Population Estimation Mining From Satellite Imagery

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by
Kwankamon Dittakan

September, 2015

Abstract

The collection of census data is an important task with respect to providing support for decision makers. However, the collection of census data is also resource intensive. This is especially the case in areas which feature poor communication and transport networks. In this thesis a number of methods are proposed for collecting census data by applying prediction techniques to relevant satellite imagery. The test site for the work is a collection of villages lying some 300km to the northwest of Addis Ababa in Ethiopia. The idea is to build a predictor that can label households according to "family" size. To this end training data has been obtained by collecting "on ground" census data and matching this up with satellite imagery. The fundamental idea is to segment satellite images so as to obtain satellite sub-images describing individual households and representing these segmentations using a number of proposed representations: graph-based, histogram based and texture based. By pairing each represented household with the collated census data, namely family size, a predictor can be constructed to predict household sizes according to the nature of each representation. The generated predictor can then be used to provide a quick and easy mechanism for the approximate collection of census data that does not require significant resource.

Contents

List of Figures

Acknowledgement

First and Foremost, I am deeply indebted and heartily thankful to my first supervisor, Professor Frans Coenen: for providing a lot of innovative ideas, his constant patience, ongoing support and encouragement since I started my Ph.D. programme of work in October 2011. His enthusiasm, constructive criticism, research ideas and his ability to always make time for discussions, have made the completion of my Ph.D. possible and highly enjoyable. I am privileged to have worked with him. I would also like to express my gratitude to: my second supervisor, Dr Rob Christley, for his assistance, suggestions and valuable comments; Judy Bettridge for assistance in supplying the "ground truth survey" data sets; and Dr Rahul Savani for first suggesting the use of regression techniques. I am also grateful to Prince of Songkla University for their financial support, and would like to express my gratitude to my research group and my friends in the Agent laboratory for helping and cheering me during my time in Liverpool. I am eternally indebted to my parents and my family who are most important in my life. I could not have completed this thesis without their constant love, understanding, support and everything. Lastly, I offer my regards and blessings to all of those not specifically named here who supported me in many ways during the completion of my study.

Chapter 1

Introduction

1.1 Introduction

A census is the procedure of acquiring and collecting information about the nature of the population of a given area, it is seen as an important mechanism whereby information can be obtained to support decision makers. Census data is widely used with respect to a variety of central and local government management and planning activities so that informed decisions can be made and budgets set. With respect to the work reported in this thesis census data is equated to population size (in many cases census data incorporates a range of additional data associated with individuals such as occupation, martial status, income bracket and so on). There are many problems associated with the collection of census data, especially in the case of national censuses. The first problem is census budget, the collection of census data requires a considerable resource in terms of money and "manpower". Another problem is the cost of processing the data after it has been collected. A third issue is that there is often a lack of good will on behalf of a population to participate in a census, even if they are legally required to do so, because people are often suspicious of the motivation behind censuses (especially when collected by government organisations) [45]. These problems are compounded in areas where there are poor communication and transport infrastructures; and/or an extensive, but sparsely populated, hinterland.

The solution argued for in this thesis is the acquisition of census data using remote sensing technology, namely satellite imagery. This is not an entirely new idea, the suggestion that satellite imagery could be used for census collection was first postulated in [26], where the main idea was to use "nightlight" satellite images to produce population census data at the "sub-district level". Most of the previously reported work on the usage of satellite imagery for census collection [103] has adopted a large scale (macro) approach, where satellite image data is typically used to estimate the population density distribution with respect areas or regions such as one kilometre "blocks". The approach proposed in this thesis is directed at the household level (thus a micro approach). The proposed micro-level based approach is particularly well suited to rural areas where satellite imagery is less "clustered". This offers the additional advantage that in rural area census collection is typically more resource intensive than in the

case of urban areas (because, as noted above, rural infrastructure tends to be less sophisticated than that of urban areas).

The main advantage of census collection using satellite imagery is reduced cost, the required satellite imagery typically is publicly available from websites such as Google Maps, Google Earth, the US National Aeronautics and Space Administration (NASA) and the US National Oceanic and Atmospheric Administration (NOAA). The data collection cost is therefore comparatively negligible. This in turn means that census collection can take place whenever data is required and not on some fixed cycle. The second advantage is that it is non-intrusive, thus overcoming the frequently encountered resistance to the collection of census data. In the context of the micro level approach, as presented in this thesis, there is an additional advantage that much more detailed census (population) data can be obtained than obtained using the macro level approaches typically reported in the literature. The general disadvantage of the use of remote sensing technology for census data collection is that it is not as accurate as in the case of "on-ground" surveys (there is always a trade of between resource reduction and accuracy); although, as will be demonstrated later in this thesis, good accuracies can be obtained. Another potential disadvantage is that it is desirable, for best results, that up to date satellite imagery is used. Typically Google Earth and the Google static map service update their imagery over a one to three years cycle depending on the region. In case of the Ethiopia data used for evaluation purpose later in this thesis a three year cycle was noted, probably because of the remote nature of the region and because rural areas feature less annual change than urban areas.

The fundamental idea presented in this thesis is thus that census (population size) data can be collected using classification and/or regression techniques applied to relevant satellite imagery. The research described is thus directed at mechanisms for the end-to-end process of building a classification or regression model that can predict household "family size" according to the nature of households captured from satellite imagery. More specifically the idea is to segment satellite images so as to obtain pixel collections describing individual households and represent these collections using some appropriate representation to which a classifier or regression model generator can be applied.

Classification is the process of building a classifier (model) describing data classes (categories). The classifier is derived using labeled training data. Classification has been successfully applied in many areas including medical diagnosis, weather prediction, credit approval, customer segmentation and fraud detection [5, 47, 202]. Regression analysis is a statistical process for estimating the relationships among variables which is widely used for prediction and forecasting. Regression analysis has also been successful applied in various application domains; examples include: medicine, economics and engineering [15, 73, 95]. The distinction is that classification is used to predict discrete and/or unordered class labels (categories) while regression is used to predict some continuous numerical data value. The accuracy of such predictive models is determined by applying it to pre-labeled test data.

The rest of this introductory chapter is organised as follows. The research objectives and

associated research issues and challenges are presented in Section 1.2. The research methodology used to address the research challenges, including the "criteria for success", is presented in Section 1.3. The contributions of the research work, including an itemisation of the published work to date arising from the research, is presented in Section 1.4. A summary of the publications [36, 37, 38, 39, 197] that have arisen out of the work presented in this thesis is given in Section 1.5. Followed by an overview of the rest of this thesis in Section 1.6 and a summary of this chapter in Section 1.7.

1.2 Thesis Objectives

From the foregoing the research domain at which this thesis is directed is thus concerned with the investigation, realisation and evaluation of algorithms and processes that can be used to build classification/regression models for the purpose of collecting population census data. This research objective is encapsulated by the following research question:

What are the most appropriate end-to-end computational processes required to collect population census data from satellite imagery using classification and regression techniques?

The resolution of this research question encompass a number of research challenges. These are articulated below in the form of a series of subsidiary research questions:

1. What are the most appropriate mechanisms for segmenting a given satellite image so that appropriate individual household sub-images (if any) can be identified?
2. Given a set of identified household images how should the content of those images be represented so that compatibility with classification and regression model generation is achieved while at the same time ensuring that key information is retained?
3. When representing household images what is the nature of the key information to be captured?
4. What are the most appropriate classification/regression techniques for predicting census data given a processed collection of household images?
5. What is the process for conducting a large scale census comprising many satellite images?
6. In the context of conducting large scale surveys how can issues associated with "overlapping" satellite images best be resolved?

The thesis sets out to provide answers to the above.

1.3 Research Methodology

To act as a focus for the work two rural areas of the Ethiopian hinterland, located some 300 km to the northwest of Addis Ababa, were used. These areas were selected because details concerning individual households had been collected by University of Liverpool field staff. This information included household size in terms of number of people normally resident and location latitude and longitude. Figure 1.1 shows part of a satellite image covering one of these districts. From the figure we can clearly observe several households. The data was collected as part of a field study, investigating the health of chickens, conducted by the School of Veterinary Science at the University of Liverpool in May 2011 and July 2012. The data was collected using a sampling process; thus, given a particular village data for only some of the households, dispersed across the village, was available. For the purpose of the research, satellite images were obtained using the Google Earth and Google Static Map services (although clearly other forms of satellite imagery could equally well have been used). The collected latitude and longitude data was used to identify appropriate satellite images that covered the geographical area of interest.

The first step in the adopted methodology was to investigate mechanisms whereby the households could be isolated using image segmentation techniques. This included mechanisms that could be adopted for image cleaning. The idea was to produce a set of "household images" that would provide the foundation for further processing. So that appropriate classification/regression models could be generated. The second step was to investigate methods for representing household images in such a way that: (i) compatibility with classification/regression model generation techniques was obtained and (ii) information loss was minimised. A review of the existing literature concerning image classification suggested three broad categories of representation technique; (i) graph-based, (ii) colour histogram based and (iii) texture based.

The next step (step three) in the proposed methodology was to consider a variety of classification and regression model generators. From the literature there are a great many of these with no clear "best" model generator. To identify the most appropriate the idea was to conduct a significant amount of evaluation combining each of the proposed representations with a number of different generators. The criteria for success in this context was prediction accuracy, comparison of predicted household sizes with known household sizes. With respect to the evaluation presented later in this thesis, in the context of the classification models, results were considered in term of a set of metrics: accuracy, specificity, sensitivity, Area Under Receiver Operating Curve (AUC) and the F-Measure; of which AUC was considered to be the most significant. Later in this thesis results are presented in the form of tables and bar charts, with the later focusing only on the AUC values from the tables. In each case the result from a Friedman significance test and a post hoc Nemeyi test is also presented. The first is to establish if the results from the table are indeed statistically significant or not (whether we can reject the null hypothesis $H0$ that there is no significant difference or not), and to present an alternative view

Figure 1.1: Example of a satellite image from the Google Static Map service.

of the data presented in the tables, and the second (where the $H0$ hypothesis has been rejected) is to attempt to identify where the statistical differences occur. In the context of the regression models the measures used were: correlation coefficient, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The next and final stage of the adopted methodology was to consider the process for conducting large scale population estimation mining using the techniques considered. To illustrate the process an entire village and its surrounding lands were considered, as opposed to individual households. For this purpose a rural area of Ethiopian, measuring 42.7 km^2, was selected that had a known (or at least reported) population size. Prior to conducting the large scale population estimation mining process it was necessary to first investigate mechanisms for acquiring large sets of satellite images (a "patchwork" of 600 satellite images was used for the study). This was followed by an investigation of the mechanisms for identifying individual households in such a way that the "overlap problem" could be addressed (the issue of the same household appearing in two or more overlapping satellite images). Evaluation was again conducted by comparing the reported (known) population size with the predicted population size using the same metrics as used previously (see above).

1.4 Contributions

The contributions of the research work presented in this thesis can be summarised as follows:

1. A novel approach for image segmentation specifically designed for segmenting individual households featured in a satellite image data set.

2. A household image representation founded on a quadtree based hierarchical decomposition of space together with a frequent subgraph mining algorithm for dimensionality reduction. The identified frequent subgraphs were arranged into a feature vector format, one vector per household, suited for input into a classification or regression model generation algorithm.

3. A household image representation founded on a colour histogram based approach. More specifically an image representation founded on multiple histograms extracted from various colour channels; a feature vector format was again used.

4. A household image representation founded on the concept of "texture" analysis. More specifically usage of Local Binary Patterns (LBPs), as before a feature vector format was again derived.

5. A detailed comparison of the proposed household image representations.

6. An analysis of a sequence of classifier generation algorithms so as to identify the most appropriate in the context of population estimation prediction from satellite data.

7. An analysis of a number of regression model generation algorithms so as to identify the most appropriate in the context of population estimation prediction from satellite data.

8. An effective mechanism for satellite image collection using the Google Static Maps service to obtain satellite image data for a specified area.

9. A novel approach for household detection specifically designed for the purpose of identifying and segmenting individual households featured in a satellite image data set covering a prescribed area.

10. A mechanism for detecting duplicated households in a given satellite image data collection so as to address the image "overlap" problem.

11. An end-to-end process for conducting large scale population estimation mining using satellite data.

12. Overall, the thesis presents an approach of population estimation founded on known techniques, but combining a new and novel methodology.

1.5 Publications

A number of research publications have arisen out of the work presented in this thesis. These are itemised below, in each case a short summary is given and a reference to where the material features in this thesis.

Kwankamon Dittakan, Frans Coenen, Rob Christley: *Towards The Collection of Census Data From Satellite Imagery Using Data Mining: A Study With Respect to the Ethiopian Hinterland*. SGAI Conf. 2012: 405-418.

This paper was the first to describe the proposed framework for remotely collecting census data using satellite imagery and data mining (classification). The main idea presented in this paper was that classifiers can be built that classify household satellite images to produce census data, provided an appropriate representation is used. The proposed representation was founded on the idea of representing segmented households using a colour histogram based formalism. The presented evaluation indicated that accurate census data can be collected using the proposed approach at a significantly reduced overall cost compared to traditional approaches to collecting such data. The work summarises some of the material presented in Chapter 5 where the detail of the proposed colour histogram based representation is presented.

Kwankamon Dittakan, Frans Coenen, Rob Christley: *Satellite Image Mining for Census Collection: A Comparative Study with Respect to the Ethiopian Hinterland*. MLDM 2013: 260-274.

As in the case of the previous paper this paper also presented the idea of using satellite imagery to generate census data from satellite imagery using a classifier for household size census prediction. The paper processes the Local Binary Pattern (LBP) representation. The presented evaluation indicated a particular variation of the LBP representation, called $LBP_{8,1}$, tended to produce the best results. The work summarises some of the material presented in Chapter 6 where the fundamental idea of texture analysis for satellite image representation, more specifically the LBP representation, is described.

Kwankamon Dittakan, Frans Coenen, Rob Christley, Maya Wardeh: *Population Estimation Mining Using Satellite Imagery*. DaWaK 2013: 285-296.

As in the case of the previous two papers, this paper also described the framework for population estimation mining (census mining) founded on the concept of applying classification techniques to satellite imagery. However, the particular note was the subgraph feature vector representation that was used to encode household imagery. The proposed framework was evaluated using test data, collected from two villages in the Ethiopian hinterland, also used in this thesis. The conducted evaluation indicated that when using a minimum support threshold of $\sigma = 10$ for the subgraph mining, good results could be obtained. The work summarises some of the material presented in Chapter 4 where the detail of the graph-based representation, using quadtree decomposition together with frequent subgraph mining, is described.

Kwankamon Dittakan, Frans Coenen, Rob Christley, Maya Wardeh: *A Comparative Study of Three Image Representations for Population Estimation Mining Using Remote Sensing Imagery*. ADMA 2013: 253-264.

This paper presents a summary of the usage of the three representation presented in this thesis for population estimation mining: (i) colour histogram based, (ii) Local Binary Pattern (LBP) based and (iii) graph-based. The presented evaluation indicated that the $LBP_{8,1}$ variation of the texture based representation produced the best overall result. The work summarises some of the material presented in Chapter 6, especially Sub-section 6.7.

Wen Yu, Frans Coenen, Michele Zito, Kwankamon Dittakan: *Classification of 3D Surface Data Using the Concept of Vertex Unique Labelled Subgraphs*. ICDM Workshops 2014: 47-54.

This paper describes the use of the concept of Vertex Unique Labelled Sub graph (VULS) mining for the use of localised classification of regions in 3D surfaces represented in terms of grid graphs. The evaluation was conducted using satellite image data where the ground surface is represented as a 3D surface with the z dimension describing greyscale value. The idea was to predict vertex labels describing land cover type. The significance of this paper is that the research presented in this thesis influenced the work in this paper.

1.6 Thesis Organisation

The organisation of the rest of this thesis is as follows. Chapter 2 provides an extensive literature review of population estimation using satellite imagery and the previous work concerning the technologies that feature in this thesis, including discussion concerning the processing of 2D image data. Chapter 3 describes the nature of the satellite image data sets, and the application domain, used as a focus for the work presented in this thesis. The nature of the necessary data preparation and image preprocessing applied to these data sets is also presented in this chapter. The three considered feature extraction approaches used for satellite image representation are described in the following three chapters. Chapter 4 presents the graph-based approach founded on a quadtree storage mechanism and a hierarchical decomposition coupled with the application of frequent subgraph mining techniques to identify frequently occurring "patterns" hidden in the identified household image data. Chapter 5 considers the proposed colour analysis based approach whereby colour histograms are used together with colour statistical features representing the identified household images. Chapter 6 presents the proposed texture analysis based approach which uses LBPs and texture statistical features to extract and represent the texture properties from identified household images. The evaluation of the three representations is also presented in Chapter 6. Chapter 7 is an investigation of the use of regression analysis for population size estimation. Chapter 8 then describes the proposed end-to-end large scale population estimation mining process and the evaluation of this process using a large scale study. Finally, in Chapter 9, the thesis is concluded with a summary, presentation and discussion of the main findings in the context of the research question and sub-questions identified above and some suggestions for future work.

1.7 Summary

This chapter has provided the necessary context and background for the research described in this thesis. In particular the motivation for the research and the thesis objectives have been detailed. A literature review of the related previous work, with respect to this thesis, is presented in the following chapter (Chapter 2).

Chapter 2

Background and Literature Review

2.1 Introduction

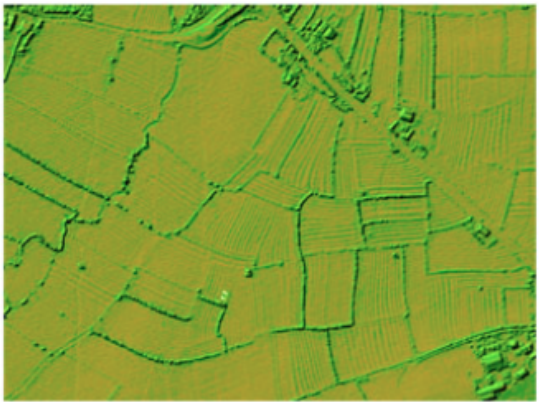
A review of the background and previous work with respect to the research presented in this thesis is presented in this chapter. The chapter starts, Section 2.2, with a review of the "population estimation using satellite imagery" application domain. The work described in this thesis is concerned with the application of data mining techniques, more specifically image mining techniques, for the purpose of population estimation using satellite image data. The main challenge in this context is how best to extract and represent image features so that mining algorithms can be applied. Image feature extraction and representation, and especially satellite image feature extraction and representation, is a central theme of this thesis. Image pre-processing is therefore discussed in Section 2.3 with a particular focus on: image enhancement, image segmentation and feature extraction. In addition, prior to the application of any data mining process, it is typically necessary, both from a computational efficiency and a computational effectiveness perspective, to reduce the dimensionality of the feature space by applying some form of feature selection. This is discussed further in Section 2.4. The chapter is then continued with Section 2.5 where some background concerning the domain of data mining is presented and more detail concerning the associated sub-domains of image mining and prediction. With respect to the latter both classification and regression are considered. A discussion concerning frequent subgraph mining is also presented because one of the representations presented uses the idea of extracting features from satellite images using graph/tree based techniques, which are then converted into a feature vector format using frequent subgraph mining. The mechanism adopted later in this thesis for the statistical comparison of different prediction models is presented in Section 2.6. Finally a chapter summary is presented in Section 2.7.

2.2 Population Estimation using Satellite Image

The collection of satellite imagery is founded on remote sensing technology, the measurement of the "energy" emanating from the earth's surface using sensors installed on an aircraft or spacecraft platform. The measured energy omitting from the earth's surface is used to form

an image. Examples of omitted energy sources include: (i) sunlight, (ii) wavelengths outside of the range of human vision, (iii) the "upwelling energy" which is radiated from the earth, and (iv) artificial sources [149]. The sensors, that may be mounted on aircraft or spacecraft, are classified into two categories: (i) passive and (ii) active. The distinction is that there is no energy source of radiation provided in the case of passive sensors such as cameras, multi-spectral scanners, thermal scanners and microwave radiometers. An example image obtained from a passive sensor is given in Figure 2.1(a). In contrast active sensors provided a built-in source of radiation, examples include Radio Detection and Ranging (RADAR) and Light Detection and Ranging (LIDAR) [91]. Figure 2.1(b) presents an example of a satellite image obtained using LIDAR technology. Satellite imagery provides an effective means of observing and quantifying the complexities of the surface of the earth.

(a) Example of an optical satellite image obtained using a passive sensor (camera) (image from www.map.google.com)



(b) Example of a satellite image obtained using an active sensor (LIDAR technology) (image from www.english-heritage.org.uk)

Figure 2.1: Example of satellite images from: (a) an active sensor and (b) a passive sensor.

Example applications where satellite images have been used include: (i) agriculture, (ii) forestry, (iii) land usage/land coverage studies, (iv) disaster management, (v) defence and security, (vi) natural resource management, (vii) climate monitoring and (viii) marine and coastal zone monitoring [6, 19, 82, 131, 162, 171, 184].

The satellite image application domain of interest with respect to this thesis is population size estimation. Knowledge of population size is an important requirement for the purposes of national planning, development and improvement processes. The principal means for a collection population size is through the mechanism of a census. Many countries across the world conduct a national population census once every ten years. In a number of countries such as Japan, Austria and Canada, a census has been taken more frequently, once every five years. The census collection activity consists of three stages: (i) planning and preparation,

(ii) census data collection and (iii) processing of the results. Traditionally there are two main approaches to census collection [134]:

1. Mailing of questionnaires to a list of household addresses, either by post or by electronic means, and asking individual householders to takes responsibility for completing and returning questionnaires.

2. Using ground staff to visit individual households in order to interview householders and obtain the desired census information.

As noted in Chapter 1, traditional population census collection (using either of the above) requires significant resource (money, staff and time) for both data collection and post-processing. In the context of the cost of census collection the UK Office for National Statistics (UKONS) reported that the UK 2011 census cost approximately £480 million; a figure that included both the cost of data collection and the cost of post processing. This is a considerable outgoing although UKONS notes that this cost "breaks down to less than £1 per person per year over the 12-year planning and operational cycle of the census" [48, 120, 133, 152]. The US 2010 census is reported to have cost $13 billion, approximately $42 per capita; by comparison, the 2010 census per-capita cost for China was about $1 and for India was $0.4 [125]. In December 2010 the US Government Accountability Office (GAO) noted that the cost of conducting US censuses had approximately doubled each decade since 1970 [164]. According to the Australian Bureau of Statistics the Australian 2011 census cost around AUD 440 million, about AUD 19 per person; whilst the 2006 census cost around AUD 300 million. Census collection is thus expensive and is becoming more so. Furthermore the resource required with respect to rural areas is typically greater to that required in urban areas because the communication and transport infrastructure in rural areas tends to be less well developed. Of course the cost has to be offset against the benefits that census data provides. UKONS argues that the cost "has to be set against its value in helping central and local government to allocate annually many billions of pounds of funding to communities" [58, 142].

Other than the cost associated with traditional census collection methods there are a number of additional disadvantages:

1. Where the intention is to collect census data using electronic means, for a variety of reasons, many people remain unconnected to the internet. In the context of the UK 2011 census UKONS reported that the most frequently cited reason for households not to have internet access was because of a "life style" decision not to. In less affluent parts of the world internet accessibility and usage is much lower (although arguably set to increase).

2. The use of questionnaires (online or otherwise) requires those completing the questionnaires to be literate, not necessarily always the case in many parts of the world.

3. Census collection is frequently viewed with suspicion.

A solution to the above disadvantages, and that advocated in this thesis, is to use satellite imagery for the purpose of population estimation. Population estimation has been a subject of researched amongst the Geographic Information Systems (GIS) and remote sensing communities for some time. From the literature we can broadly divide this research activity into two categories: (i) areal interpolation and (ii) statistical modelling [189]. In the areal interpolation category existing census information concerning some geographic area is used as an input to an interpolation algorithm to obtain a population estimation for a wider or alternative geographic area [99]. Statistical modelling in turn is concerned with the relationship between population size or density and data obtained from GIS and/or satellite imagery. The work presented in this thesis can be said to fall into the second category. The existing work on statistical modelling for the purpose of population estimation can be further categorised according to the nature of the data on which the population estimation generation is based, namely: (i) light intensity, (ii) land usage, (iii) dwelling unit count, (iv) image pixel characteristics and (v) physical or socio-economic characteristics.

The central idea on which the first category is based is that there is a functional relationship between population size and the amount of night time light emanating from an area. In [3], [26], [119] and [144] the relationship between population density and light frequency was analysed in order to convert light frequency into a population density metric using a luminous saturation measure obtained from the Defence Meteorological Satellite Program (DMSP) Operational Linesman System (OSL). In [144] the reported evaluation was directed at Japan and China, whereas in [26] and [119] it was directed at China only. In [3] the evaluation was directed at a population estimation of the Brazilian Amazon.

Work within the second category is directed at the correlation between population density and different types of land usage. The idea is to determine population densities according to land usage with respect to a set of one or more sample areas and apply this knowledge to additional areas. Land usage categories are typically identified from satellite image data. In [96] it is suggested that population densities for different types of land usage can be determined from sample surveys or census statistics. Four different types of land usage were extracted from four different cities in California, USA, and population densities computed. In [116], six types of land usage were identified in the context of Landsat TM satellite images centred on Atlanta, USA. A regression model was then applied to produce population densities for Atlanta.

The third category of approach to population size estimation using statistical modelling is to estimate the total "dwelling unit" count in a defined region and multiply this by an average number of people expected to live in a dwelling unit. There are various ways of obtaining an estimate of the dwelling unit count, but one suggested approach is to estimate this by analysing remote sensing images. In the past, when there was no effective ways of automatically identifying residential buildings within remote sensing imagery, the dwelling units were manually identified from aerial photographs (a laborious and time consuming process). With the advancement of technology and the availability of satellite imagery more advanced "fea-

ture extraction techniques" have been developed for this purpose [74]. In [2] a dwelling unit count based approach is presented using IKONOS satellite images of the Al Shaabia district in Khartoum, Sudan. The dwelling unit count approach has some similarity with respect to the work represented in this thesis.

In the fourth category, the relationship between image pixel characteristics and population densities are examined. The image pixel characteristics can be represented using a variety of mechanisms, but common examples include: mechanisms based on the spectral reflectance value of image pixels and image texture analysis mechanisms. Examples of using pixel characteristics for population estimation are presented in [87] and [103]. In [87] a system was presented whereby texture analysis was applied to Google Earth satellite images, using block sizes of 64x64 and 32x32 pixels, to estimate population densities with respect to cities in Pakistan. In [103] a variety of features were used, including: spectra signatures, principle components, vegetation indices, fraction images, texture and temperature. These features were extracted from Landsat ETM+ satellite images and used to measure population density in the city of Indianapolis, Indiana, USA.

The final category of population estimation is founded on the usage of various kinds of physical and socioeconomic information which is then interpolated to give population estimations. For example, information about demography, topography and transportation networks have all been used to estimate population size. In [114] a mechanism was presented for estimating population size by determining the correlation between the population in urban areas and the distance to the nearest Central Business Distract (CBD), distance to major roads, slope and the age of the community.

What all the above approaches to population estimation modelling have in common is that they are focussed on regions or areas rather than specific households as in the case of the work presented in this thesis. As far as the author is aware the approach presented in this thesis is entirely unique.

2.3 Image Processing

As noted above, before image mining of any form can be conducted the image set of interest must be pre-processed in an appropriate manner. In the context of the work presented in this thesis three image pre-processing operations are used: (i) image enhancement, (ii) image segmentation and (iii) feature extraction.

Image enhancement is the process of improving the quality of an image. A variety of techniques exist, the choice of technique is application dependent, a technique used for enhancing (say) brain scan images may not be appropriate for enhancing satellite images [56, 107]. With respect to the work presented in this thesis image enhancement was applied so as to improve the effectiveness of the image segmentation applied to identify households. Further detail concerning the adopted image enhancement method is given in Sub-section 2.3.1 below.

Image segmentation is the process of grouping image pixels that share some form of commonality. The groupings are usually referred to as *regions* or *objects*. The typical goal is to facilitate image analysis. Image segmentation is usually achieved by identifying objects and boundaries. In the case of the work presented in this thesis image segmentation was applied in order to isolate individual households, an overview of the adopted image segmentation process is given in Sub-section 2.3.2.

Feature extraction is the process of representing images, or image segments, according to image content. There are various content features that may be used for this purpose such as colour, texture and spatial layout [28]. Colour features are the simplest to obtained as they can be extracted directly in terms of pixel intensities. Texture is normally considered in terms of some form of prescribed spatial pattern template describing the relative position of colour, texture or shape information [179]. A review of existing work directed at feature extraction (and representation), in the context of the work presented in this thesis, is thus given in Sub-section 2.3.3.

2.3.1 Image Enhancement

A review of image enhancement processes, in the context of this thesis, is presented in this sub-section. As noted above, the objective of image enhancement is to improve the quality of an image so as to improve the effectiveness of any further processing to be applied. Typically image enhancement includes: (i) "sharpening" (de-blurring), (ii) contrast improvement, (iii) brightens adjustment and (iv) noise removal. Image enhancement has been shown to be advantageous with respect to many image analysis applications such as: medical image analysis [161], biometric authentication (such as face detection and fingerprint detection) [90, 153] and satellite image analysis [84]. The latter is of particular relevance with respect to the work presented in this thesis.

Image enhancement can be generally described as the process of transforming an input image comprised of an ordered collection of pixels R, using a function t, into an output image comprised of an ordered collection of pixels S (Equation 2.1).

$$S = t(R) \tag{2.1}$$

The challenge of image enhancement is quantifying the criterion for the enhancement; a large number of image enhancement techniques are empirical and require interactive procedures to obtain satisfactory results.

A large number of image enhancement methods have been developed which can be broadly divided into two categories: (i) spatial domain methods and (ii) frequency domain methods [56]. Spatial domain techniques are applied directly to an image, while frequency domain methods are applied to an intermediate representation (typically generated using some kind of Fourier transform) [122]. In the context of this thesis Spatial domain methods have been used. More specifically three image enhancement methods were utilised:

1. **Thresholding**, which was applied to each layer of the Hue-Saturation-Value (HSV) satellite image colour space so that the households of interest were more clearly defined and so that features such as rivers and roads were eliminated from images.

2. **Arithmetic and logic operations**, which were used to combine an input image with an another image (possibly a mask) so as to produce a single enhanced image.

3. **Histogram Equalisation**, which was used adjust the image contrast so that it was consistent across a given collection of satellite images.

Each of the above is described in further detail in the remainder of this sub-section.

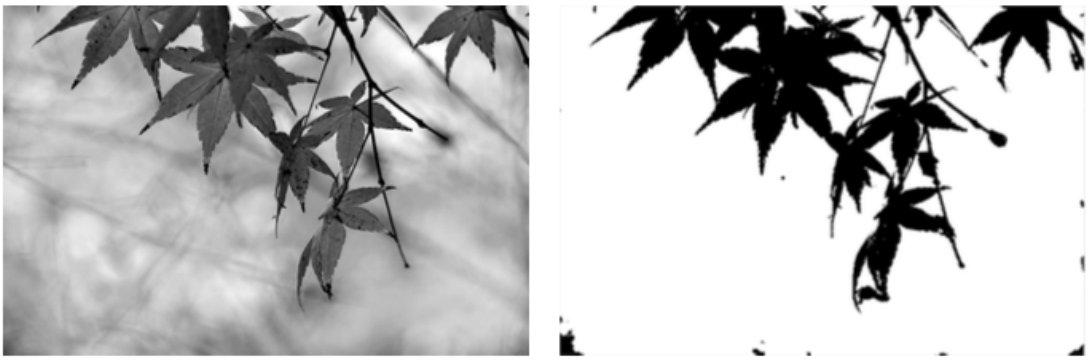
Thresholding is an image enhancement process used to separate image objects of interest (the "foreground") from the rest of an image (the "background") by converting a given image into a binary, black and white, image where white regions represent the foreground and black regions the background. The method assumes that the colour features associated with the foreground is somehow distinct from the colour features associated with the background. More formally we can define thresholding using the following:

$$s = \begin{cases} 1 & r > threshold \\ 0 & r \leq threshold \end{cases} \tag{2.2}$$

where $s \in S$ and $r \in R$.

Thresholding methods can be divided into two main categories: (i) global and (ii) local [146]. Global thresholding methods use a threshold value applicable to an entire image; the threshold value is often based on an estimation of the boundary between foreground and background as displayed in an intensity histogram (hence it is often referred to as a "point processing" operation). Local methods use an adaptive threshold whereby different threshold values are derived from local area information [72]. Figure 2.2 gives an example of thresholding. Figure 2.2(a) gives the original image while Figure 2.2(b) gives the enhanced image after thresholding has been applied.

Histogram equalisation is a widely used image enhancement technique for adjusting image contrast so that the full colour/intensity range is used. It is also used with respect to collections of images so that they conform to a shared colour/intensity range. The reasoning is as follows. Given an image where the background and foreground are both light, or both dark, the associated intensity histogram would be skewed towards one end of the greyscale, consequently all the image detail will be compressed into a small part of the histogram making it difficult to distinguish the foreground from the background. Histogram equalisation in this case will cause the compressed area to be "stretched out" to produce a more uniformly distributed histogram that allows for more effective further processing [122, 151]. Figure 2.3 presents an example image enhancement using histogram equalisation. Figure 2.3(a) shows the original image and its associated greyscale histogram in Figure 2.3(b), while Figure 2.3(c) is the enhanced image



(a) Original image (b) Enhanced image using thresholding

Figure 2.2: Example of thresholding for image enhancement

after histogram equalisation has been applied and its associated greyscale histogram in Figure 2.3(d). From the figure it can be observed that the histogram is more "dispersed" after histogram equalisation.

Arithmetic and logic operations are often used for the purpose of image enhancement by combining an input image with one or more other images so as to produce an enhanced image. Such operations are performed on a pixel-by-pixel basics, as in the case of thresholding. Typical arithmetic operations that may be applied are addition and subtraction; typical logic operations are *AND*, *OR* and *NOT* [55].

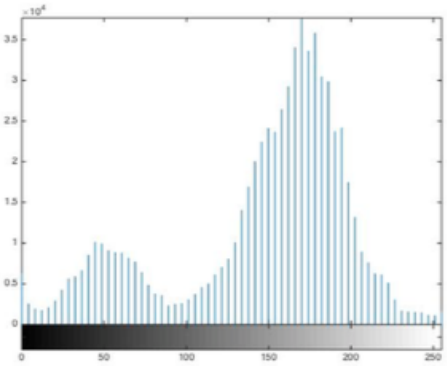
As the name suggests, using the addition operator the corresponding pixel values for two equal sized input images are each summed to produce a new image of the same size as the first two. A common variant of the addition operator is to simply add a constant value to each pixel in a single input image [124]. In the case of the subtraction operator the pixel values are subtracted from one another; alternatively a constant may be subtracted with respect to the pixels in a single input image [85].

The logical operators are used, in a similar manner to the arithmetic operations, on a pixel-by-pixel basics; they are typically used for the operation to binary valued or greyscale images. The operation of the *AND*, *OR* and *NOT* operators is as standard. To give one example the logical NOT operator is used to invert pixel values, so that dark areas in the input image become light areas in the output image and vice versa [33, 56].

Figure 2.4 gives an example of arithmetic/logic enhancement. Figure 2.4 (a) gives the original image and Figure 2.4(b) an enhanced image obtained by adding the constant value of 100 to the original image.

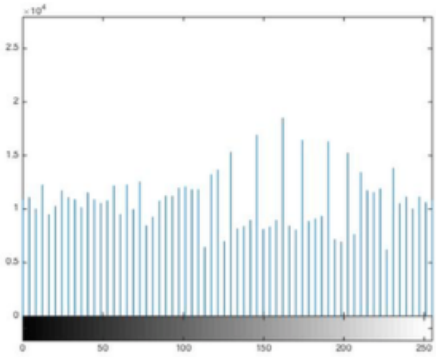
In the context of satellite image applications most of the above image enhancement techniques have been used. To give one example, in [8] histogram equalisation was applied to enhance LANDSAT satellite images.

(a) Original image



(b) Greyscale histogram of original image

(c) Enhanced image using histogram equlisation



(d) Greyscale histogram of enhanced image using histogram equlisation

Figure 2.3: Example of histogram equalisation for image enhancement

(a) Original image

(b) Enhanced image by applying an arithmetic operator

Figure 2.4: Example of the use of an arithmetic operator for image enhancement

2.3.2 Image Segmentation

An overview of image segmentation is presented in this sub-section. In the context of the work presented in this thesis segmentation was used to isolate individual households. More generally image segmentation is used in situations where some part of an image, or some region within an object, is required in the context of further image analysis [137]. In [55] image segmentation is defined as the process of partitioning a digital image into semantically interpretable regions that are more meaningful and easier to analyse with respect to a particular application then if the entire image was taken into consideration. Segmentation has been usefully employed with respect to a variety of applications [32] such as: (i) medical applications (isolating tumours and other pathologies, measuring tissue volumes, computer guided surgery, diagnosis, treatment planning and the study of anatomical structure) [46, 112, 127], (ii) geoscience (location of objects such as roads, forests and crops in satellite images) [11, 54], (iii) face recognition [13] and (iv) finger print recognition [181].

Segmentation is typically conducted according to some characteristic image feature such as colour, line, intensity or texture. There are various image segmentation methods that have been proposed. A well documented categorisation of image segmentation methods is: (i) threshold based segmentation, (ii) edge based segmentation and (iii) region based segmentation. Each is discussed in further detail below.

Threshold based segmentation is the simplest segmentation method. Note that this technique may be applied for both enhancement (see Sub-section 2.3.1) and segmentation. With respect to segmentation, thresholding is used to transform a greyscale image into a binary image. The idea of threshold based segmentation is to replace each pixel in an image with a black pixel if the image intensity of the pixel $I_{i,j}$ is less than the threshold value T (that is, $I_{i,j} < T$), or a white pixel if the image intensity is greater than that constant [192, 193]. Figure 2.5 presents an example of threshold based segmentation; Figure 2.5(a) is the original image while Figure 2.5(b) is the processed image.

In edge based segmentation the edges in an image are identified. Ideally the detected edges from the given image represent object boundaries and can thus be used to define these objects. Image segmentation using edges is typically a three step processes: (i) compute an edge image containing all edges of an original image, (ii) process the edge images so that only closed object boundaries remain, and (iii) transform the result to an ordinary segmented image by filling in the object boundaries [17, 81]. An example of edge based segmentation is shown in Figure 2.6; Figure 2.6(a) is the original image, while Figure 2.6(b) is the processed image using edge based segmentation.

Region based segmentation (see for example [4, 145, 160]) is concerned with identifying object regions in an image that share some common pixel feature (such as colour) that are in some sense homogenous. Region based segmentation methods have two basic forms of operation: (i) merging and (ii) splitting. The basic approach to image segmentation using merging (sometimes also referred to as region growing) is as follows: (i) obtain an initial segmentation

(a) Original image

(b) Segmented image using threshold based segmentation

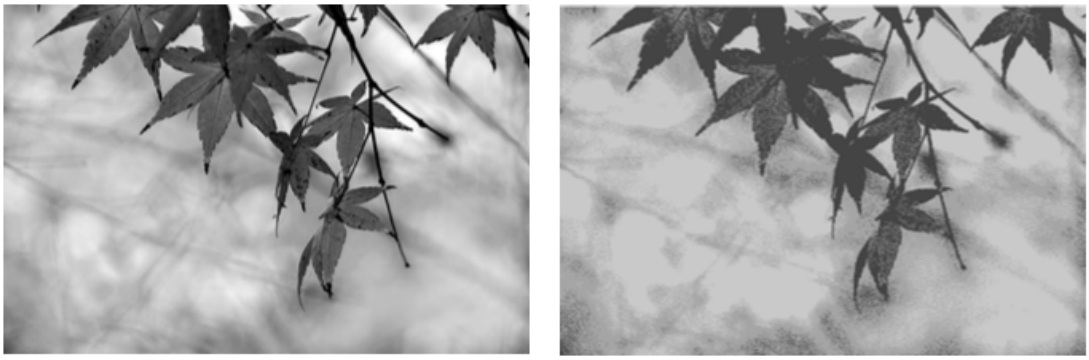
Figure 2.5: Example of threshold based image segmentation

(a) Original image

(b) Segmented image using edge based

Figure 2.6: Example of the line based image segmentation

of the image where each pixel is in its own segment, (ii) merge those adjacent segment pairs that are most similar, and (iii) repeat step (ii) until no more segments can be merged remain. The central element of the merging approach is the similarity criterion used to decide whether two segments should be merged or not. This criterion may be based on grey value similarity (such as the difference in average grey value, or the maximum or minimum grey value difference between segments), the edge strength of the boundary between the segments, the texture of the segments, or one of many other possibilities. The basic approach to image segmentation using splitting is as follows: (i) obtain an initial segmentation of the image where the entire image is in a single segment, (ii) where possible split each segment into two "homogeneous" sub-segments, and (iii) repeat step (ii) until no more splitting can take place. The criterion for the homogeneity of a segment may be the variance of its grey values, the variance of its texture, the occurrence of strong internal edges, or various other criteria. Figure 2.7 presents an example of region based image segmentation; Figure 2.7(a) is the original image, and Figure 2.7(b) is the processed image using region based segmentation.



(a) Original image

(b) Segmented image using region based segmentation

Figure 2.7: Example of the region based segmentation

The merging and the splitting approaches may be combined: the basic splitting approach is often enhanced by combining it with a merging approach, where inhomogeneous segments are split into simple geometric forms recursively. This of course creates arbitrary segment boundaries, and merge steps are included into the process to remove incorrect boundaries.

In the context of the work presented in this thesis the objective of the segmentation was to isolating individual households. With respect to the work presented in Chapters 3 to 7, this was achieved using edge based segmentation (an alternative is presented in Chapter 8, the reason for this will became clear later in this thesis). Further detail concerning the segmentation of satellite images, to identify households, with respect to the data used in this thesis for evaluation purposes, is presented in Section 3.4 and in the case of the large scale population estimation process in Chapter 8. It has already been noted that the households of interest are typically

defined by some kind of boundary, therefore line or edge segmentation is appropriated.

2.3.3 Feature Extraction

An overview of image feature extraction mechanisms is presented in this sub-section. A necessary precursor for the application of image mining is that the key properties or characteristics of the image set to be mined need to be extracted and represented so as to facilitate the desired image mining [176]. It is not possible to mine images directly because of the prohibitive amount of pixel data that would have to be considered. The most commonly used representation used with respect to prediction (classification and regression) is the feature vector representation. Feature extraction and representation is central to the research presented in this thesis, three alternatives are considered: (i) graph-based, (ii) colour histogram based and (iii) texture based, all three result in a feature vector representation. Given the significance of feature extraction with respect to this thesis a review of previous work in this area is therefore presented in this section.

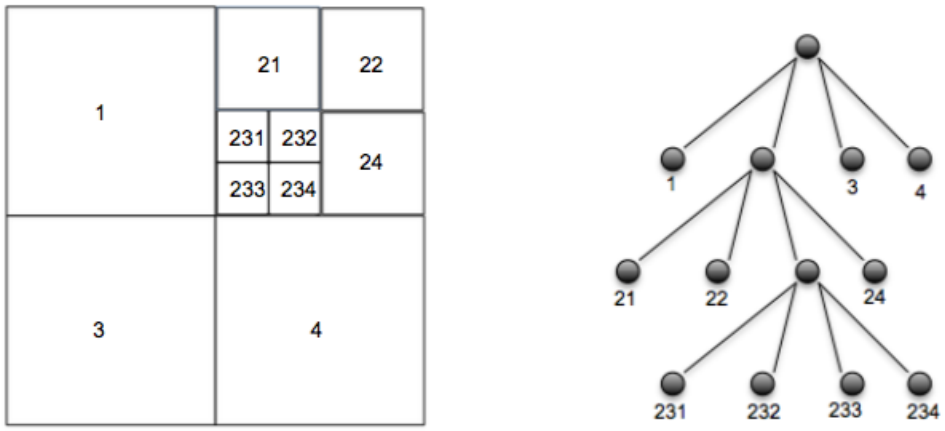
In the early work on feature extraction from images the process was not based on image content but on the textual annotation of images, see for example [21, 173]. However, the automatic generation of textual annotations for images is not a realistic one, the text-based approaches require manual annotation which is both resource intensive and challenging [117]. More recently techniques based on visual information extraction have been developed, thus content-based feature extraction instead of text based feature extraction; see for example [24, 93, 143]. In content based feature extraction the features considered are quantifiable properties of an image. These properties/features can be divided into two categories: (i) general features and (ii) domain-specific features. General features are application independent and include features such as colour, texture and spatial layout. The nature of such general features can be further divided into: (i) pixel-level features such as colour and pixel location (the colour histogram based technique for population mining from satellite images presented later in this thesis falls into this category), (ii) local features calculated over a sub-area or region of an image, and (iii) global features calculated over an entire image such as texture based feature extraction techniques (the graph and LBP based approach presented later in this thesis fall into this category). Domain-specific features are application dependent features, for example elements of the human face as used in face recognition [28]. With respect to the work presented in this thesis the general feature extraction methods used are focussed on three basic types of features: (i) spatial information, (ii) colour and (iii) texture. Each is thus briefly considered in some further detail below.

Spatial information

The first type of feature considered in the work presented in this thesis is spatial information. Given a set of identified regions and/or objects, spatial information can be used to distinguish between them. For example a blue sky region and a blue sea region may have similar colour

histograms, however the spatial locations within the image will be different. Spatial information is often represented using a graph representation [51, 157]. Examples of graph-based structural image feature representations include: Attributed Graphs (AGs), Function Describe Graphs (FDGs) and Quadtrees. The advantages of graph-based representations are their general applicability [29] and their invariance to rotation and translation [92].

One of the most common methods for spatial information feature extraction is quadtree decomposition. In a quadtree every node in the tree, apart from the leaf nodes, has four "children" labelled North-West (NW), North-East (NE), South-West (SW), and South-East (SE). The image is partitioned into four quadrants at each hierarchical level, however it is usually unnecessary to decompose all branches down to the same level. If a parent node has four children of the same value (node label) the decomposition can be stopped at the parent node. Figure 2.8 gives an example of a quadtree decomposition, Figure 2.8(a) illustrates the decomposition while Figure 2.8(b) presents the resulting tree [170].



(a) Quadtree decomposition (b) Quadtree

Figure 2.8: Example of quadtree decomposition

Note that in Figure 2.8, the coding reflects the decomposition (1 = NW, 2 = NE, 3 = SW and 4 = SE). Quadtrees have been applied widely with respect to image analysis including satellite image analysis; examples where quadtree decomposition has been used in the context of satellite imagery can be found in [195] and [128]. Quadtree decomposition was also used in the context of the work presented in this thesis; this will be discussed further in Chapter 4.

Colour

Colour features are the most frequently used feature type and thus arguably the most significant [79, 172]. Colour feature extraction offers a number of advantages: (i) simplicity of implemen-

tation, (ii) effectiveness with respect to many applications, (iii) invariance to image rotation and (iv) low storage requirement [28]. In general, colour is represented using the concept of a *colour space* (also known as colour mode, model or system). There are various existing colour spaces: (i) RGB (Red, Green, and Blue), (ii) CMYK (Cyan, Magenta, Yellow, and Key (black)), (iii) HSV (Hue, Saturation, and Value) and (iv) greyscale.

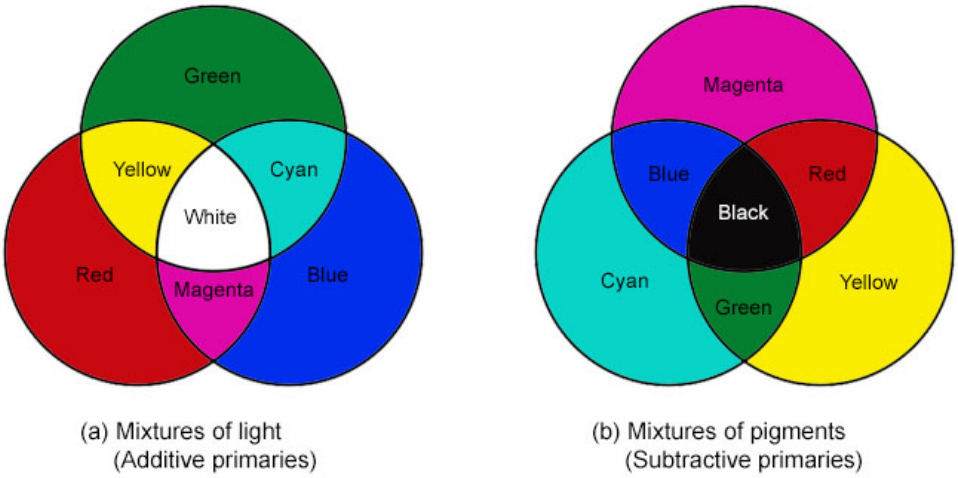


Figure 2.9: The primary colours and secondary colours for the RGB and CMY Colour Spaces.

The RGB colour space is most commonly used for images, particularly for the representation and display of images in electronic systems, such as cameras and computers. RGB is what is termed an "additive" colour space, which means that individual pixel colours are created by combing a number of light channels (see Figure 2.9(a)). In the case of the RGB colour space the channels are: Red (R), Green (G) and Blue (B); hence RGB. The wavelengths of the red, green and blue channels were adopted solely for the purpose of standardisation and not to achieve equivalence with visible colours [80]. The RGB colour space can be represented in terms of a 3D coordination system where the origin $\langle 0,0,0 \rangle$ is black and $\langle 1,1,1 \rangle$ is white. All values for red, green and blue are assumed to be in the range of 256 intensity values. Figure 2.10(a) shows the RGB colour space with the primary colours red, green and blue and the secondary colours yellows, cyan and magenta. The secondary colours are produced from the combination of two primary colours. Greyscale comprises the values along the dashed line connecting black to white.

The CMY colour space is a "subtractive" colour space, the term subtractive as used here refers to the mixing of three primary colour of pigments: Cyan (C), Magenta (M), and Yellow (Y) to give different colours as illustrated in Figure 2.9(b). In general, the CMY colour space is used in the printing processes. The devices that use coloured pigments (such as printers and copiers) require CMY data as input.

Figure 2.9(b) inducates that by combining the three primary colour pigments we get black,

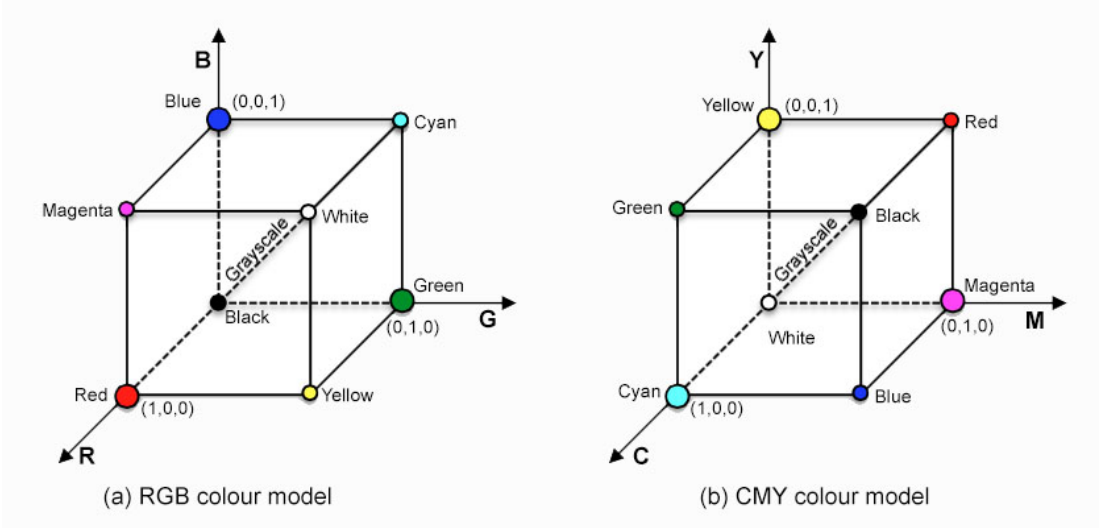


Figure 2.10: The schematic of RGB and CMY "colour cube"

but in practice combining these three colours for printing purposes does not produce pure black. To generate pure black (the predominant colour in printing), black (K) is given as a fourth colour stream leading to the CMYK colour space.

The CMY colour space can be represented by a 3D coordinate system in the same way as the RGB colour space. Again $\langle 0,0,0 \rangle$ is black and $\langle 1,1,1 \rangle$ is white. Figure 2.10(b) shows the CMY colour space with the primary colours cyan, magenta and yellow and the secondary colours red, green and blue. As in the case of the RGB colour space the secondary colours are produced from the combination of two primary colours. As before greyscale comprises the values along the dashed line connecting black to white.

The streams in the HSV colour space are the components Hue (H), Saturation(S) and Value (V) (or Hue, Saturation and Brightness in which case we refer to the colour space as the HSB colour space). HSV (HSB) is what a cylindrical coordinate representation of the RGB colour space would look like. The relationship between the RGB colour space and HSV is presented in Figure 2.11. From the figure it can be seen how the six primary and secondary colours in the RGB colour space (Figure 2.11(a)) are mapped into the hue plane. The hue of a colour refers to its dominant wavelength and can be expressed as an angle around a colour hexagon as shown in Figure 2.11(b). The saturation of a colour describes the purity of the colour measured as the distance from the V axis as presented in Figure 2.11(b); a pure red is fully saturated with a saturation level of 1, whereas tints of reds have saturations of less than 1. The value (or brightness) of a colour in the HSV system describes the "darkness" of a colour. Figure 2.11(b) shows that this value is measured along the V axis where $V = 0$ (at the end of the end of the cone) is totally black, while the $V = 1$ (at the other end of the cone at the centre of the full colour hexagon) is white [55].

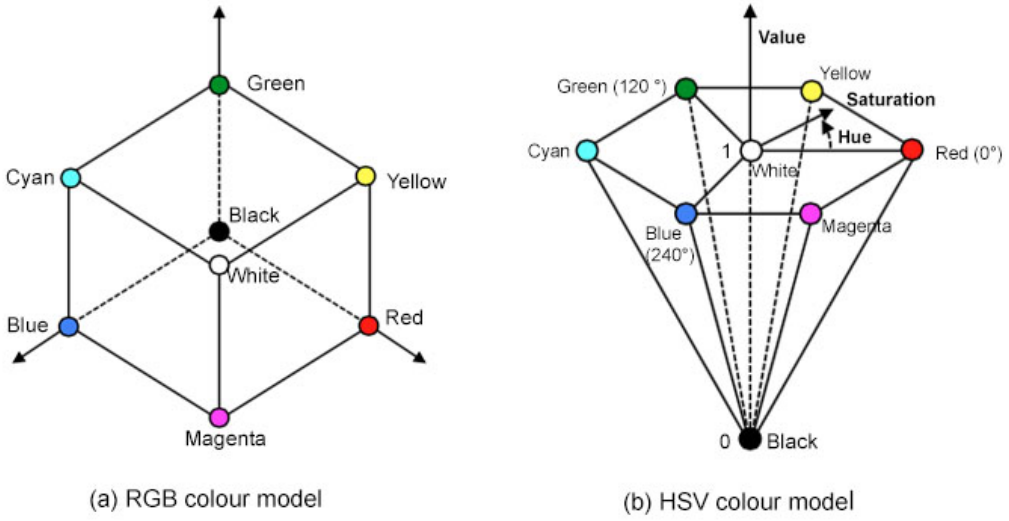


Figure 2.11: The relationship between HSV colour space and RGB colour space

Greyscale colour refers to a range of shades of grey from black to white. The intermediate shades of grey are represented by equal values for the three primary or pigments colours when using the RGB or CMY colour spaces. Because only 8 bits are required for greyscale (as opposed to 24 for RGB or CMY) the colour space is sometimes called 8-bit greyscale [158, 196].

Three of the above colour spaces (RGB, HSV and greyscale) were used with respect to pre-processing of the raw satellite image data considered in this thesis.

Texture

Feature extraction in terms of texture refers to processes whereby individual pixels are defined in terms of their neighbours. Image texture has been widely used with respect to various applications, for example content-based image retrieval, image classification, pattern recognition and computer vision [27, 52, 88]; image classification is of course of particular interest with respect to the work presented in this thesis. In more detail texture is described in terms of texture primitives or texture elements (texels), based in turn on "tone and structure". Tone is derived from pixel intensity values, while structure from the spatial relationship between pixels [170].

There are a number of techniques that have been proposed to extract texture features, these can be categorised as follows: (i) spatial texture feature extraction methods and (ii) spectral texture feature extraction methods. In the first method texture features are extracted by computing pixel statistics or by finding a local pixel structure. Spatial texture feature extraction offers the following advantages: (i) the representation is immediately meaningful, (ii) it is easy

to understand and (iii) it can be extracted from any shape without losing information. The associated disadvantage is that it is sensitive to noise and distortions. The second method, the spectral texture extraction method, transforms a given image into a frequency domain and then calculates features from the transformed image. The advantages of this second method are that it is robust and simple to compute; the disadvantages are that it has no semantic meaning and that a square image region with sufficient size is required [176].

A commonly used structure used for texture feature extraction, and used with respect to the work presented in this thesis, is the *co-occurrence matrix*. A co-occurrence matrix $C(i, j)$, holds the number of co-occurrences of individual pixel values i and j at a given distance d. The distance d is defined in terms of polar coordinates (d, θ), where d is the discrete length and θ is the orientation. Typically θ takes the values of 0°, 45°, 90°, 135°, 180°, 225°, 270°, and 315°. A number of attributes can be extracted from the co-occurrence matrix: (i) Energy, (ii) Contrast, (iii) Correlation, (iv) Homogeneity and (v) Entropy.

A commonly used spectral texture feature extraction methods, and that used with respect to the work presented in this thesis, is the Local Binary Pattern (LBP) as first proposed in [135]. In the LBP method each pixel is defined using the relative greyscale of its neighbourhood pixels [106]. Texture features tend to be robust with respect to image rotation, illumination change and occlusion [169].

Both the greyscale co-occurrence method and LBPs are used in this thesis to extract image features from satellite data featuring households. The detail of the proposed household image representation using texture analysis mechanism is presented in Chapter 6.

2.4 Feature Selection

As a result of feature extraction (as described above) a great many features are typically identified. In the context of prediction some of these features may not be relevant. The idea behind feature selection (also known as variable or attribute selection) is to prune an identified set of features, according to some criterion, so as to reduce the overall number of features to be considered and improve the effectiveness of the prediction. More specifically feature selection offers advantages with respect to: (i) prediction performance, (ii) a cost effectiveness of prediction (it typically reduces the processing time) and (iii) user understandably of prediction results [59]. An overview of feature selection is thus presented in this section.

Feature selection is essentially the process of removing redundant and/or irrelevant attributes from a given representation (such as a feature vector representation) prior to the commencement of data mining. The process is sometimes incorporated into a larger process called data cleaning [69]. Figure 2.12 presents a schematic of the feature selection process. From the figure it can be seen that feature selection is an iterative process comprised of four steps: (i) subset generation, (ii) subset evaluation, (iii) stopping criterion and (iv) result validation [31].

With reference to Figure 2.12, during subset generation a candidate subset of features is generated. This candidate subset is then evaluated and compared with the previous best subset

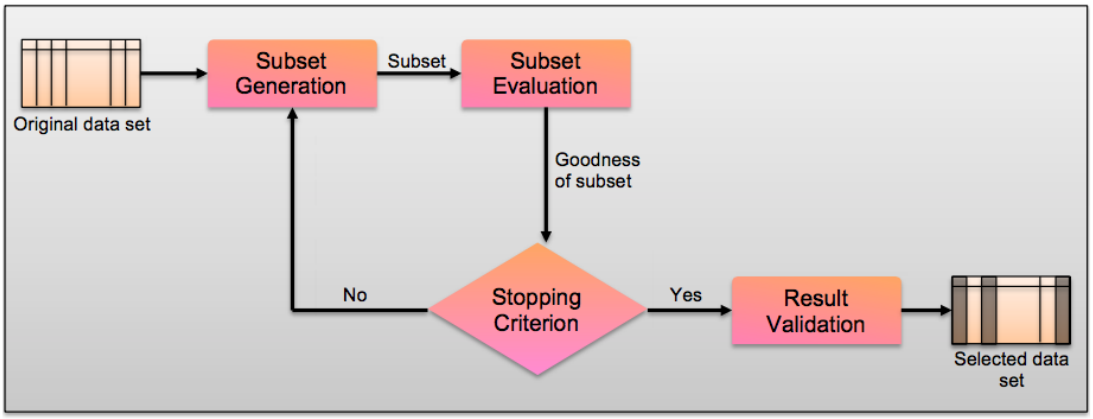


Figure 2.12: Schematic illustrating the feature selection process inspired from [31]

(if any) according to some evaluation criterion. The stopping criterion determines when the feature selection process should stop. Examples of stopping criteria include: (i) the potential set of candidate feature sets has been exhausted, (ii) arrival at a minimum number of features, (iii) arrival at a maximum number of iterations and (iv) identification of a sufficiently good subset. The final step is result validation where prior knowledge about the data and prediction outcome is used to measure the validity of a proposed subset of features. For example in [110, 111] the error rates of the predictor were compared when using the full set of features and the selected subset of features. Most feature selection mechanisms are designed to operate with classification algorithms, however there are some feature selection mechanisms directed at regression such as Correlation-based Feature Selection (CFS), an algorithm for identifying and selecting a subset of features which is highly correlated with the predictor variable [64].

Feature selection mechanism can be broadly divided into two categories: (i) ranking based and (ii) selection based [104]. In the ranking based approach individual feature are ordered according to their relevance or importance with respect to a given problem and the top k selected; whereas in the selection based approach subsets of feature are considered. The later category can be further divided into three subcategories: (i) filters, (ii) wrappers and (iii) hybrid. Using filters the general characteristics of the data are used to evaluate and select feature without reference to any particular learning algorithm. Using wrappers a proposed feature subset is evaluated, using (say) accuracy estimation, with respect to a particular learning algorithm. The hybrid is then a combination of the two [65, 111].

With reference to the work presented in this thesis, ranking based feature selection was used. Frequently reference measures used for feature ranking include: Information Gain, the Chi-Squared metric and Gain Ratio [49, 66, 187, 194]. These three measures were used with respect to the work presented in Chapters 4 to 6 which concentrate on classification, in the case of regression analysis (Chapter 7) CFS was used.

2.5 Data mining and Image Mining

The work described in this thesis is primary concerned with data mining, more specifically image mining. Thus the basic concepts of the domain of data mining and more specifically the sub-domains of image mining and prediction analysis are considered in this section. With respect to prediction analysis both classification and regression are considered. The remainder of this section provides a review of data mining in general and image mining in particular. This is followed by two sub-sections. The first is directed at prediction techniques and the second at Frequent Subgraph Mining (FSG). The significance of the latter is that FSG is the foundation for the first of the population estimation mining using satellite imagery techniques considered in this thesis, namely the Graph-Based Approach presented in Chapter 4

Data mining is a technology concerned with the extraction (mining) of interesting, but hidden, information from data. Many people are familiar with the phrase "Knowledge Discovery in Data" or KDD often used as synonym for data mining, although technically KDD describes a meta-process of which data mining is a part. Data mining can be broadly classified into two categories: predictive data mining concerned with the prediction of behaviour based on historic data and descriptive data mining concerned with the discovery of patterns in existing data that may be used to guide future decisions [178].

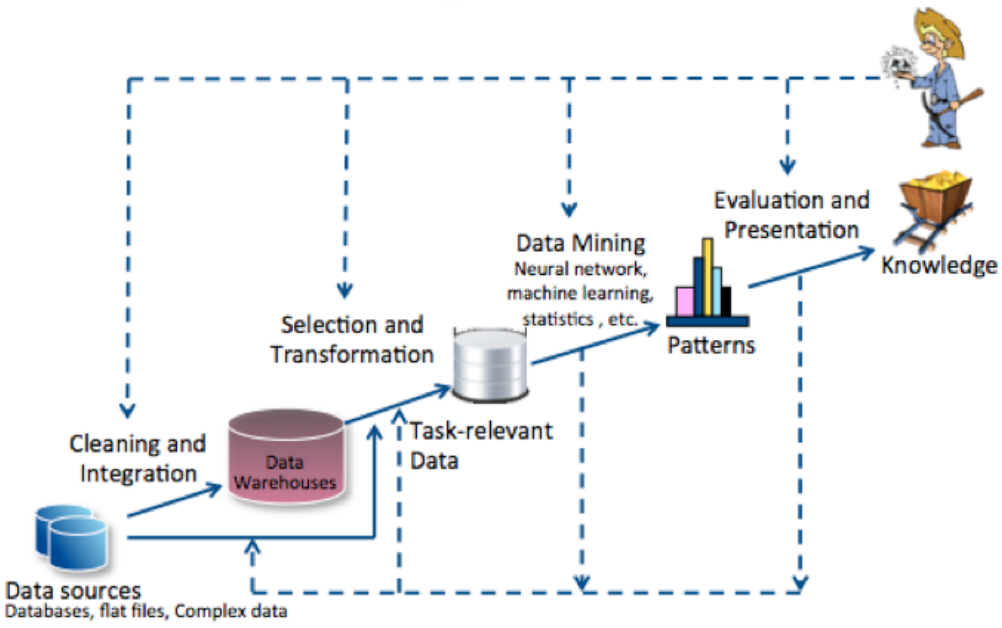


Figure 2.13: Schematic illustrating KDD meta-process inspired from [69]

A schematic of the KDD meta-process is given in Figure 2.13. From the figure it can be seen that KDD consists of four processes: (i) data cleaning, (ii) data selection (iii) data mining and (iv) evaluation (usage). In the first process the data to be mined is collected together

(possibly from many sources) and "cleaned" so as to remove noise and irrelevant data. In the context of image mining this is where image pre-processing of the form described above (see Section 2.3), is conducted. If the data comes from multiple sources an integration sub-process may also be applied (as indicated in the figure). Data selection is concerned with identifying the particular data items required (usually only a subset of the collected data is needed) and transforming or consolidating this subset into an appropriate format. This is where feature selection, as described above, will take place (Section 2.4). The next process is the actual data mining, the most significant process within the KDD meta-process where the "information discovery" takes place. In the fourth and last process the extracted information is evaluated according to some relevance criteria; this may include visualisation and/or translation into some other knowledge format so that the extracted (mined) knowledge can be presented in a more "user friendly" form to end users [69].

Data mining can be performed on any kind of data repository. Traditionally data mining has been applied to relational databases where the data is stored in a tabular format that has a two dimensional structure comprised of rows and columns [166]. The term data warehouse is sometimes used (as in the case of Figure 2.13), a subject-oriented database integrated from multiple sources in a given time period for data mining purposes. More recently data mining has been applied to a greater variety of data formats such as: spatial data, time-series data, free text, multimedia data, the World Wide Web, and video and image data [69, 71]. With respect to this thesis the last, image data, is of particular interest. Image mining is thus a form of data mining best described as the process of discovering interesting, but hidden, information within image data [163].

The nature of the data mining that may to be utilised is dependent on end user requirements. Common examples found in the literature include: (i) association rule mining, (ii) clustering and (iii) prediction analysis (typically classification and regression). The first is concerned with the discovery of what are called association rules, rules that describe relationships between attributes in the data (if x occurs y is also likely to occur). Clustering (a form of *unsupervised learning*) is the process of grouping data into "clusters" according to some notion of similarity. Prediction analysis (a form of *supervised learning*) is directed at generating a model from a data set that can then be used to predict the nature of previously unseen data [109]. The work described in this thesis is directed at the latter.

Image prediction algorithms typically operate using a pre-labeled image set, referred to as the *training set*, to generate a prediction model which can later be applied to predict labels to be associated with previously unseen data (images). Image mining has been successfully applied in many different application domains; two common application domains for image mining are healthcare and geotechnology [163]. The later is of particular relevance with respect to this thesis. The geotechnology application domain includes data mining applied to GIS, GPS and remote sensing data in the context of agriculture, forestry, environmental science and geoscience studies [30, 154].

2.5.1 Predictive Analysis

Predictive analysis is concerned with the extraction of embedded knowledge from data for the purpose of using this knowledge for prediction purposes. The main idea behind prediction analysis is to capture the correlation between what are known as "explanatory variables" and "prediction variables" from historical data [71]. A schematic of the predictive analysis process is given in Figure 2.14. From the figure it can be observed that the analysis process consists of two phases: (i) training, where a prediction model is constructed using a training set; and (ii) prediction, where the model is used to predict the nature of unseen data.

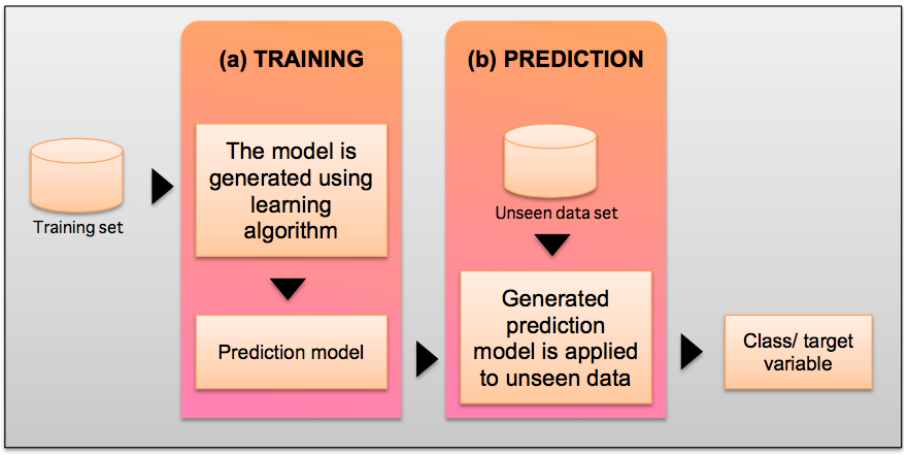


Figure 2.14: Schematic illustrating the generic predictive analysis process

From the literature two broad predictive analysis techniques can be identified: (i) classification and (ii) regression. The fundamental idea of classification is to build a model, known as a "classifier" which can then be used to label previously unseen records. The labels are referred to as "classes" and are drawn from a predefined set of classes C. The individual classes represent discrete or categorical values; the size of the set C is also typically quite small, for many applications $|C| = 2$ (binary classification). In the case of regression the predicted values are continuous numeric values. Regression is a statistical technique whereby a numerically valued function is generated [68]. Both techniques comprise a training/learning element and an application element. During the training/learning element the desired model is constructed from pre-labelled training data, hence classification and regression referred to as supervised learning techniques. Classification and regression are considered further below. Note that an important precursor for the application of a prediction model generation algorithm is feature selection, as discussed in Section 2.4 above, this has the effect of reducing runtimes and increasing accuracies.

Classification

Classification is concerned with the generation and application of a prediction model, known as a classifier, for the purpose of predicting the class labels to be associated with new records. As noted above the technique comprises two elements: (i) training and (ii) application.

Training is concerned with the construction of the desired model (classifier). This is achieved using a learning algorithm of some form, which is applied to a training set to construct the classifier. The training set comprises a set of pre-labelled records typically represented using a n-dimensional feature vector representation which in turn comprises a set of attribute values and a class label $\{a_1, a_2, \ldots a_{n-1}, c_n\}$ where a_i is an attribute value and c_n is a class label such that $c_n \in C$. The second element is the classifier application step in which a generated classifier is applied to previously unseen data so as to attach class label to each record in the new data [69].

In the context of image classification a schematic illustrating the image classification process is presented in Figure 2.15. Note that the schematic corresponds with the generic predictive analysis schematic given in Figure 2.14. The top half of the figure describes the "training" element while the lower half the "classification" element. With respect to the work presented in this thesis eight classifier generation algorithms were considered: (i) Decision Tree generators (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network.

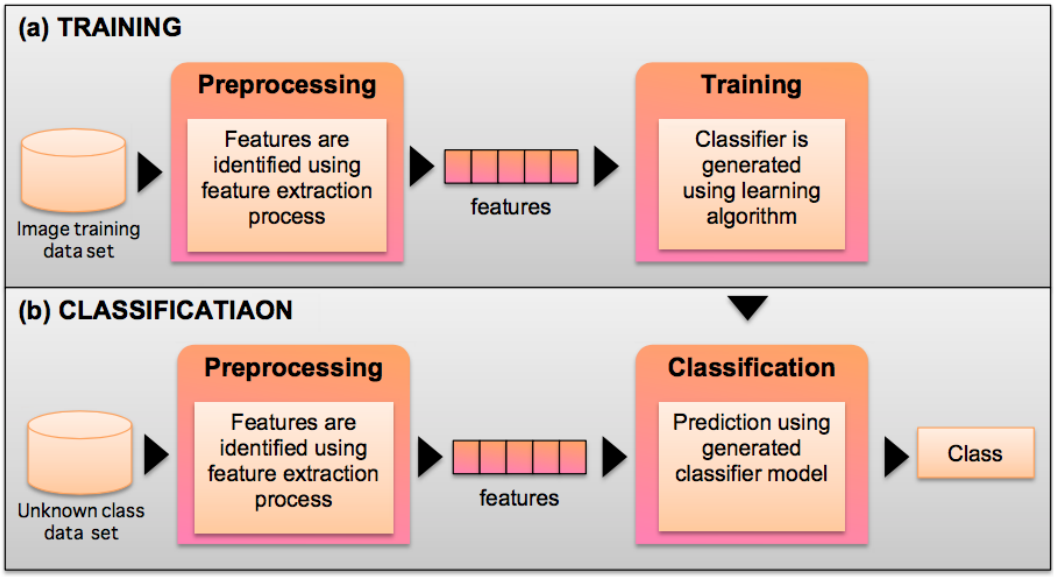


Figure 2.15: Schematic illustrating the image classification process

Regression

As noted above, classification techniques are typically used with respect to applications where the prediction is founded on the usage of discrete class labels. Again as noted above, regression is a statistical technique applicable where the prediction is concerned with real values. The idea is to mathematically model the relationship between a response variable, and one or more predictor variables, using a continuously-valued function. As in the case of classification, the model is construct using training data.

There are many existing techniques for performing regression analysis, frequently referenced techniques include simple linear regression and multiple linear regression [71, 129]. Each will be discussed in further detail below (both were used with respect to the work presented later in this thesis, the detail of regression analysis is presented in Chapter 7. In the discussion below the set of predictor variables (also known as the independent, explanatory, regress, input or exogenous variable) will be denoted by x, and the response variable (some times called as dependent, target, output or endogenous variable) by y. Note also that regression can be applied simply to study the relationship between the response variable y and one or more predictor variables x, although with respect to the work presented in this thesis the interest is in prediction.

Linear regression is the most popular regression model referenced in the literature. It is considered to be most appropriate when the correlations between x and y are almost linear. However there are some limitation to linear regression, these include: (i) their inappropriateness with respect to applications that feature non-linear relationships between x and y, (ii) linear regression analysis is limited to numerical response prediction (as opposed to class label prediction as in the case of classification), and (iii) (in common with many classification techniques) a lack of explanation about the nature of a prediction [40, 123].

The most straight forward form of linear regression is Simple linear Regression (SLR) where we have only one predictor variable x. The resulting model is presented in terms of an equation of the form [94]:

$$y = a_0 + a_1 x \tag{2.3}$$

where a_0 and a_1 are the constants for the regression model such that y increases or decreases in a linear manner as x increases. In the ideal situation the generated model describes a straight line graph expressing the relationship between y to x which is fitted to a given data set. However this ideal situation does not occur frequently with respect to real-life data. The response variable (y) will have associated errors related to each predictor variable value. The difference value between an actual response value and a predicted response value is called the residual, usually denoted as e.

$$y = a_0 + a_1 x + e \tag{2.4}$$

Multiple Linear Regression (MLR), unlike SLR, operates using a number of predictor variables. The MLR model for predicting the response value can be written as follow:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + e \tag{2.5}$$

Where: y is the response variable, $x_1, x_2, ..., x_n$ are the predictor variables, e is a random error, and $a_0, a_1, ..., a_n$ are the regression coefficients extracted from the training (input) data.

In the context of image mining the utility of regression analysis is similar to that of classification as described above. Regression analysis, when applied to an image collection, again comprises two distinct steps (as indicated by the generic prediction process presented previously in Figure 2.14): (i) training and (ii) prediction. As before the training process is applied using a training set that has been appropriately pre-processed. Once a prediction model has been generated, the model can be applied to unseen image data for response variable prediction purposes.

2.5.2 Frequent Subgraph Mining

One of the techniques considered in this thesis is based on the idea of a hierarchical decomposition of stellate images as discussed in Sub-section 2.3.3. The idea is then to use Frequent Subgraph Mining (FSM) to identify a set of features that can then be used to create a feature vector representation. The technique is fully described in Chapter 4. This sub-section provides some background material concerning FSM to support the discussion presented later Chapter 4.

FSM is essentially a graph mining technique, the process of identifying hidden information in graph data. Graph representations are widely used and are seen as a powerful and flexible mechanism for representing and/or modelling entities such as circuits, chemical compounds, protein structures, biological networks, social networks, world wild web information, workflows, xml documents and image data [68, 136]. Graph mining applications include chemical informatics, computer vision, video indexing and text retrieval. [61, 98, 148, 191].

From the literature we can identify a variety of graph mining techniques which from FSM is the most significant with respect to this thesis. Frequent subgraphs, once discovered, may be used to: characterise graph sets, discriminate between different groups of graphs, classify and cluster graphs, and facilitate similarity searches in graph databases. Example applications where FSM has been used can be found in [68] and [78] where it was used for chemical analysis. FSM van be employed with respect to one single large graph or a collection of graphs, in the context of this thesis we are interested in the latter, a collection of graphs representing household images extracted from satellite image data. Thus given a graph dataset $D = \{G_0, G_1, ..., G_n\}$, $support(g)$ denotes the number of graphs (in D) in which a subgraph g exists. A subgraph g is then said to be frequent if $support(g) \geqslant \sigma$, where σ is a minimum support threshold.

The main component of any FSM algorithm is isomorphism testing, the process of checking whether a subgraph g_i is identical to a subgraph g_j. Isomorphism testing is required with respect to candidate subgraph generation and support counting. Isomorphism testing is the main computational overhead associated with FSM. The majority of FSM algorithms seek to limit the amount of isomorphism testing that is required. One example is the Apriori-based Graph Mining (AGM) algorithm proposed in [83], which was further developed in [97] to give the Frequent Subgraph Mining (FSM) algorithm based on the idea of using what the authors refer to as the "adjacent representation" of graphs and an "edge-growing" strategy. Both AGM and FSM take advantage of the Apriori approach, presented in [1] in the context of frequent item set mining for tabular data, whereby if a k edge subgraph is not frequent none of its $k+1$ edge subgraphs will be frequent. Apriori style FSM algorithms operate in a three step manner: (i) candidate generation, (ii) support counting and (iii) pruning of graphs according to the σ threshold value. During candidate generation $k+1$ edge candidate subgraphs are generated from the frequent k edge subgraphs identified on the previous iteration, a process known as *subgraph growing*.

Algorithm 1 Frequent Subgraph Mining Process

```
1:  INPUT G = {G_1, G_2, ..., G_n}, σ = threshold;
2:  OUTPUT S = {S_1, S_2, ..., S_n};
3:  S = null;
4:  k = 1;
5:  C_k = all one edge candidate subgraph in G;
6:  loop
7:      L = set of occurrence counts for each G_i ∈ C_k obtained using an isomorphism process,
    with one to one corresponding with C_k;
8:      F = set of frequent subgraph in C_k, where for each g_i ∈ C_k   1 ∈ L < σ;
9:      S = S ∪ F;
10:     k++
11:     C_k = set of k-edge subgraphs extended from F using right most extension
12:     if (k==null) then exit
13: end loop
```

The frequent subgraph process is describing in Algorithm 1. The input consists of two variables: (i) a collection of graphs (each graph represents a image) denoted by G and (ii) the threshold value σ. The output is a collection of frequent subgraphs denoted by S. The process begins with the initiation of some parameters: (i) the set of frequent subgraph is initially an empty set, (ii) the counter k is defined as 1, and (iii) the set of one edge candidate subgraphs is assigned to C_k. The algorithm then loops (line 6 to line 13) through the following steps: (i) determine the occurrence count for each subgraph $G_i \in C_k$ using an isomorphism testing process, (ii) compare the occurrence counts of each subgraph $G_i \in G_k$ with the σ and add those G_i whose occurrence count is greater than σ to the set F, (iii) add the set F to the set S so far, (iv) increment k and generate the next size of candidate sets C_k from F and (v) repeat the process. The loop continues until no more candidate can be generated ($C_k = \emptyset$).

2.6 Evaluation measure and Statistical Significant evaluation

Once a classification or regression model has been generated it is desirable to obtain some measure of its effectiveness. This is typically achieved by applying the model to a test set whose prediction values (class labels) are known so that comparisons can be undertaken between the predicted result and the known result. A wide variety of measures have been used with which to evaluate the performance of prediction models according to their effectiveness. In the case of the evaluation presented later in this thesis five measures were used with respect to classification performance: (i) Area Under the Receiver Operation Characteristics (ROC) curve (AUC) is frequently used [9, 34, 41, 138], (ii) Accuracy (AC), (iii) Sensitivity (SN), (iv) Specificity (SP) and (v) The F-measure (FM). Of these AUC is considered to be the best for indicating the overall quality of the classifier therefore for evaluation purposes [108], as presented later in this thesis, AUC was used as the central measure with which the different proposed approaches were compared. The AUC measure is therefore discussed in further detail here. With respect to regression analysis three measures were used: (i) Correlation Coefficient (Coef), (ii) Mean Absolute Error (MAE) and (iii) Root Mean Squared Error (RMSE)

A ROC graph is a two-dimensional plot of the True Positive (TP) rate (sensitivity) versus the False Positive (FP) rate (1-specificiy). The True Positive (TP) rate measures the proportion of correctly identified positive test records. The False Positive (FP) rate measures the proportion of incorrectly identified positive test records. Conversely, the True Negative (TN) rate measures the proportions of correctly identified negative test records, while the False Negative (FN) rate measures the proportion of incorrectly identified negative test records. A ROC graph thus illustrates the relative trade off between benefits (true positives) and costs (false positives). The bottom-left corner of the graph (coordinates $\langle 0,0 \rangle$) represents the situation where we have no true positive classifications and no false positive classifications, whereas the upper right corner (coordinates $\langle 1,1 \rangle$) represents the situation where we have no true negative classifications and no false negative classification. The ideal situation is where the true positive rate equals 1 and the false positive rate equals 0. To compare ROC curves a simple approach is to consider the area under the ROC, the AUC measure, so that the nature of a ROC is described as a single number [188]. An example of a ROC curve is shown in Figure 2.16. The diagonal line represents the trade-off between the TP rate and FP rate for a random model, and has an AUC of 0.5. In practice the ROC for a well performing classifier needs to be as far to the top left corner as possible as this is where an AUC value of 1.0 will occur.

Using AUC a number of classifiers can be compared. Given a collection of classifiers and an evaluation data set we can expect that one of them will produce a best AUC value. The question is whether this is statistically significant or not? A number of approaches have been proposed to determine whether a comparison of a number of techniques is indeed statistically significant or simply a matter of chance, these include: (i) the paired t-test, (ii) the Wilconxon Signed-Rank Test, (iii) the ANOVA test and (iv) the Friedman test. The Friedman test offers the advantages over the other techniques of: (i) simplicity of calculation, and (ii) usage of a ranking

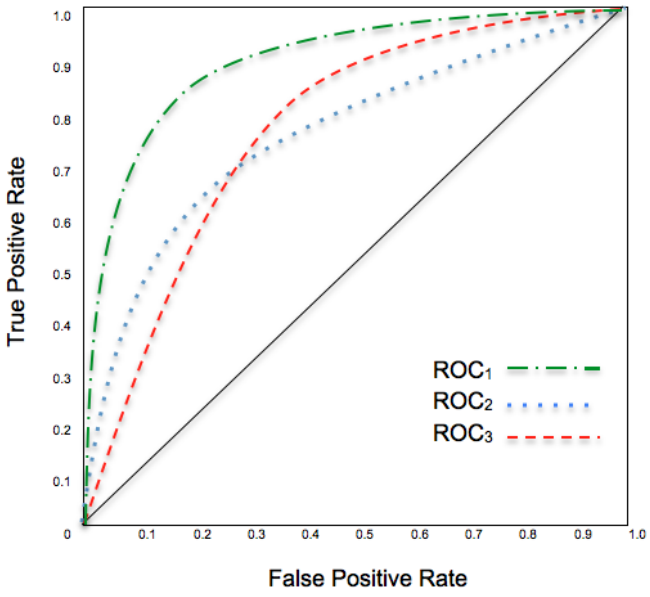


Figure 2.16: Example of ROC curves

instead of some average [53]. The Friedman test is recommended for use with "related" data sets, while the ANOVA test is recommended for use with "unrelated" data sets [35]. Therefore, the Friedman test was considered to be the most suitable statistical test for use with respect to the comparisons presented later in this thesis. The Friedman test is a non-parametric statistical test which was first introduced in [50]. The test is used to evaluate performance across a number of different models so that the null hypothesis $H0$, that there is no significant difference between the models considered, can be accepted or rejected. (The alternative hypothesis, $H1$, is that there a significant difference between the models considered.)

The Friedman test is based on the Average Ranked (AR) performance for each classification model; in this thesis performance was defined using the AUC measure described above. The Friedman test statistic χ_F^2 is then calculated as per Equation 2.6 [35, 50].

$$\chi_F^2 = \frac{12N}{K(K+1)} \left[\sum_{i=1}^{K} AR_i^2 - \frac{K(K+1)^2}{4} \right] \tag{2.6}$$

where N is the number of data sets, K is the number of classification techniques, and AR_i is the average rank for classification model i calculated as follows:

$$AR_i = \frac{1}{N} \sum_{j=1}^{N} R_{(i,j)} \tag{2.7}$$

where $R_{(i,j)}$ denotes the rank for classification model i with respect to data set j.

The assumption is that the Friedman statistical value is distributed according to the χ_F^2 distribution with $K-1$ degrees of freedom; this distribution is then used to determine whether the calculated χ_F^2 value (the Friedman statistic value) is significant or not. If the calculated value

of χ_F^2 is greater than the "null distribution", then the $H0$ can be rejected. The null distribution is a pre-determined theoretical distribution with $\alpha = 0.05$ as the most commonly used value to describe the level of significance (α) [50].

The p-value is a statistical function which is used in the context of null hypothesis testing in order to quantify the strength of the statistical significance of evidence. The p-value is the probability of the observed result assuming that the null hypothesis is true. Therefore if the p-value is equal to or greater than the significance level (α) this confirms that the $H0$ can be rejected.

On completion of the Friedman test, a post hoc Nemenyi test may be applied, if $H0$ (that there is no significant difference in operation between the models) has been rejected, so as to identify which prediction models' operation was statistically different to which other models [132]. Using the post hoc Nemenyi test differences in the performance between two or more classifiers, the Critical Difference (CD), is calculated using Equation 2.8:

$$CD = q_{\alpha,\infty,K}\sqrt{\frac{K(K+1)}{12N}} \tag{2.8}$$

where the critical value for $q_{\alpha,\infty,K}$ is based on the Studentised range statistics [86]. The performance of an individual technique is considered to be significant if the average rank differs by at least the CD value in comparison with the same other technique with which it is being compared.

2.7 Summary

This chapter has presented the relevant background and previous work with respect to the research described in this thesis. Data mining, and its sub-domain of image mining, and the challenges of satellite image categorisation were described. In the context of image pre-processing prior to the application of image mining, several image segmentation techniques that may be applied to satellite images were introduced. A review of the literature concerned with feature extraction and feature selection was also presented. The chapter was concluded with a review of the techniques used for evaluation purposes later in this thesis. In the next chapter the satellite image datasets that were used for the performance evaluation of the proposed techniques are described. The process of preparing these images for image classification/regression, including image enhancement and image segmentation, in the context of the background on image pre-processing given in this previous work chapter, will also be presented.

Chapter 3

Satellite Image Datasets

3.1 Introduction

To act as a focus for the work presented in this thesis a number of satellite image data sets were used. These data sets can be divided into two categories: (i) those used to evaluate the proposed household family size prediction techniques presented in Chapter 4, 5, 6 and 7 and (ii) that used to illustrate and evaluate the process for conducting large scale population estimation mining presented in Chapter 8. This chapter consider the first category, the second is considered in Chapter 8. This chapter commences by presenting an overview of the test sites in Section 3.2. With respect to ground truth survey, the corresponding satellite images were obtained from Google Earth, an overview of the process whereby they were obtained is presented in Section 3.3. The proposed household image segmentation process is then presented in Section 3.4. Finally the chapter is concluded in Section 3.5

3.2 Test Sites

An overview of the test sites used for evaluation purposes with respect to the work presented in Chapters 4, 5, 6 and 7 of this thesis is presented in this section. A rural area within the Ethiopia hinterland was used. More specifically the Horro district of the Oramia region of Ethiopia, which lies some 300 kilometres to the north-west of the capital Addis Ababa as shown in Figure 3.1 (in the figure the black arrow indicates the location). This area was used because a team from the University of Liverpool were operating in this area from May 2011 to July 2012 (Figure 3.2). Ground truth data sets were originally collected from two different districts: (i) Horro and (ii) Jarso (both are rural areas). Note that the priority aim for collecting the data was not to collect census data, although such data was collected as a by-product. However the quality of the available satellite imagery covering the Jarso district was found to be poor. Consequently only data from the Horro district was used to provide "ground truth" data with respect to this thesis.

The ground truth data included household geographic coordinates expressed in terms of latitude and longitude together with family size (number of people living at the household) col-

Figure 3.1: The location of Horro district, Ethiopia

Figure 3.2: Ground truth collection at one of the test sites

Table 3.1: Class label Statistics and distribution for the Site A and Site B data sets

Family Size	Minimum	Maximum	Average	Mode	Site A	Site B
Small	2	5	4.04	5	28	19
Medium	6	8	7.00	6	32	21
Large	9	12	9.80	9	10	10
Total 120	2	12	6.31	6	70	50

lected from two different sites (villages). Whatever the case, data from two sites in the Horro district was obtained, Site A and Site B, as shown in Figure 3.3, 70 households from Site A and 50 households from Site B. Site A was bounded by the parallels of latitude 9°33′44″N and 9°35′6″N, and the meridians of longitude 37°6′37″E and 37°13′37″E and Site B was by parallels of 9°43′16″N and 9°45′3″N, and the meridians of longitude 37°0′46″E and 37°8′36″E. The letters "A" and "B" in Figure 3.3 indicate the site locations. The distribution of family size over the 120 households is presented in Figure 3.4. From the figure it can be observed that a more or less normal distribution of family size was found to exist. The most frequent family sizes were found to be 6 and 8 (family sizes of 7 were not as frequent), whilst family sizes of 2 and 11 were founded to be the least frequent.

Figure 3.3: Site A and Site B locations

With respect to image classification the household collected data was separated into three classes: (i) "small family size" (2-5 people in family), (ii) "medium family size" (6-8 people in family) and (iii) "large family size" (9-12 people in family). Some statistics concerning the class distributions for the Site A and B data are shown in Table 3.1. From the table it can be seen that the minimum and maximum family size were 2 and 12 respectively, the mean was 6.31, the medium were 6 and standard deviation was 2.56. These two data sets then provided the training and test data required for our proposed census collection system. Figure 3.5 shows an example household from one of the test sites.

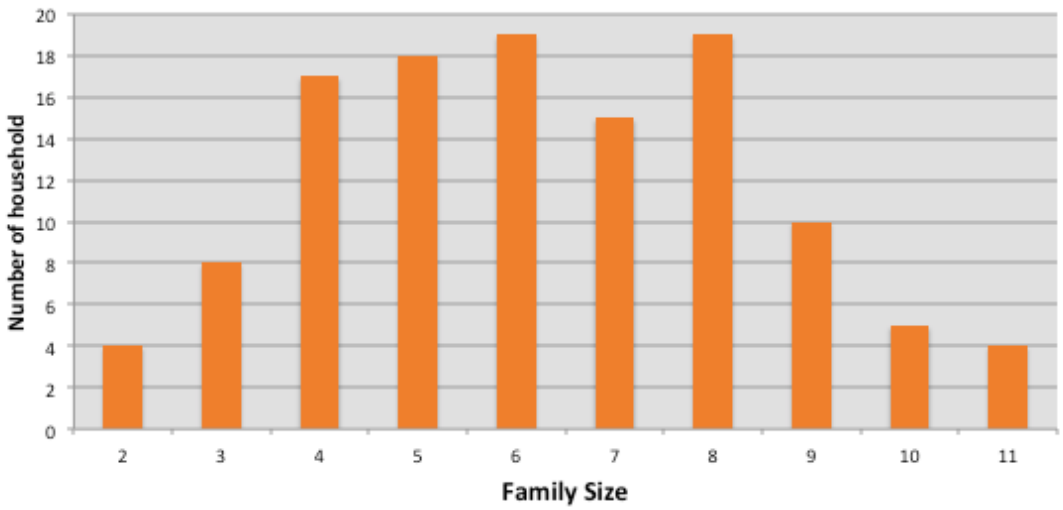


Figure 3.4: The distribution of family size over 120 households

Figure 3.5: Example of a Site A household

3.3 Satellite Image Collection

The satellite image collection process is presented in this section. The geographic coordinates (latitude and longitude) from ground truth data for each household was used to identify the households in satellite images within the Google Earth service. Figure 3.6 shows a Google Earth image indicating some of the locations of the Site A (bottom cluster) and Site B (top cluster) households used, note that in the figure households are marked and identified using an ID code.

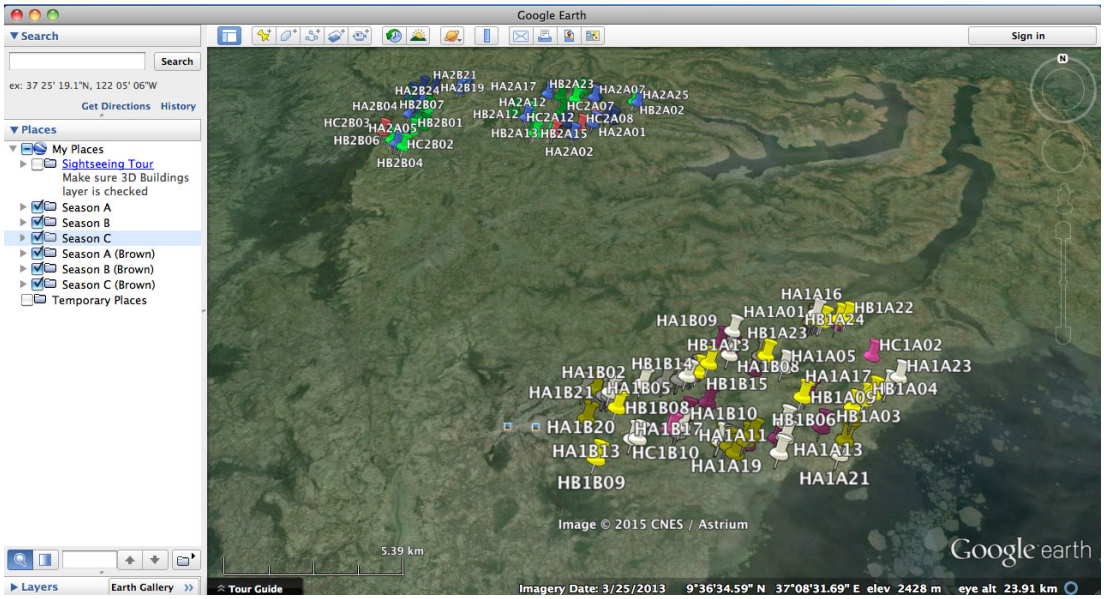


Figure 3.6: Google Earth image indicating the locations of some of the ground truth households uses

The images were originally obtained using the GeoEye satellite with a 50 centimetre ground resolution. The satellite images for Site A were released by Google Earth on 22 August 2009 (Figure 3.7(a)) and those for Site B (Figure 3.7(b)) on 11 February 2012. The Site B satellite images were obtained during the "dry season" (September to February), while the site A images were obtained during the wet season (June to August). From Figure 3.7(a) the households can be clearly identified, many of the households have tin roofs which are easy to differentiate from the (green) backgrounds, the households are less easy to identify in Figure 3.7(b) where they tend to merge into the (light-brown) background. A "close up" of a household is presented in Figure 3.8.

To act as a focus for the work presented in this thesis a number of satellite image data sets were used. These data sets can be divided into two categories: (i) those used to evaluate the proposed household family size prediction techniques presented in Chapters 4, 5, 6 and 7; and (ii) that used to illustrate and evaluate the process for conducting large scale population estimation mining presented in Chapter 8. The first comprised a collection of 120 households

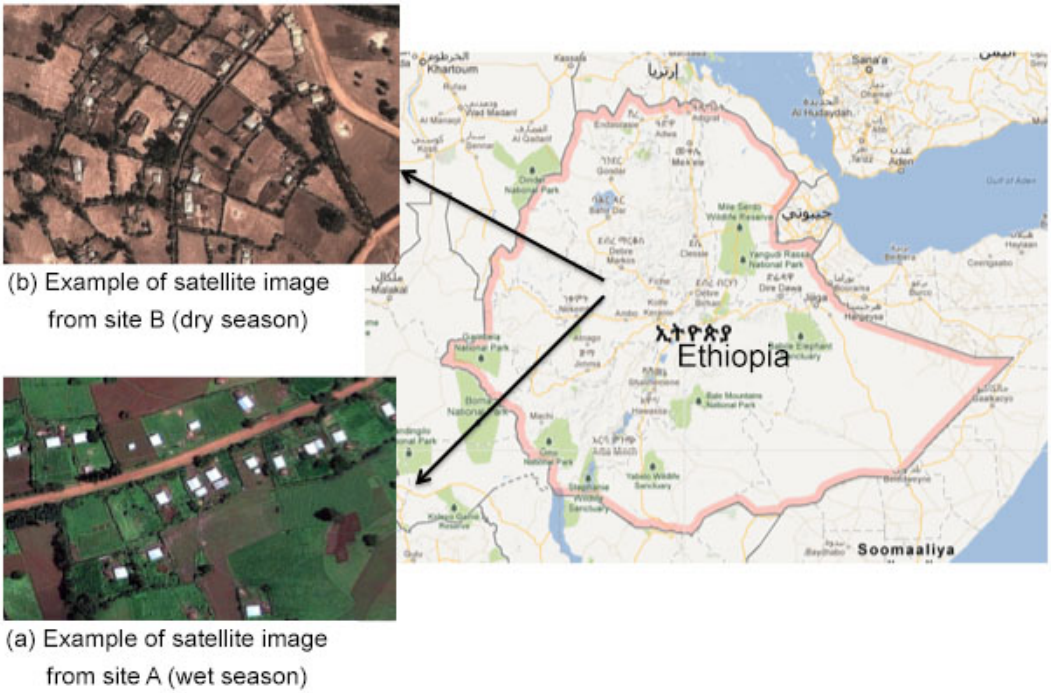


Figure 3.7: Examples of satellite images for the two different test sites (A and B)

that were obtained from Google Earth in March 2012.The second comprised a collection of 600 images that were obtained from the Google Static Map service in March 2014, both data sets were obtained with respect to the Horro district.

3.4 Household Image Segmentation

This section presents the adopted process for segmenting the Site A and Site B household image collection obtained as described above. A typical satellite image from test site A is presented in Figure 3.9. From the figure it can be seen that the households of interest are typically defined by a rough rectangular boundary in which buildings and related objects are located (although other boundary shapes may occur). Consequently an edge-based segmentation technique was adopted for the purpose of identifying (segmenting) households. The fundamental idea was to identify the rectangles that surrounding households. For this purpose a sequence of techniques was applied: (i) Canny edge detection, (ii) the Hough transform and (iii) least squares line fitting.

Canny edge detection is an edge detection technique for identifying object contours from intensity discontinuities [20], further detail of this technique is presented in Sub-section 3.4.1. The Hough transform is a segmentation technique suited for identifying imperfect instances of objects of certain predefined shapes (such as the straight lines making up a rectangle) [42],

Figure 3.8: Example of obtained satellite image

and is described in further detail in Sub-section 3.4.2. Least squares line fitting is a process of identifying the best fit line for a set of data points, and is discussed in Sub-section 3.4.3. The overall household image segmentation process is presented in Sub-section 3.4.4.

3.4.1 Canny Edge detection

The term "edge" is used to describe pixels in image analysis. Edge refers to a set of pixels where a significant local change of intensity occurs. When some significant local change in intensity occurs in pixels, they indicate a boundary between two different regions in an image. For example, a typical edge might be the border between a block of the colour blue and a block of the colour yellow. A line in an image, for instance, has two edges; one edge on each side of the line. Edge detection refers to the process of identifying and locating such pixel sets. There are numerous edge detection methods that have been developed. In the context of work presented in this thesis, as noted above, Canny edge detection was used to facilitate the process of segmenting individual households.

The Canny edge detection algorithm is an "optimal detector" and is adaptable to various environments. Assessment of Canny edge detection against the following criteria makes it optimal. First, edge detection should provide low error rate. In edge detection, failure to detect some edges and providing false positives by responding to non-edges should be minimised as much as possible. Second, an edge points should be well localised. Typically, the distance between the actual edge and located position of the edge should be minimum. Third, an algo-

Figure 3.9: Satellite image obtained from test site A.

rithm should have one response to a single edge. This criterion is included for cases when the first two substantially fail to completely eliminate the possibility of multiple responses to an edge [20].

Based on these criteria, the canny edge detector commences by applying a filter to "smooth" the image, to eliminate noise. Then it uses the intensity gradients within the image, to determine the "edge" pixel sets that feature high spatial derivatives. The algorithm then tracks along the identified edges and suppresses any pixel that is not at the maximum (non-maximum suppression). Next, the gradient array is further reduced using a hysteresis process. Hysteresis is used to track along the remaining pixels that have not been suppressed. Hysteresis uses two thresholds. If a pixel magnitude is below the first (low) threshold then the current pixel is set to zero (typically marking a non edge). If the magnitude is above the second (high) threshold then the pixel is considered to be part of an edge. If a pixel magnitude is between the two thresholds, then it is set to zero unless there is a path from this pixel to a pixel with a gradient above the second threshold [121, 170].

Canny edge detection has been applied in many application domains; example applications in the context of satellite images can be found in [75, 174, 190]. With respect to household segmentation, Canny edge detection was applied to identify the lines (edges) in a satellite image in order to facilitate the application of the Hough transform as described in the following sub-section.

3.4.2 Hough Transform

The Hough transform [42] is basically a technique for line identification in an image, but the technique has been extended so that it can be used to facilitate the identification of arbitrary shapes such as circles or rectangles [10, 199]. As mentioned before the boundary of the households of interest have a rectangular shape. Thus the Hough transform is applied to identify particular combinations of lines that form rectangles. The advantages of the Hough transform technique is that it is tolerant of gaps in feature boundary descriptions and that it is unaffected by noise.

More specifically the Hough transform is applied to transform the points in a Cartesian image space to straight lines. In general a straight line can be presented in the form shown in Equation 3.1.

$$y = mx + b \tag{3.1}$$

However, when using the Hough transform the alternative straight line equation given; in Equation 3.2 is used, where r is the distance from the origin to the closest point on the straight line, and θ is the angle between the X-axis and the line connecting the origin with that closest point.

$$\rho = xcos\theta + ysin\theta \tag{3.2}$$

The Hough parameter space for lines thus consists of two dimensions: θ and ρ, and a line is represented by a single point corresponding to a unique set of parameters (θ_0, ρ_0). The line-to-point mapping is illustrated in Figure 3.10.

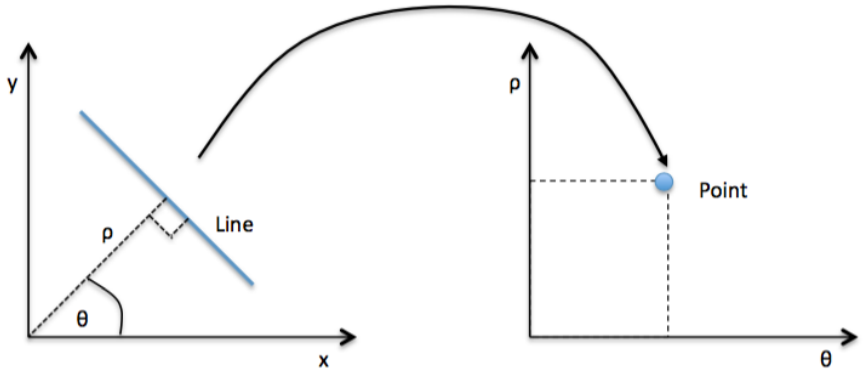


Figure 3.10: Mapping of one unique line to the Hough parameter space

To use the Hough transform some edge detection algorithm, such as Canny edge detection algorithm described above, must first be applied to determine potential edges. The edge points are then translated into the Hough parameter space and recorded in an accumulator. The Hough parameters in the accumulator are then interpreted as lines of infinite length. Finally the infinite length lines are converted to finite lines which are then overlaid with the original image.

There are many examples in the literature where the Hough transform has been applied to satellite images, in the context of a variety of application domains, example can be found in [60, 113, 147, 180].

3.4.3 Least Squares Line Fitting

The process of Curve/Line fitting is the process of finding a best fit curve/line among a collection of data points [7, 185]. In the context of the work presented in this thesis, curve fitting is applied in order to fit the straight lines identified using the Hough transform to some predefined shaped (a rectangle in our case). One technique for doing this is least squares line fitting. In this sub-section a brief overview of Least Squares line fitting [23, 130] is thus presented to aid understanding of the household image segmentation process. Given a set of ordered pairs (x_1, y_1), (x_2, y_2),..., (x_n, y_n), to find the line of best fit first the mean of the x values and the mean of the y values are calculated using Equations 3.3 and 3.4, respectively. The slope (m) of the line of best fit is then calculated using Equation 3.5

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.3}$$

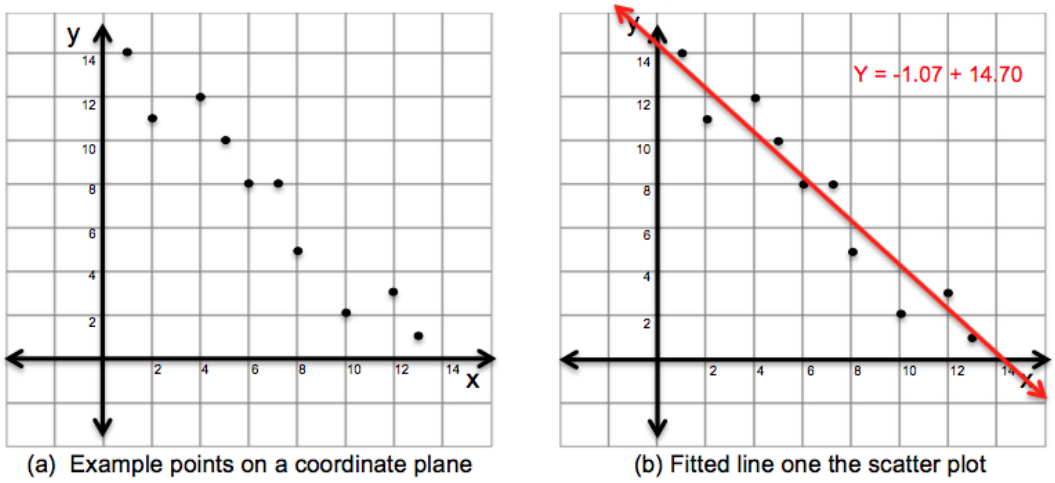
$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.4}$$

$$m = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \tag{3.5}$$

Next the $y - intercept$ of the line is calculated using the formula shown in Equation 3.6. Finally, the $y - intercept$ (b) and slope m are used to form the equation of the line (Equation 3.1).

$$b = \overline{Y} - m\overline{X} \tag{3.6}$$

Figure 3.11 shows an example of line fitting using the least squares methods. The figure shows a collection of example points in Figure 3.11(a) and a line fitted to the points by calculates the slope m and $y - intercept$ (b) in Figure 3.11 (b).



(a) Example points on a coordinate plane (b) Fitted line one the scatter plot

Figure 3.11: Example of Line fitting using the Least Squares technique

In the context of satellite image applications, an example of Least Square Line Fitting can be found in [16]. In the context of the work presented in this thesis, as already noted, Least Square Line Fitting is applied to fit the lines so as to identify the anticipated rectangular boundary of the household.

3.4.4 Household Image Segmentation Process

This sub-section describes the image segmentation processes applied to the input satellite data. Due to the complexity of the segmentation process, as described above the process is illustrated in Figure 3.12. An example of a household satellite image obtained from Google Earth of the Site A is shown in Figure 3.12 (a). However, before household segmentation can be applied it was first necessary to register and align the image so that each household was aligned in a north-south direction. The purpose of this registration and alignment was to facilitate later

feature segmentation. The result is a shown in Figure 3.12(b), where the image shown in Figure 3.12(a) has been appropriately aligned.

The next step was to convert each RGB colour represented images into an sixteen colour indexed image (as shown in Figure 3.12(c)), and then to transform this into a greyscale image to which histogram equalisation can be applied. Recall that histogram equalisation is a method for image enhancement directed at ensuring an equal colour distribution [55, 63], (Figure 3.12(d)), detail concerning histogram equalisation was presented in Chapter 2, Sub-section 2.3.1. The process commences by selecting a reference image which is then used for normalisation purposes with respect to the remaining images.

Once the image enhancement process was complete the next stage was to segment the images so as to isolate individual households. It has already been noted that the households of interest are typically defined by a rough rectangular boundary in which buildings and related objects are located. The aim was to segment these images so that these rectangular areas can be clearly isolated, however the boundaries are frequently not well defined in that the edges are not continuous. Thus, for example, region-growing segmentation techniques would be unlikely to perform well, instead line (edge) segmentation was adopted as a more suitable form of segmentation for the given application domain.

More specifically, for the purpose of image segmentation the Canny edge detection algorithm [20] and the Hough Transform [42] were applied (see details above). Prior to applying Canny edge detection and the Hough transform, contrast adjustment was applied to the images so that the household boundaries could be more readily distinguished.

Canny edge detection was then applied, the result as shown in Figure 3.12(e); in the figure the detected edges have been highlighted. As a result of applying the Hough transform, we have a collection of "lines" as shown in Figure 3.12(f). Each line is defined by a start and end point, and a ρ and θ value (length and direction).

The next part of the process is to fit a rectangle to this set of lines. This was achieved by applying a Least squares approach [18], applied to each group of lines approximating to the top, bottom, left and right sides of a rectangle (see the detail above). As a results the rectangle surrounding each household was demarcated by a pair of horizontal and a pair of vertical lines (Figure 3.12(g)). The intersections of the lines can then be found so as to delimit the surrounding rectangle in term of its four corners (Figure 3.12(h)). The boundaries found were then applied to the original image to determine the area of each individual household. The final result is a set of segmented household images such as that as shown in Figure 3.12(i).

The segmentation process as applied to a Site B household image is shown in Figure 3.13. Figure 3.13(a) is the original image. Figure 3.13(b) is the image after Figure 3.13(a) has been appropriately aligned. Figure 3.13(c) is the sixteen colour indexed image. Greyscale transformation and histogram equalisation was then applied to give Figure 3.13(d). The resulting image after Canny edge detection has been applied is presented in Figure 3.13(e). The line detection technique using the Hough transform was then used as shown in Figure 3.13(f).Figure

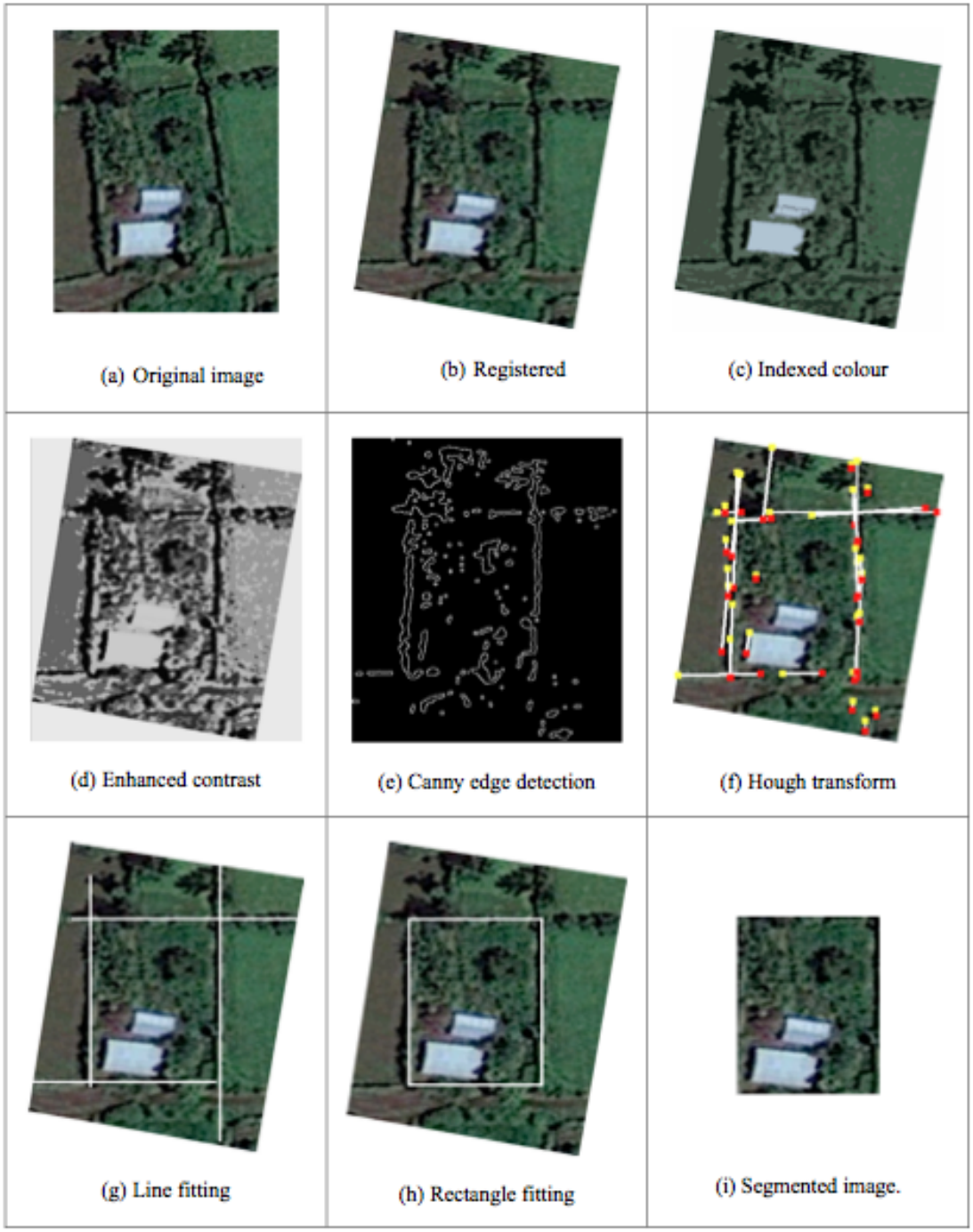


Figure 3.12: Household segmentation process for Site A data

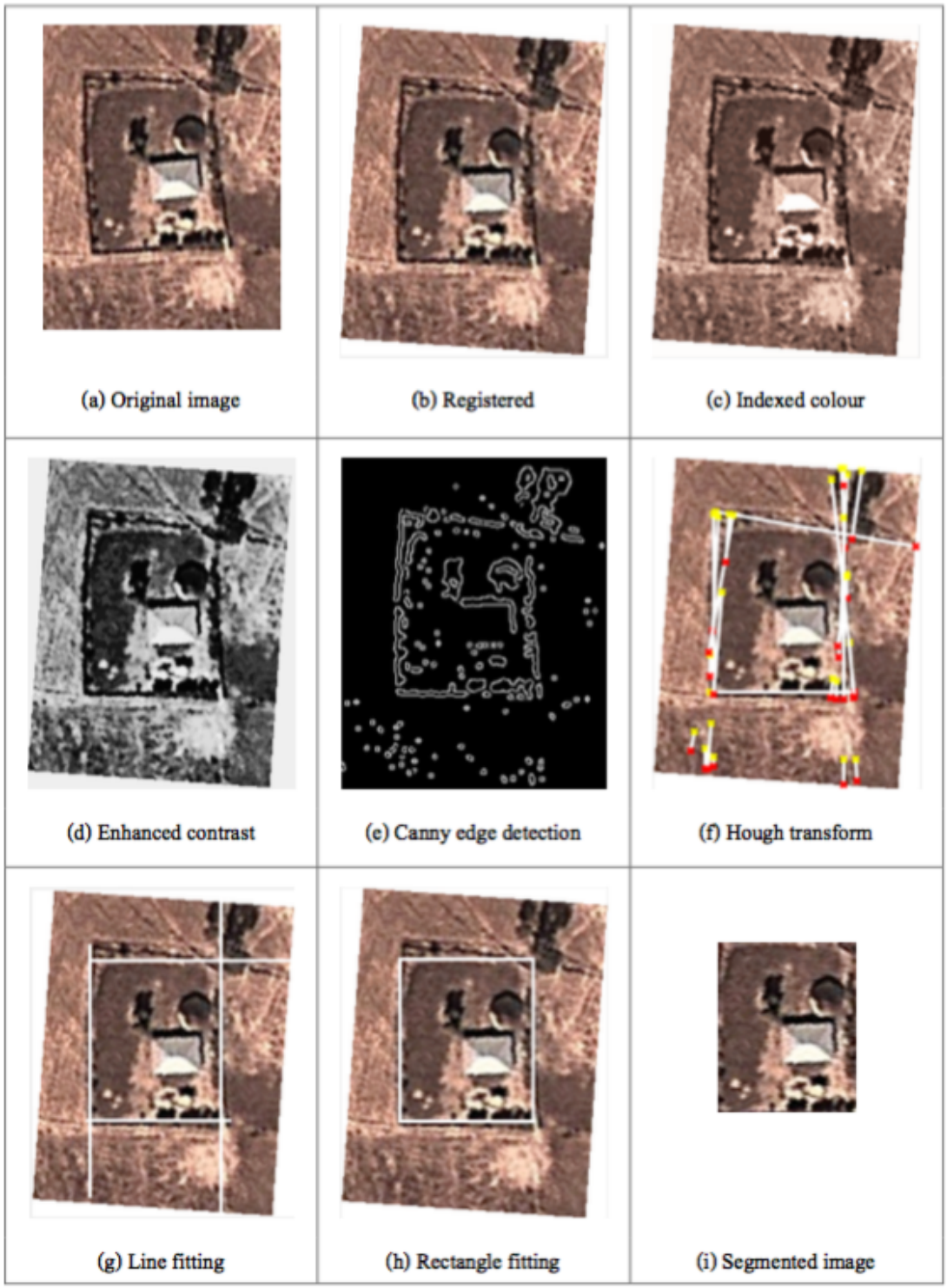


Figure 3.13: Household segmentation process for Site B data

3.13(g) shows the result when line fitting is applied using the least squares approach. Rectangle shape fitting was then applied as presented in Figure 3.13(h). The final segmented household is shown in the Figure 3.13(i).

3.5 Summary

This chapter has introduced the Site A and Site B test set data and provided the necessary context to the preprocessing of the satellite image data so as to segment the desired household areas. The segmentation process was fully described. The results were two collections of images. Note that each image has a family size (class label) associated with it. The distribution of these class labels was given in Table 3.1. Note that some initial experiments (not reported here) directed at evaluating the relationship between family size and household image size found that there was no correlation between the two. It was conjectured that this might be because poorer families lived in smaller households than richer families of the same size. Therefore, in the following three chapters three different techniques, whereby the satellite household image data can be classified are described and evaluated. The first technique considered is a graph-based technique which is presented in the following chapter.

Chapter 4

Population Estimation Mining using Satellite Imagery: The Graph-Based Approach

4.1 Introduction

This chapter considers the first of the three image representation approaches considered in this thesis, the colour histogram based and texture based representations are considered in Chapters 5 and 6 respectively. The idea promoted in this chapter is to capture the nature of each segmented household image, using a graph-based representation. In the case of the training data each graph represented household has a "family size" class label associated with it (see Table 3.1 presented in previous chapter). This training data can then be used to build a graph-based classifier that can be used to predict household sizes according to the nature of the proposed graph structure representation.

More specifically, in this chapter an image decomposition approach is considered whereby the individual households are represented using a quadtree (graph) decomposition; we refer to this as "fine segmentation" to distinguish it from the "coarse segmentation" used to identify households as described in Chapter 3. Once a set of households has been fine segmented the next stage of the data preparation phase is to translate the segmented pixel data into a form suitable for the application of a classifier. The translation needs to be conducted in such a way that all salient information is retained while at the same time ensuring that the representation is concise enough to allow for effective further processing. The fundamental idea here is to adopt a graph-based representation, more specifically a quadtree based representation (one per household). Quadtrees have been used extensively in the context of image processing (see for example [159]). However, the quadtree representation does not lend itself to ready incorporation with respect to classification algorithms. To do this we propose applying subgraph mining to the quadtree data to identify frequently occurring patterns across the data that can be used as features in the context of a feature vector representation. The patterns of interest are thus frequently occurring subgraphs. A schematic of the graph-based approach is given in Figure

4.1

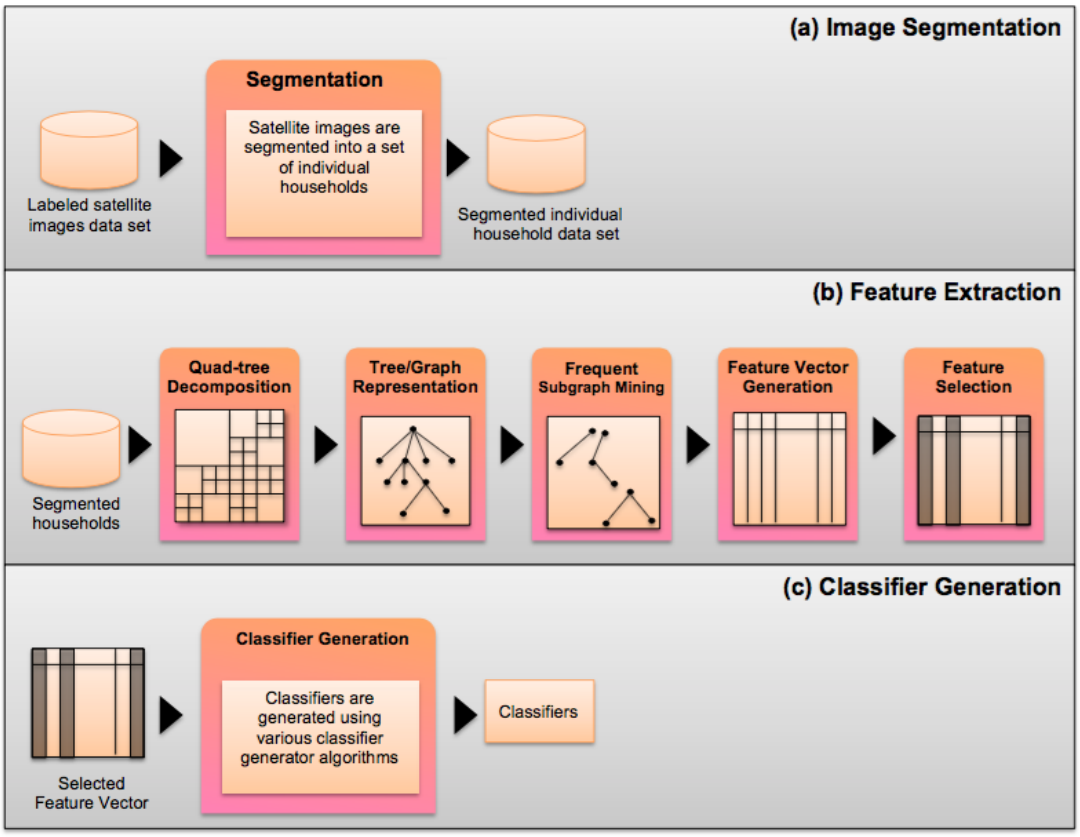


Figure 4.1: Schematic illustrating the Graph-Based Framework

A schematic of the graph-based approach is given in Figure 4.1. From the figure it can be seen that the graph-based approach for census mining consists of three processes: (a) image segmentation, (b) feature extraction and (c) classifier generation. The image segmentation process (the top rectangular box in the figure) was discussed in Chapter 3 and will thus not be considered further in this chapter. Once a set of individual segmented household has been identified, the next process, feature extraction, is used to translate the segmented pixel data into an appropriate form suitable for classifier generation (the third process within the framework). The classifier generation process is straight forward and requires little further consideration here.

The feature extraction process comprised a number of sub-processes (as shown in Figure 4.1). The fundamental idea underpinning the graph-based approach was to use a quadtree based representation (one per household). In this manner a set of subgraphs that frequently occur in the data could be identified which could then be used with respect to a feature vector representation of the form used by many classifier generation algorithms. The sub-processes

that make up the feature extraction process are: (i) quadtree decomposition, (ii) tree/graph representation, (iii) frequent subgraph mining, (iv) feature vector generation and (v) feature selection. Note that with respect to Figure 4.1 the approach presented is generic in nature and, as will be seen later in this thesis, similar approaches were used with respect to the colour histogram based and texture based approaches presented in Chapters 5 and 6 respectively.

The rest of this chapter is organised as follows. Quadtree decomposition is considered in Section 4.2, whilst Section 4.3 provides the detail of the tree/graph representation sub-process. The frequent subgraph mining and feature vector generation sub-processes are described in Section 4.4. Feature selection is then discussed in combination with classifier generation in Section 4.5. The evaluation of the proposed graph-based framework for population estimation mining using satellite imagery is then presented in Section 4.6. Finally, some further discussion and a summary are presented in Section 4.7 and 4.8 respectively.

4.2 Quadtree Decomposition

Image decomposition is a methodology for "factorising an input image into a set of components" [25]. Image decomposition has been used in the context of: (i) computer vision and computer graphic, (ii) image segmentation, (iii) image recognition and (iv) motion estimation. From the literature there are various image decomposition methods that have been proposed including quadtree, pyramid (both Gaussian and Laplacian pyramids), wavelet and scale-space representation [155, 183]. With respect to the proposed graph-based approach a quadtree representation is applied to each individual household. Quadtree decomposition is a hierarchical approach to image decomposition that naturally lends itself to a quadtree data structure. The most commonly used quadtree representation is the region based quadtree where, at each level and branch of the decomposition, a given image is decomposed into four equal regions (quadrants) [156]. The main issue with decomposition techniques is the stopping criteria to be adopted, this can be expressed in terms of some maximum level of decomposition or the homogeneity of the decomposed regions.

In the case of the household data considered in this thesis a region quadtree decomposition was used. An example is presented in Figure 4.2. With respect to this figure, Figure 4.2(a) gives the original image, Figure 4.2(b) shows a $2^3 \times 2^3$ binary array of the image where '1s' are the pixels inside the region and '0s' are the pixels outside the region. The resulting quadtree decomposition is shown in Figure 4.2(c). With respect to the quadtree storage structure typically used to encapsulate a quadtree decomposition, the root node represents the whole image, the immediate child nodes of the root nodes each represents a region quadrant and so on. The tree structure terminates with leaf nodes. In the example given in Figure 4.2 the process stops when homogenous regions are arrived at (regions comprised of all black pixels or all white pixels). The quadtree associated with the decomposition shown in Figure 4.2(c) is shown in Figure 4.2(d).

In the context of the proposed graph-based approach to population estimation mining from

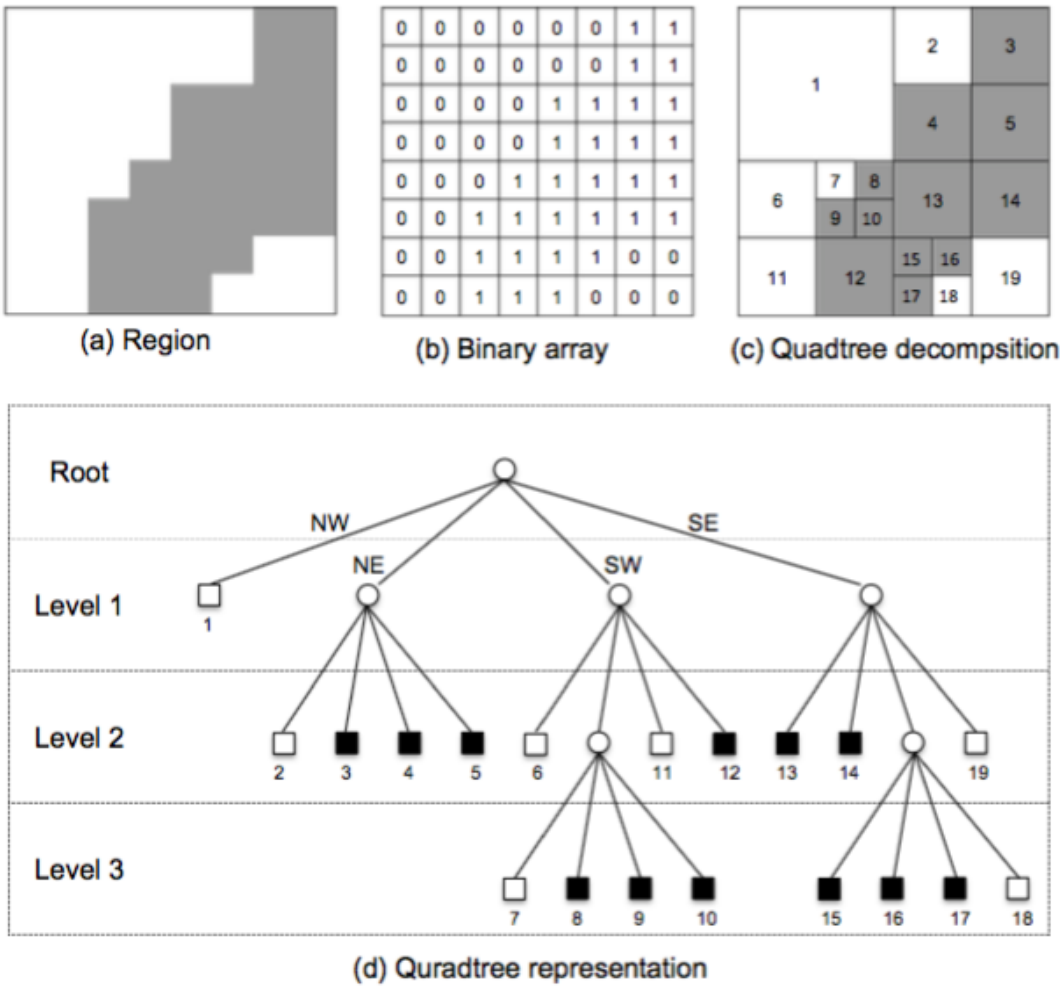


Figure 4.2: Schematic illustrating the quadtree decomposition process inspired from [156]

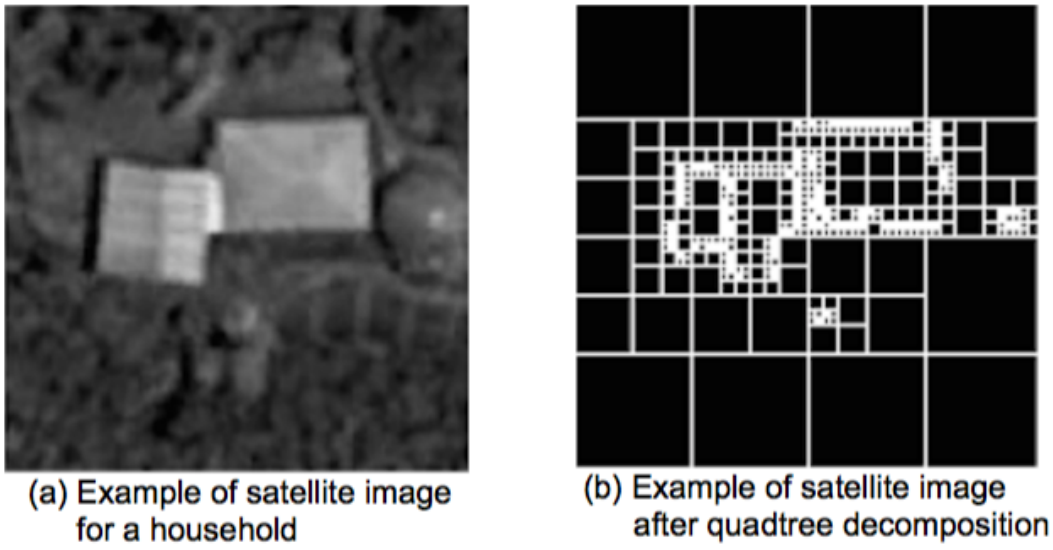


Figure 4.3: Example of a quadtree decomposition

satellite imagery, the quadtree decomposition sub-process commences by "cropping" each household image so that it is turned into a 128×128 pixel square image surrounding the main building comprising the household (this is automatically identifiable because it is the largest contiguous "white" region). Colour transformation from RGB into greyscale together with some image enhancement is then applied to each household image so that the contents of the image becomes more clear. Once the image has been enhanced, it is decomposed into sub-regions, as shown in Figure 4.3, until either: (i) uniform regions are arrived at or (ii) a maximum level of decomposition was reached. Figure 4.3(a) shows an example of a preprocessed household image while Figure 4.3(b) shows the associated quadtree decomposition. The generated decomposition was then stored in a quadtree format as described above.

4.3 Tree/Graph Representation

Once the household images have been decomposed and stored in the quadtree format, as described above, the nodes were labeled with a greyscale encoding generated using a mean intensity of the greyscale colours in each region, in this manner eight labels were derived, each describing a range of 32 consecutive intensity values. Figure 4.4 presents an example of a quadtree where the top level node (the root) represents an entire (cropped) household image, the next level (Level 1) are the root node's immediate child nodes, and so on. In the figure the nodes are labelled numerically from 1 to 8 to indicating the greyscale ranges. The edges were labelled using a set of identifiers 1, 2, 3 and 4 representing the NW, NE, SW and SE relationship between nodes reference by a particular parent node. In Figure 4.4 the number

in square brackets alongside each node is a unique node identifier derived according to the decomposition.

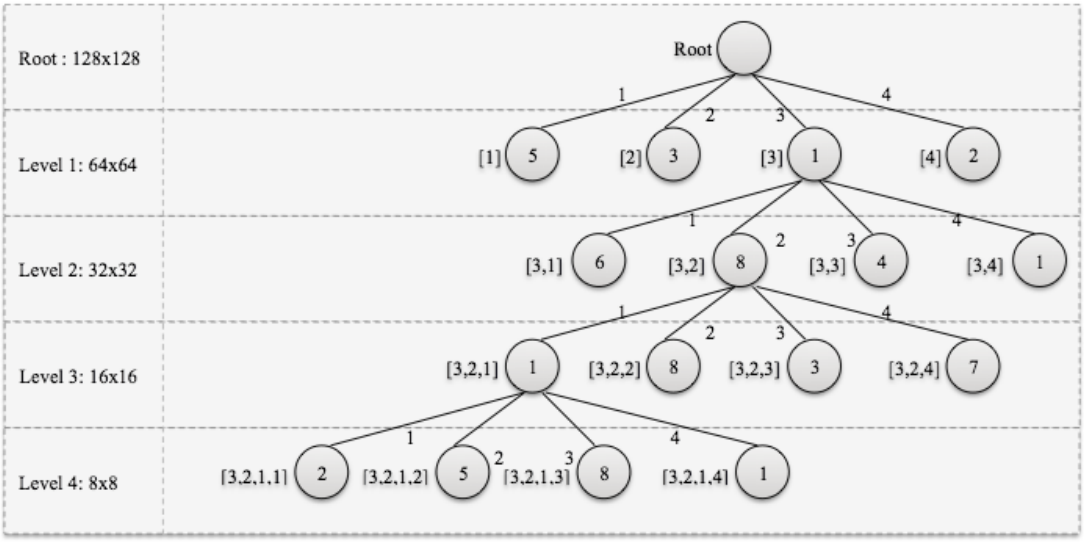


Figure 4.4: The implemented quadtree representation

4.4 Frequent Subgraph Mining and Feature Vector Generation

The quadtree (graph) based representation served to capture the content of individual fine segmented household images, although a disadvantage of the representation is the "boundary problem" where objects of interested may be located at the intersection of a decomposition. A second disadvantage is that the quadtree representation is not directly suited to the purpose of classifier generation and subsequent usage of the generated classifier. It is therefore proposed in this thesis that if frequently occurring subtrees (subgraphs) could be identified, these could be considered to be as features within a feature vector representation on the form compatible with many classifier generators. The motivation was the conjecture that such frequent subtrees would be indicative of commonly occurring features that might exist across the image set which in turn might be indicative of individual class labels. Frequent Subgraph Mining (FSM) is an established area of data mining research as discussed in Sub-section 2.5.2. Recall that given a graph data set $D = \{G_0, G_1, \ldots\}$ the support of some subgraph g is the number of occurrences of g in each graph G_i in the data set D, one count per graph. A frequent subgraph is then one whose support value exceeds some threshold σ, thus $sup_D(g) \geq \sigma$. The frequent subgraph mining problem is thus directed at finding all the frequent subgraphs in D. The value for σ is usually user specified; the lower the value of σ the greater the number of frequent subgraphs that will be discovered.

Once a set of frequently occurring subgraphs has been identified these can be arranged into

a feature vector representation such that each vector element indicates the presence or absence of a particular subgraph with respect to each household (record) as shown in Table 4.1. With reference to the table each row represents an individual household (record) numbered from 1 to m, and the columns individual frequent subgraphs represented by the set $\{S_1, S_2, \ldots, S_n\}$. The values 0 or 1 indicate the absence or presence of the associated subgraph for the record in that row. This feature vector representation is ideally suited to both the application of classifier generation algorithms and the future usage of the generated classifiers.

Table 4.1: The example of Feature Vector

Vector	S_1	S_2	S_3	S_4	S_5	...	S_n
1	1	0	1	1	1	...	1
2	1	1	0	1	1	...	0
3	1	0	1	0	1	...	1
4	0	0	1	1	0	...	0
...	...	...	...	...	...	...	...
m	0	1	1	0	1	...	1

4.5 Feature Selection and Classification

Once the feature vector generation sub-process was completed, but before classification model generation could commence, the input data was first discretised (ranged). A further challenge was the large number of features (subgraphs) identified. Note that increasing the σ threshold will address this issue, but significant features may then be missed. A feature selection strategy was thus adopted so as to reduce the number of dimensions in a manner whereby only highly discriminative features were retained. In general, feature selection algorithms are a combination of some search technique coupled with some evaluation measure for scoring features as discussed in Section 2.4 previously. With respect to the research described in this thesis three evaluation measures were considered: (i) Chi-Squared, (ii) Information Gain and (iii) Gain Ratio. On completion of the feature selection process, each household image was described in terms of a reduced number of features (a feature vector of reduced length).

Next the classification process was applied. Extensive evaluation was conducted so as to test the operation of the different parameters and their variations, however this chapter only reports the most significant results obtained (there is insufficient space to allow for the presentation of all the results obtained). With respect to the evaluation presented in the following section, eight classification generation methods were used: (i) Decision Tree generators (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network; all taken from the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench.

4.6 Evaluation

The evaluation of the proposed population estimation mining process is presented in this Section. The presented evaluation was conducted with respect to the Site A and B data sets previously introduced in Chapter 3. The overall aim of the evaluation was to provide evidence that census data can be effectively estimated using the proposed graph-based approach. To this end four sets of experiments were conducted as follows:

1. **Data Representation**: A set of experiments to identify the most appropriate support threshold, σ, for use with respect to the frequent subgraph mining (Sub-section 4.6.1).
2. **Feature Selection**: A set of experiments to examine the most appropriate feature selection algorithm (Sub-section 4.6.2).
3. **Number of attributes**: A set of experiments to analyse the most appropriate number (k) of features to retain during feature selection (Sub-section 4.6.3).
4. **Classification Generation Method**: A set of experiments to determine the most appropriate classifier generation method (Sub-section 4.6.4).

Each is discussed in further detail in the indicated Sub-sections. Ten fold Cross-Validation (TCV) was applied throughout and performance recorded in terms of: (i) accuracy (AC), (ii) Area Under the ROC curve (AUC), (iii) sensitivity (SN), (iv) specificity (SP) and (v) the F-Measure (FM), although as noted in Section 2.6 AUC was considered to be the most significant measure to be used when comparing approaches.

Table 4.2: Number of identified frequent subgraph features produced using a range of σ values with respect to the Site A and B data

σ value	Site A	Site B
$\sigma = 10$	757	420
$\sigma = 20$	149	119
$\sigma = 30$	49	60
$\sigma = 40$	24	39
$\sigma = 50$	12	19

4.6.1 Data Representation

In order to investigate the effect the value of the subgraph mining support threshold σ had on classification performance a sequence of different σ values were considered ranging from 10 to 50 incrementing in steps of 10. For the experiments the Gain Ratio feature selection technique was used to select $k = 55$ features because, as demonstrated later in this chapter, $k = 55$ was

found to produce good results. For similar reasons Bayesian Network was used for classifier generation purposes.

The number of features (subgraphs) generated in each case is presented in Table 4.2 and Figure 4.5. From the table and figure it can be seen that, as would be expected, the number of identified subgraphs decreases as the value for σ increases (and vice-versa). Note that for $\sigma = 30$, $\sigma = 40$ and $\sigma = 50$ the number of subgraphs generated is less then $k = 55$, thus in these cases all the identified subgraphs are used for classifier generation. Note also that attempts to conduct the subgraph mining using σ values of less than 10 proved unsuccessful due to the computational resource required (subgraph mining is computationally expensive).

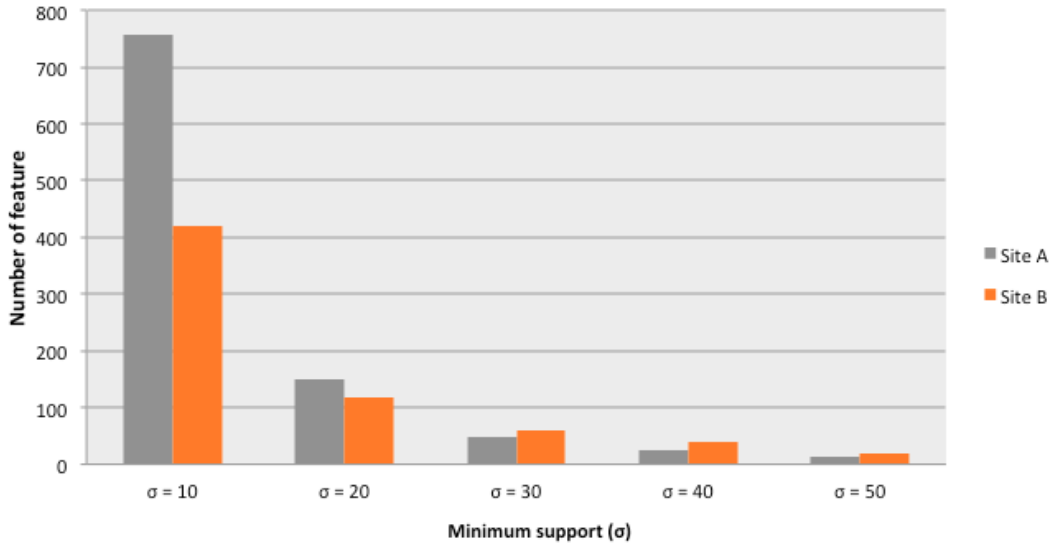


Figure 4.5: Bar graph representing the results presented in Table 4.2

The results from the experiments to determine the most appropriate value for σ, in terms of classification effectiveness, are presented in Table 4.3 (best values are highlighted in bold). From the table it can be observed that best results were obtained using $\sigma = 10$ for both Site A (wet season) and Site B (dry season); giving AUC values of 0.808 and 0.879 respectively. Note that interpretation of the table suggests that better results could be produced by decreasing the value of σ further but, as noted above, this was found to be too computationally expensive.

Figure 4.6 displays the AUC results presented in Table 4.3 in the form of a graph. The X-axis represents the σ values and the Y-axis the AUC scores obtained. From the figure it can be seen the maximum AUC value was produced when using $\sigma = 10$ for both Site A and Site B. The recorded AUC values for Site B were slightly higher when compared to Site A. However, from both sites, the graphs shows a dramatic fall in AUC when $\sigma = 20$ and $\sigma = 30$. The decrease in AUCs continues to fall, for both sites, as σ is increased further. Thus better results were obtained when using a low support threshold (σ) because with low thresholds more frequent

Table 4.3: Classification performance using a range of σ values ($k = 55$)

σ	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
$\sigma = 10$	**0.600**	**0.808**	**0.596**	**0.600**	**0.734**	**0.800**	**0.879**	**0.792**	**0.800**	**0.876**
$\sigma = 20$	0.429	0.606	0.424	0.429	0.639	0.520	0.697	0.519	0.520	0.742
$\sigma = 30$	0.329	0.427	0.336	0.329	0.600	0.380	0.535	0.382	0.380	0.666
$\sigma = 40$	0.343	0.396	0.341	0.343	0.562	0.320	0.495	0.306	0.320	0.601
$\sigma = 50$	0.371	0.446	0.354	0.371	0.557	0.320	0.461	0.312	0.320	0.603

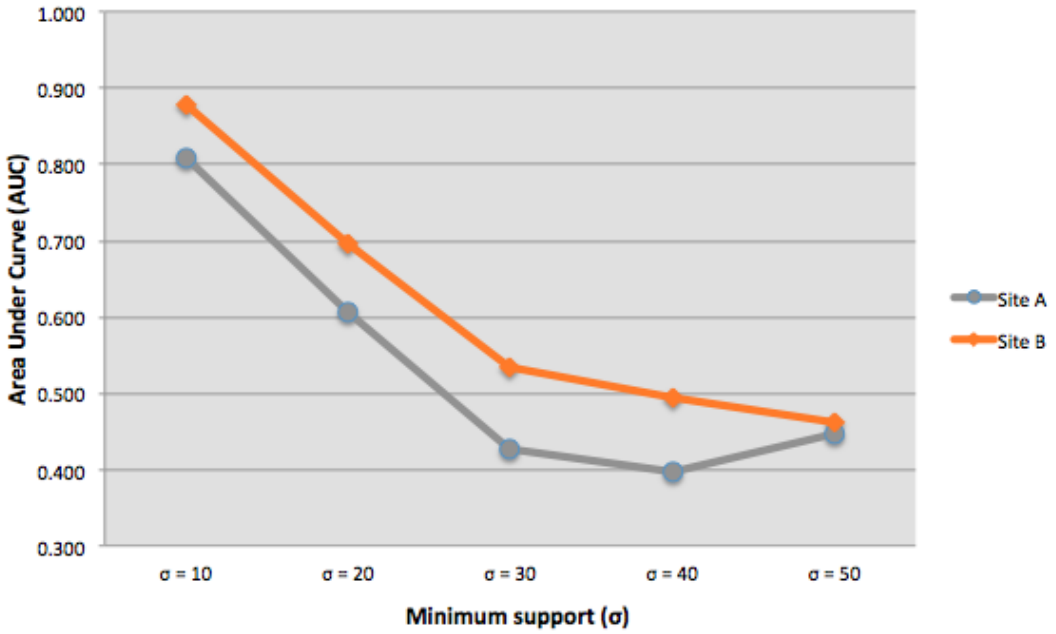


Figure 4.6: Classification performance in terms of AUC, with respect to the Site A and B data sets, over a range of σ values

subgraphs were generated, greater than $k = 55$ consequently the feature selection process was able to retain the most distinguishing subgraphs and hence better classifiers were produced.

To further evaluate the results obtained the Friedman test was applied to the AUC results obtained using all five classifiers and a range of σ values with respect to the Site A and Site B data sets as shown in Table 4.4. With reference to the table the number in parenthesises in columns two and three indicate the overall "ranking" of each individual result with respect to the two data sets. The Average Rank (AR) is given in the fourth column, this is the mean value of the rankings for each classification technique. The Friedman test statistic is based on the AR values, as presented in Section 2.6 in Chapter 2. For reference Equation 2.6 from Section 2.6 is presented again in Equation 4.1, where N is the number of data sets (two in this case) and K is the total number of classification technique considered (five in this case). The highest recorded AUC value for each data set is indicated in bold font in Table 4.3. Inspection of the table indicates that $\sigma = 10$ produced the overall best performance (AR = 1.0), while $\sigma = 40$ produced the worst overall result (AR = 4.5).

$$\chi_F^2 = \frac{12N}{K(K+1)} \left[\sum_{i=1}^{K} AR_i^2 - \frac{K(K+1)^2}{4} \right] \tag{4.1}$$

The Friedman test statistic and corresponding p value are presented in the first row of Table 4.4. The Friedman test statistic (6.80) and the significance threshold ($p < 0.1$) indicate that the null hypothesis ($H0$), that there is no difference between the techniques, can be rejected. Because $H0$ could be rejected a post hoc Nemanyi test was applied to evaluate the relative performance of the approaches. Recall from Section 2.6 that when using the post hoc Nemanyi test the performance between individual approaches can be said to be significantly different if their AR values are different by more than some Critical Difference (CD) value calculated using equation 4.2 (Equation 2.8 from Section 2.6) where the critical difference level $\alpha = 0.1$ and the value $q_{\alpha,\infty,K}$ is based on the Studentised range statistic.

Table 4.4: Friedman statistical test rankings with respect to Objective 1

Friedman test statistic = 6.80 ($p < 0.1$)			
MinSup	Site A	Site B	AR
$\sigma = 10$	0.808 (1)	0.879 (1)	1.0
$\sigma = 20$	0.606 (2)	0.697 (2)	2.0
$\sigma = 30$	0.427 (4)	0.535 (3)	3.5
$\sigma = 40$	0.396 (5)	0.495 (4)	4.5
$\sigma = 50$	0.446 (3)	0.461 (5)	4.0

$$CD = q_{\alpha,\infty,K} \sqrt{\frac{K(K+1)}{12N}} \tag{4.2}$$

The associated significance diagram is presented in Figure 4.7. In the diagram the classification techniques are listed in ascending order of ranked performance along the Y-axis; the average rank is given along the X-axis. From the figure it can be seen that the operation of the

Table 4.5: The Classification performance using different feature selection algorithms

Algorithms	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
Chi – Squared	0.629	0.708	0.626	0.629	**0.770**	0.720	0.862	0.699	0.720	0.820
GainRatio	0.600	**0.808**	0.596	0.596	0.734	**0.800**	**0.879**	**0.792**	**0.800**	**0.876**
InformationGain	**0.643**	0.797	**0.641**	**0.643**	0.743	0.740	0.871	0.735	0.740	0.846

$\sigma = 10$ approach was significant different from the $\sigma = 40$ and $\sigma = 50$ approaches; the critical difference tail for $\sigma = 10$ does not overlap with the tails for $\sigma = 40$ and $\sigma = 50$. However, there is no statistically significant difference between the $\sigma = 10$ approach and the $\sigma = 20$ and $\sigma = 30$ approaches because their CD tails overlap.

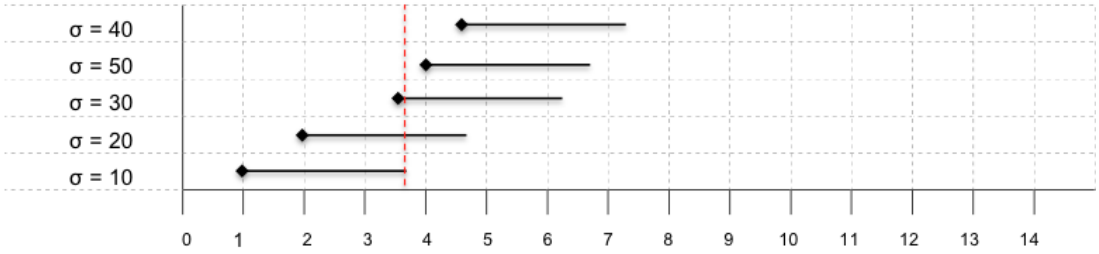


Figure 4.7: Nemenyi's post hoc critical difference diagram ($\alpha = 0.1$) for Objective 1

4.6.2 Feature Selection

This sub-section reports on the outcomes from experiments conducted to compare, in the context of classification effectiveness, the operation of the three different feature selection techniques considered: (i) Chi-Squared, (ii) Gain Ratio and (iii) Information Gain. For the experiments $\sigma = 10$ was used, because this produced the best result with respect to the experiments reported in Sub-section 4.6.1, together with $k = 55$. The obtained results are presented in Table 4.5. From the table it can be observed that best results were obtained using Gain Ratio feature selection for both Site A and Site B; giving AUC values of 0.808 and 0.879 respectively. In contrast, the Chi-Squared feature selection algorithm produced the worst results for both sites; giving AUC values of 0.708 and 0.862 respectively.

The AUC results are presented in the form of a bar chart in Figure 4.8 where the X-axis represents the three feature selection algorithms: (i) Chi-Squared, (ii) Gain Ratio and (iii) Information Gain, and the Y-axis the associated AUCs score. From the graph it can be clearly seen that: (i) the AUC values obtained for Site B were higher than the AUC values obtained for Site A, and (ii) similar results were obtained for all three feature selection algorithms. The maximum AUC value was recorded using the Gain Ratio feature selection algorithm. Information Gain was the second best performing feature selection technique and Chi-Squared feature

selection the third. The reason why the AUC values recorded with respect to the Site A data were consistently worse than the Site B data was probably because the number of feature generated from Site A was greater than that for Site B (757 versus 420), thus less complexity may produced the better performance.

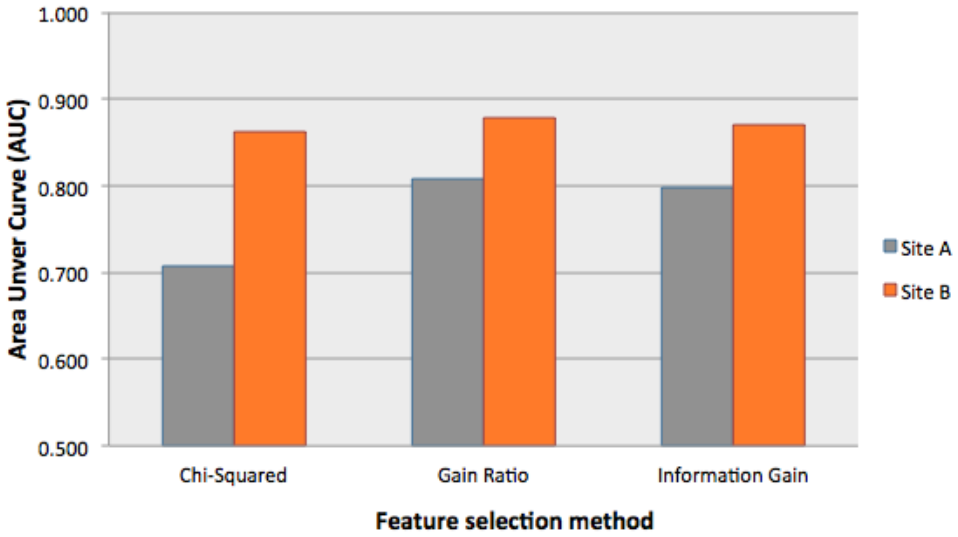


Figure 4.8: Classification performance in terms of AUC, with respect to the Site A and B data set using the three considered feature selection techniques

A Friedman test was again applied so as to determine whether the results were statistically significant or not; the associated data is given in Table 4.6. As before the overall rankings with respect to the two data sets are given in parenthesises in columns two and three. The AR of each classifier is given in the fourth column. The associated Friedman test statistic was 4.00 and the corresponding p value was $p > 0.1$; the null hypothesis, $H0$, that there is no statistical difference between the techniques, can thus be accepted. Therefore from the experiments and the Friedman test conducted it can be concluded that there was no statistical significant difference between the techniques. Thus no post hoc Nemanyi test was conducted. However it can also be concluded that Gain Ratio feature selection was the most appropriate feature selection mechanism in the context of the population estimation mining using graph-based representation considered in this chapter because it did give the best AR as presented in Table 4.6.

Table 4.6: Friedman statistical test rankings with respect to Objective 2

Friedman test statistic = 4.00 ($p > 0.1$)			
Algorithms	Site A	Site B	AR
$Chi-Squared$	0.708 (3)	0.862 (3)	3.0
$GainRatio$	0.808 (1)	0.879 (1)	1.0
$InformationGain$	0.797 (2)	0.871 (2)	2.0

Table 4.7: Classification performance over a range of values of k for Gain Ratio feature selection

Number of k	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
$k = 25$	**0.657**	0.795	**0.651**	**0.657**	0.765	0.740	0.830	0.727	0.740	0.848
$k = 30$	**0.657**	0.791	0.650	**0.657**	**0.767**	0.740	0.838	0.716	0.740	0.841
$k = 35$	0.557	0.751	0.556	0.557	0.700	0.740	0.863	0.723	0.740	0.848
$k = 40$	0.543	0.739	0.542	0.543	0.695	0.760	0.865	0.745	0.760	0.851
$k = 45$	0.600	0.748	0.598	0.600	0.739	0.780	**0.886**	0.764	0.780	0.873
$k = 50$	0.614	0.788	0.614	0.614	0.736	**0.820**	0.885	**0.816**	**0.820**	**0.892**
$k = 55$	0.600	**0.808**	0.596	0.600	0.734	0.800	0.879	0.792	0.800	0.876
$k = 60$	0.571	0.785	0.570	0.571	0.720	0.780	0.883	0.773	0.780	0.864

4.6.3 Number of attributes

From the foregoing Gain Ratio feature selection produced the most appropriate classification performance. However, this selection technique requires a parameter k. To identify the effect on classification performance of the value of k a sequence of experiments was conducted using a range of values for k from 25 to 60 incrementing in steps of 5. For the experiments $\sigma = 10$ was used because previous experiments, reported in Sub-section 4.6.1, had indicated that a value of $\sigma = 10$ produced the best performance. The Bayesian Network learning method was again adopted. The results produced are presented in Table 4.7. From the Table it can be seen that with respect to the Site A data the best results (AUC= 0.808) were obtained using $k = 55$, and with respect to the Site B data, the best results (AUC= 0.886) were obtained using $k = 45$. Thus indicating that a value of $k = 50$ (between the two) would be a good start value. The same data as in Table 4.7 is shown in graph form in Figure 4.9. In the figure the X-axis represents the k values and the Y-axis the AUC score.

The Friedman statistical test was again applied using the AUCs obtained with respect to all eights classifiers produced using different k values. The rankings are given in Table 4.8, columns two and three. The AR of each classifier is given in the fourth column. The Friedman test statistic was 4.50 and the significance threshold was $p > 0.1$. Therefore we can conclude that there is no significant different between techniques. The null hypothesis was therefore accepted and no post hoc Nemenyi test conducted. However, from the test results presented above it was concluded that $k = 55$ was the most appropriate value for k when using the Gain Ratio feature selection method in the context of the graph-based representation for population estimation mining presented in the chapter because it did give the best AR as shown in Table 4.8.

4.6.4 Classification Generation Method

To determine the most appropriate classification method eight different algorithms were considered: (i) Decision Tree (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators

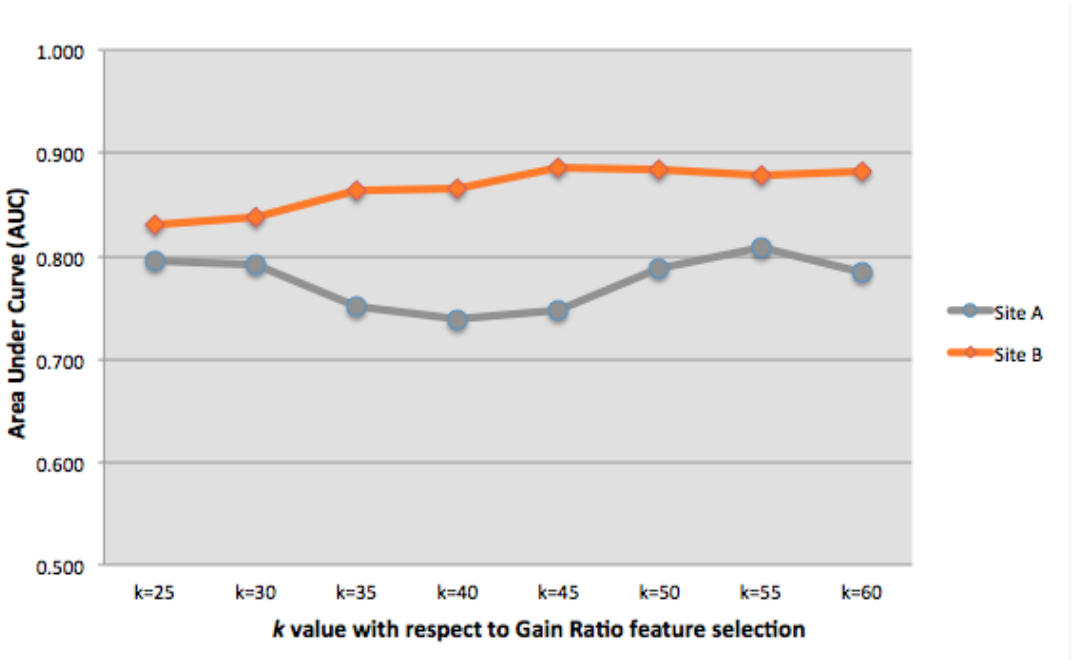


Figure 4.9: Classification performance in terms of AUC with respect to the Site A and B data sets, over a range of Gain Ratio feature selection k values

Table 4.8: Friedman statistical test rankings with respect to Objective 3

Friedman test statistic = 4.5 ($p > 0.1$)			
Number of k	Site A	Site B	AR
$k = 25$	0.795 (2)	0.830 (8)	5.0
$k = 30$	0.791 (3)	0.838 (7)	5.0
$k = 35$	0.751 (6)	0.863 (6)	6.0
$k = 40$	0.739 (8)	0.865 (5)	6.5
$k = 45$	0.748 (7)	0.886 (1)	4.0
$k = 50$	0.788 (4)	0.885 (2)	3.0
$k = 55$	0.808 (1)	0.879 (4)	2.5
$k = 60$	0.785 (5)	0.883 (3)	4.0

Table 4.9: Classification performance in terms of a a number of different classifier generators

Generator	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
*C*4.5	0.514	0.572	0.515	0.514	0.686	0.500	0.596	0.500	0.500	0.715
NaiveBayes	0.571	0.794	0.569	0.571	0.727	0.740	0.870	0.728	0.740	0.839
AODE	0.629	0.815	0.627	0.629	0.753	**0.800**	0.863	0.785	**0.800**	0.871
BayesianNetwork	0.600	0.808	0.596	0.600	0.734	**0.800**	**0.879**	**0.792**	**0.800**	**0.876**
RBFNetwork	0.614	0.728	0.615	0.614	0.746	0.660	0.796	0.660	0.660	0.820
SMO	**0.729**	0.791	**0.727**	**0.729**	**0.818**	0.620	0.733	0.610	0.620	0.781
LogisticRegression	0.671	0.810	0.672	0.671	0.789	0.560	0.729	0.560	0.560	0.771
NeuralNetwork	0.686	**0.819**	0.685	0.686	0.782	0.620	0.789	0.628	0.620	0.829

(AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network. For the experiments $\sigma = 10$ was used because this produced the best result with respect to the experiments reported in Sub-section 4.6.1. The Gain Ratio feature selection algorithm, together with $k = 55$, was adopted because experiments reported in Sub-sections 4.6.2 and 4.6.3 indicated that this produced the best result. The obtained results are presented in Table 4.9. From the Table it can be observed that with respect to the Site A data, the best results (AUC = 0.819) were obtained using the Neural Network classifier, and with respect to the Site B data, the best results (AUC = 0.879) were obtained using the Bayesian Network classifier. The C4.5 classifiers did not perform well for both Sites A and B. The Site B produced the better classification performance for five of the eight classifiers; this is a surprising result with no clear reason as to why. It was anticipated that the Site A data would (in general) produce better results because the data was produced from images obtained during the wet season which consequently featured a greater contrast between foreground and background. However, from the experiments it is clear that the graph-based representation is better at "coping" with the dry season data.

Figure 4.6.4 gives the same results from Table 4.9 in the form of a bar chart. The eight learning approaches are listed along the X-axis, and the associated AUC values along the Y-axis. The bar chart confirms the results noted above: (i) the AUCs for Site B were higher than the AUCs for Site A with respect to the C4.5, Naive Bayes, AODE, Bayesian Network and RBF Network classification models; (ii) in contrast the AUCs for Site A were higher than the AUCs for Site B in with respect to the SMO, Logistic Regression and Neural Network classification models; (iii) the lowest performance was recorded when using C4.5 for both sites; and (iv) the highest performance was obtained using Neural Network with respect to Site A and Bayesian Network with respect to Site B. Again it is interesting to note that there seems to be some distinction between the data for the two sites that leads to different groups of classifiers being more appropriate.

The Friedman statistical test was again applied to determine whether the distinctions in recorded classification performance, in terms of AUC, for all eight classifiers was indeed significant. The calculated rankings are presented in Table 4.10 (columns two and three), the AR

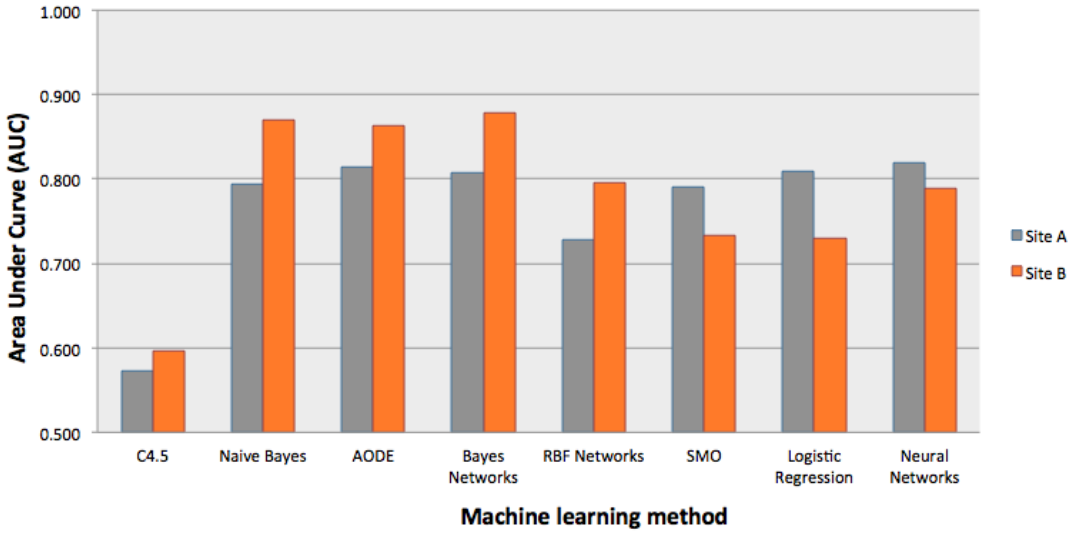


Figure 4.10: Bar graph representing the results of classification performance in term of AUCs with respect to the Site A and B using different classification generation algorithms

for each classifier is again given in column four. In this case the Friedman test statistic was 9.00 and the significance threshold was $p > 0.1$ indicating that the null hypothesis $H0$ could again be accepted, there are no statistically significant differences between the techniques. However, from the experiments conducted it can be concluded that Bayesian Network was the most appropriate classification technique in the context of the population estimation mining using the graph-based representation consider in this chapter because it produced the best AR performance as demonstrated in Table 4.10.

Table 4.10: Friedman statistical test rankings with respect to Objective 4

Friedman test statistic = 9.00 ($p > 0.1$)			
MinSup	Site A	Site B	AR
C4.5	0.572 (8)	0.596 (8)	8.0
NaiveBayes	0.794 (5)	0.870 (2)	3.5
AODE	0.815 (2)	0.863 (3)	2.5
BayesianNetwork	0.808 (4)	0.879 (1)	2.5
RBFNetwork	0.728 (7)	0.796 (4)	5.5
SMO	0.791 (6)	0.733 (6)	6.0
LogisticRegression	0.810 (3)	0.729 (7)	5.0
NeuralNetwork	0.819 (1)	0.789 (5)	3.0

4.7 Discussion

The overall classification results presented in the previous section, Section 4.6, indicated that the proposed graph-based approach, using a graph/tree representation to which frequent subgraph mining was applied, performed well over the two different satellite images datasets considered. The main findings from the four sets of experiments conducted were:

1. Classification effectiveness trended to improve as the support threshold decreased because more frequent subgraphs were identified. The reported evaluation found that $\sigma = 10$ produced the best performance.

2. The best classification performance in terms of feature selection mechanism, for both data sets (Site A and B), was obtained using Gain Ratio, follow by Information Gain, and then Chi-Squared feature selection.

3. With respect to Gain Ratio feature selection it was found that $k = 55$ produced the overall best performance.

4. The most appropriate classification generation mechanisms identified from the reported evaluation were: (i) Bayesian Network and (ii) AODE. However the average AUCs using Bayesian Network and AODE for both sites (Site A and B) was found to be 0.844 and 0.839 respectively, thus the Bayesian Network classifier produced a slightly better overall performance than the AODE classifier.

4.8 Summary

In this chapter the first of the proposed approaches to population estimation mining from satellite imagery has been presented. The proposed approach is based on a graph representation, and used a hierarchical decomposition whereby each individual household image was decomposed into a quadtree hierarchical structure by recursively partitioning the image space into quadrants. Each household image was represented using a single quadtree. A frequent subgraph mining approach was then applied so that subgraphs that frequently occur across the image set could be identified. The set of identified frequent subgraphs was then transformed into a feature vector space. A feature selection approach was applied to the feature vectors and the most discriminative features (subgraphs) selected (to which a number of classification learning methods may be applied). The reported evaluation indicated that high classification accuracy results were obtained when using a low support threshold (σ). The Gain Ratio feature selection mechanism was found to be the most appropriate feature selection mechanism with respect to both data sets (Site A and Site B). The most appropriate k value, with respect to the Gain Ratio feature selection mechanism, was found to be $k = 55$. The most suitable classification generator was found to be the Bayesian Network model. In the following chapter an alternative

approach for classifying satellite images using colour histograms and colour based statistical features is described.

Chapter 5

Population Estimation Mining using Satellite Imagery: The Colour Histogram Based Approach

5.1 Introduction

The proposed image colour based approach to population estimation mining using satellite imagery is presented in this chapter. Recall that the application of classification techniques to image data requires that the image data set under consideration is represented in a manner that captures the salient features of the data but at the same time is compatible with the classification techniques to be used. In the previous chapter a graph-based approach was suggested, in this chapter an alternative mechanism founded on the usage of image colour is proposed. The colours within an image are its most basic content; representing images in terms of this content is thus an obvious idea. One method of encapsulating image colour is to represent the distribution of colours within a given image using histograms [44, 118]. Thus, in the context of our segmented household data, the idea is to represent each household in terms of a collection of colour histograms, seven histograms per household. Of course, for classifier training purposes each collection of histograms will have a family size (class) label associated with it. In addition, the use of simple statistical information concerning the distribution of colour across an image was also considered.

A schematic of the proposed colour histogram representation approach for population estimation mining is given in Figure 5.1. From the figure it can be seen that the overall approach encompasses three processes: (a) image segmentation, (b) feature extraction and (c) classifier generation. The image segmentation process (the top rectangular box in Figure 5.1) is the same as that discussed with respect to the graph-based approach described in Chapter 4; the image segmentation process was detail in Chapter 3 and is thus not discussed further here. The feature extraction process is concerned with translating the segmented data into a form ready for

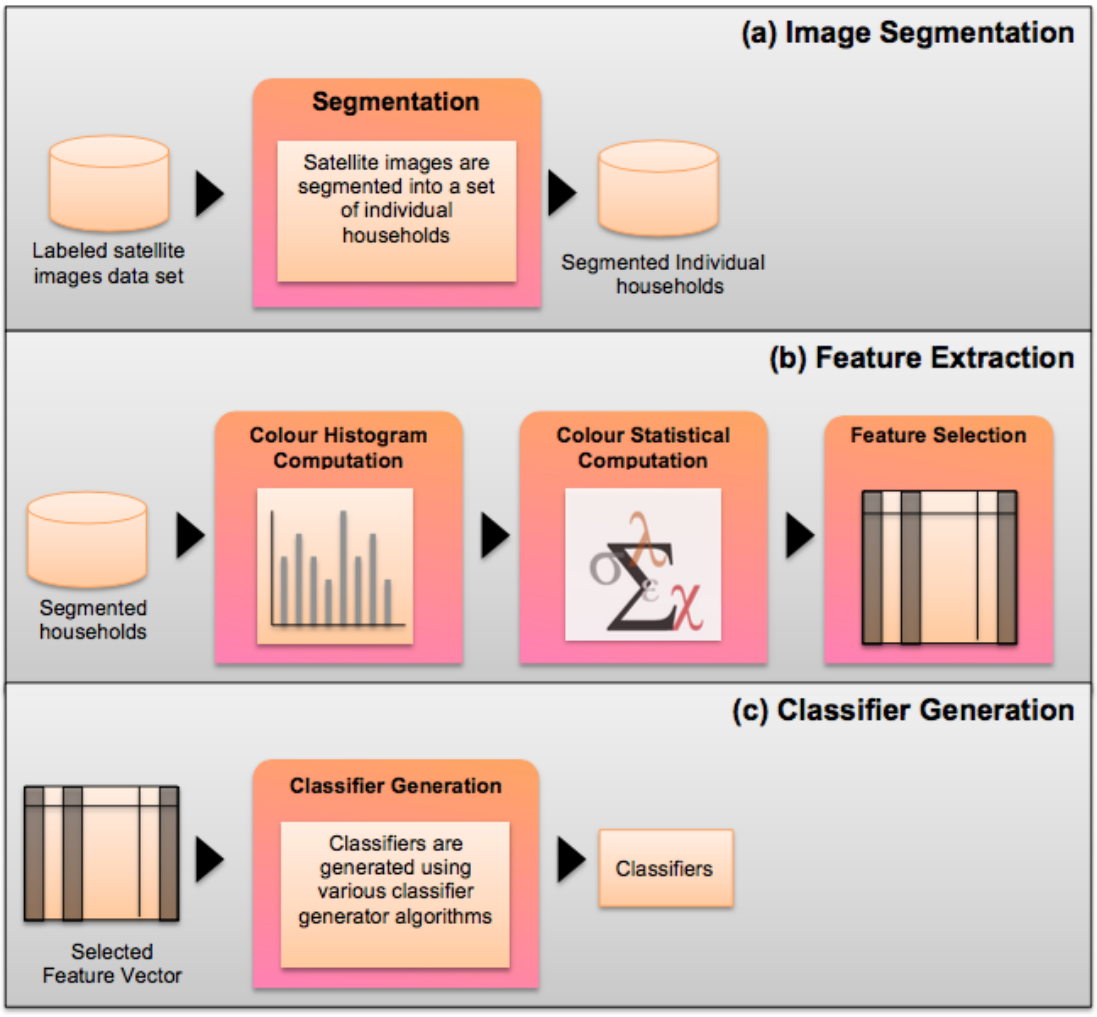


Figure 5.1: Schematic illustrating the population estimation mining approach using colour histograms

classification, the colour histogram based representation in the case of the work presented in this chapter. From the figure it can also be seen that the feature extraction process comprises three steps: (i) colour histogram computation, (ii) the identification of a number colour based statistical metrics to be included in the representation, and (iii) reduction of the feature vector dimensions using a feature selection mechanism. The third process included in Figure 5.1 is classifier generation. This is where standard classifier generation methods can be applied to build the desired classifier which can then be applied to unseen data.

The rest of this chapter is organised as follows. Section 5.2 provide detail of the proposed colour histogram representation. Section 5.3 discussed the calculation of the additional colour statistical metrics used to augment the basic colour histogram representation. The feature selection and classifier generation process is then considered in Section 5.4. Section 5.5 reports on the evaluation of the proposed approach, followed by some discussion in Section 5.6. Finally, the main findings and some associated conclusions are presented in Section 5.7

5.2 Colour Histogram Generation

The histogram is a simple mechanism for representing image content and is wildly used in computer vision and pattern recognition [70, 101]. A colour histogram serves as an effective representation of the distribution of colour within a given image. In addition the usage of colour histograms offers the following advantages: (i) they provide a useful foundation for measuring the similarity between images due to their robustness to background noise and object distortion, (ii) they have rotation and translation invariant properties, (iii) their simplicity, and as a consequence (iv) efficiency. Although colour histograms are wildly applied for many applications their main drawback is that all relative spatial information between different colour regions is lost [105, 182].

Any pixel of an image can be described in terms of the composition components of a given colour space; for example the red, green and blue components of the RGB colour space. A colour histogram can be defined for each component by counting the number of pixels for each quantised bin. For example, a RGB image can be expressed in terms of three colour histograms, each histogram representing the colour distribution for each individual colour channel (red, green and blue) [117].

With respect to the proposed population estimation mining approach using colour histograms three colour spaces were used: (i) red, green and blue (RGB), (ii) hue, saturation and value (HSV), and (iii) greyscale. Once a given satellite image was transformed into the selected colour spaces, the number bins used for the quantisation of the colour space was 32. Each colour channel is a 8-bit colour format making a palette of 256 (2^8) entries; for histograms this size is too big therefore 32 (2^5) bins (a range of 8 colours for each bin) was selected. The X-axis of each histogram thus comprises a list of bins each representing a colour range, whilst the Y-axis of each histogram represented the number of pixels falling into each bin.

For each preprocessed household satellite image seven different histograms were thus extracted: (i) three histograms from the RGB colour spaces (red, green, and blue), (ii) three histograms from the HSV colour spaces (hue, saturation, and value) and (iii) a intensity histogram using the greyscale colour space. Each of the seven histograms comprised 32 bins, giving 224 (7×32) features in total. Figure 5.2 shows seven example histograms produced using one of the identified household image used in the evaluation presented later in this chapter (Section 5.5).

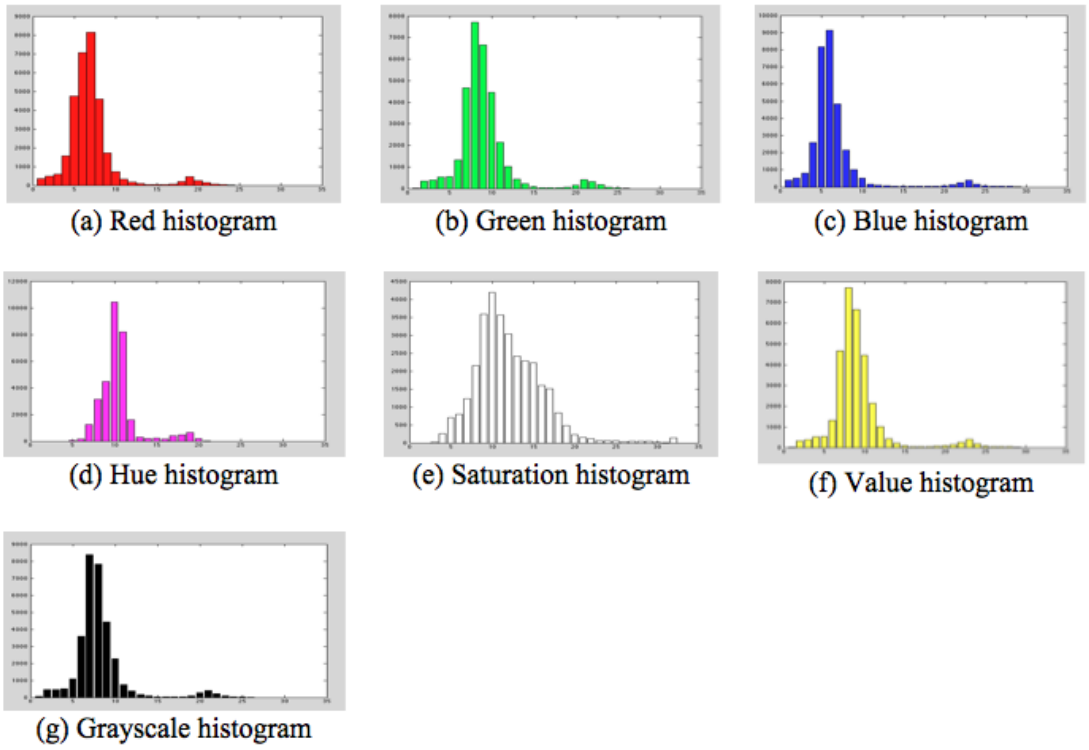


Figure 5.2: Example of the seven histogram representation for a segmented household image

5.3 Statistical Colour Metric Calculation

A simple alternative representation of the colour information in an image is to extract some simple statistical information from the image data concerning the colour distribution. The idea here was that this statistical information could be used to augment the colour histogram information (or used as a representation on its own). A total of 13 statistical features were identified: (i) 5 features describing the RGB colour channels, (ii) 5 features describing the HSV colour channels and (iii) 3 feature describing the greyscale channel. The individual features are listed in Table 5.1.

Thus on completion of the feature extraction process using both colour histograms and

Table 5.1: Additional colour based statistical features.

#	RGB colour space description	#	HSV colour space description	#	greyscale space description
1	Average red	6	Average hue	11	Average greyscale
2	Average green	7	Average saturation	12	Standard deviation of greyscale
3	Average blue	8	Average value		
4	Mean of RGB	9	Mean of HSV	13	Average of greyscale histogram
5	Standard deviation of RGB	10	Standard deviation of HSV		

the colour statistical metrics described in this section each household was represented using a feature vector of length 237 ($224 + 13 = 237$). These features were encapsulated using a feature space representation from which a collection of feature vectors could be generated, one per image.

5.4 Feature Selection and Classification

Once the 224 histograms and 13 statistical features have been identified, as described in Section 5.2 and Section 5.3 above, and before the classifier generation could be commenced, the data was first discretised (ranged) and a feature selection mechanism was applied to the feature vector space so as to reduce the overall number of dimensions so that only those features that served as good discriminators between classes were retained. As in the case of the graph-based approach described in the previous chapter three feature selection strategies were considered: (i) Chi-Squared, (ii) Gain Ratio and (iii) Information Gain.

Once the feature selection was complete the images were represented in terms of a set of feature vectors drawn from the reduced feature space. Classifier model generators could then be applied. Extensive evaluation was conducted so as to test the operation of the different parameters and their variations, however this chapter only reports the most significant results obtained (there is insufficient space to allow for the presentation of all the results obtained). Recall that with respect to the work presented in this theses eight classifier generation methods were used: (i) Decision Tree (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network. The implementations used were those available in the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench [186].

5.5 Evaluation

To evaluate the proposed colour histogram based approach the same labelled ('Small', 'Medium' and 'Large' family size) training data was used as used to evaluate the graph-based approach

described in the previous chapter. Data discretisation was again then applied to the selected features so that each continuously valued attribute was converted into a set of ranged attributes. Then the classification generator methods were applied. The overall aim of the evaluation was to provide evidence that census data can be effectively estimated using the proposed approach. To this end four sets of experiments were conducted as follows:

1. **Data Representation**: A set of experiments to determine the effect on classification performance using either histogram features only, statistical features only or a combination of the two (Sub-section 5.5.1).

2. **Feature Selection**: A set of experiments to compare the operation of the various suggested feature selection algorithms (Sub-sectionc 5.5.2).

3. **Number of attributes**: A set of experiments to analyse the effect that the number of selected attributes (k) had on performance (Sub-sectionc 5.5.3).

4. **Classification Generation Method**: A set of experiments to determine the effect on classification performance when using different classifier generation methods (Sub-section 5.5.4).

Each is discussed in further detail in Sub-sections 5.5.1 to 5.5.4 below. Ten fold Cross-Validation (TCV) was applied throughout, and performance was recorded in terms of: (i) accuracy (AC), (ii) area under the ROC curve (AUC), (iii) the F-Measure (FM), (iv) sensitivity (SN) and (v) specificity (SP). However, it should be recalled that in this thesis AUC is considered to be the most significant metric. Thus the discussion concerning the evaluation presented in this section is very much focussed in the AUC results obtained.

5.5.1 Data Representation

In order to investigate the effect of the three different forms of colour representation on classification performance the segmented household data was translated using each of the three forms to give three distinct feature spaces: (i) the Colour Histogram (CH) feature space comprised of the 224 dimensions each representing a histogram bin associated with one of the seven identified histogram representations described in Section 5.2, (ii) the Colour Statistical (CS) feature space comprised of 13 general colour statistical features (as described in Section 5.3), and (iii) the Combination of the two ($CH+CS$) hence comprising 237 ($224+13$) features. Because the size of the CH and $CH+CS$ feature spaces was significant, the Gain Ratio feature selection algorithm was applied to select the top k features together with $k = 25$ because experiments reported later in Sub-sections 5.5.2 and 5.5.3 had revealed that this was the most appropriate feature selection algorithm and the most appropriate value for k. The Logistic Regression classifier generation method was applied with respect to each of the feature spaces (feature vector data sets) because the experiments reported in Sub-section 5.5.4 had indicated that this tended to produce the most effective performance.

Table 5.2: Classification performance using different colour feat ure space representations (CH, CS and $CH+CS$)

Representations	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
CH	**0.657**	**0.822**	**0.662**	**0.657**	**0.806**	**0.640**	**0.821**	**0.633**	**0.640**	**0.798**
CS	0.329	0.522	0.339	0.329	0.617	0.420	0.573	0.418	0.420	0.685
$CH+CS$	**0.657**	**0.822**	**0.662**	**0.657**	**0.806**	0.600	0.719	0.588	0.600	0.747

The obtained results are presented in Table 5.2 (best values are shown in bold). From the table it can be seen that with respect to Site A (wet season) the CH and combined $CH+CS$ representations gave the best results in term of AUC (AUC $= 0.822$), in both cases the same results were produced indicating that the addition of the CS values had no effect.

With respect to the Site B (dry season) data set, the best performance was obtained using the CH representation (AUC $= 0.821$). The CS representation on its own did not work well. Similarly it can be argued that, overall, the CS representation did not serve to enhance the performance of the CH representation when it was combined with this representation.

Figure 5.3 shows the classification performance results given in Table 5.2 in the form of a bar graph. The horizontal axis represents the three representations: (i) Colour Histogram (CH), (ii) Colour Statistics (CS) and (iii) the Combination of the two ($CH+CS$), and the vertical axis represents the AUC score. From the graph it is clear that: (i) the CH representation gave the best results for both the Site A and Site B data sets, (ii) the worst results were produced using the CS representation (for both Site A and Site B), and (iii) with respect to the $CH+CS$ representation the performance with respect to the Site A data set was better than that recorded for the Site B data set.

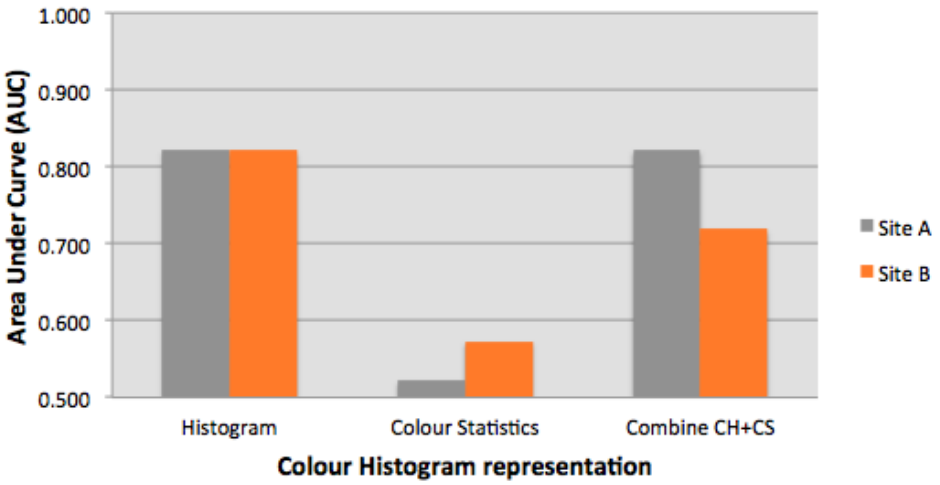


Figure 5.3: Bar graph showing classification performance in terms of AUC using different colour feature space representations (CH, CS and $CH+CS$)

A Friedman test was applied to the results in order to determine whether the results were statistically significant or not [50], the AUC results using all three colour representations with respect to the Site A and Site B data are listed in Table 5.3 (columns two and three). In each case the number in parentheses indicates the overall relative ranking of each individual result with respect to the two data sets. The Average Rank (AR) of each classifier is given in the fourth column, which is the mean value of the rankings for each representation. Recall from Section 2.6 that the Friedman test statistic is based on the *AR* values, and is calculated using equation 5.1, where N is the number of data sets (2) in this case, K is the number of classification technique considered (3) in this case.

$$\chi_F^2 = \frac{12N}{K(K+1)} \left[\sum_{i=1}^{K} AR_i^2 - \frac{K(K+1)^2}{4} \right] \tag{5.1}$$

The Friedman test statistic and corresponding p value are presented in the first row of Table 5.3. The Friedman test statistic (3.714) and the significance threshold ($p < 0.1$) indicated that the null hypothesis, that there is no statistical difference between the techniques, can be rejected. From Table 5.3 it can be confirmed that the *CH* representation produced the overall best performance (AR =1.0), while the *CS* representation produced the worst overall result (AR = 3.0).

Table 5.3: AUC values recorded for each colour feature space representation (*CH*, *CS* and *CS*+*CH*)

Friedman test statistic = 3.714 ($p < 0.1$)			
Data Representation	Site A	Site B	AR
CH	0.822(1.5)	0.821(1)	1.25
CS	0.522(3)	0.573(3)	3.00
CH+*CS*	0.822(1.5)	0.719(2)	1.75

A post hoc Nemanyi test was applied to determine the distinction in performance between the individual representations [102]. Recall from Chapter 3 that the performance of two approaches is statistically different if their ARs are different by more than a Critical Difference (CD) value calculated using equation 5.2 where the critical difference level α ($\alpha = 0.1$) and the value $q_{\alpha,\infty,K}$ is based on the Studentised range statistic [86].

$$CD = q_{\alpha,\infty,K} \sqrt{\frac{K(K+1)}{12N}} \tag{5.2}$$

The significant diagram in Figure 5.4 shows the AUC performance rank of the proposed image classification technique with Nemenyi's critical difference tails (the calculated CD for the diagram is 1.45). The diagram shows the classification techniques listed in ascending order of ranked performance on the Y-axis, and the associated AR value across both data sets along the X-axis. From the diagram it can be seen that the results produced using the *CH* representation are significantly better than the *CS* results because there is "white space" between the

Table 5.4: Classification performance using three different feature selection methods

Algorithms	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
Chi − Squared	**0.671**	0.810	**0.672**	**0.671**	0.792	0.640	0.775	0.638	0.640	0.774
GainRatio	0.657	**0.822**	0.662	0.657	**0.806**	0.640	**0.821**	0.633	0.640	0.798
InformationGain	0.657	0.796	0.657	0.657	0.799	**0.680**	0.769	**0.677**	**0.680**	**0.810**

AR values for *CS* and the end of the critical difference tail for *CH*. There was no significant difference with respect to the *CH* and *CH* + *CS* representations because there critical difference tails overlap.

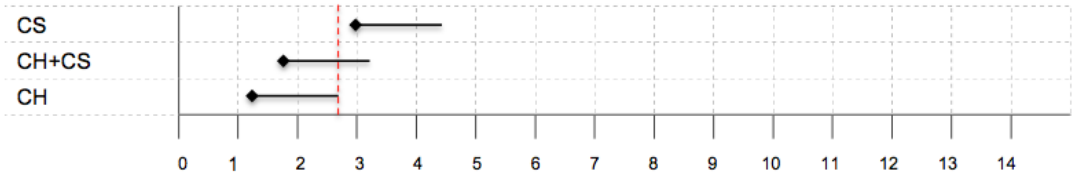


Figure 5.4: Nemenyi Significance Diagram for the different colour feature space representations ($\alpha = 0.1$)

Thus it can be concluded that colour histogram representation (*CH*) is the most appropriate representation in the context of the population estimation mining application considered in this chapter.

5.5.2 Feature Selection

Recall that three different algorithms for feature selection were considered in order to investigate the most appropriate feature selection mechanism with respect to the proposed approach. The three feature selection mechanism were: (i) Chi-Squared, (ii) Gain Ratio, and (iii) Information Gain. For the experiment used to compare the usage of these three methods the *CH* representation was used as this had been found to produce the best results as establishes in the previous Sub-section (Sub-section 5.5.1). A value of $k = 25$ was again used, together with the Logistic Regression learning method, for the same reasons as before (they produced the best results with respect to the experiments reported later in this chapter in Sub-sections 5.5.3 and 5.5.4, respectively). The results from the experiments are presented in Table 5.4. From the table it can be observed that the best results were obtained using Gain Ratio feature selection for both the Site A and the Site B data sets; giving AUC values of 0.822 and 0.821, respectively. Whereas, the Information Gain feature selection measure produced the worst results for both Sites; giving AUC values of 0.796 and 0.769, respectively.

The bar graph in Figure 5.5 illustrates the classification performances results in terms of recorded AUC value with respect to the Site A and Site B data sets using the different fea-

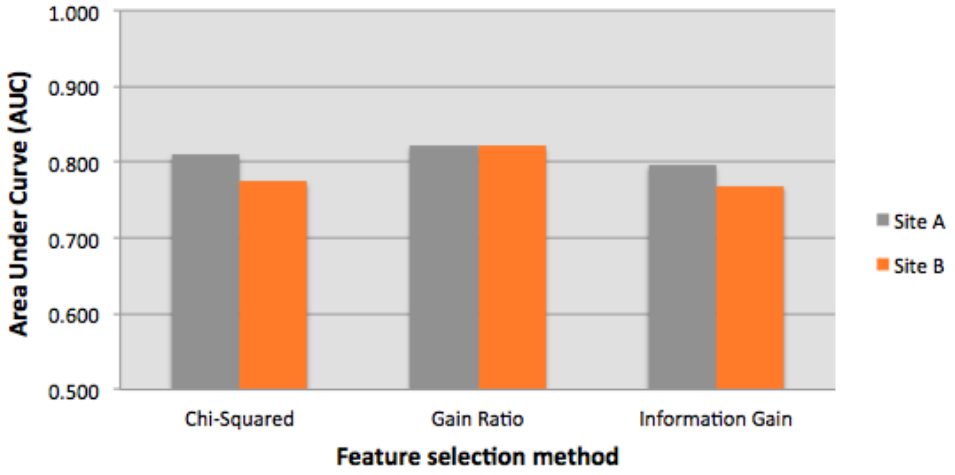


Figure 5.5: Bar graph showing classification performance in terms of AUCs using different feature selection methods

ture selection methods. The horizontal axis represents the three feature selection algorithms: (i) Chi-Squared, (ii) Gain Ratio and (iii) Information Gain. The vertical axis represents the AUC values. From the graph it can again be clearly seen that the maximum AUC values were recorded with respect to the Gain Ratio feature selection method for both sites. The AUCs recorded for the Site A data were higher than those recorded for the Site B data in all cases; this was probably because the *CH* representation with respect to the Site A (wet season) provided more distinguish colour information than for the Site B data set (dry season). Note that this is a contradictory results to that presented in Chapter 4.

Again a Friedman test was applied to determine whether there was any statistical difference in the operation associated with the usage of the three feature vector selection methods considered. The recorded AUCs using all three methods with respect to the Site A and the Site B data sets are presented in Table 5.5 (columns two and three). Again the number in parentheses in each case indicates the overall ranking of each individual method. As before the AR of each classifier is given in the fourth column.

Table 5.5: AUC values recorded for each feature selection method

Friedman test statistic = 4.00 ($p > 0.1$)			
Algorithm	Site A	Site B	AR
$Chi-Squared$	0.810(2)	0.775(2)	2.00
$GainRatio$	0.822(1)	0.821(1)	1.00
$InformationGain$	0.796(3)	0.769(3)	3.00

The Friedman test statistic and corresponding p value are presented in the first row of Table 5.5. The Friedman test statistic (4.00) and the significance threshold ($p > 0.1$) indicated

Table 5.6: Classification performance using different values for k with respect to Gain Ratio feature selection

Number of k	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
$k = 15$	0.657	0.814	0.648	0.657	0.768	0.480	0.726	0.463	0.480	0.691
$k = 20$	0.643	0.808	0.649	0.643	**0.809**	0.580	0.733	0.570	0.580	0.744
$k = 25$	0.657	0.822	0.662	0.657	0.806	0.640	**0.821**	0.633	0.640	0.798
$k = 30$	0.629	**0.826**	0.631	0.629	0.782	0.640	0.781	0.630	0.640	0.793
$k = 35$	**0.671**	0.814	**0.674**	**0.671**	0.806	0.620	0.778	0.602	0.620	0.778
$k = 40$	0.657	0.800	0.661	0.657	0.801	**0.660**	0.814	0.653	**0.660**	0.798
$k = 45$	0.586	0.739	0.590	0.586	0.756	0.640	0.781	0.635	0.640	0.791
$k = 50$	0.586	0.792	0.587	0.586	0.739	**0.660**	0.813	**0.660**	**0.660**	**0.817**

that the null hypothesis ($H0$) could be accepted, there was no statistical difference in operation between the techniques, thus no post hoc Nemenyi test was conducted. However, from the experiments conducted it can be concluded, despite the fact that the Friedman test had indicated that there was no statistically significant difference between the techniques, that Gain Ratio feature selection was the most appropriate feature selection method in the context of the population estimation mining using the colour histogram representation considered in this chapter because it did the best AR as presented in Table 5.5.

5.5.3 Number of attributes

In order to identify the effect on classification performance of the value of k, number of features to select with respect to the adopted Gain Ratio feature selection mechanism, a sequence of experiments was conduct using a range of values from $k = 15$ to $k = 50$ incrementing in steps of 5. For the experiments the colour histogram representation was again used because the experiments reported in Sub-section 5.5.1 had indicated that this produced the best performances. The Logistic Regression learning method was again adopted because the experiments reported later in Sub-section 5.5.4 had indicated that this tended to generate a best performance. The obtained results from the experiments are presented in Table 5.6. From the table it can be seen that $k = 30$ produced the best result with respect to the Site A data set with an AUC vale of 0.826. Whereas with respect to the Site B data set $k = 25$ produced the best results; best AUC value of 0.821. A value of $k = 50$ produced the worst results for both sites; AUC values of 0.582 and 0.781, respectively.

As before Figure 5.6 gives an accompanying bar graph with respect to the AUC data given in Table 5.6. The horizontal axis represents the k values and the vertical axis represents AUC values. The graph shows that, as indicated by the table, with respect to the Site A data set best performance was produced using $k = 30$ and that with respect to the Site B data set the best performance was produced using $k = 50$.

A Friedman test was again conducted. Table 5.7 gives the AUC values used for the Friedman test. The Friedman test statistic (6.578) and corresponding p value ($p > 0.1$), given in the

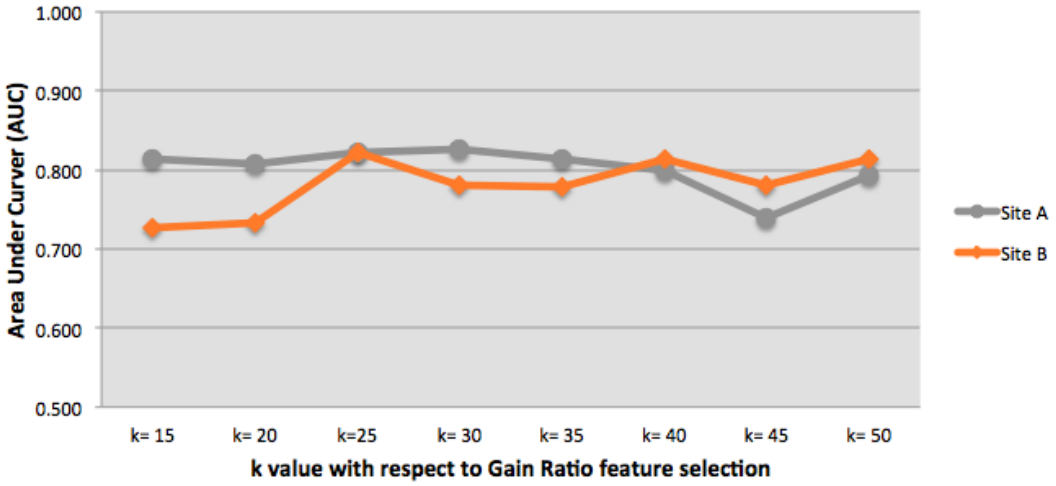


Figure 5.6: Bar graph showing classification performance using different values for k with respect to Gain Ratio feature selection

Table 5.7: AUC values recorded for each k value

Friedman test statistic = 6.578 ($p > 0.1$)			
kvalue	Site A	Site B	AR
$k = 15$	0.814(3.5)	0.726(8)	5.75
$k = 20$	0.808(5)	0.733(7)	6.00
$k = 25$	0.822(2)	0.821(1)	1.50
$k = 30$	0.826(1)	0.781(4.5)	2.75
$k = 35$	0.814(3.5)	0.778(6)	4.75
$k = 40$	0.800(6)	0.814(2)	4.00
$k = 45$	0.739(8)	0.781(4.5)	6.25
$k = 50$	0.792 (7)	0.813(3)	5.00

Table 5.8: Classifier performance using different classification models

Generator	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
*C*4.5	0.671	0.724	**0.668**	0.671	0.760	0.500	0.598	0.499	0.500	0.718
NaiveBayes	0.657	0.784	0.634	0.657	0.746	0.640	0.787	0.629	0.640	0.788
AODE	0.643	0.764	0.616	0.643	0.729	0.600	0.760	0.586	0.600	0.759
BayesianNetwork	**0.700**	0.807	0.687	**0.700**	0.782	**0.700**	0.798	**0.692**	**0.700**	**0.829**
RBFNetwork	0.557	0.705	0.566	0.557	0.761	0.580	0.676	0.574	0.580	0.759
SMO	0.629	0.733	0.627	0.629	0.780	0.600	0.734	0.592	0.600	0.762
LogisticRegression	0.657	**0.822**	0.662	0.657	**0.806**	0.640	**0.821**	0.633	0.640	0.798
NeuralNetwork	0.629	0.779	0.624	0.629	0.763	0.620	0.801	0.597	0.620	0.757

first row of the table, support the null hypothesis that there is no difference in operation between the techniques, and no post hoc Nemenyi test conducted. However, from the experiments conducted it can be concluded that with respect to Gain Ratio feature selection mechanism $k = 25$ is the most appropriate value for k in the context of the population estimation mining application considered in this chapter because it has the highest average ranking in the context of the Friedman test as shown in Table 5.7.

5.5.4 Classification Generation Method

Eight classifier generation methods were considered with respect to the experiments directed at determining the most appropriate classification method, these were the same as those considered with respect to the work presented in Chapter 4: (i) Decision Tree generators (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network. In each case the *CH* representation was used because this produced the best result with respect to the experiments reported earlier in Sub-section 5.5.1. The Gain Ratio feature selection method, with $k = 25$, was used because the experiments reported later in Sub-sections 5.5.2 and 5.5.3 had indicated that these produced the best overall results.

The results are presented in Table 5.8. According to the table it can clearly be observed that the Logistic Regression classifier generator produced the best results for both the Site A and the Site B data sets; giving AUC values of 0.822 and 0.821 respectively. The C4.5 decision tree generator produced substantially the worst performance for both sites (AUC values of 0.724 and 0.598).

Figure 5.7 shows the associated bar graph generated from the AUC data given in Table 5.8. The horizontal axis represents the eight learning approaches and the vertical axis the AUC scores. From the graph it can be observed that: (i) the classification performance with respect to Site A was better than for Site B when using C4.5, AODE, Bayesian Network, RBF Network and Logistic Regression, (ii) the classification performance with respect to Site B was better than for Site A when using Naive Bayes, SMO and Neural Network, (iii) Logistic Regression

produced the best outcomes for both Site A and Site B, and (iv) RBF Network produced the worst results for the Site A data set whereas C4.5 was produced the worst results for the Site B data set.

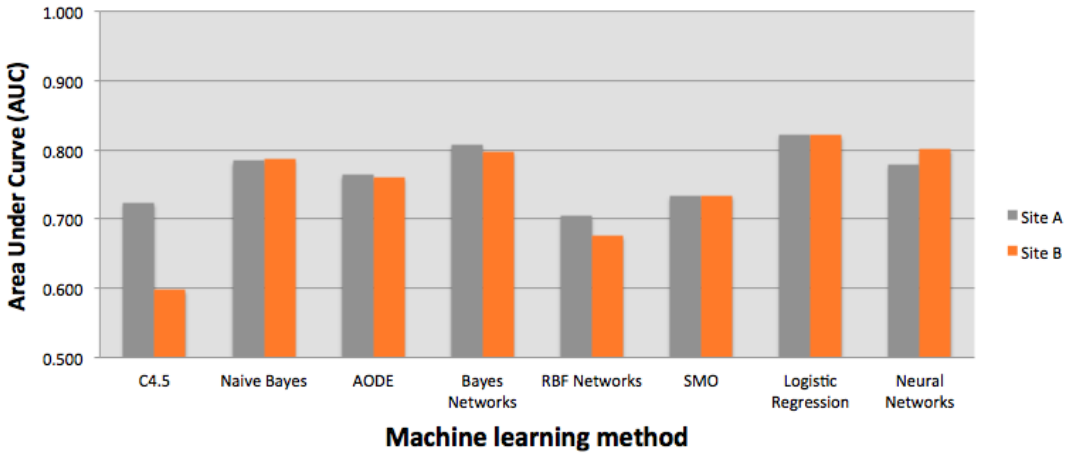


Figure 5.7: Bar graph showing classification performance using different classification models

As for the results reported with respect to the previous experiments the Friedman statistical test was again applied using the recorded AUC values as listed in Table 5.9. As before the number in parentheses in each case indicates the overall ranking of each individual result with respect to the two data sets. The AR of each classifier is given in the fourth column. The Friedman test statistic and corresponding p value are given in the first row of the table. The Friedman test statistic (13.33) and the significance threshold ($p < 0.05$) indicate that the null hypothesis can be rejected, thus there is a significant different between the techniques. A post hoc Nemanyi test was therefore conducted. The associated significant diagram is given in Figure 5.8; the calculated CD for the diagram is 4.82. The diagram demonstrates that both Logistic Regression and Bayesian Network classification models produced the best results (as already established) and that these results were statistically different from the C4.5 and RBF Network classification models.

Thus it can be concluded from Tables 5.8 and 5.9 and Figures 5.7 and 5.8 that Logistic Regression is the most appropriate classifier generator method in the context of the population estimation mining application considered in this chapter.

5.6 Discussion

The overall classification accuracies presented in the previous section, Section 5.5, indicate that the proposed population estimation mining, using a representation based on image colour

Table 5.9: AUC values recorded for each classification model

Friedman test statistic = 13.33 ($p < 0.05$)			
Generator	Site A	Site B	AR
*C*4.5	0.724(7)	0.598(8)	7.50
NaiveBayes	0.784(3)	0.787(4)	3.50
AODE	0.764(5)	0.760(5)	5.00
BayesianNetwork	0.807(2)	0.798(3)	2.50
RBFNetwork	0.705(8)	0.676(7)	7.50
SMO	0.733(6)	0.734(6)	6.00
LogisticRegression	0.822(1)	0.821(1)	1.00
NeuralNetwork	0.779(4)	0.801(2)	3.00

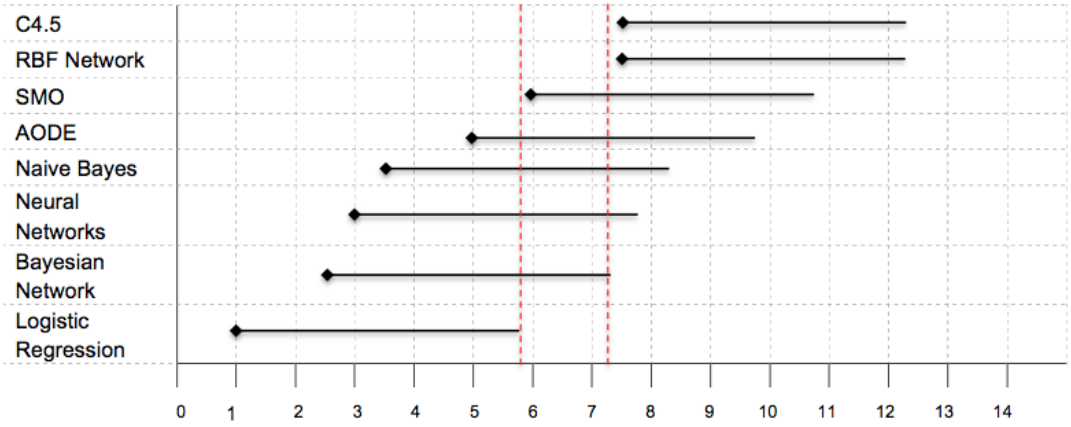


Figure 5.8: Nemenyi Significance Diagram for the different classification models

(colour histograms and some basic colour statistical information), performed well with respect to the two satellite images datasets considered (Site A and Site B). Four sets of experiments were conducted from which the main findings were as follows:

1. The most appropriate representation with respect to the application of population estimating mining for both data sets (Site A and Site B), in terms of best AUC, was the colour histogram representation (CH), followed by the combination of the colour histogram and the colour statistical representation ($CH + CS$), followed by the colour statistical (CS) representation.

2. The most appropriate feature selection mechanism with respect to both data sets (Site A and Site B) was Gain Ratio, followed by Chi-Squared, followed by Information Gain feature selection. Recall that Gain Ratio also produced the best performance with respect to the graph-based approach presented in the previous chapter.

3. With respect to the Gain Ratio feature selection approach the most appropriate value for k (number of attributes) was found to be $k = 25$. It is conjectured that lower values of k did not provide a good performance because there was insufficient data to build an effective classifier, while larger values of k resulted in overfitting.

4. A number of different classifier generator methods produced good results, however the most appropriate classification generation method from the reported evaluation was found to be the Logistic Regression mechanism (not the case with respect to the graph-based approach presented in the previous chapter).

5.7 Summary

A population estimation mining mechanism, using a representation based on image colour has been described. Three different colour representations were considered: (i) Colour Histogram (CH), (ii) Colour Statistics (CS) and (iii) a combination of the two ($CH + CS$). Experiments were conducted using the Ethiopian hinterland test data used previously. The reported evaluation indicate that high classification AUC results could be obtained when using the CH representation. The Gain Ratio feature selection mechanism was found to be the most appropriate feature selection method together with $k = 25$. Best classification results were obtained using the Logistic Regression classification model. In the following chapter an alternative approach for classifying satellite images that uses a representation founded on image texture, more specifically the usage of Local Binary Patterns, is described.

Chapter 6

Population Estimation Mining using Satellite Imagery: The Texture Based Approach

6.1 Introduction

In the previous two chapters the graph-based and colour histogram based approaches were proposed with respect to population estimation mining using satellite imagery. In this chapter an alternative mechanism founded on the usage of image texture is proposed. As noted in previous chapters, the application of classification techniques to image data requires that the image data set under consideration is represented in a manner whereby key information concerning the image content is retained while at the same time being conducive to classifier generation.

As noted in Sub-section 2.3.3 in Chapter 2, texture is an important feature with respect to both human and computer vision, one example where texture analysis has been usefully employed is with respect to pattern recognition [201]. There are three principle mechanisms that may be adopted to describe the texture in digital images: (i) statistical, (ii) structural and (iii) spectral. The statistical approach is concerned with capturing texture using quantitative measures such as "smooth", "coarse" and "grainy". Structural approaches describe image texture in terms of a set of texture primitives or elements (texels) that occur as regularly spaced or repeating patterns. In the spectral approach the image texture features are extracted by using the properties of (say) the Fourier spectrum domain so that "high-energy narrow peaks" in the spectrum can be identified [55].

One method of encapsulating image texture represented within a given image is by using Local Binary Patterns (LBPs) [44, 118]. A LBP is a texture representation method which is both statistical and structural in nature [141]. Using the LBP approach a binary number is produced for each pixel, by thresholding its value with its neighbouring pixels. LBPs offer the advantages of: (i) tolerance with respect to illumination changes, (ii) simplicity of computation and (iii) rotation invariance [106]. The LBP method has been used with respect to many applications, one example is face recognition [62, 200]. Thus the idea proposed in this chapter is to

represent each segmented household using a LBP representation. For training purposes each LBP represented household also had a family size class label associated with it. In addition the LBP information could be augmented with some statistical texture information.

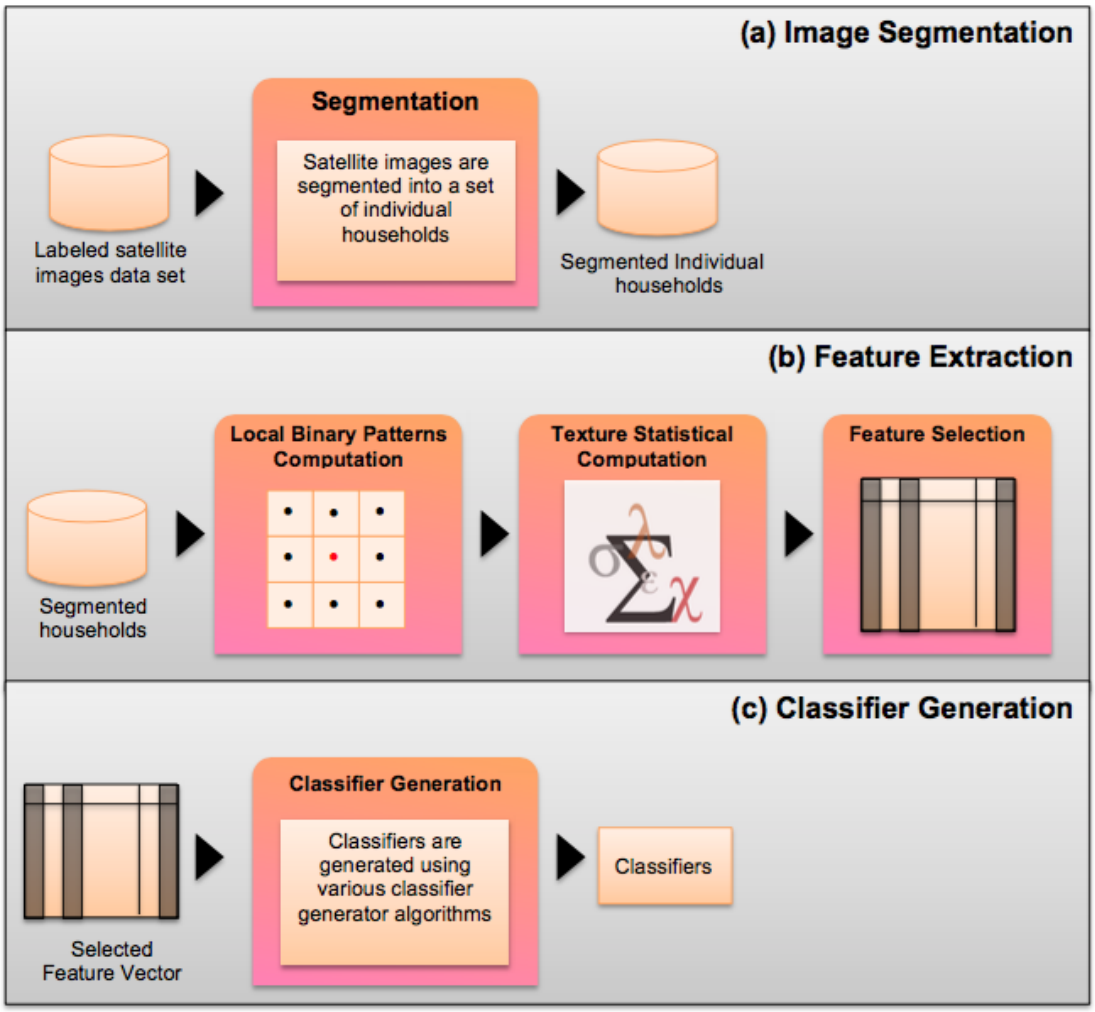


Figure 6.1: Schematic illustrating the population estimation mining approach using LBPs

A schematic of the proposed local binary pattern representation approach for population estimation mining is given in Figure 6.1. As in the case of the graph-based approach discussed in Chapter 4 and the colour histogram based approach discussed in Chapter 5, the texture based approach commences with the image segmentation process discussed in Chapter 3 and this is thus not further discussed in this chapter. The following two processes are feature extraction and classifier generation.

The feature extraction process is concerned with translating the segmented data into a form ready for classification, the texture based representation in the case of the work presented in this chapter. From Figure 6.1 it can be seen that the feature extraction process comprises three steps: (i) LBP computation, (ii) the determination of a number texture based statistical metrics

to be included in the representation, and (iii) prior to classifier generation a feature selection mechanism. The third process included in Figure 6.1 is classifier generation, where standard classifier generation methods can be applied to build the desired classifier which can then be applied to unseen data.

The rest of this chapter is organised as follows. Section 6.2 provides detail of the proposed LBP representation. Section 6.3 discusses the calculation of the additional statistical texture metrics used to augment the basic LBP representation. The feature selection and classifier generation process is then considered in Section 6.4. Section 6.5 reports on the evaluation of the proposed approach, followed by some discussion in Section 6.6. The statistical comparison of the proposed approach and the previous two approaches (detail in the previous two chapters) is presented in Section 6.7. Finally, the main findings and some associated conclusions are presented in Section 6.8

6.2 Local Binary Pattern

This Section presents an overview of the LBP image representation process. Recall that the adopted classification processes, presented in the following section, require the identified household images to be represented in a feature vector format. Therefore each household image needs to be translated into this format. The LBP concept was first introduce by Ojala [135] and, as already noted, has been widely used in the context of image texture analysis [169]. The fundamental idea is to define each pixel in an image according to its eight cardinal and sub-cardinal neighbours. In our case each household image was first converted into a greyscale image, and for each pixel its greyscale value was compared with its eight neighbouring greyscale values. If the neighbouring greyscale value was greater than the centre pixel greyscale value a '1' was recorded, otherwise a '0' was recorded. In this manner an eight digit number was defined describing each pixel according to its neighbours. The process is illustrated in Figure 6.2 [76].

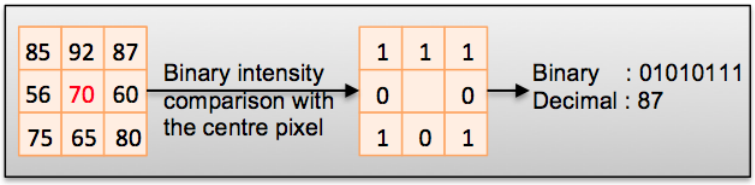


Figure 6.2: The LBP operator

Thus an LBP code is an ordered set of binary comparisons of greyscale values between the centre pixel of the window with its eight surrounding neighbourhoods. More formally the process can be expressed as shown in Equation 6.1, where i_c is the greyscale value of the centre pixel (x_c, y_c), i_n is the greyscale value of a neighbourhood pixel and $f(x)$ is defined according

to Equation 6.2.

$$LPB(x_c, y_c) = \sum_{n=0}^{7} f(i_n - i_c)2^n \tag{6.1}$$

$$f(x) = \begin{cases} 1, & if \quad x \geq 0 \\ 0, & if \quad x < 0 \end{cases} \tag{6.2}$$

In the example given in Figure 6.2 a 3×3 immediate neighbourhood is considered. However, variations of the basic LBP concept can be produced by: (i) using different *radii of neighbourhoods* (R) and/or (ii) different numbers of sampling points (P). The notation $LPB_{P,R}$ is often used to indicate different types of LBP. With respect to the work presented in this Chapter three different LBP variations were considered: (i) $LPB_{8,1}$, (ii) $LPB_{8,2}$ and (iii) $LPB_{8,3}$. All three variations are illustrated in Figure 6.3. In more detail: (i) $LPB_{8,1}$ equates to 8 sampling points within a radius of 1 (Figure 6.3(a)), (ii) $LPB_{8,2}$, equates to 8 sampling points within a radius of 2 (Figure 6.3(b)), and (iii) $LPB_{8,3}$, equates to 8 sampling points within a radius of 3 (Figure 6.3(c)).

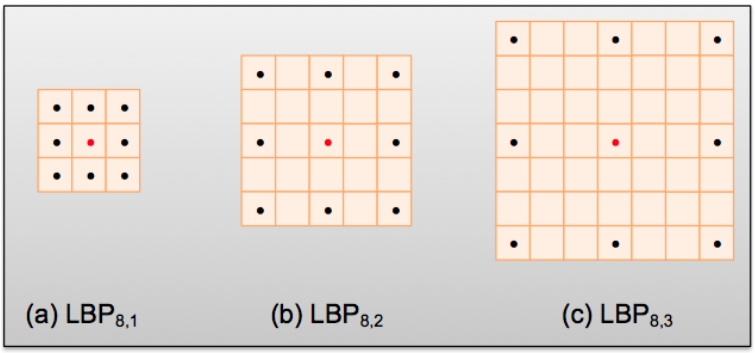


Figure 6.3: LBP variations

Using LBPs $2^8 = 256$ different texture patterns can be generated. This information can be captured in the form of a 256 element feature vector where each element holds an occurrence count for the associated LBP. An example of converting a household image into an LBP image is presented in Figure 6.4. Figure 6.4(a) shows the input image, Figure 6.4(b) the resulting LBP image.

6.3 Statistical Texture Metric Calculation

In addition, again with respect to the work presented in this chapter, each LBP representation describing a household image pixel could be augmented with a number of statistical features: (i) entropy features (E), (ii) Grey-Level Occurrence Matrix (GLOM) features (M) and (iii) wavelet transform features (W). Entropy is a statistical measure of randomness that can be

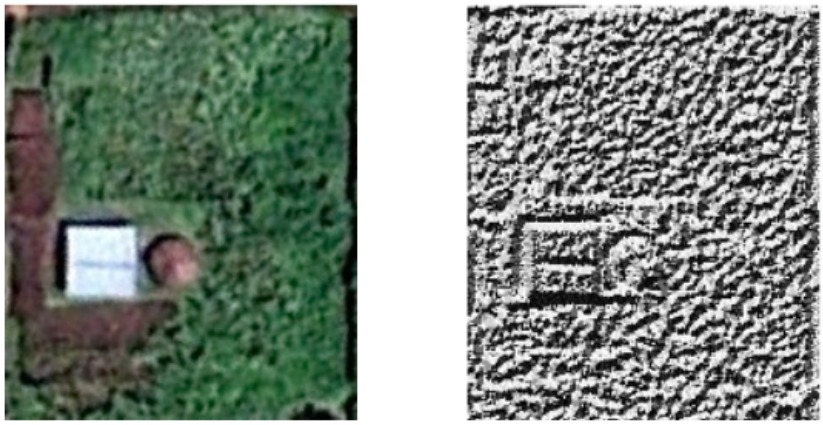


Figure 6.4: Example of LBP operator

used to characterise the texture of the given image, the entropy is defied using Equation 6.3:

$$E = -\sum(p \cdot log_2(p)) \tag{6.3}$$

where E is a scalar value representing the entropy of a greyscale image and p is the distribution of the pixel in the greyscale image.

A Grey-Level Co-Occurrence Matrix (GLCM) is a statistical tool for recording texture information that takes into consideration the spatial relationship between pixels. GLCMs are also known as grey-level spatial dependence matrices. GLCMs enable the characterisation of the texture of an image by calculating how often pairs of pixels with a specific greyscale value and a specific spatial relationship occur in an image. Once a GLCM has been created statistical measures can be extracted such as:

Contrast: A measure of the local variations in the GLCM.

Correlation: A measures of the joint probability occurrence of specified pixel pairs.

Energy: The sum of the squared elements in a GLCM, also known as the uniformity or the angular second moment.

Homogeneity: A measure of the closeness of the distribution of the elements in the GLCM to the GLCM diagonal.

The Discrete Wavelet Transform (DWT) is another technique used for texture analysis [22, 115]. By applying a DWT to a given household image a number of matrices result: the approximation coefficients matrix cA and the detailed coefficients matrices cH, cV, and cD

Table 6.1: Additional texture based statistical features.

#	Description	#	Description	#	Description
1	Entropy (E)	7	Average approximation	10	Average diagonal
2	Average Local Entropy (E)		coefficient matrix, *cA* (W)		coefficient matrix, *cD* (W)
3	Contrast (M)	8	Average horizontal		
4	Correlation (M)		coefficient matrix, *cH* (W)		
5	Energy (M)	9	Average vertical		
6	Homogeneity (M))		coefficient matrix, *cV* (W)		

(horizontal, vertical, and diagonal, respectively) [57]. In summary Table 6.1 lists the ten statistical features used to augment the LBP representation, in each case the letter in parenthesis indicates the category in which the statistic belongs: E (Entropy), M (GLCM) and W (DWT).

Thus on completion of the feature extraction process using both LBPs and the additional texture statistical metrics described above each household could be represented using a feature vector of length 266 ($256 + 10 = 266$). Alternatively the LBP representations and the texture statistical metrics (TS) can be used on their own. Thus in total seven different texture based representations were available: (i) $LBP_{8,1}$, (ii) $LBP_{8,2}$, (iii) $LBP_{8,3}$, (iv) TS, (v) $LBP_{8,1}$ + TS, (vi) $LBP_{8,2}$ + TS, and (vii) $LBP_{8,3}$ + TS. In each case the features were encapsulated using a feature space representation from which a collection of feature vectors could be generated, one per image.

6.4 Feature Selection and Classification

Once the 256 LBPs features and 10 statistical features have been identified, as described in Sections 6.2 and 6.3 above, but before the classifier generation could be commenced, the data was discretised and then a feature selection mechanism was applied to the feature vector space. As before the aim was to reduce the overall number of dimensions so that only those features that served as good discriminators between classes were retained. The same three feature selection mechanisms as considered previously with respect to the graph-based approach and the colour histogram based approach were used again with respect to the LBP approach described in this chapter: (i) Chi-Squared, (ii) Gain Ratio and (iii) Information Gain.

Once the feature selection was complete the images were represented in terms of a set of feature vectors drawn from the reduced feature space. Some form of classifier generator could then be applied. As in the case of the previous two methods extensive evaluation was conducted so as to test the operation of the proposed approach using different parameters and their variations, however this chapter only reports the most significant results obtained (there is insufficient space to allow for the presentation of all the results obtained). With respect

to the evaluation presented in this chapter the same eight classifier generation methods as in Chapter 4 and 5 were considered: (i) Decision Tree (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network. As before the implementations used were those available in the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench [186].

6.5 Evaluation

To evaluate the proposed LBP based approach the same labelled training data was used as that used to evaluate the graph-based and colour histogram approaches described in the previous two chapters. The overall aim of the evaluation was to provide evidence that census data can be effectively estimated using the proposed approach. To this end four sets of experiments were conducted as follows:

1. **Data Representation**: A set of experiments to determine the effect on classification performance using the suggested LBP variations, the texture statistics on their own and a combination of the two (Sub-section 6.5.1).
2. **Feature Selection**: A set of experiments to compare the operation of the various suggested feature selection algorithms (Sub-section 6.5.2).
3. **Number of attributes**: A set of experiments to analyse the effect that the number of selected attributes (k) in the context of feature selection, had on performance (Sub-section 6.5.3).
4. **Classification Generation Method**: A set of experiments to determine the effect on classification performance when using different classifier generation methods (Sub-section 6.5.4).

Each is discussed in further detail in Sub-sections 6.5.1 to 6.5.4 below. Ten fold Cross-Validation (TCV) was applied throughout and as in the case of earlier experiments, performance was recorded in terms of: (i) accuracy (AC), (ii) area under the ROC curve (AUC), (iii) the F-Measure (FM), (iv) sensitivity (SN) and (v) specificity (SP). As before the discussion presented in this chapter concerning the evaluation of the texture based approaches is focused on the AUC metric.

6.5.1 Data Representation

For the experiments directed at evaluating the seven different variations of the LBP representation, including the Texture Statistics (TS) representation is isolation, in the context of classification performance, the segmented household data was translated using each of the seven

Table 6.2: The seven proposed texture based representations

Representations	LBP			TS	Feature Space
	Sampling	Distance	Features		
$LBP_{8,1}$	8	1	256		256
$LBP_{8,2}$	8	2	256		256
$LBP_{8,3}$	8	3	256		256
TS				10	10
$LBP_{8,1} + TS$	8	1	256	10	266
$LBP_{8,2} + TS$	8	2	256	10	266
$LBP_{8,3} + TS$	8	3	256	10	266

Table 6.3: Classification performance using the seven proposed texture based representations

Representations	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
$LBP_{8,1}$	**0.771**	**0.881**	**0.759**	**0.771**	**0.852**	**0.720**	0.824	**0.718**	**0.720**	**0.825**
$LBP_{8,2}$	0.643	0.781	0.639	0.643	0.761	0.600	0.711	0.600	0.600	0.762
$LBP_{8,3}$	0.643	0.749	0.632	0.643	0.755	0.560	0.668	0.561	0.560	0.716
TS	0.457	0.517	0.439	0.457	0.639	0.460	0.645	0.478	0.460	0.729
$LBP_{8,1} + TS$	0.714	0.838	0.699	0.714	0.818	0.700	**0.826**	0.699	0.700	0.813
$LBP_{8,2} + TS$	0.657	0.775	0.646	0.657	0.748	0.560	0.703	0.561	0.560	0.720
$LBP_{8,3} + TS$	0.586	0.770	0.576	0.586	0.705	0.560	0.626	0.561	0.560	0.718

proposed texture based representations to give seven distinct data sets as summarised in Table 6.2.

For the evaluation directed at determining the most effective representation (Objective 1) Chi-Squared feature selection was used with $k = 40$ because additional experiments, reported on later this chapter in Sub-sections 6.5.2 and 6.5.3, had revealed that this was the most appropriate feature selection algorithm and the most appropriate value for k. For similar reasons Neural Network classification was adopted with respect to the experiments directed at Objective 1 because additional experiments, reported in Sub-section 6.5.4, had indicated that this tended to produce the most effective performance.

The obtained results are presented in Table 6.3 (best values are shown in bold font). From the table it can be observed that with respect to the Site A (wet season) data set the $LBP_{8,1}$ representation gave the best results, AUC $= 0.881$. With respect to the Site B (dry season) data set the best performance was obtained using the $LBP_{8,1} + TS$ representation, AUC $= 0.826$. The TS representation on its own did not work well.

Figure 6.5 shows the recorded classification performance, in terms of the AUC results from Table 6.3, in the form of a bar graph. The horizontal axis represents the seven representations: (i) $LBP_{8,1}$, (ii) $LBP_{8,2}$, (iii) $LBP_{8,3}$, (iv) TS, (v) $LBP_{8,1} + TS$, (vi) $LBP_{8,2} + TS$, and (vii) $LBP_{8,3} + TS$, while the vertical axis represents the AUC score. From the graph it is again clear that: (i) the $LBP_{8,1}$ representation gave the best results for the Site A data sets and $LBP_{8,1} + TS$ gave the best results for the Site B data sets; (ii) in terms of the LBP representations, using a radius of

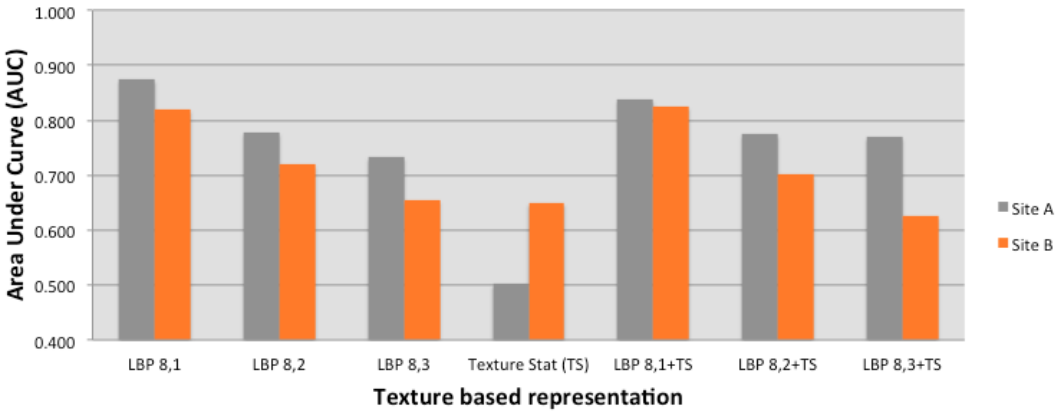


Figure 6.5: Bar graph showing classification performance in terms of AUC using the seven different texture based representations

1 gave the best results whereas a radius of 3 gave the worst results for both Sites A and B; (iii) overall performance was better with respect to Site A than Site B; and (iv) the TS performance was better with respect to Site B than Site A. However, in most cases it can be argued that the TS representation did not serve to provide any enhancement in performance. From the results given in Table 6.3 and Figure 6.5 an argument can be constructed suggesting that $LBP_{8,1}$ is the most appropriate LBP representation for population estimation mining from satellite image data.

A Friedman test was applied to the AUC results in order to determine whether the results were statistically significant or not [50]. The AUC results obtained from all seven variations of the texture based representations, with respect to both Site A and B, are listed in Table 6.4 (columns two and three). In each case the number in parentheses indicates the overall relative ranking of each individual result with respect to the two data sets. The Average Rank (AR) of each classifier is given in the fourth column, recall that this is the mean value of the rankings for each representation. Recall also that the Friedman test statistic is based on the AR values, and is calculated using equation 6.4, where N is the number of data sets (2) and K is the number of classification technique considered (7).

$$\chi_F^2 = \frac{12N}{K(K+1)} \left[\sum_{i=1}^{K} AR_i^2 - \frac{K(K+1)^2}{4} \right] \tag{6.4}$$

The Friedman test statistic and corresponding p value are presented in the first row of Table 6.4. In this case the Friedman test statistic (11.143) and the significance threshold ($p < 0.05$) indicates that the null hypothesis, that there is no statistical difference between the techniques, can be rejected. A post hoc Nemanyi test was applied to determine the distinction in performance between the individual representations [102]. Recall that the performance of two

approaches is statistically different if their ARs are different by more than a Critical Difference (CD) value calculated using equation 6.5 where the critical difference level α ($\alpha = 0.1$) and the value $q_{\alpha,\infty,K}$ is based on the Studentised range statistic [86].

Table 6.4: AUC values recorded when considering the variations of the texture based representations

Friedman test statistic = 11.143 ($p < 0.05$)			
Algorithm	Site A	Site B	AR
$LBP_{8,1}$	0.881(1)	0.824(2)	1.50
$LBP_{8,2}$	0.781(3)	0.711(3)	3.00
$LBP_{8,3}$	0.749(6)	0.668(5)	5.50
TS	0.517(7)	0.645(6)	6.50
$LBP_{8,1} + TS$	0.838(2)	0.826(1)	1.50
$LBP_{8,2} + TS$	0.775(4)	0.703(4)	4.00
$LBP_{8,3} + TS$	0.770(5)	0.626(7)	6.00

$$CD = q_{\alpha,\infty,K}\sqrt{\frac{K(K+1)}{12N}} \tag{6.5}$$

The critical difference diagram for the proposed texture based population estimation mining from satellite imagery approach in presented in Figure 6.6 (the calculated CD for the diagram is 4.10). In the diagram the classification techniques are listed in ascending order of ranked performance along the Y-axis; and the associated average rank (across both data sets) is listed along the X-axis. From the diagram it can be observed that the results produced using the $LBP_{8,1}$ and $LBP_{8,1} + TS$ representations are statistically better than the results using the $LBP_{8,3} + TS$ and TS representations (there is "white space" between them). There was no statistically significant difference with respect to the $LBP_{8,1}$, $LBP_{8,1} + TS$, $LBP_{8,2}$, $LBP_{8,1} + TS$, and $LBP_{8,3}$ representations (because the critical difference tails overlap). Thus it can be concluded that $LBP_{8,1}$ is the most appropriate representation in the context of the population estimation mining application considered in this chapter.

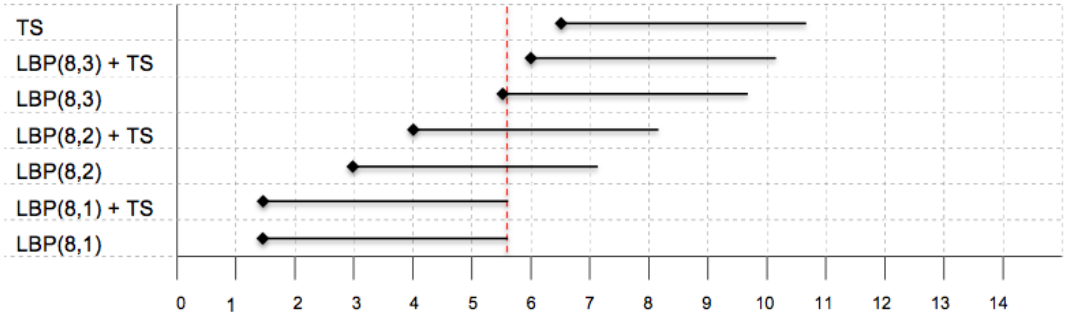


Figure 6.6: Nemenyi Significance Diagram for the different LBP representations

Table 6.5: Classification performance using the three different feature selection methods

Algorithms	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
Chi – Squared	0.771	0.881	0.759	0.771	0.852	**0.720**	**0.824**	**0.718**	**0.720**	**0.825**
GainRatio	0.743	0.862	0.723	0.743	0.825	0.620	0.728	0.621	0.620	0.759
InformationGain	**0.786**	**0.893**	**0.772**	**0.786**	**0.861**	0.540	0.723	0.540	0.540	0.725

6.5.2 Feature Selection

Recall that three different algorithms for feature selection were considered. A further set of experiments was thus conducted in order to identify the most appropriate feature selection mechanism to be used with respect to the proposed population mining framework. For the experiment the $LBP_{8,1}$ representation was used as this had been found to produce the best results as establishes in the previous sub-section (Sub-section 6.5.1). A value of $k = 40$ was again used, together with the Neural Network learning method, for the same reasons as before (they produced the best results with respect to the experiments reported on later in this chapter in Sub-sections 6.5.3 and 6.5.4 respectively).

The results from the experiments are presented in Table 6.5. From the table it can be observed that the best results were obtained using Information Gain feature selection with respect to the Site A data set (AUC = 0.893); while Chi-Squared feature selection gave the best results with respect to the Site B data set (AUC = 0.824). It was not clear as to why this might be the case. Gain Ratio was found to be the most appropriate feature selection method with respect to the graph-based and colour histogram based approaches resented in the previous two chapters.

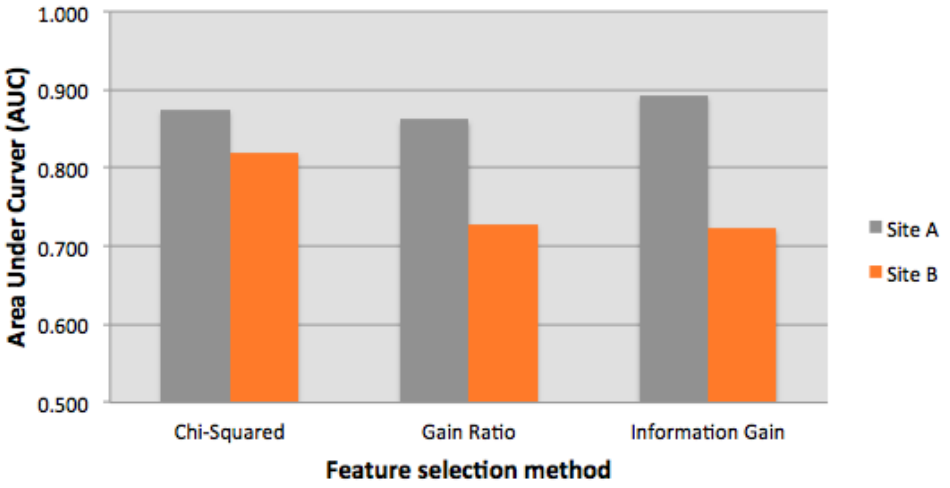


Figure 6.7: Bar graph showing classification performance in terms of AUC values using the three different feature selection methods

The bar graph in Figure 6.7 illustrates the classification performances results, in terms of only the recorded AUC values, with respect to both Site A and B. The horizontal axis represents the three feature selection algorithms: (i) Chi-Squared, (ii) Gain Ratio and (iii) Information Gain. The vertical axis represents the AUC values. The graph confirms that: (i) best AUC values were obtained using Information Gain feature selection for Site A, and Chi-Squared feature selection for Site B, and (ii) the AUCs recorded for Site A were higher than those recorded for Site B in all cases. As observed previously the later was probably because the extracted features from the wet season data provided for a better contrast between features than in the case of the dry season data.

Again a Friedman test was applied to determine whether the above results were statistically significant or not. The recorded AUCs are listed in Table 6.6 (columns two and three) with their individual ranking and their associated AR value (column four). The table again shows that the Chi-Squared feature selection method produced the best performance (AR = 1.5), while the Gain Ratio feature selection produced the worst performance (AR = 2.5).

Table 6.6: AUC values recorded for each feature selection method

Friedman test statistic = 1.00 ($p > 0.1$)			
Algorithm	Site A	Site B	AR
$Chi-Squared$	0.881(2)	0.824(1)	1.50
$GainRatio$	0.862(3)	0.728(2)	2.50
$InformationGain$	0.893(1)	0.723(3)	2.00

As before the Friedman test statistic (1.00) and corresponding p value ($p > 0.1$) value are presented in the first row of Table 6.6 and indicate that the null hypothesis, that there is no statistical difference in operation between the techniques was supported, thus no post hoc Nemenyi test was conducted. However, from the experiments reported on in this sub-section it can be concluded that Chi-Squared feature selection is the most appropriate measure in the context of texture based population estimation mining from satellite data because it had the highest average ranking with respect to the Friedman test as presented in Table 6.6.

6.5.3 Number of attributes

In order to identify the effect on classification performance of the value of k, the number of attributes selected when using Chi-Squared feature selection, a sequence of experiments was conduct using a range of values of k from $k = 15$ to $k = 50$ incrementing in steps of 5. For the experiments the $LBP_{8,1}$ representation was again used because the experiments reported in Sub-section 6.5.1 had indicated that this produced the best performances. The Neural Network learning method was again adopted because the experiments reported later in Sub-section 6.5.4 had indicated that this tended to generate a best performance.

The obtained results from the experiments are presented in Table 6.7. From the table it can be seen that $k = 25$ produced the best result with respect to Site A (AUC $= 0.883$); whereas with respect to Site B $k = 40$ produced the best result (AUC $= 0.824$). The value that produced

Table 6.7: Classification performance using different values for k with respect to Chi-Squared feature selection

Number of k	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
$k = 15$	0.786	0.875	**0.784**	0.786	**0.866**	0.600	0.697	0.600	0.600	0.745
$k = 20$	0.714	0.867	0.716	0.714	0.835	0.580	0.722	0.583	0.580	0.725
$k = 25$	0.771	**0.883**	0.768	0.771	0.859	0.600	0.718	0.593	0.600	0.754
$k = 30$	0.757	0.877	0.751	0.757	0.834	0.620	0.752	0.619	0.620	0.762
$k = 35$	0.771	0.879	0.761	0.771	0.856	0.600	0.796	0.596	0.600	0.754
$k = 40$	0.771	0.881	0.759	0.771	0.852	**0.720**	**0.824**	**0.718**	**0.720**	**0.825**
$k = 45$	**0.800**	0.862	**0.784**	**0.800**	0.861	0.660	0.809	0.660	0.660	0.786
$k = 50$	0.786	0.868	0.772	0.786	0.859	0.680	0.807	0.676	0.680	0.806

the worst result for Site A was of $k = 45$ (AUC $= 0.862$), whereas with respect to Site B $k = 15$ produced the worst result (AUC $= 0.697$). A line graph for the obtained results in terms of AUC is presented in Figure 6.8. The horizontal axis represents the k values and the vertical axis the AUC values. The graph shows, as indicated by the table, that with respect to the Site A data set the best performance was produced using $k = 25$ and that with respect to the Site B data set the best performance was produced using $k = 40$.

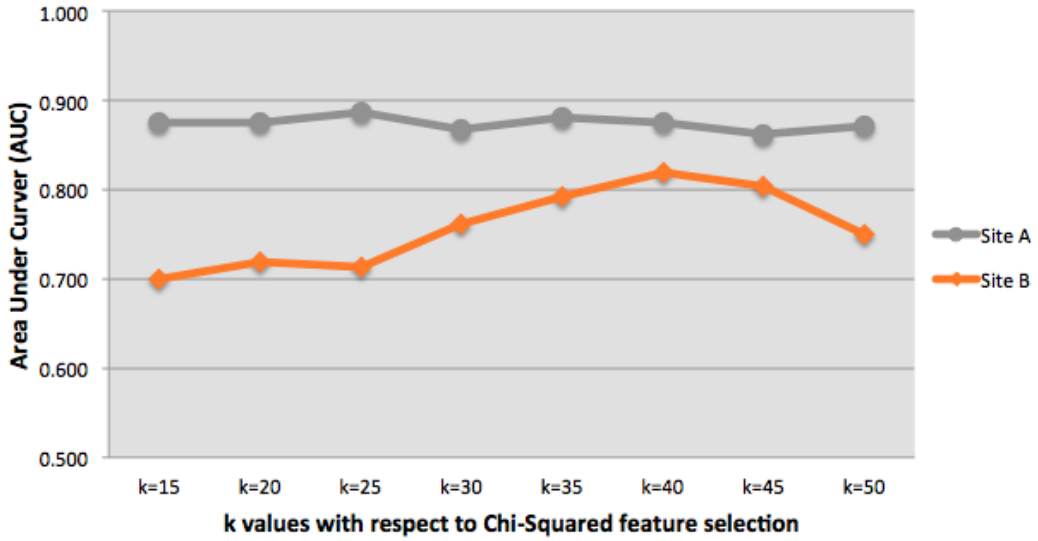


Figure 6.8: Line graph showing classification performance using different values for k with respect to Ch-Squared feature selection

A Friedman test was again conducted. Table 6.8 gives the AUC values used. The Friedman test statistic (6.167) and corresponding p value ($p > 0.1$), given in the first row of the table, indicate that the null hypothesis could be accepted; there is no statistical difference in operation between the techniques, thus no post hoc Nemenyi test was conducted. However, from the AUC

Table 6.8: AUC values recorded for each k value

Friedman test statistic = 6.167 ($p > 0.1$)			
kvalue	Site A	Site B	AR
$k = 15$	0.875(5)	0.697(8)	6.50
$k = 20$	0.867(7)	0.722(6)	6.50
$k = 25$	0.883(1)	0.718(7)	4.00
$k = 30$	0.877(4)	0.752(5)	4.50
$k = 35$	0.879(3)	0.796(4)	3.50
$k = 40$	0.881(2)	0.824(1)	1.50
$k = 45$	0.862(8)	0.809(2)	5.00
$k = 50$	0.868(6)	0.807(3)	4.50

test results presented in the Tables 6.7 and 6.8 and Figure 6.8 it was concluded that $k = 40$ was the most appropriate value for k when using Chi-Squared feature selection in the context of the texture based population estimation mining approach presented in this chapter. (This is why $k = 40$ was used with respect to the previously reported experiments.)

6.5.4 Classification Generation Method

This section presents the results obtained for the evaluation conducted with respect to the eight different classifier generators considered: (i) Decision Tree generators (C4.5), (ii) Naive Bayes, (iii) Averaged One Dependence Estimators (AODE), (iv) Bayesian Network, (v) Radial Basis Function Network (RBF Network), (vi) Sequential Minimal Optimisation (SMO), (vii) Logistic Regression and (viii) Neural Network. In each case the $LBP_{8,1}$ representation was used because this produced the best result with respect to the experiments reported in Sub-section 6.5.1. The Chi-Squared feature selection method, with $k = 40$, was used because previous experiments, reported above, indicated that this produced good results.

The results are presented in Table 6.9. From the table it can clearly be observed that the Neural Network classifier generator produced the best results for both the Site A and Site B data sets (AUC = 0.881 and AUC = 0.824, respectively). The C4.5 decision tree generator produced substantially the worst performance for both sites (AUC = 0.693 and AUC 0.652, respectively).

Figure 6.5.4 shows an associated bar graph generated from the AUC data given in Table 6.9. The horizontal axis represents the eight learning approaches; while the vertical axis represents the AUC scores. From the graph it can be observed that: (i) the classification performance with respect to Site A was better than for Site B when using C4.5, AODE, SMO, Logistic Regression and Neural Network, (ii) the classification performance with respect to Site B was better than for Site A when using Naive Bayes, Bayesian Network, and RBF Network, (iii) Neural Network produced the best overall outcomes for both Site A and Site B, and (iv) C4.5 produced the worst results for both data sets.

As for the results reported with respect to the previous experiments the Friedman statistical test was again applied using the recorded AUC values as listed in Table 6.10, together with

Table 6.9: Classifier performance using different classification models

Generator	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
C4.5	0.571	0.693	0.567	0.571	0.742	0.500	0.652	0.500	0.500	0.700
NaiveBayes	0.486	0.705	0.502	0.486	0.730	0.540	0.758	0.542	0.540	0.796
AODE	0.671	0.776	0.619	0.671	0.753	0.620	0.756	0.619	0.620	0.771
BayesianNetwork	0.557	0.716	0.567	0.557	0.761	0.520	0.762	0.525	0.520	0.782
RBFNetwork	0.600	0.718	0.593	0.600	0.768	0.600	0.722	0.595	0.600	0.776
SMO	0.700	0.777	0.680	0.700	0.789	0.680	0.761	0.682	0.680	0.789
LogisticRegression	**0.771**	0.859	**0.778**	**0.771**	**0.885**	0.680	0.756	0.679	0.680	0.803
Neural Network	**0.771**	**0.881**	0.759	**0.771**	0.852	**0.720**	**0.824**	**0.718**	**0.720**	**0.825**

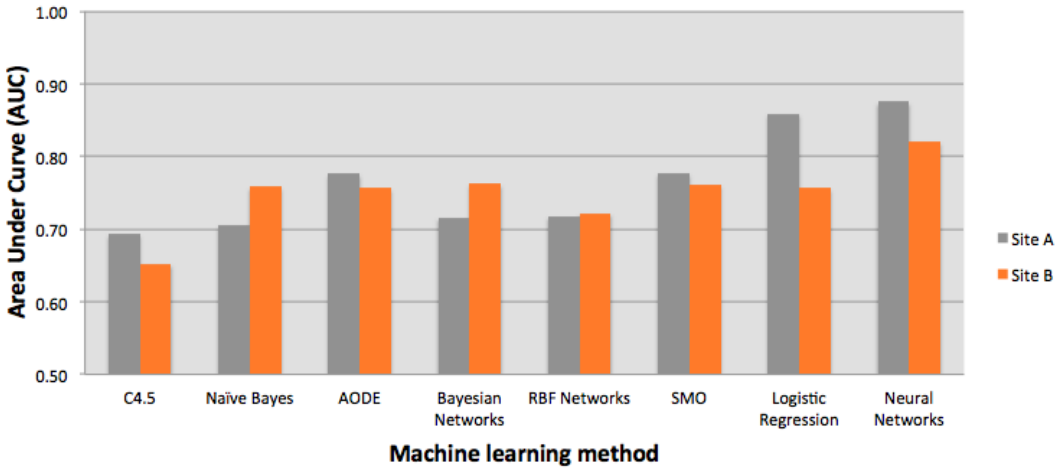


Figure 6.9: Bar graph showing classification performance using different classification models

their overall ranking with respect to the two data sets. The AR of each classifier is given in the fourth column. The Friedman test statistic (10.353) and corresponding p value ($p < 0.1$) are given in the first row of the table and indicate that the null hypothesis, $H0$ can be rejected. A post hoc Nemanyi test was therefore conducted. The associated significant diagram is given in Figure 6.5.4; the calculated CD for the diagram is 4.82. The diagram confirms that Neural Network classification produced the best results (as already established) and that the result was statistically different from the C4.5 and RBF Network classification model. There was no significant difference with respect to the Naive Bayes, AODE, Bayesian Network, RBF Network, SMO, Logistic Regression and Neural Network representations because there critical difference tails overlap.

Table 6.10: AUC values recorded for each classification model

Friedman test statistic = 10.353 ($p < 0.1$)			
Generator	Site A	Site B	AR
C4.5	0.693(8)	0.652(8)	8.00
Naive Bayes	0.705(7)	0.758(4)	5.50
AODE	0.776(4)	0.756(5.5)	4.75
Bayesian Network	0.716(6)	0.762(2)	4.00
RBF Network	0.718(5)	0.722(6)	5.50
SMO	0.777(3)	0.761(3)	3.00
Logistic Regression	0.859(2)	0.756(5.5)	3.75
Neural Network	0.881(1)	0.824(1)	1.00

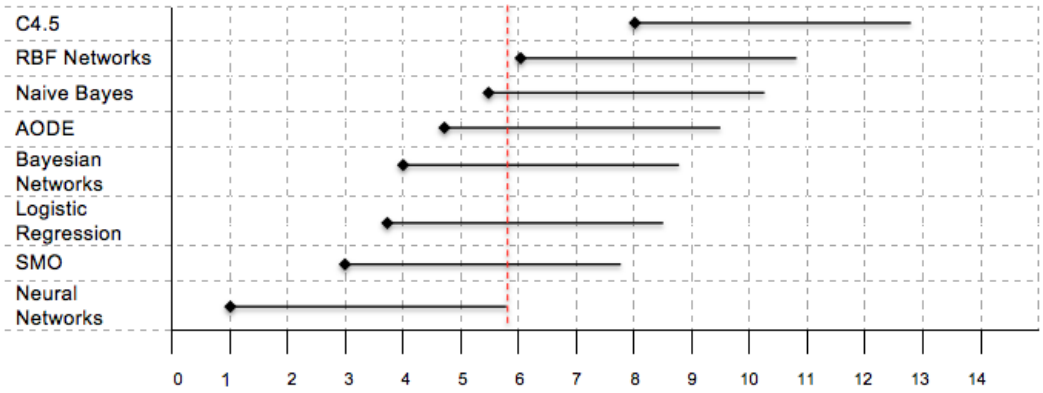


Figure 6.10: Nemenyi Significance Diagram for the different classification models

Thus, from the above, it was concluded that Neural Network classification is the most appropriate classifier generator method in the context of the texture based population estimation mining approach. (This is why Neural Network classification was used with respect to the previously reported experiments.)

6.6 Discussion

The evaluation results presented in the foregoing section indicate that the proposed population estimation mining, using a representation based on image texture (local binary patterns in some cases augmented with texture statistical information), performed well with respect to the two satellite images datasets considered (Site A and Site B). Four sets of experiments were conducted from which the main findings may be summarised as follows:

1. The most appropriate representation with respect to the application of population estimating mining for both data sets (Site A and Site B), in terms of best AUC, was the $LBP_{8,1}$ representation, follow by the combination of the $LBP_{8,1}$ and the texture statistical representation ($LBP_{8,1} + TS$). The *LBPs* with radius of 1 produced better results, in terms of classification performance, than *LBPs* with radii of 2 or 3. It was conjectured that this was because LBPs with a radius of 1 were generated using the nearest neighbour pixels, while with radius 2 and 3 the neighbourhoods were further away, hence radius 1 LBPs better served to capture the local texture.

2. The most appropriate feature selection mechanism with respect to both data sets (Site A and Site B) was found to be Chi-Squared feature selection, follow by Information Gain, followed by Gain Ratio feature selection. (Gain Ratio was also found to be the most appropriate feature selection mechanism with respect to the graph-based and colour histogram based approaches presenting in the previous two chapters.)

3. With respect to the Chi-Squared feature selection approach the most appropriate value for k (number of attributes) was found to be $k = 40$. It is conjectured that lower values of k did not provide good performance because there was insufficient data to build an effective classifier, while larger values of k resulted in overfitting.

4. A number of different classifier generator methods produced good results, however the best classification generation method, from the reported evaluation, was found to be the Neural Network method.

6.7 Statistical Comparison of the Proposed Image Classification Approaches

In this section, the evaluation and comparison of all the proposed population estimation mining approaches considered in this and the previous two chapters is presented. The comparison was undertaken in terms of classification effectiveness and run time complexity. Recall that the three different representation approaches proposed in this thesis, to classifying household satellite images according to the nature of each individual household, were: (i) graph-based, (ii) colour histogram based, and (iii) texture based. Each of these approaches will be briefly summarised below before considering their comparison.

The first representation was found on the concept of graph-based representation (see the detail from Chapter 4). This approach used a hierarchical decomposition technique together with a quadtree based representation, one graph/tree per household image. A subgraph mining algorithm, gSpan was then applied to identified frequently occurring subgraphs within a quadtree representation. The identified subgraphs were used to generate a feature space which was used to represent household image data. The reported evaluation indicated that the best classification performance in terms of AUC was obtained when using a low support threshold ($\sigma = 10$). The Gain Ratio feature selection mechanism was found to be the most appropriate feature selection mechanism with respect to this representation; and the most appropriate k value, with respect to Gain Ratio feature selection, was found to be $k = 55$. The most suitable classification generator was found to be the Bayesian Network model.

The second approach, the colour histogram based approach presented in Chapter 5, was found on the idea of representing each identified household image as a collection of colour histograms, typical seven histograms per household (red, green, blue, hue, saturation, value and greyscale) together with additional basic statistical metrics concerning the distribution of colour across an image. The reported evaluation indicated that high classification results were obtained when using the Colour Histogram (CH) representation. The Gain Ratio feature selection mechanism was found to be the most appropriate feature selection method together with $k = 25$. Best classification results were obtained using the Logistic Regression classification model.

The third approach presented in this chapter, was found on the idea of representing the segmented household images using a texture based approach. Three LBP variations were considered ($LBP_{8,1}$, $LBP_{8,2}$ and $LBP_{8,3}$), and additional simple texture statistical metrics were also used. The reported evaluation indicate that a best classification performance was obtained when using the $LBP_{8,1}$ representation. The Chi-Squared feature selection mechanism was found to be the most appropriate feature selection method to adopt together with $k = 40$. Best classification results were obtained using the Neural Network classifier generator.

For the comparison of the three approaches presented in this section a number of variations of each approach were considered as follows (fifteen distinct representations in total):

1. For the graph-based approach a range of σ values, $\{10, 20, 30, 40, 50\}$ was used together with the Gain Ratio feature selection method ($k = 55$) and the Bayesian Network classification model.

2. For the colour histogram based approach, three variations were considered: colour histogram (CH), colour statistics (CS) and combine ($CH + CS$). The Gain Ratio feature selection method, with $k = 25$, and the Logistic Regression classification model was also used.

3. For the texture based approach, all seven variations were considered: $LBP_{8,1}$, $LBP_{8,2}$, $LBP_{8,3}$, texture statistics (TS), combine $LBP_{8,1} + TS$, combine $LBP_{8,2} + TS$ and combine

Table 6.11: Classification performance using the three different proposed approaches and their variations

Approach	Site A					Site B				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
$\sigma = 10$	0.600	0.808	0.596	0.600	0.734	**0.800**	**0.879**	**0.792**	**0.800**	**0.876**
$\sigma = 20$	0.429	0.606	0.424	0.429	0.639	0.520	0.697	0.519	0.520	0.742
$\sigma = 30$	0.329	0.427	0.336	0.329	0.600	0.380	0.535	0.382	0.380	0.666
$\sigma = 40$	0.343	0.396	0.341	0.343	0.562	0.320	0.495	0.306	0.320	0.601
$\sigma = 50$	0.371	0.446	0.354	0.371	0.557	0.320	0.461	0.312	0.320	0.603
CH	0.657	0.822	0.662	0.657	0.806	0.640	0.821	0.633	0.640	0.798
CS	0.329	0.522	0.339	0.329	0.617	0.420	0.573	0.418	0.420	0.685
$CH + CS$	0.657	0.822	0.662	0.657	0.806	0.600	0.719	0.588	0.600	0.747
$LBP_{8,1}$	**0.771**	**0.881**	**0.759**	**0.771**	**0.852**	0.720	0.824	0.718	0.720	0.825
$LBP_{8,2}$	0.643	0.781	0.639	0.643	0.761	0.600	0.711	0.600	0.600	0.762
$LBP_{8,3}$	0.643	0.749	0.632	0.643	0.755	0.560	0.668	0.561	0.560	0.716
TS	0.457	0.517	0.439	0.457	0.639	0.460	0.645	0.478	0.460	0.729
$LBP_{8,1} + TS$	0.714	0.838	0.699	0.714	0.818	0.700	0.826	0.699	0.700	0.813
$LBP_{8,2} + TS$	0.657	0.775	0.646	0.657	0.748	0.560	0.703	0.561	0.560	0.720
$LBP_{8,3} + TS$	0.586	0.770	0.576	0.586	0.705	0.560	0.626	0.561	0.560	0.718

$LBP_{8,3} + TS$. The Chi-Squared feature selection method, with $k = 40$, together with the Neural Network learning method, was also used.

The results from the experiments are presented in Table 6.11. From the table it can be seen that the best results were obtained using $LBP_{8,1}$ approach with respect to the Site A data (AUC $= 0.881$); while the graph-based with $\sigma = 10$ gave the best results with respect to the Site B data (AUC $= 0.879$).

The bar graph in Figure 6.11 illustrates the classification performances results, in terms of only the recorded AUC values, with respect to both Sites A and B, using the three different approaches and their variations. The horizontal axis represents the 15 approaches. The vertical axis represents the AUC values. From the graph it can again be seen that: (i) best AUC values were obtained using the $LBP_{8,1}$ mechanism for the Site A data, and the graph-based approach with $\sigma = 10$ for the Site B data; (ii) it was conjectured that the AUC recorded for the graph-based approach for the Site B data was higher than for Site A data because the extracted features from the dry season data provided better information than the wet season data in the context of the graph-based approach; and (iii) the AUC recorded for the LBP variations of the texture based approach for the Site A data was higher than the Site B data because the generated features from the wet season data provide better texture information than the dry season data in the context of the texture based approach. In other wards the graph-based approach was better at distinguishing between dry season household images, while the LBP variation of the texture based approach was better at distinguishing between wet season household images; no such distinction was found with respect to the colour histogram based approach.

The run time complexity for each approach and its variations is presented in Table 6.12.

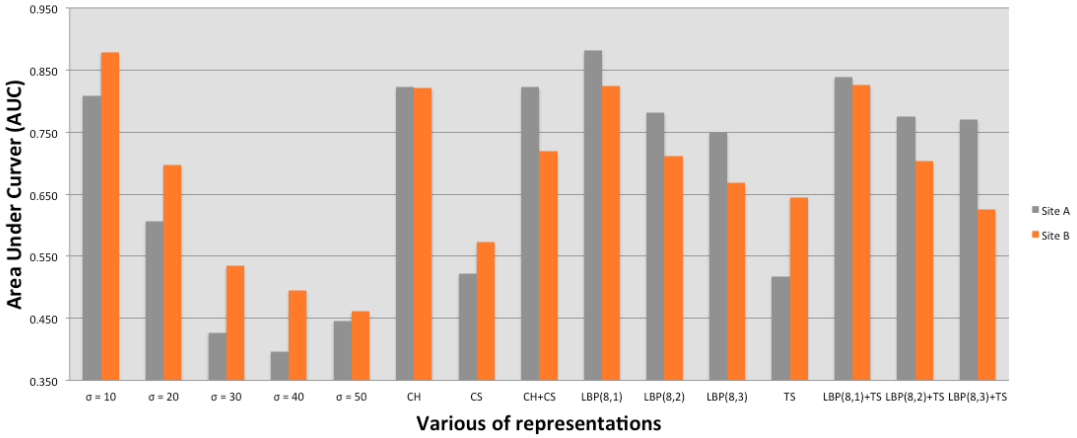


Figure 6.11: Bar graph showing classification performance in terms of AUC values using the three different proposed approaches and their variations

From the table it can be observed that: (i) The texture based approaches ($LBP_{8,1}$, $LBP_{8,2}$, $LBP_{8,3}$, $LBP_{8,1} + TS$, $LBP_{8,2} + TS$, $LBP_{8,3} + TS$) required the greatest run time of around 60 seconds with respect to the Site A data, and 40 seconds with respect to the Site B data (recall that the Site A featured 70 households while the Site B featured 50 households). The graph-based approach (σ = 40, 50) required the least run time of 0.01 seconds for both sites; (ii) the required run time for the Site A data was greater than for the Site B data in every cases, again because the number of households in the Site A data set was greater than the number of households in the Site B data set (70 versus 50); (iii) the LBP variations of the texture based approach required more run time than the other approaches because the best obtained results were obtained using Neural Network classification which required more learning time than the other algorithms considered; and (iv) the graph-based approach required less run time than the other approaches considered because the feature vectors used contained a small amount of zero-one data.

Again the Friedman test was used to compare the significance of the AUC classification performance results obtained for the different classifiers. The recorded AUCs are listed in Table 6.13 (columns two and three) with their individual ranking and their associated AR value (column four). The table confirms that the $LBP_{8,1}$ and $LBP_{8,1+TS}$ methods produced the best performance (AR = 2.00), while the $\sigma = 40$ graph-based mechanism produced the worst performance (AR = 14.50).

The Friedman test statistic (26.836) and corresponding p ($p < 0.01$) value are presented in the first row of Table 6.13 and indicate that the null hypothesis, that there is no statistical difference in operation between the techniques, can be rejected. A post hoc Nemanyi test was thus applied to analyse further the performance of the approaches. The resulting significant diagram is given in Figure 6.12 (the calculated CD for the diagram is 9.998). The diagram

Table 6.12: Run time complexity (s) using the three different proposed approaches and their variations

Approach	Site A	Site B
$\sigma = 10$	0.03	0.03
$\sigma = 20$	0.03	0.03
$\sigma = 30$	0.02	0.02
$\sigma = 40$	0.01	0.01
$\sigma = 50$	0.01	0.01
CH	0.37	0.15
CS	0.13	0.07
$CH + CS$	0.34	0.14
$LBP_{8,1}$	60.20	40.97
$LBP_{8,2}$	60.98	41.57
$LBP_{8,3}$	60.77	41.26
TS	3.50	2.47
$LBP_{8,1}$ +TS	57.71	39.73
$LBP_{8,2}$ +TS	60.44	41.04
$LBP_{8,3}$ + TS	59.98	39.67

Table 6.13: AUC values recorded using the three different proposed approaches and their variations

Friedman test statistic = 26.836 ($p < 0.01$)			
Approach	Site A	Site B	AR
$\sigma = 10$	0.808(5)	0.879(1)	3.00
$\sigma = 20$	0.606(10)	0.697(8)	9.00
$\sigma = 30$	0.427(14)	0.535(13)	13.50
$\sigma = 40$	0.396(15)	0.495(14)	14.50
$\sigma = 50$	0.446(13)	0.461(15)	14.00
CH	0.822(3.5)	0.821(4)	3.75
CS	0.522(11)	0.573(12)	11.50
$CH + CS$	0.822(3.5)	0.719(5)	4.25
$LBP_{8,1}$	0.881(1)	0.824(3)	2.00
$LBP_{8,2}$	0.781(6)	0.711(6)	6.00
$LBP_{8,3}$	0.749(9)	0.668(9)	9.00
TS	0.517(12)	0.645(10)	11.00
$LBP_{8,1} + TS$	0.838(2)	0.826(2)	2.00
$LBP_{8,2} + TS$	0.775(7)	0.703(7)	7.00
$LBP_{8,3} + TS$	0.770(8)	0.626(11)	9.50

shows the three proposed approaches and their variations listed in ascending order of ranked performance on the Y-axis, and the associated average rank (across both data sets) on the X-axis. The diagram demonstrates that the operation of the $LBP_{8,1}$, $LBP_{8,1} + TS$ and graph-based with $\sigma = 10$, was significantly different from the graph-based with σ = 30, 40, 50 (the critical difference tails do not overlap).

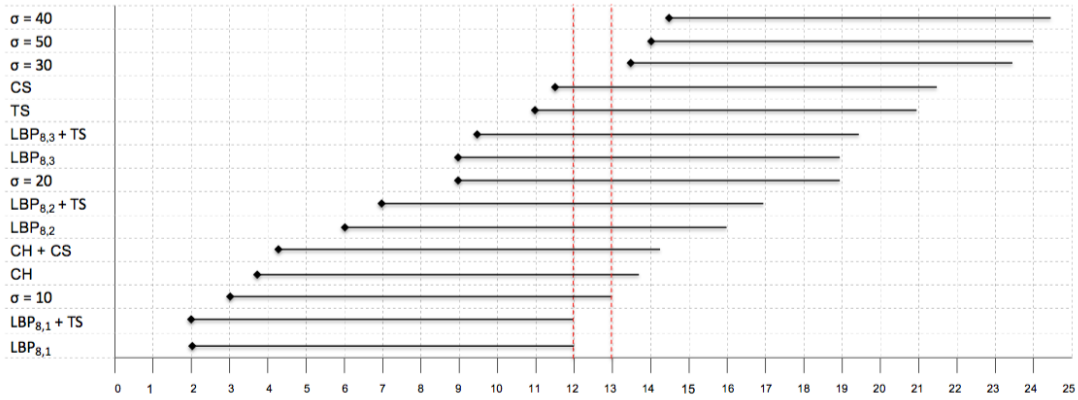


Figure 6.12: Nemenyi Significance Diagram using the three different proposed approaches and their variations

Therefore from the reported results presented in this section it can be concluded that the LBP approach is the most appropriate representation in the context of population estimation mining from satellite data.

6.8 Summary

A texture based representation focused on the used of LBPs, for capturing image texture, with respect to estimating population size from satellite imagery, has been described. Seven different texture based variations were considered: (i) $LBP_{8,1}$, (ii) $LBP_{8,2}$, (iii) $LBP_{8,3}$, (iv) TS, (v) $LBP_{8,1}$ + TS, (vi) $LBP_{8,2}$ + TS and (vii) $LBP_{8,3}$ + TS. Experiments were conducted using the Ethiopian hinterland test data used previously. The reported evaluation indicated that high AUC results could be obtained when using the $LBP_{8,1}$ representation. The Chi-Squared feature selection mechanism was found to be the most appropriate feature selection method to adopt together with $k = 40$. Best classification results were obtained using the Neural Network classification model.

A statistical comparison of all the approaches, and their variations, considered in this thesis was also presented: (i) graph-based, (ii) colour histogram based and (iii) texture based. The texture based approach was found to be the overall best approach for encapsulating household image content. Typically, the $LBP_{8,1}$ representation, together with Chi-Squared feature selection with $k = 40$ using Neural Network classification, produced the overall best result. In the

following chapter an alternative approach for predicting household size using regression analysis is presented. Because the $LBP_{8,1}$ representation produced the best results in the context of classification this is also the representation used in the context of the regression analysis discussion presented in the following chapter.

Chapter 7

Population Estimation Mining using Satellite Imagery: Regression Analysis

7.1 Introduction

As noted in Chapter 2, predictive analysis is the process of finding a predictive model which can be used in the context of new data. As also noted in Chapter 2 predictive analysis can be categorised in terms of (i) classification and (ii) regression. Recall that the difference between classification and regression is that classification models are used to predict categorical or discrete class labels, while regression models are used to predict a "numerical response variable" [71]. In the previous three chapters three alternative satellite image representations were considered and evaluated using classification algorithms from the previously presented evaluation it was established that the $LBP_{8,1}$ representation was the most effective. This chapter considers this representation in terms of regression analysis.

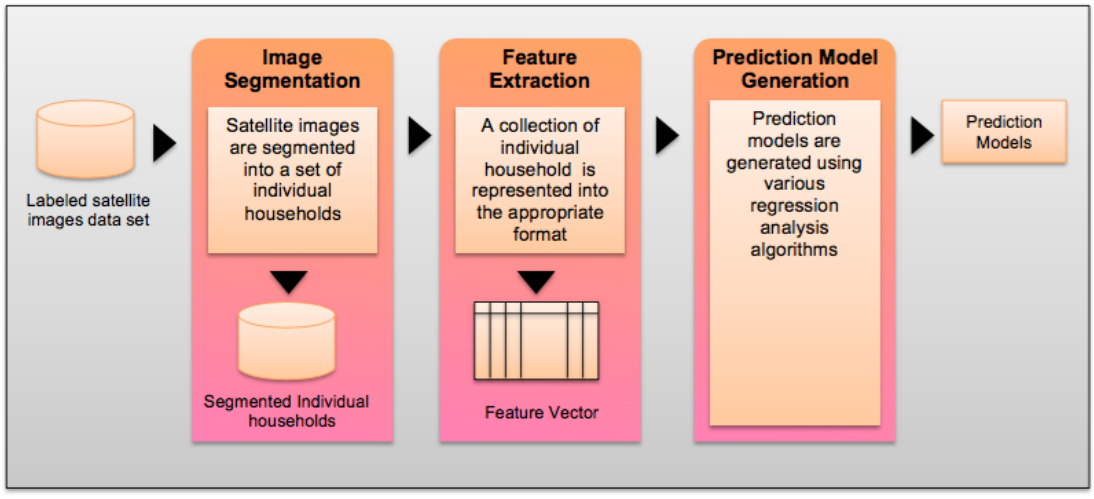


Figure 7.1: Schematic illustrating the population estimation mining approach using LBPs

A schematic of the proposed regression analysis approach for population estimation min-

ing is given in Figure 7.1. From the figure it can be seen that the overall approach encompasses three processes: (i) image segmentation, (ii) feature extraction and (iii) prediction model generation. The image segmentation process (the left rectangular box in Figure 7.1 was discussed in Chapter 3 and is thus not considered further here). The feature extraction process (the second rectangular box from the left in Figure 7.1) is concerned with translating the segmented data into a feature vector form ready for the generation of a regression model, this is where $LBP_{8,1}$ representation is applied. The third process included in Figure 7.1 (the right rectangular box in the figure) is prediction model generation; where some standard regression analysis algorithm can be applied. Once the model has been generated it can be applied to unseen data.

The rest of this chapter is organised as follows. The feature selection and prediction model generation process is considered in Section 7.2. Section 7.3 then reports on the evaluation of the proposed approach, followed by some discussion in Section 7.4. Finally, the main findings and some associated conclusions are presented in Section 7.5

7.2 Feature Selection and Prediction Model Generation

Once the input data (household satellite images) has been translated into the $LBP_{8,1}$ representation, we will have feature vectors comprise of 256 feature (as described in Sections 6.2 of Chapter 6). A feature selection mechanism is thus applied. As before the aim was to reduce the overall number of dimensions so that only those features that serve as good discriminators between classes are retained. However, because integer valued population sizes are used in the context of the regression analysis , as opposed to categorical class labels as in the case of classification, an alternative feature selection strategy was required. Correlation-based Feature Subset selection (CFS) [64] was therefore adopted for this purpose. CFS feature selection uses a wrapper method to find a best heuristic for finding the best feature subset most appropriate to regression analysis (numeric prediction).

Once the feature selection process was completed the images were represented in terms of the identified subset of the features vectors. Some form of regression model generation can then be applied. For the evaluation presented later in this chapter the following regression analysis methods were considered: (i) Linear Regression (Linear Reg), (ii) Least Median squared Linear Regression (LMedS) [150], (iii) Isotonic Regression (IsoReg) [14] and (iv) Support Vector Machine for regression (SVMreg) [165, 168]. As in the case of the classification algorithms used in the foregoing chapters the implementations used were those available in the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench [186].

7.3 Evaluation

Using the data sets presented previously in Chapter 3, a predictive model was generated with respect to each data set using the $LBP_{8,1}$ representation and the four regression model generators noted above with and without CFS feature selection. A total of eight models were therefore

Table 7.1: Comparison of the different regression analysis approaches in terms of prediction performance

Learning method	Site A				Site B			
	Coef	MAE	RMSE	RT(s)	Coef	MAE	RMSE	RT(s)
LinearReg	-0.080	2.167	2.570	0.70	0.274	1.981	2.407	0.51
LMedS	-0.288	3.262	3.894	0.05	0.215	1.952	2.353	0.04
IsoReg	-0.309	2.382	2.841	0.12	0.156	1.940	2.295	0.08
SVMreg	-0.279	3.367	3.970	0.02	0.308	1.778	2.056	0.01
LinearReg+CFS	0.084	2.145	2.550	0.01	0.400	1.727	2.093	0.01
LMedS+CFS	0.252	1.988	2.373	0.01	0.428	1.687	2.038	0.07
IsoReg+CFS	-0.202	2.287	2.706	0.01	0.109	1.912	2.282	0.04
SVMreg+CFS	0.307	1.957	2.330	0.01	0.587	0.143	1.802	0.01

used: (i) Linear regression (Linear Reg), (ii) Least Median Squared regression (LMedS), (iii) Isotonic Regression (IsoReg), (iv) Support Vector Machine for regression (SVMreg), (v) Linear Regression with CFS (Linear Reg + CFS), (vi) Linear Median Squared regression with CFS (LMedS + CFS), (vii) Isotonic Regression with CFS (IsoReg + CFS) and (viii) Support Vector Machine regression with CFS (SVMreg + CFS). For the evaluation Ten fold Cross-Validation (TCV) was again used throughout, and the following measures were recorded: (i) Correlation Coefficient (Coef), (ii) Mean Absolute Error (MAE), (iii) Root Mean Squared Error (RMSE), and (iv) Run Time (RT(s)).

The obtained results are presented in Table 7.1. From the table it can clearly be seen that the SVMreg +CFS prediction learning method produced the best performance with respect to the Site A and B data sets ($Coef = 0.307$ and $Coef = 0.587$ respectively). In contrast, the Linear Regression learner produced the worst performance with respect to the Site A data set ($CS = -0.080$), while IsoReg +CFS produced the worst performance with respect to the Site B data set ($Coef = 1.09$).

The bar graph given in Figure 7.2 gives a different perspective on the $Coef$ values recorded in Table 7.1. From the graph it can be observed that: (i) the maximum $Coef$ was produced using SVMreg + CFS learning method with respect to the Site B data, (ii) the $Coef$ values with respect to the Site B data were positive values ($Coef > 0$) in every cases, this means that the predictor and response variables were alike, and (iii) the $Coef$ values with respect to Site A in the cases where CSF feature selection was not used were negative values ($Coef < 0$), this means that predictors and response variable were not alike. The $Coef$ can be used to approximate accuracy $Acc = Coef^2$ as shown in Figure 7.3. In the figure the eight approaches are listed along the X-axis while the associated "approximate" accuracy value (apx accuracy) is given along the Y-axis. From the figure it can be seen that the SVMreg +CFS prediction learning method produced the best performance with respect to the Site B data ($Acc = 0.340$), while the Linear Reg and Linear Reg + CFS produced the best performance with respect to the Site A data ($Acc = 0.094$). IsoReg +CFS with respect to Site B data set produced the worst performance with $Acc = 0.01$.

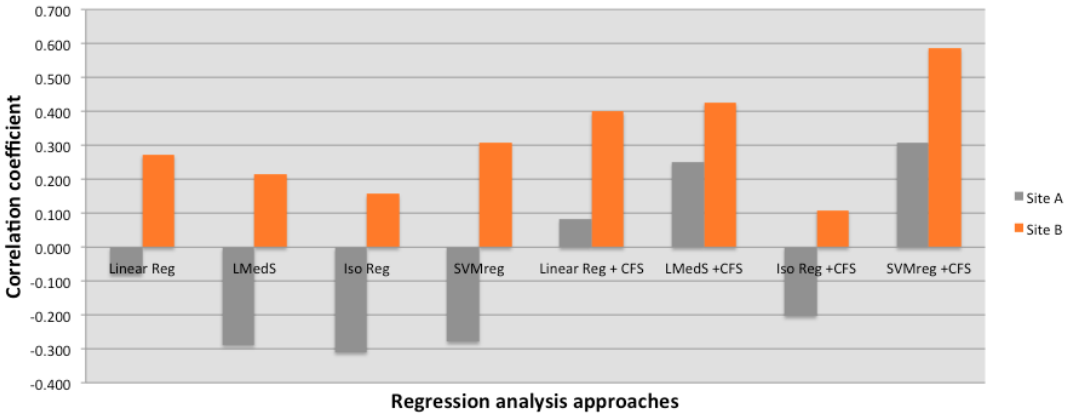


Figure 7.2: The comparison of $Coef$ results with respect to the different regression analysis approaches

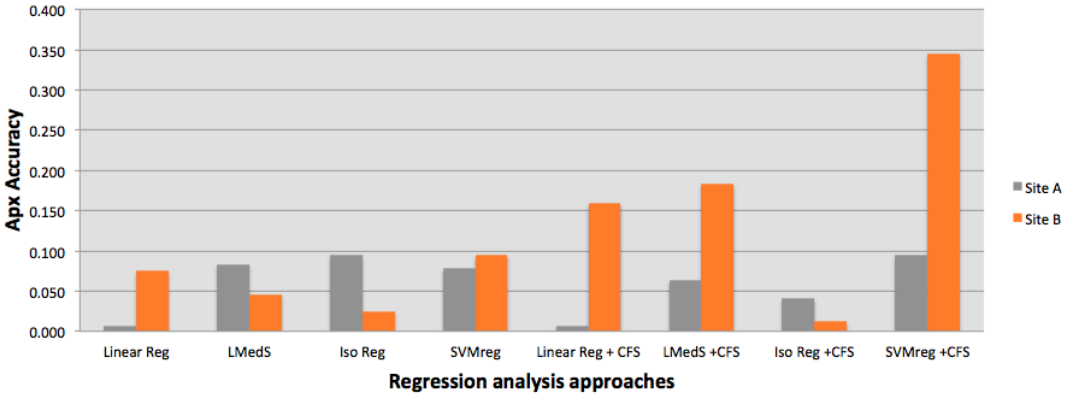


Figure 7.3: The comparison of "approximate" accuracy using $Coef$ with respect to the different regression analysis approaches

In terms of runtime complexity, from Table 7.1 it can also be seen that: (i) the run times used for generating the prediction models when the CFS feature selection mechanism was applied (predictably) were less than when CSF was not applied, because CSF reduced the number of dimensions in the feature space making it less complex; and (ii) the maximum recorded run time was with respect to the Linear Reg model (for both sites). The Linear Reg prediction model, generated using the $LPB_{8,1}$ representation and the CFS strategy, for the Site A and B data set were as shown in Equations 7.1 and 7.2, respectively.

$$\begin{aligned} \textit{Family Size} = \{ & -(0.1085)(\textit{normalised})LBP_{8,1}42 + (0.0645)(\textit{normalised})LBP_{8,1}86 \\ & +(0.2017)(\textit{normalised})LBP_{8,1}102 - (0.2070)(\textit{normalised})LBP_{8,1}106 \\ & -(0.3565)(\textit{normalised})LBP_{8,1}110 + (0.0899)(\textit{normalised})LBP_{8,1}118 \\ & +(0.1190)(\textit{normalised})LBP_{8,1}150 - (0.0453)(\textit{normalised})LBP_{8,1}155 \\ & +(0.1881)(\textit{normalised})LBP_{8,1}165 - (0.1432)(\textit{normalised})LBP_{8,1}171 + 0.4242\} \end{aligned} \tag{7.1}$$

$$\begin{aligned} \textit{Family Size} = \{ & (0.2547)(\textit{normalised})LBP_{8,1}11 + (0.0836)(\textit{normalised})LBP_{8,1}43 \\ & +(0.1408)(\textit{normalised})LBP_{8,1}44 - (0.3142)(\textit{normalised})LBP_{8,1}107 \\ & -(0.3241)(\textit{normalised}LBP_{8,1}163 + (0.4095)(\textit{normalised})LBP_{8,1}173 \\ & +(0.1126)(\textit{normalised})LBP_{8,1}211 + (0.2103)(\textit{normalised})LBP_{8,1}219 + 0.3796\} \end{aligned} \tag{7.2}$$

Equation 7.1 is the regression equation model generated using the Site A data set. The equation comprises two parts: (i) the ten CFS identified $LBP_{8,1}$ bin numbers (42, 86, 102, 106, 110, 118, 150, 155, 165 and 171) and their coefficients, and (ii) a global constant of 0.4242. Equation 7.1 is the regression equation model generated using the Site B data set. Again the equation comprises two parts: (i) the eight CFS identified $LBP_{8,1}$ bin numbers (11, 43, 44, 107, 163, 173, 211 and 219) and their coefficients, and (ii) a global constant of 0.3796.

7.4 Discussion

The overall prediction performance presented in the previous section, Section 7.3, indicated that the proposed population estimation mining, using regression analysis based on LBPs, performed well with respect to the two satellite image datasets considered (Site A and Site B). From the experiments conducted the main findings may be summarised as follows:

1. The prediction models generated using the Site B data set were more effective than when using the Site A data. This was because, as noted previously, the satellite images from site B (wet season) were more distinctive thus making the consequent prediction model more accurate than for the Site A data (dry season).

2. The run time to generate the prediction model using *CSF* is less than without *CFS*, because the feature vectors used were in a reduced form.

7.5 Summary

The main idea presented in this chapter was to considered a population estimation mining from satellite images prediction model using regression analysis based on the properties of individual households that are embedded in satellite images, which can then be used to predict the family size of each household and provided an estimate of population size for a given area. From the reported evaluation it was noted that, when using the Site B data set, the $LPB_{8,1}$ representation, coupled with the CFS strategy and SVMreg produced the best overall results. This prediction model was therefore one of the models used with respect to the large scale study described in next chapter.

Chapter 8

A Unified Process for Large Scale Population Estimation Mining Using Satellite Imagery

8.1 Introduction

from Chapter 1 that the main objective of the research presented in this thesis is to identify data mining techniques for predicting the population size to be associated with individual households identified from satellite imagery. The work presented so far has considered a number of techniques (graph-based, colour histogram based and texture based) coupled with classification and regression for predicting population size with respect to given households. The idea is of course that the techniques are applied in a much broader setting to estimate population size in a given region according to the individual households in that region. To achieve this the techniques described earlier need to be incorporated into a large scale population estimation mining process. The nature of this process is presented in this chapter.

A schematic of the proposed large scale population estimation mining process is given in Figure 8.1. From the figure it can be seen that the overall approach encompasses two parts: (i) prediction model generation and (ii) prediction for unseen data. The top half of Figure 8.1 is concerned with construction of the desired prediction model (the reader might like to compare this schematic with the generic schematic of the prediction process previously given in Figure 2.14 in Chapter 2). The second part of the overall process (bottom half of Figure 8.1) is concerned with prediction model usage. This process consists of five individual steps: (i) collation of a set of satellite images covering a prescribed area delimited by two pairs of geographic coordinates (latitudes and longitudes) describing opposing corner locations for the area of interest, (ii) segmentation of households within the satellite imagery, (iii) duplication household eliminated in situations where the same household appears in more than one image, (iv) feature extraction using the same feature set as used for model generation, and (v) application of the generated prediction model to predict family size for each household which is then later summed to give a total population size for the region. It is this second part that is of interest

with respect to this chapter.

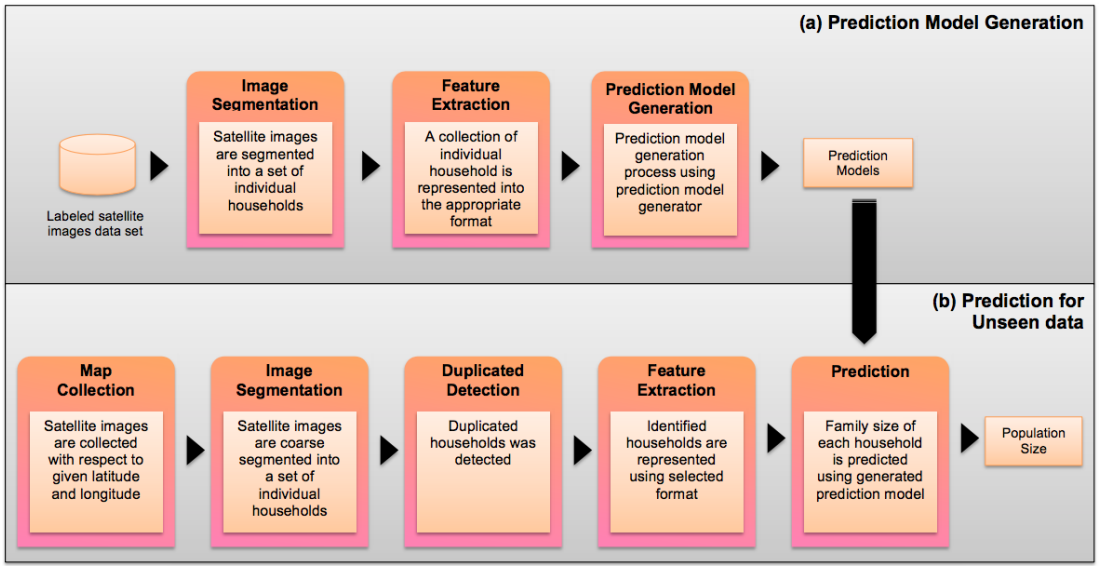


Figure 8.1: Schematic illustrating the proposed large scale population estimation mining process

The rest of this chapter is organised as follows. Steps one to five of the large scale population estimation mining process shown in Figure 8.1 are considered in Sections 8.2 to 8.6. More specifically satellite image collection is discussed in Section 8.2, segmentation in Sections 8.3, duplicate household detection and pruning in Section 8.4, household representation in Section 8.5 and prediction in Section 8.6. Section 8.7 then reports on the evaluation of the proposed large scale population estimation mining approach, followed by some discussion in Section 8.8. Finally, the main findings and some associated conclusions are presented in Section 8.9.

8.2 Satellite Image Collection (Step 1)

This section describes the satellite image collection step within the overall large scale population estimation mining process whereby the satellite image data to be used is acquired and collated. In earlier work presented in this thesis the evaluation was conducted using a set of images obtained from Google Earth (see Chapter 3). These images were hand extracted using the known locations (latitude and longitude) of households with known family size (the Site A and Site B data). However, Google Earth does not readily facilitate the automated extraction of satellite imagery, and clearly the hand-extraction of a large number of satellite images is not possible. One satellite image repository that allows automated extraction of specified images, and that is used with respect to the work presented in this chapter, is the Google Static Map service (other such repositories exist). In the case of the Google Static Map service the associated API (Application Programming Interface) allows users to download satellite imagery

using a variety of parameters. In the context of the proposed large scale population estimation mining process it is suggested that the most appropriate parameters are: (i) opposing top-left and bottom-right corner latitudes and longitudes of the area of interest, (ii) image size and (iii) "zoom level". The image size and zoom level used with respect to the data set used for evaluation purposes (as reported on in Section 8.7 of this chapter) were 1280×1280 pixels and 18 respectively. Note that the zoom level specifies the level of desired detail.

Using the Google Static Map API images were downloaded in an iterative manner commencing at the top-left corner starting with the first 1280×1280 pixel image and then continuing, in a top to bottom and a left to right manner, until the specified area was covered. A similar process can be adopted with respect other satellite image repositories. Note that so as to avoid the potential for households to be split across several satellite images an overlap of 320 pixels was used (larger than the anticipated household size including a margin for error). For this to operate correctly it was necessary to convert the top-left corner latitude and longitude of the current image into x and y pixel values, add the required 960 $(1280 - 320)$ pixel offset, to obtain the top-left x and y coordinates of the next image in the sequence and then to convert these x and y coordinates back to a latitude and longitude to be used by the Google Static Map API to retrieve the next satellite image. The distinction is that the x and y values are 2D planer values while the latitude and longitude values are spherical "surface of the earth" values. The conversion was thus a complex process conducted using the Mercator map projection equations given in Equations 8.1 to 8.4 respectively to convert: (i) from latitudes to x-pixel values, (ii) from x-pixels values to latitudes, (iii) from longitudes to y-pixels values and (iv) from y-pixels values to longitudes.

In this manner a collection of RGB encoded satellite images could be obtained forming a patch work covering a given area. An example fragment of such a patch work is presented in Figure 8.2. As noted previously households typically feature tin roofs which, in reflected sunlight, show up as small white regions within the patchwork. The households can be clearly seen in Figure 8.2. Note also that the patchwork features a significant overlap between images.

$$latitude = (2 \times tan^{-1}(e^{pixels})) - (\frac{\pi}{2}) \tag{8.1}$$

$$pixels = log(tan(\frac{\pi}{4} + \frac{latitude}{2})) \tag{8.2}$$

$$pixels = longitude \times (\frac{2^{zoomlevel} \times 256}{360}) \tag{8.3}$$

$$longitude = \frac{pixels}{(\frac{2^{zoomlevel} \times 256}{360})} \tag{8.4}$$

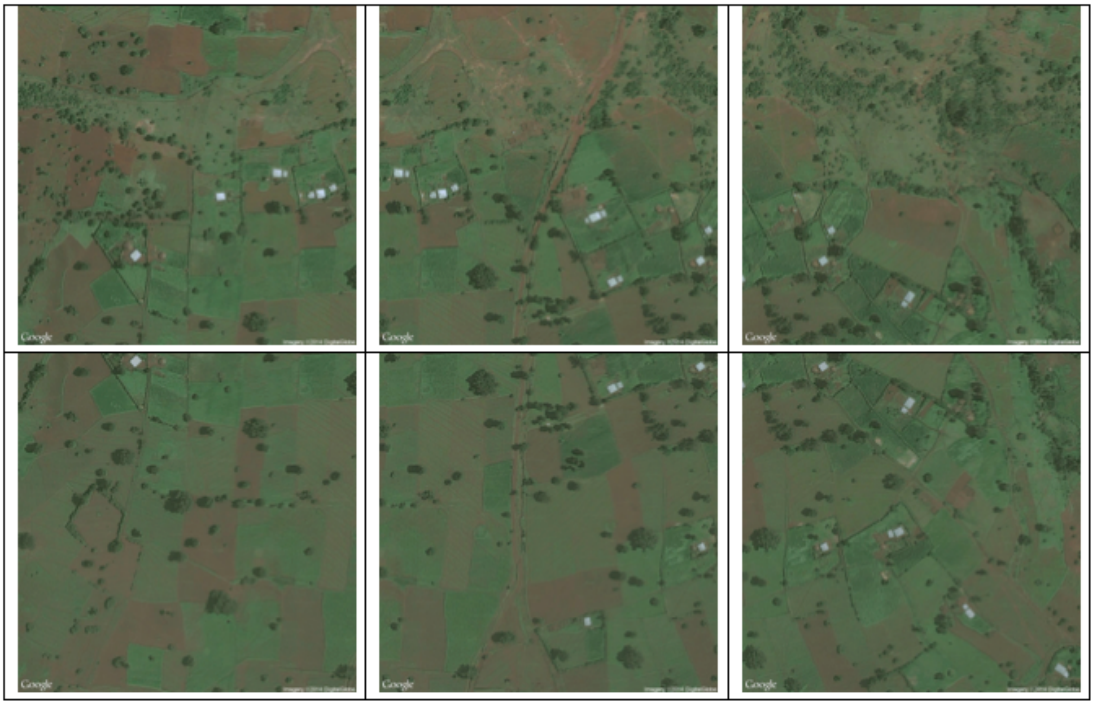


Figure 8.2: An example fragment of a collected "patchwork" of satellite image

8.3 Segmentation (Step 2)

Once a collection of satellite images for a prescribed area has been obtained, using the map collection mechanism presented above, the next step is image segmentation. As described previously the aim of segmentation is to isolate individual households within the data collection. With respect to the segmentation presented in Chapter 3 this was applied to satellite images where the location of the household of interest was known from the start. With respect to the large scale population estimation mining process presented in this chapter this information was unknown. Also any given satellite image could contain zero, one or more households. The proposed segmentation process, although directed at the same goal, therefore operated in a different manner and is thus described in some detail here. Broadly the process used a number of image masks. Experiments (not reported here) were conducted using a variety of image representations and masking techniques, a significant challenge was the illumination of roads and water ways. From these experiments it was found that masks expressed in terms of the HSV colour space produced the best results. It was thus necessary to convert the collected RGB colour space satellite images into the HSV colour space.

The proposed image segmentation algorithm is presented in Algorithm 2. The inputs to the algorithm are: (i) a collection of RGB encoded satellite images (*Images*), (ii) a collection of HSV threshold values and (iii) the top-left corner latitude (*aLat*) and longitude (*aLong*) of the area of interest. The six thresholds used were:

1. Low threshold for Hue channel (lh).
2. High threshold for Hue channel (hh).
3. Low threshold for Saturation channel (ls).
4. High threshold for Saturation channel (hs).
5. Low threshold for Value channel (lv).
6. High threshold for Value channel (hv).

Extensive evaluation was conducted (also not reported here) to identify the most appropriate HSV threshold values so as to best distinguish households from other objects within the satellite image data (such as roads and rivers). The six threshold values arrived at, and used with respect to the evaluation presented later in this chapter, were: (i) $lh = 0.35$, (ii) $hh = 0.65$, (iii) $ls = 0.05$, (iv) $hs = 0.15$, (v) $lv = 0.80$ and (vi) $hv = 1.0$. The output from Algorithm 2 is a collection of household images of the form used with respect to the work presented in earlier chapters.

Referring back to Algorithm 2 the input set of satellite images, *Image*, is processed image by image (line 14). First the top-left x and y coordinates for the current image $i \in Images$ are calculated (lines 15 and 16). The originally RGB image is then converted into a HSV format (line 17). The defined thresholds are then used to create "masked images" (images where pixels with values below or above the specified thresholds are coloured black, the rest white) for each of the three channels in the HSV representation. Thus: (i) a Hue channel masked image (line 18), (ii) a Saturation channel masked image (lined 19) and (iii) a Value channel masked image (line 20). The three masked images were then combined (line 21) to effect noise removal. All being well the resulting binary valued image will indicate household locations as white "blobs" against a "black" background. For each "blob" in the collection of households identified for image i the centroid of each blob was calculated in terms of x and y coordinates (line 24). The image is then "cut out" out off the parent satellite image (line 25) using a 200×200 pixel box centred on the identified blob centroid, to give a *household image* of the form considered earlier in this thesis. Note that:

1. This box size was selected because it more or less conforms to the maximum size of a household image at a Zoom Level of 18 as used with respect to the Google Static Map service adopted with respect to the work presented in this chapter.
2. If an alternative service and/or alternative Zoom Level were adopted these dimensions may have to be adjusted.
3. For households located at the edge of images the resulting box may be smaller than 200×200.

4. Simply imposing a pixel box over a detected centroid avoids issues reported earlier in this thesis (see Section 3.2 of Chapter 3) concerning households that were not rectangular in shape.

Referring back to Algorithm 2, the x and y coordinates of the centroid of each detected household were next converted into latitude and longitude coordinates (lines 26 and 27) using the Mercator projection equations presented earlier, for use later in the process. Each identified household image H is then added to the list of household images collected so far together with the identified latitude and longitude for its centroid (line 28).

Algorithm 2 Image Segmentation

1: **Input:**
2: $Images$ = a set of satellite images each labelled with an index
3: lh= low threshold for Hue channel
4: hh = high threshold for Hue channel
5: ls= low threshold for Saturation channel
6: hs = high threshold for Saturation channel
7: lv = low threshold for Value channel
8: hv = high threshold for Value channel
9: $aLat$ = top-left corner latitude of the area
10: $aLong$ = top-left corner longitude of the area
11: **Output:**
12: A collection of household image files, labeled with latitude and longitude of household
13: $HouseImages = \{\}$
14: **for** $i = 1$ to $|Images|$ **do**
15: $\quad mY = lat2pixel(aLat) - (640 \times (Images_i.row - 1))$
16: $\quad mX = long2pixel(aLong) + (640 \times (Images_i.col - 1))$
17: $\quad hsvImage = rgb2hsv(Images_i)$
18: $\quad hMask = mask(hsvImage, lh, hh)$
19: $\quad sMask = mask(hsvImage, ls, hs)$
20: $\quad vMask = mask(hsvImage, lv, hv)$
21: $\quad hsvMask = hMask \cap sMask \cap vMask$
22: $\quad Households = bwboundaries(hsvMask)$
23: $\quad$ **for** $j = 1$ to $|Households|$ **do**
24: $\quad\quad (x, y) = getCentre(Households_j)$
25: $\quad\quad H = cutOut(Images_i, x, y, 200, 200)$
26: $\quad\quad hLat = pixel2lat(mY - y)$
27: $\quad\quad hLong = pixel2long(mX + x)$
28: $\quad\quad HouseImages = HouseImages \cup \langle H, hLat, hLong \rangle$
29: $\quad$ **end for**
30: **end for**

Figure 8.3 illustrates the operation of Algorithm 2. Figure 8.3(a) shows a given RGB satellite image obtained from the Google Static Map service. Next the RGB image is converted into a HSV colour model image as presented in Figure 8.3(b). Figures 8.3(c), (d) and (e) then show the Hue, Saturation and Value HSV channel masked images. Figure 8.3(f) shows the result from combining the masked images from which individual households can be identified

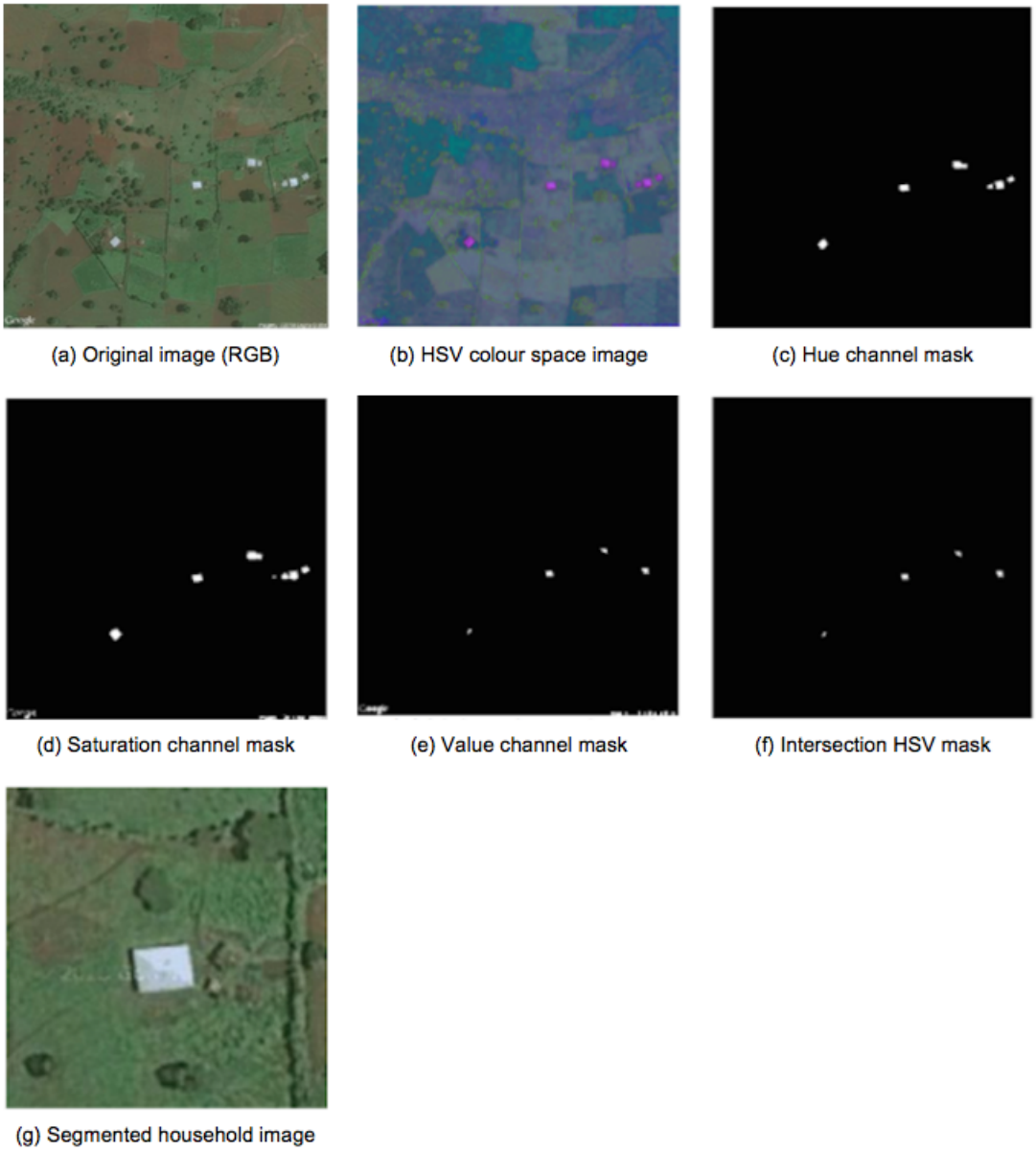


Figure 8.3: Example of the satellite image segmentation process

(white blobs) and cropped from the original satellite image. An example cropped household image is given in Figure 8.3(g).

In the above manner an entire satellite image collection can be processed and a set of household sub-images identified, each labelled with a latitude and longitude identifier. Recall that because the satellite images overlap some households may be duplicated, the process for dealing with such duplicates is discussed in the following Section.

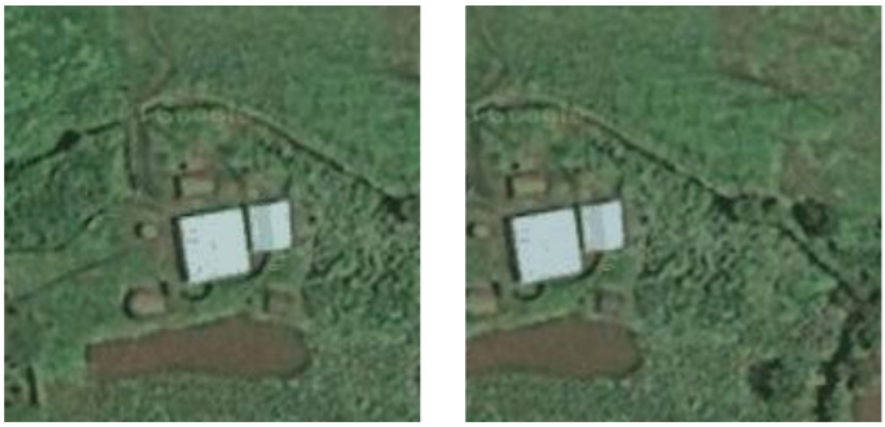
8.4 Duplicated Household Detection and Pruning (Step 3)

Given the nature of the satellite image collection mechanism, designed to ensure that for each households there exists a satellite image in which the household appears in its entirety, each household may appear in two or more images. This phenomena was illustrated in Figure 8.2. Figure 8.4 presents a example of two segmented households, from two different satellite images, which are in fact the same household. Step 3 of the large scale population estimation mining process is thus directed at identifying and pruning duplicates. The mechanism is a follows. The identified households are listed in order of latitude. This list is then processed and households with the same latitude and longitude label (within a level of tolerance) identified. If two households with the same centroid latitude and longitude both comprise 200×200 pixel boxes the later is pruned. If the boxes are unequal the household featuring the smaller sized box is pruned.

Algorithm 3 presents the duplicated household detection mechanism in more detail. The input to the algorithm is the collection of households obtained using Algorithm 2, and the latitude and longitude tolerances. The output is a collection of unduplicated households (H'). Extensive experiments (not presented here) were carried out to find the most appropriate tolerance values with which to identify duplicate households. The best value for the tolerance thresholds was found to be 0.0001. As noted above, the duplicate household detection process commences by ordering the household according to latitude (line 7). The ordered collection of households, H, is then processed household by household. Whenever a duplicate household is found the associated pixel boxes sizes are compared and if different the reference (pointer) in H to the smaller is set to *null*. Otherwise the reference in H for the second household is set to *null*. At the end of the mechanism (line 21) the list H is processed to remove all null references and create the list H' which is used in the following image representation step (Step 4).

8.5 Image Representation (Step 4)

The next stage of the processes is to represent the images in a manner compatible with prediction model generation. Three distinct household image representations were considered earlier in this thesis: (i) graph-based, (ii) colour histogram based and (iii) texture based. A statistical comparison of the three proposed representations was presented in Section 6.7 of Chapter 6. The reported evaluation indicated that the LBP representation produced the best results. This



(a) The detected household from Map_22_6

(b) The detected household from Map_23_6

Figure 8.4: Example of two duplicate households

Algorithm 3 Duplicated House Detection

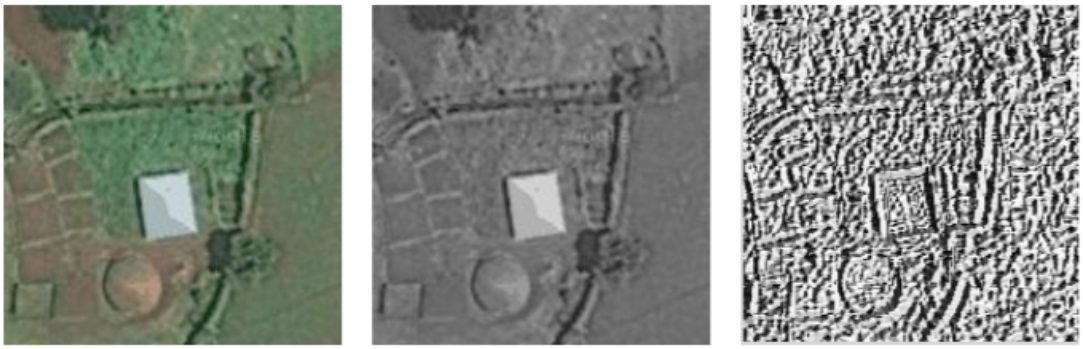
Input:
$HouseImages$ = The set of household images,labeled with latitude and longitude values, generated using Algorithm 2
$latTh$ = latitude tollerance
$longTh$ = longitude tollerance
Output:
H' =a collection of non-duplicated household images

$H =$ the $HouseImages$ set ordered according to latitude
for $i = 0$ to $i = |H| - 1$ **do**
    for $j = i + 1$ to $j = |H|$ **do**
        if $H_j \neq null$ **then**
            if $H_i.lat == H_j.lat \pm latTh$ and $H_i.long == H_j.long \pm longTh$ **then**
                if area $H_i \leq$ area H_j **then**
                    $H_i = null$
                else
                    $H_j = null$
                end if
            end if
        end if
    end for
end for
$H' = H$ with null references removed

was thus the representation recommended with respect to the large scale population estimation mining process presented in this chapter, although alternative representations could equally well have been used. For the evaluation given in Section 8.7 the graph-based representation was also considered.

In the context of the LBP representation proposed in Chapter 6 it should be recalled that to generate a set of LBPs describing an individual household images, the images were first transformed into greyscale. A 3×3 pixel window, with the pixel of interest at the centre, was then used as the basic "neighbourhood" definition with respect to the LBP representation. For each neighbourhood the greyscale value for the centre pixel was defined as the threshold value with which the surrounding eight neighbourhoods were compared. For each neighbourhood pixel a 1 was recorded if the greyscale value of the neighbourhood pixel was greater than the threshold, and a 0 otherwise. The result is an eight digit binary number. In other words 256 (2^8) different patterns can be described. Seven variations for the basic LBP concepts were evaluated in Chapter 6 from which it was concluded that $LBP_{8,1}$ was the most appropriate in the context of the population estimation mining application. This was thus the LBP representation variation used with respect to the evaluation of the proposed large scale population estimation mining process presented later in the chapter (Section 8.7).

A detailed discussed of the mechanism for converting a household image into the $LBP_{8,1}$ format was given in Chapter 6. As a reminder an example is presented here in Figure 8.5. Figure 8.5(a) shows the input image, Figure 8.5(b) the converted greyscale image and Figure 8.5(c) the resulting LBP image.



(a) Segmented household (b) Grayscale image (c) LBP image

Figure 8.5: The LBP operator

With respect to the graph-based representation described in Chapter 4 is should be recalled that the idea was to generate a quadtree for each image; using a decomposition process, and then identifying frequently occurring subgraphs (subtrees) in the graph (tree) collection. These frequent subgraphs were then used to define a feature vector representation. it should be noted that the resulting graph-based classifier should only be applied to unseen data encoded using

the same set of frequent subgraphs.

8.6 Prediction (Step 5)

The final step, Step 5, for the large scale population estimation mining process is prediction where the family size for each household is predicted using a given prediction model (classification or regression). The generation and application of various prediction models was discussed in detail earlier in this thesis. For the evaluation represented later in the chapter in Section 8.7 the best performing classification and regression models, generated using both the Site A and the Site B data, were used. Recall that the distinction between classification and regression is that the first produces categorical labels while the second produces real numbers.

8.7 Evaluation

The evaluation of the proposed large scale population estimation mining process described above is reported on in this section. The section is organised as follows. Sub-section 8.7.1 gives an overview of the test data used, including a discussion of the results from pre-processing the data so as to identify households and the representation of the households (Steps 1, 2, 3 and 4 of the above process). Sub-sections 8.7.2 and 8.7.3 then present the population estimation results obtained from the application of the prediction step (Step 5 of the above process). Sub-sections 8.7.2 considers population estimation mining in the context of classification while Sub-section 8.7.3 considers population estimation mining in the context of regression modelling. In each case the best predictor, as identified in the foregoing chapters, with respect to both the Site A (wet season) and Site B (dry season) test data, was used.

8.7.1 Test Data Collection and Pre-processing

As in the case of the Site A and Site B data, for evaluation purposes satellite image data from the Horro district in the Oramia region of Ethiopia, lying some 300 kilometres to the north-west of Addis Ababa, were again used (see Figure 8.6). This was chosen because: (i) it was the same area from which the Site A and Site B data was extracted from and thus prediction models generated using the Site A and Site B data sets were likely to be applicable, and (ii) it featured a village, and its surrounding lands, that in 2011 was reported to comprise 459 households and a population of $3,223$ (thus "ground truth" data was available). Note that the satellite image data collection was conducted in September 2013 (wet season), two years after the "ground truth" census data had been obtained. The area used measured 42.7 km^2 and was bounded by the parallels of latitude $9°32'55''$N and $9°36'23''$N, and the meridians of longitude $37°6'43''$E and $37°10'23''$E. In total a set of 600 satellite images were collected covering this area, in September 2013, using the Google Static Map API (as discussed in Sub-section 8.2 above). The processing time required to complete the download was 356.11 seconds.

Figure 8.6: Test site location for the evaluation of the proposed large scale population estimation mining process (image from Google Maps)

Table 8.1: Class label Statistics and distribution for the Site A and Site B data sets (copy of Table 3.1 from Chapter 3)

Family Size	Minimum	Maximum	Average	Mode	Site A	Site B
Small	2	5	4.04	5	28	19
Medium	6	8	7.00	6	32	21
Large	9	12	9.80	9	10	10
Total 120	2	12	6.31	6	70	50

Using the mechanism presented in Sub-section 8.3 above a total of 526 households were initially detected (including duplicates). The runtime required to process the 600 satellite images and detect households was 1,370.16 seconds (thus 2.28 seconds per satellite or 2.60 seconds per detected household).

The duplicated household detection mechanism presented in Sub-section 8.4 was then applied as a result of which 100 duplicate households were detected. In other words 426 out of the known of 459 households were detected (an accuracy of 92.8%). The suggested reasons for the discrepancy were as follows:

1. There was a two year time difference between the date of the "ground truth" survey; a period during which some households may have fallen into disuse. Manuel inspection of a proportion of the collected satellite images indicated that some households did indeed appear to be roofless thus supporting this conjecture.

2. Further inspect of the satellite imagery indicated that a small number of buildings were very poorly defined and in some cases had not been segmented correctly.

3. It was also possible that the duplicate household detection mechanism had detected some duplicates that were in fact not duplicates (although no evidence for this was found).

Whatever the case the processing time required for the duplicate household detection process was 0.40 seconds, this speed of computation was attained because the required parameters (latitude and longitude) had already been generated as part of the segmentation process applied earlier.

The collection of 426 household images were then each represented using the $LBP_{8,1}$ and graph-based representations and the set of feature vectors generated compatible with the prediction models. The processing time for these were 60.20 and 413.45 seconds, respectively.

8.7.2 Classification based Large Scale Population Estimation Mining

This sub-section considers the evaluation of the proposed large scale population estimation mining process in the context of classification. Recall from Chapter 3 that with respect to the classification models considered the household collected data was separated into three classes: (i) "small family size" (2-5 people in family), (ii) "medium family size" (6-8 people in family)

Table 8.2: Estimation of population size with respect to the Neural Network classification model generated using the Site A data set

Family Size	Average	Predicted no households	Estimated population size
Small	4.04	156	630
Medium	7.00	261	1,827
Large	9.80	9	88
	Total	426	2,545

and (iii) "large family size" (9-12 people in family). Some statistics concerning the class distributions for the Sites A and B data were presented in Table 3.1 in Chapter 3. For reasons of reader convenience these are presented again here in Table 8.1.

Previously (Sub-section 6.7) it was noted that the best performing classification models, with respect to the Site A and B data sets respectively, were: (i) the Neural Networks classification model used in conjunction with Chi-Squared feature selection (using $k = 40$) and the $LBP_{8,1}$ representation; and (ii) the Bayesian Network model couple with Gain Ratio feature selection (using $k = 55$) and the graph-based representation. These were then the two classification models used with respect to the evaluation presented in this sub-section. Recall that the test data satellite images for the evaluation of the proposed large scale population estimation mining process were originally recorded during the wet season.

Neural Networks classification with Chi-Squared feature selection and the $LBP_{8,1}$ representation (generated for Site A).

Using the Neural Network classifier generated using the Site A data the time taken for prediction was 2.00 seconds. Table 8.2 shows the results obtained. From the table it can be seen that 156 small households, 261 medium households and 9 large households were identified. Using the data presented in Table 8.2 the population estimation was conducted as follows. The average values from each class (describing by the class label statistics from Table 8.1) was used: 4.04, 7.00 and 9.80 for small, medium and large family size, respectively. As a result an estimated population size of $2,545$ was obtained, compared to a known population size of $3,223$, thus an accuracy of 78.96%. An overall processing time, including the processing time required for steps 1 to 4, of 29.49 minutes was recorded.

Bayesian Network classification with Gain Ratio feature selection and the graph-based representation (generated for Site B data).

When using any classification model, as noted previously, the representation used with respect to the data used to generate the classifier must be identical to that used to represent the unseen data to which the classifier is to be applied. In the case of the graph-based representation this means that the same global set of subgraphs needs to be used. For the evaluation presented here, so that a shared set of frequent subgraphs could be simply identified, the 50 training

Table 8.3: Estimation of population size with respect to Bayesian Network classification model generated using of the Site B data set

Family Size	Average	Predicted no households	Estimated population size
Small	4.04	226	7913
Medium	7.00	135	945
Large	9.80	65	637
	Total	426	2,495

images from the Site B data set and the 426 identified "unseen" households images for the large scale study were pooled give a data set of 476. Note that within this pooled data set only only the Site B images had family size class labels associated with them. The pooled set of images were decomposed using the quadtree decomposition described in Section 4.2 of Chapter 4. A graph-based representation was then generated as described in Section 4.3 in Chapter 4, one graph per image. Frequent subgraph mining was then applied in the same manner as described previously with $\sigma = 10$ (a subgraph is considered to be frequent if it occurs in at least 42 household graphs from the total 476 available graphs). Consequently a collection of subgraphs was identified, to which feature selection was applied and a set of feature vectors generated (one per household image). Feature selection was conducted using the Gain Ratio feature selection strategy with $k = 55$ (recall that this produced the best classification model as reported in Chapter 4 and 6.7 of Chapter 6). The feature vectors for the Site B images, the training data, were then used to generate a Bayesian Network classification model. The classification model was then tested using the Site B data again and an AUC of 0.828 obtained.

The Bayesian Network classification model was then used to evaluate the proposed large scale population estimation mining process presented in this chapter using the unseen part of the pooled data. The time taken for prediction was 2.00 seconds. The results obtained are presented in Table 8.3. From the table it can be observed that 226 small households, 135 medium households and 65 large households were identified. Using the same process as used in the case of the evaluation using the Neural Network classification model described above, an estimation population size of $2,809$ was obtained, compared to a known population size of $3,223$; thus an accuracy of 77.41%. An overall processing time of 29.37 was recorded.

8.7.3 Regression based Large Scale Population Estimation Mining

From earlier work on regression presented in Chapter 7 the most effective regression models were generated using the $LPB_{8,1}$ representation, a CFS feature selection strategy and SVMreg regression analysis for both the Site A and the Site B data sets. This the regression models were thus used with respect to the evaluations of the proposed large scale population estimation mining process presented in this sub-section.

SVMreg regression with CFS feature selection and the $LPB_{8,1}$ representation (generated

for Site A data).

The regression equation generated for the Site A data set (wet season) using SVMreg regression analysis with CFS feature selection and the $LPB_{8,1}$ representation was given in Equation 7.1 in Section 7.3 of the previous chapter, for convenience this is given again in Equation 8.5 below. Applying this regression equation the required processing time was 1.00 second. As a result an estimation population size of $2,548$ was obtained, compared to a known population size of $3,223$, therefore an accuracy of 79.06% was obtained. The overall "end-to-end" processing time was 29.48 minutes.

$$\begin{aligned} \textit{Family Size} = \{&-(0.1085)(\textit{normalised})LBP_{8,1}42 + (0.0645)(\textit{normalised})LBP_{8,1}86 \\ &+(0.2017)(\textit{normalised})LBP_{8,1}102 - (0.2070)(\textit{normalised})LBP_{8,1}106 \\ &-(0.3565)(\textit{normalised})LBP_{8,1}110 + (0.0899)(\textit{normalised})LBP_{8,1}118 \\ &+(0.1190)(\textit{normalised})LBP_{8,1}150 - (0.0453)(\textit{normalised})LBP_{8,1}155 \\ +(0.1881)&(\textit{normalised})LBP_{8,1}165 - (0.1432)(\textit{normalised})LBP_{8,1}171 + 0.4242\} \end{aligned} \tag{8.5}$$

SVMreg regression with CFS feature selection and the $LPB_{8,1}$ representation (generated for the Site B data set).

The regression equation generated for the Site B data set (dry season) using SVMreg regression analysis with CFS feature selection and the $LPB_{8,1}$ representation was presented in Equation 7.2 in Section 7.3 of the previous chapter, but for reasons of reader convenience this is again presented here in Equation 8.6, below. The run time required for the application of Equation 8.6 was 1.00 seconds. As a result an estimated population size of $2,760$ was obtained; representing an accuracy of 85.63%. An overall processing time of 29.48 minutes was recorded. Again this was an interesting result as the regression model generated using dry season data produced a best performance although applied to wet season data.

$$\begin{aligned} \textit{Family Size} = \{&(0.2547)(\textit{normalized})LBP_{8,1}11 + (0.0836)(\textit{normalized})LBP_{8,1}43 \\ &+(0.1408)(\textit{normalized})LBP_{8,1}44 - (0.3142)(\textit{normalized})LBP_{8,1}107 \\ &-(0.3241)(\textit{normalizedLBP}_{8,1}163 + (0.4095)(\textit{normalized})LBP_{8,1}173 \\ +(0.1126)&(\textit{normalized})LBP_{8,1}211 + (0.2103)(\textit{normalized})LBP_{8,1}219 + 0.3796\} \end{aligned} \tag{8.6}$$

The time taken for the application of Equation 8.6 to the $LPB_{8,1}$ feature vector represented household data was 1.00 seconds. As a result on estimated population size of $2,760$ was obtained; representing an accuracy of 85.63%. The overall processing time of 29.39 minutes was recorded.

Table 8.4: Summary of evaluation results for proposed large scale population estimation mining process

Prediction Model	Population Estimation	Accuracy (%)	Total run time (minutes)
Neural Networks classification with Chi-Squared feature selection and $LBP_{8,1}$ (Site A wet season data)	2,545	78.96	29.49
Bayesian Network classification with Gain Ratio feature selection and graph-based representation (Site B dry season data)	2,495	77.41	35.42
SVMreg regression with CFS feature selection and $LPB_{8,1}$ representation (Site A wet season data)	2,548	79.06	29.48
SVMreg regression with CFS feature selection and $LPB_{8,1}$ representation (Site B dry season data)	2,760	85.63	29.48

8.8 Discussion

A summary of the evaluation results presented above is given in Table 8.4. From the table it can be seen that when regardless of whether a classification or regression based prediction approach was adopted withy respect to the proposed large scale population estimation mining models generated using the dry season data produced the better performances than models generated using wet season data despite the fact that wet season data was used. It is thus concluded that weather wet or dry season data is used is not as significant as originally anticipated. The best performing model was the SVMreg regression when coupled with CFS feature selection and the $LPB_{8,1}$ household satellite image representation: a population size estimation of $2,760$ compared to a known population size of $3,223$, an accuracy of 85.63%. The processing time required in each case was about 30 minutes, most of this resulted from data collection and pre-processing, the choice of prediction model did not have a significant impact on processing time.

The results obtained seem to contradict the conclusions drawn at the end of Chapter 7. However it should be noted here that, as reported above, the best performing classifier had to be regenerated to take into account the frequent subgraphs occurring in the unseen test data, which might make a difference. There are number of conjectured reasons why the results obtained might not reflect the "ground truth" survey as closely as we would like (other than the limitations of the proposed large scale population estimation mining process):

1. The data from which the classification and regression models were generated might not reflect the data to which they were applied as closely as was anticipated (except in the case of the graph-based representation where the model was regenerated). Measures for determining the similarity between satellite image data sets are a subject for future work

(this is discussed further in the following concluding chapter).

2. The satellite images data use for the Site A and Site B data sets, used to train the classification/ prediction models were taken from Google Earth service while those used for the large scale study taken from Google Static Map service. This may have had some effect on prediction accuracy.

3. As already noted above, there was a two year time lag between the date of the census collection (2011) and the date of the satellite image extraction process being applied (September 2013). Manual inspection of a number of images showed signs of derelict (abandoned) households. It may thus be the case that between 2011 and 2013 depopulation had taken place and that the produced population estimates were in fact a better reflection of population size than initially thought. There have been recent reports of the depopulation in rural Ethiopia, see for example [67].

4. Census collection is often viewed with suspicion. Local authorities may suspect that it is to be used for the levying of a local tax and thus there may be an incentive to under report population size. Alternatively it may be suspected that the census is to be used for allocating development grants in which case there may be an incentive to over report population size.

Whatever the case, although (at face value) the population estimations produced were not as accurate as the "ground truth" census data (this was to be expected), the proposed method offered significant cost and time savings. Overall processing times of about 30 minutes was recorded with respect to both classification and regression approaches, as opposed to the many days that would be required to conduct the original survey using traditional methods.

8.9 Summary

A large scale population estimation mining process has been presented in this chapter. The process is founded on work presented earlier in this thesis. The results obtained suggested that the graph-based representation should be adopted couple with the use of classification model (although the proposed process can clearly also operate using other representations and prediction models). Also of note, in the context of the proposed process, are the mechanisms for collecting satellite data and pre-processing this data (duplicate detection and pruning). Evaluation was conducted using a collection of 600 satellite images covering a 42.7 km^2 square area centred over a village in rural Ethiopia for which the population size was known. For the evaluation the best performing classification and regression models from earlier work presented in this thesis were used.

Population estimations of $2,545$ and $2,495$ were obtained with respected to the classification models generated using the Site A and B data respectively. Whereas population estimations

of 2,548 and 2,809 were obtained with respected to the regression models generated using the Site A and B data respectively. These estimations compared favourably with the known population size of 3,223, accuracies of 78.96%, 77.41%, 79.06% and 85.63% respectively. The overall processing time was about 30 minutes for both techniques. The results indicated that by using the proposed framework effective population estimates can be obtained, in rural areas, at very low cost (almost zero). The following chapter concludes this thesis with a review of the work presented, the main findings and consideration of opportunities for future work.

Chapter 9

Conclusion

A summary of the proposed population estimation mining using satellite imagery approaches and processes, the main findings, the research contributions and directions for possible future work are presented in this chapter. Section 9.1 gives a summary of the proposed population estimation research presented in this thesis. The main findings and research contributions of the research work are presented in Section 9.2. Finally some directions for future research are presented in Section 9.3.

9.1 Summary

Three different approaches were proposed in this thesis to categorise satellite imagery with respect to the nature of individual household images together with a process for applying these approaches in a large scale setting. In the context of the approaches, family size prediction in terms of both classification and regression were considered. For the evaluation of the three approaches two different data sets, obtained from two different test sites, were used. In each case an image segmentation process was applied so as to isolate individual household satellite images. Because the size and amount of image data was too large to be used directly with respect to the application of household size prediction, an alternative representation was required that served to capture the key properties or characteristics of individual household images but in a reduced form. For each approach this was done in a different manner: graph-based, colour histogram based and texture based. For prediction model training purposes each represented household had a "family size" class label associated with it. This training data was then used to build a predictor of some kind. Both classification and regression prediction models were considered.

The first approach was founded on the concept of the graph mining. An image decomposition was applied whereby the individual households were represented using quadtrees. However, the quadtree representation does not lend itself to ready incorporation with respect to classification algorithms. Consequently is was proposed that some form of subgraph mining be applied to the quadtree data so as to identify frequently occurring subgraphs that can be used as features in the context of a feature vector representation. The identified frequent

subgraphs were viewed as defining a feature space which could be used to represent the image set. A given image set could thus be recast into this format so that each image is represented by a feature vector whose elements are some subset of the global set of identified frequent subgraphs making up the feature space. Standard prediction model generation techniques can then be applied to build a classification or regression model that can be applied to unseen data. The reported evaluation indicated that best classification accuracy results were obtained when using low support thresholds for identifying frequently occurring subgraphs.

The second approach was founded on the concept of image colour analysis by representing the distribution of colours using histograms. Furthermore some basic colour statistical measures were also generated to provide additional colour information. Once the histograms and colour statistical measures had been generated (typically seven histograms per household) together with an associated family size, this data could then used to construct a classifier/regression model which could then be used to predict the family size of previously unseen household images according to the nature of the colour representation of these unseen images. The reported evaluation indicated that high classification accuracy results were obtained when using only colour histograms (without any additional colour statistical information).

The third approach was founded on methods used in image texture analysis. More specifically the use of Local Binary Patterns (LBPs). A LBP is a texture representation method which is both statistical and structural in nature. Using the LBP approach a binary number is produced, for each pixel, by thresholding its (greyscale or intensity) value with its neighbouring pixels. The basic idea proposed was to use the LBP concept to represent each segmented household. As in the case of the previous approaches, for prediction model generation purposes each LBP segmented household image also had a family size associated with it. In addition the use of some statistical information concerning texture across a household image was also included. The reported evaluation indicated that high classification accuracy results were obtained when using *LBPs* with a radius of 1 than *LBPs* with radiuses of 2 and 3,

Experiments were initially conducted using classification models as the prediction mechanism, in other words using categorical labels for family sizes (small, medium, large). This established that the texture based approach produced the best overall performance. This representation was then used to evaluate the use of regression models to estimate family size from household images, in other words using real values for family sizes. Experiments indicated that regression tended to produce the effective family size prediction.

The work presented in this thesis was completed with an investigation into how the work discussed earlier in the thesis could best be applied in the context of large scale population estimation mining. A process was presented for achieving this which was tested using a 42.7 km^2 region of Ethiopia for which the population size was known. The experiments conducted indicated that good results could be obtained.

9.2 The Main Findings and Research Contributions

The research presented in this thesis was directed at providing an answer to the research question presented in Chapter 1, namely:

What are the most appropriate end-to-end computational processes required to collect population census data from satellite imagery using classification and regression techniques?

This research question had a number of research issues associated with it that required resolution before an answer to the central research question could be derived. This section presents an overview of the main findings, and the research contributions, of the work presented in this thesis with respect to the above research question and associated research issues. The section is organised by considering each of the research issues itemised in Chapter 1 in turn and then returning to the research question.

1. **What are the most appropriate mechanisms for segmenting a given satellite image so that appropriate individual household sub-images (if any) can identified?**

   Two categories of segmentation were considered in the thesis. The first was used to process satellite images, in the context of the provision of training and/or test data, where the location of the household was known. The second was used where the location was not known and first had to be established. The first was considered in Chapter 3 and comprised a complex process restricted to the identification of rectangular shaped households (it is acknowledged that this was a limitation but still effective in terms of the evaluation for which the resulting household image data was used). The second was presented in Chapter 8 and consisted of simply surrounding identified household locations with a bounding box dimensioned so as to encompass the domain of individual households regardless of shape. The second can be argued to be the "most appropriate mechanism" for segmenting satellite images so that appropriate individual household sub-images can identified because of its simplicity and its consequent general application.

2. **Given a set of identified household images how should the content of those images be represented so that compatibility with classification and regression generation is achieved while at the same time ensuring that key information is retained?**

   In Chapter 4, 5 and 6 three different representations were considered: graph-based, colour histogram based and texture based. The aim of each representation was to capture the salient elements of individual households in the best way possible so as to facilitate household size prediction (in terms of number of people). In each case the representation was eventually translated into a feature vector representation compatibility with most classification and regression generation approaches. Each representation used par-

ticular mechanisms to retain key information, the most effective representation in this context was established by conducting a series of experiments using prelabelled training and test data. The most effective representation was found to be LBP representation.

3. **When representing household images what is the nature of the key information to be captured?**

   The nature of the key information to be captured was not identified specifically. Instead a number of image representations were considered, as noted above, and whether they succeeded in capturing key information or not was established through an evaluation process. The intuition here was that the best performing representation (in terms of prediction) would also be the representation that best served to capture key household image data without specifically identifying what the key information was (if any).

4. **What are the most appropriate classification/regression techniques for predicting census data given a processed collection of household images?**

   The three proposed representations were initially evaluated using classification model generators. This evaluation established that the LBP representation was the most effective (as noted above) together with Chi-Squared feature selection and $k = 40$ and Neural Networks classification. This representation was then used in the context of regression model generation where the best performance was obtained using the LPB representation coupled with the CFS feature selection and SVMreg regression model generation. Overall, the regression was found to outperform classification, the conjectured reason for this was that classification operated using categorical labels while regression operated using real number values.

5. **What is the process for conducting a large scale census comprising many satellite images?**

   The proposed unified process for large scale population estimation mining using satellite imagery was presented in Chapter 8. This was an five steps process comprising: (i) map collection, (ii) segmentation, (iii) duplicated household detection and pruning, (iv) image representation and (v) prediction. The evaluation of the proposed process, using a region of Ethiopia where the population size was known demonstrated that the process worked well.

6. **In the context of conducting large scale surveys how can issues associated with "overlapping" satellite images best be resolved?**

   An issue with the large scale process of population estimation mining was that, so as to ensure no households were missed, it was necessary to overlap satellite images. This in turn meant that households might appear in more than one image. A process for dealing with this "overlap" was presented in Section 8.4 of Chapter 8, this was referred to as

the duplicate household detection and pruning process. Experiments indicated that the process seemed to work well.

Returning to the initial research question, the most appropriate end-to-end computational processes required to collect population census data from satellite imagery, using classification and regression techniques, is founded on the a process that encompasses: (i) a process for collecting a sequence of satellite images over a specified area, (ii) a household detection algorithm founded on the usage of masks to isolate individual households (identifiable by their distinctive roof colour), (iii) a simple segmentation technique found on a bounding box concept, (iv) application of a duplicate household detection and pruning process, (v) representation of individual households using the LBP texture based and graph-based representations (two of the three representations considered) and (vi) household "family size" prediction using classification/regression analysis. The experimental results indicated that good estimates of population size could be obtain at very little cost.

The primary contributions of the research work presented in this thesis were presented in Section 1.4 of Chapter 1, for convenience they are again presented below. Noted that in each case the relevant chapter where the contribution was establishes is given in parenthesis.

1. A novel approach for image segmentation specifically designed for segmenting individual households featured in a satellite image data set (Chapter 3).

2. A household image representation founded on a quadtree based hierarchical decomposition of space together with a frequent subgraph mining algorithm for dimensionality reduction. The identified frequent subgraphs were arranged into a feature vector format, one vector per household, suited for input into a classification or regression model generation algorithm (Chapter 4).

3. A household image representation founded on a colour histogram based approach. More specifically an image representation founded on multiple histograms extract from various colour channels; a feature vector format was again used (Chapter 5).

4. A household image representation founded on the concept of "texture" analysis. More specifically usage of Local Binary Patterns (LBPs), as before a feature vector format was again derived (Chapter 6).

5. A detailed comparison of the proposed household image representations (Section 6.7 of Chapter 6).

6. An analysis of a sequence of classifier generation algorithms so as to identify the most appropriate in the context of population estimation prediction from satellite data (Section 6.7 of Chapter 6).

7. An analysis of a number of regression model generation algorithms so as to identify the most appropriate in the context of population estimation prediction from satellite data (Chapter 7).

8. An effective mechanism for satellite image collection using the Google Static Maps service to obtain satellite image data for a specified area (Section 8.2 of Chapter 8).

9. A novel approach for household detection specifically designed for the purpose of identifying and segmenting individual households featured in a satellite image data set covering a prescribed area Section 8.3 of Chapter 8.

10. A mechanism for detecting duplicated households in a given satellite image data collection so as to address the image "overlap" problem (Section 8.4 of Chapter 8).

11. An end-to-end process for conducting large scale population estimation mining using satellite data (Section 8.7 of Chapter 8).

12. Overall, the thesis presents an approach of population estimation founded on known techniques, but combining a new and novel methodology (entire this thesis).

9.3 Future Works

The research described in this thesis has "sparked" a number of promising directions for future research. In the concluding section of this chapter, and this thesis, these future research directions are briefly presented as follows:

1. **Weighted frequent subgraph mining.**

   In the context of the graph-based representation (Chapter 4) frequent subgraph mining was applied to identified frequently occurring subgraphs (subtrees) which were then used to define feature vectors. However, the subgraph mining process operates by allocating a count of 1 if a subgraph appears in a household image graph; it takes no account of the number of times it appears, only if it appears or not. A mechanism for addressing this is to adopt what is known as weighted subgraph mining. One such algorithm is the Average Total Weighting (ATW) algorithm [89]. It is thus suggested that weighted frequent subgraph mining may enhance the population estimation mining process.

2. **Colour histogram with different numbers of bin.**

   Using the colour histogram based representation (Chapter 5) a bin size of 32 was used because intuitively this seemed like an appropriate size. However, it might be interesting to consider the effect of population estimation accuracy when using different numbers of bin sizes such as (say) 1, 2, 4, 8, 16, 64 and 128. Recall that the usage of colour histogram has some limitation; namely that no spatial information is provided. It is therefore suggested that by including spatial information in the histogram representation might have

a positive effect on prediction. Once method whereby this can be achieved is to use the Colour Coherence Vector (CCV) representation [139]. The CCV representation is more detail than the colour histogram representation. In each colour bin the image pixel are classified into two categories: *coherent* and *incoherent* pixels. Coherent pixel are pixels which are the member of large colour regions (continuous regions), while incoherent pixels are not [140].

3. **Local Binary Pattern with different extensions.**

   In the context if the texture based representation (Chapter 6) the classical LBP and its variations were applied. The classical LBP representation operates based on greyscale images only. The application of colour LBP [12] might therefore prove more effective as this might serve to capture more information colour LBP are obtained by calculating the LBP over all three channels of the RGB or HSV colour space independently, and then concatenating the results together [167]. Therefore the RGB-LBP could be applied to RGB household satellite images to extract the useful information for prediction analysis process.

4. **Compound image representation.**

   In this thesis the three representation described in Chapters 4, 5 and 6 were considered in isolation, it might be effective if the three different representations were combined to form a hybrid representation. The intuition here is that some representations might be better at describing particular features. This intuition is supported by the fact (established in Chapter 4) that the graph-based representation seemed to be able to cope well with dry season data (Site B) where the other two representations seemed to be able to cope better with the wet season data (Site A).

5. **Other image representation.** It might also be beneficial to consider alternative approaches whereby households can be represent. For example in [198] what is referred to as "Vertex Unique Labelled Subgraph" (VULS) mining was used to classify satellite images according to ground type, it might be interesting to apply the VULS concept to the representation of household images for the purpose of population estimation mining. Another candidate representation is the point series approach presented in [43] where point series are used to describe local geometries in the context of a sheet metal forming application; the technique could also be used with respect to the representation of household images for the purpose of population estimation mining.

6. **The segmentation process.**

   The segmentation process described in Chapter 3 was directed at rectangular shaped households; this was because, for reasons of expedience, a line detection technique was adopted. This in turn meant that the generated classification and regression models were trained using images representing rectangular households only, although later applied

to images featuring households of any shape. It is conjectured that more effective classification/regression models might be generated if trained using training data featuring households of any shape. To achieve this different segmentation techniques would need to be adopted, one option would be the Meanshift segmentation approach [100, 175].

7. **Experimentation with alternative data sets.**

   The evaluation described in this thesis has been focused on a rural area of Ethiopia. Although it is acknowledged that the technique is unsuited to urban areas, the technique should be equally applicable to rural areas in many other parts of the world. It would thus be desirable to conduct further experiments and evaluations with respect to rural areas in other geographical locations.

8. **Mechanisms for determining the appropriateness of a prediction model**

   In Chapter 8 it was noted that one of the reasons that an exact population estimation was not achieved, with respect to the evaluation of the proposed large scale population estimation mining process, was that the classification/regression models used (trained on rectangular household data) might not be ideally suited to the area to which it was applied. It would therefore be useful if some mechanism were available where by the "goodness of fit" of a prediction model to a previously unseen data collection could be estimated. How this would operate is a matter for further research.

9. **Explanation Generation**

   All the proposed population estimation mechanism operate in the form of a "black box"; data goes in and an answer comes out, there is no indication of how the answer was arrived at. It would be interesting to know what elements of a household image were indicative of family size, in other words provision of an explanation of the reason for a particular prediction would be useful. Explanation generation, in the context of population estimation mining from satellite imagery, would thus also be useful avenue for further work.

10. **The regularisation**

    In the context of regression analysis (Chapter 7) a collection of significant features was identified using CFS feature selection. This was essentially a form of regularisation mechanism. Typically regularisation methods are used to prevent a regression model from overfitting the training data by penalising variables with extreme parameter values [177]. An investigation of the usages of some regularisation mechanism would be a further fruitful avenue for further research.

11. **The underestimation studies**

    For all the proposed population estimation mechanisms discussed in this thesis the predicted values were found to be underestimations. Possible reasons for this were discussed

in Section 8.8 of Chapter 8, namely that it might be because of depopulation in rural areas of Ethiopia [67]. An alternative reason is some flaw in the proposed mechanisms. A study of the statistical effectiveness of the proposed estimation processes would therefore be another further potential avenue for future research.

In conclusion the work presented in this thesis has demonstrated that it is possible to effectively estimate population size within a given rural area using satellite imagery at a much reduced cost than that which would be required if a traditional form of census was conducted.

Bibliography

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the Twentieth International Conference on Very Large Data Bases (VLDB)*, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[2] A.S. Alsalman and A.E. Ali. Population estimation from high resolution satellite imagery : A case study from khartoum. *Emir*, 16(1):63–69, 2011.

[3] S. Amaral, A. V. M. Monteiro, G. Câmara, and Quintanilha. DMSP/OLS night time light imagery for urban population estimates in the Brazilian Amazon. *International Journal of Remote Sensing*, 27(5):855–870, March 2006.

[4] K. Aneja, F. Laguzet, L. Lacassagne, and A. Merigot. Video-rate image segmentation by means of region splitting and merging. In *Proceedings of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 437–442, November 2009.

[5] N.K. Anitha, G. Keerthika, M. Maheswari, and J. Praveena. A framework for medical image classification using soft set. In *Proceedings of the Second International Conference on Current Trends in Engineering and Technology (ICCTET)*, pages 268–272, July 2014.

[6] P. Arellano, K. Tansey, H. Balzter, and D.S. Boyd. Detecting the effects of hydrocarbon pollution in the Amazon forest using hyperspectral satellite images. *Environmental Pollution*, 205:225 – 239, 2015.

[7] S. L. Arlinghaus, editor. *Practical handbook of curve fitting*. CRC Press, 1994.

[8] B. Attachoo and P. Pattanasethanon. A new approach for colored satellite image enhancement. In *Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBI)*, pages 1365–1370, February 2009.

[9] B. Baesens, T.V. Gestel, S. Viaene, M. tepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

[10] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111 – 122, 1981.

[11] B. Banerjee, V.G. Surender, and K.M. Buddhiraju. Satellite image segmentation: A novel adaptive mean-shift clustering based approach. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4319–4322, July 2012.

[12] S. Banerji and C. Verma, A.and Liu. Lbp and color descriptors for image classification. In *Cross Disciplinary Biometric Systems*, volume 37 of *Intelligent Systems Reference Library*, pages 205–225. Springer Berlin Heidelberg, 2012.

[13] M. Belahcene, A. Chouchane, M. Amin Benatia, and M. Halitim. 3d and 2d face recognition based on image segmentation. In *Proceedings of the International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, pages 1–5, November 2014.

[14] M. Best and N. Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1-3):425–439, 1990.

[15] X Bin and Z. Cang. The application of multiple regression analysis forecast in economical forecast: The demand forecast of our country industry lavation machinery in the year of 2008 and 2009. In *Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining (WKDD)*, pages 405–408, January 2009.

[16] B.A. Bradley, R.W. Jacob, J.F. Hermance, and J.F. Mustard. A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite NDVI data. *Remote Sensing of Environment*, 106(2):137–145, 2007.

[17] M. Brejl and M. Sonka. Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples. *IEEE Transactions on Medical Imaging*, 19(10):973–985, October 2000.

[18] Otto Bretscher. *Linear Algebra with Applications (3rd Edition)*. Prentice Hall, July 2004.

[19] C.E. Bulgin, S. Eastwood, O. Embury, C.J. Merchant, and C. Donlon. The sea surface temperature climate change initiative: Alternative image classification algorithms for sea-ice affected oceans. *Remote Sensing of Environment*, 162:396 – 407, 2015.

[20] J Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, June 1986.

[21] S. Chang and A. Hsu. Image information systems: where do we go from here? *IEEE Transactions on Knowledge and Data Engineering*, 4(5):431–442, October 1992.

[22] T. Chang and C.-C.J. Kuo. A wavelet transform approach to texture analysis. In *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 661–664, March 1992.

[23] A. Charnes, E.L. Frome, and P.L. Yu. *The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family*. Lancaster Press, 1976.

[24] H. Chen and Z. Huang. Medical image feature extraction and fusion algorithm based on K-SVD. In *Proceedings of the Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, pages 333–337, November 2014.

[25] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 241–248, December 2013.

[26] L. Cheng, Y. Zhou, L. Wang, S. Wang, and C. Du. An estimate of the city population in China using DMSP night-time satellite imagery. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, pages 691–694, 2007.

[27] C. Chiu and C. Wu. Texture classification based low order local binary pattern for face recognition. In *Proceedings of the Eigthteenth IEEE International Conference on Image Processing (ICIP)*, pages 3017–3020, September 2011.

[28] R. S. Choras. Feature extraction for CBIR and Biometrics applications. In *Proceedings of the Seventh WSEAS International Conference on Applied Computer Science (ASC)*, pages 1–9. World Scientific and Engineering Academy and Society (WSEAS), 2007.

[29] D. J. Cook and L.B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research (JAIR)*, 1(1):231–255, February 1994.

[30] B.B. Damodaran and R.R. Nidamanuri. Dynamic linear classifier system for hyperspectral image classification for land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2080–2093, June 2014.

[31] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1âĂŽÃĎÃň4):131 – 156, 1997.

[32] R. Dass and S. Devi. Image segmentation techniques. *International Journal of Electronics and Communication Technology*, 3:66–70, January-March 2012.

[33] E.R. Davies. *Machine Vision: Theory, Algorithms, Practicalities*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[34] J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the Twenty-third International Conference on Machine Learning (ICML)*, pages 233–240, New York, USA, 2006. ACM.

[35] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research (JMLR)*, 7:1–30, December 2006.

[36] K. Dittakan, F. Coenen, and R. Christley. Towards the collection of census data from satellite imagery using data mining: A study with respect to the ethiopian hinterland. In M. Bramer and M. Petridis, editors, *Research and Development in Intelligent Systems XXIX. Incorporating Applications and Innovations in Intelligent Systems XX Proceedings of (AI-2012), The Thirty-second SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 405–418. Springer, London, UK, 2012.

[37] K. Dittakan, F. Coenen, and R. Christley. Satellite image mining for census collection: A comparative study with respect to the ethiopian hinterland. In *Proceedings of the Ninth International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, pages 260–274, New York, USA, July 2013.

[38] K. Dittakan, F. Coenen, R. Christley, and M. Wardeh. A comparative study of three image representations for population estimation mining using remote sensing imagery. In *Proceedings of the Ninth International Conference on Advanced Data Mining and Applications (ADMA)*, pages 253–264, Hangzhou, China, December 2013.

[39] K. Dittakan, F. Coenen, R. Christley, and M. Wardeh. Population estimation mining using satellite imagery. In *Proceedings of the Fifteenth International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 285–296, Prague, Czech Republic, August 2013.

[40] C Douglas and E.A. Montgomery. *Introduction to Linear Regression Analysis*. Wiley, 2013.

[41] C. Drummond and R. C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

[42] R. O. Duda and P.E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, January 1972.

[43] S. El-Salhi, F. Coenen, C. Dixon, and M.S. Khan. Predicting "springback" using 3d surface representation techniques: A case study in sheet metal forming. *The Journal of Expert Systems with Applications*, 42(1):79–93, 2015.

[44] J. Faichney and R. Gonzalez. Combined colour and contour representation using anti-aliased histograms. In *Proceedings of the Sixth International Conference on Signal Processing*, volume 1, pages 735–739, August 2002.

[45] J.A.X. Fanoe. Lessons from census taking in south africa: Budgeting and accounting experiences. *The African Statistical*, 13(3):82–109, 2011.

[46] A. Fazlollahi, N. Dowson, F. Meriaudeau, S. Rose, M. Fay, P. Thomas, Z. Taylor, Y. Gal, A. Coultard, C. Winter, D. MacFarlane, O. Salvado, S. Crozier, and P. Bourgeat. Automatic brain tumour segmentation in 18F-FDOPA PET using PET/MRI fusion. In *Proceedings of the International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 325–329, December 2011.

[47] N.J. Fesharaki and H. Pourghassem. Medical X-ray images classification based on shape features and Bayesian rule. In *Proceedings of the Fourth International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 369–373, November 2012.

[48] Office for National Statistics. National population projections, 2010-based statistical bulletin. Technical report, Office for National Statistics, 2011.

[49] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, March 2003.

[50] M. Friedman. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[51] E. Ganea and M. Brezovan. Graph object oriented database for semantic image retrieval. In B. Catania, M. IvanoviÆŠÃą, and B. Thalheim, editors, *Advances in Databases and Information Systems (ADBIS)*, volume 6295 of *Lecture Notes in Computer Science*, pages 563–566. Springer Berlin Heidelberg, 2010.

[52] M.A. Garcia and D. Puig. Improving texture pattern recognition by integration of multiple texture feature extraction methods. In *Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR)*, volume 3, pages 7–10 vol.3, Canada, August 2002.

[53] S. García, D. Molina, and F. Lozano, M.and Herrera. A study on the use of non-parametric tests for analyzing the evolutionary algorithms behaviour: A case study on the CEC 2005 special session on real parameter optimization. *Journal of Heuristics*, 15(6):617–644, December 2009.

[54] M. Ghaziani, Y. Mohamadi, A. Bugra Koku, and E.I. Konukseven. Extraction of unstructured roads from satellite images using binary image segmentation. In *Proceedings*

of the Twenty-first Conference on Signal Processing and Communications Applications Conference (SIU), pages 1–4, April 2013.

[55] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson Prentice Hall, Third edition, 2007.

[56] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, Second edition, 2001.

[57] R.C. Gonzalez, R.E. Woods, and S.L. Eddin. *Digital Image Processing Using MATLAB*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.

[58] Australian Government. Australian jobs 2013. Technical report, Department of Education, Employment and Workplace Relations, Australian Government, 2013.

[59] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[60] R.M. Hadad, A. De A Araujo, and Jr. Martins, P.P. Using the Hough transform to detect circular forms in satellite imagery. In *Proceedings of the Fourteenth Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pages 406–413. Washington DC, USA, IEEE Computer Society, October 2001.

[61] M. Haddad and H. Kheddouci. Graph based approaches for service oriented applications in ad hoc networks. In *Proceedings of the IEEE International Conference on Pervasive Services (ICPS)*, pages 431–436, July 2007.

[62] A. Hadid. The Local Binary Pattern approach and its applications to face analysis. In *Proceedings of the First Workshop on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–9. IEEE Computer Society, 2008.

[63] E. L. Hall. Almost uniform distributions for computer image enhancement. *IEEE Transactions on Computers*, 23(2):207–208, February 1974.

[64] M.A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 359–366, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[65] M.A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447, November 2003.

[66] M.A. Hall and L.A. Smith. Feature selection for machine learning: Comparing a Correlation-Based filter approach to the wrapper. In *Proceedings of the Twelfth International Conference on Florida Artificial Intelligence Research Society (FLAIRS)*, pages 235–239, Orlando, Florida, USA, May 1999.

[67] I.A. Hamza and A Iyela. Land use pattern, climate change, and its implication for food security in Ethiopia: A review. *Ethiopian Journal of Environmental Studies and Management*, 5:26–31, 2012.

[68] J. Han and M. Kamber. *Data mining : concepts and techniques*. Elsevier, Burlington, MA, USA, 2005.

[69] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Third edition, 2011.

[70] J. Han and K. Ma. Fuzzy color histogram and its use in color image retrieval. *IEEE Transactions on Image Processing*, 11(8):944–952, 2002.

[71] D. J. Hand, P. Smyth, and H. Mannila. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001.

[72] R. M. Haralick and G. Shapiro, L. *Computer and Robot Vision*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, First edition, 1992.

[73] M. Hassanzadeh and C.Y. Evrenosoglu. Power system state forecasting using regression analysis. In *IEEE Power and Energy Society General Meeting*, pages 1–6, San Diego, CA, USA., July 2012.

[74] D. Haverkamp. Automatic building extraction from IKONOS imagery. In *Proceedings of the Annual Conference on American Society for Photogrammetry and Remote Sensing (ASPRS)*, Denver, Colorado, USA, 2004.

[75] X. He, D. Li, J.and Wei, W. Jia, and Q. Wu. Canny edge detection on a virtual hexagonal image structure. In *Proceedings of the Joint Conference on Pervasive Computing (JCPC)*, pages 167–172, December 2009.

[76] G. Heusch, Y. Rodriguez, and S. Marcel. Local binary patterns as an image preprocessing for face authentication. In *Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 9–14. IEEE Computer Society, 2006.

[77] G. Holmes, A. Donkin, and Ian H. Witten. Weka: a machine learning workbench. In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361, Nov 1994.

[78] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, pages 549–552, Florida, USA, November 2003.

[79] J. Huang, S.R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Spatial color indexing and applications. *International Journal of Computer Vision (IJCV)*, 35:245–268, 1999.

[80] J.F. Hughes, A. Dam, M. McGuire, D. F. Sklar, J. D. Foley, S. K. Feiner, and K. Akeley. *Computer graphics: Principles and Practice*. Addison-Wesley Professional, Boston, MA, USA, Third edition, July 2013.

[81] G. Iannizzotto and L. Vita. Fast and accurate edge-based segmentation with no contour smoothing in 2-D real images. *IEEE Transactions on Image Processing*, 9(7):1232–1237, July 2000.

[82] M.R. Inggs and R.T. Lord. Applications of satellite imaging radarn. In *Proceedings of the Conference on South African Institute of Electrical Engineers (SAIEE)*, volume 1, pages 65–68, South Africa, 2000.

[83] A. Inokuchi, T. Washio, and H. Motoda. An Apriori-based algorithm for mining frequent substructures from graph data. In *Proceedings of the International Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 13–23, 2000.

[84] B.D. Jadhav and P.M. Patil. An effective method for satellite image enhancement. In *Proceedings of the International Conference on Computing, Communication Automation (ICCCA)*, pages 1171–1175, May 2015.

[85] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[86] D. Janez. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, December 2006.

[87] Y. Javed, M.M. Khan, and J. Chanussot. Population density estimation using textons. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, pages 2206–2209, 2012.

[88] M. Jian, L. Liu, and F. Guo. Texture image classification using perceptual texture features and gabor wavelet features. In *Proceedings of the Asia-Pacific Conference on Information Processing (APCIP)*, volume 2, pages 55–58, July 2009.

[89] C. Jiang, F. FCoenen, and M. Zito. Frequent sub-graph mining on edge weighted graphs. In *Proceedings of the Twelfth International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 77–88, Bilbao, Spain, August 2010.

[90] L. Jin, S. Satoh, and M. Sakauchi. A novel adaptive image enhancement algorithm for face detection. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR)*, volume 4, pages 843–848 Vol.4, August 2004.

[91] S. M. Jong and F.D. Van der Meer. *Remote sensing image analysis : including the spatial domain*. Remote sensing and digital image processing. Kluwer Academic, Dordrecht, Boston, USA, 2004.

[92] S. Jouili and S. Tabbone. Towards performance evaluation of graph-based representation. In *Proceedings of the Eighth International Conference on Graph-based Representations in Pattern Recognition (GbRPR)*, pages 72–81, Berlin, Heidelberg, 2011. Springer-Verlag.

[93] B. Kaur and S. Jindal. An implementation of feature extraction over medical images on open cv environment. In *Proceedings of the International Conference on Devices, Circuits and Communications (ICDCCom)*, pages 1–6, September 2014.

[94] A. K. Kaw and E. E. Kalu. *Numerical Methods with Applications*. http://www.autarkaw.com, Second edition, 2010.

[95] T. Kondo, J. Ueno, and S. Takao. Hybrid feedback gmdh-type neural network using principal component-regression analysis and its application to medical image recognition of heart regions. In *Proceedings of the Joint Seventh International Conference on Soft Computing and Intelligent Systems (SCIS) and Fifteenth International Symposium on Advanced Intelligent Systems (ISIS)*, pages 1203–1208, December 2014.

[96] S.P. Kraus, L.W. Senger, and J.M. Ryerson. Estimating population from photographically determined residential land use types. *Journal of Remote Sensing of Environment*, 3(1):35 – 42, 1974.

[97] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 313–320, 2001.

[98] Z. Lai, X. Qu, Y. Liu, D. Guo, Z. Ye, J.and Zhan, and Z. Chen. Image reconstruction of compressed sensing MRI using graph-based redundant wavelet transform. *Medical Image Analysis*, 2015.

[99] N. S. Lam. Spatial interpolation methods: A review. *The American Cartographer*, 10(2):129–150, 1983.

[100] F. Lebourgeois, F. Drira, D. Gaceb, and J. Duong. Fast integral meanshift: Application to color segmentation of document images. In *Proceedings of the Twelfth International Conference on Document Analysis and Recognition (ICDAR)*, pages 52–56, August 2013.

[101] W. K. Leow and R. Li. The analysis and applications of adaptive-binning color histograms. *The Journal of Computer Vision and Image Understanding*, 94(1-3):67–91, April 2004.

[102] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *The IEEE Transactions on Software Engineering*, 34(4):485–496, July 2008.

[103] G. Li and Q. Weng. Using Landsat ETM+ imagery to measure population density in Indianapolis, Indiana, USA. *Journal of Photogrammetric Engineering and Remote Sensing*, 71(8):63–69, 2005.

[104] H. Li, H. Guo, H. Guo, and Z. Meng. Data mining techniques for complex formation evaluation in petroleum exploration and production: A comparison of feature selection and classification methods. In *Proceedings of the Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACIIA)*, volume 1, pages 37–43, December 2008.

[105] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histograms: Image representation using markov stationary features. In *Proceedings of the IEEE International Conference on Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Anchorage, Alaska, USA, June 2008.

[106] P. Liang, S.F. Li, and J.W. Qin. Mulit-resolution local binary patterns for image classification. In *Proceedings of the Twentieth International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pages 164–169, July 2010.

[107] S.C.F. Lin, C.Y. Wong, M.A. Rahman, G. Jiang, S. Liu, N. Kwok, H. Shi, Y. Yu, and T. Wu. Image enhancement using the averaging histogram equalization (AVHEQ) approach for contrast improvement and brightness preservation. *Journal of Computers and Electrical Engineering*, 2015.

[108] C.X. Ling, J. Huang, and H. Zhang. Auc: A better measure than accuracy in comparing learning algorithms. In *Proceedings of the Sixteenth Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence*, pages 329–341, Berlin, Heidelberg, 2003. Springer-Verlag.

[109] G.S. Linoff and M.J.A. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing, Third edition, 2011.

[110] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[111] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, April 2005.

[112] J. Liu, M. Li, J. Wang, T Wu, F.and Liu, and Y. Pan. A survey of MRI-based brain tumor segmentation methods. *Tsinghua Science and Technology*, 19(6):578–595, December 2014.

[113] Q. Liu, Y. Zhang, G. Liu, and C. Huang. Detection of quasi-circular vegetation community patches using circular hough transform based on ZY-3 satellite image in the Yellow River Delta, China. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, pages 2149–2151, July 2013.

[114] X. Liu and K. Clarke. Estimation of residential population using high resolution satellite imagery. In *Proceedings of the Third International Symposium on Remote Sensing of Urban Area*, pages 153–160, June 2002.

[115] S. Livens, P. Scheunders, G. Van de Wouwer, and D. Van Dyck. Wavelets for texture analysis, an overview. In *Processing of the Sixth International Conference on Image Processing and Its Applications*, volume 2, pages 581–585, July 1997.

[116] C. Lo. Zone-based estimation of population and housing units from satellite-generated land use/land cover maps. *Remotely Sensed Cities*, pages 157–180, 2003.

[117] F. Long, H. Zhang, and D. Feng. Fundamentals of content-based image retrieval. In DavidDagan Feng, Wan-Chi Siu, and Hong-Jiang Zhang, editors, *Multimedia Information Retrieval and Management, Technological Fundamentals and Applications*, Signals and Communication Technology, pages 1–26. Springer Berlin Heidelberg, 2003.

[118] T. C. Lu and C. C Chang. Color image retrieval technique based on color features and image bitmap. *Journal of Information Processing and Management*, 43(2):461–472, March 2007.

[119] T. Ma, C. Zhou, T. Pei, S. Haynie, and J. Fan. Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: A comparative case study from China's cities. *Journal of Remote Sensing of Environment*, 124:99–107, 2012.

[120] P. Madden, J. Goodman, J. Green, and C. Jenkinson. Growing pains: Population and sustainability in the UK. Technical report, Foum for the future, 2010.

[121] R. Maini and H. Aggarwal. Study and comparison of various image edge detection techniques. *International Journal of Image Processing (IJIP)*, 3(1):1–11, February 2009.

[122] R. Maini and H. Aggarwal. A comprehensive review of image enhancement techniques. *Journal of Computing*, 2(3):8–13, March 2010.

[123] I. Maric, K. Vujic, and M. Vuksan. Prediction of Bond's Next Trade Price with Rapidminer. In *Proceedings of the International Symposium on New Business Models Sustainable Competitiveness*, pages 144–150, Serbia, June 2014.

[124] A. Marion. *Introduction to image processing*. Chapman & Hall Computing, New York, First edition, 1991.

[125] M. Mather, K. Pollard, and L.A. Jacobsen. Report on America: First results from the 2010 census. Technical report, Population Reference Bureau, Washington, DC, USA, July 2011.

[126] The Mathworks, Inc., Natick, Massachusetts. *MATLAB version 8.4.0.150421 (R2014b)*, 2014.

[127] U. Maulik. Medical image segmentation using genetic algorithms. *IEEE Transactions on Information Technology in Biomedicine*, 13(2):166–173, March 2009.

[128] S. Mesbah, M. Kholief, and A. Mahdy. A GIS vectorization model for quad-tree satellite images. In *Proceedings of the Twenty-second International Conference on Computer Theory and Applications (ICCTA)*, pages 41–46, October 2012.

[129] M.C. Mihaescu. Classification of learners using linear regression. In *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 717–721, September 2011.

[130] S. J. Miller. The method of least squares. *Mathematics Department Brown University*, pages 1–7, 2006.

[131] U. Natesan, A. Parthasarathy, R. Vishnunath, G.E.J. Kumar, and V.A. Ferrer. Monitoring longterm shoreline changes along Tamil Nadu, India using geospatial techniques. *Aquatic Procedia*, 4:325 – 332, 2015.

[132] P. Nemenyi. *Distribution-free multiple comparisons*. PhD thesis, Princeton University, New Jersey, USA, 1963.

[133] Minister of Finance and Public Service Delivery. Helping to shape tomorrow: The 2011 census of population and housing in england and wales. Technical report, UK Cabinet Office, 2008.

[134] W.P. O'Hare, Carsey Institute, and University of New Hampshire. *Rural Areas Risk Being Overlooked in 2010 Census*. Issue brief. University of New Hampshire, Carsey Institute, 2010.

[135] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(7):971–987, 2002.

[136] I. Olmos and J. A. Gonzalez. Structural graph-based representations used for finding hidden patterns. In *Proceedings of the Fourth Latin American Workshop on Non-Monotonic Reasoning (LANMR)*, Puebla, Mexico, October 2008.

[137] N.R. Pal and S.K. Pal. A review on image segmentation techniques. *Journal of Pattern Recognition*, 26(9):1277–1294, 1993.

[138] C. Parker. An analysis of performance measures for binary classifiers. In *Proceedings of the Eleventh IEEE International Conference on Data Mining (ICDM)*, pages 517–526, December 2011.

[139] G. Pass and R. Zabih. Histogram refinement for content-based image retrieval. In *Proceedings of the Third Workshop on Applications of Computer Vision (WACV)*, pages 96–102, December 1996.

[140] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *Proceedings of the Fourth ACM International Conference on Multimedia*, MULTIMEDIA1996, pages 65–73, New York, USA, 1996. ACM.

[141] M. Pietikäinen. Image analysis with local binary patterns. In *Proceedings of the Conference on Scandinavian conference on Image Analysis (SCIA)*, pages 115–118, Heidelberg, 2005. Springer-Verlag Berlin.

[142] B. Pink. Census of population and housing: Nature and content australia 2011. Technical report, Australian Bureau of Statistics, 2008.

[143] S.J.H. Pirzada and A. Siddiqui. Analysis of edge detection algorithms for feature extraction in satellite images. In *Proceedings of the IEEE International Conference on Space Science and Communication (IconSpace)*, pages 238–242, July 2013.

[144] F. Pozzi, C. Small, and G. Yetman. Modeling the distribution of human population with night-time satellite imagery and gridded population of the world. In *FIEOS Conference Proceedings*, 2002.

[145] G. Raghotham Reddy, K. Ramudu, A. Srinivas, and R. Rameshwar Rao. Region based segmentation of satellite and medical imagery with level set evolution. In *Proceedings of the IEEE Conference on Recent Advances in Intelligent Computational Systems (RAICS)*, pages 642–646, September 2011.

[146] H.K. Ranota and P. Kaur. Review and analysis of image enhancement techniques. *International Journal of Information and Computation Technology*, 4:583–590, 2014.

[147] J. Rianto. Road network detection from SPOT satellite image using Hough transform and optimal search. In *Proceedings of Conference on the Asia-Pacific Conference on Circuits and Systems (APCCAS)*, volume 2, pages 177–180, 2002.

[148] F. Riaz and K.M. Ali. Applications of Graph Theory in Computer Science. In *Proceedings of the Third International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN)*, pages 142–145, July 2011.

[149] J. Richards. *Remote Sensing Digital Image Analysis: An Introduction*. Springer Berlin Heidelberg, Secaucus, NJ, USA, Fifth edition, 2013.

[150] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York, USA, 1987.

[151] J.C. Russ. *The Image Processing Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 2006.

[152] T. Rutherford. Population ageing: Statistics. Technical report, House of Commons Library, 2012.

[153] A. Saeed, A. Tariq, and U. Jawaid. Automated system for fingerprint image enhancement using improved segmentation and Gabor wavelets. In *Proceedings of the International Conference on Information and Communication Technologies (ICICT)*, pages 1–6, July 2011.

[154] A.-B. Salberg. Land cover classification of cloud-contaminated multitemporal high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):377–387, January 2011.

[155] H. Samet. The quadtree and related hierarchical data structures. *Journal of ACM Computing Surveys (CSUR)*, 16(2):187–260, June 1984.

[156] H. Samet. Hierarchical spatial data structures. In *Proceedings of the First Symposium on Design and Implementation of Large Spatial Databases*, pages 193–212, London, UK, 1990. Springer-Verlag.

[157] A. Sanfeliu, R. Alquezar, J. Andrade, J. Climent, F. Serratosa, and J. Verges. Graph-based representations and techniques for image processing and image analysis. *Journal of Pattern Recognition*, 35(3):639 – 650, 2002.

[158] C. Saravanan. Color image to grayscale image conversion. In *Proceedings of the Second International Conference on Computer Engineering and Applications (ICCEA)*, volume 2, pages 196–199, March 2010.

[159] R. J. Schalkoff. *Digital Image Processing and Computer Vision*. Wiley, 1989.

[160] B. Senthilkumar, G. Umamaheswari, and J. Karthik. A novel region growing segmentation algorithm for the detection of breast cancer. In *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4, December 2010.

[161] N. Senthilkumaran and J. Thimmiaraja. Histogram equalization for image enhancement using mri brain images. In *Proceedings of the World Congress on Computing and Communication Technologies (WCCCT)*, pages 80–83, February 2014.

[162] J. Senthilnath, S.N. Omkar, V. Mani, R. Prasad, R. Rajendra, and P.B. Shreyas. Multisensor satellite remote sensing images for flood assessment using swarm intelligence. In *Proceedings of the International Conference on Cognitive Computing and Information Processing (CCIP)*, pages 1–5, March 2015.

[163] S. Shekhar and H. Xiong. *Encyclopedia of GIS (Springer Reference)*. Springer US, First edition, 2008.

[164] L.B. Shestha and E.J. Heisler. The changing demographic profile of the united states. Technical report, Congressional Research Service, March 2011.

[165] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, and K.R.K. Murthy. Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, 11(5):1188–1193, September 2000.

[166] A. Silberschatz, H. Korth, and S. Sudarshan. *Database Systems Concepts*. McGraw-Hill, Inc., New York, USA, Fifth edition, 2006.

[167] A. Sinha, S Banerji, and C Liu. Novel Color Gabor-LBP-PHOG (GLP) descriptors for object and scene image classification. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '12, pages 58:1–58:8, New York, NY, USA, 2012. ACM.

[168] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Journal of Statistics and Computing*, 14(3):199–222, August 2004.

[169] C. Song, P. Li, and F. Yang. Multivariate texture measured by local binary pattern for multispectral image classification. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, pages 2145–2148, 2006.

[170] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson-Engineering, 2007.

[171] A. Stumpf, J.P. Malet, P. Allemand, and P. Ulrich. Surface reconstruction and landslide displacement measurements with plâĹŽÂl'iades satellite images. *Journal of Photogrammetry and Remote Sensing (ISPRS)*, 95:1 – 12, 2014.

[172] M. J. Swain and D. H. Ballard. Color indexing. *International journal on Computer Vision*, 7(1):11–32, November 1991.

[173] H. Tamura and N. Yokoya. Image database systems: A survey. *Journal of Pattern Recognition*, 17(1):29 – 43, 1984. Knowledge Based Image Analysis.

[174] Z. Tang and D. Shen. Canny edge detection Codec using VLib on Davinci Series DSP. In *Proceedings of the International Conference on Computer Science Service System (CSSS)*, pages 221–224, August 2012.

[175] W. Tao, H. Jin, and Y. Zhang. Color image segmentation based on mean shift and normalized cuts. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1382–1389, October 2007.

[176] D.P. Tian. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8:385–395, 2013.

[177] J. Tsuruoka, Y.and Tsujii and S Ananiadou. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pages 477–485, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[178] S. Velickov and D. Solomatine. Predictive data mining: Practical examples. In *Proceedings of the Artificial Intelligence in Civil Engineering*, pages 1–17, Cottbus, Germany, March 2000.

[179] R.C. Veltkamp, M. Tanase, and D. Sent. Features in content-based image retrieval systems: A survey. In R. Veltkamp, H. Burkhardt, and H. Kriegel, editors, *State-of-the-Art in Content-Based Image and Video Retrieval*, volume 22 of *Computational Imaging and Vision*, chapter 5, pages 97–124. Kluwer Publishers, 2001.

[180] M. Wang, S. Yuan, and J. Pan. Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed Hough Transform. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, pages 508–511, July 2013.

[181] X. Wang, H. Wang, J. Fan, and F. Wang. Improved adaptive threshold algorithm for fingerprint segmentation based on multiple features. In *Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, volume 3, pages 218–222, October 2010.

[182] X. Y. Wang, J.F. Wu, and H.Y. Yang. Robust image retrieval based on color histogram of local feature regions. *Journal of Multimedia Tools and Applications*, 49(2):323–345, August 2010.

[183] Y. Wang, S. Bahrami, and S. C. Zhu. Perceptual scale space and its applications. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, pages 58–65, Beijing, China, October 2005.

[184] A. Wästfelt, T. Tegenu, M.M. Nielsen, and B. Malmberg. Qualitative satellite image analysis: Mapping spatial distribution of farming types in Ethiopia. *Journal of Applied Geography*, 32(2):465 – 476, 2012.

[185] M.K. William. *Curve Fitting for Programmable Calculators*. Syntec Inc., 1984.

[186] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Amsterdam, Third edition, 2011.

[187] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Second edition, 2005.

[188] S. Wu and P. Flach. A scored AUC Metric for classifier evaluation and selection. In *Proceedings of The International Conference on Machine Learning Society (ICML) workshop on on ROC Analysis in Machine Learning*, 2005.

[189] S.S. Wu, X. Qiu, and L. Wang. Population estimation methods in GIS and remote sensing: A review. *Journal of GIScience & Remote Sensing*, 42(1):80–96, 2005.

[190] G. Xu, G. Zhao, L. Yin, Y. Yin, and Y. Shen. A CNN-based edge detection algorithm for remote sensing image. In *Proceedings of the Conference on Chinese Control and Decision Conference (CCDC)*, pages 2558–2561, July 2008.

[191] K. Xu and F. Wang. Behavioral graph analysis of internet applications. In *Proceedings of the IEEE International Conference on Global Telecommunications Conference (GLOBECOM)*, pages 1–5, December 2011.

[192] G. Yang, K. Chen, M. Zhou, Z. Xu, and Y. Chen. Study on statistics iterative thresholding segmentation based on aviation image. In *Proceedings of the Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD)*, volume 2, pages 187–188, July 2007.

[193] J. Yang and T. Deng. Optimal binary thresholding segmentation for medical images in rough fuzzy set framework. In *Proceedings of the Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*, pages 638–643, June 2013.

[194] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[195] S. Yonghua, L. Xiaojuan, G. Huili, and Z. Chunping. Research on improving spatial resolution of RS images based on quadtree decomposition. In *Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS)*, volume 2, pages 1088–1091, July 2008.

[196] I.T. Young, J.J. Gerbrands, L.J. Vliet, K. Bibliotheek, D. Haag, Y.I. Theodore, G.J. Jacob, V. Vliet, and L. Jozef. Fundamentals of image processing, 1995.

[197] W. Yu, F. Coenen, M. Zito, and K. Dittakan. Classification of 3d surface data using the concept of vertex unique labelled subgraphs. In *Proceedings of the IEEE International Conference Workshops on Data Mining (ICDM)*, pages 47–54, 2014.

[198] W. Yu, F. Coenen, M. Zito, and S. El-Salhi. Vertex unique labelled subgraph mining. In *Research and Development in Intelligent Systems XXX. Incorporating Applications and Innovations in Intelligent Systems XXI Proceedings of (AI-2013), The Thirty-third SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 21–37, Cambridge, UK, December 2013.

[199] H. K. Yuen, J. Princen, J. Illingworth, and J. Kittler. A comparative study of Hough transform methods for circle finding. In *Proceedings of the Fifth Conference on Alvey Vision*, pages 169–174, Reading, UK, August 1989.

[200] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Proceedings of the IEEE International Conference on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(6):915–928, 2007.

[201] G. Zhao, G. Wu, Y. Liu, and J. Chen. Texture classification based on completed modeling of local binary pattern. In *Proceedings of the International Conference on Computational and Information Sciences (ICCIS)*, pages 268–271, 2011.

[202] Q. Zu and L. Wenfeng. The research of customer classification based on extended bayes model. In *Proceedings of the Third International Conference on Pervasive Computing and Applications (ICPCA)*, volume 1, pages 22–25, October 2008.

Appendix A

Additional Algorithms

A.1 Introduction

With respect to the work presented in this thesis details concerning a number of additional algorithms, not discussed earlier in the thesis, are presented in this appendix. This appendix commences by presenting the household segmentation algorithm in Section A.2. With respect to the work described in Chapter 4, the graph/tree based representation algorithms are presented in Section A.3. Section A.4 discussed the detail of the algorithm used for the colour histogram based representation which was used in Chapter 5. The texture based representation algorithm (Chapter 6) is then discussed in Section A.5. The algorithms for feature selection, classification and regression are presented in Section A.6. With respect to the large scale study (discussed in Chapter 8), the algorithm used for satellite image collection (see Section 8.2 of Chapter 8) is presented in Section A.7.

A.2 Household Segmentation Algorithm

The household image segmentation process applied to a collection of input satellite images is presented in this section. Recall from Sub-section 3.4.4 of Chapter 3 that the main objective of the image segmentation process was to isolate an individual household from given input satellite data. The process was applied using MATLAB version 2014b [126]. Algorithm 4 presents the household segmentation mechanism in more detail. The input to the algorithm is a collection of satellite images. The output is a collection of individual segmented household images (*HouseImages*).

Algorithm 4 commences by registering and aligning the original image using an image registration process (line 7). The next step is to convert the RGB image into a sixteen colour indexed image (line 8). Prior to applying histogram equalisation for image enhancement, the colour indexed images are transformed into a greyscale colour space to which histogram equalisation could be applied (lines 9-10). Once the image enhancement process has been completed the next step is to capture the household boundary in which the household building and related objects are located. To do this the image is first transformed into a binary image (line 11);

consequently Canny edge detection is applied (line 12). Once the Canny Image has been generated, the next step is to detect the straight lines, making up the boundary, using the Hough Transform (line 13). Each line is defined by start and end points, length (ρ) and direction (θ); this information is used in order to fit the horizontal and vertical lines, using least squares, to a bounding box shape (as presented in Algorithm 5 and Algorithm 6, respectively) (lines 14-15). The intersections that can be identified are then used to register the image (line 17). The cropped image was recorded as a collection of individual segmented households (line 18).

Algorithm 4 Household Segmentation

```
 1: Input:
 2: SatelliteImg = A collection of satellite images
 3: Output:
 4: HouseImages = A collection of individual segmented household image

 5: for i = 0 to i = |SatelliteImg| − 1 do
 6:     origImg= SatelliteImg[i]
 7:     registImg = register and align the original image
 8:     indexedImg = rgb2ind(registImg, 16)
 9:     greyImg = ind2grey(indexedImg)
10:     hisImg = histeq(greyImg)
11:     binaryImg = grey2binary(hisImg)
12:     cannyImg = Canny edge detection (binaryImg)
13:     lines = hough(cannyImg)
14:     [hLines, hLineTota,l minY, maxY] = hLeastSquares(origImg, lines)
15:     [vLines, vLineTotal, minX, maxX] = vLeastSquares(origImg, lines)
16:     if hLineTotal ≥ 2 and vLineTotal ≥ 2 then
17:         cropImg = crop(registedImg, [ minX, minY, (maxX - minX), (maxY - minY)])
18:         HouseImages = HouseImages ∪ cropImg
19:     end if
20: end for
```

Recall in Sub-section 3.4.3 of Chapter 3 that the Least squares mechanism was used for line/curve fitting. Algorithm 5 descibes the adopted process for line fitting using Least squares. The inputs to the algorithm are: (i) a collection of lines generated using the Hough Transform and (ii) a household image. The outputs of the algorithm are: (i) a collection of horizontal lines (*Lines*), (ii) the number of horizontal lines (*hLineTotal*), (iii) the minimum *y* value of the horizontal lines (*minY*) and (iv) the maximum *y* value of the horizontal lines (*maxY*).

Algorithm 5 starts by dividing an image into 2 equal sized horizontal regions (line 12) so that two (horizontal) lines can be captured. For each region, the start and end *y* values are calculated (lines 13-14). A number of parameters are then initialised (lines 15 to 19): (i) *lineTotal* is the number of horizontal line in the current region, (ii) *min* is minimum value of *x*, (iii) *max* is the maximum value of *x*, (iv) *y* is the average *y* value of horizontal line and (v) *yTotal* is the summary of *y* value of horizontal line. The *lineTotal* and *yTotal* were used for *y* value calculation. The *min* and *max* were used to be stat and end point of the line, respectively.

The identified lines are then processed with respect to the current region. If a line is parallel to $x - axis$ the type of the line is defined as "horizontal" and: (i) the total number of lines counter is incremented (line 22), (ii) the total y value is incremented, (iii) the minimum x value is checked and if necessary updated and (iii) the maximum x value is checked and if necessary updated (lines 20-31). If one or more lines have been discovered with respect to the current region (line 32 the current total number of lines and total y value were used to calculate the average y value (line 33). The number of horizontal lines is then incremented by one (line 34) and the minimum and maximum y for all horizontal lines calculated for output (line 35-40).

Once the horizontal line fitting has been completed vertical line fitting was conducted using Algorithm 6. The algorithm works in a very similar manner to Algorithm 5. The inputs are again: (i) a collection of lines obtained using a Hough transform and (ii) a household image. The outputs are: (i) a collection of vertical lines (*Lines*), (ii) the number of vertical lines ($vLineTotal$), (iii) the minimum x value for the vertical lines ($minX$) and (iv) the maximum x value for the vertical lines ($maxX$).

As in the case of the horizontal line fitting algorithm, Algorithm 6 begins by sub-dividing a satellite image into 2 regions (line 12). For each region, the start and end x values are calculated (lines 13-14). As before, each input line is processed so as to determine whether it is located within the current region or not. If so: (i) the total number of lines counter is incremented, (ii) the total y value is incremented, (iii) the minimum y values is checked and if necessary updated and (iv) the maximum y value is checked and if necessary updated (lines 20-31). If there is at least one line within the current region (line 32), the average x value is determined using the total number of lines together with the total x value (line 33). The number of vertical line was then incremented (lines 34) and the minimum and maximum x of all vertical lines calculated (line 35-40).

A.3 Graph-Based Image Representation

Recall from Chapter 4 that to extract the features from an image using a graph-based representation comprises five sub-processes: (i) quadtree decomposition, (ii) tree/graph representation (using the Graph Modelling Language (GML) file representation), (iii) feature subgraph mining, (iv) feature vector generation and (v) feature selection (as described in further detail on Section 4.2 to Section 4.5 of Chapter 4). With respect to graph based image representation there are three algorithms that merit further discussed in this appendix: (i) Algorithm 7 for generating the GML file format and (ii) Algorithm 8 for conducting the quadtree decomposition. To generate the GML file format (Algorithm 7), firstly a household image was decomposed using Algorithm 8 then a set of nodes and edges were generated before converted into a GML file format.

The algorithms used to generate the graph based image representation were implemented using MATLAB version 2014b [126]. The graph-based representation mechanism was used to capture the image content, however was not suitable for the purpose of classifier generation.

Algorithm 5 Least Squares and Line Fitting for Horizontal Line

```
1: Input:
2: original = A satellite image
3: Lines = A collection of lines from Hough transform
4: Output:
5: hLines = A collection of horizontal lines
6: hLineTotal = Number of horizontal lines
7: minY = Minimum y value of horizontal lines
8: maxY = Maximum y value of horizontal lines

9: hLineTotal = 0
10: minY = original.height
11: maxY = 0
12: for j =0 to j = 1 do
13:     start = j*original.height/2
14:     end = (j+1)*original.height/2
15:     hLinesj.lineTotal = 0
16:     hLinesj.min = original.width
17:     hLinesj.max = 0
18:     hLinesj.y = 0
19:     yTotal = 0
20:     for k = 0 to k = | lines | -1 do
21:         if linesk.type = horizontal and start <= linesk.start.x <= end and start <= linesk.end.x <= end then
22:             hLinesj.lineTotal = hLinesj.lineTotal + 1
23:             yTotal = yTotal + linesk.start.y + linesk.end.y
24:             if hLinesj.min > linesk.start.x then
25:                 hLinesj.min = linesk.start.x
26:             end if
27:             if hLinesj.max < linesk.end.x then
28:                 hLinesj.max = linesk.end.x
29:             end if
30:         end if
31:     end for
32:     if hLinesj.lineTotal > 0 then
33:         hLinesj.y = yTotal/(hLinesj.lineTotal*2)
34:         hLineTotal = hLineTotal + 1
35:         if hLinesj.y < minY then
36:             minY = hLinesj.y
37:         end if
38:         if hLinesj.y > maxY then
39:             maxY = hLinesj.y
40:         end if
41:     end if
42: end for
```

Algorithm 6 Least Squares and Line Fitting for Vertical Line

```
1: Input:
2: Lines = A collection of lines from Hough transform
3: original = A satellite image
4: Output:
5: vLines = A collection of vertical lines
6: vLineTotal = Number of vertical lines
7: minX = Minimum x value of vertical lines
8: maxX = Maximum x value of vertical lines

9: vLineTotal = 0
10: minX = original.width
11: maxX = 0
12: for j =0 to j = 1 do
13:     start = j*original. width /2
14:     end = (j+1)*original. width /2
15:     vLinesj.lineTotal = 0
16:     xTotal = 0
17:     vLinesj.min = original.height
18:     vLinesj.max = 0
19:     vLinesj.x = 0
20:     for k = 0 to k = | lines | -1 do
21:         if linesk.type = vertical and start <= linesk.start.y <= end and start <= linesk.end.y
    <= end  then
22:             vLinesj.lineTotal = vLinesj.lineTotal + 1
23:             xTotal = xTotal + linesk.start.x + linesk.end.x
24:             if vLinesj.min > linesk.start.y then
25:                 vLinesj.min = linesk.start.y
26:             end if
27:             if vLinesj.max < linesk.end.y then
28:                 vLinesj.max = linesk.end.y
29:             end if
30:         end if
31:     end for
32:     if vLinesj.lineTotal > 0 then
33:         vLinesj.x = xTotal/(vLinesj.lineTotal*2)
34:         vLineTotal = vLineTotal + 1
35:         if vLinesj.x < minX  then
36:             minX = vLinesj.x
37:         end if
38:         if vLinesj.x > maxX then
39:             maxX = vLinesj.x
40:         end if
41:     end if
42: end for
```

Thus Frequent Subgraph Mining (FSM) was applied. FSM is the process of identifying frequently occurring subgraphs which were considered to be features. To facilitate the operation of the FSM process, the input graph was transformed into GML format, a format compatible with the adopted FSM process.

The GML file format was generated using Algorithm 7. The input to the algorithm is a collection of household images. The output is a collection of graphs represented using the GML file format, $GmlFile$. For each household, the image was decomposed using a quadtree decomposition using a call to Algorithm 8 (line 6). Once the household images have been decomposed and recored in the quadtree format, the nodes and edges were then identified. The loop from lines 7 to 10 was for nodes and node label identification for the leaf nodes (lines 6-7). The loop from lines 11 to 23 was for nodes and edges generation the remaining nodes. One each iteration (level), each child node was used to identified the parent node (line 13). If the parent node did not exist then this parent node and node label were generated and recored in the *parentNodes* array (lines 14-18). The edge label from the child node to parent node was generated (line 19). The parent nodes were assigned to be the child nodes for next level (line 21) and the parent node array was set to empty (line 22).

Once a collection of node and edge labels were identified. The GML open and close tags were then included (lines 24-25). The GML file was then generated and saved for further usage in the context of the FSM process (lines 26-31).

The quadtree decomposition algorithm (Algorithm 8) takes as input a 2D household image and a maximum level of decomposition (a value of 4 in this case). The output of the algorithm was a quadtree whereby the whole image was represented by the root of the tree (line 5). The whole image (the root level) was then decomposed using the function *quadDecomposition* (line 6). The function from line 7 to 17 is the *quadDecomposition* function which conducts the desired quadtree decomposition process. If the current level of decomposition is not zero the current (sub) image is decomposed into four quadrants or regions (line 11). The process continues in a recursive manner until homogenous regions are reached or the maximum level of decomposition is attained (line 12-17).

A.4 Colour Histogram Based Representation Algorithm

This section presents the algorithm used to generate household images represented using the colour histogram based representation. The detail of this representation was discussed in Chapter 5. The utilised colour histogram based representation algorithm is given in Algorithm 9. The algorithm was again implemented using MATLAB version 2014b [126]. The input to the algorithm is a collection of individual household images, the output is a collection of colour histograms and other statistical colour matrices of a collection of household images.

For each household image, the RGB image was transformed into HSV colour space (line 8) and a greyscale colour space (line 9). Then two categories of colour based representation were extracted: (i) colour histograms and (ii) statistical colour metrics.

Algorithm 7 GML Generation

```
1: Input:
2: HouseImages = A collection of segmented household image
3: Output:
4: GmlFile = Graph-Based representation using Graph Modelling Language (GML)

5: for i = 0 to i =| HouseImages | - 1 do do
6:     Quadtree = quadtreeDecomposition(HouseImages[i])
7:     for i=1 to i= |Quadtree| do
8:         nodes = nodes ∪ generateNode(Quadtree[i])
9:         nodeLabels = nodeLabel + generateNodeLabel(nodes[i])
10:    end for
11:    for j=1 to j=4 do
12:        for k=1 to k= |nodes| do
13:            parentNode = calculateParentNode(nodes[k])
14:            if parentNode not exist in parentNodes then
15:                parentNode = generateParentNode(node[k])
16:                nodeLabels = nodeLabels + generateNodeLabel(parentNode)
17:                parentNodes = parentNodes ∪ parentNode
18:            end if
19:            edgeLabels = edgeLabels + generateEdgeLabel(nodes[k], parentNode)
20:        end for
21:        nodes = parentNodes
22:        parentNodes = { }
23:    end for
24:    gmlOpenTag = generateOpenTag()
25:    gmlCloseTag = generateCloseTag()
26:    GmlFile = new File()
27:    GmlFile.write(gmlHeader)
28:    GmlFile.write(nodeLabels)
29:    GmlFile.write(edgeLabels)
30:    GmlFile.write(gmlFooter)
31:    GmlFile.save()
32: end for
```

Algorithm 8 Quadtree Decomposition

```
1: Input:
2: 2D-space, max = maximum level of decomposition, 4
3: Output:
4: Quadtree

5: root = start of quadtree
6: quadDecomposition
7: function Q(u)adDecomposition(max, link, space)
8:     if max == 0 then
9:         return
10:    else
11:        decomp = {NW, NE, SE, SW} = space decomposed into quadrants
12:        for i =1 to i = 4 do
13:            link.i = decomp[i]
14:            if decomp[i] not homogeneous then
15:                quadDecoposition(max-1, link.i, decomp[i])
16:            end if
17:        end for
18:    end if
19: end function
```

The colour histograms are defined by counting the number of pixels for each quantised bin (see detail in Section 5.2 in Chapter 5). The number bins used for the quantisation of the colour space with respect to the work presented in this thesis was 32. For each household satellite image seven different histograms were generated: (i) three histograms from the RGB colour spaces (red, green, and blue) (lines 10-12), (ii) three histograms from the HSV colour spaces (hue, saturation, and value) (lines 13-15) and (iii) a intensity histogram using the greyscale colour space (line 16).

The colour metrics are a simple alternative representations for extracting the colour information from images as described in Section 5.3 in Chapter 5. For each household image 13 different statistical colour metrics were generated: (i) five features from RGB colour space (average red, average green, average blue, mean of RGB and standard deviation of RGB (lines 17-21)), (ii) five features from the HSV colour space (average hue, average saturation, average value, mean of HSV and standard deviation of HSV (lines 23-27)) and (iii) three features from the greyscale colour space (average greyscale, standard deviation of greyscale and average of greyscale histogram (lines 27-29)). Once the colour histograms and additional colour based statistical features were generated then the values were added to the *ColourFeatures* structure (lines 30-31) for further use in classification process.

Algorithm 9 Colour Histogram Based Representation

```
1: Input:
2: HouseImages = A collection of household image
3: Output:
4: ColourFeatures = A collection of colour histograms and other statistical colour measures

5: ColourFeatures = ""
6: for i = 0 to i =| HouseImages | - 1 do
7:     rgbImg = HouseImages[i]
8:     hsvImg = rgb2hsv(rgbImg)
9:     greyImg = rgb2grey(rgbImg)
10:    hisRed = imhist(rgbImg(:,:,1), 32)
11:    hisGreen = imhist(rgbImg(:,:,2), 32)
12:    hisBlue = imhist(rgbImg(:,:,3), 32)
13:    hisHue = imhist(hsvImage(:,:,1), 32)
14:    hisSaturation = imhist(hsvImage(:,:,2), 32)
15:    hisValue = imhist(hsvImage(:,:,3), 32)
16:    hisGrey = imhist(greyImage, 32)
17:    avgRed = mean2(rgbImg(:,:,1))
18:    avgGreen = mean2(rgbImg(:,:,2))
19:    avgBlue = mean2(rgbImg(:,:,3))
20:    meanRgb = mean2(rgbImg)
21:    stdRgb = std2(rgbImg)
22:    avgHue = mean2(hsvImage(:,:,1))
23:    avgSaturation = mean2(hsvImage(:,:,2))
24:    avgValue = mean2(hsvImage(:,:,3))
25:    meanHsv = mean2(hsvImage)
26:    stdHsv = std2(hsvImage)
27:    avgGrey = mean2(greyImage)
28:    stdGrey = std2(greyImage)
29:    avgGreyHis = mean2(hisGrey)
30:    ImageColourFeature = formatAsStructData(hisRed, hisGreen, hisBlue, hisHue, hisSat-
       uration, hisValue, hisGrey, avgRed, avgGreen, avgBlue, meanRgb, stdRgb, avgHue,
       avgSaturation, avgValue, meanHsv, stdHsv, avgGrey, stdGrey, avgGreyHis)
31:    ColourFeatures = ColourFeatures + ImageColourFeature
32: end for
```

A.5 Texture Based Representation Algorithm

The algorithms for texture based representation are presented in this section. The detail of the texture based analysis was discussed in Chapter 6. Two algorithms are discussed in this Appendix section: (i) the proposed texture based representation algorithm (Algorithm 10) and (ii) the algorithm for calculating LBPs (Algorithm 11). The algorithms were again implemented using MATLAB version 2014b [126].

The inputs to Algorithm 10 are: (i) a collection of individual household image and (ii) a radius R (recall the notation $LBP_{P,R}$). The output from the algorithm is a set of Local Binary Patterns (LBPs) and additional texture statistical measures. For each household image, the RGB image was transformed into the greyscale colour space and then the two categories of texture representation generated: (i) LBPs and (ii) statistical texture metrics. For the LBP representation the parameter R was given to define the radius for generating LBPs, then Algorithm 11 was applied for LBP generation (line 9) (Algorithm 11 is presented in further detail below). For the second category was discussed in detail in Section 6.3 of Chapter 6. Recall that ten different statistical colour metrics were generated: (i) two entropy metrics, entropy (line 10) and average local entropy (lines 11-12); (ii) four metrics produced using a Grey-Level Co-Occurrence Matrix (GLCM) (contrast, correlation, energy and homogeneity) (lines 13-14) and (iii) four features produced using a Discrete Wavelet Transform (DWT) (average approximation coefficient matrix (cA), average horizontal coefficient matrix (cH), average vertical coefficient matrix (cV) and average diagonal coefficient matrix (cD)) (lines 15-19)). Once both the statistical texture metrics and the LBPs had been generated the values were appended to $TextureFeatures$ (lines 20-21) for further use in the classification process.

Algorithm 11 presents the process for generating LBPs. The inputs to this algorithm (passed from Algorithm 10) are: (i) an individual household image and (ii) a radius R; the output for the algorithm is a collection of LBP values (lbpCount). The algorithm begins by determining the width and height of the given image (line 6). For each image pixel the LBP value was then generated (lines 7-34) (see Section 6.2 of Chapter 6 for further detail). As noted previously, with respect to the LBPs, $2^8 = 256$ different texture patterns could be generated. Once the LBPs had been calculated, this information could thus be conceptualised in the form of a 256 element feature vector where each element holds an occurrence count for the associated LBP (lines 35-37).

A.6 Feature Selection, Classification and Regression Algorithm

The algorithms associated with the following three distinct processes: (i) feature selection, (ii) classification and (iii) regression analysis, are presented in this section. The algorithms were used as discussed within the context of Chapters 4 to 7. The implementation of these algorithms was that available within the Waikato Environment for Knowledge Analysis (Weka) [77], version 3.6.12.

Algorithm 10 Texture Based Representation

Input:
HouseImages = A collection of household image
R = A radius number
Output:
TextureFeatures = LBP and other texture statistical measures

TextureFeatures = “”
for i = 0 to i =| HouseImages | - 1 **do**
    greyImg = rgb2grey(HouseImages[i])
    lbpValues = lbp(HouseImages[i],R,8)
    entropy = entropy(HouseImages[i])
    localEntropy = entropyfilt(greyImg)
    avglEntropy = mean2(localEntropy)
    glcMatrix = greycomatrix(greyImg)
    glcProps =greycoprops(glcMatrix,’Contrast’,’Homogeneity’,’Correlation’,’Energy’)
    [cA, cH, cV, cD] = dwt2(HouseImages[i], ’db1’)
    avg_cA = mean2(cA)
    avg_cH = mean2(cH)
    avg_cV = mean2(cV)
    avg_cD = mean2(cD)
    ImageTextureFeature = formatAsStructData(entropy, avglEntropy, glcProps.contrast, glcProp.energy, glcProps.correlation, glcProps.homogeneity, avg_cA, avg_cH, avg_cV, avg_cD, lbpCount)
    TextureFeatures = TextureFeatures + ImageColourFeature
end for

Algorithm 11 Local Binary Pattern

Input:
HouseImg = A household image
R = A radius number with respect to notation $LBP_{P,R}$
Output:
lbpCount = A collection of LBP values

```
[width, height] = size(HouseImg)
for i=1 to i= height do
    for j=1 to j=width do
        centre = HouseImg[i,j]
        lbp = 0
        if centre > HouseImg[i-R,j-R] then
            lbp = lbp + 128
        end if
        if center > HouseImg[i-R,j then
            lbp = lbp + 64
        end if
        if center > HouseImg[i-R,j+R] then
            lbp = lbp + 32
        end if
        if center > HouseImg[i,j+R] then
            lbp = lbp + 16
        end if
        if center > HouseImg[i+R,j+R]  then
            lbp = lbp + 8
        end if
        if center > HouseImg[i+R,j-R] then
            lbp = lbp + 2
        end if
        if center > HouseImg[i,j-R] then
            lbp = lbp + 1
        end if
        lbps = lbps ∪ lbp
    end for
end for
for i=0 to i=255 do
    lbpCount[i] = count number of i in lbps
end for
```

Table A.1: Feature selection algorithm

No	Evaluator	WEKA library name	Search algorithm
1	Chi-squared feature selection	ChiSquaredAttributeEval	Ranker
2	Gain Ratio feature selection	GainRatioAttributeEval	Ranker
3	Information Gain feature selection	InfoGainAttributeEval	Ranker
4	Correlation-based feature subset selection	CfsSubsetEval	BestFirst

Table A.2: Classification learning method

No	Method	WEKA library name
1	Decision tree (C4.5)	J48 (with binary split)
2	Naive bayes	NaiveBayes
3	Averaged one-dependence estimators	AODE
4	Bayesian network	BayesNet
5	Radial basis function network	RBFNetwork
6	Sequential minimal optimisation	SMO
7	Logistic regression	Logistic
8	Back propagation neural network	MultilayerPerceptron

Feature selection is the process of identifying a subset of relevant features for usage in model construction (classification and regression) as discussed in Section 2.4 of Chapter 2. Recall that two categories of feature selection mechanism were used with respect to the work presented in this thesis: (i) feature section for classification and (ii) feature selection for regression analysis. For the first category three atlernatives were considered: (i) Chi-squared, (ii) Gain Ratio and (iii) Information Gain. These algorithms were used to select relevant attributes/features to facilitate the classification model learning process described in Chapters 4 to 6. For the second category the Correlation-based Feature Subset (CFS) selection algorithm was applied to identify the relevant subset of attributes/features to facilitate the regression analysis process described in Chapter 7. Table A.1 presents some further derail concerning the four foregoing algorithms. From the table it can be seen that each algorithm is identified using the Weka library name.

Once revised feature vectors had been generated (feature vectors with features deemed irrelevant removed) classifier generation could commence. Recall that classification is the process of classifier generation and application using various classification learning methods. The classification process was described in Sub-section 2.5.1 of Chapter 2. Eight different classification methods were considered with respect to the work presented in Chapters 4 to 6, these are listed in Table A.2; note that each generator is again identified by its Weka library name.

A similar process to the above was followed with respect to regression analysis once revised feature vectors had been generated. Recall that regression analysis is the process of regression model generation using various regression analysis methods. The detail of the regression process was presented in Sub-section 2.5.1 of Chapter 2. Recall also that four different regression

Table A.3: Regression analysis method

No	Method	WEKA library name
1	Linear regression	LinearRegression
2	Least median squared linear regression	LeastMedSq
3	Isotonic regression	IsotonicRegression
4	Support vector machine for regression	SMOreg

prediction method were considered with respect to the work presented in Chapter 7; these are listed in Table A.3 (again the Weka library names are used as the identifiers).

A.7 Satellite Image Collection Algorithm

This section presents the algorithm used for downloading a satellite image collection. The algorithm was used with respect to the large scale study presented in Section 8.2 of Chapter 8. The proposed map downloader algorithm is presented in Algorithm 12. This algorithm was implemented using the Eclipse Java EE IDE. The inputs to the algorithm are: (i) the top-left corner latitude of the area (*tlLat*), (ii) the top-left corner longitude of the area (*tlLong*), (iii) the bottom-right corner latitude of the area (*brLat*), (iv) the bottom-right corner longitude of the area(*brLong*) (v) the Google API key, (vi) the mapSize, (vii) the scale and (viii) the zoom level. The Google API key is a code which is used to identify a Google account. Recall that the Mercator map projection equations given in Equations 8.1 to 8.4 respectively were used to convert: (i) from latitudes to x-pixel values, (ii) from x-pixels values to latitudes, (iii) from longitudes to y-pixels values and (iv) from y-pixels values to longitudes.

Algorithm 12 then proceeds as follows. First a URL is defined for use with the Google API (lines 12 to 16). This is made up of five elements: (i) the API url, (ii) the zoom level, (iii) the scale, (iv) the API key and (v) size of image (line 12-16). The last four are obtained from the input. Note that to capture a satellite image using the Google API, the downloader algorithm requires the centre of each image. The collection of satellite images is then obtained by processing the region from top to bottom starting at the top left corner. The latitude and longitude (*long*) for the first satellite image to be obtained is calculated and the image downloaded. The latitude is then incremented with *mapSize* (line 24) and the next image downloaded, and so on until an entire column of images has been obtained. The longitude is then incremented with *mapSize* (line 26), the latitude reset (line 19), and the next column of satellite images downloaded. The process continues until images for the entire region have been obtained.

Algorithm 12 Map downloader

```
1: Input:
2: tlLat = The top-left corner latitude of the area
3: tlLong = The top-left corner longitude of the area
4: brLat = The bottom-right corner latitude of the area
5: brLong = The bottom-right corner longitude of the area
6: apiKey = The Google API Key
7: mapSize = The map size in pixels
8: scale = The scale to be used
9: zoom = The zoom level to be used
10: Output:
11: MapImages = A collection of satellite image files

12: url = "http://maps.googleapis.com/maps/staticmap?format=
    jpg&maptype=satellite&sensor=false/"
13: url = url + "&zoom=" + zoom
14: url = url + "&scale"= + scale
15: url = url + "&key=" + apiKey
16: url = url + "&size=" + mapSize + "x" + mapSize
17: long = pixelToLong(longToPixel(tlLong) + (mapSize/2))
18: while long < brLong do
19:     lat = pixelToLat(latToPixel(tlLat) + (mapSize/2))
20:     while lat > brLat do
21:         mapUrl = url + "&center=" + lat + "," + long
22:         mapImage = download(mapUrl)
23:         MapImages = MapImages ∪ mapImage
24:         lat = pixelToLat(latToPixel(tlLat) + (mapSize))
25:     end while
26:     long = pixelToLong(longToPixel(tlLong) + (mapSize))
27: end while
```