Estimation of the Spatial Distribution of Urban Population using Remotely Sensed Satellite Data in Riyadh, Saudi Arabia

Alahmadi, M., Atkinson, P. and Martin, D.
School of Geography, University of Southampton, Southampton, SO17 1BJ, UK
Tel. (+44 (0) 23 8059 3808), Fax (+44 (0) 23 8059 3295)
mh_alahmadi@yahoo.com

Summary

This paper investigates the potential use of Landsat ETM+, remotely sensed height data (RSHD), ward-level population and dwelling units to provide population estimates at parcel-level. The paper applies a conventional method of population estimation and then refines this using the RSHD. Further refinements are applied using an advanced image classification algorithm. The model is calibrated at the block-level to downscale population to the parcel-level. The results indicate that the RSHD and the advanced image classification algorithm do improve the accuracies. However, low and high density areas are still difficult to predict accurately.

KEYWORDS: Downscaling population, land cover, remotely sensed height data, regression.

1. Introduction

Timely and accurate information about small-area population distribution is important. For example, it is crucial in decisions about when and where to construct public and private facilities such as mosques, schools, libraries, and markets. In addition, it is an important input in land use, transportation, and emergency planning. Census population is accurate but labour- and time- consuming in collection and updating and available only at coarse geographical units.

Since the 1950s, many methods of population estimation have been developed using remotely sensed satellite data and aerial photography. Population distribution is related to factors such as land use, transportation networks and distance from the city centre, which can be used to assist in determining where people should be allocated within the study area.

Holt et al. (2004) used Landsat to generate a population density surface. Lo (2003) employed an allometric growth model utilizing Landsat TM to estimate population and dwelling units. Langford (2006) provided a detailed comparison of global and regional regression models and dasymetric mapping. Lu et al. (2010) used Light Detection and Ranging (LiDAR) with satellite data and promising results were achieved.

The aim of this study is to develop reliable small-area population estimates. This will be achieved through applying a conventional method and then gradually refining this. A conventional method refers to use of coarse resolution satellite data to derive urban information such as built-area and then regression with population.

The study area is Um Alhamam ward, Riyadh, with a total population of 45,991. It has been selected as reflecting the variety of land use found across the whole area of Riyadh (Figure 1).

Census population, dwelling units, vector boundaries, Landsat ETM+, digital terrain model (DTM) and digital surface model (DSM) data have been obtained from different agencies in Saudi Arabia.

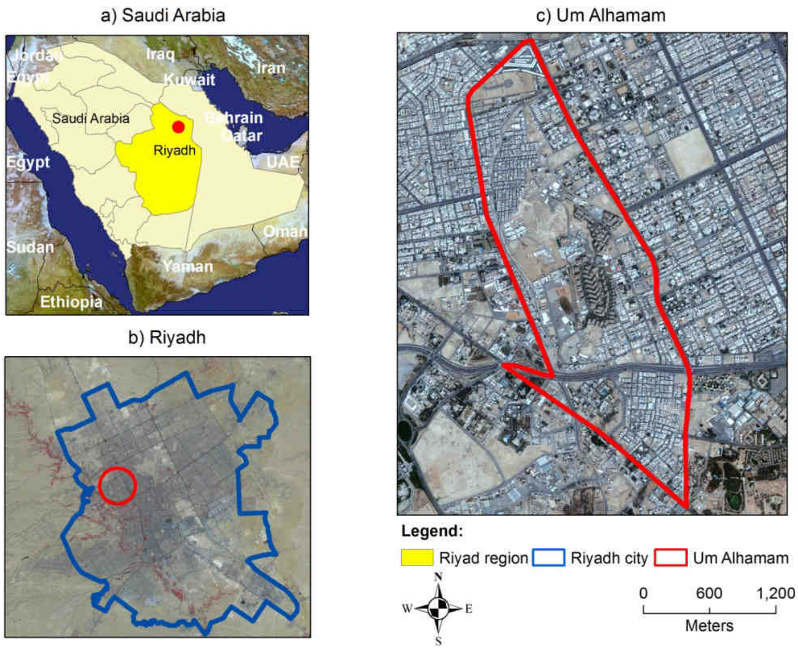

Figure 1. Site location.

2. Method

The census population is only available at ward-level in Saudi Arabia. Counts of dwelling units are available for smaller geographic units such as parcels. As expected, the correlation between population and dwelling units is very strong at 0.98. Thus, the density of dwelling units will be predicted using remotely sensed data, which could then be used to estimate population counts.

1- Iterative self-organizing data analysis technique (ISODATA) and support vector machine (SVM) are used to derive land covers.
2- RSHD is used to refine the conventional method.
3- An ordinary least square model is used to predict the relationship between the variables.

$$D_{ns} = a + bA \tag{1}$$

where D_{ns} is density of dwellings, A is built-area and $a\ and\ b$ are intercept and slope terms respectively.

4- Accuracy assessment is based on overall relative error (ORE), mean absolute error (MAE) and root mean square error (RMSE).

$$ORE = \frac{(\hat{D} - D)}{D} \times 100 \tag{2}$$

$$MAE = \frac{\sum_{k=1}^{n} |\hat{D}_i - D|}{n} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n} (\hat{D}_i - D_i)^2}{n}} \tag{4}$$

where $\widehat{D}$ and D are the total estimated and reference dwellings across Um Alhamam respectively and n is the number of zones. The smaller the values of ORE, MAE and RMSE, the better the model performance.

The built-area is represented in two ways. The first is to use built-area proportion (Figure 2a) and the second to intersect the built-area with non-zero height data (Figure 2b).

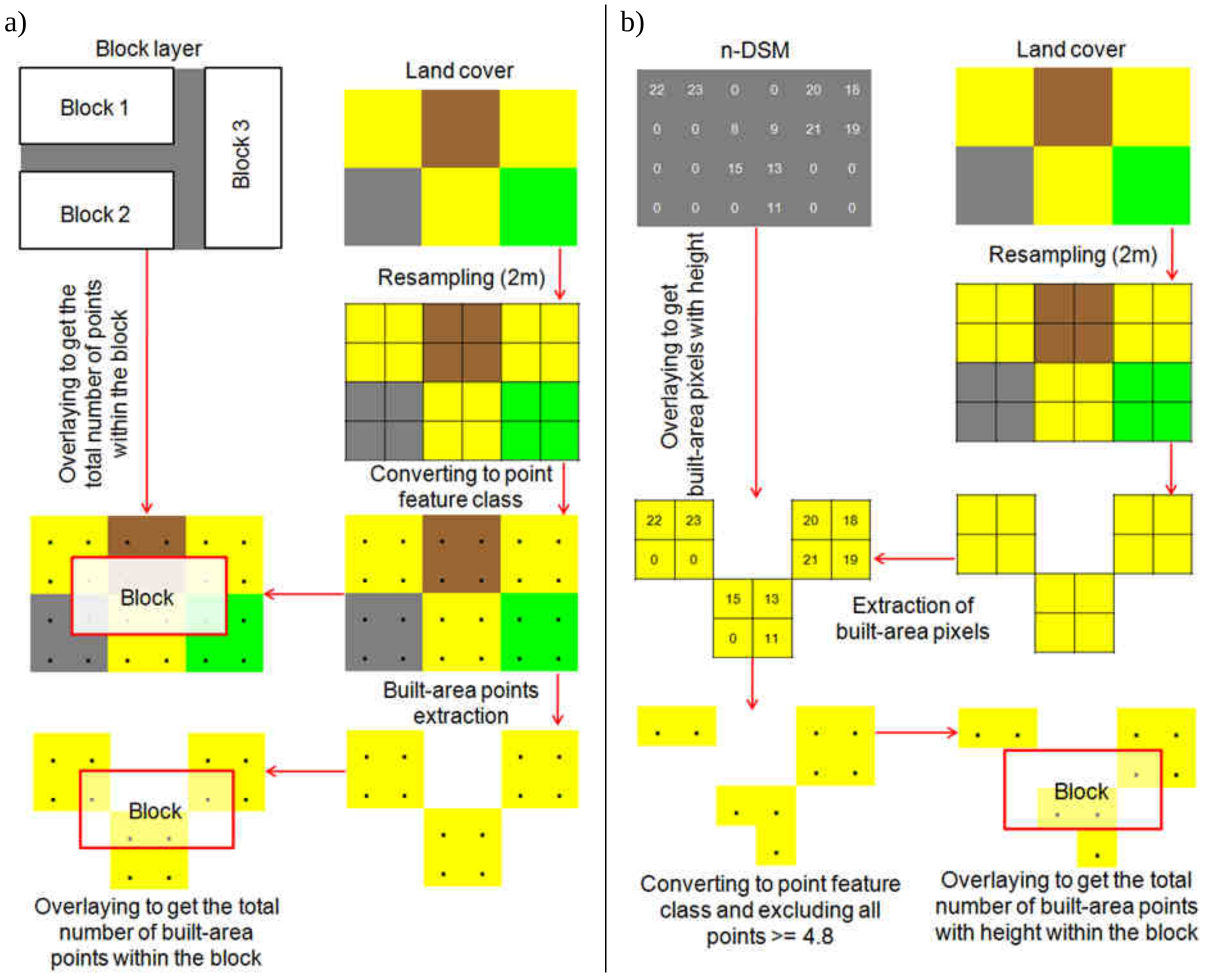

Figure 2. Independent variables: a) built-area proportion; b) built-area proportion with height**.**

The number of floors derived from the RSHD (Figure 3) provides further independent variables to refine the conventional method.

The building height = (number of floors * floor height) + average roof height

If the floor height is **3 m** = (2 * 3) + 1.8 = 7.8m

If the floor height is **4 m** = (2 * 4) + 1.8 = 9.8m

If the floor height is **5 m** = (2 * 5) + 1.8 = 11.8m

That means the building height of **2 floors** ranges from **7.8 to 11.8m**

Figure 3. Determining the height of a building with 2 floors based on the minimum and maximum floor height in Riyadh.

3. Results and discussion

3.1 ISODATA

Figure 4 show the relationship between the variables. The Pearson r of built-area proportion with height is higher than the built-area proportion (0.54 compared with 0.22 respectively). Unfortunately, exponential, growth and power models could not be applied as a result of the zero values in some blocks which have entirely non-residential uses. However, a square root transformation of the y-variable does slightly improve the Pearson r values. Thus, it was used.

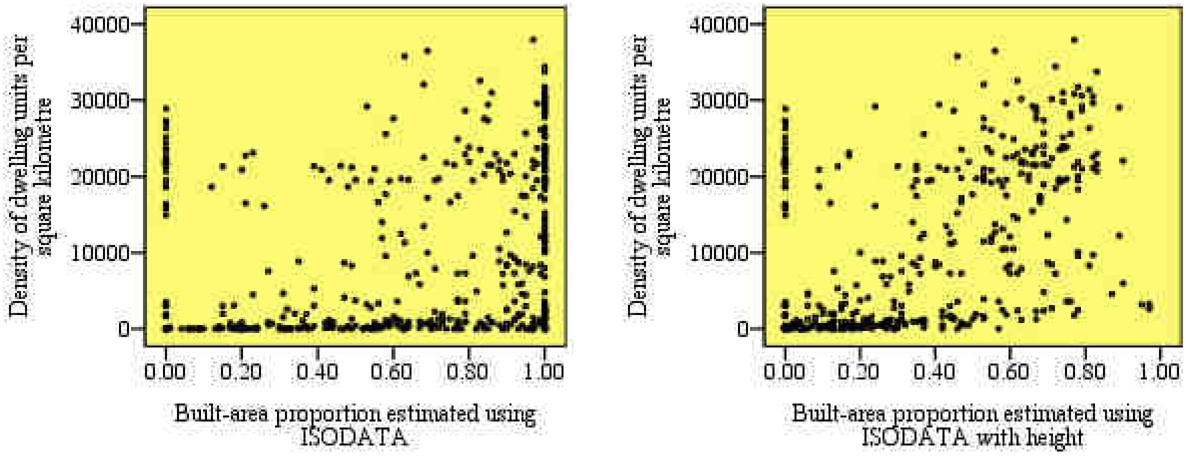

Figure 4. The relationship between the variables at the block-level.

All blocks were utilized for developing the model which will be used to downscale dwellings to the parcel-level.

Based on the conventional model (Model 1):

$$Density\ of\ dwelling\ unit = 5568.79 + 6534.73 * B_{ISO} \quad (5)$$

where B_{ISO} is built-area proportion estimated using ISODATA.

Based on the refined conventional model (Model 2):

$$SQRT\ (Density\ of\ dwelling\ units) = 1.48 + 129.57 * B_{ISO} + 22.67 * LRB + 90.53 * HRB \quad (6)$$

where $SQRT$ is a square root, LRB and HRB are low rise and high rise built areas respectively estimated from the RSHD.

The results of Models 1 and 2 are shown in Table 2.

Table 2. Results using ISODATA.

Parameters	Results	
	Model 1	Model 2
Real number of dwelling units	7,771	7,771
Estimated number of dwelling units	23,068	10,968
ORE	197%	41%
MAE	7.68	2.93
RMSE	409	156
True population	45,991	45,991
Estimated population	138,391	65,816

From Table 2, it can be clearly seen that utilizing the RSHD to refine the built-area and provide dummy variables substantially affects the accuracy. For example, the ORE and MAE in Models 1 and 2 are (197%, 7.68) and (41%, 2.93) respectively.

Residual errors of Models 1 and 2 are mapped at the parcel level in Figure 5a and 5b. Previous studies have reported that extremely high and low population density is difficult to predict using remotely sensed data alone. This problem is somewhat solved using Model 2. However, some parcels located in high density sub-wards have larger underestimates whereas some parcels located in low density sub-wards have larger overestimates. Figure 5c and 5d show the population distribution of Models 1 and 2.

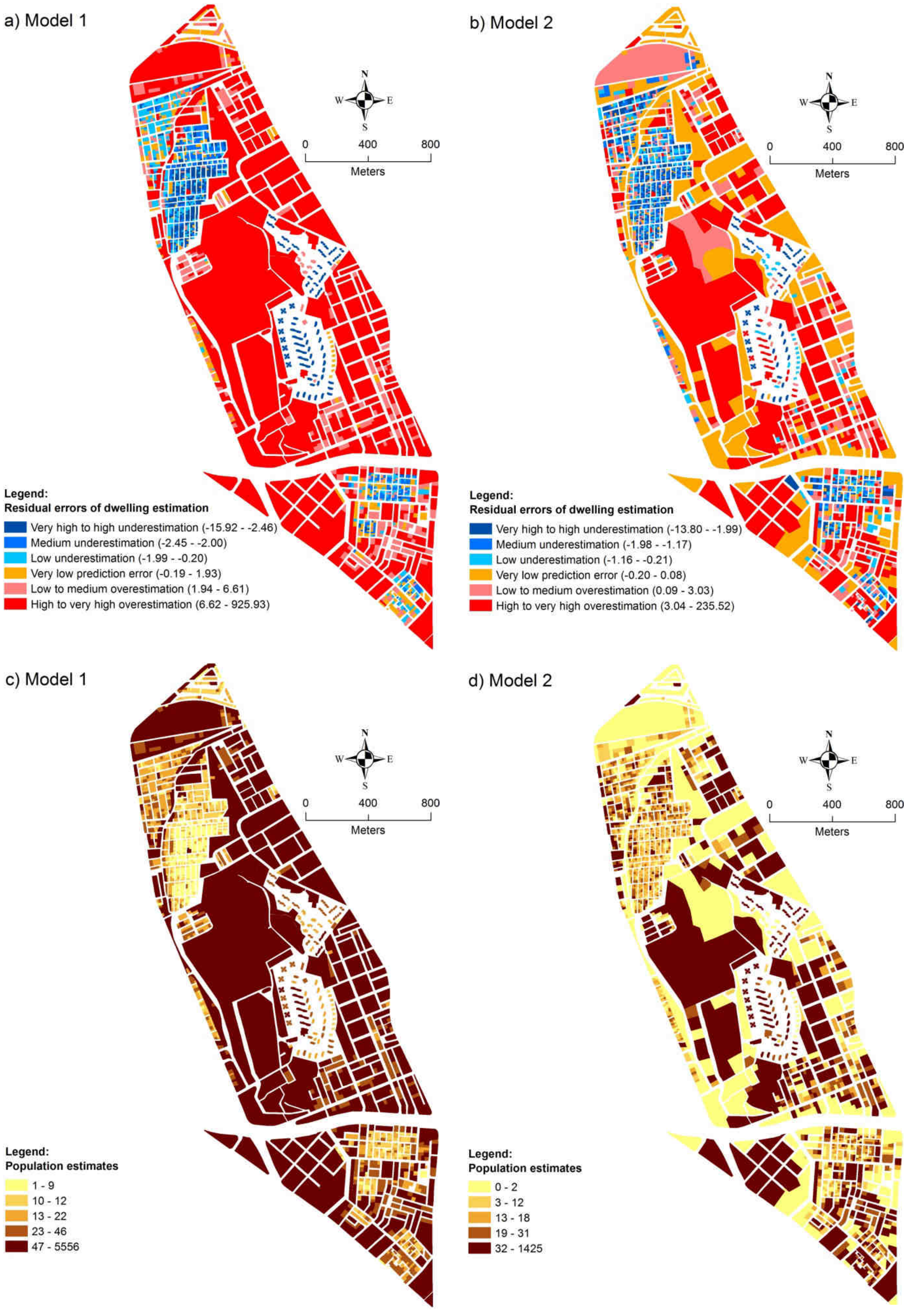

Figure 5. Models 1 and 2: dwelling unit residuals and population estimates.

3.2 SVM

There is no noticeable difference in the relationships between the different variables shown in Figure 4 when SVM is used. However, the land cover accuracies obtained from SVM are slightly more accurate than from ISODATA.

It has been seen how the RSHD affects prediction accuracies using ISODATA. We can now compare the effect of using the more advanced SVM algorithm.

Based on the further refined conventional model (Model 3):

$$\text{SQRT}\ (Density\ of\ dwelling\ units) = 1.48 + 127.74 * B_{SVM} + 16.6 * LRB + 78.7 * HRB \quad (7)$$

where B_{SVM} is built-area proportion estimated using SVM.

The results of Models 2 and 3 are shown in Table 3.

Table 3. Comparison results between Model 2 (ISODATA) and Model 3 (SVM).

Parameters	Results	
	Model 2	Model 3
Real number of dwelling units	7,771	7,771
Estimated number of dwelling units	10,968	10,254
ORE	41%	32%
MAE	2.93	2.68
RMSE	156	143
True population	45,991	45,991
Estimated population	65,816	61,495

From Table 3, it can be observed that utilizing an advanced algorithm such as SVM compared to ISODATA does improve the accuracy. For example, the ORE and RMSE in Models 2 and 3 are (41%, 156) and (32%, 143) respectively. There is no discernible difference between the pattern of errors from Model 2 (Figure 5b and 5d) and Model 3. Thus, maps of Model 3 are not presented.

4. Conclusion

The RSHD and SVM do improve the accuracies compared with the conventional method and they represent the first steps towards a range of possible refinements. However, overestimations and underestimations could be the result of (1) the different variances of the variables at the block- and parcel-levels; (2) the different residential uses within the built-area which are not considered; and (3) land cover classification errors. Thus, the authors plan to apply other solutions such as (1) providing three separate models for: low, medium and high density areas, (2) using an adjustment of built-area based on habitable volume, (3) applying dasymetric and volumetric dasymetric mapping, and then (4) using high spatial resolution satellite data such as QuickBird to derive a more detailed classification.

5. Acknowledgements

Thanks for King Abdulaziz City for Science and Technology for financial support and High Commission for the Development of Arriyadh and Space Dimension Technology for data.

6. References

Holt J B, Lo C and Hodler T W (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science,* 31(2)**,** 103-121.

Langford M (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems,* 30(2)**,** 161-180.

Lo C (2003). Zone-based estimation of population and housing units from satellite-generated land use/land cover maps. *In:* Mesev V (ed.) *Remotely Sensed Cities.* London: Taylor and Francis. 157-180.

Lu Z, Im J, Quackenbush L and Halligan K (2010). Population estimation based on multi-sensor data fusion. *International Journal of Remote Sensing,* 31(21)**,** 5587-5604.