Locally Optimized Mapping of Slum Conditions in a Sub-Saharan Context: A Case Study of Bamenda, Cameroon

Julius Anchang
University of South Florida, yuh@mail.usf.edu

Follow this and additional works at: http://scholarcommons.usf.edu/etd
Part of the Geographic Information Sciences Commons, and the Urban Studies and Planning Commons

Scholar Commons Citation
Anchang, Julius, "Locally Optimized Mapping of Slum Conditions in a Sub-Saharan Context: A Case Study of Bamenda, Cameroon" (2016). Graduate Theses and Dissertations.
http://scholarcommons.usf.edu/etd/6447

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Locally Optimized Mapping of Slum Conditions in a Sub-Saharan Context:

A Case Study of Bamenda, Cameroon

by

Julius Yuh Anchang

A dissertation submitted in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
in Geography and Environmental Science and Policy
School of Geosciences
College of Arts and Sciences
University of South Florida

Co-Major Professor: Ambe Njoh, Ph.D.
Co-Major Professor: Ruiliang Pu, Ph.D.
Joni Downs (Firat), Ph.D.
Fenda Akiwumi, Ph.D.
Jeffrey Cunningham, Ph.D.

Date of Approval:
November 1, 2016

Keywords: multi-index slum measurement, variable clustering, unsupervised learning, satellite imagery, analytic regionalization

Copyright © 2016, Julius Yuh Anchang
DEDICATION

To my parents, John Ngong Anchang (RIP) and Joanna Kess Yuh (RIP). My only regret is that they are not here today to witness the fruit of their labor.
ACKNOWLEDGMENTS

My foremost gratitude goes to my major advisor, Prof. Ambe Njoh and co-Major advisor, Dr Ruiliang Pu, for the mentorship they so willingly proffered for this dissertation. I am fortunate to have benefitted simultaneously from two domain experts, permitting me to approach this dissertation in a holistic manner. I am also grateful for the patience and time they awarded me to get the job done.

This dissertation research could not have been completed without critical data, namely satellite imagery and household survey data. To this end, my profound gratitude goes to the Digital Globe Foundation for providing high resolution GeoEye imagery as part of an image grant award in 2013. I will also like to heartily thank the CAMGIS PLC Chief, Mr Zachary Fru Nsutebu, who was so generous as to let me have full access to recently collected household survey data for the city of Bamenda in Cameroon.

I am indebted to the School of Geosciences, for providing me a high quality academic environment for this dissertation research. My experiences here have been very positive as I have benefitted a lot from faculty, staff and colleagues. Specials thanks to Dr Joni Downs-Firat, Dr Fenda Akiwumi and Dr Jeffrey Cunningham (from Engineering) for accepting to serve on the dissertation committee.

My entirely family, my fiancée, Mefor, and close friends have all been rock solid in their support for me during this challenging but rewarding period. I would certainly not have made it this far without them. I cannot thank them enough.
# TABLE OF CONTENTS

LIST OF TABLES .......................................................................................................................... vi

LIST OF FIGURES ........................................................................................................................ viii

LIST OF ABBREVIATIONS .......................................................................................................... xi

ABSTRACT .................................................................................................................................. xii

1. INTRODUCTION .................................................................................................................... 1
    1.1 Background .......................................................................................................................... 1
    1.2 Statement of the problem .................................................................................................... 3
    1.3 Study aims and research questions .................................................................................. 9
    1.4 Description of the study area ........................................................................................... 10
    1.5 Definition of key terms ..................................................................................................... 13
        1.5.1 Defining slums ........................................................................................................... 13
        1.5.2 Difference between slums and informal settlements .............................................. 15
        1.5.3 Other slum synonyms ............................................................................................ 15
    1.6 Organization of chapters .................................................................................................. 16

2. A REVIEW OF RELATED LITERATURE ........................................................................... 18
    2.1 Chapter overview .............................................................................................................. 18
    2.2. Historical and theoretical foundations of the slum phenomenon .................................. 18
        2.2.1 The slum problem as seen by the West ................................................................. 18
        2.2.2 The slum problem in developing countries ............................................................. 20
            2.2.2.1 The slum as a domain of “hope” or “despair” .............................................. 20
            2.2.2.2 The slum as a byproduct of modernization .................................................... 21
            2.2.2.3 The slum as a product of “disjointed” modernization .......................... 22
        2.3 A review of slum indicators ......................................................................................... 25
            2.3.1 A review of household indicators ....................................................................... 25
2.3.2 The theoretical links between household slum indicators .............................. 28
2.3.3 A review of physical settlement indicators .................................................. 29
2.4 Quantifying slum conditions: from indicators to indices ............................... 30
2.5 A review of key slum mapping literature ...................................................... 32
  2.5.1 Studies using field data ............................................................................ 32
  2.5.2 Studies using remote sensing data (imagery) ............................................. 35
  2.5.3 Hybrid studies that combine image and non-image data ........................... 36
2.6 Machine learning as a potential tool for predicting slum conditions ............... 38
2.7 Analytic regionalization to derive urban spatial units ................................. 39
2.8 Identified gaps in literature and restatement of research objectives ............... 43

3. METHODOLOGY ............................................................................................ 45
  3.1 Chapter overview ......................................................................................... 45
  3.2 Data collection ............................................................................................. 45
    3.2.1 Remote sensing data .............................................................................. 45
    3.2.2 Household data and derivation of spatial units ....................................... 46
    3.2.3 Additional GIS data collection and processing ....................................... 48
  3.3 Image classification to derive land use/cover information ............................ 50
    3.3.1 Vegetation extraction and masking ......................................................... 50
    3.3.2 Object based extraction of non-vegetation land use/cover classes .......... 53
      3.3.2.1 Image segmentation ....................................................................... 53
      3.3.2.2 Selection of optimal features/attributes ............................................ 54
      3.3.2.3 Image object classification ............................................................... 56
    3.3.3 Post classification processing and accuracy assessment ......................... 57
  3.4 Variables used in the study ........................................................................... 58
    3.4.1 Household survey variables aggregated per statistical unit .................. 58
    3.4.2 Variables derived from image classification ........................................... 59
    3.4.3 Additional GIS variables ....................................................................... 59
  3.5 Objective I: To explore a multi-index (multi-factor) measurement approach in
  mapping household slum conditions ............................................................... 59
    3.5.1 Hierarchical variable clustering (HVC) of household indicators .......... 60
    3.5.2 Quantifying derived latent factors as multiple slum indices ................. 64
    3.5.3 Validating indices using location theory ............................................... 65
3.6 Objective II: To investigate the potential of unsupervised learning for mapping slum conditions using physical settlement attributes ................................................................. 66
  3.6.1 Classifying units into unlabeled physical types/categories ....................... 66
  3.6.2 Comparing physical types in terms of household and location attributes ............................................................. 67

3.7 Objective III: To determine neighborhood scale for mapping slums ........ 68
  3.7.1 Spatially constrained clustering of spatial units ................................... 69
  3.7.2 Comparing derived regions in terms of household attributes ................. 71

4. RESULTS AND DISCUSSION - MAIN OBJECTIVE I ........................................ 72
  4.1 Chapter overview ..................................................................................... 72
  4.2 Summary statistics of selected household variables .................................. 72
  4.3 Deriving slum factors and multiple binary slum indices through variable clustering and conceptual interpretation ................................................................. 73
    4.3.1 Results of hierarchical clustering of household slum variables ............ 73
    4.3.2 Results of multi-index construction using slum factors ....................... 76

4.4 Validating derived slum indices using location theory ............................ 80
    4.4.1 Relationship between Occupancy Disadvantage Index (Occ_D) and settlement location attributes ............................................................... 80
    4.4.2 Relationship between Utilities Disadvantage Index (UT_D) and settlement location attributes ............................................................... 82

4.5 Discussion ............................................................................................... 84
    4.5.1 The importance of multi-index measurement of slums ....................... 84
    4.5.2 Validating multiple indices using location theory ............................... 86

4.6 Summary ............................................................................................... 87

5. RESULTS AND DISCUSSION - MAIN OBJECTIVE II .................................. 88
  5.1 Chapter overview ..................................................................................... 88
  5.2 Results of image classification and derivation of physical settlement attributes .......................................................... 88
    5.2.1 Image classification results .................................................................. 88
    5.2.2 Derived physical settlement attributes .............................................. 90

5.3 Results of mapping slum conditions using physical attributes of statistical units ......................................................................................... 92
LIST OF TABLES

Table 1.1 Summary of some housing characteristics of Bamenda ................................................. 12
Table 2.1 UN-HABITAT designated slum indicators and their recommended benchmarks ......... 26
Table 2.2 Key physical slum indicators ..................................................................................... 30
Table 3.1 Description of object feature attributes ..................................................................... 55
Table 3.2 Description of training sample used for object classification ..................................... 55
Table 3.3 Description of samples selected for accuracy assessment ........................................... 58
Table 4.1 Summary of selected household indicators aggregated at statistical unit level .......... 73
Table 4.2 Bi-variate correlation coefficients (Spearman) between selected household variables ........................................................................................................ 74
Table 4.3 Variable groups and potential factors derived from clustering .................................. 76
Table 4.4 Location attributes of units after group balancing ....................................................... 81
Table 4.5 T-test comparing units of different Occ_D levels in terms of location attributes........ 82
Table 4.6 Summary of location attributes of units after UT_D group balancing ......................... 83
Table 4.7 T-test comparing units of different UT_D levels in terms of location attributes .......... 83
Table 5.1 Classification accuracy assessment ............................................................................. 90
Table 5.2 Description of principal components used for AHC ................................................ 92
Table 5.3 Derived physical types in terms of original variables .................................................. 94
Table 5.4 Kruskal-Wallis tests comparing distribution of 6 household variables across all four physical types .............................................................................................. 96
Table 5.5 Kruskal-Wallis tests comparing distribution of 3 location variables across all four physical types ....................................................................................................... 99
Table 6.1 Analytic regions in terms of physical attributes.................................................................111

Table 6.2 Kruskal-Wallis tests comparing distribution of 6 household variables across all 5 regions ........................................................................................................................................112
LIST OF FIGURES

Figure 1.1 Map of Study Area: A) Cameroon, B) North West Region, C) 2012 GeoEye Image of Bamenda (Courtesy of Digital Globe Foundation) ................................................. 11

Figure 1.2 Physical growth of Bamenda from 1978 to 2007 (Courtesy of USGS) .................... 12

Figure 0.1 Simple illustration of regionalization (4 regions from 16 units) ................................. 40

Figure 3.1 Base map of statistical units: A) original units B) selected units (ESRI, 2015) ....... 49

Figure 3.2 Map showing additional GIS layers: CBD, market places, roads, elevation (ESRI, 2015) ........................................................................................................ 49

Figure 3.3 Distance rasters used to derive location variables (ESRI, 2015) .............................. 50

Figure 3.4 Satellite image processing workflow ........................................................................ 51

Figure 3.5 Object feature selection based on feature importance score (IBM, 2012) ............. 56

Figure 3.6 Image subset showing selection of test areas/pixels for accuracy assessment (EXELIS, 2012) ............................................................................................... 58

Figure 3.7 Specified options in SPSS used for hierarchical clustering of variables (IBM, 2012) ............................................................................................................... 62

Figure 3.8 Hypothetical example showing hierarchical clustering of variables ..................... 63

Figure 3.9 Depiction of MST algorithm: a) Connectivity graph; b) Minimum spanning tree after graph pruning (Assunção et al., 2006, p. 800) .......................................................... 70

Figure 3.10 Tool input parameters for first run of spatially constrained clustering using ArcMap (ESRI, 2015) ...................................................................................................... 71

Figure 4.1 Results of hierarchical clustering of variables .......................................................... 75

Figure 4.2 Map of derived slum indices (Occ_D and UT_D) .................................................... 77

Figure 4.3 Occ_D expressed in terms of % of non-owner households ..................................... 78
Figure 6.5 Box Plot: Mean % households without drainage by region ...........................................113
Figure 6.6 Pair wise comparisons of Regions: % non-owner households.................................114
Figure 6.7 Pair wise comparisons of Regions: % households with >2 residents per room ..........115
Figure 6.8 Pair wise comparisons of Regions: % households without drainage ......................115
Figure 6.9 Map of derived regions overlaid with vernacular neighborhood names ..................119
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHC/HC</td>
<td>Agglomerative hierarchical clustering</td>
</tr>
<tr>
<td>AMOEBA</td>
<td>A multidirectional optimal ecotope-based algorithm</td>
</tr>
<tr>
<td>AR</td>
<td>Analytic regionalization</td>
</tr>
<tr>
<td>AZP</td>
<td>Automated zoning procedure</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>HSRS</td>
<td>High Spatial Resolution Satellite</td>
</tr>
<tr>
<td>IPUMS</td>
<td>Integrated Public Use Microdata Series</td>
</tr>
<tr>
<td>KW</td>
<td>Kruskal-Wallis</td>
</tr>
<tr>
<td>LOSH</td>
<td>Local spatial heteroskedasticity</td>
</tr>
<tr>
<td>MDG</td>
<td>Millennium Development Goals</td>
</tr>
<tr>
<td>MST</td>
<td>Minimum Spanning Tree</td>
</tr>
<tr>
<td>REGION</td>
<td>Image derived analytic neighborhood</td>
</tr>
<tr>
<td>SL</td>
<td>Supervised learning</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical package for the social sciences</td>
</tr>
<tr>
<td>SSA</td>
<td>Sub-Saharan Africa</td>
</tr>
<tr>
<td>UL</td>
<td>Unsupervised learning</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>UNDATA</td>
<td>United Nations Data</td>
</tr>
<tr>
<td>UNEP</td>
<td>United Nations Environment Program</td>
</tr>
<tr>
<td>UN-HABITAT</td>
<td>United Nations Human Settlement Program</td>
</tr>
<tr>
<td>UNSTAT</td>
<td>United Nation Statistics</td>
</tr>
<tr>
<td>ESRI</td>
<td>Environmental Systems Research Institute</td>
</tr>
<tr>
<td>IBM</td>
<td>International Business Machine</td>
</tr>
<tr>
<td>ENVI</td>
<td>Environment for Visualizing Images</td>
</tr>
</tbody>
</table>
ABSTRACT

Despite being an indicator of modernization and macro-economic growth, urbanization in regions such as Sub-Saharan Africa is tightly interwoven with poverty and deprivation. This has manifested physically as slums, which represent the worst residential urban areas, marked by lack of access to good quality housing and basic services. To effectively combat the slum phenomenon, local slum conditions must be captured in quantitative and spatial terms. However, there are significant hurdles to this. Slum detection and mapping requires readily available and reliable data, as well as a proper conceptualization of measurement and scale. Using Bamenda, Cameroon, as a test case, this dissertation research was designed as a three-pronged attack on the slum mapping problematic. The overall goal was to investigate locally optimized slum mapping strategies and methods that utilize high resolution satellite image data, household survey data, simple machine learning and regionalization theory.

The first major objective of the study was to tackle a "measurement" problem. The aim was to explore a multi-index approach to measure and map local slum conditions. The rationale behind this was that prior sub-Saharan slum research too often used simplified measurement techniques such as a single unweighted composite index to represent diverse local slum conditions. In this study six household indicators relevant to the United Nations criteria for defining slums were extracted from a 2013 Bamenda household survey data set and aggregated for 63 local statistical areas. The extracted variables were the percent of households having the following attributes: more than two residents per room, non-owner, occupying a single room or studio, having no flush toilet,
having no piped water, having no drainage. Hierarchical variable clustering was used as a surrogate for exploratory factor analysis to determine fewer latent slum factors from these six variables. Variable groups were classified such that the most correlated variables fell in the same group while non-correlated variables fell in separate groups. Each group membership was then examined to see if the group suggested a conceptually meaningful slum factor which could quantified as a stand-alone "high" and "low" binary slum index. Results showed that the slum indicators in the study area could be replaced by at least two meaningful and statistically uncorrelated latent factors. One factor reflected the home occupancy conditions (tenancy status, overcrowded and living space conditions) and was quantified using K-means clustering of units as an ‘occupancy disadvantage index’ (Occ_D). The other reflected the state of utilities access (piped water and flush toilet) and was quantified as utilities disadvantage index (UT_D). Location attributes were used to examine/validate both indices. Independent t-tests showed that units with high Occ_D were on average closer to nearest town markets and major roads when compared with units of low Occ_D. This was consistent with theory as it is expected that typical slum residents (in this case overcrowded and non-owner households) will favor accessibility to areas of high economic activity. However, this situation was not the same with UT_D as shown by lack of such as a strong pattern.

The second major objective was to tackle a "learning" problem. The purpose was to explore the potential of unsupervised machine learning to detect or "learn" slum conditions from image data. The rationale was that such an approach would be efficient, less reliant on prior knowledge and expertise. A 2012 GeoEye image scene of the study area was subjected to image classification from which the following physical settlement attributes were quantified for each of the 63 statistical areas: per cent roof area, percent open space area, per cent bare soil, per cent paved road
surface, per cent dirt road surface, building shadow-roof area ratio. The shadow-roof ratio was an innovative measure used to capture the size and density attributes of buildings. In addition to the 6 image derived variables, the mean slope of each area was calculated from a digital elevation dataset. All 7 attributes were subject to principal component analysis from which the first 2 components were extracted and used for hierarchical clustering of statistical areas to derive physical types. Results show that area units could be optimally classified into 4 physical types labelled generically as Categories 1 – 4, each with at least one defining physical characteristic. Kruskal Wallis tests comparing physical types in terms of household and locations attributes showed that at least two physical types were different in terms of aggregated household slum conditions and location attributes. Category 4 areas, located on steep slopes and having high shadow-to-roof ratio, had the highest distribution of non-owner households. They were also located close to nearest town markets. They were thus the most likely candidates of slums in the city. Category 1 units on other hand located at the outskirts and having abundant open space were least likely to have slum conditions.

The third major objective was to tackle the problem of "spatial scale". Neighborhoods, by their very nature of contiguity and homogeneity, represent an ideal scale for urban spatial analysis and mapping. Unfortunately, in most areas, neighborhoods are not objectively defined and slum mapping often relies in the use of arbitrary spatial units which do not capture the true extent of the phenomenon. The objective was thus to explore the use of analytic regionalization to quantitatively derive the neighborhood unit for mapping slums. Analytic neighborhoods were created by spatially constrained clustering of statistical areas using the minimum spanning tree algorithm. Unlike previous studies that relied on socio-economic and/or demographic information, this study innovatively used multiple land cover and terrain attributes as neighborhood homogenizing
factors. Five analytic neighborhoods (labeled Regions 1-5) were created this way and compared using Kruskal Wallis tests for differences in household slum attributes. This was to determine largest possible contiguous areas that could be labeled as slum or non-slum neighborhoods. The results revealed that at least two analytic regions were significantly different in terms of aggregated household indicators. Region 1 stood apart as having significantly higher distributions of overcrowded and non-owner households. It could thus be viewed as the largest potential slum neighborhood in the city. In contrast, regions 3 (located at higher elevation and separated from rest of city by a steep escarpment) was generally associated with low distribution of household slum attributes and could be considered the strongest model of a non-slum or formal neighborhood. Both Regions 1 and 3 were also qualitatively correlated with two locally recognized (vernacular) neighborhoods. These neighborhoods, "Sisia" (for Region 1) and "Up Station" (for Region 3), are commonly perceived by local folk as occupying opposite ends of the socio-economic spectrum.

The results obtained by successfully carrying the three major objectives have major implication for future research and policy. In the case of multi-index analysis of slum conditions, it affirms the notion the that slum phenomenon is diverse in the local context and that remediation efforts must be compartmentalized to be effective. The results of image based unsupervised mapping of slums from imagery show that it is a tool with high potential for rapid slum assessment even when there is no supporting field data. Finally, the results of analytic regionalization showed that the true extent of contiguous slum neighborhoods can be delineated objectively using land cover and terrain attributes. It thus presents an opportunity for local planning and policy actors to consider redesigning the city neighborhood districts as analytic units. Quantitively derived neighborhoods are likely to be more useful in the long term, be it for spatial sampling, mapping or planning purposes.
1. INTRODUCTION

1.1 Background

Our planet is more urban today than it has ever been before. To quote from a 2007 online article featured by The Economist: "Homo sapiens has become Homo urbanus" (Grimond, 2007). Statistically speaking, the percentage of urban dwellers in the world has increased from about 20% to over 50% over the last century (UN-HABITAT, 2013, p. 25). At the same time, there has been a reversal of roles between the developed and developing regions, with the latter overtaking the former in terms of total contribution to the global urban pool. Countries in developing regions such as Sub-Saharan Africa (SSA) currently exhibit the highest annual urban growth rates (Njoh, 2003; UN-HABITAT, 2010). Take for example central Africa, a region normally considered among the least urbanized parts of the world. At least five countries therein - Angola, Cameroon, Congo, Gabon, São Tomé e Príncipe - currently have urban majorities (UN-HABITAT, 2014).

The causes and consequences of this rapid urbanization, and its relation with development, is a well-known subject matter of academic inquiry (Fox, 2012; Henderson, Storeygard, & Roberts, 2013; Njoh, 2003). Despite being generally indicative of macro-economic growth at the cross-country level, urbanization in developing regions such as SSA is also intertwined with poverty (Njoh, 2003). A significant proportion of new urban residents in developing countries are poor migrants from rural areas succumbing to “push/pull” migration factors (Jahan, 2012). This has contributed to the process known as the “urbanization of poverty,” a notion that the urban setting has either already or will soon outpace the rural setting as the primary domain of the poor. This
rise in urban poverty has been labeled by the World Bank as “significant and politically explosive” (Shi, 2000, p. 14).

A tangible evidence of urban poverty is found in the housing sector, specifically through the prevalence of slums (Arimah, 2010). Though often lacking consistency and clarity in definition, the slum phenomenon has come to be universally recognized as the embodiment of poor urban residential conditions that dominate developing countries. In SSA for instance, provision of housing and services has proven largely insufficient in the face of an ever increasing demand by a mostly poor urban populace (Fox, 2014). A significant portion of the new urban caste are unskilled and under-educated rural-urban migrants restricted to the low income informal economic sector (De Soto, 2000). Upon arrival at the urban scene, they intuitively seek to minimize cost of living via cheaper but otherwise inadequate housing. Limited upward socio-economic mobility means the status quo persists longer than intended, leading to the establishment of permanent slum areas.

The classical notion of slums calls to mind congested central city areas that have fallen into disrepair and disrepute. Their continued existence is sustained by their attractive location as residents prioritize access to central areas of the city where economic activity is most intensive. Landlords and property owners can thus maintain profits by renting out units with minimal investment in repair and maintenance. This is typical of the Western understanding of slums, as a product of industrialization, migration and urban flight. However, another iteration of the slum phenomenon permeates the modern urban scene, particularly in cities of the global South. While traditional ‘central’ slums still occur in these regions, there is the also the proliferation of mostly new areas that consist of unplanned and even outright illegal settlements also known as informal or squatter settlements.
The slum problem is arguably one of the most daunting urbanization-related global issues of the 21st century. Apart from being a visible testament to urban poverty, slum incidence has been positively correlated with public health ills like higher child mortality rates and disease incidence (Hanchett, Akhter, Khan, Mezulianik, & Blagbrough, 2003). About a third of urban residents worldwide (i.e., almost a billion people) currently reside in areas considered as slums (UN-HABITAT, 2003, p. 5). Again, just as with the general issue of urbanization, the slum problem is quite profound in SSA. Whereas other developing regions have managed to curb their overall slum incidence to some extent, such improvement is far less visible in SSA, where a healthy majority continue to reside in slums (Fox, 2014; UN-HABITAT, 2008, 2010). Considering that a significant chunk of future human population growth is expected to be absorbed by cities in such regions (Martine, 2012), we can safely assume that, if left unaddressed, slums will be the dominant feature of future urban landscape.

A further evidence of the major international concern surrounding the slum phenomenon can be found within the Millennium Development Goals (MDGs) framework. Target 11 under MDG Number 7 specifically aims to achieve a “significant improvement in the living conditions of 100 million slum dwellers by 2020” (UN-HABITAT, 2003, p. 7). The United Nations Human Settlement Program (UN-HABITAT) is the agency specifically charged with monitoring the progress towards achieving this goal. Efforts in this regard have so far resulted in the development of a universal framework for defining and measuring slums, focusing on objectively measured household attributes such as: access to water, sanitation, secured tenure, living space and a durable dwelling (UN-HABITAT, 2003).

1.2 Statement of the problem

To implement effective slum improvement efforts, the slum phenomenon itself must first
be adequately captured in quantitative and spatial terms. In other words, local slum conditions must be measured and mapped to determine priorities for geographically targeted intervention efforts.

The first obvious step is to have a proper understanding of what slums are. This would allow for a proper definition and hence quantification of the conditions. Unfortunately, the slum phenomenon has always been too complex to nail down with a simple definition. There is a universal understanding of what slums represent (i.e., poor housing and lack of services) but this does not easily lend itself to local nuances. Despite their ubiquitous nature, slums often manifest in different ways locally with potentially very contrasting traits. As such different slums categories have been derived with names such as “inner-city/central slums”, “slum islands”, “peripheral squatters”, etc. (UN-HABITAT, 2003, p. 85). This plurality no doubt complicates the development of a generalizable mapping framework and requires that efforts be customized for local situations.

There are practical problems that complicate effective slum mapping efforts. A major practical concern stems from the required data. There are two broad types of slum related data: 1) field data and; 2) remote sensing data. Traditional field instruments such as censuses and surveys can be used for collecting information about household assets and socioeconomic attributes. If properly georeferenced, this information can then be input into a Geographic Information System (GIS) database to allow for direct mapping of slum indicators/index (Weeks, Hill, Stow, Getis, & Fugate, 2007). In some cases, such data may not even directly measure slum attributes but proxies such as poverty or other forms of socio-economic deprivation (Baud, Sridharan, & Pfeffer, 2008; Baud, Pfeffer, Sridharan, & Nainan, 2009; Davis, 2003). Field data can also be qualitative such as those obtained using participatory surveys, interviews and focus groups. These can take into account local as well as expert knowledge and perceptions on slum conditions (Karanja, 2010;
The prohibitive cost of obtaining field data however limits their usefulness (Kohli, Sliuzas, Kerle, & Stein, 2012; Owen & Wong, 2013). In most cases, such field data are not collected frequently enough to match the temporal dynamics of urban settlements and slum areas under study. For example, since independence, only three household level censuses (conducted in 1975, 1987, and 2005) have been administered in Cameroon since the country gained independence in 1960 (Minnesota Population Center, 2015). More so, this census data are often only georeferenced to the top 3 - 4 administrative levels of the country: region, division, sub-division and possibly commune. Except for a few local cases, there is little or no household level data at the finer sub-city levels, despite the importance of such data for detailed slum mapping.

Remote sensing (RS) data, such as earth observation satellite imagery, provides a viable alternative to field data. RS allows for the collection of unbiased, spatially extensive and temporally consistent information on the human physical environment (Patino & Duque, 2012). The last decade and a half has seen the deployment of space-borne instruments such as Digital Globe’s IKONOS, GeoEye, and World-View. These instruments can acquire optical imagery of anywhere on the earth’s surface at up to sub-meter resolutions. This means RS data can now match the detailed level of field-work at much lower relative cost. The usefulness of imagery in social scientific studies stems from the fundamental hypothesis that human behavior shapes and is also shaped by the immediate physical environment (Rashed, Weeks, Stow, & Fugate, 2005). The general question that needs to be answered is: how do we confidently associate the image data with a phenomenon of interest? In the context of slums, the primary question is: how do we associate non-image slum indicators, such as access to water and sanitation, with image-derived settlement indicators? Commercial satellite image coverage has steadily improved in recent years.
and costs have generally dropped. As such, the technical challenge facing regions such as SSA today is not RS data availability. The challenge is having the requisite tools to process such data. Such tools need to be efficient, inexpensive and have a high potential for automation to be useful in operational circumstances.

Besides the data constraints, there are other methodological difficulties that are faced when attempting to map slum conditions. These difficulties stem from different aspects of the conceptual framework such as definition, measurement, and determination of spatial scale. Some questions that still need to be properly answered are: how do we quantify slum conditions for a specific local context? How do we learn or predict slum conditions from associated data? How can we determine the appropriate spatial scale for mapping slum conditions?

Quantitative slum measurement is typically done by employing as exhaustively as possible, observed indicators defined by theory or expert knowledge. These include indicators such as those laid down by United Nations (UN) (e.g. access to water and sanitation) (UN-HABITAT, 2003). These indicators are easily measured at the household level and can also be aggregated over various spatial scales. A common trend in slum mapping research so far has been to combine multiple indicators into a single composite slum or proxy variable, such as a binary index (e.g. Graesser et al., 2012; Owen & Wong, 2013); or one calculated on a continuous or ranked scale (e.g. Baud et al., 2008; Weeks et al., 2007). This facilitates the mapping of global slum conditions in an area. However, crucial information and context can be lost, such as how individual indicators or groups of indicators behave locally.

When using image data for slum mapping, the objective is usually to discern or "learn" the slum status of a settlement based from information contained in the image scene. Heuristics as well as empirical evidence strongly suggest that slum areas have recognizable physical attributes
that can be used to distinguish likely slum from non-slum areas. For example, in several case studies, slums have been shown to have higher building densities, little or no vegetation, irregular lay out structure, hazardous location, amongst others (Hofmann, Strobl, Blaschke, & Kux, 2008; Kit, Lüdeke, & Reckien, 2012; Owen & Wong, 2013). While these attributes can be easily derived from imagery, the problem however lies with the inconsistency in how they capture slums in different local settings. For example there is no universal threshold for settlement size and building density above which a settlement should be considered a slum (Gulyani & Bassett, 2010). Therefore, non-image information related to local slum conditions is often required to support any image based mapping approach.

The combination of non-image and image data takes place under the framework of supervised learning, a philosophy that emphasizes the \textit{a priori} specification of a target (response) variable of interest (Kotsiantis, Zaharakis, & Pintelas, 2007). In simple terms, this means prior information on slum conditions for at least a sample of mapping units is needed to inform and validate image based approach. Supervised learning has been used in slum mapping research by employing techniques such as regression modeling and decision trees (Graesser et al., 2012; Stoler et al., 2012; Weeks et al., 2007). Unfortunately, supervised methods are heavily user-reliant as they depend on the correct definition and measurement of slums. More so, in the case where there is no readily available and up to date secondary slum data, as is common in many developing countries, costly field work will be needed.

Another major methodological challenge faced when mapping slum is how to determine the appropriate scale and boundaries of spatial units. More frequently than not, slum research utilizes arbitrary mapping units (such as census tracts or administrative units) for which indicators can be conveniently aggregated (Sluizas & Kuffer, 2008). This leads to numerous spatial analytic
problems, most notable of which is the modifiable areal unit problem (MAUP) (Openshaw & Openshaw, 1984). The MAUP is a perennial problem for the spatial analyst. Arbitrary spatial units such as those derived from administrative divisions may have extents and boundaries which do not reflect the spatial distribution of the phenomenon being investigated or mapped (Davis, 2003; Páez & Scott, 2005). In the case of slums, it means they may not reflect the true boundaries of contiguous slum areas. Unfortunately, census/survey data that are frequently used for slum mapping often have to be aggregated into such units, for operational convenience and to protect the identity of respondents (Kohli et al., 2012).

The importance of scale is highlighted in the existing definition of slums by the UN, which includes the assertion that a slum must be a "contiguous" area with "a minimum settlement size" (UN-HABITAT, 2003, p. 11). The determination of this "size" however remains ambiguous and subjective at best. In India for example, a slum is considered to be “a compact area of at least 300 people or about 60-70 households” (according to Census of India (2001) as cited in Kit et al. (2012)). Contrast this with Bangladesh where a slum is considered to be made up of “five or more households” (Hanchett et al., 2003, p. 44). This discrepancy in size moving from one locality to another is clearly a source of constraint for developing a generalizable mapping framework.

More clearly understood is the aspect of contiguity. This means an individual slum must have an unbroken spatial extent. This proves problematic in certain situations. Some localities, for example, may have homogenous settlement attributes with easily discernible neighborhood boundaries, while others may have a mix of different settlement characteristics within the same area (Schneider-Sliwa & Bhatt, 2008; Taubenböck & Kraff, 2014; Turner, 1969). In the case of the latter, the conceptualization of slums as contiguous homogenous settlement areas will no doubt be very challenging and will require more flexibility in determining scale and boundaries.
1.3 Study aims and research questions

The preceding paragraphs demonstrate clearly that slum mapping is challenged on both conceptual and practical fronts. As such, this dissertation research was conducted with the following primary aim: to investigate methodologies that integrate geospatial information (remote sensing and GIS) and basic machine learning for locally optimized mapping of slum conditions. This was achieved using household survey data, very high resolution satellite imagery and ancillary GIS data. The study emphasized the use of data driven and less user reliant methods to achieve mapping outcome.

The specific objectives and the associated research questions were as follows:

1. **Objective**: To explore a multi-index approach for local slum mapping.
   
   **Questions**: Are there multiple underlying and uncorrelated factors that represent the variability of aggregated household indicators across local spatial units? How do these factors relate with broader settlement theory such as location?

2. **Objective**: To investigate the potential of unsupervised learning for determining slum conditions from remote sensing (image) data.
   
   **Questions**: What are the image-derived physical types into which areal units can be classified? Do differences in physical typology reflect differences in aggregated household conditions within areal units? Do they relate to unit location attributes?

3. **Objective**: To investigate the potential of analytic regionalization for determining optimal neighborhood units for mapping slums.
   
   **Questions**: What are the optimal neighborhood boundaries in the study area based on land cover and terrain homogeneity? Do these analytically derived neighborhoods represent significantly different levels of household slum conditions?
1.4 Description of the study area

After establishing the research problem and objectives, it is important to situate this research in the context of the study area. The study was carried out within the City extent of Bamenda, located approximately 6°N and 10°E (Figure 1.1). Bamenda is the capital of one of Cameroon’s ten administrative regions, namely the North-West Region. The built-up area extends approximately 10 kilometres in both cardinal directions with an estimated population of about 270,000 (Minnesota Population Center, 2015). In term of the size of urban areas in Cameroon, Bamenda is the third largest (albeit a distant third) after Douala and Yaoundé, both of which are essentially primate cities.

Bamenda lies within the Western Highlands geographic region of Cameroon, and is characterized by elevations ranging between 1000 - 2000 meters above sea level. The region originally boasts a mixture of Guinea-Savannah and montane forest. While this is still visible in surrounding areas, little or no native vegetation survives within the city itself. The current urban vegetation in Bamenda is dominated by non-native trees, shrubs, grass fields and plots of farmland. In addition to the above attributes, Bamenda has a network of rivers and streams, principal among which are the Mezam River and its tributaries that follow a dendritic flow pattern. Several parts of the area are thus prone to natural hazards such as seasonal flash floods (Guedjeo et al., 2013; Nyambod, 2010)

There are several factors that make Bamenda a suitable test candidate for this study. The city’s population increased by almost 400% between 1975 and 2005, from 50,000 to 270,000 (Minnesota Population Center, 2015). However, being well below 500,000, it would still be considered a small-size city. It is worth highlighting the fact that such a profile is quite common among African urbanities and accounts for almost half of the urban population worldwide (UN,
Physically, Bamenda has also experienced significant expansion in an outward radial pattern over the past few decades (see Figure 2). Morphologically it contains the main structural elements of a typical African city as described in De Blij’s model (De Blij, 1964). This includes a historic "Old Town" neighborhood, a Central Business District (CBD), and new residential settlements emerging at the periphery. Typical to common urban growth models, new development in the area tends to follow the trunk road network.

Figure 1.1 Map of Study Area: A) Cameroon, B) North West Region, C) 2012 GeoEye Image of Bamenda (Courtesy of Digital Globe Foundation)
Table 1.1 Summary of some housing characteristics of Bamenda

<table>
<thead>
<tr>
<th>Living space</th>
<th>toilets</th>
<th>Sewage</th>
<th>Roof</th>
<th>Wall</th>
<th>tenancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Averagely</td>
<td>18.5% have flush toilet, 10% have private septic tank, 90% have no direct sewage disposal system</td>
<td>89 % use corrugated</td>
<td>virtually all (99.3%) use stone and brick</td>
<td>44% owner occupied, 52 % tenant occupied</td>
<td></td>
</tr>
<tr>
<td>2.24 persons per bedroom</td>
<td>80.5% use latrine</td>
<td>Virtual use</td>
<td>Virtually all (99.3%) use stone and brick</td>
<td>44% owner occupied, 52 % tenant occupied</td>
<td></td>
</tr>
</tbody>
</table>

Source: Summarized from 2005 Census microdata (Minnesota Population Center, 2015)
Table 1.1 (above) summarizes housing conditions in the city, extracted from the 2005 National Census of Cameroon dataset. This was downloaded using IPUMS International census data portal (Minnesota Population Center, 2015). This provides direct evidence of existence of universal slum conditions in the area and as well insight on how different Bamenda may be from other cities. For example, while relatively few people have modern flush toilets and sewage facilities, the clear majority of buildings are constructed with relatively durable materials. It must be mentioned here however that there are potential inaccuracies particularly with regards to building materials. There is a common local practice of constructing houses using cheaper adobe (mud) blocks followed by cement mortar plastering before painting. To the naked eye of the census or survey administrator, these will not be discernible from modern cement block constructions. However, given Bamenda’s relatively cold and dry climate, there is every indication that mud bricks are suitable enough, although it may also matter if the house is located on a geologically unstable site.

1.5 Definition of key terms

1.5.1 Defining slums

A simple internet dictionary query reveals that the word “slum” has been used since the early 19th century to describe "an area of a city where poor people live and the buildings are in bad condition" (Merriam-Webster, 2016). This definition captures the classical picture of slums, i.e., as congested decadent inner-city areas that were a common feature in European and North American cities prior to the 20th Century. Though initially used in mostly informal and derogatory language (Dyos, 1967), the term “slum” overtime became part of formal/academic lexicon, with increasing usage in urban social studies. The following are some examples of early definitions:

"... an area of physical deterioration in which people live out of economic compulsion or
to escape discipline and control of standards imposed in more stable neighborhoods…” (Lind, 1930, p. 207).

“a building, groups of buildings, or area characterized by overcrowding, deterioration, unsanitary conditions or absence of facilities or amenities which endanger the health, safety or morals of its inhabitants or the community” (Gutkind, 1960, p. 131).

The common denominator of such early definitions is that they included qualitative assertions about the disreputable nature of slums. Such definitions looked beyond mere housing characteristics, highlighting the prevalence of social, economic and health-related ills within slum communities. While this may be viewed as offering a more complete picture, it made the actual measurement of the slum phenomenon more problematic from a practical perspective.

More recent definitions are offered by the UN. For example, the glossary section of the United Nations Information Portal (UNDATA) defines slums as:

“areas of older housing that are deteriorating in the sense of their being underserviced, overcrowded and dilapidated” (UNDATA, 2016).

Another good example is that proposed and adopted by participants in a United Nations Expert Group Meeting held in October 2002 in Switzerland. They defined a slum as:

"a contiguous settlement where the inhabitants are characterized as having inadequate housing and basic services ” (UN-HABITAT, 2002, 2003).

Contemporary slum definitions like the two mentioned above have a more surgical precision. The clear emphasis is on housing, specifically the quantifiable aspects of dwelling units. Largely ignored are the socio-economic and cultural attributes of the households and communities, which though relevant, are generally more difficult to measure objectively.
1.5.2 Difference between slums and informal settlements

The terms “slums” and “informal settlements” are frequently used interchangeably in modern literature. However, typical definitions of informal settlements come from an understanding that is conceptually different from the traditional notion of slums. A pair of definitions of informal settlements are provided in the glossary section of the United Nations Statistics Division (UNSTAT) as follows:

" 1. areas where groups of housing units have been constructed on land that the occupants have no legal claim to, or occupy illegally;

2. Unplanned settlements and areas where housing is not in compliance with current planning and building regulations (unauthorized housing)’’ (UNSTAT, 2016).

Thus, in the simplest sense, informal settlements are residential areas comprising of units whose existence is unsanctioned or illegal either in terms of the land on which they are located or the way they have been constructed. This is irrespective of whether said units suffer from traditional slum attributes such as being underserviced or being overcrowded. It is important to mention “squatter settlements” as another commonly used term, suggesting a more severe type of informal settlements. Squatter settlements are residential areas completely void of de jure status. Squatter residents commonly occupy land such as vacant public land or designated hazard prone areas. The high levels of tenure insecurity and risk associated within squatter settlements means they receive little or no infrastructural investment. In the long-term, this translates into a high level of deprivation common with traditional slum areas (UN-HABITAT, 2003).

1.5.3 Other slum synonyms

The terms “slum”, “informal” or “squatter” are the most widely used in academic literature and in English speaking countries. However, they do not exhaust the nomenclature associated with
the slum phenomenon. Another commonly used term for squatter settlement is the “shanty town” which conjures up a mental picture of a congested area in an urban periphery, where people mostly live in small shacks constructed with very crude materials. Other, mostly vernacular, terms have been deployed to describe largely similar though not always identical settlement conditions in different local and regional settings. In predominantly Francophone countries, the terms “bidonville” or “taudis” are commonly used to describe urban slums (Sietchiping, 2004). “Barrios” is a generic term for neighborhoods or municipal divisions across the Spanish speaking world (Siembieda & Moreno, 1998). However, it has also come to specifically mean peripheral squatter settlement in cities like Santiago, Chile and Caracas, Venezuela (Handelman, 1975). Other country-specific terms include “favelas” (Brazil) and “Bustee” (India) (UN-HABITAT, 2003).

1.6 Organization of chapters

This dissertation is divided into seven chapters. Chapter One is an introduction to the study and begins with background information on the topic of slums. This is followed by a statement of the research problem and a presentation of objectives and associated research questions. The chapter also includes a description of the study area and definitions of key terms.

Chapter Two provides a review of relevant literature. It establishes the historical and theoretical foundations of the slum phenomenon. This is followed by an examination of slum household and physical indicators. A review of recent literature on slum mapping and associated methodologies is also included in this chapter, which concludes with the identification of gaps in literature to be addressed by research objectives.

Chapter Three focuses entirely on methodology. It provides a detailed description of all methodological steps undertaken in this study: data collection and processing, derivation of analysis units and variables, and analyses. For each research objective, the set of analysis tools
chosen is described in detail with embedded justification.

Chapters Four, Five and Six present and discuss the findings of carrying out the first, second and third main objectives respectively. Chapter Four concerns the exploration of a multi-index approach to measure and map local slum conditions. Chapter Five focuses on the use of unsupervised learning or classification for determining slum conditions from satellite image data. Chapter Six looks at the use of analytical regionalization in determining optimal spatial units for slum mapping.

Chapter Seven provides a conclusion to the study. This includes a recant of objectives and main findings and a discussion on the broader implication of the study, particularly for policy. A statement on the limitations of the study and the avenues for future research concludes the chapter.
2. A REVIEW OF RELATED LITERATURE

2.1 Chapter overview

This chapter provides relevant theoretical background to this dissertation. This includes an examination of theories about the origins and causes of the slum phenomenon in both Western and Developing World contexts. The chapter also sheds light on the multidimensional nature of slums and how they are captured through a variety of physical and non-physical indicators. This is followed by a review of recent slum mapping literature as well as a look at potential methodological strategies that can be used to optimize local slum mapping. The chapter ends by identifying gaps in the literature to be addressed by the study.

2.2. Historical and theoretical foundations of the slum phenomenon

2.2.1 The slum problem as seen by the West

The slum phenomenon as we know it today can trace its roots to the Victorian Era in Europe and North America. Evidence of urban squalor from this period is found in literature published at the time (Engels & Kelley, 1887; London, 1904; Riis, 1901). “The People of the Abyss” by investigative journalist Jack London is a good example. This was a documentation of his firsthand experiences in the famous “East End” slums of London (London, 1904). Even fictional work from the same era bore testament to the widespread occurrence of slums. In this regard, there is no better example than perhaps the most famous Victorian writer, Charles Dickens. In his serialized long novel, “Bleak House,” Dickens had the following to say about such a locale:

“It is a black, dilapidated street, avoided by all decent people...these ruined shelters have
bred a crowd of foul existence that crawls in and out of gaps in walls and boards...”
(Dickens, 2016, ch 16, par 8).

There are many factors at the root of slums in the West, although one main driver, namely industrialization, is dominant. The Industrial Revolution that engulfed Western Europe between the 18th and 19th centuries, and the economic growth that came with it, instigated unprecedented levels of migration to urban areas. Over time this led to congestion in the cities and hence increased demand for housing. As wages remained low especially among unskilled workers, more people could not afford better quality housing and were, therefore restricted to declining areas of the city (Dyos & Dyos, 1982; Haw, 1899; London, 1904).

A similar situation unfolded in North America especially in cities like Boston, New York and Chicago. By the mid-19th century, “The Five Points” neighborhood in Lower Manhattan was already an established slum with a global reputation. Again, high levels of migration and population growth, driven by industry and commerce, were clearly the root cause. “Five Points” in particular owed its origins to the establishment of “slaughterhouses” and “tanneries” in the vicinity of the now non-existent ‘Collect Pond’ (Anbinder, 2001).

One thing consistent with slums from this era is the social stigma that was attached to the phenomenon at the time. Slums were not just viewed as areas suffering from poverty and housing deprivation. They were also the domains of the most undesirable ills of society such as crime, violence and immorality (Dyos, 1967).

By the mid-20th century, slums had all but disappeared in Europe and North America. However, that is not to say that housing deprivation does not exist in these regions today. Perhaps the closest thing that can be likened to a slum in the West today is the “colonias” that exist along the United States - Mexico Border areas. These are unregulated Hispanic-dominated settlements
found in or near border towns in Texas and New Mexico. Once again, the origins lie predominantly in economically driven migration. Decades of cross-border migration and the rise of industries in these areas created border settlements that essentially served as transportation and trade hubs. Income differential, especially between the north and south sides of the border meanwhile drove migration to these areas. The result has been the establishment of unregulated semi-rural settlements. These settlements suffer from problems such as poor building materials, lack of potable water and sanitation facilities, and the lack of roads (Dabir, 2001; Donelson & Esparza, 2010).

2.2.2 The slum problem in developing countries

By the mid-point of the 20th century, slum phenomenon was a problem mostly associated with urban centers in the global South. Within the developing world, Latin America was perhaps the earliest pacesetter and helped inform contemporary slum theory. Interestingly, Latin America has also witnessed the greatest reduction of slum incidence and today, other regions like South/South East Asia and sub-Saharan Africa have taken over the relay baton (UN-HABITAT, 2013). The following sub-sections shed light on different though often overlapping theoretical perspectives on the origins and continued prevalence of slums in developing countries.

2.2.2.1 The slum as a domain of “hope” or “despair”

One of the earliest theories that explain the occurrence of slums in developing countries was proposed by Charles Stokes in his work "A Theory of Slums" (Stokes, 1962). Using a simple model with examples from Latin American cities such as Lima and Caracas, Stokes broadly identified two types of settlements based on the psycho-socioeconomic aspects of the residents. The first category, ‘slums of hope,” was coined in poignant reference to areas of the city settled by new urban migrants. Residents in these areas adopted a positive psychological attitude filled with
expectation to eventually progress up the socio-economic ladder (Stokes, 1962). Under ideal circumstances, residence in the “slums of hope” will be temporary, with settlers moving on to integrate with formal parts of the city as their economic and social disposition improves. “Slums of hope” depended on the presence of an “escalator class” of people who possessed the “ability” (e.g. through education and training) to make the upward climb. Many newly formed slums and informal settlements sprouting in the cities of developing countries today can be considered as prime examples (Magigi & Drescher, 2010).

The second category of Stokes’ model offered a diametrically opposite perspective. In this case, slums are regarded centers of “despair.” The residents often lacked the "ability" (i.e. required education or skills) and hence the financial means to escape the slum (Stokes, 1962; UN-HABITAT, 2003). This may be compounded by broader issues such poor macro-economic performance and failure of institutional/government housing policy (Fox, 2014). The result was an extended and possibly indefinite stay in slum areas. It is instructive to note that “slum of despair” typically arose from “slums of hope,” marked by a significant change in attitude from hopeful to bleak as residents became resigned to the status quo.

2.2.2.2 The slum as a byproduct of modernization

In his seminal work, “Freedom to Build,” John F.C. Turner, a prominent 20th century British architect and urbanist, discussed urban practices and policies, including the problem of uncontrolled urban settlements in places such as Latin America. Turner offered a more sanguine explanation of the slum problematic, one that was infused with modernization theory. He viewed uncontrolled urban growth and resultant informal settlements as a necessary and unavoidable by-product of urbanization and modernization in developing countries. However, he was also critical of attempts to stifle the growth of unplanned settlements through to the top down imposition of
rigid Western style housing regulations. Such measures were to be viewed as counter-productive and only serving to keep the housing market beyond the means of the majority urban poor. Turner coined the phrase “housing as a verb” to suggest that it was more efficient for housing to be promoted as a locally originated flexible activity rather than as an inflexible object or “noun” that must meet certain preconceived minimum standards. As such, unplanned and unregulated urban growth was not to be viewed as a problem but as an opportunity for grassroots-engineered solutions to housing problems. The negative aspects associated with traditional slums could be kept at bay by guiding the development path of these areas through the promotion of incremental self-improvement as well as providing external support through service upgrading. To this effect, a settlement typology was needed that reflected not only the current state but the direction/trend of the settlement development. Within this framework, settlements could be classified as incipient, semi legal, provisional and their trends could be labeled as improving, stagnating or deteriorating (Turner, 1969; Turner & Fichter, 1972a).

2.2.2.3 The slum as a product of “disjointed” modernization

Modernization ideas like Turner’s might have had profound influences on housing policy, particularly in regions like SSA. Notable among these was the widespread adoption of the policy of benign neglect or “laissez-faire” (Arimah, 2010; Sietchiping, 2004). This policy meant that authorities initially turned a blind eye to the problem of slums and informal settlements in the hope that they would disappear in time aided by sustained economic growth and prosperity that was supposed to come with modernization. Fox (2014) however challenged this theory and instead coined the term “disjointed modernization” as an explanation of the variation of slum incidence across countries in SSA. He empirically demonstrated using regression analysis that a combination of population growth, colonialism, failed economics and institutional policy all contributed to the
present slum situation. Therefore, although these factors had individual merits, they did not, on their own, suffice as explanations.

As demonstrated with Western urbanization and slums, economically driven rural-urban migration is a key process which feeds into population increase in cities. However, using contrasting examples of Phoenix (Arizona, USA) and Accra (Ghana), Fox (2014) resisted the tendency of focusing on population growth as the sole cause for slums and informal settlements. Population increase had to be interpreted in the broader context of economic growth and development. In regions like SSA, migration towards cities is driven more by negative conditions that compel people to move out of rural areas even when such movement is not backed by sufficient evidence of economic prosperity in the city. Conflict, the looming threat of climate change and environmental degradation are common examples of “push” factors that have directly impacted the livelihood of rural residents notably in the agricultural sector (Hare, 1999).

The macro-economic environment is also an obvious driver of urbanization and hence a slum causing factor. Although Western slums were ultimately associated with poverty and deprivation, it was economic prosperity at the macro-economic level that was at the root. In regions like SSA, however, the process is slightly different. It is the limited (or lack of) economic growth at the macro-level that is responsible for slum prevalence. Empirical evidence of this was offered by Arimah (2010) who undertook a study to explain the variation in slum incidence across countries in 3 different regional settings: Africa, Asia and Latin America. The macroeconomic environment was a key component of the model operationalized through a series of predictor variables such as GDP per capita, GDP growth, level of indebtedness and Gini (inequality) index. GDP per capita could be seen in particular to have profound negative relationship with slum prevalence across all three regions (Arimah, 2010).
Another component of the “failed modernization” theory by Fox (2014) is the impact of colonialism on sub-Saharan urban processes and hence slum development. Most of today’s urban areas in SSA have colonial foundations. To put it in the words of Obudho and Aduwo (1989), urban centers “were created to serve the interest of the expatriate Europeans.” There are several mechanisms by which colonial policies may have influenced the development process of cities in SSA. One can consider, for example, the segregation/exclusion policies which barred the colonial subjects (natives) from participating in urban life. This was commonplace in East Africa where towns, for example, were seen primarily as areas of European and Asian activity. Native Africans, who were for the most part temporary migrant laborers, were forced to settle on the urban fringes, separated from the well planned urban life (Gutkind, 1960). Another mechanism to consider is the continuation of the centralized top down colonial era policies and laws by African leaders and elite after independence. Some of these laws, such as those related to land acquisition and tenure, are still at odds with local customs and practices (Fox, 2014).

It is important to note that not all evidence is unanimous in the indictment of colonialism as promoter of slum incidence in SSA. In fact, a very recent study by Njoh (2015) proved just the opposite. In the study, the intensity of colonialism was examined as an explanatory factor for cross-country slum incidence in SSA. Aspects of colonialism that were considered included “duration, extent of colonial investment, extent of colonially-induced urbanization and administrative strategy” (Njoh, 2015, p. 108). The study showed that more intense colonization rather resulted in lowering overall slum incidence. As Njoh contends, a possible explanation for this is the fact that more intensely colonized areas received significantly more investment in infrastructure as the Europeans sought to create urban areas comfortable enough by their own standard. Therefore, although natives were excluded from urban life during the colonial era, the infrastructure was there
to be inherited and benefited from in the post-colonial era.

2.3 A review of slum indicators

2.3.1 A review of household indicators

The definition of slums by the UN emphasized measurable attributes such as physical and legal characteristics of homes, as well as whether they had access to basic services. Under their criteria, a slum was to be considered as a settlement area in which fifty percent of the households were lacking in any of five specific indicators: access to water, sanitation, living space, durable structure and secure tenure status (UN-HABITAT, 2003). These indicators are described in detail in Table 2.1.

Adequate access to clean water and safe sanitation are important indicators of the quality of human life, backed by their proven correlation with issues such as health and mortality (Hanchett et al., 2003). A significant reason for high child mortality rates in developing countries is because of diseases such as diarrhea and cholera which are directly linked with unsafe water, poor sanitation and hygiene (Isunju, Schwartz, Schouten, Johnson, & van Dijk, 2011). Level of accessibility and cost of water both help to determine “access”. For example, despite being essentially the same commodity, piped water delivered conveniently through residential plumbing has completely different access level compared to that provided by community “stand pipes.” Similarly cost of water can range from almost free in the case of public funded community projects to being unbearably costly when said water resources are subject to privatization and market economics, even in the informal sector. Recent research in the areas of development and informal settlements has established that a participatory approach may sometimes yield strong positive outcomes that balance cost, durability and quality of water supply (Ananga, Njoh, Anchang, & Akiwumi, 2016; Njoh, 2006).
Table 2.1 UN-HABITAT designated slum indicators and their recommended benchmarks

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>Household members must have sufficient access to clean water for domestic use at affordable price and without extreme effort in the collection process</td>
<td>At least 20 liters of water per individual per day acquired at no more than 10% household income. Piped connection, bore-hole or rain are acceptable sources</td>
</tr>
<tr>
<td>Sanitation</td>
<td>Household members must have sufficient access to a proper human waste disposal system</td>
<td>Toilets can be both internal or external. If external must be well constructed and be shared between a reasonable number of households. This includes improved pit latrines</td>
</tr>
<tr>
<td>Living space/overcrowding</td>
<td>Household members must have access to sufficient living space, measured in terms of average number of people using a single bedroom</td>
<td>Not more than 2 persons occupying a single bedroom in a home</td>
</tr>
<tr>
<td>Tenure security</td>
<td>Household must have a secure tenure status that offers sufficient legal protection from eviction</td>
<td>Possession of de jure legal documents such as title deeds and building permit. Otherwise, possession of documents that show at least de facto protection from eviction by local authorities. In case of non-owners, possession of legally binding tenancy agreement or document authorizing occupation of unit by legal owner</td>
</tr>
<tr>
<td>Durability of the dwelling</td>
<td>The dwelling unit must be constructed using durable permanent materials and located on non-hazardous land</td>
<td>Permanent wall materials include stone, cement blocks, bricks, tiles. Roof materials include bricks, tiles, corrugated metal. Buildings cannot be located near highways, landfills, unstable slopes, flood zones, etc.</td>
</tr>
</tbody>
</table>

(Arimah, 2010; UN-HABITAT, 2003)

When it comes to the evaluating progress towards achieving Millennium Development Goals (MDGs), sanitation is one area that lags significantly behind. One aspect of concern in regions like SSA is that sanitation facilities are frequently shared, an unfortunate consequence of multi-household units (Tumwebaze, Orach, Niwagaba, Luthi, & Mosler, 2013).

Acceptable sanitation facilities range from in-house modern flush toilet systems with
proper sewage disposal to outside pit latrines. The latter are normally a feature of the sub-Saharan rural household but are increasingly common in urban areas which are largely underserviced or even cut off from central water and sewage networks. In well serviced areas, latrines may provide a backup option in cases of disrepair or outages. However, whether a pit latrine is safe or sufficient depends on several factors: how it is constructed, where it is located and how many households share it (UN-HABITAT, 2003).

Security (or insecurity) of tenure is an indicator that has become very important in how we define a slum household or neighborhood today. It essentially measures the legal rights of an individual or household to occupy a unit or land on which that unit is built. As such, it is the defining trait of modern day informal/squatter settlements which are generally considered to be of illegal origins. Taking into account the widespread application of slum clearance policies in recent decades (Arimah, 2010), it is understandable why security of tenure has become such a key issue in the slum discourse.

It is impossible to discuss security of tenure without the context of land tenure laws and systems. In regions like SSA, it is important to distinguish *de jure* and *de facto* land tenure (Durand-Lasserve & Royston, 2002b). The former refers to what is considered legal by the State and legal authorities (e.g. possession of a land title) while the latter recognizes the customary and local claims to tenure (e.g. through inheritance, private conveyance of deed, or locally brokered land transactions). Informal land transactions may benefit from *de facto* recognition and depending on local enforcement policy, that may represent sufficient security.

Tenure security is perhaps the most challenging slum indictor to measure. Unlike with the other indicators, it is difficult to ascribe specific benchmarks to it. This is because tenure security is best understood as a fuzzy continuum rather than a simple categorization (Durand-Lasserve &
Royston, 2002a; Van Gelder, 2009). It can range from most insecure status, such as illegally squatting on environmentally sensitive or public land, to the most secure dweller who possesses a full complement of legal owner or renter documents. Along this continuum may arise situations with different levels of security (Durand-Lasserre & Royston, 2002a). For example, *ceteris paribus*, an owner will typically enjoy more security than a tenant. Similarly, a tenant will have more security if they possess a legally binding tenancy agreement than if they don’t, or if they have just an informal agreement. Ultimately, protection from eviction depends heavily on the local enforcement attitudes and laws. In the light of the difficulties associated with measuring tenure, thus, some studies have chosen to employ indirect proxies such as tenancy status (i.e., owner/renter) to represent security of tenure (Getis, 2015; Weeks et al., 2007).

2.3.2 The theoretical links between household slum indicators

Gulyani and Bassett (2010) used a 4-dimensional construct ("the living conditions diamond") as a holistic mechanism to define modern slum theory. The four dimensions comprised of tenure, infrastructure, unit quality, and location/neighborhood. Each dimension could be further broken down into sub levels. Tenure for example not only embodied legal protection from eviction but was also characterized by issues such as owner/tenant composition and duration of residence. Infrastructure was not only about by whether residents had access to a service but also how reliable and affordable the service was. Unit quality reflected both building materials and overcrowded conditions and location/neighborhood embodied both the physical and social environment of the area. The theoretical links between the different dimensions are obvious. For example, tenure security has an impact on infrastructure as illegal areas are less likely to be serviced. Location on a hazardous site will also attract less investment in infrastructure due to high risks of eviction. There is also evidence of intersection between overcrowding (tenure) and sanitation
(infrastructure) conditions. The same type of sanitation facility, such as a pit latrine, may prove sufficient or insufficient depending on how many households actively share it (Gulyani & Bassett, 2010).

2.3.3 A review of physical settlement indicators

So far, our definition of slums has focused solely on individual household attributes. However, slums also have a “physical face.” In fact, the physical appearance of an area is more frequently used than not as evidence by both experts and lay people to quickly ascertain if it represents a likely slum or not.

The physical characterization of slum settlements has its basis in urban land use/cover classification. Ridd (1995) described a model which asserts that urban surfaces can be classified into one of three major land cover categories: vegetation, impervious surface and soil (V-I-S). Consequently different land uses such as residential, commercial, and industrial were known to have different compositions of these basic cover types (Ridd, 1995). The VIS concept was applied in a study to determine if slum areas in Accra Ghana could be identified through their land cover compositions (Weeks et al., 2007).

Kohli et al. (2012) suggested three conceptual levels at which the physical characteristics of area can be viewed for slum determination: the object, settlement area and environment. The object level considered individual structures such as buildings and their attributes (e.g. shape, height, size and roof materials). The settlement level covered the wider area and considered attributes such as building density, form and road networks. The environment focused on environmental components of the settlement such as terrain, vegetation and proximity to waste dumps or landfills sites.

There is an overwhelming body of evidence that slums physically differ from non-slum
areas in terms these physical indicators. This has led to attempts to develop a physical slum ontology (Kohli et al., 2012). This is a knowledge based framework which collates information on physical attributes that can be used to generalize slum mapping. Table 2.2 summarizes a selection of physical indicators that can contribute towards such an ontology.

2.4 Quantifying slum conditions: from indicators to indices

The slum phenomenon is multi-dimensional as evidenced by the diversity of indicators that can be empirically measured. However, for practical purposes such as mapping, it is important to be able to capture such complexity in much simpler form. This requires the derivation of composite measures, also known as indices, that represent dimensionally reduced summaries of the original information.

Table 2.2 Key physical slum indicators

<table>
<thead>
<tr>
<th>Conceptual level</th>
<th>Indicator</th>
<th>Conditions in slums and informal areas</th>
<th>Cited work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geomorphology of terrain</td>
<td>Steeper slope, unstable areas, more concave slopes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Settlement level</td>
<td>Texture measures</td>
<td>Higher settlement entropy and contrast, less homogeneity</td>
<td>Stasolla and Gamba, 2007</td>
</tr>
<tr>
<td>Road characteristics</td>
<td>Less elongated, narrower widths, lower per cent of paved road surfaces</td>
<td>Hofmann et al, 2008; Sliuzas et al, 2008</td>
<td></td>
</tr>
<tr>
<td>Building density and roof cover</td>
<td>High building density, high per cent roof coverage</td>
<td>Baltasavias &amp; Mason, 1997;</td>
<td></td>
</tr>
<tr>
<td>Object level</td>
<td>Dwelling characteristics (shape, size, roof materials, placement, road access)</td>
<td>Small average dwelling footprint, simpler shapes (4-sidedness), precarious placement, limited road network access</td>
<td>Baltasavias &amp; Mason, 1997; Hofmann et al, 2008;</td>
</tr>
</tbody>
</table>

Partly adapted from Owen and Wong (2013)

Empirical index construction and use is widespread in the socio-economic domain, as evidenced by the multitude of indices employed at local and international levels for both research
and policy. Some notable examples are the Socio-economic status (SES) index (Stevens & Cho, 1985), the Human Development Index (HDI) (Anand, 1994), and the Poverty Index (Sen, 1976). The methods used to construct indices can range from robust multivariate statistical methods to simple averaging or summation techniques (UN-HABITAT, 2003). The latter is quite common within slum/deprivation mapping literature (Baud et al., 2008; Weeks et al., 2007). Whatever the method used, the result is mostly in the form of a single composite measure that reflects, as much as possible, the variation of original contributing variables across observation units.

A popular multivariate statistical technique used for index construction is principal component analysis (PCA) (Jolliffe, 2002; Vyas & Kumaranayake, 2006). Unlike indices created through simple adding and averaging, each principal component derived through PCA is a weighted composite (linear combination) of all the original variables (using Eigen vectors from correlation matrices). PCA is a descriptive technique used to summarize multidimensional data into fewer dimensions that are laid in orthogonal directions. If used for the creation of a single index, PCA works best when most variables in the data are correlated and exhibit high variance (Vyas & Kumaranayake, 2006).

Exploratory factor analysis (EFA) (Cudeck, 2000) is another multivariate technique commonly used for robust index construction. Unlike PCA, EFA carries an inherent hypothesis about the nature/structure of the data. Specifically, it assumes that there are subsets of highly correlated variables within the data suggesting the presence of unseen or latent factors (Cudeck, 2000). Unlike principal components, the factors derived using EFA are more than just statistical summaries of original variables. These factors must also have theoretical/conceptual validation. While PCA is intuitively useful for creating a single index, the latent factors derived from EFA allow for constructing multiple diverse indices from the same data.
The use of factor analysis within urban research owes its prominence to the discipline of factorial ecology, also called social areal analysis (Shevky & Bell, 1955). This involves integration of factor analysis with (urban) social ecology to understand the distribution of multiple underlying socio-economic conditions across urban spatial units. With advancements in computer technology by the mid-20th century, and the creation of large census data sets, factor analysis came to the fore as the main method for understanding urban spatial structure. The initial focus was on understanding the spatial organization of Western cities, resulting in the Shevky-Bell typology which described cities as being structured based on 3 three separate dimensions: socio-economic status ("social rank"), family conditions (e.g. marital status, housewives, having children) and ethnic segregation (Anderson & Bean, 1961; Shevky & Bell, 1955). More recently, and with new information, this typology has been revised with new factors that reflect contemporary urban realities (UN-HABITAT, 2003).

2.5 A review of key slum mapping literature

2.5.1 Studies using field data

A traditionally reliable approach to mapping urban deprivation involves the direct measurement of quantitative indicators using field data obtained through a census or survey. Such data can be used to generate slum indices directly or related proxies such as poverty or deprivation indices. These can then be integrated into a GIS database for mapping. A good example is by Baud et al. (2008) who used a 'multiple deprivation index' derived from census data to map urban poverty at electoral ward level in Delhi, India. Central to the study was the recognition of multiple dimensions of deprivation under the "livelihood's" approach, where each dimension reflected a different form of capital (human, financial, physical, social, natural). Among the key questions they asked was whether different hotspots emerged because of concentration of the different
deprivation categories and which dimensions were contributing most at the level of each census unit. Notable results indicated there was diversity in terms of the underlying factors creating poverty hotspots which were not always restricted to known slum areas (Baud et al., 2008).

Building on previous work, Baud et al. (2009) compared the spatial patterns of deprivation across three Indian mega-cities: Delhi, Mumbai and Chennai. They examined both the intra and intercity variation in poverty hotspot distribution determined using multiple deprivation index. A key result was that while cities were similar in mean deprivation, they did vary in the distribution. For example, in Chennai, "outliers" indicating presence of extreme deprivation were observed.

Using Rosario, Argentina as a case study, Martinez (2009) explored the mapping of poverty and deprivation in the form of intra-urban inequality. The study combined both census and administrative data to derive GIS-based indicators useful for mapping different dimensions of needs. One aspect of the methodology was the use of gap analysis to highlight the difference between highly unequal units. The study also used different spatial aggregation levels such as the city, district, and neighborhood to understand the effect of scale (Martínez, 2009).

In the sub-Saharan region, Weeks et al. (2007) developed an index to map slum conditions in Accra, Ghana, using census data. Basing on a 5-point UN criteria for defining slums, they developed a quantitative housing quality index for each census enumeration area and mapped slum conditions for the entire city.

Quantitative field data allows for objectivity in measurement and mapping of slum conditions, especially using indices. This is especially useful for studies that cover large geographic extents. However, the use of such data assumes generalizability and adherence of the phenomenon of interest to predetermined theory. This is evidenced by the pre-selection of what indicators to observe and measure. This is problematic as slums do not manifest in the same way
in all local situations.

The above studies mostly relied on quantitative information to develop measures for mapping. Other studies have included qualitative information, such as information obtained from expert interviews or through the participatory input of local citizens. Qualitative data when integrated with a GIS can be used for understanding socio-economic, demographic and physical characteristics of local settlements including slums (Joshi, Sen, & Hobson, 2002; Karanja, 2010). A good example is the study by Martinez (2009) who used expert interviews to investigate the relevance of pre-selected indicators for mapping inequality in Rosario, Argentina. They were able to determine that overcrowding, education, employment, water connection and accessibility to schools were the most important as perceived by local policy makers (Martínez, 2009).

In their attempt to characterize and monitor slums in Addis Ababa, Ethiopia, Lemma et al. (2006) employed a rapid appraisal technique that integrated local participatory knowledge with GIS. They used focus group discussions at different city levels to establish locally relevant slum indicators based the knowledge of local experts/professionals. They also complimented this with direct field observation of physical environment of slum areas to validate knowledge gained from experts and to identify areas of poor water supply. Through focus groups, for example, they were able to discover that poor sanitation, poor structural housing quality and low accessibility were the dominant slum indicators in Addis Ababa (Lemma et al., 2006).

In general, field data (both qualitative and quantitative) provide certain critical advantages for the mapping of slum settlements, especially when obtained at the household (micro) level. This helps to mitigate the ecological fallacy problem (Piantadosi, Byar, & Green, 1988). When integrated with GIS, field data can allow for testing independent hypotheses about the spatial distribution of slums and related phenomena.
The above studies show that qualitative data provides crucial local context and depth to slum analysis and can be used to fill knowledge gaps. However, the subjectivity associated with them indicates they may suffer from reliability and hence have limited use outside of local area or at different times. As such, they will be difficult to fit in an operational framework that not only maps but monitors slums across large extents.

2.5.2 Studies using remote sensing data (imagery)

Aerial photographs have been for urban social studies since the mid-20th century (Green, 1956; Monier & Green, 1957). The underlying hypothesis to this is the belief that human behavior shapes, and is shaped by, the physical environment in which they live (Rashed et al., 2005). High resolution satellite imagery can now provide detailed repetitive and unbiased information on the urban physical environment (Jensen & Cowen, 1999; Patino & Duque, 2012).

Image-based slum mapping commonly relies on the physical distinctiveness of slum areas. This distinction can be made using identifiable physical settlements attributes such as building density (Mason & Baltsavias, 1997), texture (Kit et al., 2012), vegetation (Weeks et al., 2007). Details of these indicators and relevant cited works have been provided elsewhere in this report (see section 2.3.3).

From a methodological point of view, settlement information can be extracted from imagery using both qualitative and quantitative image analysis methods. Qualitative image analysis entails visual interpretation techniques. This is often done by domain experts to identify slum areas. For example, to supplement their field data collection, Lemma et al. (2006) used manual interpretation of imagery to build additional knowledge on slums in Addis Ababa. They focused on the irregularity of settlement, the size of buildings, the road networks and lack of open space. This was used to fill in knowledge gaps that could not be filled with focus groups and other
field methods (Lemma et al., 2006).

Angeles et al. (2009) embarked on mapping slums in Bangladesh as part of an exercise to develop a proper sampling frame for studying slum-related health issues. They used visual assessment of IKONOS imagery of different cities in the country to predetermine likely slum locations, focusing on building density and roof materials. Their use of imagery was thus an efficient and less costly initial step to identify locations that will be subject to subsequent field validation through street level surveys. With visual assessment of imagery, they were able to correctly capture 70% of slum locales, with the error largely due to time lapse between image collection and field validation, and the lack of consistency in physical appearance within and between cities (Angeles et al., 2009).

Quantitative image based slum analysis employs algorithmic image classification and derivation of land use/cover thematic maps, from which settlement attributes can be quantified. A commonly used technique is object based image classification which focuses on the extraction and classification of "image objects" rather than pixels. Image objects represent locally homogenous regions on an image for which additional non-spectral attributes (such as shape, texture, context) can be quantified (Blaschke, 2010). At their most primitive level, image objects represent real world objects such as individual road fragments, tree crowns, roof sections, amongst others. This makes them ideal for slum mapping at least as the object which can then scaled up hierarchically to settlement area levels (Hofmann et al., 2008; Kohli et al., 2012).

2.5.3 Hybrid studies that combine image and non-image data

It would be ideal to map slums using only image based indicators, to fully benefit from the advantages of Remote sensing technology and avoid the challenges associated with field work. That is why there have been efforts geared towards the development of a physical slum ontology
based solely on imaged-derived characteristics (Hofmann et al., 2008; Kohli et al., 2012). However, image-based slum mapping often needs to be complimented with non-image data that is directly linked to the theoretical slum definition or a closely related proxy (e.g. socio-economic deprivation).

Niebergall et al. (2008) used an integrative analysis approach by combining information derived from object oriented classification of Quickbird data and socio economic survey data. Using associations between both domains, they could determine measures such as population density, average building size and water consumption per person. These were then used to identifying vulnerable areas in Delhi (Niebergall, Loew, & Mauser, 2008). Following up on their previous work (Baud et al., 2008; Baud et al., 2009), Baud et al. (2010) investigated the potential of combining image data, 'ground truthing' and local knowledge for detecting different types of low quality residential areas in Delhi, India (Baud, Kuffer, Pfeffer, Sliuzas, & Karuppannan, 2010). Both Stow et al. (2013) and Stoler et al. (2012) managed to find significant correlation between land cover information with a slum index in Accra Ghana. In the case of the latter, they utilized the normalized difference vegetation index (NDVI) for vegetation extraction from both ASTER and Quickbird data, which was significantly correlated with slum index at the level of each enumeration unit in Accra, Ghana (Stoler et al., 2012; Stow et al., 2013).

In a recent study conducted by Owen and Wong (2013) in selected neighborhoods in Guatemala City, Guatemala, 26 image-derived geospatial measures were evaluated for their ability to distinguish formal from informal settlements. The determination of informal/formal status was done a priori using local knowledge. By applying a discriminant function analysis and classification by regression trees, they found that entropy of road structure, amount of vegetation, degree of asphalt road surface and percentage of bare soil were the most significant in
differentiating formal and informal neighborhoods (Owen & Wong, 2013).

### 2.6 Machine learning as a potential tool for predicting slum conditions

Machine learning is a sub-field in Artificial Intelligence. In simple terms, it refers to the ability of computer programs to acquire information (i.e. ‘learn’) from existing data. Mitchell (1997) formally defines it as a process whereby a computer program “learns from experience with respect to some class of tasks and performance measure, with its performance improving with the said experience” (Mitchell, 1997). Through the acquisition of knowledge or experience (i.e. from input data), this system can perform tasks such as prediction, data mining, visualization, and construction of the expert systems for decision making.

There are two basic machine learning philosophies: supervised and unsupervised learning. In supervised learning (SL), the system requires input data to have prior labels (i.e. there must be a clear response variable). Examples of SL algorithms include regression models, artificial neural networks, classification and regression trees (CART), and support vector machines (Kotsiantis et al., 2007). SL has been used in slum and informal settlement research, such as using CART to predict whether a settlement was informal or formal based on image derived indicators (Graesser et al., 2012; Owen & Wong, 2013).

Unsupervised learning (UL) on the other hand, allows for knowledge acquisition in the absence of a clearly stated response variable. In the context of slums, such information may be non-existent or difficult to obtain in specific local contexts. UL includes the use of common clustering algorithms (e.g. iso-clustering, k-means, hierarchical clustering) whereby observations are grouped into homogenous clusters based on a similarity measure (Barlow, 1989). This is very useful in cases where defining a response variable is problematic. Hierarchical clustering in particular is a deterministic approach that allows for cluster merging or partitioning at different
hierarchical levels giving greater flexibility for controlling the number and characteristics of clusters (Jain, Murty, & Flynn, 1999). UL is mainly used for computer visualization, pattern recognition and data mining while supervised learning lends itself more to predictive modeling.

2.7 Analytic regionalization to derive urban spatial units

The UN definition of slums emphasizes the aspect of contiguity (UN-HABITAT, 2003). The means a slum must be an unbroken areal unit within the city that is homogenous with respect to some form of housing deprivation. The combined aspects of contiguity and homogeneity clearly suggest the idea of the neighborhood unit as the most suitable platform for mapping slum conditions.

The neighborhood as a basic spatial unit of the city is a foundational concept urban planning. The generalized "neighborhood unit" idea was proposed by Clarence Perry in 1929. It described the neighborhood as an urban spatial unit with defined boundaries and population size. Residents within a neighborhood would share a common dependence on the same centrally located facilities (educational, religious, commercial and recreational) (Patricios, 2002)

This conceptualization of the neighborhood obviously is not applicable in the context of present unplanned settlements and slums in non-Western settings. Sub-Saharan city neighborhoods are typically not the product of judicious urban planning. Mostly they are a sections of the city that have evolved organically overtime to achieve homogeneity through forces such as culture and migration

A more common type of neighborhood in SSA is the "vernacular neighborhoods" (Getis, 2015; Weeks, Getis, Hill, Agyei-Mensah, & Rain, 2010; Weeks et al., 2012). These are neighborhood units derived essentially from local place knowledge. A vernacular neighborhood would typically have unifying ethnic or some other socio-cultural feature as well as an easily
recognizable local name (Weeks et al., 2010). Despite the practical convenience of using them, however, vernacular neighborhoods can be problematic in that they often lack clear formal or official boundaries. Even among locals, knowledge of the neighborhood may not extend beyond the name and general geographic area. The factors that define a particular vernacular neighborhood (e.g., ethnic homogeneity) may change with time as dynamic forces continue to alter the urban fabric (Getis, 2015). In such a case, the original neighborhood becomes obsolete as a mapping unit.

In the face of such challenges, the alternative of deriving neighborhood units objectively and quantitatively has been greatly explored. One approach that has found success is the use of analytic regionalization (AR) techniques (Duque, Ramos, & Suriñach, 2007; Duque, Royuela, & Noreña, 2012b). AR is derived from regionalization, a classical approach to studying geographic phenomena (Haggett, 1967). The general concept is that large geographic space can be divided up (or that smaller geographic units can be combined) to form contiguous areal units (or regions), such that each is homogenous with respect to some attribute (or sets of attributes) while being quite different from the neighboring regions (Figure 2.1).

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.1 Simple illustration of regionalization (4 regions from 16 units)

AR specifically employs an algorithmic approach to achieve region creation, by optimizing an "objective function. “This function is usually the homogeneity of the region with respect to an
attribute of interest (Openshaw & Baxter, 1977). While regionalization algorithms differ greatly in their mathematical/technical details, the typical framework is as follows: beginning with smallest available geographic subdivision of an area, areal units with attribute measurements are subjected to some form of spatially constrained clustering. This is clustering that uses neighborhood relations such as contiguity or connectivity in addition to attribute values to determine whether units belong in the same larger group or region. The region forming process is iterative and continues until the maximum (or at least a threshold) of intra-region homogeneity and inter-region disparity is achieved. Some of the prominent algorithms developed to date include the automated zoning procedure (AZP) (Openshaw & Baxter, 1977), spatial clustering based on the minimum spanning tree (Assunção, Neves, Câmara, & da Costa Freitas, 2006), a multidirectional optimal ecotope-based algorithm (AMOEBA) (Aldstadt & Getis, 2006), local spatial heteroscedasticity (LOSH) (Ord & Getis, 2012), and the max-p regions solution (Duque, Anselin, & Rey, 2012a).

AR leads to the creation of “intelligent” or "smart" geographic units rather than arbitrary ones. If combined with the proper choice of relevant socio-economic, cultural and physical attributes, it can be used for a rigorous determination of urban structure. The optimized homogeneity function means a derived region is a statistically accurate simulation of the theoretical neighborhood. It also means AR helps to mitigate two major problems perennially associated with spatial analysis: aggregation problems (such as modifiable areal unit problem) and spatial dependency problems (such as spatial autocorrelation) (Openshaw, 1978; Openshaw & Openshaw, 1984). With respect to the latter, analytically derived regions provide a platform for proper spatial sampling as any observation within a derived region is not expected to be strongly correlated with those in adjacent regions. Finally, AR often employs the unsupervised learning principle, which is useful for cases where there are no prior labels on neighborhood structure.
(whether vernacular or formal).

When it comes the specific case of slum mapping, a few studies have used AR to determine mapping units. In their efforts to map slums in Accra Ghana, Weeks et al. (2007) employed an analytic regionalization technique developed by Duque Cardona (2004) to redefine over 1700 census enumeration units into a 214 contiguous neighborhoods. The method employed a linear approach, whereby an objective decision was made to attach and grow neighboring units based on a threshold of similarity/dissimilarity. The main variable used for this was a slum index derived by averaging household level measures of slum indicators per original census area. Through this approach, they could tackle spatial aggregation problems associated with the original units by reducing the Moran's I associated with the slum index to a non-significant level.

Duque et al. (2012a) applied the max-p regions technique (developed by Duque et al. (2012a)) as part of multi-step approach to determine the best scale for mapping poverty and slum areas in Medellin, Columbia. One unique aspect of the max-p regions algorithm was that it allowed for internalizing a regional parameter, such as size/shape of derived regions. The authors initially applied spatial smoothing and factor analysis to a set of socio-economic variables. From this, the first four extracted factors alongside a contiguity matrix and minimum size criterion were fed into the algorithm to derive 139 new homogenous regional units from 243 original administrative neighborhoods. The new units were then used to map a slum index derived by aggregating five dummy variables related to UN-Habitat slum criteria. Mapping results showed regions associated with critical housing (structural) conditions were in more peripheral areas while those associated with critical personal/social characteristics were scattered across the city.

One of the most recent studies to apply AR to understand urban structure in SSA was by Getis (2015). This study employed the AMOEBA algorithm (previously developed by Aldstadt
and Getis (2006)) to organically construct neighborhoods from census enumerations areas in Accra Ghana. The AMOeba method employed an iterative decision making process by adding neighboring units to a seed unit or groups of units based on the Getis-Ord \( (G_i) \) (local spatial-autocorrelation) statistic. When the addition of a new contiguously connected unit no longer resulted an increase in \( G_i \), then it was determined that the most homogenous region had been formed from the already aggregated units. The authors used 3 variables as separate singular input for the algorithm: a slum index, a housing quality index and vegetation index.

2.8 Identified gaps in literature and restatement of research objectives

Urban social and environmental research in SSA is generally skeletal. Even more rare are studies involving the spatial analysis/mapping of slums. As determined from the literature, no slum study in the region has thoroughly investigated the multi-dimensional nature of slums in a local context, instead of just simply aggregating slum factors into a single index. The theoretical intersections between different slum indicators is a well-established fact (Gulyani & Bassett, 2010). However, some indicators may not be statistically correlated enough to be lumped together as part of a single combined measure. Thus, there is a need for a multi-index measurement approach to capture the diversity of slum conditions in a local area. The first objective of this study is to address this issue by factor analyzing selected household slum indicators. This would help determine if there are multiple underlying aggregate factors that capture different perspectives of slum conditions within the study area.

As observed in the literature, studies using image data for slum mapping rely on a beforehand specification of the response variable (e.g. slum index/status) in sample data to determine if they can be predicted from image derived attributes. However, this \textit{a priori} information is often unavailable, unreliable, and costly to obtain in many local cases. Unsupervised
learning is a potentially powerful tool for mining local data but its use in image based slum mapping research is scant. As such, the second objective of this study is to explore a more inductive approach in learning slum conditions from imagery; by using unsupervised learning theory to define settlement physical types which could reflect slum status.

Finally, one of the major gaps revealed in the surveyed literature on slum mapping is the general failure to use appropriate non-arbitrary spatial units for mapping slum conditions. This means most slum mapping results are vulnerable to spatial analytic problems like the MAUP. Analytic regionalization has potential to correct this issue and few studies have been able to derive analytic units for mapping using relevant survey or census data available at for the smallest possible sub-city level (Duque et al., 2012a; Weeks et al., 2007). Unfortunately, census data at such tiny geographic scales is rare to come by. The third major objective of this dissertation research is to investigate the potential of using only image derived attributes for deriving suitable analytic units that could serve as proxy neighborhoods for mapping slums.
3. METHODOLOGY

3.1 Chapter overview

The purpose of this chapter is to describe in detail all the methodologies used in this study. This includes a description of all the data and how they were collected, as well as a description of all variables derived from the data. A whole section is devoted to the image analysis procedures used to extract physical settlement information from imagery. Finally, the chapter focuses on each main research objective, describing in detail the specific set of analyses undertaken to realize said objective.

3.2 Data collection

3.2.1 Remote sensing data

This study made use of very high resolution multispectral satellite imagery as the primary remote sensing data source, obtained from the Digital Globe Foundation. Specifically, the study employed a GeoEye image, acquired on December 20, 2012, and covering the entire city of Bamenda and its immediate environs (Figure 2). The image data set is multispectral comprised 4 bands (blue, green, red and near infra-red) with a ground sampling distance (spatial resolution) of 1.62m. In addition, there was a single panchromatic band with a 0.5m ground sampling distance.

The 4 individual bands were corrected for geometric distortion using ENVI software (EXELIS, 2012) with vendor provided Rational Polynomial Coefficients (RPCs) and an ASTER Digital Elevation dataset. The raw brightness values were converted to at-sensor radiance values and input into ENVI's FLAASH atmospheric correction module to derive surface reflectance.
values. Finally, a Gram Schmidt pan sharpening technique was performed in ENVI on the GeoEye image data to enhance the spatial resolution of each of the original 4 multispectral bands using the panchromatic band information. This led to the derivation of sub meter spatial resolution (0.5m). While there are many methods for enhancing the resolution of imagery, the Gram Schmidt method is generally considered more accurate as the algorithm is designed to take into account the different spectral characteristics of specific sensor types (Laben & Brower, 2000). As this study employs multi-sensor data, it seems to be the logical method for image enhancement.

To facilitate further image processing, the pan-sharpened image was resized by clipping to study area extent and resampling to 1m spatial resolution, which is still a more than sufficient resolution for detailed settlement level studies. This significantly cut down processing times of image analysis procedures.

3.2.2 Household data and derivation of spatial units

Household survey data were obtained as secondary data from CAMGIS Plc, a local consultancy firm operating in Bamenda, Cameroon. Under the auspices of a project funded by the Cameroon-European Union Cooperation, CAMGIS has been actively formulating a new Five Year Strategic Development Plan for the city of the Bamenda¹. As part of the project, a detailed household survey was undertaken in the city in the summer of 2013. To facilitate this, and considering the absence of any official neighborhood boundary maps, CAMGIS created a base map of statistical units in Bamenda using the following features:

- Visible natural barriers;
- Major roads and transportation networks;

¹ Project funded by the European Development Fund (Project no. 8 ACP CM 17/8 ACP TPS 052)
• Homogeneity of land use and land development;
• Tentative limits of existing quarters.

(CAMGIS, 2014)

A total of 102 units were originally created (Fig. 3.1A). For this dissertation research, we eliminated units that fell outside of the visible built up extent of the city and those that were dominated by public or commercial land uses. The latter included amongst others: The Catholic Cathedral, the central commercial district, prominent public and private schools. This resulted in the retention of only 63 of the original units to be used as spatial units for quantifying household survey data (Fig. 3.1B). The field survey made use of systematic street level sampling of every 10 household within each statistical unit. This resulted in over 4000 individual households surveyed (CAMGIS, 2014).

For this study, only the most relevant variables related to UN-Habitat slum definition were included. The overwhelming majority of the dwelling units in Bamenda (>95%) are built with relatively durable material and roofed with corrugated metal. Because of this lack of variability, it was counter intuitive to include variables related to durability of the building structure. Also, data on security of tenure (such as having a building permit, land title deed) were not considered reliable enough to be included in the study due to large number of missing responses. Many tenant households in Bamenda have absentee landlords and less than formal tenancy agreements. So, they do not have reliable information on the legality of the original land ownership and building construction. As such the tenancy status (whether owner or renter) was included in this study to serve as a proxy of security of tenure.
3.2.3 Additional GIS data collection and processing

An ASTER-derived 30m resolution Digital Elevation dataset (Figure 4) was downloaded for the study area from the United States Geologic Survey (USGS) using the Earth Explorer tool. This was used both for geometric correction of the GeoEye imagery and to derive terrain/surface GIS variables. The digital elevation dataset was processed using ENVI’s terrain modeling tool to calculate slope and surface curvature (Wood, 1996).

In addition to elevation data, additional GIS data was derived to quantify important attributes related to slum/settlement theory. Specifically, these are locational attributes that reflect important settlement concepts like accessibility and centrality, both of which have been to categorize slums (UN-HABITAT, 2003). Using manual digitization methods, the following GIS vector layers were extracted from the GeoEye Image (Figure 3.2):

- Central Business District or CBD (which also includes the Main Town Market)
- Secondary town markets (Nkwen and Ntarinkon Markets)
- Main Access Road Network (Trunk line and main roads, all paved)

Using ArcMap Spatial Analyst Distance tools (ESRI, 2015), these features were used to create distance rasters (Figure 3.3) in order to quantify location attributes of statistical units. The path distance function was used instead of the Euclidean distance as the former incorporates elevation data to calculate the actual surface distance.
Figure 3.1 Base map of statistical units: A) original units B) selected units (ESRI, 2015)

Figure 3.2 Map showing additional GIS layers: CBD, market places, roads, elevation (ESRI, 2015)
3.3 Image classification to derive land use/cover information

Figure 3.4 is a schematic of the methodology applied to extracting land use/cover information from Geo Eye Image. The purpose was to extract land use/cover information which could be used to quantify physical appearance of statistical units. A stepwise/hierarchical image classification approach similar to that used by Pu and Landry (2012), was adopted. This involved extracting green vegetation cover initially, which was then used as mask for subsequent extraction of other urban surfaces using object based image classification.

3.3.1 Vegetation extraction and masking

In this study, a vegetation index was used to reduce original 4-band image into a single dimension. Vegetation indices enhance the information content of imagery to highlight vegetation areas. They are a normalized 1-dimensional representation of 2 original spectral bands, with higher values associated with vegetated pixels. The Normalized Difference Vegetation Index or NDVI (Rouse et al, 1971) is the most commonly used vegetation index, but for this study we used the
Soil Adjusted Vegetation Index (Mason & Baltsavias), which is designed to account for the attenuating influence of background soil brightness on vegetation spectra (Huete, 1988; Pu, Landry, & Yu, 2011). SAVI is more suitable for urban environments especially in Sub-Saharan Africa where bare soil surfaces are common and influence vegetation spectra.

![Diagram of satellite image processing workflow](image)

Figure 3.4 Satellite image processing workflow

Using the Band Expression tool in ENVI software (EXELIS, 2012), the soil adjusted vegetation index (Mason & Baltsavias) was calculated from the red and near infra-red (NIR) bands of the pan-sharpened GeoEye imagery as follows:

\[
SAVI = \left( \frac{\text{Band 4} - \text{Band 3}}{\text{Band 4} + \text{Band 3} + L} \right) \times (1 + L),
\]

(1)  

Where Band 3 = red channel, Band 4 = near infra-red channel and L = soil brightness factor
(0.5 used in this case). The output was a raster layer with values ranging from -1.5 to +1.5. This was rescaled to a 0 - 1 range using standard deviation stretch method.

A combination of Jenks Natural Breaks (Taubenböck & Kraff) classification and agglomerative hierarchical clustering) was implemented as a novel unsupervised classification technique to derive vegetation class image from SAVI data (Anchang, Ananga, & Pu, 2016). The NB method was used for classification of SAVI data into ten (10) unlabelled classes. The purpose of this was to determine if the natural clusters in SAVI data space can effectively separate vegetation from other land cover. To classify a continuous variable \(x\) into \(k\) number of classes, the NB algorithm randomly selects a set of \(k-1\) values within the range of ordered values of \(x\) and uses them as initial class boundaries. The means for each class and the sum of squared deviations of class values from class means is calculated. The total sum of squared deviations (SSDT) for all classes is noted. Then values are moved to neighbouring classes if they are closer to those class means and thus reducing the overall SSDT. This process continues iteratively until SSDT is at minimum or falls below a threshold value (De Smith, Goodchild, & Longley, 2007). The NB method was used to create 10 unlabelled classes for which class statistics (means and variances) were calculated and stored in a signature file.

In order to derive vegetation/non-vegetation binary layer, the dendrogram tool in ArcMap (ESRI, 2015) was used to produce a hierarchical clustering scheme for merging the classes derived from initial NB classification of SAVI data. The underlying procedure can be explained as follows: First, using class statistics stored in the signature file generated from the NB classification, the Euclidean distance, \(D\), between all possible pairs of initial SAVI classes is calculated,

\[
D = \sqrt{(\mu_m - \mu_n)^2},
\]

where \(\mu_m\) and \(\mu_n\) are the means of any 2 classes (m and n) respectively.
The algorithm then proceeds to merge closest pair for classes (in terms of D). After merging, D is updated between all classes and then the next closest pair is merged. The process continues iteratively until all classes have been merged. The resulting dendrogram shows hierarchy of merged pairs from which any desired number of classes can be obtained. Using the hierarchy displayed in the output dendrogram, two final classes were obtained and labelled intuitively as vegetation/non-vegetation.

3.3.2 Object based extraction of non-vegetation land use/cover classes

After obtaining the vegetation class image, a mask was created for the original multispectral GeoEye image to exclude vegetation from to further image classification. The next classification process was aimed at deriving eight (8) non-vegetated land cover/land use classes: roofs, paved road, dirt road, bare soil, building shadow, tree shadow and non-vegetated/bare fields. Prior extensive visual assessment of the scene data revealed that in addition to vegetation, this was an exhaustive list of land cover/land use types within the study area. This is also in support of the assertion that human settlements essentially comprise buildings, roads and open spaces (Pesaresi, Ehrlich, Gamba, & Herold, 2009). Image classification was done using object based methodology to facilitate the extraction of settlement features from high resolution imagery (Hofmann et al., 2008). This employs not only spectral information but also spatial/contextual information. In total, there were 3 main steps: image segmentation to derive image objects, selection of optimal image objects features and image object classification.

3.3.2.1 Image segmentation

The edge detection segmentation algorithm provided within the ENVI Feature Extraction Module was used (EXELIS, 2012). This algorithm uses the Sobel method (Sobel & Feldman, 1968) which assumes that intensity gradient is highest at the edges. By taking the first derivative
of the image in both x and y direction, the maxima and minima (edge) can be determined.

Five bands were stacked and used for image segmentation: the four original spectral bands (Blue, Green, Red and Near Infra-Red) and a SAVI band. The inclusion of the SAVI layer was not for vegetation extraction but because it was believed it was still useful for separating some other classes, such as bare soil and tilled agricultural fields. The segmentation process was controlled by two key user-determined parameters: the scale parameter and the merge parameter, both of which can take any value from 0 - 100. The scale parameter determines the size of image objects to be created while merge parameters allows for adjacent image objects to be merged based on similarity of their attributes. Using the sample visualization window provided by the software and trying multiple values, it was determined that a scale parameter of 15, a merge parameter of 80, and 5 x 5 filter window for texture, produced good results in terms of segmenting features of interest.

3.3.2.2 Selection of optimal features/attributes

The segmentation process resulted in created of image objects for which 54 image object features (attributes) were calculated automatically by the algorithm (Table 3.1). These include 20 spectral feature (4 for each band), 20 textural features (4 for each band) and 14 spatial/geometric features. This increase in dimensionality necessitated a feature selection step to minimize classification errors.

First, by overlaying image object outlines over original composite image, a training sample of image objects was manually selected. Table 3.2 shows the classes and the number of image objects. Classes were determined by careful and extended visualization of all areas within the image. In order to select optimal attributes for differentiating classes, an ensemble decision tree model, the Extra Trees classifier (Geurts, Ernst, & Wehenkel, 2006) was implemented on the training sample. This was obtained using the python Machine Learning (Scikitlearn) library. This
technique is a non-parametric classification technique that can output feature importance scores which ranks features per their importance in separating classes of interest. Using the mean feature importance score as a cut-off, 19 of the original 54 features were selected and retained for image classification (see Figure 3.5).

Table 3.1 Description of object feature attributes

<table>
<thead>
<tr>
<th>Object Feature categories</th>
<th>Feature label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral features (20 features)</td>
<td>savg_b_n</td>
<td>Mean of band n</td>
</tr>
<tr>
<td></td>
<td>sstd_b_n</td>
<td>Standard deviation of band n</td>
</tr>
<tr>
<td></td>
<td>smin_b_n</td>
<td>Minimum of band n</td>
</tr>
<tr>
<td></td>
<td>smax_b_n</td>
<td>Maximum of band n</td>
</tr>
<tr>
<td>Textural features (20 features)</td>
<td>tavg_b_n</td>
<td>Mean (texture) of band n</td>
</tr>
<tr>
<td></td>
<td>tvar_b_n</td>
<td>Variance (texture) of band n</td>
</tr>
<tr>
<td></td>
<td>tent_b_n</td>
<td>Entropy (texture) of band n</td>
</tr>
<tr>
<td></td>
<td>tran_b_n</td>
<td>Range (texture) of band n</td>
</tr>
<tr>
<td>Spatial/Geometric features (14 features)</td>
<td>fx_area</td>
<td>Object area</td>
</tr>
<tr>
<td></td>
<td>fx_length</td>
<td>Object length</td>
</tr>
<tr>
<td></td>
<td>fx_convex</td>
<td>Object convexity</td>
</tr>
<tr>
<td></td>
<td>fx_solid</td>
<td>Object solidity</td>
</tr>
<tr>
<td></td>
<td>fx_round</td>
<td>Object roundness</td>
</tr>
<tr>
<td></td>
<td>fx_formfac</td>
<td>Object form factor</td>
</tr>
<tr>
<td></td>
<td>fx_elong</td>
<td>Object elongation</td>
</tr>
<tr>
<td></td>
<td>fx_rectfit</td>
<td>Object rectangular fit</td>
</tr>
</tbody>
</table>

n = spectral band number (1 - 4)

Table 3.2 Description of training sample used for object classification

<table>
<thead>
<tr>
<th>Training class</th>
<th>Number of objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>bare soil</td>
<td>114</td>
</tr>
<tr>
<td>Dirt road</td>
<td>150</td>
</tr>
<tr>
<td>fields</td>
<td>100</td>
</tr>
<tr>
<td>paved road</td>
<td>145</td>
</tr>
<tr>
<td>roof</td>
<td>122</td>
</tr>
<tr>
<td>roof (dark)</td>
<td>71</td>
</tr>
<tr>
<td>roof (dust covered)</td>
<td>113</td>
</tr>
<tr>
<td>Shadow (building)</td>
<td>120</td>
</tr>
<tr>
<td>Shadow (other)</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>1015</td>
</tr>
</tbody>
</table>
3.3.2.3 Image object classification

A multiclass support vector machine (SVM) (Cortes & Vapnik, 1995), available within ENVI Feature Extraction module, was trained and used to classify all image objects into land cover classes, using the 19 most important features obtained from feature selection process. SVMs are robust supervised classifiers that have been successfully applied to land cover classification from remote sensing data (Huang, Davis, & Townshend, 2002; Pal & Mather, 2005).

Using training information, a SVM classifies data points into 2 groups by implementing an optimal boundary (or hyperplane) that maximizes the separation between these groups (Mountrakis, Im, & Ogole, 2011). Although nominally a linear classification technique, SVMs can
be made to perform non-linear classification using kernel functions. The Gaussian radial basis function (RBF) (Shmilovici, 2005) was used with a gamma kernel value of 0.03, bias value of 1.00 and a penalty parameter value of 100.

### 3.3.3 Post classification processing and accuracy assessment

Post Classification processing was performed on the classification output image by sieving and merging classes to derive the following six final classes: open space, roofs, dirt road, paved road, bare soil, and building shadow. Merging of classes minimizes confusion between closely related classes, and thereby improves on over all classification accuracy. In this case, open space class was created by merging vegetation and fields while the roof class was created by merging the three types of roof surfaces (bright, dark and dust covered).

The final classification image was exported as a raster layer, followed by accuracy assessment. As the original GeoEye image used in this study had no time-coincident field "ground truth" data, test pixels for accuracy assessment were selected by direct observation of the original pan sharpened image. At 1m spatial resolution, there was enough detail for an experienced image analyst to visually pick out representative samples of the classes of interest.

A combination of steps was taken to minimize any potential bias/errors in creating test samples. This made use of a random sampling technique. The image extent was subdivided into 5 X 5 (i.e. 25) zones and within each zone a set of 3 random points were created, giving a total of 75 points. For each point, a Region of Interest (ROI) polygon was created within the closest patch belonging to one of the six classes of interest (Figure 3.6). The result was 18 ROIs per zone (450 ROIs in total, 75 ROIs per class). Table 6 summarises the ROI test polygons and the constituent number of pixels. Classification accuracy assessment was done by using a confusion (error) matrix. This is a table that reports individual class classification accuracies, as well as the overall accuracy.
It also reports the Kappa coefficient of classification scheme, which is measure of agreement between observed (test) and predicted (classified) samples (Lillesand, Kiefer, & Chipman, 2004).

Figure 3.6 Image subset showing selection of test areas/pixels for accuracy assessment (EXELIS, 2012)

Table 3.3 Description of samples selected for accuracy assessment

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of ROIs</th>
<th>Actual number of pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roofs</td>
<td>75</td>
<td>20,273</td>
</tr>
<tr>
<td>Open Space/vegetation</td>
<td>75</td>
<td>20,082</td>
</tr>
<tr>
<td>Paved road</td>
<td>75</td>
<td>18,670</td>
</tr>
<tr>
<td>Dirt Road</td>
<td>75</td>
<td>10,088</td>
</tr>
<tr>
<td>Bare Soil</td>
<td>75</td>
<td>20,482</td>
</tr>
<tr>
<td>Building Shadow</td>
<td>75</td>
<td>10,162</td>
</tr>
</tbody>
</table>

3.4 Variables used in the study

Based on the theory, and considering the housing and physical attributes of Bamenda, the following variables were extracted for each statistical unit and used to answer the three main research questions:

3.4.1 Household survey variables aggregated per statistical unit

1. percentage of households without access to piped-borne water (both home and external)
2. percentage of tenant households (non-owner)
3. percentage of households without access to flush toilets
4. percentage of households with more than two persons per bedroom
5. percentage of households living in a single room or studio
6. percentage of households without external drainage

3.4.2 Variables derived from image classification
1. Percent roof cover
2. Percent of paved road surface cover
3. Percent of dirt road surface cover
4. Percent of bare surface (non-road) cover
5. Percent of building shadow cover
6. Percent open space cover (green vegetation and fields)

3.4.3 Additional GIS variables
7. Mean Elevation
8. Mean Slope (percent)
9. Mean distance to closest main access road (meters)
10. Mean distance to central commercial district (CBD) (meters)
11. Mean distance to nearest town market (meters)

3.5 Objective I: To explore a multi-index (multi-factor) measurement approach in mapping household slum conditions

The first main objective of this dissertation research was to determine if there were multiple dimensions (factors) that underlay household indicators and if these dimensions reflected independent settlement theory. A sub-objective was to determine if original household slum indicators could be grouped in separate unobserved but meaningful factors that summarize the
variation of indicators across spatial units of analysis. In other words, for $k$ household indicators, are there $n$ latent factors ($1<n<k$) that effectively capture the joint variation of original indicators?

The second sub-objective was to validate this construct by determining the relationship between these latent factors and independent theoretical settlement attributes.

### 3.5.1 Hierarchical variable clustering (HVC) of household indicators

Hierarchical variable clustering (HVC) was used to derive potential latent slum factors from household indicator variables. The application of data clustering to variables (as opposed to observations) has been proposed as a viable and potentially advantageous alternative to traditional exploratory factor analysis (Vigneau & Qannari, 2003).

Using SPSS software (IBM, 2012) (Figure 3.7), agglomerative hierarchical clustering (AHC) algorithm was used to perform HVC, leading to the derivation of homogenous groups of household variables. The algorithm groups data elements (in this case variables) based on their similarity using a bottom up logic generalized as follows:

1. each element is first considered a single member group.
2. the similarity/distance between all possible pairs of groups is determined.
3. The pair with highest similarity is merged.
4. Similarities between all new possible pairs are updated.
5. Steps 3 and 4 are repeated until all the original elements belong to one group

The following 6 aggregated household variables (elements) were subject to the procedure:

- Percentage of households without flush toilet
- Percentage of households without access to pipe-borne water
- Percentage of households who are renters/tenants
- Percentage of households living in a single room or studio
- Percentage of households with more than 2 persons per bedroom
- Percentage of households without external on site drainage

Although these variables already represent well known indicators of the slum phenomenon (using the UN-Habitat 5-point criteria), the objective here was to determine a fewer number, k (where k> 1), of potentially meaningful but diverse factors that describe the slum conditions within the area.

The standard output of a hierarchical clustering procedure is a dendrogram. This shows all the merged cluster pairs with lower levels nested under the higher levels. As the objective, here was variable grouping and not observation grouping, the similarity between groups was determined using Pearson correlation coefficient and the between-groups (average) linkage function. Although metric distance measures (such as Euclidean) are commonly used to measure similarity, correlation is more appropriate for clustering variables as it uses covariance (which is the objective of factor analysis) and does not require elements to be standardized to the same measurement scale.

Some of the advantages of this approach over traditional exploratory factor analysis are as follows:
- Hierarchical clustering is a completely deterministic process that provides full structure of variable associations within the data set.
- A dendrogram is easy to interpret (e.g. using only visual subjective analysis) to determine a potential latent factor.
- The reliability of traditional factor analysis is significantly impacted by data attributes such sample size and number of variables.
- Variable clustering overcomes the orthogonal constraints of traditional PCA/EFA: only a specific subset variables belong or contribute information to a specific latent factor

Figure 3.7 Specified options in SPSS used for hierarchical clustering of variables (IBM, 2012)

It is important to note that HVC was just an exploratory tool used for data mining and visualization. It informed on the correlations between variables and the optimal number of homogenous subgroups into which these variables could be classified. It did not automatically help identify the potential latent factors nor quantify them. To achieve this, a combination of visual assessment of the resulting dendrogram and a subjective conceptual interpretation of the original indicators was used.

Visual interpretation of factors made use of the horizontal axis which represents the scaled similarity/dissimilarity between hierarchies of merged pairs. SPSS outputs this on a unit-less scale of 0 - 25. This allows for easy visual determination of best cluster solutions by observing relative distances between hierarchical merge levels. The best solution (number of clusters) is the one that has maximum dissimilarity between merged pairs. For example, assume we are given 5 hypothetical variables (v1 - v5) and an output dendrogram as shown below (Figure 11). To determine if these indicator measurements can be organized into latent factors, we determine at
what point of the hierarchy \((a - d)\), the dendrogram should be "cut" or "pinched" to produce the optimal number of groups of variable. This will be point which results in the highest dissimilarity (i.e., lack of correlation) between different groups. Using this hypothetical example, this will be around \(d\), which will result in variables being classified into 2 homogenous subgroups: group 1 (made up of \(v1\) and \(v2\)) and group 2 (made up of \(v3\), \(v4\) and \(v5\)) (Figure 3.8).

As each derived cluster represents a set of highly correlated variables, they can serve as an indication of the presence of unseen factors. However, for this to be conclusive, the variables belonging to the same cluster (group) must be associated in meaningful way. If majority of variables within a group share a similar conceptual denomination (i.e. a common theme), this will enable proper labeling of the latent factor that is to be derived. For example, if it can be determined that most members of a group of correlated variables were directly related to access to basic household utilities, then the group can be said to suggest the existence of a "utilities" related factor.

![Figure 3.8 Hypothetical example showing hierarchical clustering of variables](image)

Figure 3.8 Hypothetical example showing hierarchical clustering of variables
3.5.2 Quantifying derived latent factors as multiple slum indices

The previous section explains how potential latent factors were deduced from a set of original indicator measurements using variable clustering. However, unlike traditional factor analysis which produces factors scores as part of a linear regression fitting, variable clustering merely results in the classification of variables into homogenous groups. Each group may be construed as indicative a factor but does not directly quantify these factors for further quantitative analyses.

In this study, for each group of positively correlated variables (i.e., each derived latent factor), a slum binary index (high/low) was created using k-means classification (Hartigan & Wong, 1979). This was applied to the 63 statistical units (cases) using only variables within the suggested factor group. K-means is a popular and efficient clustering algorithm that works by classifying observations based on the Euclidean distance measure. Unlike AHC, K-means is one-step/flat clustering procedure that requires the desired number of output groups to be specified a priori. It then iteratively minimizes the total sum of squared differences within groups and maximizes the separation between groups.

In this study, the variables belonging to each derived factor group were all positively correlated. When K-means classification is applied to only positively correlated variables, it is analogous to clustering a single variable, resulting in groups that could be intuitively labeled as "high" or "low" in terms of the contributing variables. Considering the nature of original variables used in this research (e.g. per cent of households lacking a facility), "high" indicated greater disadvantage/deprivation while "low" indicated less disadvantage/deprivation with respect to the latent factor. The choice of binary (2-group) classification was to allow for this simple conceptualization of relative "high" and "low" and to facilitate subsequent analyses using
This approach of index construction can be considered superior to creating simple summation/averaging index as the latter simplistically places equal importance among contributing variables. The K-means classification output, on the other hand, is more influenced by the variables with greater variance than those with less. In other words, the resultant classification of high and low is weighted in terms of original variables. Also, it avoids the use of any arbitrary, theoretical or subjective threshold, but utilizes only the data to determine relatively what is "high" and what is "low" in local context.

3.5.3 Validating indices using location theory

The binary slum indices were validated by investigating their relationship with three theorized settlement location attributes. Using SPSS software (IBM, 2012), independent t-tests were conducted to determine if units with different index values (i.e. "high" and "low") were also significantly different with respect to settlement location. The generalized null hypothesis could be stated thus: for any variable (v) measuring unit location, there was no difference between units of “high” or “low” slum index categories in term of v. The locational variables tested were as follows:

- Mean distance (m) to central commercial district (indicating centrality)
- Mean distance (m) to nearest town market (indicating both centrality and closeness to major source of informal employment)
- Mean distance (m) to closest major access road (indicating accessibility to transport network)

Centrality, accessibility and proximity to source of income are all theoretically established location attributes that can help to profile slum residents (Gulyani & Talukdar, 2008; UN-
HABITAT, 2003). Statistical inference was made at significance level of 0.05, and the general null hypothesis was that units of "high" and "low" classification (as per derived binary indices) were not significantly different with respect to location attributes.

3.6 Objective II: To investigate the potential of unsupervised learning for mapping slum conditions using physical settlement attributes

The second major objective of the dissertation research was to determine if slum conditions could be inferred by unsupervised/inductive learning from physical settlement attributes derived primarily from imagery. To realize this objective, the following research questions needed to be answered: what are the different physical categories/types into which statistical units can be classified? Do these types reflect differences in household slum conditions and settlement location attributes?

3.6.1 Classifying units into unlabeled physical types/categories

A multistep approach was employed to derive the best classification of statistical units into physically similar categories. The following 6 settlement variables derived from land use/cover classification of satellite imagery were used:

- Percent open space area (green vegetation and fields)
- Percent roof area (building density/coverage)
- Building shadow area to roof area ratio (additional indicator of building density and sizes)
- Percent unpaved surface area (dirt road plus bare soil surfaces)
- Percent paved road surface area (as a fraction of total road surface area)
- Mean Slope

First of all, a principal component analysis was performed using SPSS software (IBM, 2012) on the above 6 variables to reduce noise in the data and maximize variance. This was
expected to lead to best separation of potential clusters. The first two derived principal components (with eigen values > 1) were extracted and used for unsupervised classification of statistical units. The hierarchical clustering (AHC) algorithm provided by SPSS software was used to classify units into groups based on physical similarities. The general description of the algorithm has previously provided (see section 3.4.1). However, in this case the objective was to classify observations (statistical units) and not variables and made use of squared Euclidean distance between observations and clusters, assuming complete (furthest neighbor) linkage.

The squared Euclidean distance has an optimization edge over normal Euclidean distance by emphasizing the contrast between distant clusters more than closer ones. Complete linkage also leads to the finding of more compact and less chained clusters. A range of solutions (2 - 10 clusters) was specified as output and the dendrogram was visually inspected to determine the best cluster solution, using scaled distance between merged pairs. The best clusters were deemed to be the one with largest average inter cluster distance.

3.6.2 Comparing physical types in terms of household and location attributes

Using SPSS software (IBM, 2012), Kruskal-Wallis (KW) tests (Kruskal & Wallis, 1952) were carried out to test the null hypotheses that physical categories of units were not different in terms of household slum indicators and location attributes. The KW test is the non-parametric equivalent analysis of variance (ANOVA) and was chosen as the data did not fully meet the necessary assumptions such as normality of distribution within groups and homogeneity of variances. Unlike ANOVA, the KW test does not employ distribution-dependent measures such as means and variances. It works by ranking observation in each category and tests if the general distribution is the same across categories. Inference was made at a significance level of 0.05.

The following dependent variables were used to determine statistically significant
differences between derived neighborhoods (UN-HABITAT, 2003):

*Household variables:*

- Percent of households without pipe borne water
- Percent of households without flush toilet
- Percent of non-owner (tenant) households
- Percent of households living in a single room or studio
- Percent of households with > 2 persons per bedroom
- Percent of Households without storm water drainage

*Location variables:*

- Mean distance to CBD
- Mean distance to nearest town market
- Mean distance to closest main access road

### 3.7 Objective III: To determine neighborhood scale for mapping slums

The third major objective of this research was to tackle the scale and boundary problem related to slum mapping. Analytic regionalization, also known spatially constrained clustering (Patino & Duque, 2012; Weeks et al., 2007), was employed in order to determine the optimal neighborhood scale and boundaries for mapping slum conditions. The rationale for this objective was based on the following premise:

- The statistical units used in this study were subject to aggregation problems such as the modifiable areal unit problem (MAUP).
- Most of the units were smaller than what is typically considered a city neighborhood, thus allowing them to serve as building blocks.
- The physical attributes of statistical units reflected slum conditions within them and hence
could be used to delineate contiguous slum neighborhoods.

Under this objective, the following research questions were addressed: can physically homogenous and contiguous neighborhoods be formed from adjacent statistical units? Are these neighborhoods different with respect to household slum conditions?

3.7.1 Spatially constrained clustering of spatial units

A minimum spanning tree (MST) algorithm was used to group statistical units into contiguous neighborhoods (regions). This made use of image derived attributes (similarly used for classifying units into physical types – see section 3.6.1). In addition to mean slope, mean elevation was included as a terrain measure. This was meant to help separate units located at the top of the Bamenda Escarpment (an area known locally as "Hill Top Station") as these cannot be considered contiguous with the rest of the city.

A detailed explanation of the MST algorithm is provided by Assunção et al. (2006). In summary, the technique uses a connectivity graph to represent spatial units such that each edge in the graph represents a link between adjacent features (Figure 3.9a). Associated with each edge is a cost function that reflects the dissimilarity between the connected units (i.e., their Euclidean separation in attribute space). Edges associated with a high dissimilarity are pruned reducing the graph to edges that only link very similar regions (Figure 3.9b). This process can continue iteratively, each time resulting in a simpler graph. The objective is to achieve the least complex graph or a minimum spanning tree that will lead to the formation of larger regions that are homogenous within themselves and quite distinct from each other with respect to the attributes of interest (Assunção et al., 2006).
Figure 3.9 Depiction of MST algorithm: a) Connectivity graph; b) Minimum spanning tree after graph pruning (Assunção et al., 2006, p. 800)

The Grouping Analysis Tool provided within the ArcMap Spatial Analyst tool box (ESRI, Redlands, CA) was used for this objective. It utilizes the MST algorithm for spatially constrained clustering. Figure 3.10 shows the tool input parameters. The dissimilarity measure used was the Euclidean distance. K-nearest neighbors were specified as spatial constraint instead of contiguity edges. The latter were not being applicable as there were existing gaps due to the omission of certain areas from the study (see section 3.2.2 and Figure 3.1 for details).

Just like with most non-hierarchical clustering algorithms, the tool requires the arbitrary specification of the number of regions that is desired. However, as part of the output, it can be programmed to generate a Pseudo-F statistics plot for a possibility of 2-15 groups. The number of groups with the highest Pseudo-F statistic value is considered the best solution from a statistical clustering point of view, i.e., the solution with the best separation between groups. For this study, the tool was run first by specifying a random number of desired groups just to derive Pseudo-F plot. Then a second run was carried out by specifying the optimal number of groups as determined
3.7.2 Comparing derived regions in terms of household attributes

Kruskal Wallis tests were conducted, at a significant level of 0.05, to determine if at least 2 analytically derived city regions or neighborhoods were significantly different with respect to household slum conditions. KW tests were used after determining that assumptions for ANOVA were not met. The following dependent variables were used to determine statistically significant differences between derived neighborhoods:

- Percent of households without pipe borne water.
- Percent of households without flush toilet.
- Percent of non-owner (tenant) households.
- Percent of households living in a single room or studio.
- Percent of households with > 2 persons per bedroom.
- Percent of Households without storm water drainage.
4. RESULTS AND DISCUSSION - MAIN OBJECTIVE I

4.1 Chapter overview

This chapter focuses on the presentation, interpretation and discussion of results obtained in carrying out the first main objective of this dissertation research. The primary objective was to explore an alternative multi-index approach to mapping slum conditions from household indicators. This was achieved through variable clustering (as an alternative to exploratory factor analysis), followed by multi-index construction using 6 slum indicators measured at the household level and aggregated for each spatial/statistical unit. The resultant indices were then independently validated using settlement location theory. Specific research questions were as follows:

- Are there multiple (i.e., more than 1) latent and uncorrelated slum factors that capture the joint variance within subsets of slum indicators?
- What is the relationship between multiple indices derived from latent slum factors and independent attributes derived from slum location theory?

In the following sections, the results are presented, interpreted and discussed in detail. The chapter ends with a recap of the main highlights.

4.2 Summary statistics of selected household variables

Table 4.1 shows the summary statistics of the 6 household variables selected for developing multiple-index framework for mapping slums conditions. Each variable was chosen for its direct link to one of the five major UN-Habitat slum criteria. Some variables that directly measure security of tenure (such as possession of land title and building permits) could not be included as
they suffered from large number of missing cases. Such information can only be reliably obtained from owners many of whom are 'absentee landlords' in Bamenda. Also excluded were variables related to durability of building structure (such as roof and wall materials) since the study area displays little variability in this regard.

Table 4.1 Summary of selected household indicators aggregated at statistical unit level

<table>
<thead>
<tr>
<th>Slum Indicator (as % of households)</th>
<th>Number of statistical units</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>without pipe borne water</td>
<td>63</td>
<td>8.64</td>
<td>9.09</td>
</tr>
<tr>
<td>with no flush toilet</td>
<td>63</td>
<td>80.06</td>
<td>13.34</td>
</tr>
<tr>
<td>non-owner (tenant)</td>
<td>63</td>
<td>42.15</td>
<td>18.53</td>
</tr>
<tr>
<td>with more 2 people per bedroom</td>
<td>63</td>
<td>38.30</td>
<td>33.30</td>
</tr>
<tr>
<td>living in single room or studio</td>
<td>63</td>
<td>32.83</td>
<td>15.24</td>
</tr>
<tr>
<td>without drainage</td>
<td>63</td>
<td>37.74</td>
<td>17.78</td>
</tr>
</tbody>
</table>

4.3 Deriving slum factors and multiple binary slum indices through variable clustering and conceptual interpretation

4.3.1 Results of hierarchical clustering of household variables

Table 4.2 shows the bi-variate correlations (Spearman, 1904) between household indicator variables. The strongest positive correlation was between the percentage of households that are non-owners and the percentage living in a single room or studio (p< 0.01). Both variables were also significantly correlated with the percentage with more than 2 residents per room (res_per_ro) (p<0.01 and p<0.05 respectively). Also statistically significant (p<0.05) was the correlation between percent of households without a flush toilet and percent without piped water. Percent of households without a storm water drainage was not significantly correlated with any of the other
variables. There were no significant negative correlations between variables.

Figure 4.1 displays the dendrogram resulting from the hierarchical clustering of the indicators variables. The vertical axis of the dendrogram represents the variables to be clustered while the horizontal axis represents the similarity between variables (and cluster of variables) as a scaled distance measure (standardized on a unit less scale of 0 - 25). Just as was seen with the bi-variate correlations, it can be observed that the most similarity was between the percentage of households that are non-owners (tenant) and the percent living in a single room or studio (room_studi) (separated by distance measure of <5). This constituted the first merged pair or cluster of variables in the hierarchy. The next most similar variable to this merged pair is the percent of households with more than 2 persons per room (res_per_room). Similarity, though relatively less strong, closeness can be observed between percent of households without a flush toilet (no_flush) and percent of households without pipe-borne water (non-pipebo).

Table 4.2 Bi-variate correlation coefficients (Spearman) between selected household variables

<table>
<thead>
<tr>
<th></th>
<th>% with &gt; 2 per room</th>
<th>% living in single room/studio</th>
<th>% without flush toilet</th>
<th>% without pipe water</th>
<th>% Without drainage</th>
<th>% non-owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>% with &gt; 2 per room</td>
<td>1.000</td>
<td>0.401**</td>
<td>-0.118</td>
<td>0.007</td>
<td>-0.174</td>
<td>0.321*</td>
</tr>
<tr>
<td>% in single room or studio</td>
<td>0.401**</td>
<td>1.000</td>
<td>-0.007</td>
<td>-0.006</td>
<td>0.025</td>
<td>0.428**</td>
</tr>
<tr>
<td>% without flush toilet</td>
<td>-0.118</td>
<td>-0.007</td>
<td>1.000</td>
<td>0.271*</td>
<td>-0.087</td>
<td>-0.203</td>
</tr>
<tr>
<td>% without pipe water</td>
<td>0.007</td>
<td>-0.006</td>
<td>0.271*</td>
<td>1.000</td>
<td>0.171</td>
<td>0.037</td>
</tr>
<tr>
<td>% without Drainage</td>
<td>-0.174</td>
<td>0.025</td>
<td>-0.087</td>
<td>0.171</td>
<td>1.000</td>
<td>0.064</td>
</tr>
<tr>
<td>% non-owner</td>
<td>0.321*</td>
<td>0.428**</td>
<td>-0.203</td>
<td>0.037</td>
<td>0.064</td>
<td>1.000</td>
</tr>
</tbody>
</table>

* = sig. at 0.05, ** = sig. at 0.01
Figure 4.1 Results of hierarchical clustering of variables

Using the dendrogram, the slum indicators were classified into 3 groups based on similarity:

- Group 1: percentage of non-owner households, percentage of households living in a single room or studio and percentage of households with more than 2 persons per room.
- Group 2: percentage of households without flush toilet and without pipe-borne water.
- Group 3: percentage of households without drainage.

To determine if a group suggested a potential factor, the variables belonging to the group were qualitatively judged if they related when interpreted on a conceptual basis. Of the three variables classified as group 1, two (res_per_ro and room_studi) directly measured amount of
living space or accommodation while the third variable (tenant) measured tenure conditions. As such, the entire group was determined intuitively as measuring occupancy conditions. Similarly, group 2, made of variables measuring access to both water and sanitation (non-pipebo and no_flush), was determined to collectively measure a utilities access. Since group 3 was only occupied by a single observed variable (no drainage), for this research, it was not considered as indicative of an underlying factor. This does not mean that it could not contribute to other factors as there is potential for additional variables to be used that also be meaningfully correlated with lack of drainage. Therefore, only 2 "unseen" or latent factors, hereafter referred to as slum factors, were deduced: occupancy factor and utilities factor (Table 4.3).

Table 4.3 Variable groups and potential factors derived from clustering

<table>
<thead>
<tr>
<th>Indicator variable</th>
<th>Group</th>
<th>Slum Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>% living in single room/studio</td>
<td>1</td>
<td>OCCUPANCY</td>
</tr>
<tr>
<td>% with &gt; 2 persons/bedroom</td>
<td>1</td>
<td>OCCUPANCY</td>
</tr>
<tr>
<td>% non-owner</td>
<td>1</td>
<td>OCCUPANCY</td>
</tr>
<tr>
<td>% without flush toilet</td>
<td>2</td>
<td>UTILITIES</td>
</tr>
<tr>
<td>% without pipe-borne water</td>
<td>2</td>
<td>UTILITIES</td>
</tr>
<tr>
<td>% without drainage</td>
<td>3</td>
<td>N/A*</td>
</tr>
</tbody>
</table>

*not associated with variables, thus discarded

4.3.2 Results of multi-index construction using slum factors

Binary indices quantifying each potential slum factor were generated by conducting an unsupervised (K-means) classification of statistical units using the variables clustered within each factor group. Since each group consists only of positively correlated variables, this classification approach resulted in groups labeled relatively as “high” and "low” with respect to original contributing variables. Thus, following two slum related indices were derived and labeled intuitively:

- an occupancy disadvantage/deprivation index (Occ_D) where in statistical units were
classified as having low disadvantage ("0") or high disadvantage ("1").

- A utilities disadvantage/deprivation index (UT_D) again with units classified as having low ("0") or high ("1") disadvantage.

Figure 4.2 shows both indices mapped separately over the study area while figures 4.3 – 4.7 are box plots that summarize the distribution of original household variables within the derived indices and justify the choice of "low" (0) and "high" (1) labels. In total 46 units were classified as having "low" Occ_D (i.e. "0") while 17 were classified as having "high" Occ_D ("1"). On the other hand, 17 units were classified as having "low" UT_D while 46 were classified as having high UT_D (see Figure 4.2).

![Map of derived slum indices (Occ_D and UT_D)](image)

Figure 4.2 Map of derived slum indices (Occ_D and UT_D)
Figure 4.3 Occ_D expressed in terms of % of non-owner households

Figure 4.4 Occ_D expressed in terms of % households with >2 residents per room
Figure 4.5 Occ_D expressed in terms of % of households living in a single room or studio

Figure 4.6 UT_D expressed in terms of % of households without flush toilet
4.4 Validating derived slum indices using location theory

4.4.1 Relationship between Occupancy Disadvantage Index (Occ_D) and settlement location attributes

Independents sample tests (t-test) were conducted to compare units of high and low Occ_D with respect to three location attributes. The attributes, theorized to be indicative of slum settlements, were as follows: mean distance to central business district (CBD), mean distance to nearest town market, mean distance to nearest main access routes. For each test, it was determined a priori if the assumptions for normality and equal variances across groups was violated or not.

Levene's tests of homogeneity of variances across groups were significant for the most of dependent variables (see Appendix I), meaning there was a significant case of unequal variances between groups. SPSS software provides an automatic correction of the t-test for this, known as Welch's t-test (Welch, 1947). However, this is usually unreliable when both groups are heavily
unbalanced in terms of size, as was the case here (46 against 17). As such further sampling was done to balance the groups - by randomly selecting 17 of the 46 units originally classified as "low" Occ_D (Occ_D = 0). This resulted in both groups (i.e., “high” and “low” Occ_D) having 17 units each. This not only reduced the instances of unequal variances, but allowed for a more reliable interpretation of test results where variances remained unequal.

Table 4.4 shows the summary statistics of test variables after balancing Occ_D groups. On the average, units of "high" Occ_D were closer (i.e. smaller mean distance) to nearest town markets, main roads and the CBD.

Table 4.4 Location attributes of units after group balancing

<table>
<thead>
<tr>
<th></th>
<th>Occ_D value</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to markets</td>
<td>0</td>
<td>17</td>
<td>1636.10</td>
<td>846.09</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>948.51</td>
<td>543.62</td>
</tr>
<tr>
<td>Distance to CBD</td>
<td>0</td>
<td>17</td>
<td>2139.77</td>
<td>1111.24</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>2006.38</td>
<td>915.57</td>
</tr>
<tr>
<td>Distance to access roads</td>
<td>0</td>
<td>17</td>
<td>509.89</td>
<td>316.92</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>278.44</td>
<td>172.51</td>
</tr>
</tbody>
</table>

Occ_D = Occupancy Disadvantage Index (0 = low, 1 = high), N = number of units

Table 4.5 shows the results of the independent t-tests after group sizes have been balanced. Except for distance to roads, for which the Welch t-test value was used, variances were assumed equal across Occ_D groups (0 and 1). Occ_D groups were significantly different in terms of mean distance to nearest town market (p<0.01) and mean distance to nearest main access roads (p<0.5). Units of high Occ_D (Occ_D = 1) were on the average located more than 500 meters closer to markets places and roughly 200 meters closer to main access roads when compared to units of low Occ_D (Occ_D = 0) (Table 4.4.). On the other hand, both groups did not differ much in terms of distance to the CBD.
Table 4.5 T-test comparing units of different Occ_D levels in terms of location attributes.

<table>
<thead>
<tr>
<th></th>
<th>t value</th>
<th>degrees of freedom</th>
<th>p-value</th>
<th>Mean Difference</th>
<th>Standard Error of Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to market places</td>
<td>2.819</td>
<td>32</td>
<td>0.008</td>
<td>687.58</td>
<td>243.91</td>
<td>190.75 1184.42</td>
</tr>
<tr>
<td>Distance to CBD</td>
<td>0.382</td>
<td>32</td>
<td>0.705</td>
<td>133.39</td>
<td>349.21</td>
<td>-577.93 844.71</td>
</tr>
<tr>
<td>Distance to access roads</td>
<td>*2.645</td>
<td>24.72</td>
<td>0.014</td>
<td>231.44</td>
<td>87.51</td>
<td>51.10 411.79</td>
</tr>
</tbody>
</table>

*Levene's test significant at 0.05 level, equal variances not assumed, Welch t-test statistic is used instead

4.4.2 Relationship between Utilities Disadvantage Index (UT_D) and settlement location attributes

Independent t-tests were also carried out to determine if UT_D groups also demonstrated differences with respect to settlement attributes. As was the case with Occ_D, three settlement location attributes were used as test variables and *a priori* tests were conducted to determine if test assumptions were met. Levene's tests showed unequal variances for test variables across groups UT_D groups (see Appendix I), and hence random sampling was done to balance group size and make the use of Welch's t-test more reliable.

Table 4.6 summarizes test variables within balanced groups. On the average, units of high UT_D (UT_D = 1) were located further away (greater mean distance) from markets places, central commercial distract and access roads.
Table 4.6 Summary of location attributes of units after UT_D group balancing

<table>
<thead>
<tr>
<th>UT_D value</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to market places</td>
<td>0</td>
<td>17</td>
<td>1183.80</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>1604.24</td>
</tr>
<tr>
<td>Distance to CBD</td>
<td>0</td>
<td>17</td>
<td>1749.77</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>2231.64</td>
</tr>
<tr>
<td>Distance to access roads</td>
<td>0</td>
<td>17</td>
<td>329.68</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>530.50</td>
</tr>
</tbody>
</table>

UT_D = Utilities Disadvantage Index (0=low, 1=high), N = number of units

Table 4.7 shows the results of the comparing UT_D groups in terms of their location attributes. In all cases, variances were assumed equal (i.e., after group balancing). Of all the test variables, only the mean distance to access roads showed significance (p<0.05) in distinguishing units of low UT_D and high UT_D. Units of high UT_D (UT_D = 1) were on the average located 200 meters more distant from access roads when compared to units of low UT_D (UT_D = 0).

Table 4.7 T-test comparing units of different UT_D levels in terms of location attributes

<table>
<thead>
<tr>
<th>t value</th>
<th>degrees of freedom</th>
<th>p-value</th>
<th>Mean Difference</th>
<th>Standard Error of Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.684</td>
<td>32</td>
<td>0.102</td>
<td>-420.44</td>
<td>249.63</td>
<td>-928.91 - 88.03</td>
</tr>
<tr>
<td>-1.280</td>
<td>32</td>
<td>0.210</td>
<td>-481.86</td>
<td>376.37</td>
<td>-1248.51 - 284.78</td>
</tr>
<tr>
<td>-2.347</td>
<td>32</td>
<td>0.025</td>
<td>-200.82</td>
<td>85.57</td>
<td>-375.12 - 26.51</td>
</tr>
</tbody>
</table>

*Levene's test significant at 0.05 level, equal variances not assumed, Welch t-test statistic is used
4.5 Discussion

4.5.1 The importance of multi-index measurement of slums

The objective discussed in this chapter was based on the idea that slum conditions could be measured/represented using multiple diverse indices as opposed to the common approach of constructing a single index from available indicators. It is true that a single index allows for the local slum conditions to be quantified as a whole and for all units to be compared directly and ranked on the same scale (UN-HABITAT, 2003). However, while this may be sufficient to quantify the overall magnitude of the problem, information is lost on the descriptive side. In such a case, additional questions remained unanswered. For example, which indicators or sets are most pertinent in an area? Which indicators allow for better separation of local spatial units?

The results presented above revealed 2 things: 1) Six household slum indicators in Bamenda could be jointly represented by at least 2 slum factors or indices; and 2) as these factors/indices are uncorrelated, they represent unique and diverse perspectives of the slum phenomenon in the local context. From a statistical point of view, multiple indices constructed from subsets of homogenous variables do not only quantify but also provide more clarity to the picture. Such an approach achieves better separation of spatial units and helps to reveal clearer patterns in spatial distribution. As is seen in this study, the 2 major slum factors revealed by the data were neither statistically nor spatially correlated. Hence, it was more informative to depict and explain them separately rather combining them into a single slum index.

The existence of the derived factors could also be explained logically. For example, the existence of utilities related factor could be explained in that the infrastructure of piped water and flush toilets overlap as both services require in-house plumbing and drainage systems. It was highly improbable for a residence to have flush toilets without piped water connection. However,
there can be piped water without flush toilets as in the case of external public stand pipes and toilets (latrines). Piped water connection also depended heavily on being able to access the wider trunk water network (Gulyani & Bassett, 2010). Meanwhile flush toilets were only limited to onsite sewage disposal through septic tanks, as Bamenda does not have central/collective sewerage, not even at local neighborhood level.

As for the occupancy related factor, it made sense than an area dominated by single room/studio residences would also have a high proportion of overcrowded households. Similarly, owner-occupied units were less likely to have issues of space because of the self-building culture that is common in Cameroon. Owners households are more likely to have considered their individual family size in the process of construction. This follows Tuner’s theory that flexibility of self-built homes may actually lead to positive housing outcome (Turner, 1969; Turner & Fichter, 1972b). Meanwhile, tenant households would not have the same control as they only feature the post-building stage. In economically challenged urbanities such as Bamenda, the tenant households, including those with large families, would most likely consider rental cost and location before living space.

Multidimensional measurement is common feature in urban social research, and has been used extensively for poverty/deprivation and slum mapping (Baud et al., 2008; Weeks et al., 2007). A good example is the livelihood assets framework: assets represent diverse types of capital (human, social, financial, physical and natural) and the extent to which individuals, households or areas possess these assets can be quantified and used to measure different types of deprivation (Baud et al., 2008). Similarly, the 5-point multi-criteria developed by the UN and used to define slums (UN-HABITAT, 2003) can be also used to automatically derive multiple slum indices. Unlike these however, the 2 indices derived in this study were the product of a localized data driven
methodology. Hence they represent a more accurate picture of the study area. In other words, we could surmise that 2 slum factors derived here were more efficient in describing the slum conditions in Bamenda than would the standard UN slum indicators.

4.5.2 Validating multiple indices using location theory

The results also showed that the derived slum indices were reliable measures of slum conditions, by reflecting broader slum settlement theory, particularly centrality and accessibility. This was especially true for the Occupancy disadvantage index: areas of “high” Occ_D (i.e. with higher percentage of overcrowded and tenant households) were consistently closer to town market and access roads when compared areas of “low” Occ_D. This could be interpreted in the sense that the former was more likely to feature households involved in the informal economy with less access to private means of transportation (hence being close to main roads). The fact that “high” and “low” Occ_D categories were not significantly different in terms of distance to CBD should not dilute the above assertion. The concept of centrality has to be extended beyond the traditional "downtown" or CBD to the all areas of intensive economic activity (UN-HABITAT, 2003). In the case of Bamenda, this meant considering all local town markets, of which there are 4: two main markets located in the CBD area and 2 secondary markets (Nkwen and Ntarinkon) located away from the city center. These areas provided ideal employment opportunities for the low-skilled and under-educated residents who typically make up slums. Proximity to the main non-seasonal access roads also ensured accessibility for those who could not obtain housing within the vicinity of the market place.

The Utilities disadvantage (UT_D) index did not reflect settlement location attributes in the same way. Areas having “high” UT_D (i.e., higher proportion of households lacking flush toilets and piped water) instead seemed to be further away from roads. Meanwhile there was no
meaningful relationship between UT_D and distance to markets or CBD.

The relative lack of utilities in less road-accessible areas could be attributed due to the limited nature of the general network infrastructure (e.g. water distribution pumps and pipelines) and the extra burden of extending service to remote areas. However, in the case of Bamenda, these areas also featured, albeit in isolated cases, individual well-constructed and serviced homes housing middle-high income families seeking to benefit from greater space and relatively lower cost of land. These residents were also likely to enjoy superior modes of transport such as ownership of private vehicles (Turner, 1969; UN-HABITAT, 2003).

Thus, unlike the case with the occupancy conditions, lack of access to utilities in the study area (particularly piped water) potentially cut across areas with different socio-economic profiles and was not very helpful in separating local areas as slums even though it is a universally recognized slum indicator. This is not uncommon in literature. For example, based on the perception of focus group discussion participants, Lemma et al. (2006) decided to exclude water supply as a key indicator for mapping slums in Addis Ababa as it was considered a widespread problem.

4.6 Summary

The purpose of this chapter was to present and discuss the results obtained from conducting Main Objective I of this dissertation research, namely: to explore a multi-index approach to measuring slum conditions in Bamenda. Slum indicators were analyzed using variable clustering to derive latent slum factors which were quantified as 2 diverse/uncorrelated binaries ("high" and "low") indices: Occupancy Disadvantage and Utilities Disadvantage index. Both indices were validated by determining if they reflected broader settlement theory, particularly in terms of the proximity to town markets, CBD and major roads.
5. RESULTS AND DISCUSSION - MAIN OBJECTIVE II

5.1 Chapter overview

This chapter is devoted to presenting, interpreting and discussing the results obtained from carrying the second main objective of this dissertation research. The objective was to investigate the potential of unsupervised machine learning for learning slum conditions using image derived settlement attributes. The specific research questions were as follows:

- What are the physically distinct types of local settlement units?
- Do physically distinct groups of units differ with respect to aggregate household slum indicators?
- Do they differ with respect to location attributes?

To answer these questions, a methodological approach was adopted that combined the following: image classification to derive land use/land cover, quantifying said land cover per statistical unit, unsupervised classification of statistical units into unlabeled physical types, and using tests of significance to compare these physical groupings with respect to household and location attributes.

5.2 Results of image classification and derivation of physical settlement attributes

5.2.1 Image classification results

Figure 5.1 displays the thematic land use/cover maps for a section of the study area, as derived from the classification of GeoEye multispectral imagery. Figure 5.2 shows the effect of
merging classes to derive main classes of interest.

Figure 5.1 Output map showing land use/cover (left) for a portion (right) of the study area.

Figure 5.2 Output maps before (left) and after (right) merging classes.
Table 5.1 Classification accuracy assessment

<table>
<thead>
<tr>
<th>Class</th>
<th>PA (%)</th>
<th>UA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roof</td>
<td>99.61</td>
<td>78.87</td>
</tr>
<tr>
<td>Open space</td>
<td>100.00</td>
<td>97.33</td>
</tr>
<tr>
<td>Bare Soil</td>
<td>93.70</td>
<td>94.16</td>
</tr>
<tr>
<td>Paved Road</td>
<td>96.41</td>
<td>100.00</td>
</tr>
<tr>
<td>Dirt road</td>
<td>85.62</td>
<td>90.08</td>
</tr>
<tr>
<td>Shadow</td>
<td>54.53</td>
<td>100.00</td>
</tr>
</tbody>
</table>

OA = 91.87%, Kappa coefficient = 0.90

PA = producer accuracy, UA = user accuracy, OA = overall accuracy

Table 5.1 (above) shows the key results of the assessment of classification accuracy. The detailed error matrix is provided in Appendix II. The overall accuracy of image classification was high (91.87 %), same as the Kappa Coefficient (0.90). The producer accuracies for each class represents the percentage of test pixels correctly classified per their observed class. The user accuracies on the other focus on the classified pixels and determines what per cent of them had been correctly labeled as such. Therefore, the producer accuracy evaluates the performance of the classification technique while the user accuracy evaluates the quality of the output classification map. Except for dirt roads (85%), and building shadows (54%), most classes achieved very high producer accuracies (>90%), with roof and open spaces being highest. Also, the user accuracies were also high for all classes except for roof class (78%) (Table 5.1).

5.2.2 Derived physical settlement attributes

Figure 5.3 displays choropleth maps of 6 physical attributes within each statistical unit, derived directly from image classification. Each attribute has been further classified for map purposes into five classes using the Jenks natural breaks classification (functionality provided
within ArcMap software). Units of high building coverage (roof area) appeared to predominate central areas while open space was more common in the peripheral settlements. The widest and paved roads seemed to be more prevalent in the city center as well while dirt road and bare soil surfaces spread out more evenly from center to the periphery.

Figure 5.3 Choropleth maps showing physical attributes of statistical units
5.3 Results of mapping slum conditions using physical attributes of statistical units

5.3.1 Unsupervised classification of areal units to derive physical types

Statistical units were grouped into unlabeled settlement types (physical types) based on their physical similarities using Agglomerative Hierarchical Clustering (AHC). First, a principal component analysis of 7 physical settlement attributes quantified for each statistical unit was performed. These included 6 image-derived attributes (see Figure 5.3) as well as slope in Appendix III. The first 2 non-rotated components (with Eigen values > 1, Table 5.2) were extracted and used for clustering. Figure 5.4 shows the result (dendrogram) of hierarchical clustering of units using these 2 principal components. Visual examination of the dendrogram suggested an optimal classification with 4 distinct settlement types (labeled 1 - 4), hereafter referred to as physical types or physical categories. Using scaled similarity measure provided on the horizontal axis, this solution represented the largest total inter cluster Euclidean separation between clusters (Figure 5.4). Table 5.3 describes the said physical types in terms of the means and deviations of original physical variables (not the principal components used for clustering). This is useful in identifying the key attribute(s) defining each physical type or category. Figure 5.5 is the map result that shows spatial distribution of the physical types in the study area

Table 5.2 Description of principal components used for AHC

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigen values</th>
<th>Extraction Sums of Squared Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>% of Variance</td>
</tr>
<tr>
<td>1</td>
<td>4.80</td>
<td>68.58</td>
</tr>
<tr>
<td>2</td>
<td>1.03</td>
<td>14.68</td>
</tr>
</tbody>
</table>
Figure 5.4 Dendrogram showing hierarchical clustering of units using physical attributes
Table 5.3 Derived physical types in terms of original variables

<table>
<thead>
<tr>
<th>Physical settlement attribute</th>
<th>settlement type (physical category)</th>
<th>1 (n=14)</th>
<th>2 (n=34)</th>
<th>3 (n=8)</th>
<th>4 (n=7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>percent roof area</td>
<td>11.59</td>
<td>6.65</td>
<td>28.02</td>
<td>6.89</td>
<td>41.69</td>
</tr>
<tr>
<td>percent open space area</td>
<td>80.46</td>
<td>10.24</td>
<td>53.84</td>
<td>10.89</td>
<td>32.39</td>
</tr>
<tr>
<td>Percent bare soil area</td>
<td>2.67</td>
<td>1.25</td>
<td>6.68</td>
<td>1.73</td>
<td>5.94</td>
</tr>
<tr>
<td>percent total road surface</td>
<td>3.58</td>
<td>2.11</td>
<td>8.42</td>
<td>2.90</td>
<td>17.42</td>
</tr>
<tr>
<td>percent of road surface that is paved</td>
<td>27.48</td>
<td>19.68</td>
<td>28.56</td>
<td>15.19</td>
<td>70.88</td>
</tr>
<tr>
<td>shadow area to roof area ratio</td>
<td>0.33</td>
<td>0.05</td>
<td>0.31</td>
<td>0.04</td>
<td>0.18</td>
</tr>
<tr>
<td>mean slope</td>
<td>16.81</td>
<td>4.52</td>
<td>12.48</td>
<td>1.68</td>
<td>13.14</td>
</tr>
</tbody>
</table>

SD = standard deviation, n = number of units in each category (total = 63)

Figure 5.5 Map showing Classification of units into physical types/categories

Statistical units in the first physical type (Category 1) dominated the peripheral regions. They had the highest mean per cent open space cover (>80%), and were lower than the other
categories in terms of per cent cover of roof area, per cent total road surface area and per cent bare soil surface area. Category 2 comprised units that had moderate per cent roof and open space area, but with the highest mean per cent bare soil and unpaved road surfaces (>10%). Category 3 formed the central core of city, with largest mean per cent roof area and smallest mean open space cover. This category was also endowed with the highest percent of paved road surface. Category 4 contained units located on the steepest slopes along the Bamenda escarpment (mean slope > 25%). Although with only moderate mean percent roof area, units in this category were characterized by the highest building shadow-to-roof area ratio, indicating the presence of more numerous smaller buildings though with relatively less density when compared to Category 3.

5.3.2 Relationship between physical types and household slum indicators

To validate derived physical types as potentially indicating slum conditions, independent samples Kruskal-Wallis tests were conducted to compare them with respect to the following household variables: percent without pipe borne water, percent without a flush toilet, percent with more than 2 persons per bedroom, percent living in single room or studio, percent of non-owner(tenant) households and per cent without external drainage on the premises (see Appendix IV for details). Unlike its parametric counterpart, ANOVA, the Kruskal-Wallis test looks for differences in the distribution from one group to another, by ranking all observations. It is less susceptible to the failure of the data to meet assumptions such as normality within groups and equality of variances across groups.

Table 5.4 shows the main highlights of Kruskal-Wallis tests conducted using SPSS software. At least 2 physical categories/types were significantly different in their distribution of the per cent of non-owner households and per cent of households with greater 2 residents per room. There was no significant difference between categories with respect to percent of households not
having piped water, flush toilet, drainage and living in a single room or studio.

Figure 5.6 shows the distribution of the per cent of non-owner households across all four physical types while Figure 5.7 shows the pair wise comparisons between individual physical types. The per cent of non-owner households was generally higher for units in category 4 when compared to the units in the other categories. In terms of paired wise comparisons, there was significant difference between categories 1 and 4 ($p < 0.05$) and less so between categories 2 and 4 ($p < 0.1$).

Figures 5.8 - 5.9 show respectively the description and pair wise comparisons of physical types in terms of the per cent of households with more than 2 persons per room. Units in category 4 once again generally had a higher per cent of overcrowded households while those in category 1 had the lowest. In terms of paired wise comparison, there was a significant difference between categories 1 and 4 and between categories 2 and 4 ($p < 0.01$) in both cases.

Table 5.4 Kruskal-Wallis tests comparing distribution of 6 household variables across all four physical types

<table>
<thead>
<tr>
<th>Test variable</th>
<th>Sig. ($p$-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% non-owner</td>
<td>0.029*</td>
</tr>
<tr>
<td>% with &gt;2 residents per room</td>
<td>0.000**</td>
</tr>
<tr>
<td>% living in single room or studio</td>
<td>0.199</td>
</tr>
<tr>
<td>% without flush toilet</td>
<td>0.677</td>
</tr>
<tr>
<td>% without pipe water</td>
<td>0.291</td>
</tr>
<tr>
<td>% without drainage</td>
<td>0.528</td>
</tr>
</tbody>
</table>

*significant at 0.05, **significant at 0.001, n = 63
Figure 5.6 Distribution of % of non-owner households across physical types

Figure 5.7 Pair wise comparisons of physical types: % of non-owner households

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000-2.000</td>
<td>-2.003</td>
<td>5.223</td>
<td>-3.84</td>
<td>.701</td>
<td>1.000</td>
</tr>
<tr>
<td>1.000-3.000</td>
<td>-11.150</td>
<td>7.645</td>
<td>-1.458</td>
<td>.145</td>
<td>.868</td>
</tr>
<tr>
<td>1.000-4.000</td>
<td>-26.828</td>
<td>9.717</td>
<td>-2.761</td>
<td>.006</td>
<td>.035</td>
</tr>
<tr>
<td>2.000-3.000</td>
<td>-9.146</td>
<td>8.046</td>
<td>-1.137</td>
<td>.256</td>
<td>1.000</td>
</tr>
<tr>
<td>2.000-4.000</td>
<td>-24.825</td>
<td>10.035</td>
<td>-2.474</td>
<td>.013</td>
<td>.080</td>
</tr>
<tr>
<td>3.000-4.000</td>
<td>-15.679</td>
<td>11.484</td>
<td>-1.365</td>
<td>.172</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.
Figure 5.8 Distribution of % overcrowded households across physical types

Each node shows the sample average rank of settlement type (physical attributes).

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000-3.000</td>
<td>-11.743</td>
<td>7.605</td>
<td>-1.544</td>
<td>.123</td>
<td>.735</td>
</tr>
<tr>
<td>1.000-2.000</td>
<td>-18.697</td>
<td>5.195</td>
<td>-3.599</td>
<td>.000</td>
<td>.002</td>
</tr>
<tr>
<td>1.000-4.000</td>
<td>-30.422</td>
<td>9.666</td>
<td>-3.147</td>
<td>.002</td>
<td>.010</td>
</tr>
<tr>
<td>3.000-2.000</td>
<td>6.954</td>
<td>8.004</td>
<td>.869</td>
<td>.385</td>
<td>1.000</td>
</tr>
<tr>
<td>3.000-4.000</td>
<td>-18.679</td>
<td>11.424</td>
<td>-1.635</td>
<td>.102</td>
<td>.612</td>
</tr>
<tr>
<td>2.000-4.000</td>
<td>-11.725</td>
<td>9.983</td>
<td>-1.174</td>
<td>.240</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 5.9 Pair wise comparisons of physical types: % of overcrowded households
5.3.3 Relationship between physical types and location attributes

To further reinforce the hypothesis that physical types derived through unsupervised classification of statistical units reflect possible slum conditions within them, further tests were conducted to compare them with respect to 3 settlement location attributes, typically associated with slum residents. In each case, at least 2 physical categories were significantly different in terms of the distribution of all three location attributes: mean distance to Central Business District (CBD), mean distance to main roads and mean distance to town market (Table 5.5).

Units in Category 3 were generally closest to the CBD while units in Category 1 were furthest away (Figure 5.10). Significant differences occurred between categories 1 and 3 (p<0.001) and between categories 1 and 2 (p<0.05) in terms of proximity to the CBD (Figure 5.11). In terms of proximity to markets, units in Category 1 were significantly from all the other 3 categories (p<0.05) (Figures 5.12 and 5.13). Finally, units in categories 2 and 3 were closest to main roads, and were significantly different from units in Category 1 in that regard (Figures 5.14 and 5.15).

Table 5.5 Kruskal-Wallis tests comparing distribution of 3 location variables across all four physical types

<table>
<thead>
<tr>
<th>Test variable</th>
<th>Sig. (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance to town market</td>
<td>0.000**</td>
</tr>
<tr>
<td>Distance to CBD</td>
<td>0.000**</td>
</tr>
<tr>
<td>Distance to main road</td>
<td>0.000**</td>
</tr>
</tbody>
</table>

**significant at 0.001, n = 63
Figure 5.10 Distribution of distance to CBD across physical types

Each node shows the sample average rank of settlement type (physical attributes).

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000-2.000</td>
<td>19.85</td>
<td>8.050</td>
<td>2.466</td>
<td>.014</td>
<td>.082</td>
</tr>
<tr>
<td>3.000-4.000</td>
<td>-29.250</td>
<td>11.489</td>
<td>-2.546</td>
<td>.011</td>
<td>.065</td>
</tr>
<tr>
<td>3.000-1.000</td>
<td>35.125</td>
<td>7.649</td>
<td>4.592</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>2.000-4.000</td>
<td>-9.400</td>
<td>10.040</td>
<td>-.936</td>
<td>.349</td>
<td>1.000</td>
</tr>
<tr>
<td>2.000-1.000</td>
<td>15.275</td>
<td>5.225</td>
<td>2.923</td>
<td>.003</td>
<td>.021</td>
</tr>
<tr>
<td>4.000-1.000</td>
<td>5.875</td>
<td>9.721</td>
<td>.604</td>
<td>.546</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 5.11 Pair wise comparison of physical types: distance to CBD
Figure 5.12 Distribution of distance to nearest town market across physical types

Figure 5.13 Pair wise comparison of physical types: distance to nearest town market
Figure 5.14 Distribution of distance to closest main road across physical types

Each node shows the sample average rank of settlement type (physical attributes).

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000-4.000</td>
<td>-20.607</td>
<td>11.489</td>
<td>-1.794</td>
<td>.073</td>
<td>.437</td>
</tr>
<tr>
<td>3.000-2.000</td>
<td>20.957</td>
<td>8.050</td>
<td>2.603</td>
<td>.009</td>
<td>.055</td>
</tr>
<tr>
<td>3.000-1.000</td>
<td>37.201</td>
<td>7.649</td>
<td>4.864</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>4.000-2.000</td>
<td>.350</td>
<td>10.040</td>
<td>.035</td>
<td>.972</td>
<td>1.000</td>
</tr>
<tr>
<td>4.000-1.000</td>
<td>16.594</td>
<td>9.721</td>
<td>1.707</td>
<td>.088</td>
<td>.527</td>
</tr>
<tr>
<td>2.000-1.000</td>
<td>16.244</td>
<td>5.225</td>
<td>3.109</td>
<td>.002</td>
<td>.011</td>
</tr>
</tbody>
</table>

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 5.15 Pair wise comparison of physical types: distance to closest access roads
5.4 Discussion

5.4.1 Understanding physical types

Using the physical appearance of an area to qualitatively judge household conditions is a well-known efficient practice (Kohli et al., 2012). More so, physical information can be easily observed on acquired imagery, eliminating the need for field visits. However, it also assumes the observer already has the necessary expertise or prior slum data to make such judgments. In such cases, mapping is usually achieved through supervised learning, by quantifying the physical attributes and modeling their relationship with previously acquired slum data using techniques like regression or classification trees (Graesser et al., 2012; Owen & Wong, 2013).

As the result here demonstrates, unsupervised learning from imagery achieve the same mapping outcome with reasonable success and without the need for prior local slum data. This was rooted in the idea that, at least locally, slum areas always tend to look different from non-slum areas, and thus could be highlighting simple data clustering using relevant physical indicators. Such indicators were easily derived from the theoretical and were part of generalized physical slum ontology (Kohli et al., 2012). Also, unsupervised learning allowed for output of more than two simple slum/non-slum categories. This was important in the light of potentially significant intra local physical deviations.

The unsupervised approach was an inductive learning process, allowing for statistical units to be placed into initially unlabeled but eventually meaningful categories based on physical attributes. The use of a deterministic algorithm such as hierarchical clustering, coupled with relevant land cover attributes of the study area, allowed for the creation of the "best possible" number of physical types from the data point of view. In order words, the classification was most representative of the "local context."
Per the results, statistical units in Bamenda could be optimally classified into four unique physical types (labeled as Categories 1 - 4). Although these were initially unlabeled in terms of slum conditions, their physical descriptive properties could be used to profile them more profoundly. One category (Category 1) comprised units with lots of open space and very low roof coverage, characteristics which according literature makes them less likely to feature slum conditions and more likely to consist of affluent or formal housing units (Weeks et al., 2007). However, drawing this conclusion was not straightforward as this category also suffered from lack of sufficient road access, particularly, a low proportion of paved road surface. Another category (Category 2) had relatively moderate levels of roof density and open space cover, but registered high proportion of unpaved surfaces (bare soil and dirt roads). The latter attribute (dirt surfaces) has been shown to be key feature of slum/informal areas (Owen & Wong, 2013). A third category (Category 3) had the highest roof coverage/building density of all the physical types. Combined with the relatively central location of most units in this type, theory would suggest this to be likely representative of traditional central slums (UN-HABITAT, 2003). However, as seen with validation tests comparing all four physical types, this category did not necessarily register the worst household conditions, another evidence of how local contest matters in slum mapping. Category 4 units had two distinguishing physical features: 1) location on the very steep slopes along part of the Bamenda Escarpment, 2) highest shadow area to roof area ratio. Despite displaying only moderate levels of most land cover attributes, comparison tests showed this physical type to be associated with most deprived living conditions.

The shadow area to roof area ratio was is an innovative measure introduced in this study to intuitively describe settlement structure. Shadow information is normally considered an undesirable element in imagery that should be optimally removed during image processing (Dare,
2005). But in this case, building shadows were classified and retained in the output map to gain more insight to built-up characteristics of the statistical units. The clear majority of residential buildings in the study area had similar heights. This means they could not easily cast shadows on each other but on the ground or low profile open spaces in between them. Therefore, an area with numerous smaller but more spaced buildings would have more shadow coverage than an area with larger buildings but closely packed building. Thus, the shadow area/roof area ratio for the area provided additional information on the number of the buildings and their sizes as opposed to simple building or roof area. This was very useful in this study as algorithmic image classification methods could not be relied on to extract clean building outlines/footprints for accurately calculating density. Manual counting and digitization methods would have been impractical considering the sheer size of the study area. The shadow-roof ration helped to explain why Category 3 units, despite having the highest percentage of roof coverage, had less severe household compared to Category 4. The latter, though dense, featured smaller size buildings, which is a well-known attribute of slums and informal settlements (Hofmann et al., 2008).

5.4.2 Physical typology as an indicator of household slum attributes

The physical types discussed above were arrived at through an unsupervised classification process that did not assume pre-existing knowledge on the distribution of slum conditions in the study area. However, the results from the Kruskal Wallis tests comparing them showed that the they could be used to represent slum conditions across the study area. Two aggregated household slum attributes were significantly associated \textit{a posteriori} with the physical typology of areal units: 1) the proportion of non-owner/tenant households and 2) proportion of overcrowded households.

These household attributes were on the average highest amongst units in Category 4 and lowest amongst units in Category 1. The former was mostly located on very steep slopes and had
numerous smaller buildings (high shadow to roof area ratio) while the latter represented suburban areas with very low building density and large amounts of open space. On this evidence, "Category 4" would represent the "physical face" of the overcrowded and legally insecure areas while "Category 1" will represent the face of most desirable/formal housing within the study area. Physically, Categories 2 and 3 did exhibit slum like symptoms namely abundant dirt surfaces and high building density respectively. However, they did not have the worst slum conditions, although they did have higher slum attributes compared to Category 1.

It can be surmised that in the context of the study area, steep slope and small building sizes were the most "telling" physical traits indicative of slum conditions while open space and low density were most indicative of non-slum conditions. An attempt can be to made understand the direct links between the physical domain and the household slum attributes. The small size of the buildings places a physical constraint that would translate into fewer bedrooms and hence overcrowded households. The hazardous location (steep slopes) indicates likely illegal or unauthorized construction. The fact that Category 4 areas were dominated by non-owner (tenant) households is further confirmation of their potential lack of tenure security and informal/slum status. Although being a tenant is not automatically an indicator of insecurity, empirical evidence has shown that slums tend to be dominated by rent-paying tenant households (Gulyani & Talukdar, 2008).

The location attributes of the different physical types also provided useful insight into whether they could be profiled as slums or not. Category 3 units which had very high roof area coverage (typical of many slums) happened be the most centrally located; with shortest mean distance to the "Down Town" Area (CBD), the nearest town market and the closest access road. On the other hand, Category 1 units (profiled to consist of mostly formal housing) were generally
furthest from these commercial and transport hubs. On likely reason is that these areas featured residents working in the formal sector with daily activities not heavily tied to places such as town markets. Another possibility is that they were affluent enough to own private modes of transport which afforded them to take up residence in more remote areas with the city.

5.5 Summary

The purpose of this chapter was to present and discuss results pertaining to realizing Main Objective II of this dissertation research. The objective's focus was on how to infer slum conditions of an area from its physical attributes through an unsupervised learning process. Using hierarchical clustering, it was determined that area units in Bamenda could be classified into 4 physical types or categories, each with different defining physical characteristics. Based on a posteriori evidence, it was determined that physical types could serve as proxy labels for varying household slum conditions within the study area.
6. RESULTS AND DISCUSSION - MAIN OBJECTIVE III

6.1 Chapter overview

This chapter presents, interprets and discusses the results of the third major objective of this dissertation research. This overall objective was as follows: to determine optimal scale and boundaries for mapping slum conditions. This was achieved through analytic regionalization (spatially constrained clustering) of statistical area units. Unlike Main Objective II, the focus here was not the unsupervised physical classification of units but to determine whether adjacent units were similar enough to be merged into larger contiguous and physically homogenous city regions. Such regions would be more representative of neighborhood scale for mapping slum conditions. Thus, the two research questions addressed in this Chapter as follows: 1) what are the analytically derived physically homogenous regions in the city; 2) Can these regions be used to identify contiguous slum/non-slum neighborhoods?

6.2 Results of deriving contiguous slum neighborhoods/regions through analytic regionalization

6.2.1 Physically derived analytic neighborhoods (regions)

To generate analytic regions/neighborhoods within the study area, all 63 statistical units under consideration were subject to spatially constrained clustering. This made use of mainly physical attributes (land cover and terrain measures). Elevation was an additional terrain variable (besides slope) to help separate neighborhoods located at the top the Bamenda escarpment (an area locally referred to as Hill Top Station Bamenda).
Spatially constrained clustering was by the minimum spanning tree algorithm, with k-nearest neighbors selected as the spatial/contiguity constraint. Pseudo F-statistics were used to decide the optimal number of physically derived analytic neighborhoods (hereafter referred to simply as regions) based on the selected attributes. Figure 6.1 shows the resultant regions as groups of adjacent units while Figure 6.2 is Pseudo F-statistic Plot used to determine the desired number of regions (from a possible range of 2-15) specified at onset of clustering.

Figure 6.1 Map of analytically derived regions
Based on the Pseudo F statistics, the study area could best be separated into 5 regions that are physically homogenous. Table 6.1 describes each region in terms of physical variables used for regionalization and helps to identify the homogenizing traits of each region. Region 1 was located on the steepest slopes and had the highest mean shadow-to-roof area ratio. Region 2 (centrally located) was characterized by the highest percent roof area and paved road surface area. Regions 3 and 4 (both at the southern periphery of the city) were quite similar (low building density and large open spaces) but with Region 3 located at higher elevation (top of escarpment) and having better paved road coverage. Region 5 was not very distinctive in terms of any attributes but had relatively high percent of bare soil coverage with a low proportion of paved roads surfaces.
Table 6.1 Analytic regions in terms of physical attributes

<table>
<thead>
<tr>
<th>Physical settlement attribute</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent roof area</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>percent open space area</td>
<td>22.48</td>
<td>8.21</td>
<td>40.87</td>
<td>3.99</td>
<td>7.11</td>
</tr>
<tr>
<td>Percent bare soil area</td>
<td>61.56</td>
<td>13.29</td>
<td>33.77</td>
<td>6.04</td>
<td>6.04</td>
</tr>
<tr>
<td>percent total road area</td>
<td>5.58</td>
<td>2.71</td>
<td>16.87</td>
<td>2.65</td>
<td>4.21</td>
</tr>
<tr>
<td>proportion of road surface that is paved</td>
<td>48.59</td>
<td>10.03</td>
<td>69.16</td>
<td>4.21</td>
<td>44.51</td>
</tr>
<tr>
<td>shadow area to roof area ratio</td>
<td>0.47</td>
<td>0.09</td>
<td>0.18</td>
<td>0.04</td>
<td>0.34</td>
</tr>
<tr>
<td>mean slope</td>
<td>28.95</td>
<td>4.80</td>
<td>13.43</td>
<td>2.43</td>
<td>20.17</td>
</tr>
<tr>
<td>mean elevation</td>
<td>1292.29</td>
<td>12.95</td>
<td>1240.12</td>
<td>9.36</td>
<td>1511.12</td>
</tr>
</tbody>
</table>

SD = standard deviation, n = number of original units in each region

6.2.2 Comparing regions in terms of household slum conditions

Kruskal-Wallis tests were used to compare all 5 regions in terms of six aggregated household slum indicators. The purpose was to determine if at least 2 of these 'intelligently' delineated spatial units were significantly different enough and if at least one could be clearly labeled as a contiguous slum neighborhood. The basic null hypothesis was that the distribution of slum variables was the same across all Regions.

Table 6.2 summarizes the Kruskal-Wallis test results of comparing the 5 regions in terms of slum indicators. Three of the six household slum indicators had distributions that could separate regions. Overall, at least 2 regions were significantly different ($p<0.05$) in terms of the distribution of the following: per cent of non-owner households, per cent of overcrowded households (i.e. more than 2 residents per room), and per cent of households without external drainage.

Figures 6.3 - 6.5 show the distribution (mean) of the three important household indicators across the Regions. Region 1, located on steep slopes, had the highest mean of most indicators: per cent of households that are non-owners and with more than 2 residents per room. Region 2, the centrally located high density area, had mostly moderate levels of slum indicators but ranked second highest in terms of mean per cent of non-owner households. Regions 3 and 4 generally
ranked low in slum conditions. Region 3 was lowest in terms of non-owner households and those without drainage, while Region 4 was lowest in terms of overcrowded households (i.e., households with more than 2 persons per room). Region 4 was a conundrum as it also ranked highest in terms of households without drainage. Region 5 characterized by relatively high amount of bare soil surfaces also had moderate mean levels of the slum indicators. However, as opposed to Region 2, it exhibited extremely large standard deviations of these indicators.

Table 6.2 Kruskal-Wallis tests comparing distribution of 6 household variables across all 5 regions

<table>
<thead>
<tr>
<th>Test variable</th>
<th>Sig. (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% non-owner</td>
<td>0.044*</td>
</tr>
<tr>
<td>% with &gt;2 residents per room</td>
<td>0.012*</td>
</tr>
<tr>
<td>% living in single room or studio</td>
<td>0.317</td>
</tr>
<tr>
<td>% without flush toilet</td>
<td>0.768</td>
</tr>
<tr>
<td>% without pipe water</td>
<td>0.355</td>
</tr>
<tr>
<td>% without drainage</td>
<td>0.027*</td>
</tr>
</tbody>
</table>

*significant at 0.05, n = 63

Figure 6.3 Box Plot: Distribution of (%) non-owner households by region
Figure 6.4 Box Plot: Distribution of (%) overcrowded households by region

Figure 6.5 Box Plot: Mean % households without drainage by region
Figures 6.6 - 6.8 show which pairs of regions were significantly different with respect to the specific household indicators. In terms of per cent of non-owner households, Region 1 was significantly different (highest distribution) from Region 3 (lowest distribution) (Figure 6.6). Region 1 also had a significantly higher distribution of the overcrowded households (with more than 2 residents per room) when compared to Region 4 (Figure 6.7). Finally, Regions 3 and 4 were significantly different in terms of the distribution households without drainage (Figure 6.8).
Figure 6.7 Pair wise comparisons of Regions: % households with >2 residents per room

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.000-3.000</td>
<td>9.292</td>
<td>8.037</td>
<td>1.156</td>
<td>.248</td>
<td>1.000</td>
</tr>
<tr>
<td>4.000-2.000</td>
<td>12.958</td>
<td>9.113</td>
<td>1.422</td>
<td>.155</td>
<td>1.000</td>
</tr>
<tr>
<td>4.000-5.000</td>
<td>-16.193</td>
<td>6.170</td>
<td>-2.625</td>
<td>.009</td>
<td>.087</td>
</tr>
<tr>
<td>4.000-1.000</td>
<td>34.208</td>
<td>10.523</td>
<td>3.251</td>
<td>.001</td>
<td>.012</td>
</tr>
<tr>
<td>3.000-2.000</td>
<td>3.667</td>
<td>9.606</td>
<td>.382</td>
<td>.703</td>
<td>1.000</td>
</tr>
<tr>
<td>3.000-5.000</td>
<td>-6.901</td>
<td>6.877</td>
<td>-1.003</td>
<td>.316</td>
<td>1.000</td>
</tr>
<tr>
<td>3.000-1.000</td>
<td>24.917</td>
<td>10.953</td>
<td>2.276</td>
<td>.023</td>
<td>.229</td>
</tr>
<tr>
<td>2.000-5.000</td>
<td>-3.234</td>
<td>8.199</td>
<td>-0.399</td>
<td>.690</td>
<td>1.000</td>
</tr>
<tr>
<td>2.000-1.000</td>
<td>21.250</td>
<td>11.785</td>
<td>1.806</td>
<td>.071</td>
<td>.709</td>
</tr>
<tr>
<td>5.000-1.000</td>
<td>10.015</td>
<td>9.696</td>
<td>1.064</td>
<td>.062</td>
<td>.624</td>
</tr>
</tbody>
</table>

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 6.8 Pair wise comparisons of Regions: % households without drainage

<table>
<thead>
<tr>
<th>Sample1-Sample2</th>
<th>Test Statistic</th>
<th>Std. Error</th>
<th>Std. Test Statistic</th>
<th>Sig.</th>
<th>Adj.Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000-2.000</td>
<td>11.633</td>
<td>9.646</td>
<td>1.227</td>
<td>.220</td>
<td>1.000</td>
</tr>
<tr>
<td>3.000-5.000</td>
<td>-12.693</td>
<td>6.905</td>
<td>-1.838</td>
<td>.066</td>
<td>.660</td>
</tr>
<tr>
<td>3.000-4.000</td>
<td>-23.917</td>
<td>8.070</td>
<td>-2.964</td>
<td>.003</td>
<td>.030</td>
</tr>
<tr>
<td>3.000-1.000</td>
<td>26.833</td>
<td>10.998</td>
<td>2.440</td>
<td>.015</td>
<td>.147</td>
</tr>
<tr>
<td>2.000-5.000</td>
<td>-0.859</td>
<td>8.142</td>
<td>-0.106</td>
<td>.916</td>
<td>1.000</td>
</tr>
<tr>
<td>2.000-4.000</td>
<td>-12.083</td>
<td>9.151</td>
<td>-1.320</td>
<td>.187</td>
<td>1.000</td>
</tr>
<tr>
<td>2.000-1.000</td>
<td>15.000</td>
<td>11.813</td>
<td>1.270</td>
<td>.204</td>
<td>1.000</td>
</tr>
<tr>
<td>5.000-4.000</td>
<td>11.224</td>
<td>6.195</td>
<td>1.812</td>
<td>.070</td>
<td>.700</td>
</tr>
<tr>
<td>5.000-1.000</td>
<td>14.141</td>
<td>9.706</td>
<td>1.457</td>
<td>.145</td>
<td>1.000</td>
</tr>
<tr>
<td>4.000-1.000</td>
<td>2.917</td>
<td>10.566</td>
<td>.276</td>
<td>.783</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.
6.3 Discussion

6.3.1 Analytically derived neighborhoods as ideal spatial units for mapping slum

This chapter demonstrated the outcome of combining smaller statistical areal units to derive larger contiguous units (regions) that are homogenous within themselves and quite different from each other in at least one physical attribute. The Regions were meant to simulate city neighborhoods which would be more ideal for mapping slum conditions. The original units had been created only for the practical convenience of collecting disaggregated household data. While each original unit could be individually characterized in terms of image and household attributes (a portrayed in Chapter V), they did not capture the true spatial configuration of conditions within the city nor represent the true extent of slum areas. They would have been vulnerable to spatial aggregation problems such as the modifiable area unit problem (Openshaw & Openshaw, 1984).

Although the "neighborhood" concept more often than not leans on the socio-economic and cultural segregation within a city (Weeks et al., 2010), it can also be defined based on physical parameters (Getis, 2015). In this study, the latter was achieved by maximizing the homogeneity function among adjacent units with regards to several images derived land cover and terrain attributes. If we consider the simple notion that the urban physical environment is a reflection or product of the people living within, then the physically homogenous regions could simulate neighborhoods in social terms as well. By employing image derived attributes that are specifically associated the slum conditions, it was also possible by extension to determine the slum profile each region. Region 1 stood apart as having significantly the highest proportion of overcrowded and non-owner households. This region was made up of four adjacent statistical areas located close to the escarpment. It the most evident contiguous slum area or neighborhood within the city. In the same vein, Region 3 could be the least likely slum neighborhood or area within the city.
The results also showed that not all Regions could be cleanly labeled as a slum or non-slum neighborhood. While it was possible to distinguish extremities (i.e. the most and least slum-like Regions such as Regions 1 and 3 respectively), others (such as Regions 2 and 5) were not so easily defined. Region 5 was a large contiguous area extending across almost half of the city. Though it had moderate mean levels of slum indicators, the large range and high standard deviations within made it difficult to distinguish it from both high slum and low-slum areas. Due to its considerable size, Region 5 could not be set apart as a useful neighborhood unit in the real sense of the term. It was more likely a combination of adjacent areas that could not be adequately differentiated in the physical sense. The fact that such a large physically uniform area could not be labeled easily as a slum/non-slum was evidence that the study featured a 'cheek by jowl' arrangement, where households of vastly different socio-economic profiles and housing standards lived in close proximity (Lees, 2008).

It is also evident that the individual image derived attributes were not of equal importance in determining if a region should be profiled as a slum neighborhood or not. For example, the main distinguishing features of Regions 2 and 5 was high roof density and a high proportion of dirt roads respectively. These traits would be considered elsewhere as indicating slum status (Baud et al., 2010; Hofmann et al., 2008; Owen & Wong, 2013). However, Regions 2 and 5 did not portray the highest (or lowest) mean levels of any household variable suggesting that these characteristics were not particularly relevant for slum identification in the local context. For example, the high amount of dirt roads surfaces could be attributed to the local area’s history of slow execution of public works projects such as residential road construction. In fact, outside of the major/trunk road network, very little of the road surface in Bamenda is paved. Contrast this with the main homogenizing attribute for Region 1, i.e., location on steep slopes. This attribute clearly more
important in the local context for delineating slum neighborhood (Region 1 had the most overcrowded and unsecure tenure conditions). A similar result was obtained by Duque et al. (2012b) where analytic regions associated with the most critical housing conditions were also located on steep slopes at periphery of Medellin Colombia.

6.3.2 Explaining derived regions and slum conditions in the context of local place knowledge

In many urbanities in SSA, pre-existing knowledge of city neighborhoods mostly exists in administrative (Duque et al., 2012b) or vernacular terms (Getis, 2015; Weeks et al., 2007). This is equally the case with Bamenda where local names such as "Azire", "Bayelle", "Sisia", "Ntamulung", "Old Town", "Meta Quarters", "New Lay Out", "Up Station", amongst others, are used to identify residential neighborhoods (Figure 6.9). As the names suggest, most of these neighborhoods are derived from the pre-colonial/pre-urban indigenous settlements in the area. Though these vernacular neighborhoods have persisted over time, they are of limited utility for rigorous urban mapping due to lack of objectivity associated with their delineation, ambiguous boundaries and rigidity in the face of dynamic forces that have altered the urban social and physical fabric over time (Getis, 2015).
Despite their limitations, however, vernacular neighborhood units proved very useful in the qualitative validation of the quantitatively derived regions. This was done by simply determining the extent to which local knowledge reflected what had been quantitatively deduced. A cursory Internet (Google) search with the key words "Slums" and "Bamenda" was used to find out what public knowledge was there about slums in Bamenda. This returned a few relevant results which, despite not offering a comprehensive academic appraisal of the subject, still provided some useful information. All the returned online articles were consistent in highlighting one specific neighborhood, namely the "Sisia" (also "Sesia" or "Sessia") quarter as a prominent slum area in the city. Looking back at Figure 6.16, “Sisia” neighborhood spatially correlates with the Region 1

Figure 6.9 Map of derived regions overlaid with vernacular neighborhood names

2 http://bamendaonline.net/blog/bda-sesia-earmarked-for-major-urban-upgrading/
http://allafrica.com/stories/200810160942.html
http://www.cameroonpostline.com/un-habitat-govt-to-eradicate-bamenda-iii-slums/
thereby validating the earlier analytic delineation of the latter as slum neighborhood. Similarly, "Up station" is known locally as a government residential area located at the top of the Bamenda Escarpment. Some local features that are a testament to this are: Governor's Residence, Army Camp, Court houses, Government offices, high quality residential units. It is not surprising therefore that this area corresponded spatially with Region 3 which showed over all the least slum like conditions. Region 5 however correlates with no less than 10 vernacular neighborhoods which also explains why it exhibited a wider range of household attributes.

6.4 Summary

This chapter focused on the issue of scale. The objective was to quantitatively determine optimal spatial units for slum mapping slum conditions. Analytic neighborhoods were created by spatial clustering of statistical units, using the minimum spanning tree algorithm and select land cover and terrain attributes. The derivatives were contiguous areas or regions that were most homogenous in at least one physical attribute. Five such regions were created and Kruskal Wallis tests revealed that they were significantly different in terms of aggregated household indicators. Regions 1 and 3 stood out and could be clearly labeled as a slum and non-slum neighborhoods respectively. They were also spatially correlated with well-known vernacular neighborhoods that are locally perceived as occupying opposite ends of socio-economic and household deprivation spectrum
7. CONCLUSION

7.1 Chapter overview

The purpose of this chapter is to provide a fitting conclusion to this dissertation research report. The chapter reviews the research problem and how the main findings contributed to filling the relevant gaps in literature. This is followed by a discussion on the broader implications of these findings, especially how they may impact slum related policy and practices. The last section states the limitations of the study and suggests avenues for future research.

7.2 The contribution of key findings

This dissertation research constituted a multi-pronged attack on the local slum mapping problematic. The general objective was to investigate various improved strategies and techniques for correctly identifying local slum conditions in Bamenda, Cameroon. The rationale was borne from three problems: 1) existing methods used for slum measurement, such as the use of single aggregate indices, were too simplified and could not account for the diversity that is common in most localities; 2) slum prediction from image data relied too heavily on prior ground based data; and 3) slums were too often mapped using arbitrary spatial units which did not always represent their true extent. Relying on both image and survey data, this research attempted to tackle the above problems by using variable clustering (for factor analysis), hierarchical clustering (for learning slums conditions from imagery) and analytic regionalization (for neighborhood unit delineation).

The findings from carrying out Objective I revealed that multiple household indicators
could be collapsed into fewer latent slums factors, each reflecting a different dimension of the slum problem. When quantified as binaries, these factors were validated using settlement location theory. The results also demonstrate the effectiveness of using hierarchical variable clustering as an unconventional but effective and easier to understand surrogate for exploratory factor analysis (EFA). This is particularly important as traditional EFA may not always be eligible given certain data conditions.

The results from carrying Objective II revealed that local areal units could be classified using hierarchical clustering into settlement types that had distinct physical features. The differences in physical attributes could be used to infer differences in household slum conditions. For example, the “steep slope” physical type had significantly higher mean household slum conditions. This means that physical settlement attributes can be relied on to profile local housing conditions even when housing data is unavailable.

Objective III was designed specifically to tackle the problem of scale effects on mapping. The results revealed that the original spatial units could be grouped intelligently, using image derived land cover attributes and spatial constraints, to derive larger contiguous units. These derived areas were each homogenous with regards to some physical attribute and could serve thus as proxies for the “neighborhood” scale. Direct comparison revealed that derived ‘neighborhoods’ were also significantly different from each other in term so household conditions. This was used to identify the largest possible contiguous slum or non-slum neighborhood within the city.

7.3 Policy implications of the study

The overall strength of the methods investigated in this research lay in their data-driven nature. This means that the results were fully sensitive to the local conditions, in this case Bamenda, Cameroon. This is ideal for mapping a phenomenon such as slums, which inherently
suffers from high heterogeneity.

The findings of this research have significant implications for local urban planning policy and practice. The first and most obvious area that this research impacts is in optimizing the provision and improvement of basic services such as water. The multi-index approach to mapping slums allows for a spatially diverse perspective of household conditions in an area. This is a more accurate and informative way of depicting local slum conditions, as opposed to using a single “tell-all” index that would simply measure the totality of slums attributes. This is akin to how factorial ecology helped in the understanding of the spatial structure of cities from a socio-economic perspective.

On the policy front, the multi-index measurement approach could allow for intervention efforts to be compartmentalized. Areas dominated by a specific service problem, e.g. lack of piped water, could receive focused attention of a nature that is tailored only to that problem. A good example could be the establishment of more stand pipes in a locale or improving the service network infrastructure. This would be more resource-efficient from the service provider’s point of view. In the same vein, an area dominated by problems such as overcrowding would receive a different but suitable type of intervention.

Slum clearance has been on the rise on recent decades and often has significant consequences for the livelihoods of many urban residents in developing countries. A further implication of multi-index mapping is that, by being able to pin-point areas with unique critical issues, clearance activities can now be designed to be more surgical, affecting fewer people than would be the case of blanket targeting of slum residents. This would help dilute the overall negative impacts of such controversial practices.

Public health can also be impacted by multi-index mapping. The ability to pin-point areas
that have the most critical water and sanitation issues means it is possible to determine urban residents most vulnerable to a public health crisis. Outbreaks of diseases such as cholera are frequent in developing countries and are highly dependent on the quality of water and sanitation facilities. Health officials can tailor sensitization efforts in relevant areas and can make sure local health clinics and centers are equipped for such eventualities.

In terms of urban planning practice, this research’s findings can lead to improvement in operational efficiency, such as in the monitoring of slum areas. First, remote sensing data is not only cheaper to obtain, but provides better space-time coverage. It is also immune to political machinations, falsification, and bias. When it comes to mapping of slums from image data, the unsupervised classification approach is an inductive learning process, not unlike natural human learning. It does not require prior knowledge of slums in the area nor does it depend on rigid slum stereotypes. From a practical perspective, this means unsupervised classification offers greater efficiency and automation in the mapping process. It also means the possibility of undertaking rapid and frequent slum appraisals at city wide scales without the need for ground survey. This could be useful very useful for resource challenged urban centers like Bamenda. With the high dynamism within slum areas and the high frequency with which imagery can be acquired, this approach can form the platform for developing a cost-effective monitoring tool.

This research can significantly impact urban spatial planning. In this regard, the implications of quantitatively deriving contiguous neighborhood-like units are huge. This is especially true if said derived units can be easily labeled as slums or not. Analytic regionalization helps to mitigate problems like the MAUP by revealing the true spatial extent of the slum hotspots. This would render geographic targeting much more accurate and effective. Not only is it possible to determine the general location of deprivation, but also the true extent of it. The use of analytic
neighborhoods is superior to the use of communes (too coarse) or vernacular neighborhoods, both of which are currently the only officially recognized forms of spatial organization within the city of Bamenda. From the results in this study, it can for example be suggested that local authorities in Bamenda redefine/redesign the city neighborhood boundaries to reflect the analytic framework. This can in turn positively impact several domains.

In terms of public health, it would allow for optimal allocation of local clinics and health centers, each tailored to the needs of the immediate and relatively homogenous population. Similarly, non-critical service industry entities such as banks and credit unions can deploy local branches offices with services and offerings tailored to the local community. Furthermore, the analytic regions can serve future purposes such as provide a better basis for spatial sampling for surveys and censuses. Future studies can now include for example, a “slum sample” in their analyses.

Climate change mitigation and disaster management are areas that can also be significantly impacted by this study. Derived analytic regions for example can also serve as a platform of spatial understanding of environmental justice and equity issues. Climate change mitigation actions such as tree planting could be guided by careful consideration of urban structure in term of physical and housing attributes. This framework can help determine if existing tree planting efforts are biased in favor of more affluent or less deprived neighborhoods. The ability to define slum areas based on physical attributes can also be harnessed for disaster planning. For example, areas located near steep slopes and escarpments are most vulnerable to landslides. If such areas also happen to have the worst housing conditions, as is the case with Bamenda, this would further magnify the problem and hence need for action.

Finally, the study could significantly impact upon urban security. Though slums are today
largely defined purely in terms of housing and settlement attributes, it is a well-known fact that slum areas around the world constitute some of the most insecure and outright dangerous places for human habitation. This is compounded by the high densities and lack of roads which limits accessibility to such areas. The ability to quickly profile and map urban areas based on their physical attributes can be a useful tool for law enforcement and social workers to prioritize and monitor areas that have high potential for becoming hotspots for crime and unrest.

7.4 Limitations to the study and avenues for future research

One limitation in this study is the use of only 6 original slum variables for multi-index slum measurement. For convenience, these variables were chosen as they closely fit the UN-Habitat point criteria for defining slums. However, it must be acknowledged that there are many more slum related indicators that could be measured at the household level and that could lead to the discovery of more latent slum factors. It is therefore recommended that additional variables be included in future studies. Such variables do not have to be limited to universally recognized slum indicators and may even be derived intuitively or from qualitative local surveys.

The results of image based unsupervised classification of areal units into physical slum types, or their combination into analytic neighborhoods, is limited to the specific case of Bamenda and the time of image acquisition (i.e., December 2012). Considering the dynamic and varied nature of urban areas, it is recommended that the evidence established in this research be extended by examining additional urban areas in the region. This will confirm the general usefulness of these methods. It is also recommended that longitudinal element be added by using multi-date imagery.

In the case of deriving analytics neighborhoods, this study only made use of one algorithm (the minimum spanning) as it was conveniently part of the ArcGIS software suite. However, there are many algorithms in existence with different mathematical approaches and hence potentially
different mapping outcomes. It may be informative to undertake a study that looks at the impact of using different regionalization techniques on the ability to accurately delineate slum boundaries.
REFERENCES


Arimah, B. C. (2010). *The face of urban poverty: Explaining the prevalence of slums in*


mathematical modeling, 265-296.


Duque, J. C., Royuela, V., & Noreña, M. (2012b). A Stepwise Procedure to Determine a Suitable Scale for the Spatial Delimitation of Urban Slums *Defining the Spatial Scale in*
Modern Regional Analysis (pp. 237-254): Springer.


based characterization of formal and informal neighborhoods in an urban landscape.


http://www.economist.com/node/9070726


Haw, G. (1899). *No Room to Live, the Plaint of Overcrowded London*: "The Daily news" office.


UN-HABITAT. (2002). Defining slums: Towards an operational definition for measuring slums. Background Paper 2, Expert group meeting on slum indicators, October. *October. UN Habitat, Nairobi.*


Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35.

### APPENDIX I – INDEPENDENT T TESTS RAW OUTPUT

#### Occ_D groups by location

<table>
<thead>
<tr>
<th>Location</th>
<th>Levene's Test for Equality of Variances</th>
<th>Independent Samples Test for Equality of Means</th>
<th>( t )-test for Equality of Means</th>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
<td>Df</td>
<td>Sig. (2-tailed)</td>
<td>Mean</td>
<td>Std. Error</td>
</tr>
<tr>
<td>to_mkts</td>
<td>Equal variances assumed</td>
<td>6.071</td>
<td>017</td>
<td>2.576</td>
<td>61</td>
<td>012</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to_cbd</td>
<td>Equal variances assumed</td>
<td>3.023</td>
<td>087</td>
<td>479</td>
<td>61</td>
<td>634</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to_rds</td>
<td>Equal variances assumed</td>
<td>6.649</td>
<td>012</td>
<td>2.196</td>
<td>61</td>
<td>032</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### UT_D groups by location

<table>
<thead>
<tr>
<th>Location</th>
<th>Levene's Test for Equality of Variances</th>
<th>Independent Samples Test for Equality of Means</th>
<th>( t )-test for Equality of Means</th>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>95% Confidence Interval of the Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
<td>Df</td>
<td>Sig. (2-tailed)</td>
<td>Mean</td>
<td>Std. Error</td>
</tr>
<tr>
<td>to_mkts</td>
<td>Equal variances assumed</td>
<td>5.719</td>
<td>020</td>
<td>-1.133</td>
<td>61</td>
<td>262</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to_cbd</td>
<td>Equal variances assumed</td>
<td>2.900</td>
<td>094</td>
<td>-1.557</td>
<td>61</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to_rds</td>
<td>Equal variances assumed</td>
<td>2.048</td>
<td>157</td>
<td>-1.385</td>
<td>61</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX II – IMAGE CLASSIFICATION ACCURACY

Error/Confusion Matrix

<table>
<thead>
<tr>
<th>Class</th>
<th>roof</th>
<th>bare soil</th>
<th>paved road</th>
<th>dirt road</th>
<th>shadow</th>
<th>open space</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>roof</td>
<td>99.61</td>
<td>2.15</td>
<td>3.32</td>
<td>3.17</td>
<td>39.67</td>
<td>0</td>
<td>25.67</td>
</tr>
<tr>
<td>bare soil</td>
<td>0.3</td>
<td>93.7</td>
<td>0</td>
<td>11.21</td>
<td>0</td>
<td>0</td>
<td>20.43</td>
</tr>
<tr>
<td>paved road</td>
<td>0</td>
<td>0</td>
<td>96.41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18.05</td>
</tr>
<tr>
<td>dirt road</td>
<td>0.1</td>
<td>4.15</td>
<td>0.27</td>
<td>85.62</td>
<td>0.3</td>
<td>0</td>
<td>9.6</td>
</tr>
<tr>
<td>shadow</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>54.53</td>
<td>0</td>
<td>5.55</td>
</tr>
<tr>
<td>open space</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.41</td>
<td>100</td>
<td>20.68</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Commission and Omission errors

<table>
<thead>
<tr>
<th>Class</th>
<th>Commission (%)</th>
<th>Omission (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>roof</td>
<td>21.13</td>
<td>0.39</td>
</tr>
<tr>
<td>bare soil</td>
<td>5.84</td>
<td>6.3</td>
</tr>
<tr>
<td>paved road</td>
<td>0</td>
<td>3.59</td>
</tr>
<tr>
<td>dirt road</td>
<td>9.92</td>
<td>14.38</td>
</tr>
<tr>
<td>shadow</td>
<td>0</td>
<td>45.47</td>
</tr>
<tr>
<td>open space</td>
<td>2.67</td>
<td>0</td>
</tr>
</tbody>
</table>
SPSS output of Principal Component Analysis on physical settlement attributes used for Hierarchical Clustering

Communalities

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Initial</th>
<th>Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>% paved road</td>
<td>1.000</td>
<td>0.680</td>
</tr>
<tr>
<td>% dirt road</td>
<td>1.000</td>
<td>0.894</td>
</tr>
<tr>
<td>% bare soil</td>
<td>1.000</td>
<td>0.863</td>
</tr>
<tr>
<td>% roof</td>
<td>1.000</td>
<td>0.967</td>
</tr>
<tr>
<td>shadow/roof ratio</td>
<td>1.000</td>
<td>0.664</td>
</tr>
<tr>
<td>% open</td>
<td>1.000</td>
<td>0.989</td>
</tr>
<tr>
<td>slope</td>
<td>1.000</td>
<td>0.770</td>
</tr>
</tbody>
</table>

Extraction Method: PCA

Total Variance Explained

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigenvalues</th>
<th>Extraction Sums of Squared Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>% of Variance</td>
</tr>
<tr>
<td>1</td>
<td>4.800</td>
<td>68.577</td>
</tr>
<tr>
<td>2</td>
<td>1.027</td>
<td>14.677</td>
</tr>
<tr>
<td>3</td>
<td>0.754</td>
<td>10.769</td>
</tr>
<tr>
<td>4</td>
<td>0.360</td>
<td>5.136</td>
</tr>
<tr>
<td>5</td>
<td>0.031</td>
<td>0.443</td>
</tr>
<tr>
<td>6</td>
<td>0.028</td>
<td>0.397</td>
</tr>
<tr>
<td>7</td>
<td>9.162E-007</td>
<td>1.309E-005</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis.

Component Matrix

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>paved_pct</td>
<td>.657</td>
<td>.498</td>
</tr>
<tr>
<td>dirt_pct</td>
<td>.914</td>
<td>-.244</td>
</tr>
<tr>
<td>bare_pct</td>
<td>.892</td>
<td>-.260</td>
</tr>
<tr>
<td>roof_1</td>
<td>.968</td>
<td>.175</td>
</tr>
<tr>
<td>Shad</td>
<td>.804</td>
<td>.133</td>
</tr>
<tr>
<td>Open</td>
<td>-.975</td>
<td>-.197</td>
</tr>
<tr>
<td>slope</td>
<td>.453</td>
<td>.751</td>
</tr>
</tbody>
</table>

Extraction Method: PCA

a. 2 components extracted.
APPENDIX IV – KRUSKAL WALLIS TESTS RAW OUTPUT

Physical type by household slum variable

<table>
<thead>
<tr>
<th>Hypothesis Test Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Hypothesis</td>
</tr>
<tr>
<td>1. The distribution of non_owner is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>2. The distribution of GT2_residents_per_room is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>3. The distribution of room_or_studio is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>4. The distribution of no_flushtoilet is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>5. The distribution of no_pipedown_water is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>6. The distribution of no_drainage is the same across categories of settlement type (physical attributes).</td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.
**Physical type by location attribute**

<table>
<thead>
<tr>
<th>Hypothesis Test Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Hypothesis</td>
</tr>
<tr>
<td>The distribution of distance to marketplace is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>The distribution of distance to CB is the same across categories of settlement type (physical attributes).</td>
</tr>
<tr>
<td>The distribution of distance to roads is the same across categories of settlement type (physical attributes).</td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.

**Regions by household slum variable**

<table>
<thead>
<tr>
<th>Hypothesis Test Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Hypothesis</td>
</tr>
<tr>
<td>The distribution of non_owner is the same across categories of city neighborhoods (regions).</td>
</tr>
<tr>
<td>The distribution of G72_residents_per_room is the same across categories of city neighborhoods (regions).</td>
</tr>
<tr>
<td>The distribution of room_or_studio is the same across categories of city neighborhoods (regions).</td>
</tr>
<tr>
<td>The distribution of no_flushtoilet is the same across categories of city neighborhoods (regions).</td>
</tr>
<tr>
<td>The distribution of no_pipable_water is the same across categories of city neighborhoods (regions).</td>
</tr>
<tr>
<td>The distribution of no_drainage is the same across categories of city neighborhoods (regions).</td>
</tr>
</tbody>
</table>

Asymptotic significances are displayed. The significance level is .05.