A GIS APPROACH TO POPULATION ESTIMATION IN A COMPLEX URBAN ENVIRONMENT USING SPOT MULTISPECTRAL IMAGES

C. P. Lo
Professor of Geography
University of Georgia
Athens, GA 30602, U.S.A.

ABSTRACT

Population estimation was carried out using high-resolution (20 metres) multispectral satellite data (SPOT) in the Kowloon metropolitan area of Hong Kong, which is characterised by high population density and multipurpose high-rise buildings. The smallest areal unit for census data -- Tertiary Planning Unit (TPU) -- was used to aggregate population. With the aid of a raster GIS (IDRISI) and a vector GIS (pcARC/INFO), the satellite image data were registered with the TPU boundaries. A supervised image classification was carried out to extract high-density and low-density residential uses. Spectral radiance values of each band for each TPU were also extracted. Based on a sample of 12 representative TPUs, three-variable and one-variable linear regression models were developed, linking spectral radiance values with residential population densities. Number of living quarters estimation was also carried out in the same way. Residential land use area extraction from classified image was also employed to estimate population at the TPU level. The results showed over-estimation of population in all cases, but corrections to these estimates based on the proportion of high-density residential use in TPU, could be applied to those predominantly non-residential TPUs to improve the results of the estimates. It was concluded that the linear regression modelling and residential land use extraction could be efficiently combined under a raster GIS to obtain population estimates from high-resolution satellite image data at the very fine areal unit level. Further improvement of the accuracy of the population estimates requires research on methods of image enhancement, image classification, and the development of a knowledge-based expert system for a more realistic determination of population density for each census unit.

KEY WORDS: Population Estimation, GIS, High-Resolution Satellite Images, Regression Models, Image Classification.

1. INTRODUCTION

The methods of population estimation from aerial photographs have been well established, and the strengths and weaknesses of each method are quite clearly understood (Lo, 1986). In general, the most accurate approach requires the correct identification and enumeration of dwelling units from large-scale aerial photographs and an exact knowledge of household sizes of different dwelling types (Watkins and Morrow-Jones, 1985). However, this approach is only suited for use in small areas where the photograph coverage is not excessively large and the task of dwelling unit count is not unduly onerous. With the increased availability of computer technology, such an approach must be viewed as outdated. The author has experimented with the use of high-altitude aerial photographs (National High Altitude Photography at 1:80,000 scale) and space photographs (Large Format Camera photography at 1:756,000 scale) for the dwelling unit count approach to estimate population with the aid of a raster GIS (Lo, 1989). While this has facilitated the generation of estimated population data at the census tract level for the whole city in a short time, the accuracy of the result is not as good as that generated from large-scale aerial photographs because of the limitation of resolution in these high-altitude and space photographs. The raster GIS, however, is very effective and could have performed better if the input image data were in digital format. It becomes clear from these research endeavours that high-resolution satellite image data in raster format may be useful to automate population estimation for large areas with the help of GIS.

In previous research, low-resolution satellite images, notably Landsat MSS images with a spatial resolution of 79 metres, have been successfully employed to estimate the population of cities in China with the aid of the allometric growth model (Lo and Welch, 1977). However, for small area population estimation, only Ilisaka and Hegedus (1982) attempted to use the spectral radiance of the four MSS bands to estimate population within a grid of 500 metres by 500 metres dimension. Although they found strong correlation between population densities and spectral radiance mean values within the grid, and developed linear regression equations relating population densities with spectral mean values of four MSS bands, it was not made clear by the authors whether the population densities derived from these equations would produce accurate population estimates. Because of the low spatial resolution of the Landsat MSS data, the spectral radiance represents only the average of reflectance of different cover types. The residential use reflectance, on which population estimation has to be based, is highly diluted. This certainly will affect the accuracy of the population estimates.

Since February 21, 1986, the first SPOT imaging satellite has launched, making available 20-metre spatial resolution multispectral scanner data in three spectral bands (0.50-0.59 μm, 0.61-0.68 μm and 0.79-0.89 μm). The use of these higher resolu-tion satellite data for population estimation has not yet been fully tested. Because of the availability of these data in digital form, they can be readily imported to a computer and analyzed by a raster-based GIS. In this paper, such an evaluation based on a model 386/33 MHz personal computer, the IDRISI raster-based GIS, and the vector-based pcARC/INFO GIS is reported.

2. STUDY AREA AND DATA

The study area selected for this research is the metropolitan area of Hong Kong which comprises the twin cities of Victoria located on Hong Kong Island and Kowloon located on Kowloon peninsula. The harbor separates the two cities. Only the Kowloon metropolitan area will form the subject of this study. In 1986 the Kowloon metropolitan area had a population of 2,299,458 or about 42.6 percent of the total population of Hong Kong (Census and Statistics Department, 1987). Because of the shortage and high cost of land for construction, people live in overcrowded conditions in high-rise buildings. Kowloon contains the highest population density district of Hong Kong known as Mong Kok with a density of 139,600 people per square kilometre in 1986. As a typical colonial city, Kowloon displays an extremely mixed land use. A multistoried building is often multi-functional in nature. Residential use is often mixed with commercial and even industrial uses in the same building. This highly mixed building use renders dwelling units extremely difficult to identify. The Kowloon metropolitan area was chosen mainly because of its complexity which provides a genuine challenge to the population estimation methodology. Hong Kong has reliable population statistics acquired by the Census and Statistics Department through its decennial full censuses and intervening sampled censuses. These census data are available at the smallest areal units known as Tertiary Planning Unit (TPU) in Hong Kong. By aggregating the TPUs, the Secondary Planning Unit (SPU) is formed. This can go up to the Primary Planning Unit (PPU) for the first-order administrative districts of Hong Kong. The Tertiary Planning Unit population data for 1986 provide the "ground truths" for this evaluation. Finally, the author's familiarity with the Hong Kong urban environment helped him in the analysis of the image data.

The SPOT HRV mutlispectral image for the scene K286-J305 covering Hong Kong on 14th January, 1987 was used for this research. Coverage of SPOT data over Hong Kong is extremely limited, and is only available for the relatively cloud free winter season. The quality of the data is good, and the Kowloon metropolitan area only occupies about 4 percent of the total area of the 50 km by 50 km scene. The total number of pixel covering the study area is no more than 350 rows by 450 columns, thus making it suitable for analysis with a model 386 personal computer.

3. METHODOLOGY

In maximizing the advantages of a GIS approach, the population estimation methodology has been specially oriented towards the use of spectral radiance values of the three SPOT HRV bands and the population density approach. The SPOT data of the study area was first downloaded from tapes to floppy disks which were then input to the personal computer. A raster-based GIS-IDRISI -- was used to store the data as one byte pixels in band sequential format. Before these data could be used for image analysis, they had to be geometrically rectified for tilt and other image distortion. Five control points were picked by the SPOT multispectral images and their UTM coordinates were read from a 1:20,000 topographic map of Kowloon published by the Hong Kong Government. By selecting a first order polynomial fit in rectification and the nearest neighbor approach in resampling, each of the three spectral bands of Kowloon was rectified. All this was done under the RESAMPLE command of IDRISI.

The smallest areal units -- Tertiary Planning Units (TPUs) -- were employed as the basis for population aggregation. As has been indicated above, the study area was windowed out from the original SPOT scene of 3005 scan lines by 3171 pixels, and covered only 350 scan lines by 450 pixels because of graphics and storage limitations of the personal computer. A total of 44 complete TPUs of varying sizes (from a low of 56,785.4 square metres for TPU number 220 to a high of 1,819,583 square metres for TPU number 263) (Figure 1). The TPU boundaries were digitized with the aid of the ADS command in the vector-based pcARC/INFO GIS programs. Through data conversion in pc ARC/INFO, the digitized TPU boundary data were first converted to ERDAS "GIS" file which was then converted into IDRISI image format. Thus, the TPU boundary data were converted into the raster format. The TPU image file had to be rectified and resampled using the same five control points employed in the SPOT multispectral image rectification and resampling. This ensured that the TPU image could be registered perfectly with the SPOT multispectral images.

The EXTRACT command in IDRISI was used to extract the mean spectral radiance values from each band of the SPOT multispectral data for each TPU. The area of each TPU in terms of the number of pixels was also obtained by the AREA command of IDRISI. Population and living quarter data of each TPU acquired by the 1986 By-Census were also entered into the computer using the SAS package. By dividing these data with the pixel area of each TPU, population densities and living quarter densities by TPU were obtained. Pearson correlations among the TPU spectral radiance values, the population and living quarter densities were computed. These revealed a strong negative correlation of the mean spectral radiance values for Band 3 image (0.79-0.89 μm near infrared band) with the population density (-0.62 at 0.0001 level of significance) and with the living quarter density (-0.64 at 0.0001 level of significance). Based on this revelation, a sample of 12 TPUs were selected so that the full range of the mean spectral radiance values of Band 3 image (which varied from 20 to 45) were covered. In this sample, the correlation of the spectral radiance values of Band 3 image with population densities and living quarter densities were found to be very strong, respectively -0.91 and -0.92 at 0.0001 level of significance. The negative correlation indicated that higher spectral radiance values (light tones) in the image were related to lower population and living quarter densities. This sample provided a good basis to develop linear regression models to estimate population densities and living quarter densities at TPU level using spectral radiance values. The following sets of linear models were developed, one set for using spectral radiance values of all three bands of SPOT image and the other set for using spectral radiance values for Band 3 only of the image:

$$POPDEN = 372.7919-16.9422X_1+14.6943X_2-3.3255X_3 \quad \text{——— (1a)}$$

$$POPDEN = 129.0261-2.7390X_3 \quad \text{——— (1b)}$$

$$QUARDEN = 68.4092-2.6821X_1+2.3068X_2-0.7053X_3 \quad \text{——— (2a)}$$

$$QUARDEN = 29.3461-0.6188X_3 \quad \text{——— (2b)}$$

where POPDEN is population density per pixel, QUARDEN is living quarter density per pixel, and X_1, X_2 and X_3 are respectively mean spectral radiance values for Bands 1, 2 and 3 of the SPOT image at TPU level.

These sets of equations were applied to estimate the population and the number of living quarters for all the 44 TPUs in the Kowloon study area.

In addition to using the spectral radiance values, a more traditional approach to population estimation was also adopted. The IDRISI program allows image classification to be carried out. With the aid of a detailed land use map of Kowloon and Hong Kong produced by the Hong Kong Government at the scale of 1:30,000 in 1987, a supervised classification (maximum likelihood method) was employed to classify the SPOT multispectral image data of the study area into eight classes of land use/cover: (1) high density residential, (2) low density residential, (3) water, (4) industrial, (5) vacant, (6) woodland, (7) government and (8) recreational. For the purpose of population estimation, only the first two classes of land cover/use were of interest. The accuracy of the classification of the residential uses was determined by comparing with the Hong Kong-Kowloon land use map which was also digitized using the ADS program of pcARC/INFO. This was found to be 86 per cent. The resulting residential land use image was also rectified and resampled for registering with the TPU image so that areas of residential use at both high and low densities were extracted for each TPU of the Kowloon study area using IDRISI. From the digitized Hong Kong land use map and the population data of the 1986 By-Census, the residential population density for each TPU was determined. The residential densities varied from a low of 4.3 persons per residential area pixel for TPU #213 to a high of 128.2 persons per residential area pixel for TPU #281. By finding the upper and lower quartile values, the residential population densities for high-density residential use and low-density residential use were determined as 103 and 30 respectively. These two figures were used to compute the population size for each TPU based on the areas of high and low density residential areas of high and low density residential uses obtained from the supervised classification of the SPOT image data. The result of this approach to population estimation was then compared with that obtained by the use of the linear models of spectral radiance values.

4. RESULTS OF ANALYSIS

4.1 Population Estimation Based on Regression Models

The result of population estimation based on the spectral radiance values for the three SPOT image bands (Equation 1a) and that based on the spectral radiance value for Band 3 only (Equation 1b) are shown in Table 1. The results are not particularly accurate in both cases. The mean relative error for the estimates from the three-variable regression is 735.10 per cent while that for the one-variable regression is 854.93 per cent. These are caused by large over-estimation errors in some small-size TPUs with a predominantly non-residential use, such as, TPU #213 (railway terminal district with only 140 residential population), TPU #215 (a commercial district with only 868 residential population), TPU #216 (the tip of Kowloon peninsula called Tsim Sha Tsui which is predominantly commercial with only 560 residential population) and TPU #256 (an industrial area with only 1820 residential population). There are excellent estimates for a number of TPUs, notably, TPU#242 (with a relative error of -2.42 per cent), TPU #243 (with a relative error of -2.9 per cent), and TPU # 268 (with a relative error of -4.26 per cent), which are more purely residential districts. The population estimates from the single-variable equation exhibit similar magnitude and variations in the error pattern. In other words, the two linear equations do not make too much a difference, implying that Band 3, the infrared band, is an excellent estimator for population densities.

In order to correct for the problem caused by the high occurrence of non-residential use in the TPU, the per cent of high density residential use pixels classified by the maximum likelihood method of supervised classification is applied to the population estimates in those cases where the percentage is below 20 per cent. This correction helps to reduce the excessive population estimated for these essentially non-residential TPUs (Table 1).

4.2 Living Quarters Estimation Based on Regression Models

Estimation of the number of living quarters may be more appropriate than estimating population directly from the SPOT multispectral data. After all, the spectral radiance values in each band of the image data reflect the density of the built-up area in the city. The results of the estimation based on the three-variable regression model (Equation 2a) and the one-variable regression model (Equation 2b) are shown in Table 2. The mean relative errors are respectively 647.75 per cent and 718.17 per cent, which are slightly smaller in magnitude than those for the population estimation. The pattern of errors of living quarters estimation by TPU is identical to that for population estimation. The errors arise because of the occurrence of non-residential structures mixing with residential structures. Again, a correction factor using the percentage of high-density residential use is applied to those TPUs with less than 20 per cent high residential use. This helps to reduce the excessive estimation in these non-residential TPU. At present, the estimation of the number of living quarters does not seem to offer a better accuracy than the more direct population estimation approach because it is still necessary to establish household sizes in these living quarters before population can be estimated. Estimation of household sizes can give rise to errors.

4.3 Population Estimation Based on Residential Land Use Areas Extracted from SPOT Multi-spectral Images

Estimation of population based on residential land use areas requires a knowledge of the residential population density in each TPU. This can be achieved with reference to the population data of a sample of TPUs obtained from the 1986 By-Census. These data are then related to the actual area of residential land use provided by the digitized land use map of Hong Kong Government. Because of the great variations in residential density, the upper quartile figure of 103 persons/residential pixel for high-density residential use and the lower quartile figure of 30 persons/residential pixel for low-density residential use represent at best a comprise. The results of population estimation based on this method bear this out (Table 3). The mean relative error is 516.30 per cent. This represents only a marginal improvement in accuracy than that obtained by the use of linear regression models reported in sections 4.1 and 4.2 above. Examination of the error for each TPU reveals again the same problem TPUs where non-residential use predominates. The error is mainly caused by the inaccuracy inherent in the digital image classification as well as the use of a single population density for high-residential use and a single population density for low-residential use. Improvement in the method of digital image analysis and the use of different population densities for different TPUs in computing population from high- and low-density residential use pixels should give much better accuracy.

5. CONCLUSIONS

Despite the poor accuracy of the methods of population estimation using SPOT image data, this research points out a very useful GIS approach to automate population estimation at a very fine areal unit level. Normally, satellite data are good only for estimating population in much larger areal units. If we add up the estimated population by TPU in Table 3, the figure is 1,365,464 compared with the actual population of 1,323,056, thus giving an acceptable relative error of only 3.2 per cent.

An important finding from this research is the usefulness of Band 3 (0.79-0.89 μm) -- the near infrared band -- for population estimation. The raster GIS approach has been found to be most efficient in rasterizing, rectifying and resampling vector data of census district boundaries for registering with the satellite images. Extraction of residential land use from a classified image can be quickly carried out to produce another image for overlaying with population density per pixel. By multiplying the two images together, population estimates for each census district can be easily obtained.

Population estimation based on radiance values of the three spectral bands is obviously limited by the small range of the spectral values and by the large number of mixed pixels of residential and non-residential use. The complex urban environment of Kowloon exhibits high-rise buildings with multipurpose functions, thus making it difficult to estimate residential population accurately. Further research should concentrate on stretching and "purifying" the pixels through image enhancement techniques.

Despite the poor accuracy, the error pattern reveals some degree of regularity based on the percentage of high-density residential use pixels that corrections can be applied to those TPUs with predominantly non-residential use. Future research should focus on the development of a set of correction equations.

Finally, the research suggests that population estimation from high-resolution satellite images at small areal unit level should combine the regression model approach with residential land use extraction by digital image processing in a knowledge-based expert system within the GIS framework. The expert system will allow more realistic population density in each areal unit to be determined.

6. REFERENCES

Census and Statistics Department, Hong Kong, 1987. Hong Kong 1986 By-Census: Tertiary Planning Unit Summary Tables Kowloon and New Kowloon. Government Printer, Hong Kong.

Ilisaka, J. and Hegedus, E., 1982. Population estimation from Landsat imagery. Remote Sensing of Environment, 12:259-272.

Lo, C. P., 1986. Applied Remote Sensing. Longman, New York, pp. 40-70.

Lo, C. P., 1989. A raster approach to population estimation using high-altitude aerial and space photographs. Remote Sensing of Environment, 27:59-71.

Lo, C. P. and Welch, R., 1977. Chinese urban population estimates. Annals of the Association of American Geographers, 67:246-253.

Watkins, J. F. and Morrow-Jones, H. A., 1985. Small area population estimates using aerial photography. Photogrammetric Engineering and Remote Sensing, 51:1933-1935.

Table 1. Accuracy of Population Estimation Based on Regression Models by TPU.

TPU	Actual Population	Estimated Population Based on 3-Variable Equation		Estimated Population Based on 1-Variable Equation	
	No.	No.	Relative Error %	No.	Relative Error %
211	21000	86857	313.60	66259	215.52
212	21308	54366	155.00	62005	190.99
213	140	35268	25091.10	39069	27806.62
214	8505	9256	8.83	10885	27.99
215	868	23339	2588.65	26020	2897.73
216	560	15281	2628.72	16015	2759.87
221	108759	121553	11.76	118576	9.03
222	31535	40989	29.98	43761	38.77
220	3829	9035	135.95	9075	137.01
225	53431	62721	17.39	64224	20.20
226	10269	-87674	-953.77	47022	357.90
227	28910	33895	17.24	31827	10.09
228	22960	19835	-13.61	18179	-20.83
229	33222	37089	11.64	35013	5.39
231	6517	19741	202.92	19418	197.97
232	3892	20105	416.57	24648	533.31
233	8134	23224	185.51	26661	227.77
234	13797	26491	92.01	33143	149.22
235	18158	30046	65.47	31237	72.03
236	53263	28903	-45.73	40531	-23.90
237	19838	23227	17.08	20408	2.87
241	55762	47784	-14.31	48802	-12.48
242	65751	64161	-2.42	65182	-0.87
243	34314	33320	-2.90	37030	7.91
244	22204	35746	60.99	37291	67.95
245	27503	9073	-67.01	14285	-48.06
246	16205	14722	-9.16	20206	24.69
255	50309	55534	10.39	62012	23.26
264	71323	90181	26.44	82954	16.31
265	24045	31006	28.95	28265	17.55
266	63273	53135	-16.02	48406	-23.50
267	45969	42261	-8.07	35432	-22.92
285	47894	33583	-29.88	32380	-32.39
256	1820	23496	1190.97	28160	1447.25
261	13916	28213	102.74	47555	241.73
262	40271	10520	-73.88	22374	-44.44
263	83314	5937	-92.87	36762	-55.88
268	27069	25917	-4.26	29592	9.32
271	18277	59913	227.81	71005	288.49
272	12467	8366	-32.90	26765	114.69
281	196651	116155	-40.93	90219	-54.12
282	47236	32317	-31.58	14743	-68.79
283	68187	42597	-37.53	40610	-40.44
284	16870	47808	183.39	43065	155.27
Overall			735.10		854.93

Table 2. Accuracy of Living Quarters Estimation Based on Regression Models By TPU

TPU	Actual No. of Living Quarters	Estimated No. of Living Quarters Based on 3-Variable Equation		Estimated No. of Living Quarters Based on 1-Variable Equation	
	No.	No.	Relative Error %	No.	Relative Error %
211	5719	18473	223.00	15275	167.10
212	6349	13114	106.55	14264	124.67
213	35	8335	23713.15	9027	25692.01
214	2338	2245	-3.97	2500	6.93
215	273	5554	1934.29	6003	2099.05
216	420	3536	741.96	3689	778.35
221	24871	27642	11.14	27187	9.31
222	7455	9620	29.04	10065	35.01
220	1736	2077	19.66	2079	19.78
225	11956	14501	21.29	14703	22.98
226	4088	-10125	-347.67	10952	167.92
227	6902	7633	10.59	7281	5.50
228	5033	4434	-11.90	4156	-17.42
229	7217	8364	15.89	8011	11.00
231	1540	4513	193.05	4460	189.63
232	1078	5055	368.96	5712	429.92
233	2688	5650	110.21	6182	129.99
234	3528	6605	87.21	7675	117.54
235	4795	7012	46.24	7173	49.61
236	12026	7507	-37.58	9422	-21.65
237	4151	5226	25.90	4799	15.62
241	13181	11040	-16.25	11200	-15.03
242	14504	14752	1.71	14986	3.33
243	8190	7887	-3.70	8527	4.11
244	4592	8246	79.57	8608	87.46
245	6811	2349	-65.51	3399	-50.09
246	3696	3804	2.93	4691	26.91
255	13769	13246	-3.80	14282	3.72
264	16114	20056	24.46	19227	19.32
265	5586	6856	22.74	6613	18.38
266	13251	11835	-10.69	11117	-16.10
267	10514	9277	-11.77	8185	-22.15
285	12124	7614	-37.20	7481	-38.30
256	553	5977	980.86	6587	1091.05
261	4410	8578	93.16	11174	153.38
262	9912	3631	-63.37	5343	-46.09
263	20965	4886	-76.69	9218	-56.03
268	6874	6709	-2.39	7248	5.45
271	4130	15404	272.99	16864	308.33
272	3437	3520	2.43	6454	87.79
281	50603	25617	-49.38	21811	-56.90
282	11865	7009	-40.93	4009	-66.21
283	17159	9725	-43.32	9548	-44.35
284	3689	10634	188.26	9905	168.49
Overall			647.75		718.17

Table 3. Accuracy of Population Estimation Based on Extracted Residential Land Use Areas from SPOT

TPU	Actual Population	Estimated Population*	
	No.	No.	%
211	21000	47167	124.60
212	21308	61348	187.91
213	140	12047	8505.00
214	8505	9934	16.80
215	940	15997	1601.81
216	560	5471	876.96
220	3829	9437	146.46
221	108759	116323	6.95
222	31535	44355	40.65
225	53431	61077	14.31
226	10269	39678	286.39
227	28910	36861	27.50
228	22960	21321	-7.14
229	33222	37865	13.98
231	6517	12727	95.29
232	3892	11791	202.95
233	8134	14387	76.87
234	13797	24994	81.16
235	18158	18821	3.65
236	53263	28852	-45.83
237	19838	14031	-29.27
241	55762	43682	-21.66
242	65751	57114	-13.14
243	34314	31678	-7.68
244	22204	28040	-26.28
245	27503	11934	-56.61
246	16205	13738	-15.22
255	50309	43563	-13.41
256	1820	15974	777.69
261	13916	37960	172.78
262	40271	16532	-58.95
263	83314	38856	-53.36
264	24045	19631	-18.36
266	63273	58297	-7.86
267	45969	48619	5.76
268	27069	29451	8.80
271	18277	55088	201.41
272	12467	20574	65.03
282	47236	21318	-54.87
283	68187	25455	-62.67
284	16870	29106	72.53
285	47894	17454	-63.56
Overall			516.30

*Based on a high-density residential density of 103 persons/pixel and a low-density residential density of 30 persons/pixel.