kernel width σ has much stronger effect than C in the performance of RBF kernel (Lin et al., 2008; Amami et al., 2015).

If the choice of C is too large, the classification accuracy will be relatively high in the training phase, but relatively low in the test phase (Lin et al., 2008). If the value of C is too small, it will get a undesirable classification accuracy, therefore, the model is useless (Lin et al., 2008). Another adjustable parameter σ plays a significant role in the performance of RBF (Amami et al., 2013). If σ is overrated, the kernel will performs linearly; if σ is underrated, the regularization will lost and the decision boundary will be sensitive to the outliers in training data (Amami et al., 2013).

2.4. Feature Extraction

Actually, there are a plenty of image features that can be extracted from an image, different studies analyse various image features for identifying informal settlements. The grey-level co-occurrence matrix (GLCM) is one of the most popular textural features in many studies (Pesaresi, Gerhardinger, & Kayitakire, 2008; Kuffer, Pfeffer, Sliuzas, et al., 2016). Wurm et al., (2017) used GLCM and differential morphological profiles to classify slums in Mumbai, and interpreted the results based pixel-based and patch-based accuracy. Graesser et al., (2012) and Engstrom et al., (2015) analyse the role of spatial, structural and contextual features (e.g. GLCM, Histogram of Oriented Gradients, Line Support Regions, Lacunarity, NDVI) to characterize informal settlements for four cities. Ella, Bergh, Wyk, & Wyk, (2008) compared textural feature algorithms for classifying urban settlements, finding GLCM has a good performance while Local Binary Pattern have a small advantage for classification. Taubenböck & Kraff, (2014) identified the physical features of informal settlements to analyse structural homogeneity and heterogeneities of informal settlements and detect informal settlements at three sites in Mumbai.

2.5. Feature Selection

As mentioned in section 2.3, the performance of SVM is impacted by the two parameters C and σ, in addition, with regard to the remote sensing imagery with high-dimensional features, the correlation between features also affects the classification accuracy (Lin et al., 2008, Wang et al., 2017). Feature selection (FS) allows to reduce the number of features, being similar to dimensionality reduction (DR) (Janecek, Gansterer, Demel, & Ecker, 2008). Feature selection is an approach to select a subset of relevant features (Fauvel, 2007). The aim of this process is to remove some inoperative or redundant features in the dataset to reduce the processing time as well as to improve the classification accuracy (Lin et al., 2008). While, dimensionality reduction uses algorithms and techniques to create new combinations of attributes to reduce the data set (Janecek et al., 2008).

In general, feature selection contains three different approaches: filters, wrappers, and embedded approaches (Janecek et al., 2008). Filters are classifier which independently rank features or feature subsets (Kotsiantis, 2011). They select features based on some predefined criteria, such as mutual information, independent component analysis, class separability measure or variable ranking (Kotsiantis, 2011). Wrappers are feedback methods which use classifier to assess the quality of a feature set (Janecek et al., 2008), training one learning machine based on each feature subset considered (Kotsiantis, 2011). However, wrapper approaches are time-consuming and conditioned to the type of classifier (Kotsiantis, 2011). Embedded approaches perform feature selection and elimination as part of the learning system (Sta & Jain, 2014), the advantage is that they make better use of the available data, avoid to separate training set to training and validation data (Kotsiantis, 2011).

To exclude highly correlated and redundant features from classification analysis, some studies employed feature selection strategies to choose the most important image features for classification (Chen, Li, & Gu, 2014; Pal & Foody, 2010; Kotsiantis, 2011). Chen et al., (2014) developed a supervised multi-view feature selection (SMFS), features are decomposed into different feature subsets, each feature subset represents a view and each view contains an image characteristic. The process of feature evaluation and selection are implemented in each view (Chen et al., 2014). Pal & Foody, (2010) applied four feature selection methods (SVM Recursive Feature Elimination (SVM-RFE), Correlation-Based Feature Selection (CFS), Minimum-Redundancy–Maximum-Relevance ((mRMR), Random Forest) to hyperspectral image classified by SVM and compared the resulting classification accuracies.

2.6. Social Media Data for Informal Settlement Identification

Social media as a novel data source has been recently combined with remote sensing in many fields such as land use mapping (Liu et al., 2017), transportation damage assessment (Hwang, Evans, & Hanke, 2017), disaster mapping (Fohringer, Dransch, Kreibich, & Schröter, 2015) etc. Social media data plays an important role in filling the gap in remote sensing, because it is possible to reflect socioeconomic characteristics, which cannot be captured by remote sensing techniques.

As for the application of social media and RS-based approach for informal settlements identification, Klotz et al., (2017) presented a methodology for slum and non-slum classification in Mumbai, India. Quantifying the intensity of geolocated tweets on building blocks through the city to extract the clustered neighbourhoods which are more or less digital oriented (Klotz et al., 2017).

Formula 9 introduced by Anselin & Williams, (2016) presents a relative location-based social network population location quotient (LQ), which can be used to highlight the clusters of geolocated tweets related to the population in the entire study area:

$$\mathrm{LQ} = \frac{\mathrm{M_{block}}/\mathrm{M_{total}}}{\mathrm{P_{block}}/\mathrm{P_{total}}} = \frac{\mathrm{MQ}}{\mathrm{PQ}} \quad (9)$$

MQ presents the relative geolocated tweets and PQ presents the relative population per block, respectively. To aggregate geolocated tweets and number of population to the level of building blocks, Klotz et al., (2017) first delineated the building blocks from HR image based on the morphological homogeneity and closely surrounded road network. Then they use spatial autocorrelation to explore the clusters concentrated by blocks, which have more similar values on LQ at the city scale. After acquiring the digital hot spots (more Twitter users) and digital deserts (less Twitter users), they compared them with the slums identified by image analysis (spatial perspective). The results of this study show that geolocated tweets are not only limited in slums but also not prevail in other residential neighbourhoods. However, 44.7% of the slum areas detected from satellite image are digital deserts, 51.5% of slums did not show significant neighbourhood characteristics. While, only 3.8% of slums were recognized as digital neighbourhoods. Thus, in this study, social media data can be a referential factor for slums, but it is unable to substitute morphological slum identification method. However, the combination of a RS-based approach and social media data has a high analytical potential and should be explored in further research (Klotz et al., 2017).

2.7. Morphological Informal Settlements Mapping Combines with Socio-economic Data

Except for social media data, a lot of studies combined remote sensing data and social economic data to analyse the relationship between them (Wurm & Taubenböck, 2018; Sandborn & Engstrom, 2016; Taubenbock et al., 2009). Wurm & Taubenböck, (2018) performed an assessment to detect the ability of VHR image to identity slums in the city of Rio de Janeiro by using morphological characteristics. In this study, the authors visually delineated morphological slums from VHR images (1m and higher) based on

the physical characteristics like building density, size, height, arrangement, location, the road network as well as construction materials comparing with the surrounding neighbourhoods (Wurm & Taubenböck, 2018). The remote sensing-based outcome is morphological slums which is used to compare with census slums derived from social-economic data. Because in Rio de Janeiro, census data provide the information about the household income in each census block (Wurm & Taubenböck, 2018). Therefore, census slums could be identified in terms of the poverty line (Wurm & Taubenböck, 2018). Official data provided by the Brazilian Census were used as spatial reference to assess the ability of morphological slum mapping, the result shown 93.7% morphological slums agree with true slum (Wurm & Taubenböck, 2018). Therefore, this high accuracy could acknowledge that morphological mapping is suitable for slum identification. While only 45% of morphological slum blocks are coincide with by census slums. As a result, the authors revealed that morphologically identified slums could be related to urban poverty, however, because of the wider physical features of slums, a method based solely remote-sensing has difficulties to distinguish slums and formal areas with similar structure.

3. STUDY AREA AND DATA DESCRIPTION

3.1. Study Area

The location of study area is urban zone of Dar es Salaam. The study area approximately occupies 316 km^2, covering all the planned residential areas and dense informal settlements around the city centre. Dar es Salaam is the largest city in Tanzania, located in the eastern part in Tanzania bordering the Indian Ocean (figure 3.1). The entire administrative boundary of the city covers 1630 km^2 and had 5,465,400 inhabitants in 2016. The city has approximate 100 informal settlements and almost 75% of total residential housing is located in informal areas (Rasmussen, 2013). Figure 3.2 shows an overview of the informal settlements in Manzese, Dar es Salaam

According to UN-Habitat (2010), the characteristics of informal settlements in Tanzania can be concluded as (1) tenure security: people who live in informal settlements have some 'perceived' tenure security because the government supports informal settlements upgrading and will pay full and fair compensation to people whose ownership is abolished or interfered. (2) The quality of housing: due to the households in informal settlements have perceived tenure security, they prefer to use permanent and modern materials for construction (UN-Habitat, 2010). This case occurs in both rural and urban areas, particularly, the use of modern materials has been the highest in urban areas in Dar es Salaam (UN-Habitat, 2010). (3) Residents living in informal settlements: All income groups can be found in the same informal settlement (UN-Habitat, 2010). Due to informal land market is easier to get access to land for housing and residents are encouraged to use permanent building materials on the land in informal markets which is not controlled by local authorities (UN-Habitat, 2010).

In Dar es Salaam, middle and low income families live in informal settlements next to each other (Rasmussen, 2013). Some residents have full-time jobs such as the teacher in university or staff in government, while others work in informal sectors like street vendors (Rasmussen, 2013). In general, poor infrastructures, unreliable services and hard maintenance highlight the disparity between formal settlements and informal settlements.

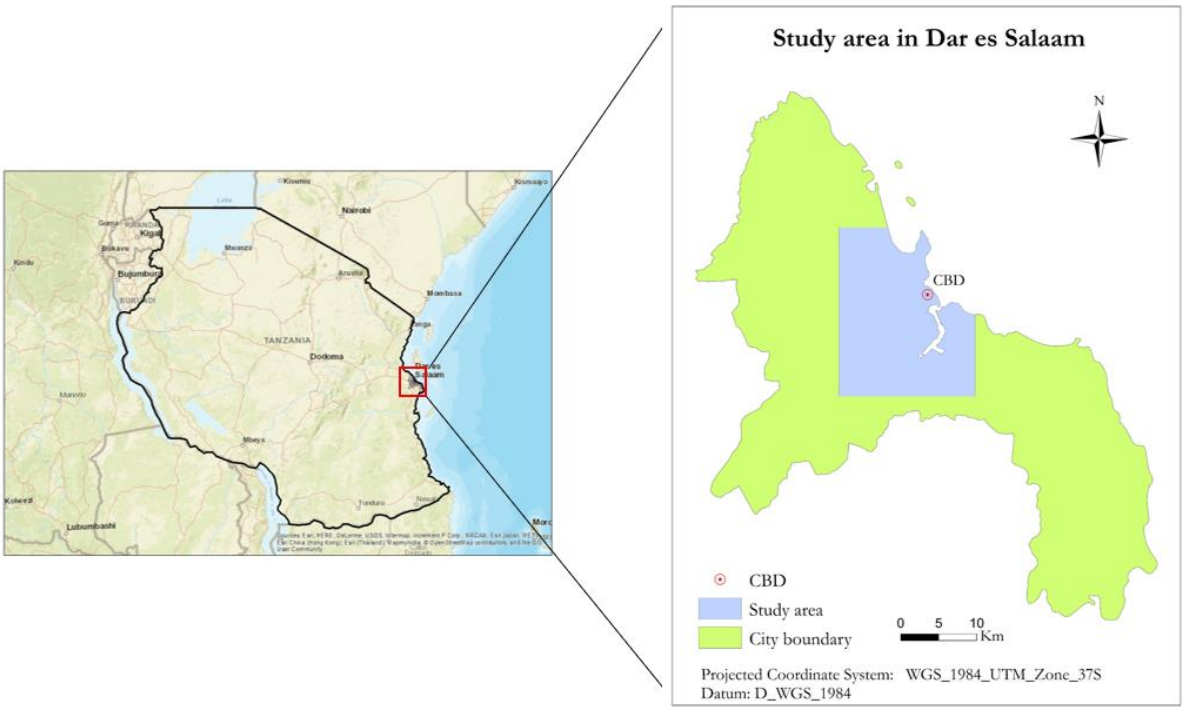


Figure 3.1 Study area.

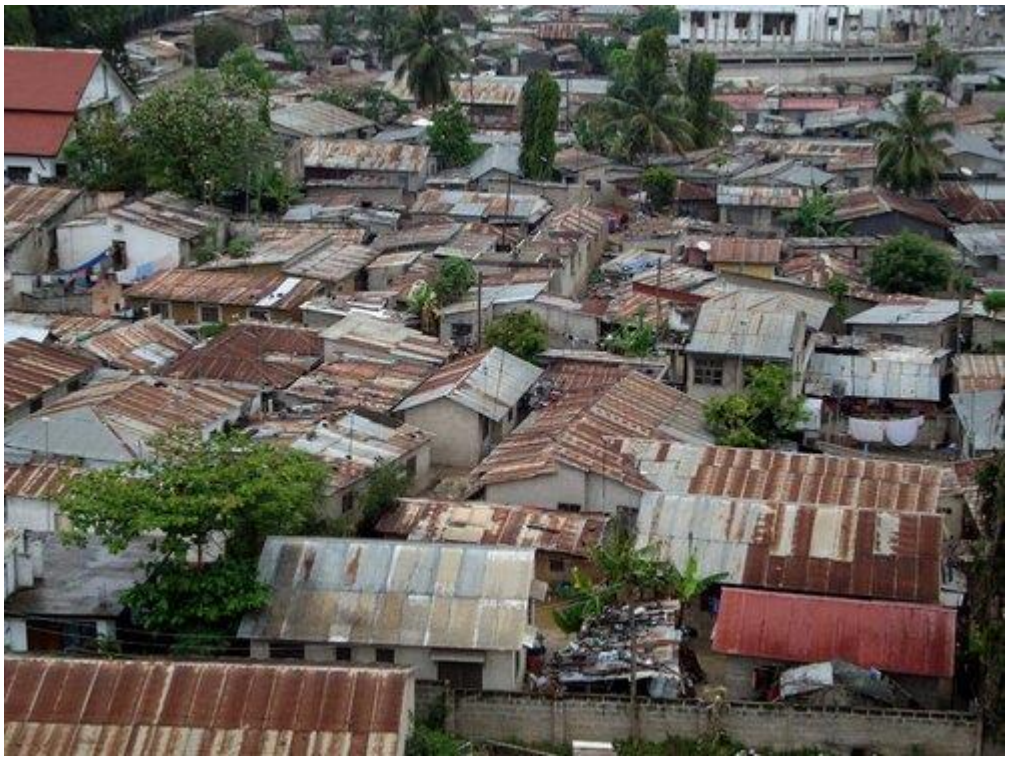


Figure 3.2 Informal settlements in Manzese, Dar es Salaam.

(Source: http://msongo.blogspot.nl/2013/04/dar-es-salaam-most-dangerous-areas.html)

3.2. Data Description

The data needed in this study are secondary data, include two satellite images (Pleiades-1A image and Sentinel-2A image), land use, building footprints, city boundary, subwards boundary, census data as well as geolocated tweets. The description of data is presented in Table 3.1.

3.2.1. Satellite Images

The study involves two satellite images with different spatial resolution. The high-resolution image is a Sentinel-2A[1] which was launched in 2015. This satellite contributes to provide a full and systematic coverage of the land surface at global scale. It contains 13 spectral bands at a spatial resolution of up to 10 m (the resolution of R, G, B and NIR bands is 10 m, 13 bands information shows in appendix 1). The frequency of revisit time is 5 days and swath width is 290 km. Therefore, Sentinel-2A has a good potential covering a large area and capturing moderate image details at the same time (Wurm et al., 2017).

Another image is captured from Pleiades[2], which has very high resolution (2 m) and delivers extremely high level of detail. The images obtained from Pleiades have four spectral bands (R, G, B, NIR), this satellite is capable of providing orthorectified colour data and offering a daily revisit to any location on the earth. Pleiades image has four spectral bands Therefore, such VHR images are effective for urban mapping.

3.2.2. Twitter Data

Geolocated tweets are extracted from twitter's public application programming interface (API) as another main dataset in this research. Twitter is a popular social networking site used for information sharing and conversation. Users can share their daily life through micro-blogging with a maximum of 280 characters[3] (up from 140 characters in Nov, 2017) per message. Additionally, twitter offers function that it can record the geographical location of users at each time when a tweet is created (Frias-Martinez et al., 2012). These locations are formed in two manners with quite different precision level (Li et al., 2013). One option is the users can set their location which is automatically captured by a built-in Global Positioning System (GPS) receiver in the mobile equipment, recording the user's position in the form of longitude and altitude. Another option is users can select the location from a set of place names offered by Twitter. In this case, Twitter records the estimated location of the mobile device or the Internet Protocol (IP) address of the computer, providing several possible locations (e.g. a building, a neighbourhood, a city, even a country) based on geocodes for selection (Li et al., 2013). However, these two manners provide different location accuracy. Location recorded by GPS has a relatively precise accuracy at magnitude of several meters (Li et al., 2013). While the users select the location provided by Twitter, the accuracy just in general, ranges from 30 to 3000 meters (Li et al., 2013). The location generated by the IP address is also not precise, the positional accuracy depends on the method which is used to convert the IP address to coordinates (Li et al., 2013). This study uses the information in tweets dataset include user ID, created time, longitude and altitude. Other private information such as user description, user name, the number of followers, the number of friends, and text are excluded.

[1] Sentinel-2A data is obtained from European Space Agency (ESA). Image is captured on July 10, 2016. Sentinel-2A data is free to download from http://sentinel-pds.s3-website.eu-central-1.amazonaws.com/

[2] Pleiades-1A data was provided by ESA. Image is acquired on May 7, 2016. The available image can be found on http://www.intelligence-airbusds.com/en/4871-browse-and-order

[3] https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/

3.2.3. Other Secondary Data

In addition, the study also needs land use data, city boundary, subwards boundary, building footprints as well as census data. Land use data is used to select training and test sets for classification. Building footprints[4] download from Open Street Map (OSM) are needed to extract residential buildings, combining with census data in subwards to estimate population number in spatial unit.

Table 3.1 Data description.

Dataset	Source	Year	Format	Description
Sentinel-2A	ESA	2016	Raster	Contains 13 bands (use R, G, B and NIR (10 m resolution) in this study)
Pleiades-1A	ESA	2016	Raster	2 m resolution, 4 bands (R, G, B and NIR)
Land use	ITC	2010	Vector	Include Agriculture, Commercial, Education, Forest, Industrial, Informal settlements, Recreational, Residential, Special Purpose, Transportation, Water Body
Geolocated tweets	Twitter API	2015	Vector	Shown in points
Building footprints	OSM	2017	Vector	Describe the distribution of buildings
City boundary	OSM	2017	Vector	Boundary of Dar es Salaam
Subwards boundary	ITC	2012	Vector	Show subwards in Dar es Salaam
Census data	ITC	2002 & 2012	Vector & Document	Demographic data in each subward

[4] OSM data were downloaded on 24-09-2017 from http://download.geofabrik.de/africa/tanzania-latest-free.shp.zip

4. METHODOLOGY

4.1. General Approach

Overall, this research applied an exploratory design to analyse the relationship between informal settlements extracted from VHR image and the distribution of social media data. The aim is to determine whether the use of social media data is assisting in mapping informal settlements in Dar es Salaam. The analysis of social media data is still new in studying informal settlements, and its potential needs to be tested. Therefore, conducting an exploratory design can provide a broader view about the combination of a ML algorithm (SVM) using VHR image and HR image, and social media data for informal settlements identification.

This research is based on a quantitative approach, achieved by EO technique and location-based social activities analysis. Firstly, image features will be selected based on the physical characteristics of informal settlements in Dar es Salaam and previous studies, which have been done in image features selection. Then the most important features will be selected by a feature selection algorithm. This process allows to capture the most significant geospatial information and reduce redundant features before classification. After that, the SVM classifier is applied for informal settlements identification.

To explore the potential of social media data for informal settlements identification, this research employs geolocated tweets to analyse their distribution. This process generates a map with hot and cold, showing locations with more (digital hot spots) or less tweets (digital deserts). Thus, comparing the relationship between informal settlements extracted from satellite images and digital deserts. The results show whether social media data allow mapping of informal settlements. The general methodology is shown in figure 4.1, details are provided in the following subsections.

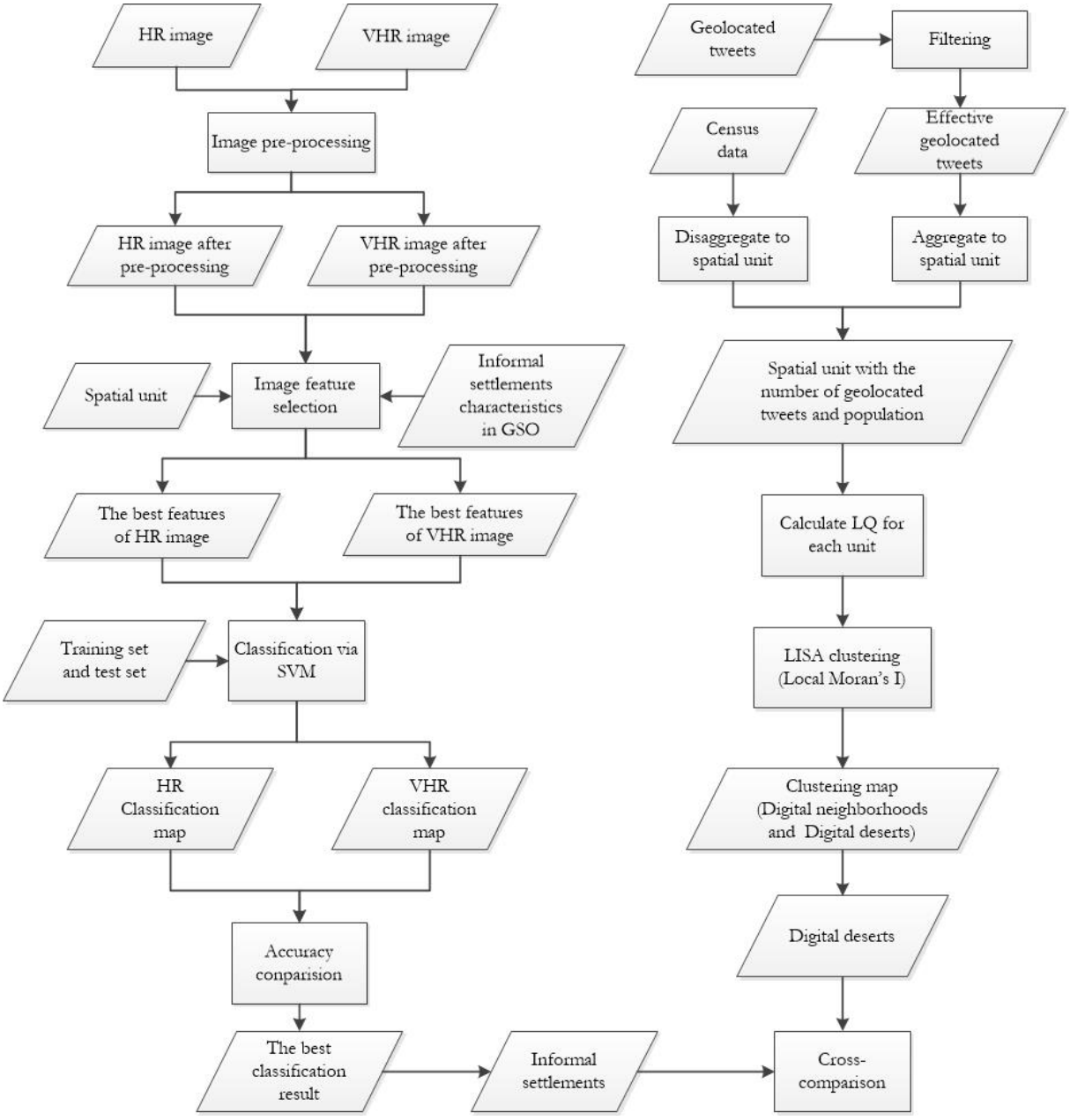


Figure 4.1 Flowchart of methodology.

4.2. Classification of HR and VHR Satellite Image

4.2.1. Satellite Image Pre-processing

This research uses two different satellite images. Image pre-processing is performed before image classification. This includes atmospheric correction as well as cloud removal are required in the beginning.

4.2.1.1. Sentinel-2 Image Pre-processing

This research acquires Sentinel-2 data (R, G, B and NIR bands) in Level-1C. Atmospheric correction can be achieved by converting Level-1C to Level-2A using sen2cor module in ESA snap toolbox. The output of this process is an orthoimage Bottom-Of-Atmosphere (BOA) corrected reflectance product (Wurm et al., 2017).

Since the Sentinel-2 data extracted from ESA had approximate 0.2% thick clouds and cloud-shadows, which also alter the ground local luminance, the cloud affected areas were restored by cloud removal methods. As no available cloud-free image cover Dar es Salaam, this study uses multitemporal images[5] as reference to generate a cloud-free image. The cloud cover areas could be replaced with the corresponding areas in the reference images, and a histogram matching approach is used to generate cloud-free image with proper brightness, this process was done by Mosaic in ArcMap.

4.2.1.2. Pleiades-1A Image Pre-processing

The Pleiades-1A image has a resolution 2 meters (R, G, B and NIR bands). ASTRIUM, (2012) indicates that all Pleiades-1A images have been corrected for radiometric and geometric based on internal parameters, ephemeris and attitude measurements. In addition, atmospheric correction is required to retrieve the surface reflectance from Pleiades-1A image by removing the atmospheric effects, this step is essential for enhancing the image classification.

4.2.2. Data Updating

As mentioned in 3.2.3, this study employs land use data as reference for training set preparation. Satellite images were captured from 2016, but the latest available land use data dates from 2010. Inevitably, land use changes over time as a result of anthropogenic activities (Gómez, White, & Wulder, 2016). Therefore, the 2010 land use data need to be updated before training set selection. The adopted procedure was as follows:

- convert the land use shapefile to a kml file and import to Google Earth
- load 2016 imagery and overlay land use kml
- identify changed land use areas and modify land use polygons accordingly
- convert the updated land use kml to shapefile

For example, figure 4.2 shows an example with informal settlements: blue areas are unchanged informal settlements, yellow areas are new informal settlements interpreted and delineated in Google Earth.

[5] Multitemporal reference images are obtained from European Space Agency (ESA). The date of three Sentinel-2A image are 30-06-2016, 30-07-2016, 07-03-2017. Free to download from http://sentinel-pds.s3-website.eu-central-1.amazonaws.com

Figure 4.2 Informal settlements changes from 2010 to 2016 (blue is original informal settlements in 2010, yellow represents new informal settlements in 2016).

4.2.3. Spatial Unit Selection

Previous studies have stated that an aggregation to a larger spatial units (beyond pixels) is useful for informal settlements analysis (Duque et al., 2016; Taubenböck & Kraff, 2014). The commonly used spatial unit is building blocks (Duque et al., 2016). In general, building blocks can be visually delineated based on morphological homogeneity and the close meshed road network. OSM is the source to download the road network layers which can be used for generating building blocks. The OSM data for Dar es Salaam provide a very detailed street network, however, due to the informal settlements cover more than 75% of the entire city and its complexity, the road network in this city is quite dense and unclosed. Therefore, the delineation of building blocks would need visual interpretation and manual digitalization for arranging roads (Duque et al., 2016), but this approach requires considerable processing time and workload, which was not feasible within the time constraints of this MSc research.

Duque et al., (2016) suggested an alternative spatial unit, which is automatically generated based on regular grids to perform informal settlements detection. Such a regular grid can be created in any GIS software, it only requires setting the boundary of the study area and the size of grid cells. Comparing with the size of actual urban blocks, we decide to use 200m×200m, 100m×100m, 75m×75m, and 50m×50m fishnet to extract image features and to test which gird size could get the best classification accuracy. In this stage,

we pick a small but mixed area that covers 31.2 km² to test different grid sizes. Appendix 2 shows different grid sizes covering the VHR and HR images.

4.2.4. Image Feature Extraction

From a spatial perspective, informal settlement can be distinguished obviously from other constructions, high resolution satellite imagery can capture unique spatial characteristics of informal settlements. However, image classification is not a straightforward process (Graesser et al., 2012), because an urban area is a heterogeneous landscape, different classes may have similar spectral information (e.g. glass and trees, formal settlements and informal settlements). Therefore, the use of only spectral features is not sufficient for informal settlements identification.

In this study, we use FETEX 2.7 to extract image features from three perspectives: spectral features, structure features as well as texture features. FETEX 2.7 is an interactive computer package for object-oriented feature extraction (Ruiz, Recio, Fernández-Sarría, & Hermosilla, 2011). This tool was written in IDL 6.2 and it can be performed in ENVI 4.2 of higher (Ruiz et al., 2011). FETEX 2.7 extracts image features from images by processing pixels in each block, without changing resolution and pixel value (Duque et al., 2016). Spectral features reflect colour characteristics, while structure and texture features present the spatial arrangement for each interest entities in the image (Duque et al., 2016).

Spectral features:
Spectral features directly reflect the spectral response of objects, depending on different land cover types, state of vegetation, soil composition, roof materials, etc (Ruiz et al., 2011). These features are helpful to distinguish spectrally homogeneous objects (Ruiz et al., 2011). Unlike the original RGB bands, spectral features provide a summary statistics for each grid based on the value of inside pixels (Duque et al., 2016). For each band included in the input image, FETEX 2.7 can calculate 7 features including the values of mean, standard, deviation, minimum, maximum, range, sum and majority in each grid (Ruiz et al., 2011).

Texture features:
Texture features present the spatial distribution of intensity values in the image (Ruiz et al., 2011). FETEX2.7 extracted texture features based on pixels inside each grid. The first order histogram features are the simplest approach to provide texture characteristics, FETEX 2.7 extracted kurtosis and skewness which manifest the distribution of value of the histogram in a grid. The most popular texture features are second order histogram features, based on the grey level co-occurrence matrix (GLCM) (Ruiz et al., 2011). GLCM is a statistical approach to measure the relationship between pixels and their direct spatial neighbourhood (Wurm et al., 2017). The extent of interpretation is defined by matrix size (Wurm et al., 2017) and the matrix can be calculated for four directions (0°, 45°, 90°, 135°). FETEX 2.7 is available to measure 7 GLCM features including contrast, uniformity, entropy, variance, covariance, inverse difference moment and correlation (Ruiz et al., 2011). Since FETEX 2.7 is a block-based feature extraction tool, these textures features are calculated as one value in each grid. As for the directions of the matrix, FETEX 2.7 measured the average value of four orientations (0°, 45°, 90°, 135°) to avoid the effect of orientation of the element in the grid (Ruiz et al., 2011). Edgeness factor is another helpful texture feature which reveals the density of edges present in a neighbourhood and has good performance in different landscape classification problems (Ruiz et al., 2011). FETEX 2.7 measures the mean and the standard deviation of edgeness factor for each grid.

Structure features:
Structure features reflect the spatial distribution of elements in each grid based on the randomness or regularity of the arrangement of elements (Ruiz et al., 2011). FETEX 2.7 calculates structure features using the experimental semivariogram method. The semivariogram is a useful approach to characterize regular

patterns, it not only measures the spatial associations of the value of a variable, but also calculate the degree of spatial correlation between pixels (Ruiz et al., 2011). FETEX 2.7 calculate the mean of the semivariogram by measuring six orientations (from 0°, 30°, 60°, 90°, 120°, 150°) in each grid. After that, using Gaussian filter to smooth each semivariogram curve so that avoid experimental fluctuations (Ruiz et al., 2011). FETEX 2.7 extracts semivariogram features based on the zonal analysis which is defined by a set of singular points on the semivariogram, like the first minimum, the first maximum, and the second maximum.

4.2.5. The Dataset

The shape of study area is irregular, thus a small part of ocean has to be added as one category. This study employed land use data as a reference to manually label grids into four categories: formal settlements, informal settlements, other (include industrial, commercial, educational, forest, green space as well as transport facilities), as well as ocean.

The entire study area is covered by 39738 grids in total. In total, 7129 samples were labelled for land area and the ocean. Figure 4.3 shows the distribution of samples. The next step is to randomly split all the samples into a training set and test set. Thus, 60% of the labelled samples were used for training and tuning the SVM classifier and 40% of labelled samples to test and evaluate the predictive ability of SVM classifier. Table 4.1 shows details of the dataset.

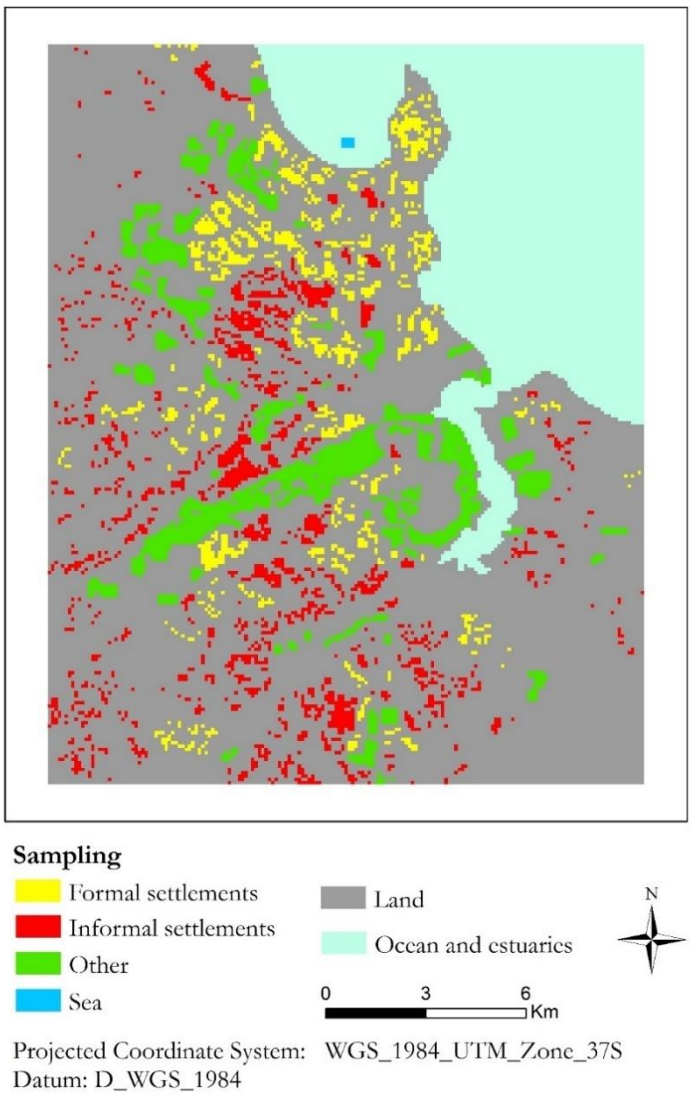


Figure 4.3 The distribution of samples.

Table 4.1 Composition of the dataset.

Number of grids	Labelled	No labelled	Formal settlements	Informal settlements	Other	Ocean	Training	Test
39738	5936	33802	1394	2059	2471	12	3512	2424

(Total grids include 8292 grids are ocean and estuaries, 31446 grids are land area)

4.2.6. Image Feature Selection

Feature selection is an important step before classifying the image. The main goal of a feature selection algorithm is to select a subset of significant features and reduce the running time in classification. Hilbert-Schmidt Independence Criterion (HSIC) is a kernel method which measures the nonlinear dependence between features and classes (Camps-valls, Member, Mooij, & Schölkopf, 2010). This filter approach calculates the Hilbert-Schmidt norm of the cross-covariance operator of the samples in the corresponding Hilbert space (Camps-valls et al., 2010; Gretton, Bousquet, Smola, & Schölkopf, 2005). This approach starts with the full feature set, and the output are the ranked features. Sequential Forward Selection (SFS) is the simplest greedy feature selection algorithm which starts from an empty feature set (Ladha & Deepa,

2011). SFS gradually add feature x^+ and combine with previous selected features Y_k, the aim is to achieve the highest objective function $J(Y_k+x^+)$ (Ladha & Deepa, 2011). In each iteration, the algorithm selected the most outstanding feature among the remaining available features which have not been selected to the feature set (Marcano-Cedeño, Quintanilla-Domínguez, Cortina-Januchs, & Andina, 2010). Following functions describe each stage in SFS.

Step 1. Start from an empty feature set

$$Y_0 = \{\emptyset\} \tag{10}$$

Step 2. Select the next outstanding feature

$$X^+ = argmax[J(Y_k + X)]; x \notin Y_k \tag{11}$$

Step 3. Update

$$Y_{k+1} = Y_k + X^+; k = k + 1 \tag{12}$$

Step 4. Return to step 2

The procedure starts from step 1, repeats step 2 to step 4 until the number of selected features reach the predefined number or the predictive result does not improve any further (Schaffernicht, Möller, Debes, & Gross, 2009).

4.2.7. Classification Accuracy Assessment

For supervised classification, the common way to assess accuracy is using the confusion matrix. The confusion matrix is also called error matrix, it is a table showing the degree of misclassification among all the classes (Gómez & Montero, 2011). Another popular measure is the overall accuracy which is obtained from the confusion matrix and it shows the percentage of cases correctly classified (Gómez & Montero, 2011). But the overall accuracy in unable to differentiate the number of correct labels of different classes (Sokolova, Japkowicz, & Szpakowicz, 2006)

In addition, recent studies introduce other methods for evaluation measures, such as recall, precision and F-score. Table 4.2 shows a confusion matrix for binary classification, and formula 13, 14, 15 present recall, precision and F-score, respectively, based on the values of the confusion matrix (Sokolova & Lapalme, 2009).

Table 4.2 Confusion matrix for binary classification (Sokolova & Lapalme, 2009).

Data class	Classified as positive	Classified as negative
Positive	*true positive (tp)*	*false negative (fn)*
Negative	*false positive (fp)*	*true negative (tn)*

$$\text{recall} = \frac{\text{tp}}{\text{tp+fn}} \tag{13}$$

$$\text{precision} = \frac{\text{tp}}{\text{tp+fp}} \tag{14}$$

$$\text{F} - \text{score} = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}} \tag{15}$$

Recall also refer to as Sensitivity, shows the effectiveness of a classifier to identify positive labels (Sokolova & Lapalme, 2009). Precision represents the class agreement of the data labels with positive labels derived from the classifier (Sokolova & Lapalme, 2009). F-score denotes the harmonic mean of recall and precision (Sokolova & Lapalme, 2009). In this study, the accuracy of classification will be measured by confusion matrix, overall accuracy, recall, precision and F-score.

4.3. Social Media Data Analysis

4.3.1. Twitter Data Filtration

Since this study is interested in comparing informal settlements identified from satellite images and digital desert derived from geolocated tweets, we need tweets with highly granular geolocation. Therefore, we only collect geolocated tweets whose location is automatically recorded as coordinate by Twitter via GPS on mobile equipment, not manually selected by twitter users. These geolocated tweets with coordinates are presented as points on the map, the coordinates are resolved to five decimal places, approximately 1 meter (Li et al., 2013), and we expect GPS in mobile equipment has a good accuracy which could be several meters, so the location information could precisely presented.

The original twitter dataset contains 127,006 geolocated tweets covering the study area from January 1st to December 31st in 2015. However, there are a considerable quantity of geolocated tweets has fixed location, these might be created by non-mobile terminals but with GPS. Frias-Martinez et al., (2012) supposed that some mobile advertising companies would send a large number of daily tweets to terminals. These tweets with fixed coordinate cannot provide relevant information about the location of mobile users. Therefore, to eliminate these tweets, we adopted the filtering rule proposed in Frias-Martinez et al., (2012): any GPS location generates more than 20 tweets per user in one day should be deleted. Furthermore, we observed some geolocated tweets located on the roads. In this case, users might send tweets on the road to destination, but we cannot distinguish these users live in informal settlements or formal settlements. Therefore, we also removed geolocated tweets cover all the roads, otherwise, these irrelevant tweets would increase the number of geolocated tweets in the grids which cover the roads. After data filtration, the total number of geolocated tweets was reduced to 114,736.

4.3.2. Twitter Data Aggregation

Since the geolocated tweets are distributed anywhere at city scale, it is difficult to visually estimate which location has more or less Twitter users. Thus, we need to build a block for aggregating geolocated tweets, exploring the relationship between the total number of geolocated tweets and population in each block. Review the approach presented in Klotz et al., (2017), they used building blocks which are visually described from image based on morphological homogeneity and adjacent road network. But in this work, it was infeasible to generate building blocks, so the geolocated tweets were aggregated into regular grids which were used as the spatial unit for image classification (refer to section 4.2.3).

4.3.3. Estimation and Disaggregation of Census Data

In this step, the challenge is to capture population numbers in each grid. Census is a complicated and wide-range activity which commonly take once in one decade. In Dar es Salaam, the smallest unit of census data is in subwards and the latest available data was from 2012 and 2002. As the number of population is a changeable variable and the twitter data used in this work were captured in 2015, the population number for 2015 was estimated for the subwards based on the trend from 2002 to 2012.

The next step was to disaggregate population number from subwards to grids. Building footprints from OSM are used to generate estimated the number of population in grids. Following shows the approach to disaggregate population grids based on census data and building footprints (Martin, 2015).

Step 1. Select and export all building footprints that are used for residential purposes.
Step 2. Calculate the area for each residential building. a_i
Step 2. Calculate the area of all the residential building in each subward.

$$S = \sum_{a=1}^{n} a_i \quad (16)$$

Step 3. Calculate the share (R) of all the residential building area for each residential building in each subward.

$$R = \frac{a_i}{S} \tag{17}$$

Step 4. In each subward, multiply the population data (P) with the share of each residential building. O is the estimated population number in each residential building.

$$\mathrm{O} = \mathrm{P} \times \mathrm{R} \tag{18}$$

Step 5. Assign the estimated population data of the residential building footprints to each grid.

According to the number of people and the number of geolocated tweets in each grid, LQ mentioned in 2.6 can be calculated based on formula 9.

4.3.4. Clustering Analysis

After obtaining the distribution of LQ, this research is interested in investigating the relationship between each grid and their surrounded grids. Therefore, this step will employ Local Indicators of Spatial Association (LISA) to detect clusters from spatial distribution of LQ. LISA is a statistical method which satisfies the following two requirements:

1. the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation (Anselin, 1995);
2. the sum of LISAs for all observations is proportional to a global indicator of spatial association (Anselin, 1995).

For a variable y_i which is observed at location i, the LISA (L_i) can be expressed as following:

$$L_i = f(y_i, y_{J_i}) \tag{19}$$

function f probably contains additional parameters, and y_{J_i} represents the values observed in the neighbourhood J_i of i.

Klotz et al., (2017) explained that clusters are assembled by blocks, which have similar values with regard to their surrounded blocks at city scale as compared to total randomness. Local Moran's I is one of the forms which analyses the degree of spatial variation between each block and their surrounded blocks.

For each observation i, the local Moran's I (Anselin, 1995) can be expressed as

$$\mathrm{I} = \left(\frac{n}{\sum_i \sum_j w_{ij}}\right) \sum_i \sum_j w_{ij} z_i z_j / \sum_i z_i^2 \tag{20}$$

where z_i and z_j are the deviations from the mean and the summation over another observation j, only neighbouring values j$\in J_i$ are included (Anselin, 1995). w_{ij} is a distance weighting between z_i and z_j, it also can be defined as the inverse of the distance. The positive I_i means the cluster has similar values (high or low) while negative value indicates a cluster of dissimilar values.

In this work, the grids with different LQ value were summarized as table 4.3. If grids with high LQ surrounded by grids with high/low LQ, the clusters aggregated by these grids are hot spots, represent digital neighbourhood. On the contrary, the grids with low LQ surrounded by grids with high/low LQ, these clusters are clod spots, represent digital desert.

Table 4.3 Characteristics of four types of cluster (Klotz et al., 2017).

Characteristics	Label	Cluster type
Blocks with **above** average LQ surrounded by blocks with an **above** average LQ	HH	Digital Neighbourhoods
Blocks with **above** average LQ surrounded by blocks with a **below** average LQ	HL	
Blocks with **below** average LQ surrounded by blocks with an **above** average LQ	LH	Digital Deserts
Blocks with **below** average LQ surrounded by blocks with a **below** average LQ	LL	

4.4. Relationship Analysis

To quantitatively confirm these observations, this step used cross-comparison matrix to analyse how do the digital neighbourhoods and digital deserts quantitatively related to classes observed by remote sensing approach. The comparison was described based on the grids. Since we were not interested in ocean and estuaries, and no geolocated tweets occurred in this area. So, it was excluded in the comparison.

5. RESULTS

5.1. The Important Features and Classification Results Comparing VHR and HR Images

5.1.1. Local Characteristics of Informal Settlements Based on GSO

Informal settlements in the world mostly have common spatial characteristics: buildings are small and clustered with high density, lack open space around a household and the roads are irregularly arranged (Kuffer & Barros, 2011). Informal settlements in Dar es Salaam also share these properties, and the following paragraphs provide their details in terms of the GSO.

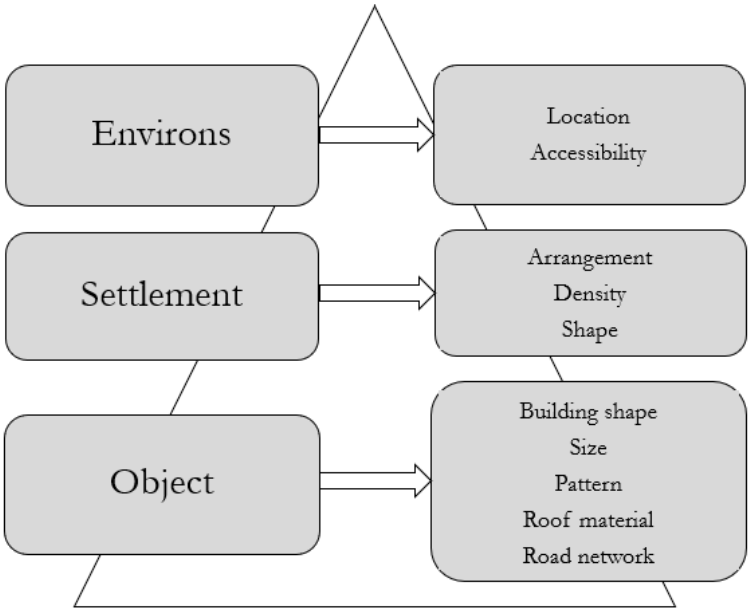


Figure 5.1 GSO in three levels with corresponding indicators in Dar es Salaam

Figure 5.1 provides an overview of GSO in three levels with corresponding indicators in Dar es Salaam. At object level, informal settlements in Dar es Salaam can be defined by building size, shape, pattern as well as roof type. A visual interpretation of the VHR satellite image show that the size of the houses in informal settlements are varying. In general, houses have a rectangular shape with size from 50 m^2 to 150m^2. Some houses have larger size depends on the family size (Sirueri, 2015). Physically, one of the major characteristics of informal residential building is that they are single storey houses which follow the traditional Swahili house (Rasmussen, 2013). The most popular roof material of residential buildings is corrugated iron sheets (Kuffer, 2003). The road network can be described as narrow and labyrinthine streets, being commonly not accessibly for motor vehicles (Rasmussen, 2013).

At settlements level, neighbourhoods present a compact arrangement within an organic urban structure. High densities make the buildings very crowded, generally densification increases in neighbourhoods near the city centre (Rasmussen, 2013). Another obvious characteristic is that houses were constructed along the main road (Rasmussen, 2013), therefore, the shapes of informal settlements depend on the surrounding streets.

From an environ perspective, living environments vary a lot – poor households are often located haphazard areas (Rasmussen, 2013). Access to basic infrastructure is a problem in most areas. For

instance, it is common to see pit latrines or shower rooms are placed facing sidewalks (Rasmussen, 2013). In addition, the narrow roads cause serious problems when the pit latrines need to be cleaned (Rasmussen, 2013).

5.1.2. Extracted Features Which Describe the Local Ontology for Informal Settlements

This study aims to classify VHR image and HR image into 3 classes: formal settlements, informal settlements and other which includes industrial, commercial, educational, park, green space as well as transportation. To distinguish different classes, spectral information is the simplest way to identify different objects with various colour. In a spectral feature set, we extracted mean, standard deviation and majority statistic for each R, G, B, and NIR band from VHR image and HR image respectively, these three features contain more information about the spectral variances for different classes than other summary statistics (minimum, maximum, range and sum) (Duque et al., 2016). However, as mentioned in section 2.1, people living in informal settlements in Dar es Salaam preferred to use permanent and modern materials for construction, so it might be difficult to differentiate informal settlements and formal settlements based on the roof colour. Therefore, the use of texture features might be helpful to fill this gap.

In Dar es Salaam, informal settlements are aggregated in large areas, show irregularly ordered and small house size. Within some small informal settlements, no road network could be detected, while in larger informal settlements, several irregular and narrow paths could be discovered. Therefore, the texture of informal settlements looks more dense and irregular than formal settlements. Furthermore, structure features also can help to describe the difference between informal settlements and formal settlements, especially detecting the regular patterns like formal settlements. Table 5.1 illustrates spectral, texture, structure features extracted from FETEX 2.7.

Table 5.1 Image features extracted from FETEX 2.7.

Group	Variable name	Description	Total
Spectral features	MEAN R, G, B, NIR	Mean of pixel values in R, G, B, NIR band	14
	SD R, G, B, NIR	Standard deviation of pixel values in R, G, B, NIR band	
	MAJ R, G, B, NIR	Majority of pixel values in R, G, B, NIR band	
	MEAN NDVI	Mean of NDVI	
	SD NDVI	Standard deviation of NDVI	
Texture features	UNIFOR	GLCM uniformity	11
	ENTROP	GLCM entropy	
	COVAR	GLCM covariance	
	CONTRAS	GLCM contrast	
	IDM	GLCM inverse di erence moment	
	CORRELAC	GLCM correlation	
	VARIAN	GLCM variance	
	ME	Mean of the edgeness factor	
	SDE	Standard deviation of the edgeness factor	
	SKE	Skewness value of the histogram	
	KUR	Kurtosis value of the histogram	
Structure features	RVF	Ratio variance at first lag	10
	RSF	Ratio between semivariance values at second and first lag	
	FDO	First derivative near the origin	
	SDT	Second derivative at third lag	
	MFM	Mean of the semivariogram values up to the first maximum	
	VFM	Variance of the semivariogram values up to the first maximum	
	DMF	Difference between the mean of the semivariogram values up to the first maximum and the semivariance at first lag	
	RMM	Ratio between the semivariance at first local maximum and the mean semivariogram values up to this maximum	
	SDF	Second order difference between first lag and first maximum	
	AFM	Area between the semivariogram value in the first lag and the semivariogram function until the first maximum	
Total			35

5.1.3. Comparison Between Different Grid Size

To select the best grid size for block-based classification, four different grid sizes were tested on a small part of VHR and HR images. This process selected the top 10 features derived by SFS. Table 5.2 shows the overall accuracy of different grid sizes. For HR image, it is impossible to extract features based on a 50 m grid because too less pixels are contained in such a small grid. The results show 100 m grid for both VHR and HR images outperformed the other three grid sizes to identify formal settlements, informal settlements and other land cover. Therefore, 100 m grid is the best spatial unit for VHR and HR images classification, and it is also the unit for the following twitter data analysis.

Table 5.2 Grid size comparison.

Satellite image	Pleiades-1A				Sentinel-2A			
Grid size (Square cell)	200 m	**100 m**	75 m	50 m	200 m	**100 m**	75 m	50 m
Overall accuracy	84.9%	**93.0%**	87.4%	85.2%	80.6%	**85.5%**	76.5%	N/A

5.1.4. The Most Important Features for VHR and HR Images Classification

To better understand the discriminating capacity of the image features for the distinction between formal settlements, informal settlements and other land cover. SFS as a powerful feature selection algorithm singled out the best 10 features for VHR image and HR image.

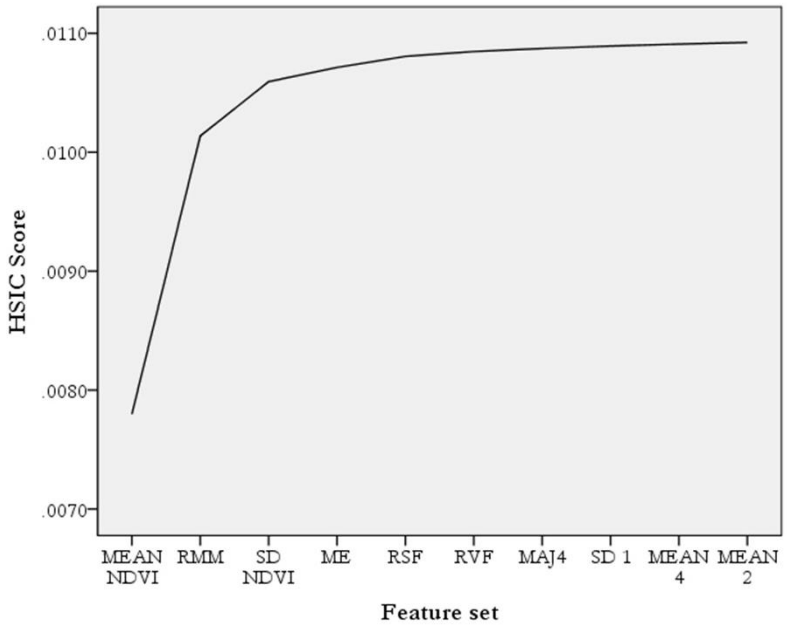


Figure 5.2 The best feature set, showing the HSIC score for VHR image (Score is achieved by cumulative features).

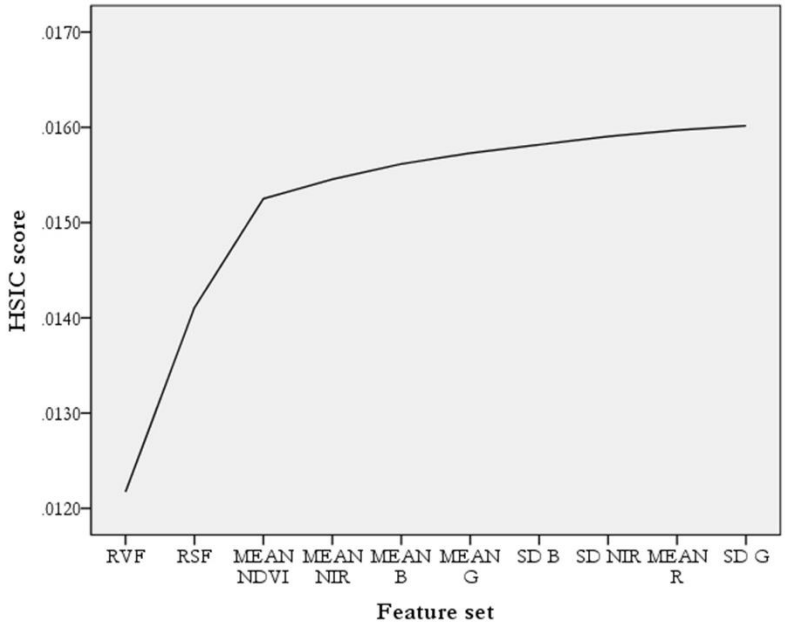


Figure 5.3 The best feature set, showing the HSIC score for HR image (Score is achieved by cumulative features).

The best feature sets and corresponding HSIC score for both VHR and HR images are illustrated in figure 5.2 and 5.3. For both two images, the HSIC score rapidly increases when the best three features are selected, subsequently, it grows very slowly. The detail of features sets is shown in table 5.3. As can be seen from the table, spectral features are comparatively important for both VHR and HR image classification. It is very interesting that three of the top 10 features (MEAN NDVI, MEAN NIR, MEAN G) are significant for both images which indicates that R, G and NIR bands can reflect the spectral discrepancy between residential areas and other land cover. NDVI also has a good performance to identify informal settlements because less green spaces exist in their surroundings. About structural features, RVF is an indicator that provides the information about the relationship between spatial correlation at long and short distances (Balaguer, Ruiz, Hermosilla, & Recio, 2010). RSF describes the changes in the variability of data at short distances. Whereas, RMM complements the information provided by RVF, considering the effect of the total variability of the data (Balaguer et al., 2010). One of the differences between formal settlement and informal settlements is that informal settlements are always randomly arranged (at neighbourhood level), while formal settlements are regularly arranged. These three structural features describe the variability of spatial arrangement in each grid. However, for both images, only ME as a texture features contributes to the accuracy improvement of VHR image. ME measures the density of the edges shown in the image, the value of ME in informal settlements is higher than formal settlements and other land cover because a lot of buildings with small sizes, narrow roads and limited open space lead to such dense settlements.

Table 5.3 Composition of the best feature sets.

Image	Spectral features	Texture features	Structural features
VHR image	mean NDVI, SD NDVI, majority NIR, SD R, mean NIR, mean G	ME	RMM, RSF, RVF
HR image	mean NDVI, mean NIR, mean B, mean G, SD B, SD NIR, mean R, SD G,		RVF, RSF

5.1.5. Comparison of Classification Accuracies

This step used the selected top 10 features shown in 5.1.4 and the samples presented in 4.2.5 as input using a SVM classifier with RBF kernel to identify formal settlements, informal settlements as well as other for the VHR and HR image. Figure 5.4 shows the classification results and table 5.4, 5.5 and 5.6 provides the confusion matrix, recall, precision and F-score for the VHR image and HR image.

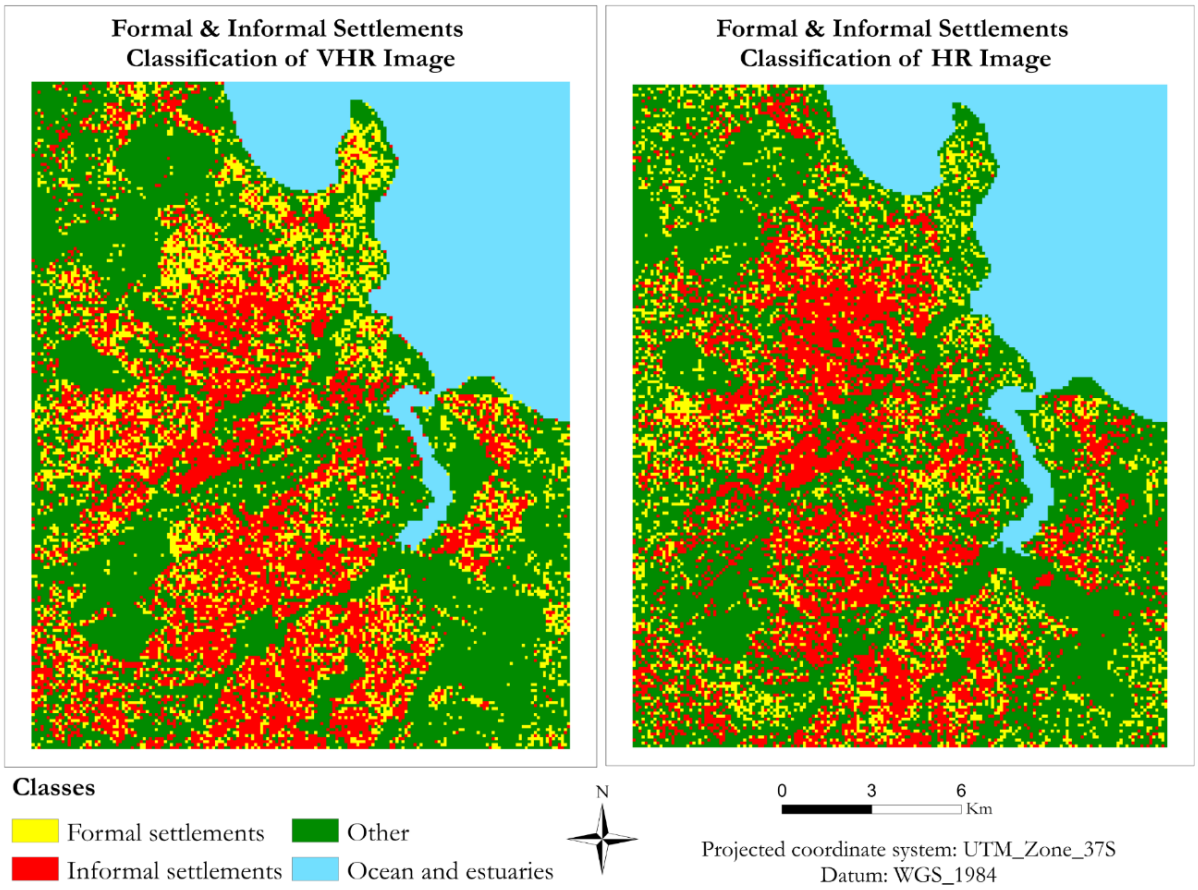


Figure 5.4 The classification maps.

The overall accuracy of HR image is around 72.5%, while the higher accuracy was derived from VHR image which is approximately 82.5%. Comparing the classification map of VHR image and HR image, it shows that the HR image identified more informal settlements than VHR image. About 58.4% of residential areas are identified as informal settlements in VHR image, while approximate 63.1% of residential areas are recognized as informal settlements in HR image. It seems that a lot of formal settlements were misclassified as informal settlements in HR image classification. The recall of formal settlements of HR image also proved only 30.2% of formal settlements were accurately classified while 48.1% was misclassified as informal settlements. For the VHR image, however, only 66.7% of formal settlements is correctly predicted and 36.4% of formal settlements is misclassified as informal settlements. In summary, these results show that the VHR image achieves a higher accuracy than HR image and the misclassification between formal settlements and informal settlements exists in both images. Appendix 3 shows a sector from the VHR image and the classification results of VHR and HR images.

Table 5.4 Accuracy assessment of VHR image.

		Ground truth					
		Formal settlements	Informal settlements	Other	Sea	Total	Precision
Classes	Formal settlements	274	90	47	0	411	66.7%
	Informal settlements	187	768	38	0	993	77.3%
	Other	53	9	947	0	1009	93.9%
	Ocean	0	0	0	9	9	100.0%
	Total	514	867	1032	9	2422	
	Recall	53.3%	88.6%	91.8%	100.0%		**82.5%**

Table 5.5 Accuracy assessment of HR image.

		Ground truth					
		Formal settlements	Informal settlements	Other	Sea	Total	Precision
Classes	Formal settlements	155	52	111	0	318	48.7%
	Informal settlements	247	779	107	0	1133	68.8%
	Other	112	36	814	0	962	84.6%
	Ocean	0	0	0	9	9	100%
	Total	514	867	1032	9	2422	
	Recall	30.2%	89.9%	78.9%	100%		**72.5%**

Table 5.6 Summary of F-score of each class for VHR and HR image.

Images	Classes			
	Formal settlements	Informal settlements	Other	Sea
VHR image	59.30%	82.60%	92.80%	100%
HR image	37.30%	77.90%	81.70%	100%

5.2. Twitter Data Analysis

5.2.1. Twitter Data Aggregation and Census Data Disaggregation

We utilized 100 m grids as unit to aggregate randomly distributed geolocated tweets and disaggregate census data, figure 5.5 shows the population density in subwards in 2015, and provides an overview to show the residential buildings and 100 girds in Kisiwani. We calculated the number of geolocated tweets in each grid and the twitter density map is shown in figure 5.6. The geolocated tweets show a concentration in the areas around the city centre, especially commercial, educational area and formal settlements present in general higher twitter density. As for informal settlements, there are small number of grids with high twitter density, but generally most of informal settlements revealed low tweets density or no tweets. Figure 5.7 shows a 3D map of the twitter density combined with classes identified from VHR image. The peaks of twitter density appear in formal settlements and other which near the seaside, while only few of grids identified as informal settlements show higher twitter density. Figure 5.8 shows twitter density within residential areas, it can be seen clearer that more formal settlements match with the peaks of twitter density.

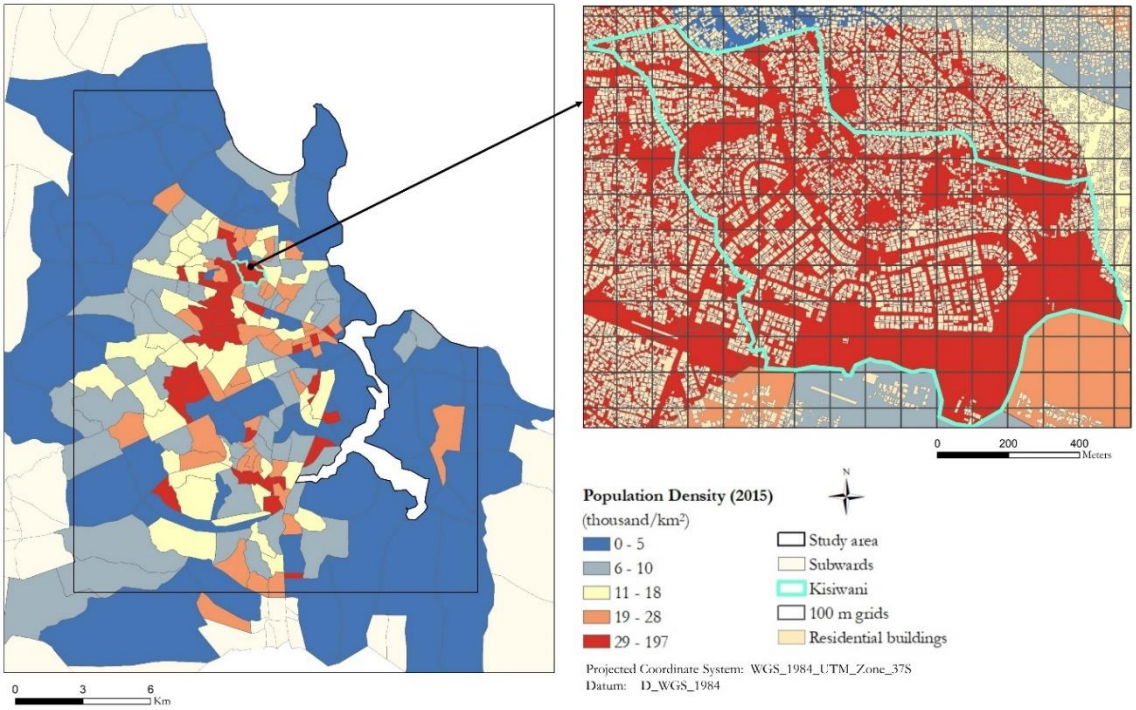


Figure 5.5 Population density in subwards (left); An overview of residential buildings and 100 grids shown in one subward (Kisiwani) (right).

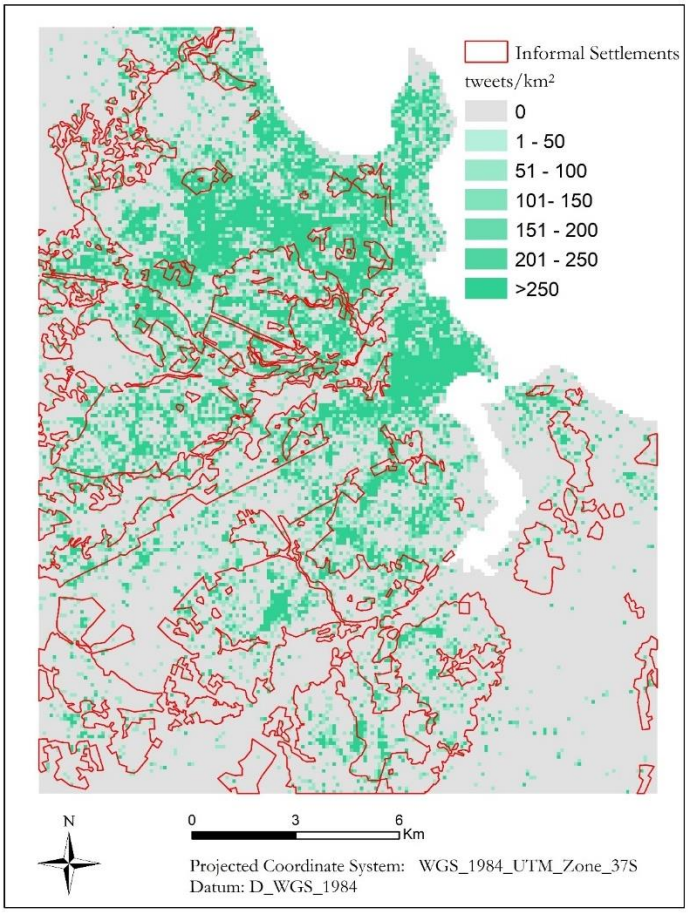


Figure 5.6 Twitter density map (informal settlements shown as red border polygon are derived from land use data).

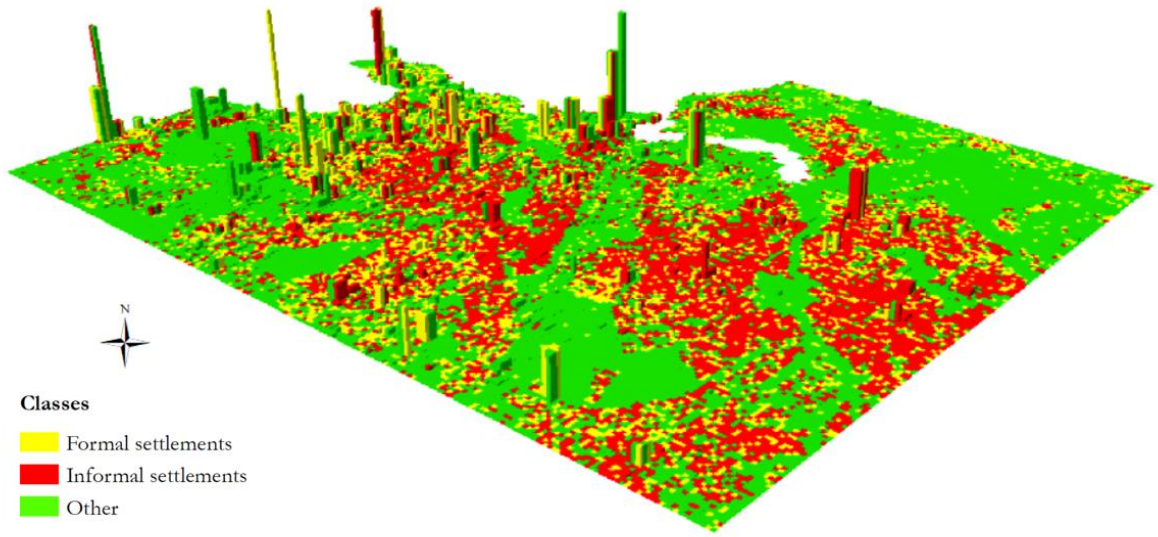


Figure 5.7 3D map – Twitter density within classes identified from VHR image (height represents twitter density).

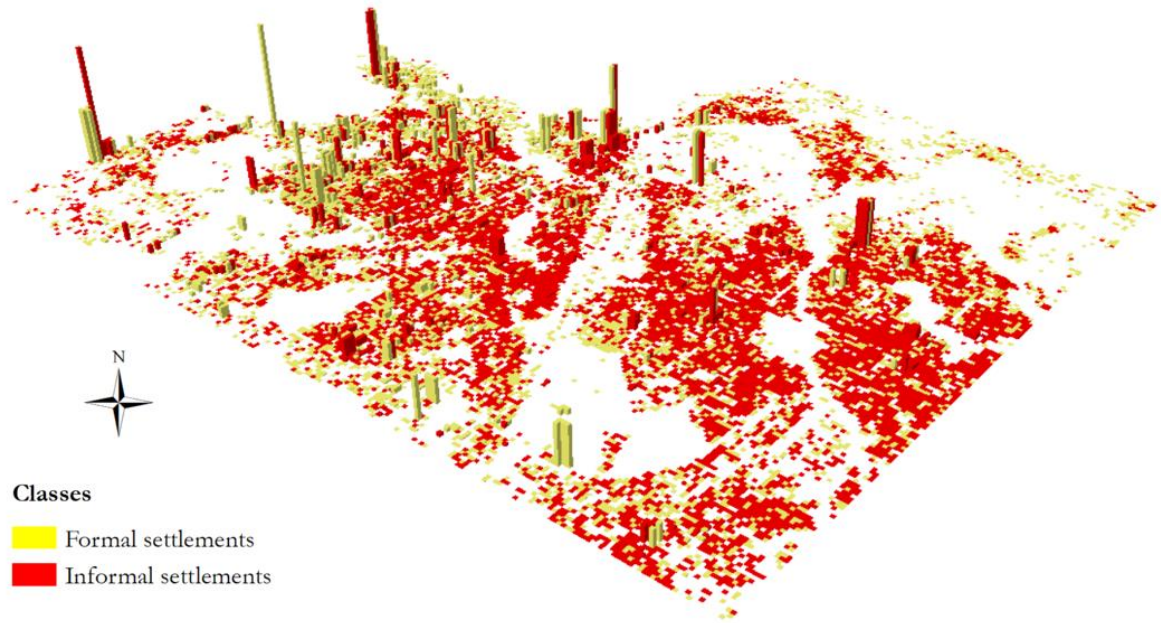


Figure 5.8 3D map - Twitter density within informal settlements and formal settlements identified from VHR image (height represents twitter density).

Figure 5.9 presents population density in each grid. We employed residential building footprint as reference to estimate population data. Therefore, the areas with dense residential buildings represent more people living at these locations. It can be seen from figure 5.9, there is a large area without population located in the middle dividing the study area into two parts. This area is used for industrial and transportation purposes. The informal settlements located on both sides of this area show higher population density. Formal settlements and informal settlements which are located far away from this area show relatively low population density. Factories always attracts low income people to settle nearby because of more affordable house prices, more job opportunities and the accessibility to city centre. Appendix 4 provides population density map to show the number of people per hectare.

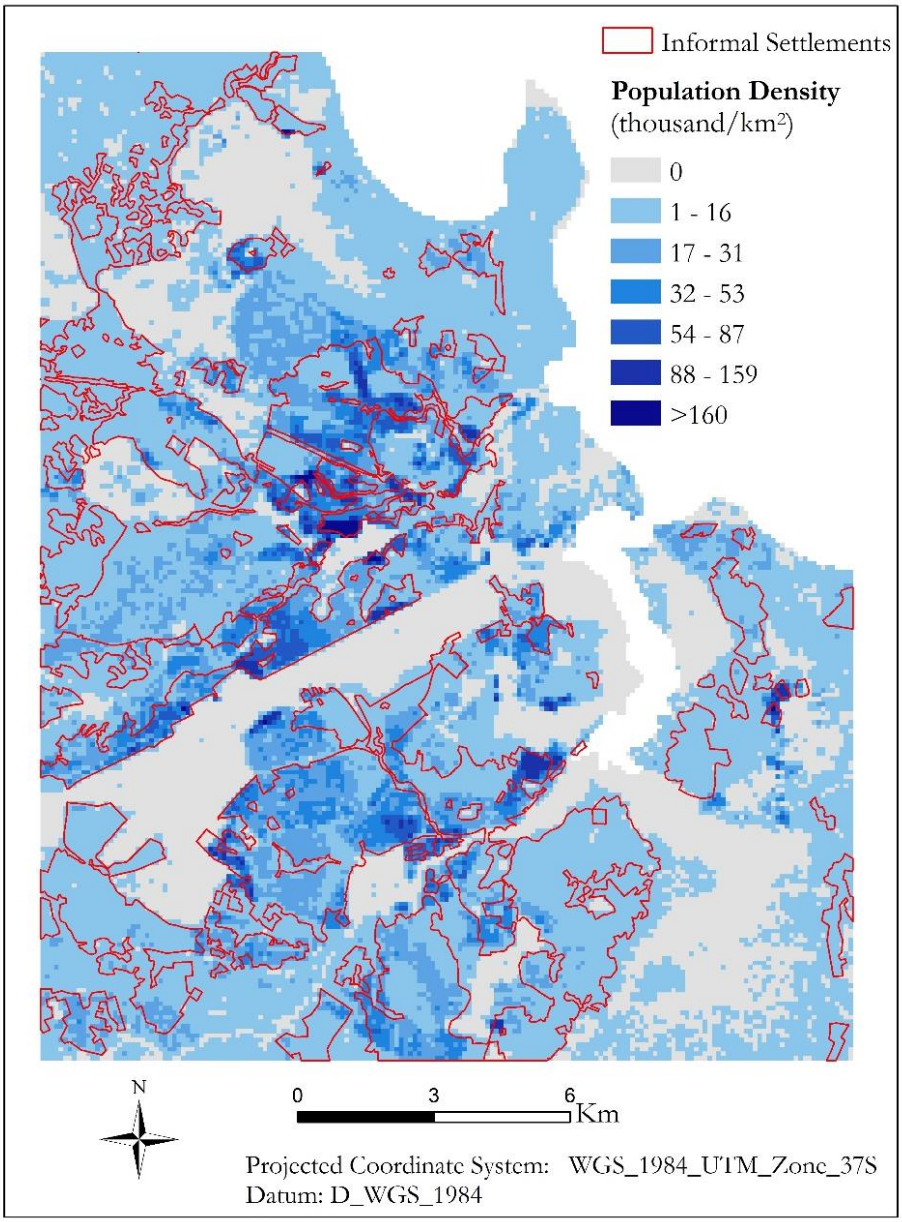


Figure 5.9 Population density in 2015.

5.2.2. The Distribution of Hot and Cold Spots

In the previous section, we already showed the number of population and the number of geolocated tweets in each grid. Thus, for each grid, the value of LQ was derived based on formula 9. In total, 15,317 grids located in residential areas have no recorded tweets. Therefore, these grids could be regarded as natural digital deserts (NT, LQ=0) (Klotz et al., 2017). In addition, 7,897 grids have no population, which means these grids are not residential areas. Therefore, in order to obtain meaningful results, the grids without tweets (NT) and the grids without population (NP) are excluded from subsequent clustering (Klotz et al., 2017). Finally, we transferred LQ to logarithmic form (figure 5.10) as input for clustering to reduce passive influence of highly skewed distributions on clustering results (Klotz et al., 2017).

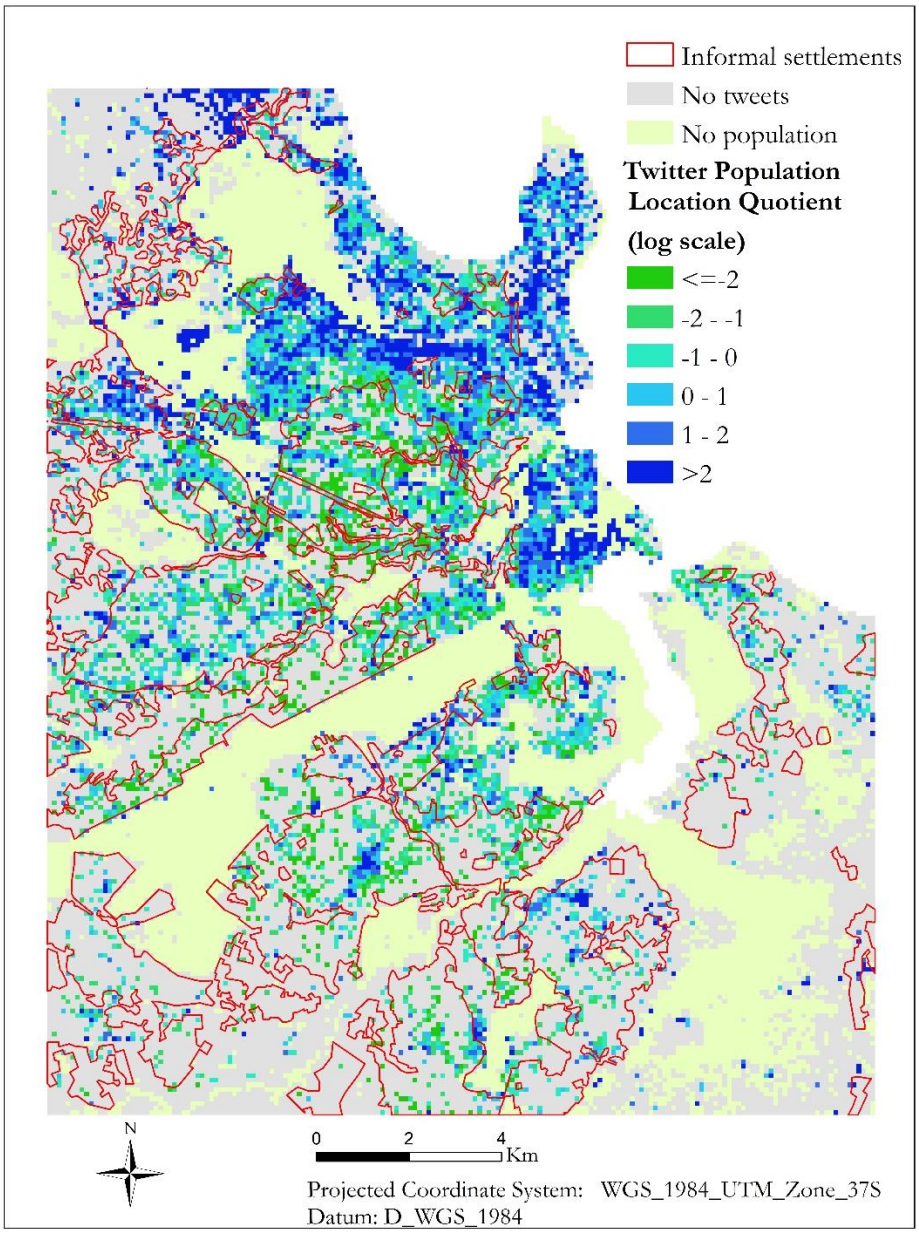


Figure 5.10 Log-transformed Twitter Population Location Quotient Map.

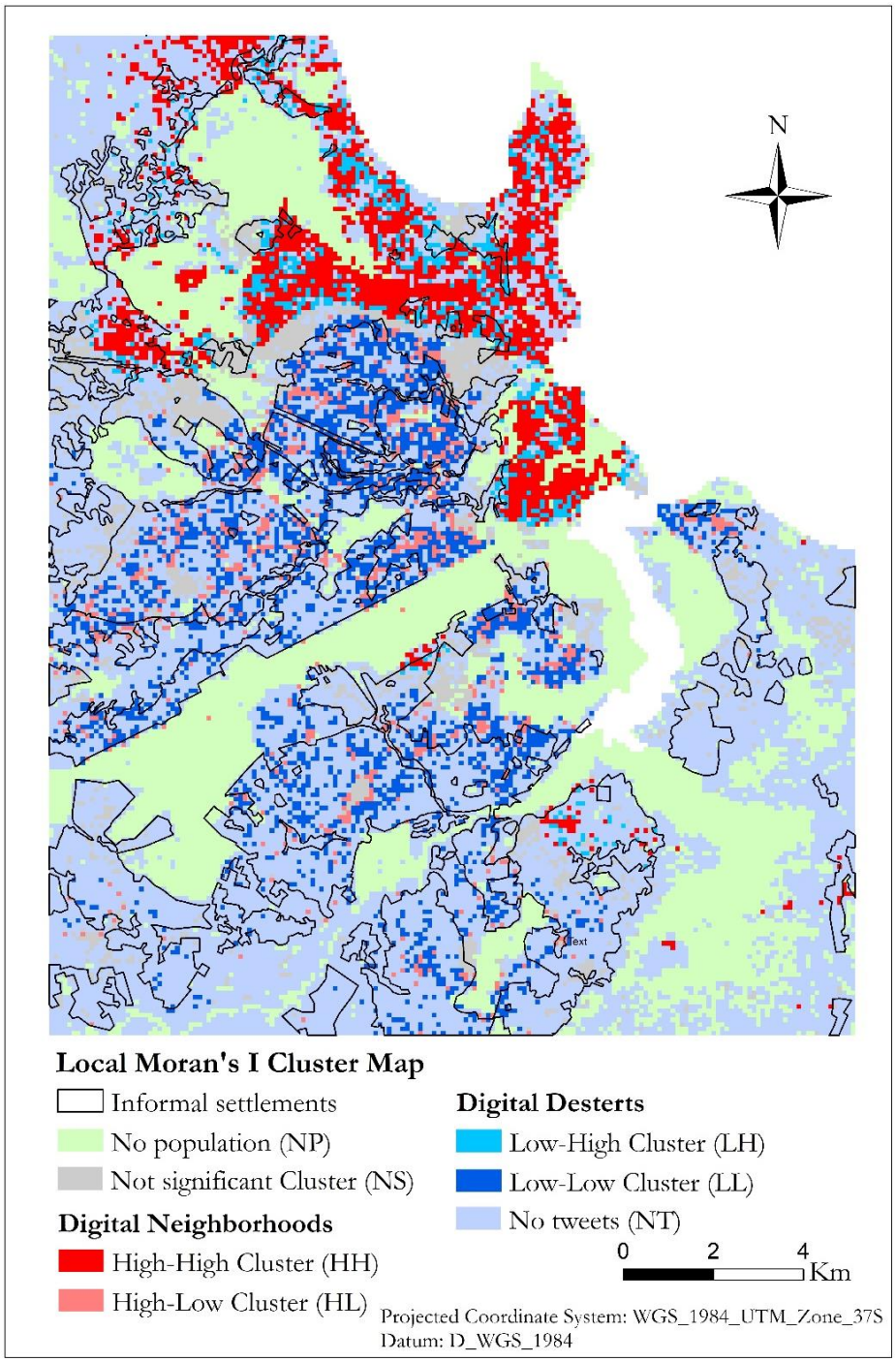


Figure 5.11 Local Moran's I Cluster Map.

To detect the hot spots and cold spots based on the spatial distribution of log-transformation of LQ, we utilized the Local Moran's I to cluster the grids which have more similar value with respect to the neighbours in the study area. Figure 5.11 spatially shows the clustering results and highlights the hot spots (digital neighbourhoods) and cold spots (digital deserts) in red/pink and blue/wathet respectively. The hot spots (HH) are the grids with an above average LQ encircled by grids with an above average LQ. On the contrary, grids with a below average LQ surrounded by grids with below average LQ constitute cold spots (LL). Outliers HL and LH are also regarded as hot and cold spots, respectively. In addition, the cluster map also includes NS (insignificant grids) and NP. As figure 5.11 shown, the digital neighbourhoods (hot spots) were concentrate in the north-eastern part around city centre (formal settlements or commercial

and residential mixed area). While digital deserts (cold spots) are located in the middle of the study area. It is interesting that some small digital deserts are located in or next to the digital neighbourhoods. Analogously, some small digital neighbourhoods distribute in the digital deserts. Although a large number of grids are digital deserts, which coincides well with the locations of informal settlements far away from the city centre.

5.3. The Relationship Between Informal Settlements and Digital Deserts

The final step of this study is to quantitatively analyse how the classes identified from VHR image related to neighbourhoods clustered via twitter data. Table 5.7 shows the statistical results in a cross-tabulate. We use 8232 grids (all the grids with tweets and population, LQ ≠ 0) as input in cluster detection, 36.4% are detected as digital neighbourhoods and 42.0% are detected as digital deserts, while other 21.6% grids are insignificant for clustering. Appendix 5 shows the cross-comparison matrix in the number of grids.

Table 5.7 Quantitative cross-comparison matrix (%).

	Digital neighbourhoods		Digital deserts				
	HH	**HL**	**LH**	**LL**	**NT**	**NS**	**NP**
Formal settlements	34.5%	23.1%	41.0%	23.6%	21.2%	28.3%	4.4%
Informal settlements	19.7%	36.2%	30.8%	56.5%	34.4%	32.3%	1.8%
Other	45.7%	40.7%	28.2%	19.9%	44.4%	39.4%	93.8%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

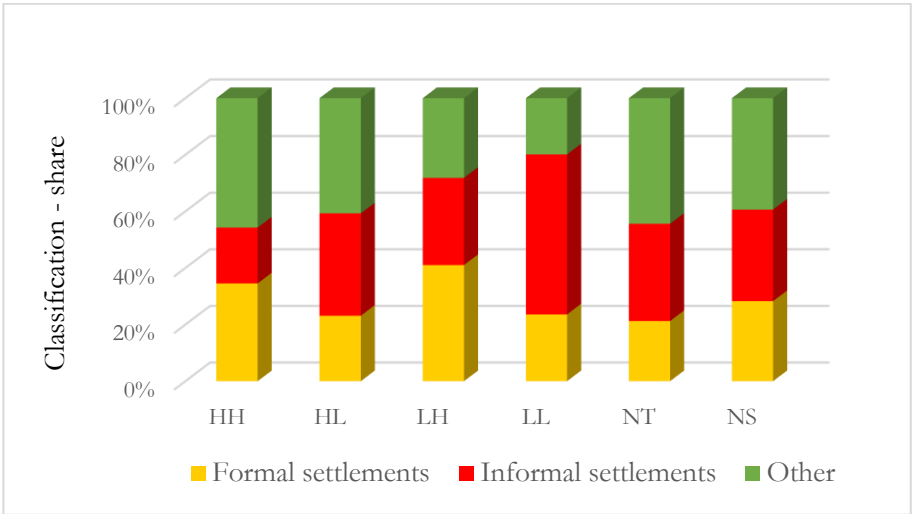


Figure 5.12 Share of classification results per cluster.

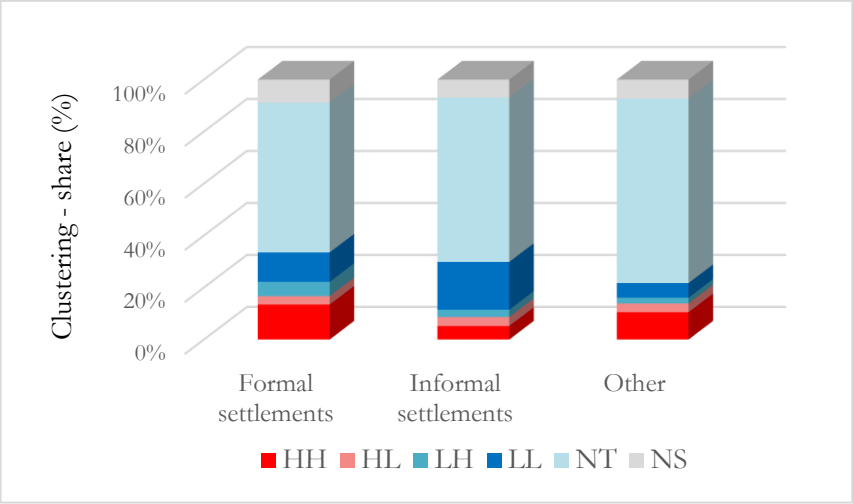


Figure 5.13 Share of clustering results per class

Figure 5.12 illustrates the percentage that the classes identified from VHR image share to the clusters detected based on twitter. About 20% HH and 36% HL clusters match with informal settlements classified from VHR image. 722 grids (include HH and HL) match with classified informal settlements, but after checked with the 2015 land use map, 180 grids located in formal settlements) Almost 57% of LL are informal settlements, however 41% of LH overlays with formal settlements. Other shares more natural digital deserts (44.4%) than informal and formal settlements and 39.4% of NS (insignificant grids) match with other. About 4.4% and 1.8% of NP (no people grids) are recognized as formal settlements and informal settlements, respectively. Because a small number of residential grids were misclassified to other. In principle, NP should match with other because it is no residential covered areas. Combining this with figure 5.13 which provides complementary information, for formal settlements, informal settlements and other, Natural digital deserts exists in every class but also occupy a large percentage (54% in formal settlements, 62% in informal settlements and 40% in other). Formal settlements have a significant share (12%) marked as HH (more than informal settlements which share 5.1% marked as HH). And informal settlements indeed have a better match with digital deserts than formal settlements and other.

5.4. The Exploration Between Informal Settlements and Digital Neighbourhoods

According to appendix 5, 722 girds are classified as informal settlements but recognized as digital neighbourhoods (434 grids are HH and 288 grids are HL). In order to explore whether these grids are real informal settlements or a number of them are misclassified as informal settlements, we compare these grids with the updated land used map. We find 258 (35.7%) grids are completely within formal settlements in land use map, 262 (36.3%) grids are mixed by formal settlements and informal settlements, while only 202 grids are located in informal areas in land use map. Therefore, it means only 202 (28.0%) grids are real informal settlements and recognized as digital neighbourhoods. Appendix 7 shows the spatial distribution of misclassified formal settlements, mixed residential areas and real informal settlements. Finally, we select 6 sectors (appendix 8) clustered by the grids which are real informal settlements and recognized as digital neighbourhoods, and check with Google Map to see the surroundings around these sectors. We find 5 sectors have schools (primary school, second school or university) in the surroundings, one sector has two bars, pub and health center nearby, and one sector in the north near a software company and government office. These sectors have good accessibility to education or other services, they might be better-off places as part of the informal settlements.

6. DISCUSSION AND LIMITATION

6.1. The Features Used to Classify Satellite Images

In this study, we first extracted 35 features based on the characteristics of local informal settlements and other literatures, which focus on block-based informal settlements classification. These features reflected characteristics in terms of three aspects: spectral, texture and structure. In Dar es Salaam, residents in formal settlements or informal settlements use similar materials as roof, therefore, solely spectral information is not sufficient to distinguish formal settlements and informal settlements. In this study, different feature sets were selected in different images, mean NDVI, mean NIR, mean G, RSF and RVF are common important features for both images. While only one texture feature ME was selected in the feature set of VHR image, no texture feature was selected for HR image.

In recent studies, texture features are very popular in informal settlements identification (Shabat & Tapamo, 2017; Ella et al., 2008; Wurm et al., 2017). Texture measure the variation or repetition of image elements based on spectral intensity to detect differences for informal settlements (Owen & Wong, 2013). Texture of informal settlements commonly exhibits higher entropy, higher contrast as well as lower homogeneity (Owen & Wong, 2013). GLCM are the most commonly used texture features, the most important parameter contain orientation, displacement and moving window size (Zhou et al., 2017). However, in our results the top 10 features did not include any GLCM measures for VHR and HR images. This study employed 100 m square grids as units extracting features. However, it might not be a good size for GLCM extraction. Mboga, Persello, Bergado, & Stein, (2017) utilized SVM classifier with GLCM measures to identify informal settlements in Dar es Salaam. They used multispectral images with 0.6 m spatial resolution that cover a small area and derived GLCM features for a window size of 165 pixels, which showed the best accuracy in SVM classifier. However, they tested their methodology on small tiles (1.2 km × 1.2 km). The window size is an important parameter in GLCM extraction, a small grid might result in the sparsity and instability of GLCM (Franklin, Wulder, & Gerylo, 2001), while larger window sizes could overlap different features and bring spatial errors (Anys, Bannari, He, & Morin, 1994).

6.2. Miscalssification Between Informal Settlements and Formal Settlements

VHR and HR images were classified based on the same training and test samples, using the top 10 image features followed by the same features selection strategy – SFS. However, the accuracies derived from these two images are quite different (82.4% in VHR image, 72.5% in HR image). For both images, the misclassification mainly occurred between formal settlements and informal settlements. As we can see in table 5.4 and table 5.5, a lot of formal settlements were predicted as informal settlements. The selected features might not strongly reflect the different characteristics of formal settlements and informal settlements. In addition, the design of the training set might be another reason causing the misclassification. More samples of informal settlements than formal settlements were used to train the SVM, this unbalanced problem made the classifier tend to be overwhelmed by informal settlements and ignore the formal settlements (Wang & Xue, 2014). Because informal settlements cover a much larger area than formal settlements in Dar es Salaam, it is hard to select a large number of samples for formal settlements as same as the number of informal settlements.

6.3. Accuracy of Land Use Data

Land use data was used to provide reference information for training and test sets preparation. The 2016 land use data was obtained by visually updating and manually digitizing the 2010 land use data. However, comparing the updated 2016 land use and ground objects in VHR image, still discrepancy can be noticed.

For instance, the planned houses which are located far away from city centre surrounded by large green space is recorded as formal settlements in land use data, but they would be classified as 'Other' because some houses are too small and green space occupies larger area in one grid (example shown in figure 6.1 (1)). The formal settlements located near the city centre have more green space surround houses, the green space would be classified as 'other' in the VHR image classification, but it was recorded as formal settlements in land use data (example shown in figure 6.1 (2)). Moreover, a large number of mixed blocks (different land use in one block) cannot be avoided. Therefore, the mismatching between land use and classification results is discussed in this section.

To detect how the VHR image classification result is related to land use data, we randomly selected 10,000 grids with labels 'Formal settlements', 'Informal settlements' as well as 'Other' which were classified in VHR image, and utilized the same grids with the same labels but derived from updated 2016 land use data to generate a cross-comparison result (table 6.1). About 53% grids were correctly classified (same as the land use type). The result shows a lot of formal settlements (50.3%) were misclassified into 'Other' which proves the findings explained in last paragraph. However, although VHR images obtained the best classification accuracy, definitely some grids were misclassified into other classes. Therefore, the result in table 6.1 estimates more than 50% probability that the land use data is related to the VHR image classification result. Only 26.8% of formal settlements from land use map match with classified formal settlements, around half of informal neighbourhoods in land use map correspond with classified informal settlements. While, other land cover with 82.6% accuracy is the best. Appendix 9 provides comparison between classification result and land use data in number of grid.

Table 6.1 Grid comparison between classification result and land use data

		Land use map		
	Categories	Formal settlements	Informal settlements	Other
Classification result (VHR image)	Formal settlements	26.8%	21.4%	9.4%
	Informal settlements	22.9%	51.3%	8.0%
	Other	50.3%	27.3%	82.6%

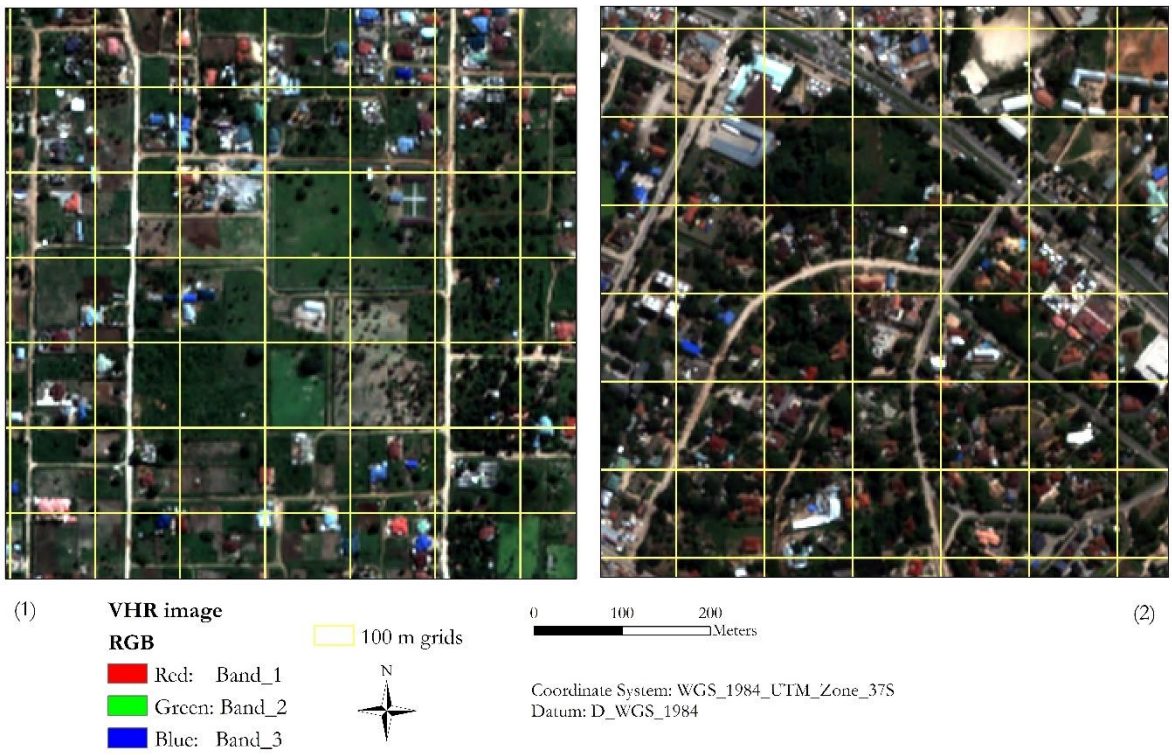


Figure 6.1 Planned houses surrounded by large green space ((1) formal settlements far away from city centre; (2) formal settlements near city centre)

6.4. The Rules for Twitter Data Filtration

To reduce the noise of geolocated tweets, this study proposed two rules for twitter data filtration. One is allowing only 20 tweets generated at the same GPS location in one day by one user. Another is to remove all the tweets located on the roads. However, these two rules are not rigorous enough and inherent biases still existed during twitter data analysis. For example, some grids in informal settlements or formal settlements recorded a large number of geolocated tweets (the highest is 4 tweets/person in informal settlements, 9 tweets/person in formal settlements), these tweets were generated by one or several users, but less than 20 tweets for one user per day on the same GPS location. However, we could not recognize these tweets that were created by the normal users or by companies for advertising purposes. Therefore, Klotz et al., (2017) proposed severe filtering rules that delete all the geographic duplicates and restrict the amount of tweets created by one twitter account per day in each building block. Therefore, they reduce a large number of tweets, from 128,000 reduce to around 16,150 (covering 18 weeks). In order to select significant geolocated tweets, (Huang & Wong, 2015) also provided a strategy that randomly selected a few twitter users with a lot of geolocated tweets and manually investigated the users' records to prove the users are not companies or organizations.

6.5. OSM Dataset Accuracy

In this study, building footprints is an important tool to disaggregate census data and estimate population number per grid cell. It is necessary to consider the accuracy of building footprints as it would affect the accuracy of population number in each grid. This section qualitatively assess building footprints data in OSM based on three criteria: completeness, position accuracy and shape accuracy (Fan, Zipf, Fu, & Neis, 2014).

Completeness measure the lack or excess of data (Fan et al., 2014). For example, the building was not recorded in OSM but it can be found in VHR image or the building was recorded in OSM but it is not a real building in VHR image. Position assess how the coordinate value of building footprints relates to the real building in VHR image (Fan et al., 2014). Shape accuracy measures the similarity of building shape (Fan et al., 2014) between building footprints in OSM and real buildings in VHR image. Appendix 6 shows three different errors in OSM corresponding to these criteria. In appendix 6 (1), building footprints in OSM mostly cover the buildings visible in the VHR image, but several buildings were omitted and not recorded in OSM. Appendix 6 (2) shows the constant shift of building footprints in OSM (in the same projection), the position does not cover the position of real buildings. It is clear to see in appendix 6 (3), there are several houses located in the middle, but the shape of building footprint is too small to match with the real building.

Building footprints data was digitized based on Unmanned Aerial Vehicle (UAV) with resolution of 0.37 m (Kaaya, 2017). These errors might be attributed to the equipment during data collection or other reasons when digitizing. However, manually modifying the data would take a long time and need a lot of labour force. So, we did not correct these problems. However, it is a limitation that the errors negatively affect the accuracy of estimated population number in some grids.

6.6. Census Data Disaggregation

In this work, we used the area of residential building to estimated population number in each grid. Therefore, if the grid contains more residential areas means more people live there. However, this simple and understandable approach has a bias, which could be attributed to the different population density in formal settlements and informal settlements. In Dar es Salaam, almost 70% of people live in informal settlements (World Bank 2002), these residents crowd in downtown areas or large unplanned areas. High population density is common in informal settlements. If two houses with the same area are located in formal settlements and informal settlements, respectively, two houses have the same population number calculated by the method in this work. However, the fact is that formal settlements is not as crowded as informal settlements, residents live in formal settlements should have more living space than people settled in informal settlements. Therefore, the house located in formal settlements might has less residents than the house in informal settlements, although they have the same house size. But this study did not consider this limitation in the step of population disaggregation.

This work utilized regular girds as unit to estimate population number per gird based on the residential area in each grid. However, the intersection between grid boundary and some building footprints cannot be avoided (blue buildings in figure 6.2 (2)). In this way, the area of intersecting building footprints outside a gird would also be added into the total residential area in this grid. In this case, the residential area in each grid would be more than reality. Therefore, the estimated population number per grid would be higher because more residential area calculated in grids. While, another spatial unit which is building blocks (shown in figure 6.2 (1)) derived based on morphological homogeneity and road network, all the building footprints were completely surrounded by building blocks. It could be an alternative to deal with this defect in future research.

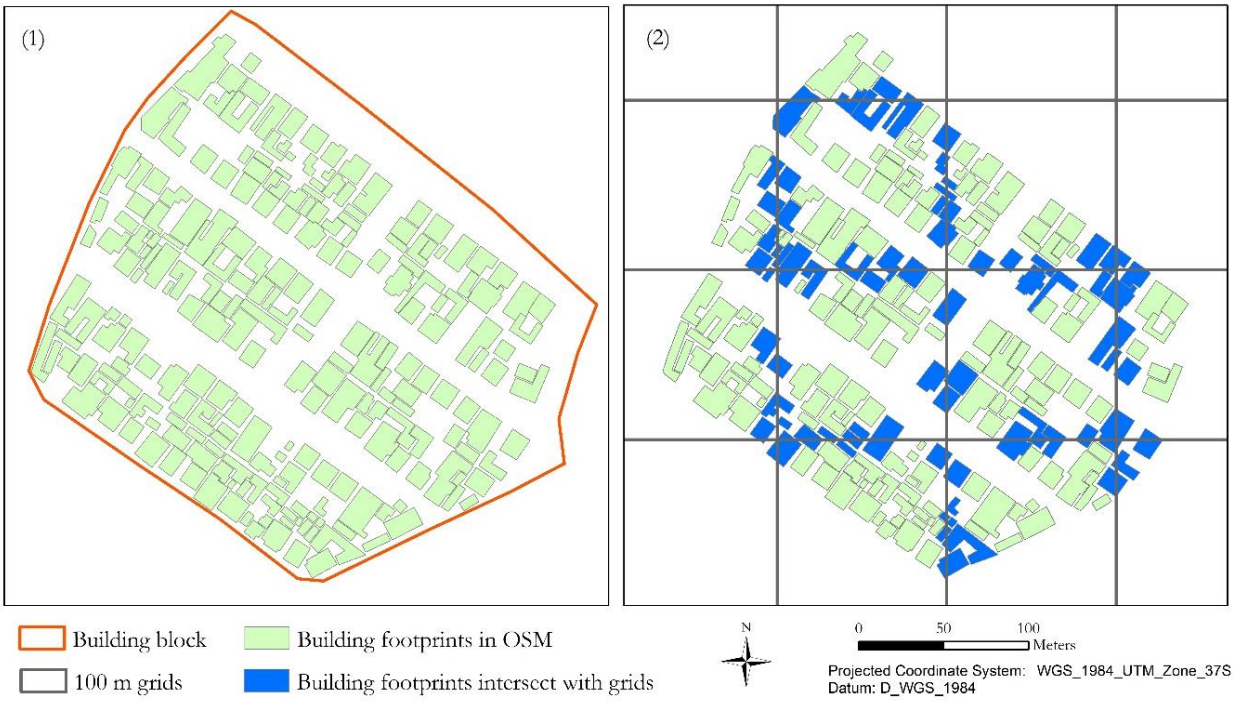


Figure 6.2 Building footprints in building block and regular grids

6.7. Comparison Between Informal Settlements Classified from VHR Image and Digital Deserts Derived from Twitter Data

The comparison between informal settlements and digital deserts conveys that digital deserts mainly prevail in informal settlements but also appear in some formal settlements. Similarly, some informal settlements are not digital deserts, it is possible to find some twitter users who live in informal settlements. However, formal settlements, informal settlements and other were identified based on machine learning algorithm, the VHR image classification result presents that about 18% of grids among the entire study area were misclassified.

Previous studies(Wurm & Taubenböck, 2018); Klotz et al., 2017) recognized unplanned areas by visual interpretation of very high resolution image. According to the morphological characteristics such as built-up density, building height and so on, unplanned areas could be recognized on building block level. This approach might be more accurate than machine learning methods, however, it is an expensive approach because researchers have to spend a lot of time to do fieldwork, and need sufficient local knowledge.

6.8. Transferability Assessment

As we know, some informal areas are better-off in Dar es Salaam, many medium income residents also live in informal settlements. Based on the result in section 5.4, only 2.4% (202 grids are real informal settlements and recognized as digital neighbourhoods, 8463 grids are total number of classified informal settlements) of classified informal settlements are digital neighbourhoods, while 97.6% of classified informal settlements are regarded as digital deserts. Therefore, the combination of EO data and twitter data allows to analyse the socio-economic disparities in informal settlements, the methodology used in Mumbai is transferable to Dar es Salaam.

7. CONCLUSION AND RECOMMANDATIONS

7.1. Conclusion

The main goal of this study is to analyze the combination of ML algorithm using VHR and HR images and social media data for informal settlements identification in Dar es Salaam. To accomplish this main objective, this study first utilized a ML approach to identity informal settlements from VHR and HR images, then used twitter data to detect areas with less geolocated tweets, which are called 'digital deserts'. Finally, the spatial relationship between informal settlements and digital deserts was compared. Four specific objectives were achieved by answering nine research questions. The following provides concise conclusion of each specific objective.

The first objective was to single out the most significant image features to classify informal settlements comparing VHR image and HR image. After reviewing several literatures, the physical characteristics of informal settlements were presented using the conceptualization of GSO. Rectangular shape with size from 50 m^2 to 150 m^2, single storey with corrugated iron sheet roof and narrow road network could describe the characteristics on object level. At settlements level, informal settlements could be characterized as compact arrangement with high building density, the shape of the neighbourhoods depend on the surrounded roads. Move to environ level, informal settlements commonly located at haphazard areas. Based on these characteristics, 35 image features (contain three aspects: spectral features, texture features and structural features) of VHR image and HR image were extracted in 100 m grids from FETEX 2. Finally, SFS was employed to select the best 10 features. The results show (table 5.3) spectral features were more important for classifying both two images. ME was the best texture feature for VHR image classification and HR image did not have texture features in the best feature set. RSF and RVF were the common structural features for both images.

The second objective was to employ SVM to identity informal settlements from VHR and HR image. After comparing several grid sizes, 100 m square grids had the best performance to classify both images. We used the same training set and test set, the respective best 10 features of VHR and HR image as input in SVM classifier. The results showed VHR image obtained the best classification accuracy of 82.4% while HR image got accuracy of 72.5%. VHR image provide more details than HR image (resolution is 2 m and 10 m respectively), it is much easier to recognize objects and extract texture and structural features from VHR image. These features are helpful to distinguish informal settlements and formal settlements. While, HR image is coarse with less information which is hard to get better classification results. In conclusion, the VHR image outperformed HR image in informal settlements identification.

The third objective was to analyse the utility of twitter data to depict the spatial distribution of digital deserts. The geolocated tweets were aggregated in 100 square grids to get the number of geolocated tweets in each block. Then census data were disaggregated from subswards to 100 m square grids according to the percentage of residential area in each grid. Finally, we applied Local Moran's I to detect the spatial distribution of digital deserts and digital neighbourhoods. The results (figure 5.12) showed digital deserts distributed in most of the study area. Manzese, Tandale, Makumbusho, Magomeni, Ndugumbi, Makurmla, Mabibo, Mburahati, Mzimuni, Kigogo, Ilala, Buguruni, Tabata as well as Segerea detected more LL clusters. While some wards distant from city centre were also recognized as digital deserts. Some wards (Msasani, Kinondoni, Upanga Magharibi, Upanga Mashariki, Kariakoo, Kisutu) near city centre were recognized as digital neighbourhoods. Several wards (Mikocheni, Mwananyamala, Kijitonyama, Sinza, Ubungo, Kawe, Jangwani and Kivukoni) were mixed by digital deserts and digital neighbourhoods.

The last objective was to explore the relationship between informal settlements extracted from VHR image and digital deserts derived by twitter data. About 56.5% of regional lows (LL) contributing to classified informal settlements and 19.7% of regional highs matching (HH) with classified informal settlements. Moreover, LL clusters also appeared in formal settlements (23.6%). Overall, although the informal settlements are very diverse in Dar es Salaam, the results in this study have shown the methodology used in Mumbai to map informal settlements by twitter data is transferable to Dar es Salaam. The main difference is that this study utilized regular grids instead of building blocks as spatial unit to detect digital deserts and digital neighbourhoods and ML approach to identify informal settlements.

7.2. Recommandations

For future study, this work has some scope for further improvement and also possible to try a different approach to identity informal settlements from satellite image. Several recommendations are provided:

1. Exploring more image features, which could contribute to VHR image classification accuracy.

The result in this study illustrated spectral features performed better than texture features and structural features. Therefore, other spectral features such as Soil Adjusted Vegetation Indices (SAVI), Infrared Percentage Vegetation Index (IPVI), Ratio Difference Vegetation Index (RDVI), Normalized Difference Infrared Index (NDII) and so on are worth to explore in future research.

2. Using ML approach to identify informal settlements on pixel level or object level.

Further identification of informal settlements could be achieved by pixel-based classification. In this way, the samples could be selected more precise and obtain better classification results. Because block-based classification has a lot of mixed blocks which would cause misclassification. Another suggestion is to use algorithms for object recognition, it can also improve the classification results.

3. Using rigorous rules for twitter data filtration

In future study, the rigorous rules could be used to reduce some inherent biases of twitter data. For example, deleting the geographic duplicates (Klotz et al., 2017) to avoid the geolocated tweets are generated by twitter bots for advertising purposes. And limit the amount of geolocated tweets per gird sent by the same user per day (Klotz et al., 2017), because some users are twitter fans and they would send a lot of tweets per day in the same area.

4. Disaggregating census data using dasymetric mapping approach.

Dasymetric mapping is a technique to refine the spatial distribution of census data by considering additional parameters like land use/cover (Weichselbaum & Papathoma, 2005) or structural characteristics of residential buildings (Klotz et al., 2017) that influence the spatial distribution of population number. Dasymetric mapping is a good solution for the problem mentioned in section 6.6.

LIST OF REFERENCES

Amami, R., Ayed, D. Ben, & Ellouze, N. (2013). Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition. *International Journal of Digital Content Technology and Its Applications*, *7*(9), 418–424. https://doi.org/10.4156/jdcta.vol7.issue9.50

Anselin, L. (1995). Local Indicators of Spatial Association -LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., & Williams, S. (2016). Digital neighborhoods. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, *9*(4), 305–328. https://doi.org/10.1080/17549175.2015.1080752

Anys, H., Bannari, A., He, D. C., & Morin, D. (1994). Texture analysis for the mapping of urban areas using airborne MEIS-II images. In *Proceedings - International Airborne Remote Sensing Conference and Exhibition* (p. III-231-III-246). Strasbourg, France. Retrieved from https://eurekamag.com/research/018/437/018437508.php

ASTRIUM. (2012). Pléiades Imagery User Guide. Retrieved from http://blackbridge.com/geomatics/upload/airbus/Pleiades User Guide.pdf

Balaguer, A., Ruiz, L. A., Hermosilla, T., & Recio, J. A. (2010). Definition of a comprehensive set of texture semivariogram features and their evaluation for object-oriented image classification. *Computers & Geosciences*, *36*(2), 231–240. https://doi.org/10.1016/J.CAGEO.2009.05.003

Camps-valls, G., Member, S., Mooij, J., & Schölkopf, B. (2010). Kernel Dependence Measures. *IEEE Geoscience and Remote Sensing Letters*, *7*(3), 587–591.

Cervone, G., Schnebele, E., Waters, N., Moccaldi, M., & Sicignano, R. (2017). Using Social Media and Satellite Data for Damage Assessment in Urban Areas During Emergencies. In P. Thakuriah, N. Tilahun, & M. Zellner (Eds.), *Seeing Cities Through Big Data* (pp. 443–457). Springer, Cham. https://doi.org/10.1007/978-3-319-40902-3

Chehata, N., Orny, C., Boukir, S., Guyon, D., & Wigneron, J. P. (2014). Object-based change detection in wind storm-damaged forest using high-resolution multispectral images. *International Journal of Remote Sensing*, *35*(13), 4758–4777. https://doi.org/10.1080/01431161.2014.930199

Chen, S., Ouyang, Y. C., & Chang, C. (2012). Weighted radial basis function kernels-based support vector machines for multispectral image classification. *Geoscience and Remote Sensing Symposium*, 4339–4342. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6351707

Chen, X., Li, H., & Gu, Y. (2014). Multiview Feature Selection for Very High Resolution Remote Sensing Images. *2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, 539–543. https://doi.org/10.1109/IMCCC.2014.116

Duque, J. C., Patiño, J. E., & Betancourt, A. (2016). *Exploring the Potential of Machine Learning for Automatic Slum Identification From VHR Imagery*. Retrieved from http://www.scioteca.caf.com/bitstream/handle/123456789/975/Duque%2C Patino %26 Betancourt %282016%29.pdf?sequence=1&isAllowed=y

Duque, J. C., Patino, J. E., Ruiz, L. A., & Pardo-Pascual, J. E. (2015). Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data. *Landscape and Urban Planning*, *135*, 11–21. https://doi.org/10.1016/j.landurbplan.2014.11.009

Ella, L. P. A., Bergh, F. van den, Wyk, B. J. Van, & Wyk, M. A. van. (2008). A Comparison of Texture Feature Algorithms for Urban Settlement Classification. In *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International* (p. III-1308-III-1311). Boston, MA, USA: IEEE. https://doi.org/10.1109/IGARSS.2008.4779599

Engstrom, R., Sandborn, A., Yu, Q., Burgdorfer, J., Stow, D., Weeks, J., & Graesser, J. (2015). Mapping Slums Using Spatial Features in Accra, Ghana. In *Urban Remote Sensing Event (JURSE), 2015 Joint*. Lausanne, Switzerland: IEEE. https://doi.org/10.1109/JURSE.2015.7120494

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, *28*(4), 700–719. https://doi.org/10.1080/13658816.2013.867495

Fauvel, M. (2007). *Spectral and Spatial Methods for the Classification of Urban Remote Sensing Data*. University of Iceland.

Fohringer, J., Dransch, D., Kreibich, H., & Schröter, K. (2015). Social media as an information source for rapid flood inundation mapping. *Natural Hazards and Earth System Sciences*, *15*(12), 2725–2738. https://doi.org/10.5194/nhess-15-2725-2015

Franklin, S. E., Wulder, M. A., & Gerylo, G. R. (2001). Texture analysis of IKONOS panchromatic data

for Douglas-fir forest age class separability in British Columbia. *International Journal of Remote Sensing*, *22*(13), 2627–2632. https://doi.org/10.1080/01431160120769
Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing Urban Landscapes using Geolocated Tweets. In *2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing* (pp. 239–248). Amsterdam, Netherlands: IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.19
Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *116*, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008
Gómez, D., & Montero, J. (2011). Determining the accuracy in image supervised classification problems. In A. Press (Ed.), *Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology* (pp. 342–349). Aix-les-Bains, France. https://doi.org/doi:10.2991/eusflat.2011.103
Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J., & Bright, E. (2012). Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(4), 1164–1176. https://doi.org/10.1109/JSTARS.2012.2190383
Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. In S. Jain, H. U. Simon, & E. Tomita (Eds.), *Algorithmic Learning Theory* (pp. 63–77). Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/11564089_7
Gruebner, O., Sachs, J., Nockert, A., Frings, M., Khan, M. M. H., Lakes, T., … Hostert, P. (2014). Mapping the Slums of Dhaka from 2006 to 2010. *Dataset Papers in Science*, *2014*, 1–7. https://doi.org/10.1155/2014/172182
Hofmann, P. (2001). Detecting Informal Settlements From Ikonos Image Data Using Methods of Object Oriented Image Analysis - An Example Form Cape Town (South Africa). In J. Carsten (Ed.), *Remote Sensing of Urban Areas/ Fernerkundung in urbanen Räumen* (Vol. 35, pp. 107–118). Regensburg.
Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, *54*, 240–254. https://doi.org/10.1016/j.compenvurbsys.2015.09.001
Huang, Q., & Wong, D. W. S. (2015). Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. *Annals of the Association of American Geographers ISSN:*, *105*(6), 1179–1197. https://doi.org/https://doi.org/10.1080/00045608.2015.1081120
Hwang, S., Evans, C., & Hanke, T. (2017). Using Social Media and Satellite Data for Damage Assessment in Urban Areas During Emergencies, 427–439. https://doi.org/10.1007/978-3-319-40902-3
Iannelli, G. C., Lisini, G., Dell'Acqua, F., Feitosa, R. Q., da Costa, G. A. O. P., & Gamba, P. (2014). Urban area extent extraction in spaceborne HR and VHR data using multi-resolution features. *Sensors (Switzerland)*, *14*(10), 18337–18352. https://doi.org/10.3390/s141018337
Ishtiyaq, M., & Kumar, S. (2011). Typology of Informal Settlements and Distribution of Slums in the NCT, Delhi. *Space and Society, Hiroshima University*, *1*, 37–46. Retrieved from http://home.hiroshima-u.ac.jp/hindas/PDF/2010/Ishtiyaq_and_Kumar(2011).pdf
Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008). On the Relationship Between Feature Selection and Classification Accuracy. In Saeys (Ed.), *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008* (Vol. 4, pp. 90–105). Antwerp, Belgium.
Kaaya, D. A. (2017). *Explaining Variations in Informal Neighborhoods' Consolidation Levels in Dar es Salaam, Tanzania*. University of Twente.
Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, *11*(5), 352–359. https://doi.org/10.1016/j.jag.2009.06.002
Klotz, M., Wurm, M., Zhu, X., & Taubenböck, H. (2017). Digital deserts on the ground and from space An experimental spatial analysis combining social network and earth observation data in megacity Mumbai. In *Joint Urban Remote Sensing Event JURSE 2017*. Dubai. https://doi.org/10.1109/JURSE.2017.7924562
Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). An ontology of slums for image-based classification. *Computers, Environment and Urban Systems*, *36*(2), 154–163. https://doi.org/10.1016/j.compenvurbsys.2011.11.001
Kotsiantis, S. B. (2011). Feature selection for machine learning classification problems: a recent overview. *Artificial Intelligence Review*, *42*(1), 157–176. https://doi.org/10.1007/s10462-011-9230-1

Kuffer, M. (2003). Monitoring the Dynamics of Informal Settlements in Dar Es Salaam by Remote Sensing: Exploring the Use of Spot, Ers and Small Format Aerial Photography. In M. Schrenk (Ed.), *Proceedings of CORP 2003* (pp. 473–483). Vienna, Austria. Retrieved from http://www.corp.at/archive/CORP2003_Kuffer.pdf

Kuffer, M., & Barros, J. (2011). Urban morphology of unplanned settlements: The use of spatial metrics in VHR remotely sensed images. *Procedia Environmental Sciences*, *7*, 152–157. https://doi.org/10.1016/j.proenv.2011.07.027

Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from space-15 years of slum mapping using remote sensing. *Remote Sensing*, *8*(6), 1–29. https://doi.org/10.3390/rs8060455

Kuffer, M., Pfeffer, K., Sliuzas, R., & Baud, I. (2016). Extraction of Slum Areas From VHR Imagery Using GLCM Variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(5), 1830–1840. https://doi.org/10.1109/JSTARS.2016.2538563

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering (IJCSE)*, *3*(5), 1787–1797. Retrieved from http://journals.indexcopernicus.com/abstract.php?icid=945099

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Lin, S.-W., Lee, Z.-J., Chen, S.-C., & Tseng, T.-Y. (2008). Parameter determination of support vector machine and feature selection using simulated annealing approach. *Applied Soft Computing*, *8*(4), 1505–1512. https://doi.org/10.1016/j.asoc.2007.10.012

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Wang, H. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, *31*(8), 1675–1696. https://doi.org/10.1080/13658816.2017.1324976

Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. In *IECON Proceedings (Industrial Electronics Conference)* (pp. 2845–2850). Glendale, AZ, USA: IEEE. https://doi.org/10.1109/IECON.2010.5675075

Martin. (2015). Using OSM building footprints to disaggregate OGD census data | gicycle. Retrieved February 19, 2018, from https://gicycle.wordpress.com/2015/08/17/using-osm-building-footprints-to-disaggregate-ogd-census-data/

Mboga, N., Persello, C., Bergado, J. R., & Stein, A. (2017). Detection of Informal Settlements from VHR Satellite Images using Convolutional Neural Networks. *Remote Sensing*, *9*(1106). https://doi.org/10.3390/rs9111106

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Murray, H., Lucieer, A., & Williams, R. (2010). Texture-based classification of sub-Antarctic vegetation communities on Heard Island. *International Journal of Applied Earth Observations and Geoinformation*, *12*(3), 138–149. https://doi.org/10.1016/j.jag.2010.01.006

Naorem, V., Kuffer, M., Verplanke, J., & Kohli, D. (2016). Robustness of rule sets using VHR imagery to detect informal settlements - a case of Mumbai, India. In N. Kerle, M. Gerke, & S. Lefevre (Eds.), *Proceedings of GEOBIA 2016 : Solutions and synergies*. Enschede: University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC). https://doi.org/10.3990/2.416

Owen, K. K., & Wong, D. W. (2013). An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics. *Applied Geography*, *38*(1), 107–118. https://doi.org/10.1016/j.apgeog.2012.11.016

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(5), 2297–2307. https://doi.org/10.1109/TGRS.2009.2039484

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by anisotropic rotation-invariant textural measure. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *1*(3), 180–192. https://doi.org/10.1109/JSTARS.2008.2002869

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., … Zanchetta, L. (2013). A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(5), 2102–2131. https://doi.org/10.1109/JSTARS.2013.2271445

Rasmussen, M. I. (2013). The power of Informal Settlements . The Case of Dar Es Salaam , Tanzania. In *Planum, The Journal of Urbanism* (Vol. 1, pp. 1–11). Milan.
Richards, J. ., & Jia, X. (2006). *Remote Sensing Digital Image Analysis* (5th ed.). Springer Heidelberg New York Dordrecht London. https://doi.org/10.1007/3-540-29711-1
Ruiz, L. A., Recio, J. A., Fernández-Sarría, A., & Hermosilla, T. (2011). A feature extraction software tool for agricultural object-based image analysis. *Computers and Electronics in Agriculture*, *76*(2), 284–296. https://doi.org/10.1016/j.compag.2011.02.007
Sandborn, A., & Engstrom, R. N. (2016). Determining the Relationship between Census Data and Spatial Features Derived from High-Resolution Imagery in Accra, Ghana. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(5), 1970–1977. https://doi.org/10.1109/JSTARS.2016.2519843
Schaffernicht, E., Möller, C., Debes, K., & Gross, H.-M. (2009). Forward feature selection using Residual Mutual Information. In *ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning* (pp. 583–588). Bruges, Belgium. Retrieved from http://www.citeulike.org/user/jjrodriguez/article/6082471
Shabat, A. M., & Tapamo, J.-R. (2017). A comparative study of the use of local directional pattern for texture-based informal settlement classification. *Journal of Applied Research and Technology*, *15*(3), 250–258. https://doi.org/10.1016/J.JART.2016.12.009
Sirueri, F. O. (2015). *Comparing Spatial Patterns of Informal Settlements Between Nairobi and Dar Es Salaam*. University of Twente.
Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In A. Sattar & B. Kang (Eds.), *Advances in Artificial Intelligence* (Vol. 4304, pp. 1015–1021). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11941439_114
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
Sta, U., & Jain, L. C. (2014). *Feature Selection for Data and Pattern Recognition*. (S. Urszula & L. C. Jain, Eds.) (1st ed., Vol. 584). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-45620-0
Taubenböck, H., & Kraff, N. J. (2014). The physical face of slums: A structural comparison of slums in Mumbai, India, based on remotely sensed data. *Journal of Housing and the Built Environment*, *29*(1), 15–38. https://doi.org/10.1007/s10901-013-9333-x
Taubenbock, H., Wurm, M., Setiadi, N., Gebert, N., Roth, A., Strunz, G., … Dech, S. (2009). Integrating remote sensing and social science. In *2009 Joint Urban Remote Sensing Event* (pp. 1–7). Shanghai, China: IEEE. https://doi.org/10.1109/URS.2009.5137506
UN-Habitat. (2003). *The Challenge of Slums-Global Report on Human Settlements*. UN-Habitat. Retrieved from https://unhabitat.org/books/the-challenge-of-slums-global-report-on-human-settlements-2003/
UN-Habitat. (2010). *Informal Settlements and Finance in Dar es Salaam, Tanzania*. (X. Q. Zhang, Ed.). Nairobi: UN-HABITAT.
UN-Habitat. (2015). *Habitat III Issue Papers 22 – Informal Settlements. United Nations Conference on Housing and Sustainable Urban Development* (not edited, Vol. 2015). New York: UN-Habitat. https://doi.org/http://dx.doi.org/10.3402/gha.v5i0.19065
UN-Habitat. (2016a). *SLUM ALMANAC 2015/2016 Tracking Improvement in the Lives of Slum Dwellers*. Nairobi.
UN-Habitat. (2016b). *World City Report*. Retrieved from http://wcr.unhabitat.org/wp-content/uploads/2017/02/WCR-2016_-Abridged-version-1.pdf
United Nations. (2015). The Millennium Development Goals Report. *United Nations*, 72. https://doi.org/978-92-1-101320-7
Veljanovski, T., Kanjir, U., Pehani, P., Oštir, K., & Kovačič, P. (2012). Object-Based Image Analysis of VHR Satellite Imagery for Population Estimation in Informal Settlement Kibera-Nairobi, Kenya. In B. Escalante-Ramirez (Ed.), *Remote Sensing – Applications* (pp. 407–434). https://doi.org/10.5772/37869
Wang, M., Wan, Y., Ye, Z., & Lai, X. (2017). Remote sensing image classification based on the optimal support vector machine and modified binary coded ant colony optimization algorithm. *Information Sciences*, *402*, 50–68. https://doi.org/10.1016/j.ins.2017.03.027
Wang, Z., & Xue, X. (2014). Multi-Class Support Vector Machine. In Y. Ma & G. Guo (Eds.), *Support Vector Machines Applications* (pp. 23–49). Springer, Cham. https://doi.org/10.1007/978-3-319-02300-

Weichselbaum, J., & Papathoma, M. (2005). Sharpening census information in GIS to meet real-world conditions – the case for Earth Observation. *Transactions on Ecology and the Environment, Sustainable Development and Planning II*, *84*, 143–152. https://doi.org/10.2495/SPD050141

Wurm, M., & Taubenböck, H. (2018). Detecting social groups from space – Assessment of remote sensing-based mapped morphological slums using income data. *Remote Sensing Letters*, *9*(1), 41–50. https://doi.org/10.1080/2150704X.2017.1384586

Wurm, M., Weigand, M., Schmitt, A., Geiß, C., & Taubenböck, H. (2017). Exploitation of textural and morphological image features in Sentinel-2A data for slum mapping. In *Urban Remote Sensing Event (JURSE), 2017 Joint* (pp. 17–20). Dubai, United Arab Emirates: IEEE. https://doi.org/10.1109/JURSE.2017.7924586

Zhou, J., Yan Guo, R., Sun, M., Di, T. T., Wang, S., Zhai, J., & Zhao, Z. (2017). *The Effects of GLCM parameters on LAI estimation using texture values from Quickbird Satellite Imagery*. https://doi.org/10.1038/s41598-017-07951-w

Zuo, R., & Carranza, E. J. M. (2011). Support vector machine: A tool for mapping mineral prospectivity. *Computers and Geosciences*, *37*(12), 1967–1975. https://doi.org/10.1016/j.cageo.2010.09.014

Appendix 1

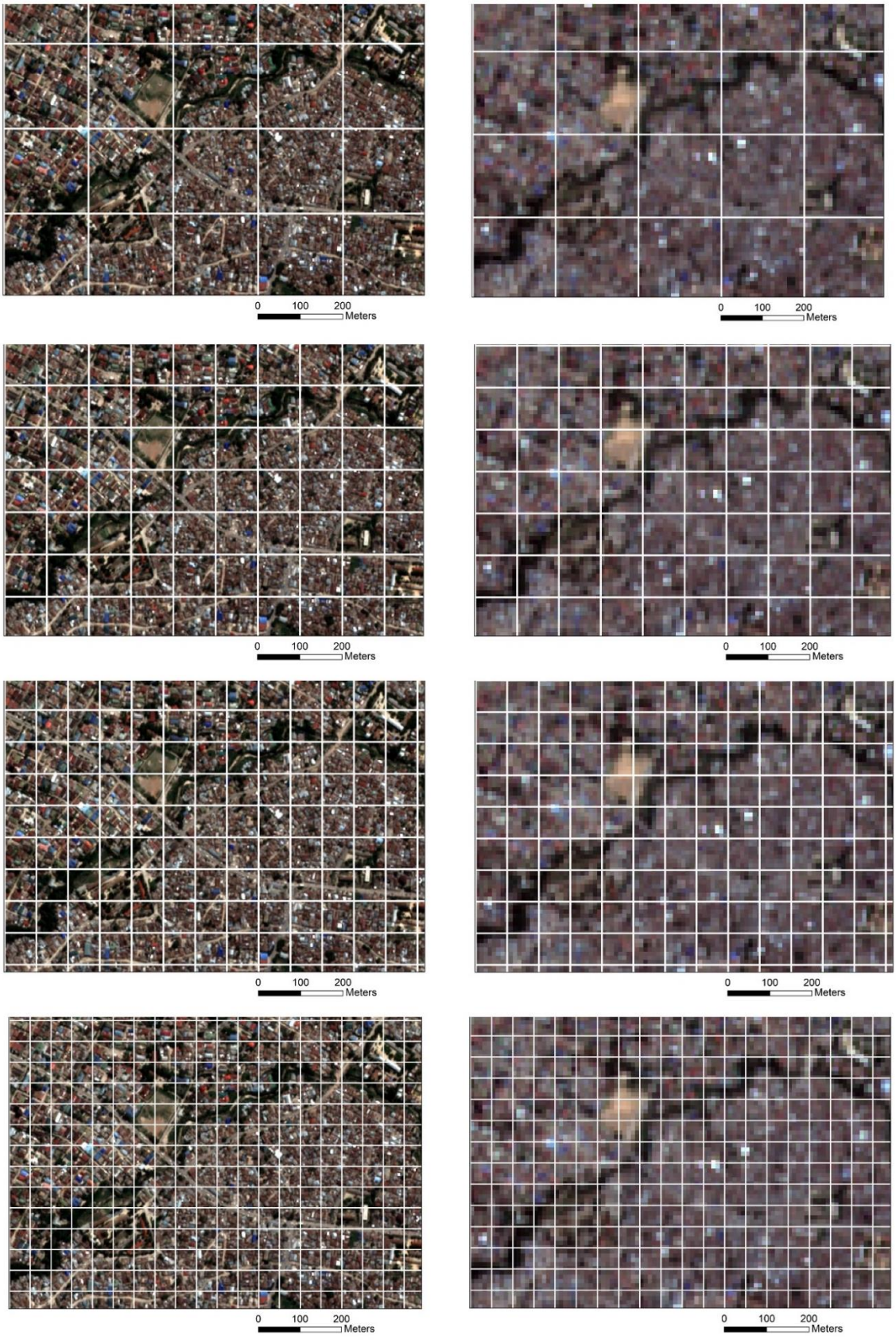
Sentinel-2A bands information.

Band name	Resolution (m)	Central wavelength (nm)	Band width (nm)	Purpose
B01	60	443	20	Aerosol detection
B02	10	490	65	Blue
B03	10	560	35	Green
B04	10	665	30	Red
B05	20	705	15	Vegetation classification
B06	20	740	15	Vegetation classification
B07	20	783	20	Vegetation classification
B08	10	842	115	Near infrared
B08A	20	865	20	Vegetation classification
B09	60	945	20	Water vapour
B10	60	1375	30	Cirrus
B11	20	1610	90	Snow / ice / cloud discrimination
B12	20	2190	180	Snow / ice / cloud discrimination

(Source: http://www.gdal.org/frmt_sentinel2.html)

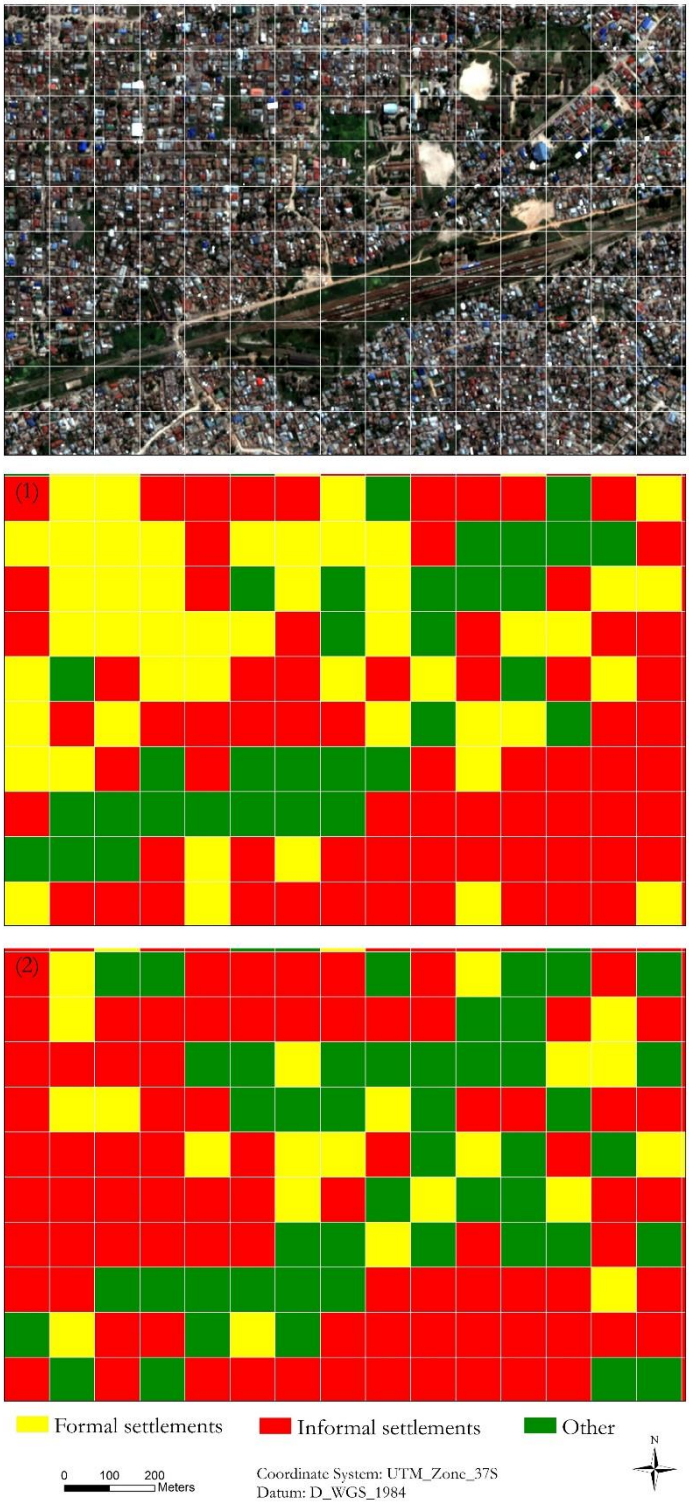
Appendix 2

Regular grids over images (right side: Pleiades image, left side: Sentinel image; grid sizes from top to bottom: 200 m, 100 m, 75 m, 50 m).



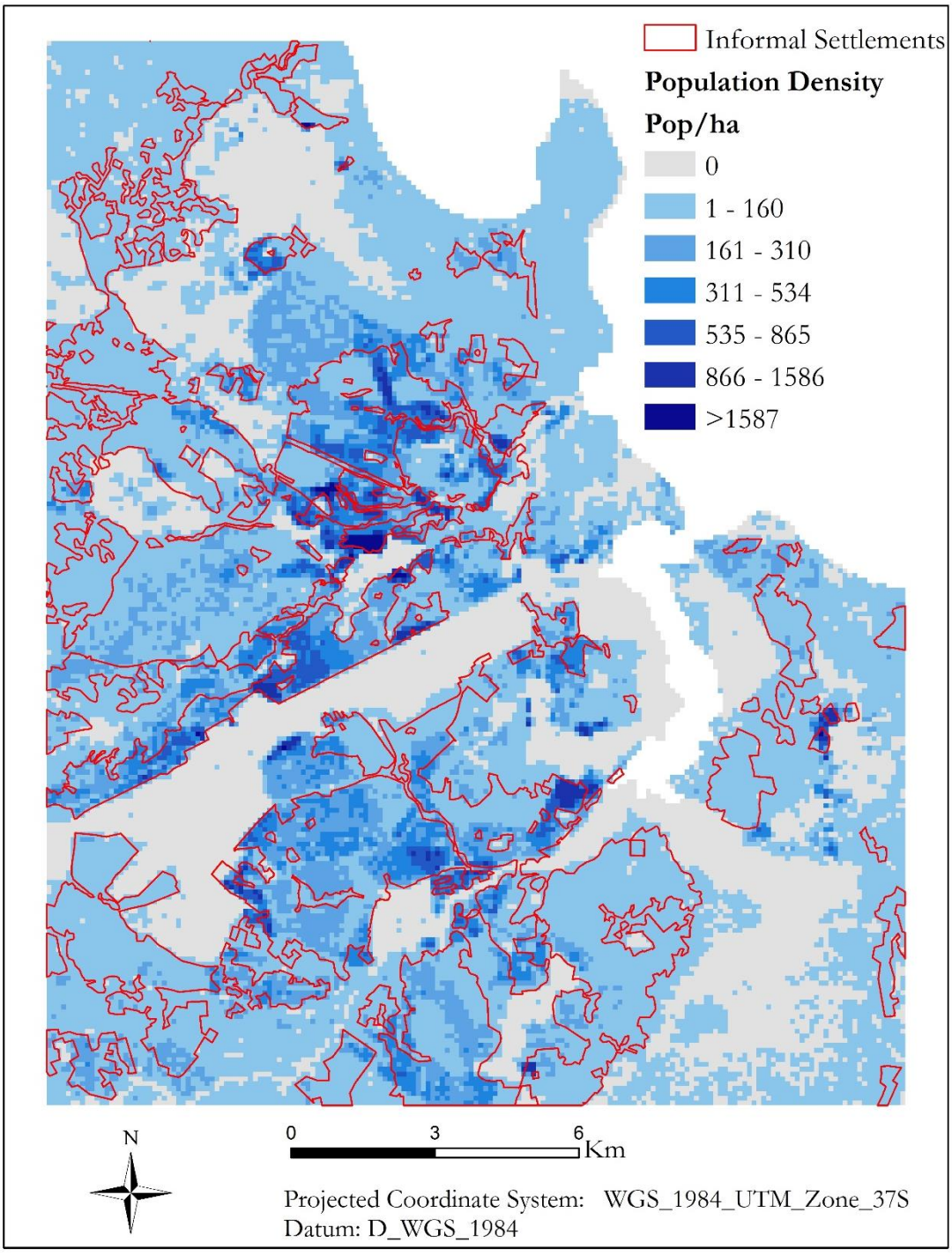
Appendix 3

Classification results for VHR (1) and HR (2) images of selected sector.



Appendix 4

Population density map in 2015 (Number of people per hectare).



Appendix 5

Quantitative cross-comparison matrix.

	Digital neighbourhoods		Digital deserts					
	HH	**HL**	**LH**	**LL**	**NT**	**NS**	**NP**	**Total**
Formal settlements	760	184	309	639	3254	502	347	5995
Informal settlements	434	288	232	1531	5265	574	139	8463
Other	1006	324	212	538	6798	699	7411	16988
Total	2200	796	753	2708	15317	1775	7897	31446

Appendix 6 Errors in OSM

(1) The lack of building footprints in OSM.

VHR image
RGB
Red: band 1
Green: band 2
Blue: band 3

Building footprints in OSM

0 50 100
Meters

N

Projected Coordinate System: WGS_1984_UTM_Zone_37S
Datum: D_WGS_1984

(2) The shifting of building footprints in OSM.

VHR image
RGB
Red: band 1
Green: band 2
Blue: band 3

Building footprints in OSM

0 40 80
Meters

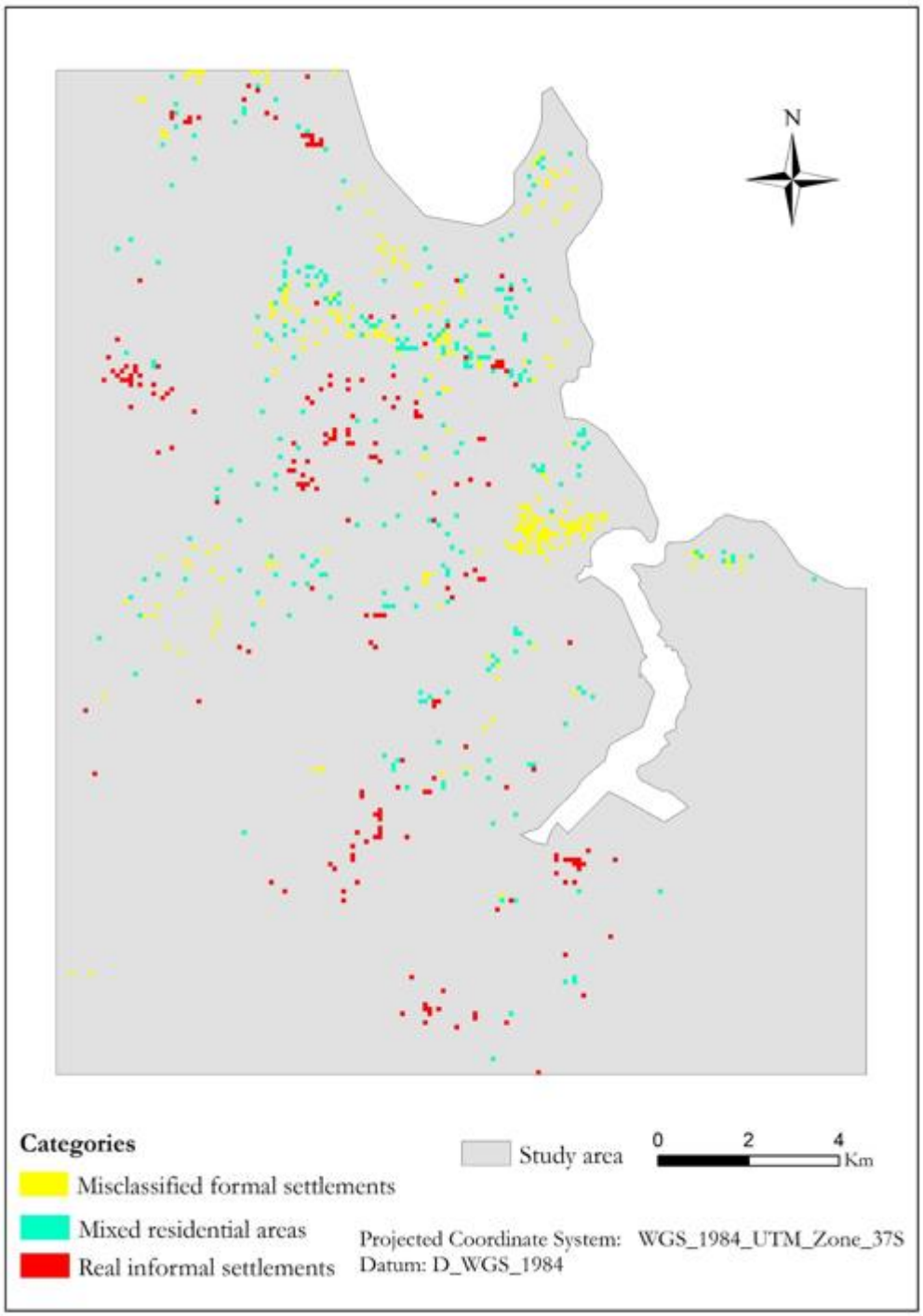
N

Projected Coordinate System: WGS_1984_UTM_Zone_37S
Datum: D_WGS_1984

(3) The shape error of building footprints in OSM.

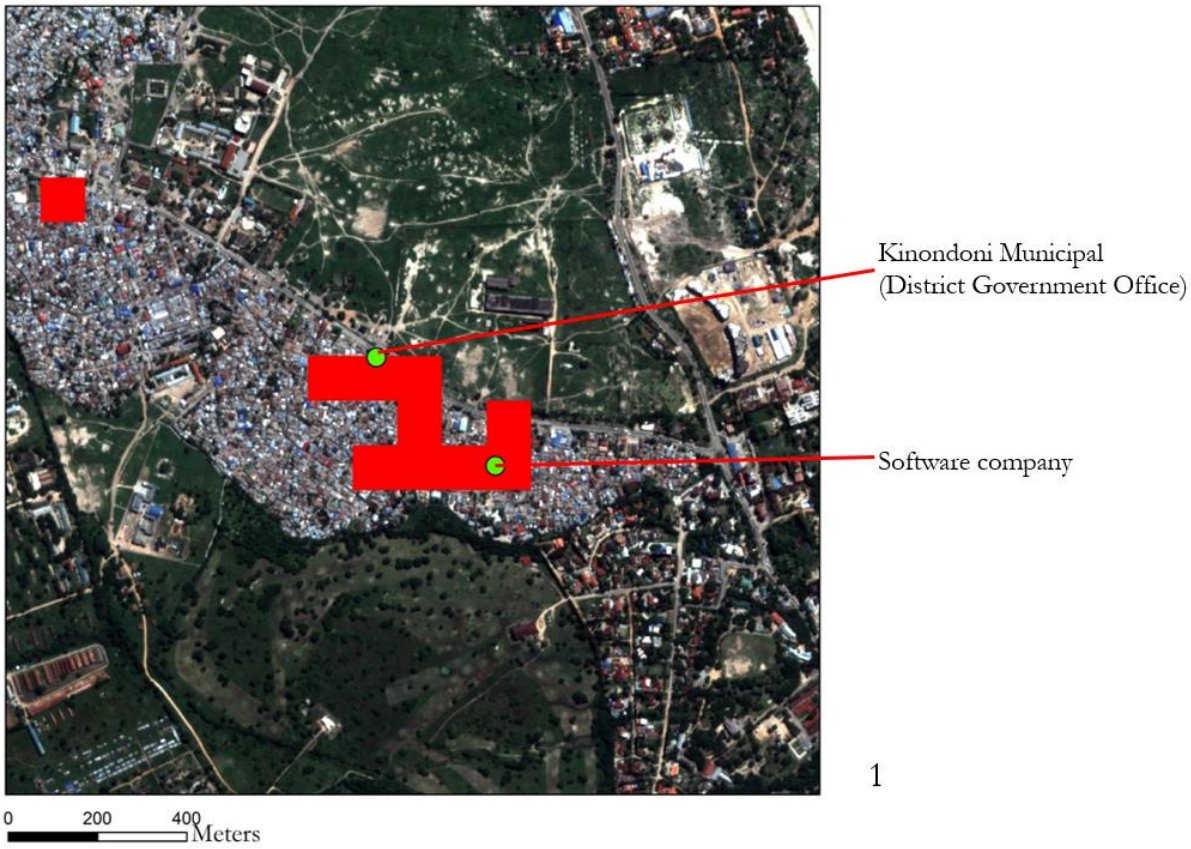
Appendix 7

Grids were classified as informal settlements and recognized as digital neighbourhoods.

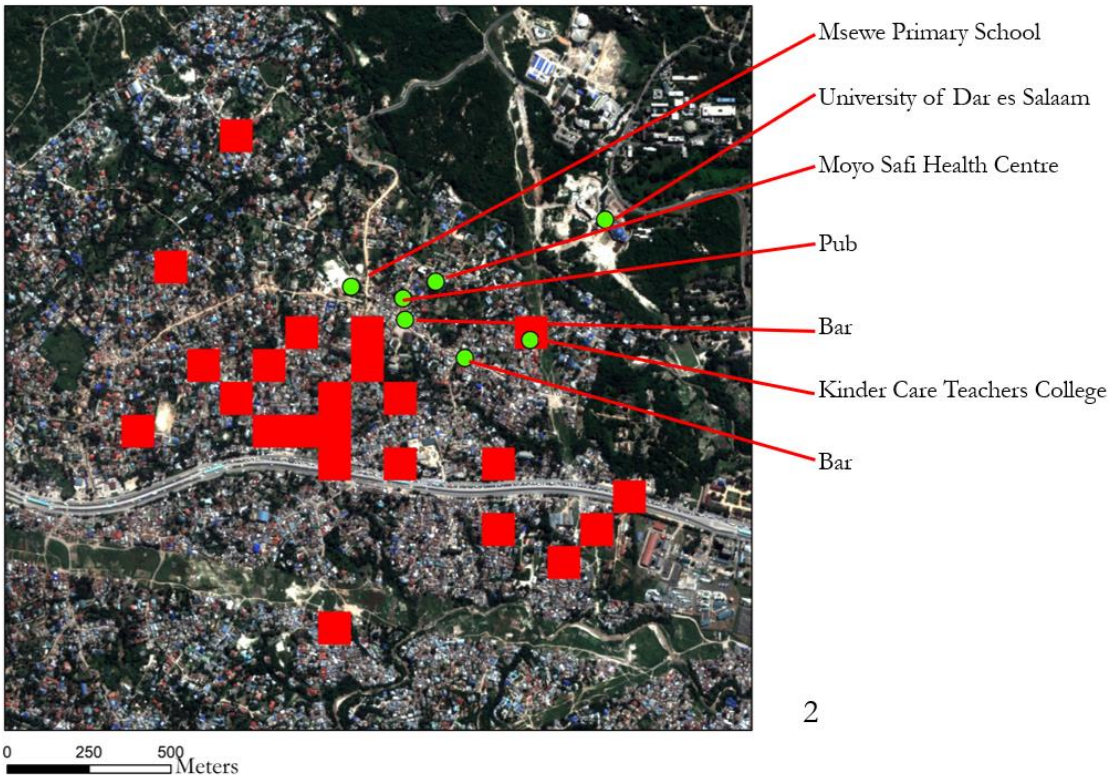


Appendix 8

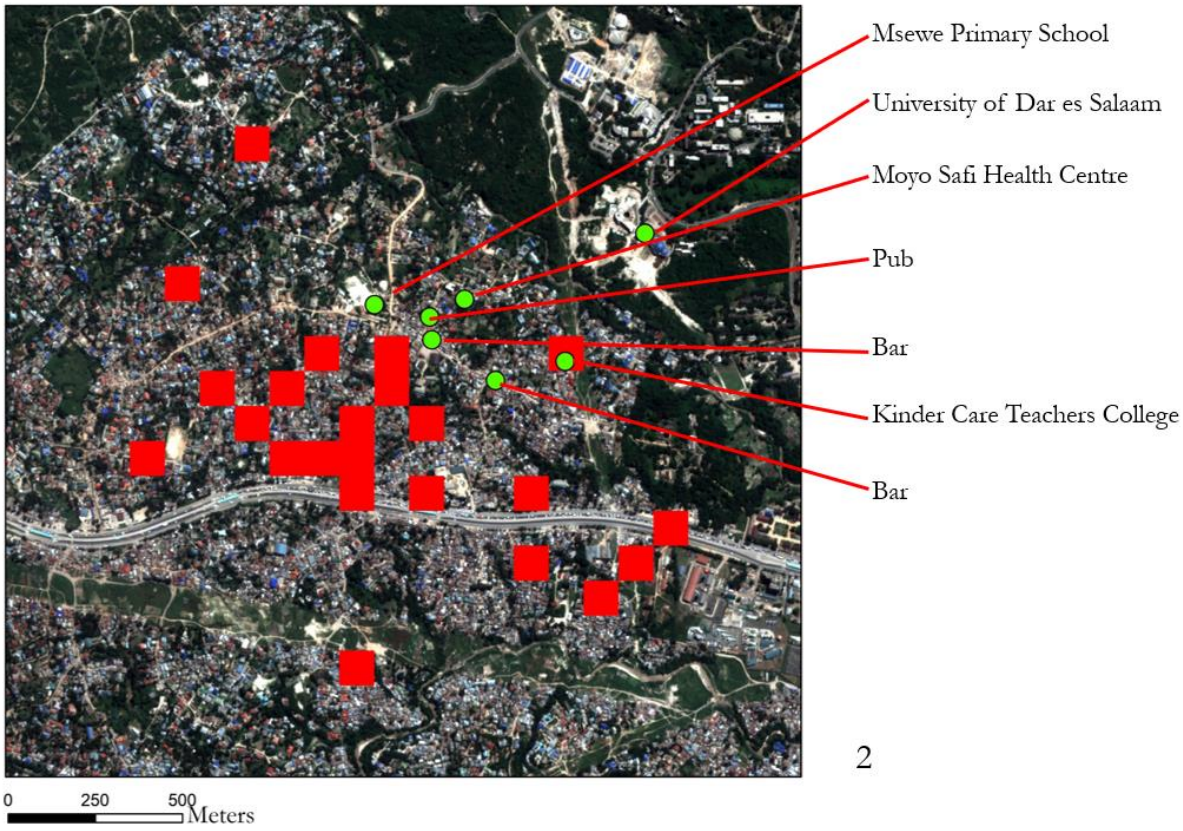
Sector 1



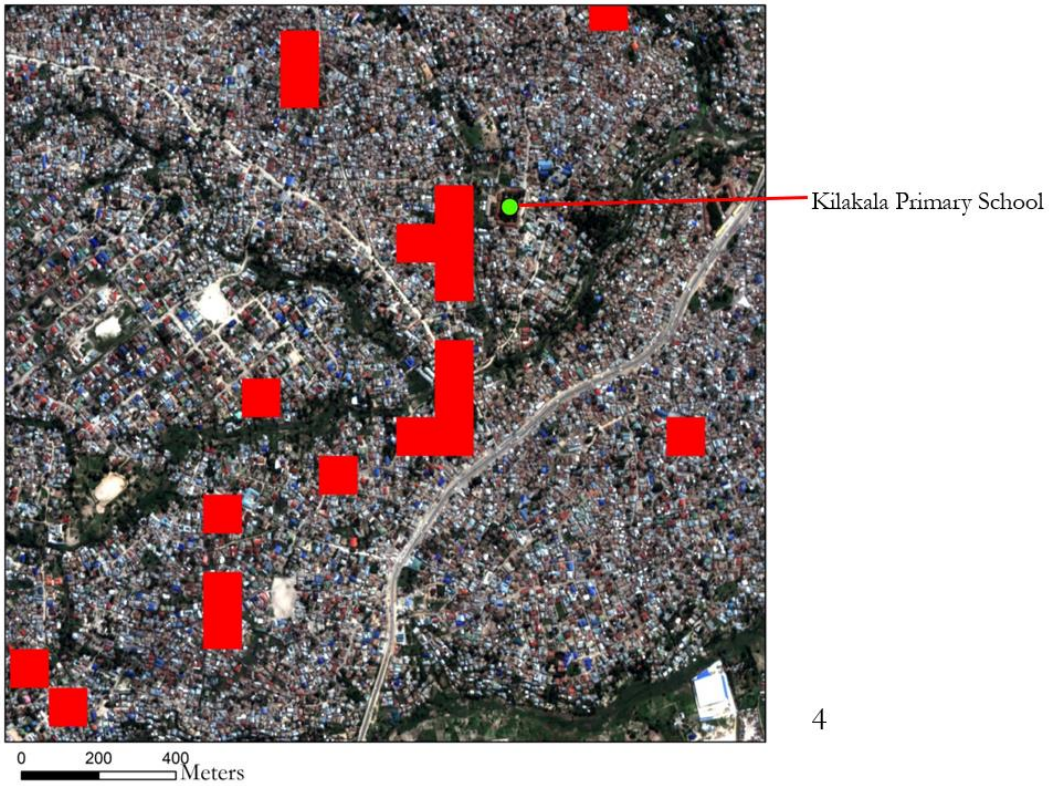
Sector 2



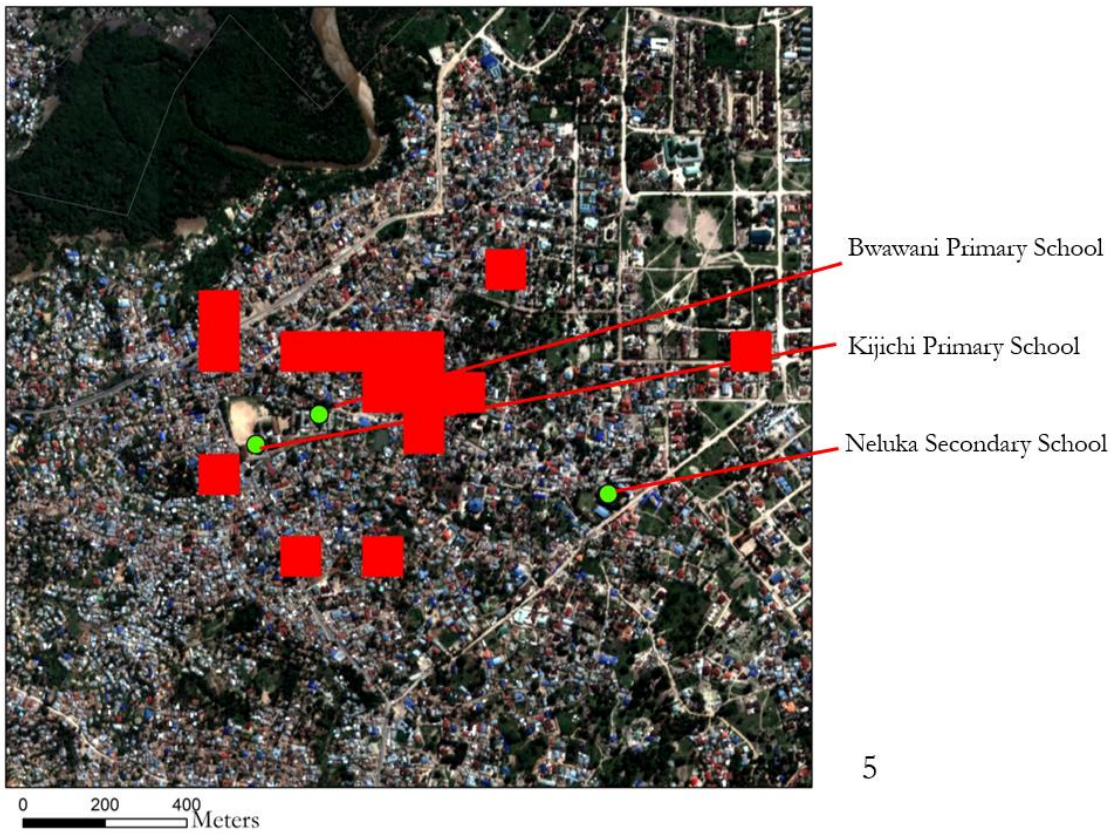
Sector 3



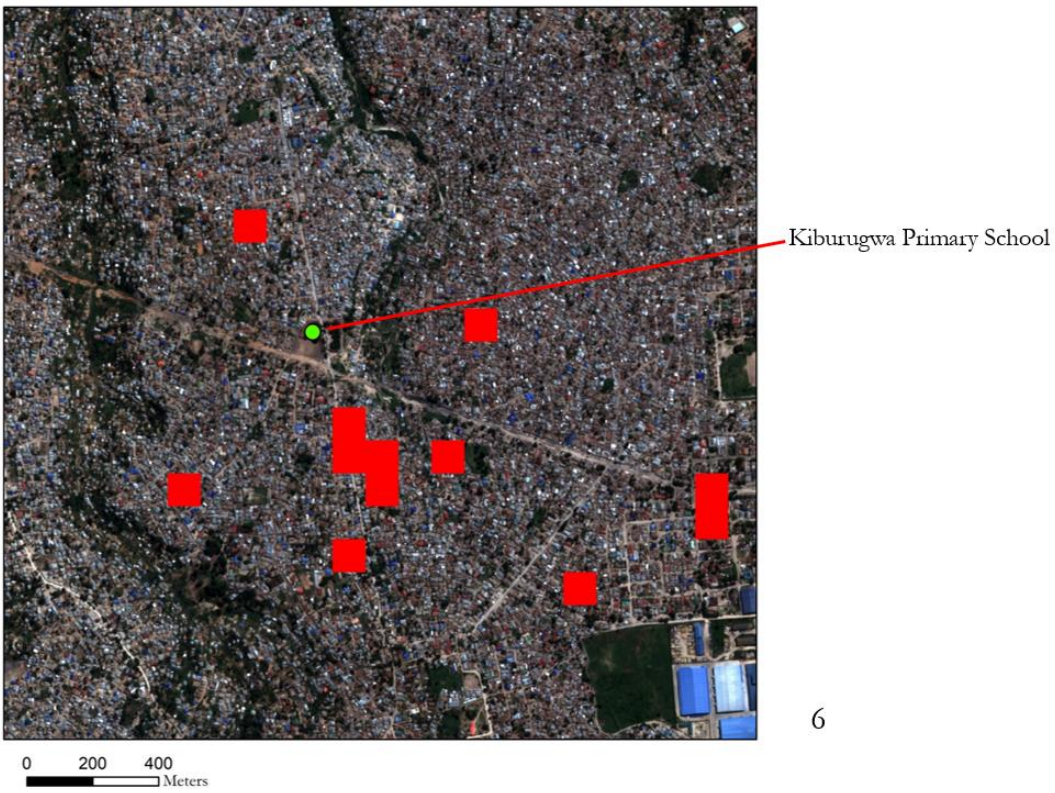
Sector 4



Sector 5



Sector 6



Appendix 9

		Land use map		
	Categories	Formal settlements	Informal settlements	Other
Classification result	Formal settlements	964	658	312
	Informal settlements	824	1579	266
	Other	1806	841	2750