Census Estimation Using Histogram Representation of 3D Surfaces: A Case Study Focusing on the Karak Region

Subhieh El-Salhi
Department of Computer Information Systems
The Hashemite University
Zarqa, Jordan
Email: subhieh@hu.edu.jo

Safaa Al-Haj Saleh
Department of Software Engineering
The Hashemite University
Zarqa, Jordan
Email:safan@hu.edu.jo

Frans Coenen
Department of Computer Science
University of Liverpool
Liverpool, UK
Email: coenen@liverpool.ac.uk

***Abstract*—National and regional infrastructure planning is founded on the use of many factors, of which population size can be argued to be the most fundamental. Population size is typically acquired through a census. However, manual census collection is an expensive and resource intensive process; especially in regions that are poorly connected. Computer-aided population estimation, when done accurately, therefore offers significant benefit. This paper presents a comprehensive framework for estimating the population size of a region of interest by applying classification techniques to terrain data. Central to the proposed framework is a novel histogram representation technique designed to support the generation of appropriate and effective classifiers central to the operation of the framework. The presented work uses the Karak region, in Jordan, as a case study for population size estimation. The proposed framework and the representation technique have been evaluated using a variety of classification mechanisms and parameter settings. The reported evaluation of the proposed representation technique demonstrates that good results can be obtained with regard to estimate the population size.**

***Index Terms*—Histogram representation; Geographic Information System (GIS); Population estimation; 3D surface; Satellite images; Data Mining; Classification technique; Karak region**

I. Introduction

Population size plays an essential role in a number of important governmental decision making strategies directed at the process of urban development. This is especially the case with respect to critical sectors such as health, education, transport and urban services. Governments usually obtain population size data using a census [1]. This is typically collected in either a door-to-door manner or by requiring individuals to complete a form. Both entail considerable cost, particularly in areas which feature a poor communication infrastructure; form filling also assumes a high degree of literacy amongst the population of interest, not necessarily always the case. A solution is to adopt some form of population estimation mechanism. However, it is important that the estimation is as accurate as possible so as to best inform strategic planning and decision making activities.

In the past, population estimation was conducted using "Aerial Photography"; however in the modern day aerial photography has been replaced with satellite imagery. The advantages of using satellite imagery over methods involving aerial photography can be summarised as follows: (i) extensive and regular coverage, (ii) time and cost efficiency and (iii) accessibility to regions which would otherwise be difficult to access [2]. Satellite images provide a rich data source accurately depicting terrain features such as mountains, hills, valleys, plateaus, plains, deserts and, importantly, population centres of all kinds. The key challenge is then how best to extract population estimations from such imagery? In times gone-by, using aerial photography, this was done by visual inspection; an undesirable approach because of the resource required and the subjectivity involved. Using satellite imagery the process can be automated; the challenge is how best to implement this automation. In both cases the approach is particularly effective in rural areas, areas where traditional census collection is the most resource intensive.

The proposed solution to the population estimation from satellite imagery problem, presented in this paper, considers the problem in terms of a 3-D surface where the x and y coordinates are the x-y coordinates of the centre point of each grid squares and the z coordinate is the number of dwellings (households) within the grid square (the value we wish to predict); the value at each 3-D location is then a terrain type. The idea is to represent individual locations in terms of their terrain type and neighbouring terrain types. The intuition is that population size, in a given region, is defined by terrain type. The basic idea is to use a machine learning approach whereby large-scale terrain image is input to a population estimation system. In more detail the input image is divided into a set of grid squares; with each grid square assigned a terrain type. A prediction model is then applied to predict the number of households in each grid square from which a population estimation can be obtained by multiplying the predicted number of households by an average household size. The challenge is the nature of the prediction model to be adopted, and how this model is to be generated. Most prediction (and classification) models take a feature vector as input. It is easy to see how this would work given tabular data,

but not as clear given image data. How the image data should be represented so that a good prediction model is generated is therefore an issue. The proposed solution is to adopt a histogram-based technique, an idea founded on the concept of local binary patterns [3].

To train the prediction model a dataset encompassing the Karak region of Jordan was used together with the known location of households within the test region. The Karak region was selected because it was considered to be a good exemplar region because of its geographical diversity.

The average household size data used in this case was obtained from a statistical study conducted by the Jordanian Department of Statistics [4], [5] and The United Nations [6]. The main contributions of the paper are:

1) A comprehensive framework to estimate (predict) population size (census) using terrain image data.
2) A novel technique to support the framework and represent the 3D surfaces depicted in 2D terrain image.

The rest of the paper is organised as follows. In section II a brief overview of related work is presented. The characteristics of the selected dataset are described in Section III. In Section IV the proposed framework for census prediction is introduced followed by its usage in Section V. The evaluation of the proposed framework is reported in Section VI using a variety of parameters. Finally some conclusions are discussed in Section VII.

II. Overview of related work

Automatic census estimation has recently emerged as a novel research area to overcome the drawbacks of traditional census collection approaches; namely the expensive of collecting the data, in terms of time and cost, especially in rural areas where the transportation infrastructure is typically insufficient. Generally speaking, automatic census estimation is inappropriate for inner city areas or commercial areas as it is difficult to distinguish between residential buildings and other types of building, and residential building shared by multiple families (such as tower blocks) and buildings occupied by a single family. Automatic census estimation is therefore most effective in rural areas where the main residential units are detached houses and the population is sparsely distributed [7], [8]; these are also the areas where traditional census data requires the greatest resource. There are a number of reports of research directed at census estimation founded on geographical features and labelled example data to produce a model that can then be used for census estimation purposes. The main challenge that most of the proposed census estimation techniques seek to address is how best to represent data extracted from GIS images so that classification techniques can be applied effectively. The need for a convenient representation arises from the fact that the extracted image data tends to be too large to be used in its raw form; so a representation is required that encapsulates the data in a reduced form.

The different data representation techniques presented in the literature include: colour histograms [7]–[10], local binary patterns [8]–[10] and graph based structures [9]–[12]. Colour histograms are used to represent the colour distribution in the image set. The X-axis of the histogram records "bin" identifiers, each bin representing a colour range; while the Y-axis represents the number of pixels falling into each bin. This representation offers computational simplicity and tolerance against changes such as rotation and scaling. Local Binary Patterns (LBPs) are essentially a texture representation technique that can be used more generally. The main idea of LBPs is to divide the examined image into cells and then compare the pixel values in each cell to each of its eight neighbours; if the value is greater than neighbouring value a "0" is recorded, otherwise a "1" is recorded. The resulting eight binary digits are then formed into an eight-bit number representing the cell. This representation has been widely used as it is easy to generate and robust to illumination changes. Graph based techniques have been used to represent the structure images, a frequently used representation is the Quad-tree representation that has been frequently used in the context of census estimation applications. Given a collection of quad trees frequent sub-graph mining techniques have been applied to identify frequently occurring sub-graphs that can then be used subsequently to generate feature vectors.

In addition to satellite imagery, various other forms of image have been used for census estimation. These include various types of Geographic Information System (GIS) map [7]–[10], [12]–[15], 3D satellite images (which include a Z dimension with which to represent 3D surfaces) [11], Ikonos images (high-resolution satellite images) [16] and aerial images (images taken from aircraft or other forms of flying object) [17].

Given a suitable feature vector representation different classifiers can be applied so as to generate a census estimation model, these include: K-Nearest Neighbour, Decision Tree, Naive Bayes, Logistic Regression and Support Vector Machine. The k-Nearest Neighbor (k-NN) technique is a lazy classification model that classifies a previously unseen object according to the majority class associated with the k nearest neighbours to the object [18], a variety of similarity (distance) measures can be used such as: Euclidean distance, Manhattan distance and Mahalanob distance. A Decision Tree is a decision support structure based on a tree-like graph of decisions and their possible consequences. Each internal node represents a test on an attribute, each branch represents the output of the test and each leaf node represents a class label [18]. Naive Bayes is a simple probabilistic classifier that assumes that there are no dependencies amongst features [18]. Logistic Regression is a classification model where the dependent variable is categorical. It can be binomial, ordinal, or multinomial. The goal of this classifier is to find the best fitting model to describe the relationship between the dependent variable and the independent variables [19]. A Support Vector Machine (SVM) is a supervised, non-probabilistic, binary linear classifier. By representing the training data as points in a space the SVM operates by separating the points with a dimensional hyperplane that divides the two categories by a gap that is as wide as possible [19].

The presented previous work directed at automated census

estimation has demonstrated that estimating census automatically has number of advantages, such as low cost and high speed of collection, compared to traditional census collection methods; although automatic estimation is less accurate than the real process of census collection. In the work presented by Reis et al. [13] a method for population estimation was developed based on demographic data extracted from satellite images where linear regression models were used to represent the extracted data. The developed method was evaluated using dataset representing districts of Belo Horizonte city in Brazil, the evaluation demonstrated that census results obtained using linear models were similar to these obtained using more complex models. Javed et al. [14] proposed a methodology for population density estimation using texton-based classification. Satellite images were split up into blocks and then textons for each block were extracted using Leung Malik (LM) filter bank. Extracted textons (texture features) were then used as training features for a k-NN classifier. Different satellite images of cities in Pakistan were used to test the proposed methodology. In the work presented in [16] the authors proposed a methodology for population estimation using Ikonos images of a district in Khartoum city where two categories of population unit were considered: dwelling units and residential areas. The presented approach can be summarised as follows. Residential buildings were separated from non-residential building using a number of selected features such as shape, size, pattern and texture. After that, all residential buildings were represented as polygons, after which vector layers for the represented polygons were generated. Consequently, residential buildings were classified into one of the categories and areas belonging to the residential category identified. An estimated population for each residential category area was then derived and a total population estimation computed. Kurtz et al. [15] proposed a multi-resolution method to generate clusters for the purpose of mapping homogeneous patterns of urban elements. The proposed methodology worked by clustering two types of images representing the same scene, Medium Spatial Resolution (MSR) images and High Spatial Resolution (HSR) images, by combining a per-pixel-based analysis with region-based analysis. The use of HSR images, which have a finer resolution than a MSR images, enabled the detection of urban objects such as houses, gardens and roads and their construction material (for example houses with green roof tiles).

There have been a number reports where classification models have been used to label households according to family size so as to predict number of people living in a previously unseen household; the households are typically segmented individually and represented using some feature vector representation. For example Yu et al. [11] presented a technique for classifying households described in terms of 3D surface images using Vertex Unique Labelled Subgraphs (VULSs). A VULS is a sub graph within some larger graph that has a unique vertex labeling. A Backward Match Voting (BMV) algorithm was used for classifying the vertex labels. Several census estimation frameworks have been proposed by Dittakan et al., several that used a histogram representation [7], another that used a quad-tree representation [12] and a third that used LBPs [8]. Dittakan et al. produced a fourth framework that used a combination of representations colour histograms, LBPs and a graph-based representation [9].

It is worth mentioning that the histogram techniques proposed in the previous works were based on generating different "color histograms" -including red green blue, hue, saturation, value and gray scale- and Local Binary Patterns to represent segmented households to generate feature vectors that were used to build classification models. While in the presented work a terrain image is divided into grids; whereas for each grid a "histogram representation" of the terrain type associated with the neighbours of the grid is generated.

III. Census Dataset

As noted in the introduction to this paper the Karak region was selected as the case-study region with which to generate the desired household prediction model. The Karak is rural region located at the southern part of Jordan (approximately 140 kilometres to the south of Amman) as shown in Figure 1. The Karak governorate's topography is characterised by a diverse terrain of desert (in the eastern part), mountains (the Mo'ab heights in the western part) and agricultural areas (known as the Ghour and Semi-Ghour). The Karak governorate was selected with respect to the work presented in this paper because of the wide range of different types of geographical terrain that it covers, compared to other Jordanian governorates. It is also sparsely populated, it has lowest population density with respect to the other governorates in Jordan (90.6 inhabitants per square kilometer [20]) which means that it is particularly well suited to the application of population estimation using satellite imagery. This was also a region where the number of households was known; information required to for training the number of households prediction model and for evaluation purposes.

Fig. 1. The Karak governorate in Jordan.

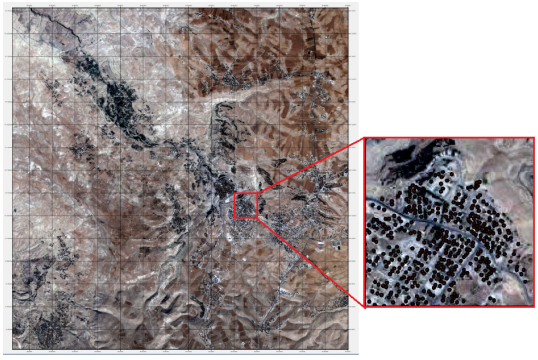

Fig. 2. Satellite image of the Karak region (left-hand side) and detailed showing household locations (right-hand side).

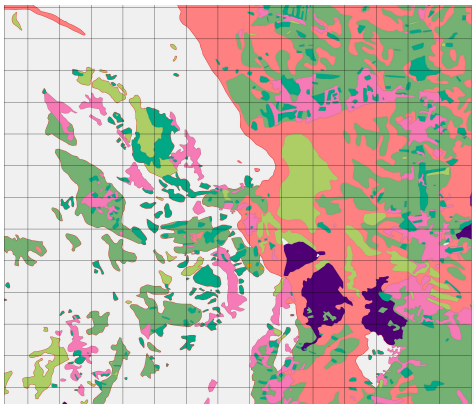

Fig. 3. Terrain map for the Karak region.

Figure 2 shows a 2D satellite image of the Karak test region with a grid superimposed. The example satellite image was released in April 2015 during the dry season and obtained from the Royal Jordanian Geographic Center (RJGC). The satellite image comprises $42,171,195$ pixels ($6,465 \times 6,523$). It covers a region bounded by the parallels of latitude $31°12'05''$ N (31.201388 N) and $31°25'00''$ N (31.416666 N) and the meridians of longitude $35°62'0''$ E (36.034722 E) and $35°75'00''$ E (36.250000 E). The figure also includes a detail with households marked with black dots. The household location data was also provided by the RJGC.

Figure 3 shows a terrain map corresponding to the same area covered by the satellite image given in Figure 2. The Terrain map highlights eight terrain types were considered: (i) desert, (ii) rocky land, (iii) orchards, (iv) green house areas, (v) bushes cover, (vi) cultivated area, (vii) uncrowded urban areas and (viii) crowded urban areas (a more detailed description of each is given in Section IV-A). Each terrain type is identified by a unique colour and a corresponding numeric code ($\{1, 2, \ldots, 8\}$). Both the terrain image and the corresponding satellite image of houses' distribution are used as input to the proposed population estimation system.

IV. Census Estimation Framework Generation

The generation of the proposed framework is presented in this section. The generation process comprises of the following phases (as shown in Figure 4): (i) a Pre-processing, (ii) a Surface Representation and (iii) a Classification. More detailed discussion concerning each phase in the following sections.

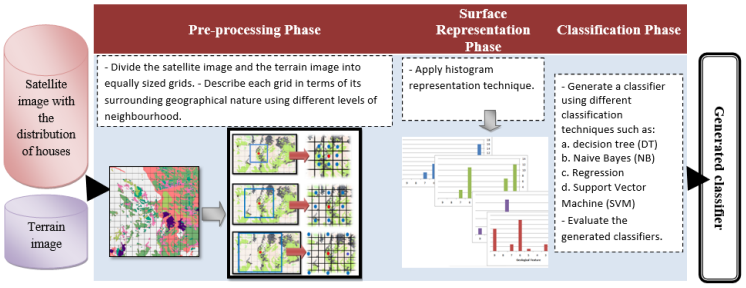

Fig. 4. The Histogram Representation Framework in the Context of Census Prediction.

A. Pre-processing Phase: Gird Dataset Generation

The inputs to the pre-processing phase are: (i) a terrain image and (ii) a corresponding satellite image of houses' distribution. Example satellite and terrain images of the Karak region were given in Figures 2 and 3. The two images are divided into equally-sized grids and for each grid the dominant terrain type is extracted and then the grid is labelled with the number of houses located within it. In more detail:

1) Using the given house "coordinate" data each house location is marked on the satellite image.
2) The marked-up satellite image is then converted into binary image so that house locations appear as white dots while the rest of the image appears as a black region. Figure 5 gives an example.
3) The terrain image and the corresponding marked-up binary satellite image are then divided into a set of equally-sized grids of size g pixels; for example if $g = 50$ the grids will measure 50×50 pixels. For the evaluation reported in Section VI a range of different values for g was experimented with. The generated grids are numbered sequentially row by row starting from the upper left corner and ending at the bottom right corner and thus the neighbours of each grid can be easily identified.
4) The grid squares in both images are then used to define a 3D surface where the z-dimension is the number of houses and the x and y dimensions are the x and y coordinates for the centre of each grid square. We can think of the surface as being a 3D lattice connecting grid square centres.
5) For each grid the dominant terrain type is extracted and then the grid is labelled with the number of houses located within it. The dominant geographic terrain type for each grid square is identified from the terrain image and is indicated by the dominant RGB colour in each

Fig. 5. Detail of the satellite image, with house locations marked, converted into a binary image.

grid. The number of houses in each grid is calculated by extracting the number of white dots for the grid in the associated binary image.

6) The number of surrounding neighbours of a given grid square is defined according to a neighbourhood distance parameter q. The maximum number of neighbouring grids is defined by $(2q+1)^2$. For example $q = 1$ will define neighbourhood of 9 grid squares, a distance of $q = 2$ neighbourhood of 25 grid squares, and so on. Experiments were conducted using a range of values for the parameter q, the results of these experiments is reported on in Section VI. A number of example neighbourhoods for $q = \{1, 2, 3\}$ are given in Figure 6.
7) Collate the grids data into a single data file. A fragment of such a file, using $q = 1$ is given in Figure 7. The fragment is taken from a set of $17,030$ grid squares. The first and last records in the fragment are corner gird squares so only have three neighbours, the remaining records included in the fragment are grid squares running along the top and bottom of the region and thus all have five neighbours. Looking at the first record in more detail, the dominant terrain type is "rocky land" (terrain identifier 2) and the number of household included is 2; and so on. The same fragment of the grid square dataset is presented in Figure 8 where the dominant geographic terrain type of each neighbour is shown in the last column.

B. Surface Representation Phase: Histogram Representation Technique

From the forgoing the generated grid data represents a 3D surface. However this does not readily lend itself the prediction model generation; prediction model generation is typically founded on a feature vector representation. The idea, for each grid square, is to generate a histogram of the terrain type associated with the grid square in question and its neighbours as defined by the distance parameter q. As noted earlier, eight terrain types were identified, which relate to eight bins (x-axis of the histogram). The y-axis of the histogram is then the number of occurrences of each terrain type in the current grid and its neighbours. Figure 9 gives an example histogram where $q = 3$ and consequently the size of the neighbourhood equates to 49 ($(2q+1)^2 = (6+1)^2 = 49$).

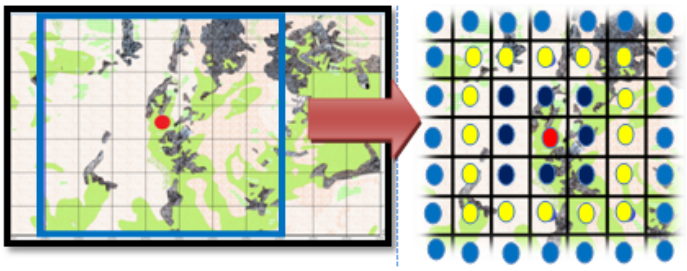

Fig. 6. Example neighbourhood for a given location (red dot) for $q = 1$ (dark blue dot), $q = 2$ (yellow dot) and $q = 3$ (light blue dots) superimposed onto a terrain image.

Grid Number	Dominant Terrain Type	Number of Houses	First Layer Neighbors
1	2	2	2 131 132
2	2	0	1 3 131 132 133
3	1	0	2 4 132 133 134
4	1	0	3 5 133 134 135
5	2	0	4 6 134 135 136
6	1	0	5 7 135 136 137
7	2	5	6 8 136 137 138
8	2	0	7 9 137 138 139
9	1	1	8 10 138 139 140
10	1	0	9 11 139 140 141
11	1	0	10 12 140 141 142
12	1	0	11 13 141 142 143
13	1	3	12 14 142 143 144
14	1	0	13 15 143 144 145
15	1	0	14 16 144 145 146
16	1	0	15 17 145 146 147
17	1	3	16 18 146 147 148
18	1	1	17 19 147 148 149
19	1	0	18 20 148 149 150
20	1	4	19 21 149 150 151
21	1	0	20 22 150 151 152
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
17026	1	0	16895 16896 16897 17025 17027
17027	6	13	16896 16897 16898 17026 17028
17028	6	12	16897 16898 16899 17027 17029
17029	1	2	16898 16899 16900 17028 17030
17030	1	0	16899 16900 17029

Fig. 7. A fragment of the grid square dataset for the Karak region using a grid size of $g = 50$ (50×50) neighbourhood of $q = 1$.

C. Classification Phase

Each histogram, describing a grid square (see above) can then be converted into a feature vector of the form $V = \{v_1, v_2, \ldots, v_8\}$. For the purpose of prediction model generation a prediction value p needs to be added to the feature vector. The collection of feature vectors then forms the input data set D to the prediction model generation. Experiments were conducted, see Section VI, using five different prediction model learning techniques.

Normalisation and *discretisation* processes are two main preprocessing stages required to operate the functionality of the classification phase successfully and efficiently.

Normalisation is a preprocessing stage used to maintain all the attributes of different ranges (scales) in the dataset in the same range and to assure that all the attributes of the dataset are treated equally (in our case the attributes of the population size and the number of surrounding neighbours

Grid Number	Dominant Terrain Type	Number of Houses	First Layer Neighbors
1	2	2	2 1 1
2	2	0	2 1 1 1 1
3	1	0	2 1 1 1 1
4	1	0	1 2 1 1 1
5	2	0	1 1 1 1 1
6	1	0	2 2 1 1 1
7	2	5	1 2 1 1 1
8	2	0	2 1 1 1 2
9	1	1	2 1 1 2 2
10	1	0	1 1 2 2 1
11	1	0	1 1 2 1 1
12	1	0	1 1 1 1 1
13	1	3	1 1 1 1 1
14	1	0	1 1 1 1 1
15	1	0	1 1 1 1 1
16	1	0	1 1 1 1 1
17	1	3	1 1 1 1 1
18	1	1	1 1 1 1 1
19	1	0	1 1 1 1 1
20	1	4	1 1 1 1 1
21	1	0	1 1 1 1 1
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
17026	1	0	1 1 1 1 6
17027	6	13	1 1 6 1 6
17028	6	12	1 6 1 6 1
17029	1	2	6 1 1 6 1
17030	1	0	1 1 1

Fig. 8. The same fragment of the grid square dataset for the Karak region presented in Figure 7 with the dominant geographic terrain type of neighbourhood.

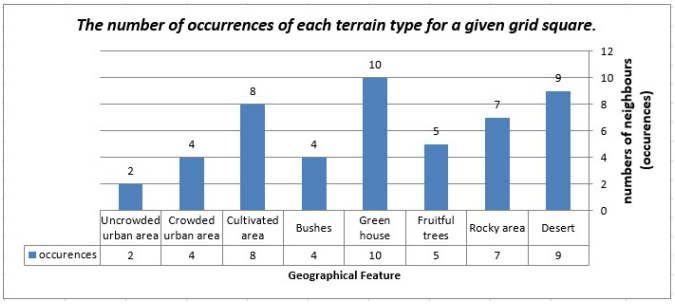

Fig. 9. Example histogram presents the number of occurrences of each terrain type for a given grid square and its neighbours with ($q = 3$).

of each predefined geographical features are derived from different scales) [21], [22]. There are different techniques of Normalisation such as Min-Max Normalisation, Z-Score Normalisation and Decimal Point Normalisation. However, empirical results indicated that Min-Max Normalisation outperforms other normalisation techniques in terms of simplicity, computational time and accuracy. Thus, in this paper the reported experiments were conducted using Min-Max Normalisation technique.

Discretisation is the process of transforming the continuous attributes into nominal attributes in such a way that the real values are mapped into one of a predefined set of intervals [23]. Nominal data types are essential for some classification techniques to operate efficiently such as Decision Tree (C4.5) and Naive Bayes (NB) [24]. Equal Width Binning (EWB) and Equal Frequency Binning (EFB) are the most popular *unsupervised* techniques of discretisation. Equal Width Binning divides the continuous data into set of intervals of equal size while the Equal Frequency Binning divides the continuous data into set of intervals of equal number of values (further details about discretisation can be found in [25], [26]). However, earlier empirical experiments of the presented work indicated that there was no significant difference between using either of them, therefore EFB was adapted for the entire set of experiments in accordance with Min-Max Normalisation technique.

V. Census Estimation Framework Usage

The census generation framework presented in the foregoing section, once trained, can then be used to predict the number of households in a given area. By multiplying this by the average number of inhabitants for a household. This will vary between geographic regions but the average number inhabitants for a Jordanian household, obtained from the Jordanian Department of Statistics (JDS), was 4.8 persons [4], [5]. Thus given a new region for which a population is to be estimated all that is required is a terrain map. This has to be segmented according to the same value of g used to train the system and feature vectors generated using the same value for q as originally used. Terrain maps are available for most geographic regions, the challenge will be converting these terrain maps so that they relate to the eight terrain types identified above. It may be possible to train a new model, using the process described in Section IV, but information concerning household locations would be required. An alternative is to automatically identify household locations. A mechanism for doing this was presented in [27], [28] with respect to a region of Ethiopia where the standard roofing material was corrugated iron which was relatively easy to detect in satellite imagery.

VI. Evaluation

This section describes the results obtained from the evaluation of the proposed Census Estimation Framework. A sequence of experiments were conducted to evaluate the effectiveness of the Framework using the Karak region test data. Experiments were conduct using five different prediction model generation mechanism: (i) k-NN coupled with Dynamic Time Warping (DTW), (ii) Decision Tree (C4.5), (iii) Naive Bayes (NB), (iv) Logistic Regression and (v) Support Vector Machine (SVM). A range of values for g from 50 to 400 pixels increasing in steps of 50, and three of values for q from 1, 2 and 3, were used. The framework was implemented using MATLAB $R2013a$, the JAVA programming language and Waikato Environment for Knowledge Analysis (WEKA)[1]. The experiments were run using a 2.4 GHz Intel Core $i5$ PC with 4 GB RAM and 64-bit operating system, running Windows 7 (SP1). The reported experiments were all conducted using real data obtained from RJGC for the Karak region where the total number of households is 9581. Ten fold Cross-Validation (TCV) (90% training set and 10% testing set) [29] was used throughout. The reported results are presented in terms of accuracy, Area Under the ROC Curve (AUC) and run time (in seconds). AUC and accuracy were selected as these are

[1]http://www.cs.waikato.ac.nz/ml/weka/

the most popular mechanisms used to assess the effectiveness of classifiers in machine learning and data mining research.

The objectives of the evaluation were:

- To identify the most appropriate prediction model generation mechanism to support the proposed Census Estimation Framework.
- To identify the most appropriate grid size with respect to the proposed histogram representation technique.
- To identify the most appropriate level of neighbourhood to be used (including whether a neighbpurhood should be considered at all).
- To analyse the run time requirements for the proposed framework.

The results are presented in graph form in Figures 10 to 15. Figures 10 to 12 give AUC values, and Figures 13 to 15 accuracy values. Each of the above objectives is discussed further, with respect to the reported results, in the following four sub-sections, Sub-sections VI-A to VI-D respectively.

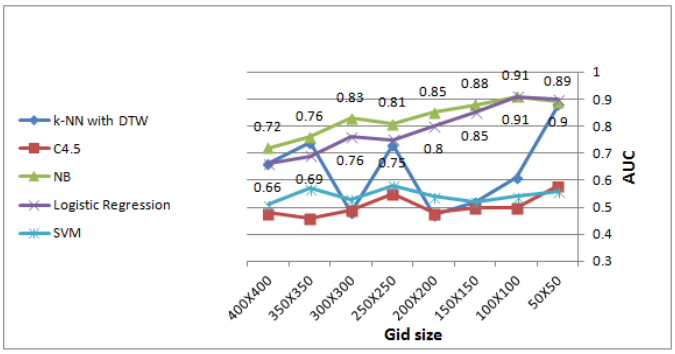

Fig. 10. The AUC results $q = 1$.

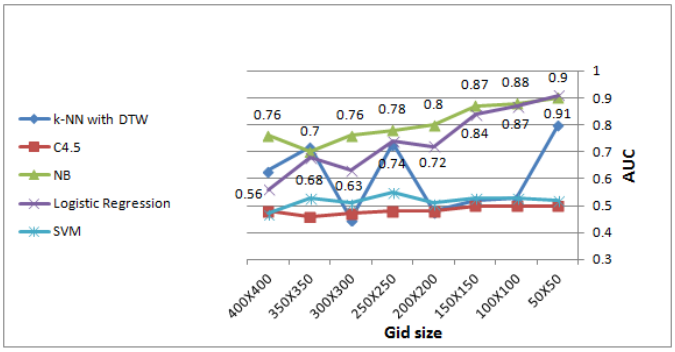

Fig. 11. The AUC results $q = 2$.

A. Prediction Model Generation Mechanism

From the results presented in Figures 10 to 15 the following observations can be made with respect to the most appropriate prediction model generation mechanism. The k-NN technique achieved the "best" accuracy results among all classification techniques considered, on occasion reaching 1.00 (100%) regardless of the values of g and q used. However, accuracy does not take into account any "clas bias", whereas the AUC

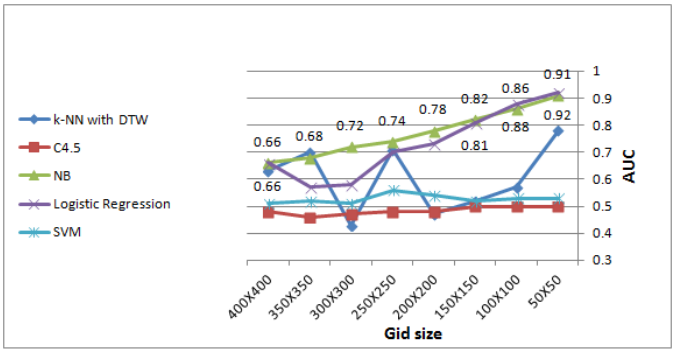

Fig. 12. The AUC results $q = 3$

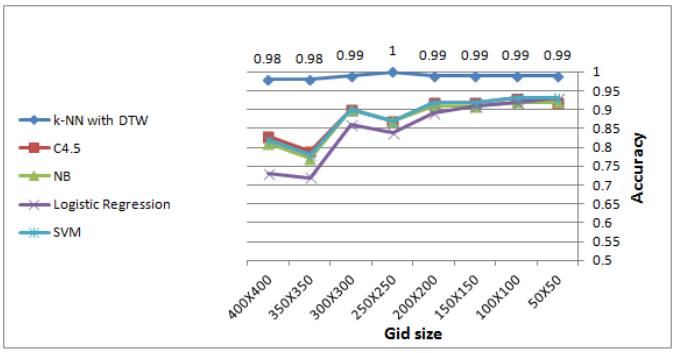

Fig. 13. The accuracy $q = 1$.

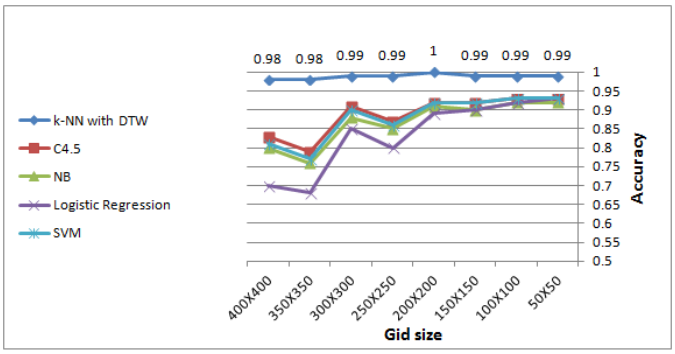

Fig. 14. The accuracy results $q = 2$.

Fig. 15. The accuracy results $q = 3$.

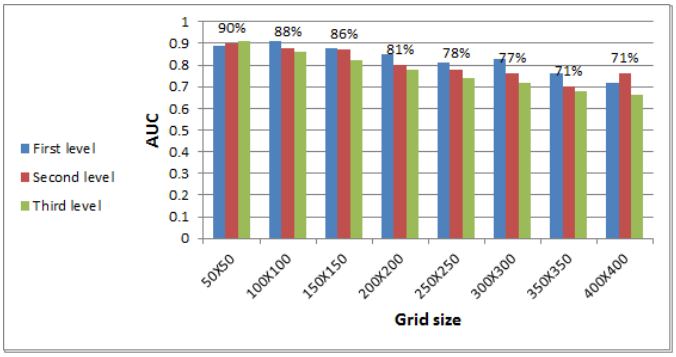

Fig. 16. The average of the AUC results of the NB technique for the first, second and third neighbourhood levels using different grid sizes.

metric does; the k-NN technique did not perform well in terms of the AUC metric. Note that an AUC value of 0.5 is equivalent to a guess while anything less than 0.5 is worse than a guess; hence both k-NN and C4.5 can be dismissed. Out of the remainder Naive Beyes produced the best overall performance in terms of the AUC metric; a best AUC value of 0.91. Although Logistic Regression also produced good results, again a best AUC value of 0.91, it was concluded that Naive Beyes was the most appropriate prediction model. A more detailed review of the results obtained using Naive Beyes is given in Figure 16.

B. Grid Size

With reference to the results presented in Figures 10 to 15 it can be seen that low grid sizes seem to be beneficial in the majority of cases; in the remainder of cases grid size had little effect. In terms of the best performing prediction mechanism, Naive Beyes, grid size has a clear impact with best performance being recorded when $g = 50$. The likely reason for this is that larger grid sizes will cover a larger area of terrain which may not appropriate description with respect to the current grid square. Therefore it can be concluded that the grid size of $g = 50$ (50×50) was the most appropriate.

C. The Analysis of Neighbourhood Levels

There is a degree of inter-operation between g and q, as g and q are increased the area "covered" will also increase. However, where g defines grid size q defines how many terrain descriptors are considered by the forecast model. Figures 17 to Figure 26 provide an alternative view of the results presented in Figures 10 to 15 so as to highlight the effect of using different values for q. Inspection of Figures 17 to Figure 26 indicates that there is no significant difference between the three different values for q considered ($q = 1$, $q = 2$ and $q = 3$); although in terms of AUC and using the DT technique $q = 1$ produced the best performance. Using Naive Bayes, the best performing forecast model as established in Sub-section VI-A, $q = 3$ produced best results but only in a very marginal manner. Therefore, it can be concluded that the first level of neighbourhood associated with smaller grid sizes is the "best" level of neighbourhood, it was conjectured that this was because the immediate neighbourhood of the current grid square gives a more accurate description of the nature of the current grid square.

An additional set of experiments was conducted with the neighbourhood distance set to zero, $q = 0$, using the Naive Beyes forecast model and $g = 50$ (because this pairing had been shown to produce best results), and comparing the AUC values obtained with the values obtained previously using $q = 1$, $q = 2$ and $q = 3$. The effect of using $q = 0$ is that the neighbourhood of a given grid square is not considered, on the the terrain value in the current grid square. The results are summarised in Figure 27. Inspection of the figure clearly indicates that inclusion of a neighbourhood in the forecast model is beneficial.

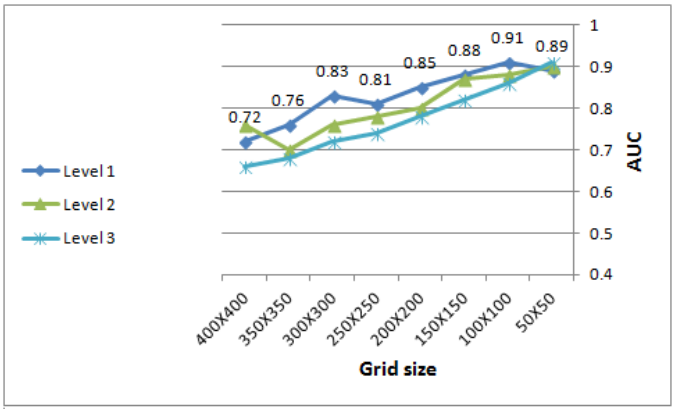

Fig. 17. The AUC results of the Naive Beyes technique using different neighbourhood distances (q) and grid sizes (g).

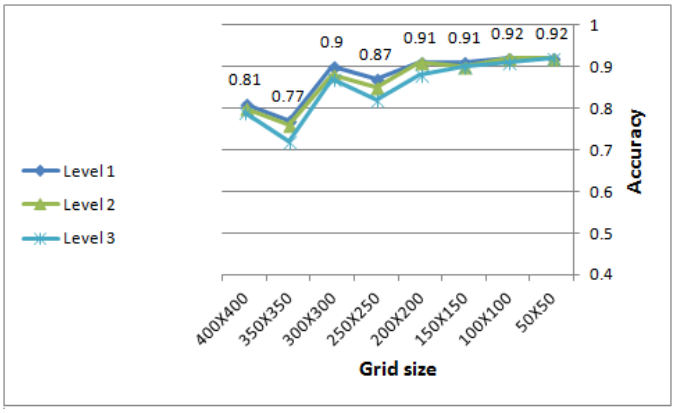

Fig. 18. The accuracy results of the NB technique using different neighbourhood distances (q) and grid sizes (g).

D. Run Time Analysis

Figures 28 present 32 give the runtime recorded for each experiment. The run times are average runtime over ten folds

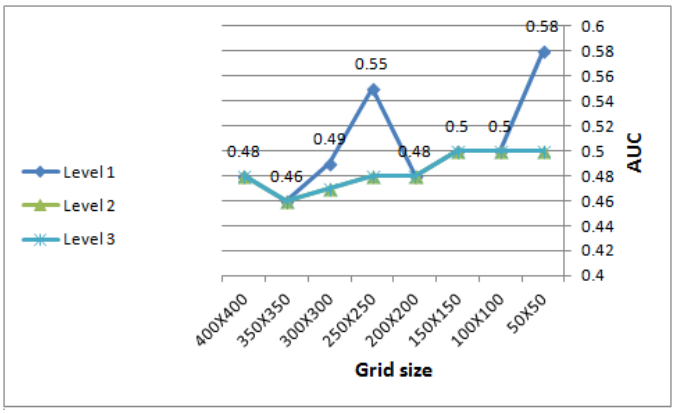

Fig. 19. The AUC results of the DT technique (C4.5) using different neighbourhood distances (q) and grid sizes (g).

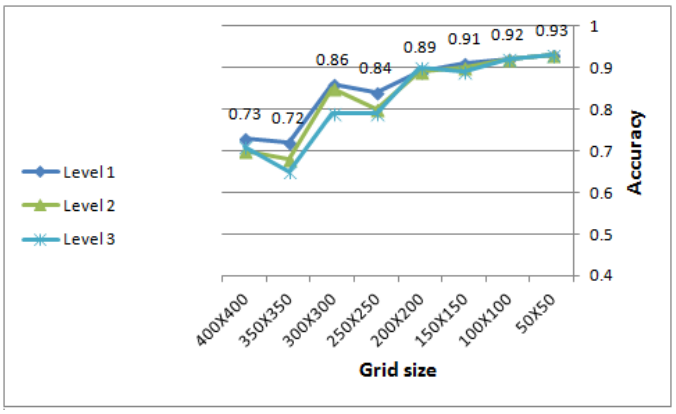

Fig. 22. The accuracy results of the Linear Regression technique using different neighbourhood distances (q) and grid sizes (g).

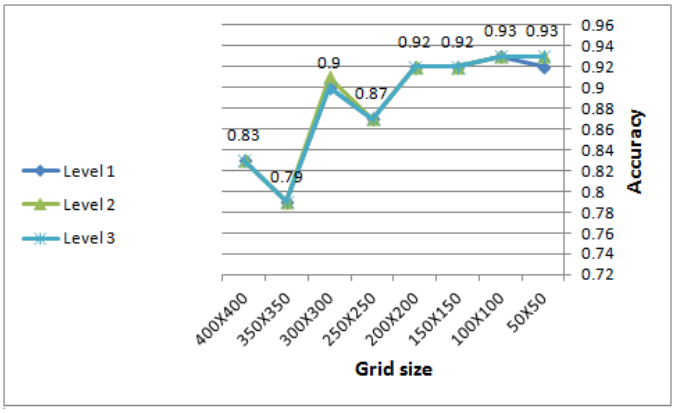

Fig. 20. The accuracy results of the DT technique (C4.5) using different neighbourhood distances (q) and grid sizes (g).

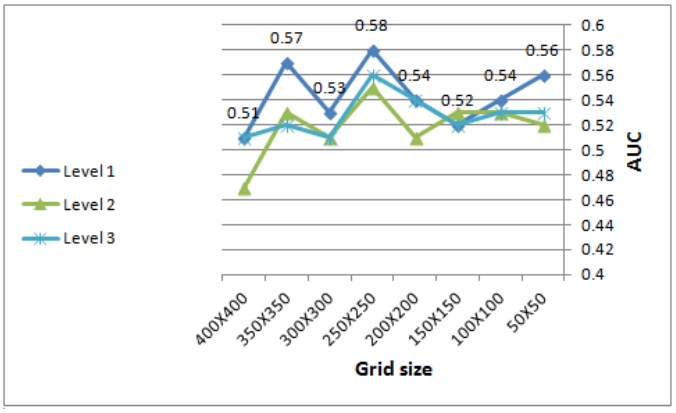

Fig. 23. The AUC results of the SVM technique using different neighbourhood distances (q) and grid sizes (g).

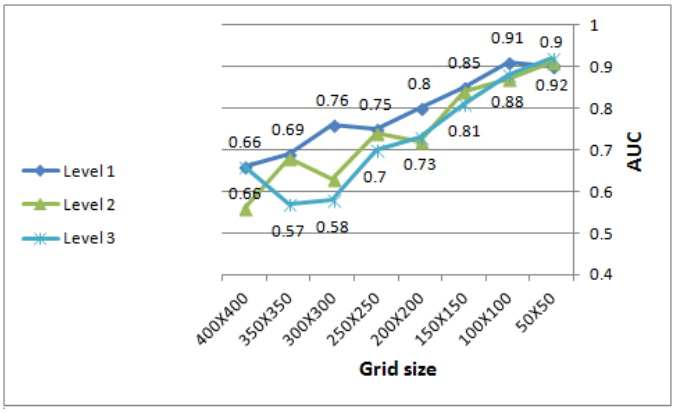

Fig. 21. The AUC results of the Linear Regression technique using different neighbourhood distances (q) and grid sizes (g).

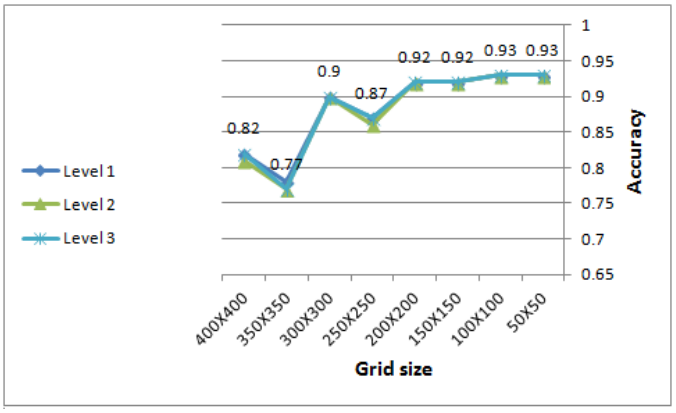

Fig. 24. The accuracy results of the SVM technique using neighbourhood distances (q) and grid sizes (g).

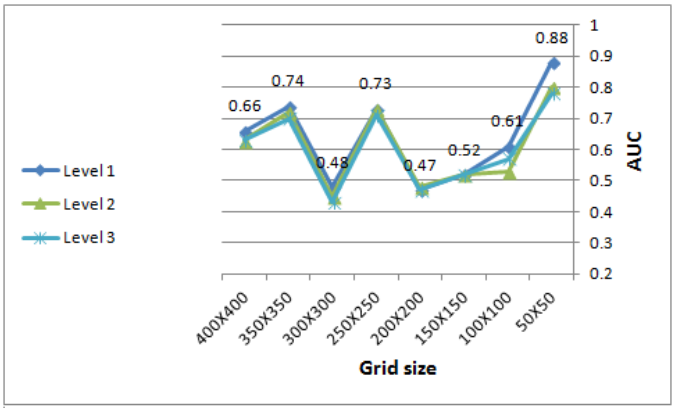

Fig. 25. The AUC results of the k-NN technique using different neighbourhood distances (q) and grid sizes (g).

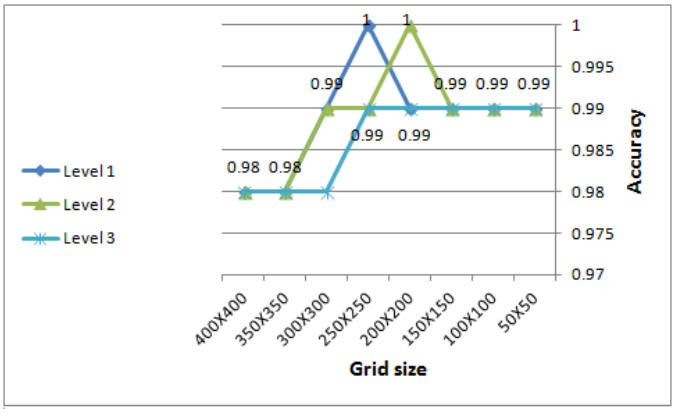

Fig. 26. The accuracy results of the k-NN technique using different neighbourhood distances (q) and grid sizes (g).

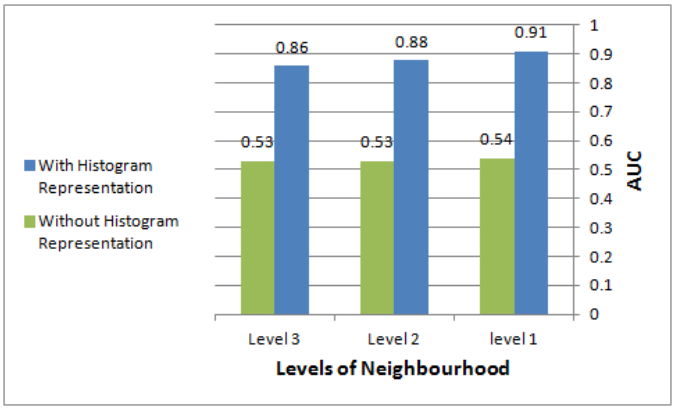

Fig. 27. The AUC results obtained, using Naive Bayes and $g = 50$, comparing $q = 0$ (no neighbourhood) with neighbourhood distances of $q = 1$, $q = 2$ and $q = 3$.

of cross validation. As anticipated, as the value of g was increase the runtime decreased, this was because as g was increasing the number of grid squares in the test region decreased. This is illustrated in Table I. Further inspection of Figures 28 present 32 indicates that the longest (worst) runtimes were recorded using Logistic Regression prediction models, and that the shortest (best) runtimes were recorded using Naive Bayes prediction models. The neighbourhood size tended to have little impact on the recorded runtime.

TABLE I
NUMBER OF GRID SQUARES OBTAINED USING DIFFERENT VALUES FOR g (GRID SIZE) WITH RESPECT TO THE KARAK TEST REGION.

g (grid size)	$\|D\|$ in terms of number of records (number of grid squares)
50	17030
100	4290
150	1936
200	1089
250	702
300	484
350	361
400	289

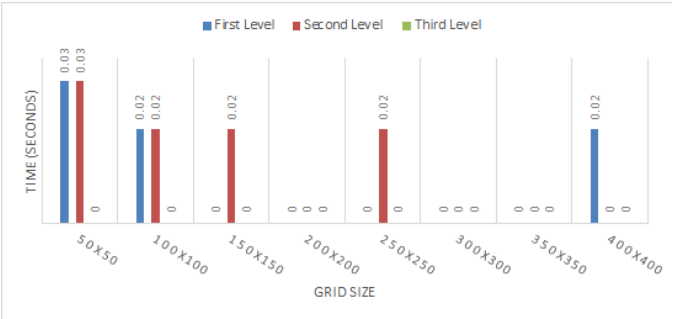

Fig. 28. The run time analysis of NB technique using different grid sizes (g) and neighbourhood distances (q).

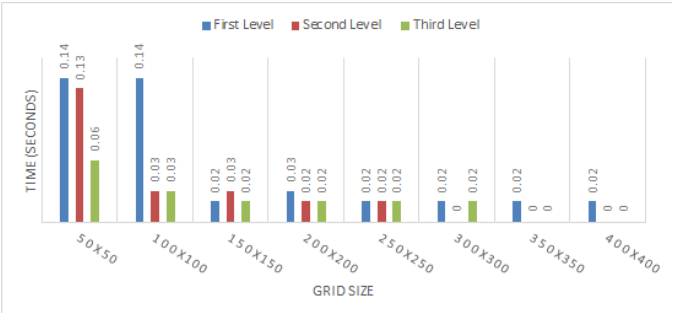

Fig. 29. The run time analysis of DT technique using different grid sizes (g) and neighbourhood distances (q).

VII. CONCLUSION

This paper has presented a comprehensive Census Estimation Framework based on a novel *histogram represen-*

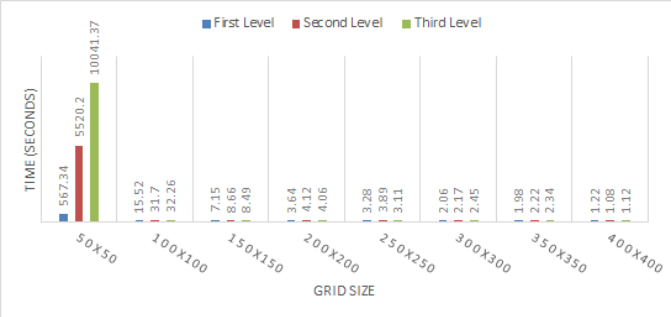

Fig. 30. The run time analysis of Regression using different grid sizes (g) and neighbourhood distances (q).

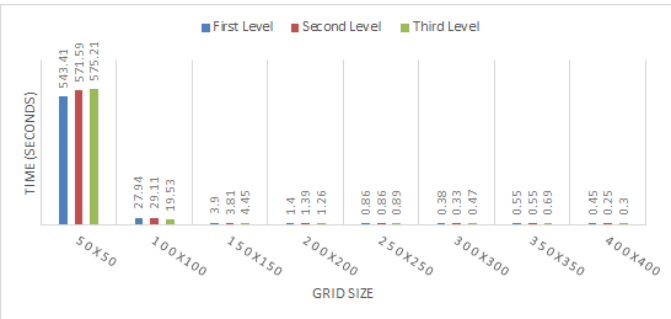

Fig. 31. The run time analysis of SVM technique using different grid sizes (g) and neighbourhood distances (q).

tation technique. The histogram technique incorporate with the framework to predict the population size based on the geographical nature of a given satellite image using data mining techniques, and more specifically classification techniques. The novelty of the histogram technique is the ability to capture the geographical nature of the local neighbourhood zone associated with the 3D surfaces of interest. The ultimate goal of the Census Estimation Framework is to generate an accurate population predictor (classifier).

The major strength of the described framework is the ability of the proposed histogram representation to effectively encapsulate the characteristics and the features of the 3D surfaces so that they can support the classification phase while constructing the desired classifiers. The evidence of the effectiveness of the proposed framework is demonstrated by: (i) the classification results obtained using a real dataset of Karak city in Jordan and (ii) the run time analysis of the framework using different classification techniques.

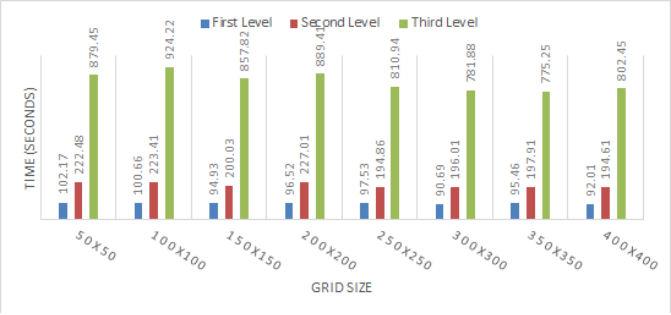

Fig. 32. The run time analysis of k-NN with DTW technique using different grid sizes (g) and neighbourhood distances (q).

The significant contributions of the work presented in this paper are discussed with respect to the theoretical and practical contributions. The theoretical contributions offered by the presented work can be listed as follows: (i) a novel 3D surface representation technique is proposed to express the local features effectively to support the process of classifier generation and thus the application of census prediction, (ii) a comparative study of the performance of different classification technique with respect to the 3D representation technique and finally (iii) a significant analysis of the Census Estimation Framework performance when the 3D surface representation technique (histogram representation) is employed. However, the practical contributions offered by the work presented in this paper may be clearly seen in the application of the proposed framework to generate census prediction classifier using satellite images, given a previously unseen region on the satellite image, where by the foreseen population that is likely to be in a certain location can be predicted correctly. Consequently, the population prediction would guide the decision-making process of policy makers to (i) plan the future uninhabited regions easily, (ii) guide the resource allocation process effectively and (iii) manage the development process carefully and this is of course with respect to the geographical distribution.

Some outstanding results were obtained as a consequence of a very comprehensive evaluation process (presented above) for employing the histogram representation in the framework. Naive Bayes technique along with the grid size of 50×50 and using the first level of neighborhood succeeded to achieve the best overall results. Furthermore, the remarkable results give an indication that the histogram representation technique enhances the operation of the proposed framework.

In the context of future work, further investigations for alternative representation mechanisms are needed. Moreover, additional experiments are required to determine the possibility to generating a generic classifier to predict the population using satellite images, not limited to the images obtained for regions in Jordan.

ACKNOWLEDGMENT

The authors would like to thank The Hashemite University for the continuous help and endless support. Authors also would like to acknowledge the Royal Jordanian Geographic Center (RJGC) for providing the data for this research. The comments and suggestions of the referees and editors are highly appreciated.

REFERENCES

[1] J. Mennis and T. Hultgren, "Intelligent dasymetric mapping and its application to areal interpolation," *Cartography and Geographic Information Science*, vol. 33, no. 3, pp. 179–194, 2006.

[2] C. Lo, "Automated population and dwelling unit estimation from high-resolution satellite images: a gis approach," *Remote Sensing*, vol. 16, no. 1, pp. 17–34, 1995.

[3] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, Nov 2011.

[4] Department of Statistics, "Jordan in Figures 2016," http://dosweb.dos.gov.jo/products/jordan-in-figures-2016/, [Online; Accessed: July 12, 2018].

[5] M. Ghazal, "Population stands at around 9.5 million, including 2.9 million guests," January 2016.

[6] U. Nations, "Household size and composition around the world 2017." Data booklet, http://www.un.org/en/development/desa/population/publications, 2017.

[7] 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 *SGAI Conf.* Springer, 2012, pp. 405–418.

[8] K. Dittakan, F. Coenen, and R. Christley, "Satellite image mining for census collection: a comparative study with respect to the ethiopian hinterland," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2013, pp. 260–274.

[9] 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 *ADMA (1)*, 2013, pp. 253–264.

[10] F. Coenen, "Mining satellite images for census data collection: A study using the google static maps service." in *KDIR*, 2016, pp. 7–8.

[11] W. Yu, F. Coenen, M. Zito, and K. Dittakan, "Classification of 3d surface data using the concept of vertex unique labelled subgraphs," in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 2014, pp. 47–54.

[12] K. Dittakan, F. Coenen, R. Christley, and M. Wardeh, "Population estimation mining using satellite imagery," in *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 2013, pp. 285–296.

[13] I. A. Reis, V. L. Silva, and E. A. Reis, "Adjusting population estimates using satellite imagery and regression models," *Anais XV Simpsio Brasileiro de Sensoriamento Remoto, SBSR*, pp. 830–837, 2011.

[14] Y. Javed, M. M. Khan, and J. Chanussot, "Population density estimation using textons," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. IEEE, 2012, pp. 2206–2209.

[15] C. Kurtz, N. Passat, P. Gancarski, and A. Puissant, "Multi-resolution region-based clustering for urban analysis," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5941–5973, 2010.

[16] A. S. Alsalman and A. E. Ali, "Population estimation from high resolution satellite imagery: A case study from khartoum," *Emirates Journal for Engineering Research*, vol. 16, no. 1, pp. 63–69, 2011.

[17] A. E. Ali, "Population estimation from aerial photographs: a case study from sudan," *Geography*, pp. 132–136, 1988.

[18] A. Ashari, I. Paryudi, and A. M. Tjoa, "Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool," *Int. J. Adv. Comput. Sci. Appl*, vol. 4, no. 11, pp. 33–39, 2013.

[19] N. Pochet and J. Suykens, "Support vector machines versus logistic regression: improving prospective performance in clinical decision-making," *Ultrasound in Obstetrics & Gynecology*, vol. 27, no. 6, pp. 607–608, 2006.

[20] "Karak Governorate," https://moi.gov.jo/EN/ListDetails/Governorates_and_Sectors/57/4, [Online; Accessed: October 18, 2020].

[21] L. A. Shalabi and Z. Shaaban, "Normalization as a preprocessing engine for data mining and the approach of preference matrix," *2006 International Conference on Dependability of Computer Systems*, pp. 207–214, 2006.

[22] S. G. K. Patro and K. Sahu, "Normalization: A preprocessing stage," *CoRR*, vol. abs/1503.06462, 2015.

[23] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning." in *IJCAI*, 1993, pp. 1022–1029.

[24] M. K. Ismail and V. Ciesielski, "Design and application of hybrid intelligent systems," A. Abraham, M. Köppen, and K. Franke, Eds. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2003, ch. An Empirical Investigation of the Impact of Discretization on Common Data Distributions, pp. 692–701.

[25] J. Han, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[26] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[27] F. Coenen and K. Dittakan, "Image representation for image mining: A study focusing on mining satellite images for census data collection," in *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. Springer, 2016, pp. 3–27.

[28] F. "Coenen and K. Dittakan, ""image representation for image mining: A study focusing on mining satellite images for census data collection"," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Cham: Springer International Publishing, 2019, pp. 3–27.

[29] C. Schaffer, "Selecting a classification method by cross-validation," *Machine Learning*, vol. 13, no. 1, pp. 135–143, Oct 1993.