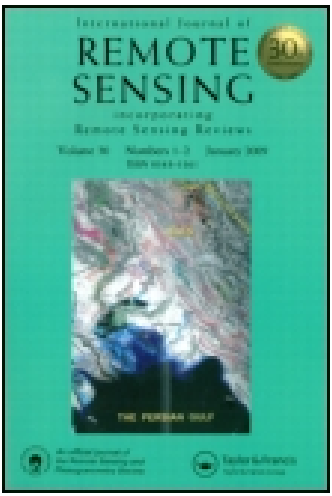


International Journal of Remote Sensing

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tres20

# Estimating census district populations from satellite imagery: Some approaches and limitations

J. T. Harvey

Published online: 25 Nov 2010.

To cite this article: J. T. Harvey (2002) Estimating census district populations from satellite imagery: Some approaches and limitations, International Journal of Remote Sensing, 23:10, 2071-2095, DOI: 10.1080/01431160110075901

To link to this article: http://dx.doi.org/10.1080/01431160110075901

INT. J. REMOTE SENSING, 2002, VOL. 23, NO. 10, 2071–2095



# Estimating census district populations from satellite imagery: some approaches and limitations

J. T. HARVEY

School of Information Technology and Mathematical Sciences, University of Ballarat, P.O. Box 663 Ballarat 3353, Australia; e-mail: j.harvey@ballarat.edu.au

(*Received 10 November 2000; in final form 29 May 2001*)



**Abstract.** Small-area population densities and counts were estimated for Australian census collection districts (CDs), using Landsat TM imagery. A number of mathematical and statistical refinements to previously reported methods were explored. The robustness of these techniques as a practical methodology for population estimation was investigated and evaluated using a primary image for model development and training, and a second image for validation.

Correlations of up to 0.92 in the training set and up to 0.86 in the validation set were obtained between census and remote sensing estimates of CD population density, with median proportional errors of 17.4% and 18.4%, respectively. Total urban populations were estimated with errors of +1% and −3%, respectively. These results indicate a moderate level of accuracy and a substantial degree of robustness.

Accuracy was greatest in suburban areas of intermediate population density. There was a general tendency towards attenuation in all models tested, with high densities being under-estimated and low densities being over-estimated. It is concluded that the level of accuracy obtainable with this methodology is limited by heterogeneity within the individual CDs, particularly large rural CDs, and that further improvements are in principle unlikely using the aggregated approach. An alternative statistical approach is foreshadowed.

## 1. Introduction

The task of population estimation differs in three important respects from many other applications of orbitally acquired data. Firstly, satellite data relate to land surface and land cover characteristics. Whilst it is true that in many remote sensing applications the variables of interest are aspects of land use which are only indirectly linked to the measured characteristics, in the case of population the link is even more tenuous and conjectural. For example, reflectances alone capture little if any information about the height and type of urban structures, which presents calibration problems at high densities. Secondly, estimating a quantitative variable like population across the spatial dimensions of an image is inherently more ambitious than the more usual qualitative objectives such as segmentation or classification. If the methodology is also required to be broadly applicable and robust to differences in season, geographical location, culture, etc., then the undertaking is doubly difficult. And thirdly, ground reference population data for model development is generally

*International Journal of Remote Sensing*
ISSN 0143-1161 print/ISSN 1366-5901 online © 2002 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/01431160110075901

only available at a much lower resolution than the remote sensing data itself, usually for census-related areal units.

Lo (1986) distinguished four different approaches to population estimation from remotely sensed imagery, based on:

1. counts of dwelling units;
2. measurement of areas of urbanization;
3. measurement of areas of different land-use; and
4. automated digital image analysis.

The first three methods had long been applied to the visual interpretation of analogue images from aerial photography at various scales. The fourth, particularly applied to orbitally acquired imagery, was seen as an emergent and radically different methodology.

In fact, in the fifteen years since Lo's categorization, most of the population-related research reported has not involved new approaches that are specific to digital image analysis. Most has involved digital implementations of Lo's second and third categories, though Lo's first approach too is now becoming feasible with the general availability of orbitally acquired imagery of adequate resolution.

The second of Lo's approaches, applicable at large regional scales with low resolution imagery, is based on the 'allometric' modelling of a direct mathematical relationship between the population of an urban area and its size. This approach was demonstrated using satellite imagery by Ogkrosky (1975) for cities in the US state of Washington, Lo and Welch (1977) for Chinese cities, and Stern (1984) for Sudanese villages. Similar large-scale correlations have been reported between population and indicators from later orbital sensors, such as night-time illumination (Welch and Zupko 1980, Sutton *et al*. 1997, Imhoff *et al*. 1997) and synthetic aperture radar (SAR) measurements (reviewed by Henderson and Xia 1997).

Lo's third approach, applicable at small to medium scales with medium resolution imagery, has been applied in various ways to population and related demographic characteristics. Increases in areas of urbanization have been monitored in many studies using techniques for change detection and land-use classification (for example Jensen and Toll 1982, Martin 1989, Martin and Howarth 1989, Quarmby and Cushnie 1989, Haack *et al*. 1996, Kressler and Steinocher 1996, Yeh and Li 1996, Kwarteng and Chavez 1998, Chen *et al*. 2000, Masek *et al*. 2000, Yeh and Li 2001).

Land-use classification has also been indirectly applied to small-area population estimation, with remote sensing imagery being used to provide ancillary data for the planning of censuses and demographic surveys (National Aeronautics and Space Administration 1978, Stern 1984, Olerunfemi 1986, Davreau *et al*. 1989, Booz-Allen and Hamilton 2000), or to enable the disaggregation of census counts at a finer spatial resolution (Fisher and Langford 1995, 1996, Yuan *et al*. 1997). The most extensive use of such a hybrid methodology is the LandScan Global Population Project (Dobson *et al*. 2000), in which census data, night-time illumination, land cover derived from various types of remotely sensed imagery, and other information about demography, topography and transportation networks, have all been combined to produce ambient population estimates for most of the world at a 30- × 30-second (approx 1- × 1-km) resolution.

A small amount of published research realises Lo's fourth category, by using modelling techniques which are only available within an automated digital image analysis framework, to directly estimate population at the small-area scale (in the

order of 1 km$^2$ or less) from satellite imagery alone. The present study falls into this category. Its direct methodological precursors are now reviewed.

In their seminal 1982 study, Iisaka and Hegedus used regression models based on Landsat Multispectral Scanner (MSS) data to estimate small-area populations in residential sections of suburban Tokyo. The image was resampled to 50 m × 50 m pixels co-registered with a 500 m grid for which census-based population data was available. The explanatory variables were mean reflectances of the four MSS bands calculated over the 10 × 10 pixel grid squares. Two equations, obtained by stepwise linear regression, expressed population as a linear function of the mean radiances of three MSS bands, with coefficients of variation ($R^2$) of 0.70 and 0.59 for the two census years studied. In this context $R^2$ can be interpreted either as the square of the correlation between the census and remote sensing estimates, or as the proportion of the spatial variation in the census figures associated with differences in spectral characteristics.

Langford *et al.* (1991) used a classification approach to estimate the populations of 49 census wards in northern Leicestershire. The explanatory variables were the numbers of pixels in each of five land use categories (industrial/commercial, dense residential, ordinary residential, uninhabited, agricultural), obtained by supervised classification of a Landsat TM image. $R^2$ values of 0.85, 0.82 and 0.75 were obtained for multiple regression models with all five explanatory variables, two explanatory variables (dense and ordinary residential), and one explanatory variable (all residential) respectively. However, some models were forced through the origin, and the basis of the reported $R^2$ values was not clear.

Lo (1995) used a mixture of both types of predictor (mean reflectances and counts of pixels in classes) to estimate the population and dwelling unit numbers in 44 tertiary planning units (TPUs) in Kowloon, Hong Kong, using multispectral SPOT imagery. Five different regression models were reported for each. In four cases, the form was linear and the dependent variable was population or dwelling density. The explanatory variables were: means of SPOT bands 1, 2 and 3; mean of SPOT band 3 alone; mean population per pixel in high and low density residential classes; proportion of pixels in high density residential class. In the fifth case, the form was logarithmic, the dependent variable was population or dwelling count, and the explanatory variable was the number of pixels in the high density residential class. In each case the models were estimated using 12 TPUs and then validated by application to the full set of 44 TPUs. $R^2$ values were reported for only the fourth and fifth of these models in the training phase, the values being 0.88 and 0.77, respectively. Results for the full set of TPUs were summarized in terms of the relative error in the total estimated population, which ranged from −5.3% to +5.3%, and the mean of the absolute values of the proportional errors for individual TPUs, which after deletion of four extreme outliers ranged from 64% to 99% for the different models. Corresponding results for dwelling unit estimation varied from −10.1% to +5.0% and from 50% to 77%. Whilst the overall totals were estimated with reasonable accuracy, the figures for individual TPUs were not, illustrating the difficulty of applying remote sensing methodology to an area of very high population density including many multi-storey and multi-functional structures.

Webster (1996) developed models for estimating dwelling densities in the 47 suburbs of Harare, Zimbabwe. The explanatory variables, derived from both SPOT and TM images and based on a subsample of pixels within each suburb, were characterized as measures of tone (six TM bands); measures of texture (three measures

derived from a classification of pixels into urban and non-urban: urban pixel density, homogeneity and entropy); and one measure of context (distance from the city centre). Results from five models were reported, one based on each of the three texture variables in turn, and two (with two and three explanatory variables) selected using stepwise regression. Reported $R^2$ values were in the range 0.69 to 0.81.

Webster also reported a more extensive analysis for estimating the numbers of dwellings in 65 grid squares on a transect through Cardiff, Wales, which were co-registered with a dwelling count database. Of a reported 70 texture statistics investigated, the seven chosen by stepwise regression were described as measures of 'edginess' and 'ripple', generated using line detection algorithms and Fourier and Laplace transform methodology. $R^2$ values for linear and logarithmic models were reported as 0.86 and 0.97, respectively. The latter value was inflated by a forced zero intercept (as pointed out by the author) and probably also by the logarithmic form of the model. Absolute and relative errors for each grid square were listed but no summary statistics were reported.

These studies demonstrated that there is a degree of correlation between population or dwelling counts and various remote sensing indicators. However, if models based on remote sensing data are to be seriously considered as a practical operational methodology for population estimation, the robustness of the relationships must be demonstrated by validation on data other than that used for model development. Lo (1995) carried out some such validation, though the results were not compared with those for the training data. Langford *et al.* (1991) carried out a more limited form of cross-validation by deriving estimates for a second set of areal units overlaid on the same geographical area as the training set.

The aim of the research reported in this paper was twofold: to apply and evaluate some mathematical and statistical refinements to the regression methodology reported in the cited papers; and to more thoroughly evaluate validity and robustness, by using one image to identify and train models and applying these models to a second image of a nearby culturally and demographically similar area, obtained on the same date.



## 2. Study areas and data

### 2.1. *Study areas*

The locations of the study areas are shown in figure 1. The primary study area was Ballarat Statistical District (BSD), an inland region of some 634 $km^2$ in extent, centred on the provincial city of Ballarat, 110 km west of Melbourne, Victoria, Australia. The Australian Bureau of Statistics (ABS) conducts a Census of Population and Housing every five years. The smallest geographical unit for which Australian census figures are published is the Census Collection District (CD). For the 1986 census, BSD comprised 138 CDs, of which 122 were classified as urban, using a population density threshold of 200 persons/$km^2$ supplemented by contextual rules aimed at reducing fragmentation (ABS 1998). Urban CDs ranged in area from 0.09 $km^2$ to 3.28 $km^2$, with a mean area of 0.54 $km^2$. Rural CDs ranged in area from 8.63 $km^2$ to 118.93 $km^2$, with a mean area of 34.23 $km^2$.

The secondary study area was Geelong Statistical District (GSD), a similarly mixed urban/rural area of some 352 $km^2$ in extent, centred on the port city of Geelong, 90 km south east of Ballarat. GSD comprised 225 CDs, of which 214 were classified as urban. Urban CDs ranged in area from 0.05 $km^2$ to 3.74 $km^2$, with a

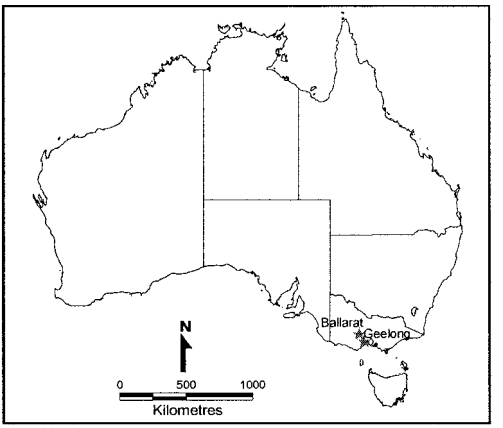


Figure 1. Location of study areas.



mean area of 0.50 km$^2$. Rural CDs ranged in area from 0.97 km$^2$ to 45.05 km$^2$, with a mean area of 22.25 km$^2$.

2.2. *Ground reference population data*

Estimated resident population (e.r.p.) figures are produced annually by ABS for larger geographical areas called Statistical Local Areas (SLAs), but not for individual CDs. In non-census years, mathematical models are used to estimate population changes from the census baseline, employing a range of data such as births, deaths, school enrolments and building approvals.

An interpolation and apportioning procedure (Harvey 1999) was used to estimate a ground reference population for each CD on the date of acquisition of the TM image, 14 February 1988. Essentially, the rate of change of each CD during the intercensal period 1981–1986 was compared with that of the SLA in which it fell, and the differential was then used, together with the annual SLA estimates for June 1987 and June 1988, to estimate the population of each CD on the required date. Overall, the estimated population of BSD was 79 179, of which the urban area contributed 70 222. The estimated population of GSD was 147 910 of which the urban area contributed 142 250.

2.3. *Census collection district boundaries*

Digitized 1986 census boundaries of the CDs in the two study areas were obtained from the Australian Survey and Land Information Group (AUSLIG).

2.4. *Landsat Thematic Mapper data*

The satellite data consisted of two subsets of the Landsat Thematic Mapper (TM) scene path 93 row 86F, of 14 February 1988. The two subscenes were rectified and co-registered to the digitized CD boundaries. The rectified subscenes were parallelograms within a $1412 \times 1008$ pixel ($42.4 \times 30.2$ km) rectangle and a $1119 \times 1174$ pixel ($33.6 \times 35.2$ km) rectangle for Ballarat and Geelong respectively. Each image pixel was assigned to the CD within which its centroid fell. A rudimentary haze-correction was applied to each image by threshholding. Figure 2 shows quasi-natural colour images of the two areas.

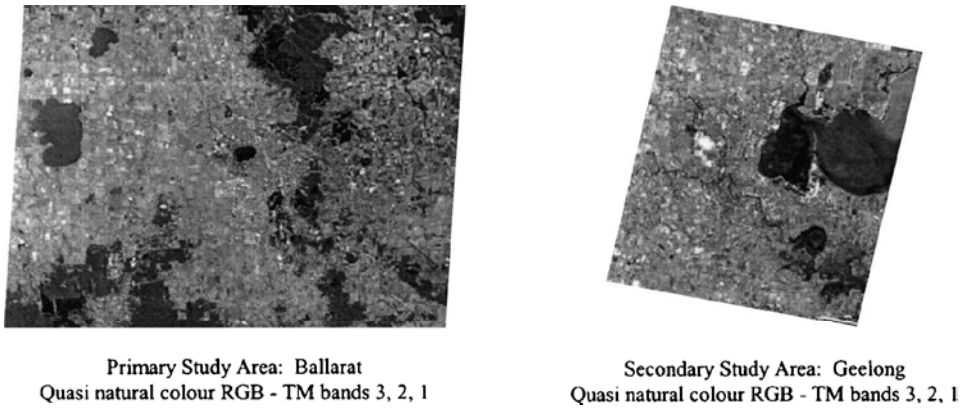


Figure 2. Images of the two study areas.



2.5. *Computing methods*

Image analysis was performed using ERMapper™ software on a Sun SPARCstation 2. Maps of the study areas were produced using CDATA96™ census mapping software and Mapinfo™ GIS software. Mapinfo™ was also used to locate ground control points.

Because remote sensing and GIS software packages are quite limited in their capacity to undertake anything but the most straightforward statistical analysis (Mesev 1998), most statistical analysis was performed using Minitab™, SPSS™ and Excel™.

## 3. Methodology

3.1. *Form and conceptual basis of models*

The base geographical unit for this study was the Census Collection District (CD). Population and population density are scale-dependent and time-dependent concepts. As the size of the areal units for which it is calculated becomes smaller, population density becomes more grainy and variable. Of course, instantaneous population density also changes throughout the day, but for most demographic and planning purposes, population is assigned by place of residence. This conceptualization also has the advantage that an individual's residence is unambiguously and discretely located within a particular CD.

Although the obvious probability model for population counts is the Poisson distribution, the Gaussian or normal distribution provides a simpler and adequate approximation when the counts are large, as is the case for CDs. Ordinary least squares (OLS) models with normal errors were used in the present study and in all of the studies cited, although Langford *et al.* (1991) also obtained very similar results using Poisson regression models.

The simplest OLS regression models for population estimation are additive linear models of the form:

$$p_i = \beta_0 + \sum_{j=1}^{n} \beta_j r_{ij} + \varepsilon_i \quad (1)$$

where $p_i$ = population or population density of CD $i$, $r_{ij}$ = remote sensing indicator $j$ for CD $i$, $n$ = number of explanatory variables or indicators; and $\beta_i$ are parameters to be estimated from the data.

The random errors $\varepsilon_i$ represent the variation in population unexplained by the remote sensing indicators, and are assumed to be independent and identically normally distributed with constant variance. In the simplest models, $r_{ij}$ is the mean reflectance of CD $i$ in band $j$, but the basic form can be extended to incorporate other derived variables such as measures of variability or band-to-band ratios. Mathematical transformations of the dependent population variable can also be incorporated. No provision is made for spatial autocorrelation in the errors; whilst in principal this is possible, in practice the irregularity in the size and the shape of the spatial units makes it difficult to implement.

If the explanatory variables are relatively scale-invariant measures such as averages of spectral characteristics or proportions of pixels in different classes, then the natural dependent variable is population density. If the explanatory variables are scale dependent quantities such as pixel counts, then the natural dependent variable is the total population of the areal unit. When the units are equal in size, as with grid squares, the distinction is immaterial. In this research the CDs were irregular and unequal in size, and population density was used as the dependent variable.



3.2. *Model evaluation: measures of bias, consistency and accuracy*

The particular statistical performance indicators reported in this study were chosen either to facilitate comparisons with the other research cited, or for their robustness to outliers resulting from occasional substantial violations of model assumptions.

The terms *bias*, *consistency* and *variance* have precise technical meanings when applied to statistical estimators. In a conceptually related sense, if the true values of population or population density are consistently underestimated or overestimated, we can say there is bias in the model or procedure. Whether or not there is bias, the estimation error for individual cases varies, to a degree that may be large or small in comparison to any bias present. This can be called *variability* or conversely *consistency*. The general term *accuracy* covers both aspects. Accuracy implies both consistency and lack of bias. Inaccuracy may be due to either bias, or variability, or both.

When OLS models are fitted to training data, the estimated values for the training cases are always unbiased in the sense described. However, if population density is modelled for units of unequal area, it does not follow that the population estimates (calculated by weighting the density estimates by area) will be unbiased, and hence sum to the correct total population for the whole training area. Furthermore, whenever an estimation equation is applied to other cases beyond the training set, there is no guarantee that the estimates of either population densities or populations will be unbiased. In the results that follow, the *relative error in the total urban population* is used as an indicator of overall bias.

The coefficient of variation $R^2$, which can be interpreted as the square of the correlation between observed and fitted values, is commonly cited as an indicator of the strength of a linear regression relationship. However, $R^2$ must be interpreted with caution. For models with a forced zero intercept, an $R^2$ value calculated in the usual way from sums of squares cannot be interpreted in the usual way; it will be misleadingly high unless appropriately adjusted (Myers 1990). Similarly, with a transformed dependent variable, $R^2$ relates to the linearity of the relationship between the transformed dependent variable and the explanatory variables, and may give a misleading impression about the accuracy of estimates of the untransformed

dependent variable obtained by back-transformation (Harvey 1999). In this paper, to facilitate broad comparability, modified $R^2$ values based on back-transformed estimates are quoted where appropriate.

Furthermore, $R^2$ indicates nothing about bias—estimates can be highly correlated with ground reference values and yet be consistently high or consistently low.

The relative or proportional error of estimation

$$RE = \frac{r-g}{g} \times 100\% \tag{2}$$

where $r$ is a remote sensing estimate and $g$ is a ground reference value, is a measure of performance which is the same for both population and population density. Since the errors of estimation for individual cases may be positive or negative, an average relative error is an indicator of bias. An average of the absolute values is an indicator of overall accuracy, including both variability/consistency and bias aspects. Two such measures have been used here: the *mean absolute proportional error* (abbreviated as *mean % error*) and the *median absolute proportional error* (or *median % error*):

$$\text{Mean \% error} = \frac{\sum_{k=1}^{m} |RE_k|}{m} \tag{3}$$

where $k$ indexes the $m$ CDs under consideration.

$$\text{Median \% error} = \text{50th percentile of the set of } m \text{ absolute values } |RE_k| \tag{4}$$

Means and medians were used in preference to root mean square averages because of the lack of robustness of the latter to the distorting influence of a few extreme outlying values.

### 3.3. *Validity and robustness*

Validation of regression models can be considered at two levels, which are here referred to as internal and external validation:

*Internal validation*. Given some training data from which some estimation procedure or function is derived, the term internal validation refers to the accuracy with which the dependent variable is estimated within the training set.

*External validation*. For an estimation procedure to have any genuinely predictive value (and potential operational use), its applicability must be demonstrated over some extended domain. In particular, when building complex models using variable selection procedures such as stepwise regression, the chosen model is inevitably tailored to the particular characteristics of the training data, and unlikely to work as well with any other data (the phenomenon of *capitalization on chance*).

*Robustness* is almost synonymous with external validity, but it carries the connotation of validity over a broader domain. A procedure trained on a sample or subset of a particular image may be externally validated by applying it to other samples or subsets from the same image or a similar image. If it also works for other rather different images it is more likely to be described as robust.

In this research, population estimation equations were externally validated by applying them to similar CDs in another section of the same TM scene.

## 4. Results and discussion for the primary study area

### 4.1. *Selection of dependent and explanatory variables*

Regression analysis was performed on data for the 132 CDs in the primary study area. The dependent variable was the average population density of each CD. The explanatory variables included band means (Iisaka and Hegedus 1982) together with a variety of standard spectral and spatial transformations of the TM data, including a number used by Forster (1980, 1981, 1983) in the context of urban land use. Two variants of the basic form of model were also considered, with logarithmic and square root transformations being applied to the dependent variable. At each stage, one or more models were selected as the 'best' and examined further. The six models chosen are summarized in table 7. Details of the selection process are now given.



### 4.2. *Models based on average reflectances*

To establish a benchmark comparable with the results of Iisaka and Hegedus (1982), CD population density $D$ was initially regressed on average reflectances in the six TM bands.

Population density ranged from 4.5 to 5142 persons $km^{-2}$, with a mean value of 1556.9 persons/$km^2$. The lowest densities were generally found in the large rural CDs. The average reflectances of the six TM bands were all positively correlated, the highest correlations being between the visible bands 1, 2, and 3. The strongest correlations with population density were for band 5 ($-0.52$), band 4 ($-0.40$) and band 1 ($+0.34$). This suggests that the presence of human population is weakly related to high reflectance in the short visible (blue) wavelengths, perhaps associated with paved surfaces and some roofing materials, and relatively low reflectance in the near infrared, associated with relative lack of vegetation.

A multiple regression model utilising all six bands resulted in $R^2 = 0.539$. Stepwise selection resulted in the 4-variable model identified as Model 1 in table 7. The reduction in fit was negligible ($R^2 = 0.537$). As is often the case, the best multivariate model did not utilize just the variables exhibiting the highest individual correlations with the dependent variable.

### 4.3. *Transformations of dependent and explanatory variables*

The fit for the benchmark model was poor in comparison to those of Iisaka and Hegedus ($R^2 = 0.59$ and 0.70 for two different years). This confirmed the expectation that population estimation would be more difficult for CDs of irregular shape and size in a mixed urban/rural area in regional Australia, where the degree of local heterogeneity within each CD would be higher than for suburban residential areas of metropolitan Tokyo.

One obvious problem was that the values of the dependent variable ranged over three orders of magnitude. An analysis of the residuals ($D - \hat{D}$) revealed pronounced positive skew, increasing variance with increasing $D$, and some evidence of concave-upward curvilinearity, all of which are commonly encountered when a dependent variable has a large range. A routine statistical strategy is to transform the dependent variable by taking the logarithm or the square root. Both of these were investigated.

A second approach to the representation of non-linear relationships within a linear framework is to transform the explanatory variables. The pool of potential predictors was enlarged by applying to the CD means a range of transformations:

firstly the squares of the six basic band means, then the 15 band-mean-to-band-mean cross-product terms, the 15 pairwise band-to-band ratios, and finally the 15 pairwise difference-to-sum ratios.

The full set of variables, and the models selected by stepwise regression, are summarized in tables 1 and 2. The variables *pij*, *rij* and *dij* are 'functions of means', not 'means of functions'; each variable is the product or ratio of mean values derived from CD aggregate figures, rather than the mean of a product or ratio calculated for each individual pixel. Table 2 shows that, beginning from a base $R^2$ value of around 0.5, the incorporation of squares, cross-products or ratios increased $R^2$ to

Table 1. Summary of regression variable nomenclature.

| Generic name | Number of variables | Description |
|---|---|---|
| bi | 6 | Mean of TM band i |
| si | 6 | Square of bi |
| pij | 15 | Cross-product bi × bj |
| rij | 15 | Ratio bi/bj |
| dij | 15 | Difference-to-sum ratio (bi − bj)/(bi + bj) |
| Total | 42 | i, j = 1,2,3,4,5,7 |

Table 2. Population density models based on CD means summary of stepwise regression results.*

| | Urban and rural areas ($n = 138$) | | | | | |
|---|---|---|---|---|---|---|
| | Dependent variable $D$ | | Dependent variable $\sqrt{D}$ | | Dependent variable $\ln D$ | |
| Potential predictors (number) | Selected predictors | $R^2$ | Selected predictors | $R^2$ ($R^2_b$) | Selected predictors | $R^2$ ($R^2_b$) |
| bi (6) | **b5 b7 b4 b3** | **0.537** | **b5 b7 b4 b3** | **0.652 (0.557)** | b5 b7 b3 b2 | 0.684 (0.343) |
| bi si (12) | s5 b7 s1 b5 b4 s4 s7 | 0.735 | **s1 b1 s7 b4 b5 s4** | **0.847 (0.755)** | s5 b7 s1 b1 s7 b5 | 0.861 (0.687) |
| bi si pij (27) | s5 p47 p57 p45 | 0.730 | s2 b4 p24 s7 b5 | 0.845 (0.757) | s5 p14 b4 s1 p47 | 0.855 (0.656) |
| bi si pij rij (42) | r57 r14 r37 | 0.696 | r57 r15 r17 r47 | 0.844 (0.757) | p45 s1 r17 p35 p37 p13 s3 r37 | 0.901 (0.704) |
| bi si rij dij (42) | r57 r14 d35 d14 d37 r25 | 0.753 | **r57 r14 d14 r15** | **0.846 (0.762)** | r13 r14 r15 s4 s1 s7 | 0.910 (0.719) |

*Models selected for further investigation are shown in bold typeface. Predictors in each model are listed in the order of selection. Parenthesized $R^2_b$ values are based on back-transformation. They are the $R^2$ values obtained when the dependent variable $D$ is regressed on the estimate of $D$ (obtained by inverse transformation of the regression estimates for the transformed $D$ values), and hence provide a more realistic indication of the predictive accuracy of the model than does the raw $R^2$ value.

around 0.7. Application of either the square root or the logarithmic transformation to $D$ resulted in a further increase in $R^2$ to within the range 0.8 to 0.9.

Many of the potential predictors were quite highly correlated, so that the number of variables retained by the stepwise procedure was reasonably small, ranging from 1 to 8. Also in most cases the specificity of the chosen set of variables is not high; when all possible sets of predictors are examined ('best subsets' regression, Myers 1990) there are alternative sets which perform almost as well. For example in the basic 4-band model, band 1 or band 2 can replace band 3 without much loss, since all three are highly correlated.

Two general points are noteworthy. Firstly, except in the case of the logarithmic models, when band ratios (variables whose names begin with 'r' or 'd') were included they completely displaced the other predictors (see last two rows of table 2). Secondly, band 2 only appears relatively infrequently in table 2. This suggests that visible green provides the weakest spectral signature of human habitation.



The residuals from these models were examined in the light of the demographic characteristics of the individual CDs (Ballarat and Western Victoria Regional Information Bureau 1989). The residual distributions tended to be positively skewed, with the extreme positive values being consistently associated with the same few CDs. The CD whose population was consistently underestimated by the greatest amount contained a large multi-storeyed geriatric institution. The populations of six CDs consisting predominantly of high-density public housing were also substantially underestimated throughout. Conversely, whilst overestimation was not so extreme in most models, populations did tend to be overestimated in older established areas where there were relatively high proportions of small households in disproportionately large houses.

There are also two technical issues of interpretation. Firstly, in models with a transformed dependent variable, the $R^2$ value exaggerates the precision of estimation. In this situation, the adjusted values based on back-transformation and designated $R_b^2$ give a more meaningful indication of comparative predictive performance. Values of this statistic are included in parentheses in table 2 and subsequent tables. For the square root transformation they are in general marginally higher than the $R^2$ values for the untransformed model, and for the logarithmic transformation they are substantially lower. This indicates that the increase in $R^2$ is largely illusory; the population density estimates produced by the transformed models are not substantially more accurate, and in some cases less so.

Secondly, as with any stepwise regression procedure, there is the aspect of capitalization on chance. This will be considered in the context of external validation in §5.

The fitted values and residuals from the three types of model were examined graphically. On the basis of goodness of fit, parsimony, and the statistical characteristics of the residuals, the square root transformation consistently outperformed the logarithmic transformation. This conclusion was later borne out for more complex models.

From table 2, the four models in bold type were retained for further examination: the base model and three that utilized the square root transformation. The first square root model was selected as a more statistically appropriate base model than the untransformed base model. All other square root models showed considerable improvement over the base model. The two chosen for further consideration were based on: band means and their squares; and ratios of band means. The regression equations are shown in table7.

4.4. *Models based on measures of spatial variability*

Researchers such as Forster (1993), Barnsley and Barr (1996) and Webster (1996) have suggested that image variability or texture can provide indicators of urban land use. To incorporate this aspect, the variance, standard deviation and coefficient of variation (standard deviation/mean) of each TM band were calculated for each CD.

The results of stepwise regression analyses using the 6-band means and the 18 ($3 \times 6$) variation measures are shown in table 4. In this and subsequent tables, the suffix notation shown in table 3 has been used.

Table 4 shows that a linear model based on the means and standard deviations of the basic TM bands within each CD produced somewhat better population density estimates ($R^2 = 0.751$) than the models based on means and functions of means. The addition of variance or coefficient of variation terms produced little further improvement. A similar pattern was evident with the square root models. Again the logarithmic models tended to incorporate more variables but did not perform as well. On balance, the preferred prediction model from table 4 was the final square root transformation model, shown in boldface, for which the regression equation is shown in table 7.

4.5. *Models based on spectral transformations*

The band-to-band ratios discussed in §4.3 were calculated using the CD averages for each band. The alternative is to calculate band-to-band ratios for each pixel and

Table 3. Variable suffix nomenclature.

| Suffix | Meaning | Example |
|---|---|---|
| none | Mean | b5 |
| v | Variance | b5v |
| s | Standard deviation | b5s |
| c | Coefficient of variation | b5c |

Table 4. Population density models based on CD means and spatial variation measures: summary of stepwise regression results.*

| Potential predictors (number) | Dependent variable $D$ | | Dependent variable $\sqrt{D}$ | | Dependent variable $D$ | |
|---|---|---|---|---|---|---|
| | Selected predictors | $R^2$ | Selected predictors | $R^2$ ($R^2_b$) | Selected predictors | $R^2$ ($R^2_b$) |
| Mean (6) | b5 b7 b3 b4 | 0.537 | b5 b7 b4 b3 | 0.652 (0.557) | b5 b7 b3 b2 | 0.684 (0.343) |
| Mean, std dev (12) | b5 b1s b7 b4 b4s | 0.751 | b5 b7 b1s b4 b4s b7s | 0.802 (0.766) | b5 b7 b4 b7s b3 b4s b2 | 0.763 (0.497) |
| Mean, std dev, coeff of var (18) | b5 b1c b4c b7 b4s b5c | 0.759 | b5s b5 b7c b4c b7 b3 b4s | 0.821 (0.769) | b5 b7 b4 b7c b3 b4s b5c b3s | 0.791 (0.533) |
| Mean, std dev, coeff of var, variance (24) | b5 b1c b4c b7 b4v b1 | 0.768 | **b5 b7c b4c b7v b4 b7 b1v b1c b5c** | **0.840 (0.780)** | b5 b7 b3 b4 b4v b7c b5c b3s | 0.794 (0.524) |

*See explanatory notes for table 2.

then average the results for each CD. Of the many spectral transformations that could be applied to the six TM bands at the individual pixel level, 14 were identified by preliminary visual screening as having some capacity to discriminate between residential and other land uses (Harvey 1999). These are listed in table 5.

These 14 variables were calculated for each pixel in the study area, and then means, standard deviations, coefficients of variation and variances were derived for each CD. The 56 resulting variables were combined with the 24 variables based on individual TM bands, and a number of stepwise regression analyses were then applied to various subsets of the 80 variables. The resulting models are summarized in table 6, using the suffix notation of table 3.

The models with the untransformed population density $D$ as the dependent variable fall into two groups. For the first model and the last four models (which utilize only the spectrally transformed predictors), $R^2$ values in the range 0.779 to 0.825 were obtained. In the second, third and fourth models, as happens from time to time with any incremental sub-optimal search, the stepwise procedure 'stopped short' with relatively few variables selected and relatively low $R^2$ values reached.

Once again the logarithmic transformation produced apparent improvement in $R^2$ at the expense of parsimony, but the improvement was not maintained after back-transformation. However the models utilizing the square root of population density as the dependent variable again performed very well. The preferred model from table 6 (shown in bold face) is a very stable one, in the sense that it emerged from six different analyses. The regression equation is shown in table 7.

### 4.6. *Comparative evaluation of the regression models*

The set of six models chosen to represent the range of options tested for estimating population density of CDs are summarized in table 7. In all but the first model, the equation predicts the square root of the population density. Population density estimates are found by squaring the predicted value for each CD.

Table 7 is in two sections. The first section shows the regression equation used to estimate the transformed dependent variable. By progressive enhancement of the set of predictors, the value of the coefficient of determination ($R^2$) was increased from around 0.54 to just over 0.90.

Table 5. Selected spectral transformations.

| Variable | Description |
|---|---|
| nb1 | Normalized band 1 |
| nb2 | Normalized band 2 |
| rl4 | Ratio: band 1 to band 4 |
| rl5 | Ratio: band 1 to band 5 |
| r25 | Ratio: band 2 to band 5 |
| r57 | Ratio: band 5 to band 7 |
| ds15 | Difference/sum ratio: bands 1, 5 |
| ds25 | Difference/sum ratio: bands 2, 5 |
| ds35 | Difference/sum ratio: bands 3, 5 |
| ds57 | Difference/sum ratio: bands 5, 7 |
| ch123 | Hue: bands 1, 2, 3 (cylindrical co-ordinates) |
| ch125 | Hue: bands 1, 2, 5 (cylindrical co-ordinates) |
| rh123 | Hue: bands 1, 2, 3 (rectangular co-ordinates) |
| rh125 | Hue: bands 1, 2, 5 (rectangular co-ordinates) |

Table 6. Population density models based on CD means and spatial variation of selected spectral transformations: summary of stepwise regression results.*

| Potential predictors (number) | Dependent variable $D$ | | Dependent variable $\sqrt{D}$ | | Dependent variable $D$ | |
|---|---|---|---|---|---|---|
| | Selected predictors | $R^2$ | Selected predictors | $R^2$ ($R^2_b$) | Selected predictors | $R^2$ ($R^2_b$) |
| TM bands + transformations: means (20) | rh123 r14 ds35 r57 rh125 ch125 | 0.794 | rh123 r14 r57 ds35 rh125 ch125 | 0.898 (0.832) | r14 r57 ds35 ch123 ds15 b5 b7 | 0.880 (0.774) |
| TM bands + transformations: means, std devs (40) | rh123 b1s | 0.729 | **rh123 r14 r57 ds35 rh125 rh125s** | **0.904 (0.843)** | rh123 rh123s b3 b4s b1 b7s b1s b5 b7 | 0.897 (0.706) |
| TM bands + transformations: means, std devs, coeffs of var (60) | rh123 b1s | 0.729 | **rh123 r14 r57 ds35 rh125 rh125s** | **0.904 (0.843)** | rh123 rh123s b3 b4s b1 b7s b1s b5 b7 | 0.897 (0.706) |
| TM bands + transformations: means, std devs, coeffs of var, variances (80) | rh123 b3v b2s | 0.758 | **rh123 r14 r57 ds35 rh125 rh125s** | **0.904 (0.843)** | rh123 rh123s b3 b4s b1 b5v b7c b3v ds15 rh125s b5s rh123c | 0.916 (0.694) |
| Transformations only: means (14) | rh123 r14 ds35 r57 rh125 ch125 | 0.794 | rh123 r14 r57 ds35 rh125 ch125 | 0.898 (0.832) | r14 r57 ds35 ch123 ds15 | 0.867 (0.741) |
| Transformations only: means, std devs (28) | r14s r25 r57 r14 rh125s rh123s | 0.825 | **rh123 r14 r57 ds35 rh125 rh125s** | **0.904 (0.843)** | rh123 r14 rh123s r57 ds35 ds25 rh125s | 0.894 (0.764) |
| Transformations only: means, std devs, coeffs of var (42) | rh123 r14s r25 rh125 r57c | 0.779 | **rh123 r14 r57 ds35 rh125 rh125s** | **0.904 (0.843)** | rh123 r14 rh123s r57 ds35 ds25 rh125s rh123c | 0.898 (0.746) |
| Transformations only: means, std devs, coeffs of var, variances (56) | rh123 r14s r25 rh125 r57c | 0.779 | **rh123 r14 r57 ds35 rh125 rh125s** | **0.904 (0.843)** | r14 rh123s ch123v r57 ds35 rh125s ch123s rh123c rh123v ch123 | 0.928 (0.771) |

*See explanatory notes for table 2.

The second section shows the relationship between the values of the dependent variable (the reference population densities) and the back-transformed estimates. This shows that in terms of predictive accuracy, the effective coefficient of variation ($R^2_b$) was increased from around 0.54 to around 0.84. This corresponds to an increase in the correlation between the remote sensing and ground reference figures from around 0.75 to around 0.92.

However, correlation is not the only criterion to be considered: a high correlation implies a linear relationship, but not necessarily accuracy of estimates. In the first model fitted to the untransformed population density, the estimates obtained were necessarily unbiased, so that the regressing the ground reference data on the estimates resulted in an intercept of 0 and a slope of 1. In all other models the data was transformed, which could have led to biases in the form of offsets (zero errors) or errors of scale (slope errors). However, table 7 shows that in all cases, the unforced

Table 7. Summary of selected models for population density based on TM data aggregated over census collection districts.*

| Model type | Class of predictors | Dependent variable | Number of predictors | Regression equation | $R^2$ | $D$ vs. $\hat{D}$ Regression coeffts.** (unforced & forced) | $R_b^2$ | $s$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Band mean | $D$ | 4 | 72.3 − 135.6 b5 + 332.0 b7 − 151.0 b3 + 61.6 b4 | 0.537 | 0 + 1.00; 1.00 (figure 3A) | 0.537 | 739 |
| 2 | Band mean | $\sqrt{D}$ | 4 | − 1.4 − 2.33 b5 + 5.37 b7 + 1.22 b4 − 2.04 b3 | 0.652 | 81 + 1.01; 1.05 | 0.557 | 739 |
| 3 | Mean, (mean)$^2$ | $\sqrt{D}$ | 6 | − 171.34 − 0.140 s1 + 9.220 b1 + 0.0344 s7 + 4.874 b4 − 1.952 b5 − 0.0359 s4 | 0.847 | − 58.1 + 1.07; 1.04 | 0.755 | 550 |
| 4 | Ratios & difference to sum ratios | $\sqrt{D}$ | 4 | 345.35 − 68.41 r57 − 275.56 r14 + 226.89 d14 + 120.30 r15 | 0.846 | 1.7 + 1.03; 1.03 | 0.762 | 541 |
| 5 | Mean, std dev, variance, coefft of variation | $\sqrt{D}$ | 9 | 75.20 − 2.19 b5 − 245.20 b7c − 70.36 b4c + 0.171 b7v + 0.851 b4 + 2.88 b7 − 0.124 b1v + 69.57 b1c + 70.37 b5c | 0.840 | 116 + 0.95; 1.01 | 0.780 | 521 |
| 6 | Mean & std dev of spectral transformations at pixel level | $\sqrt{D}$ | 6 | 530.10 + 0.278 rh123 − 92.34 r14 − 60.81 r57 + 165.91 ds35 − 1.308 rh125 − 0.370 rh125s | 0.904 | 9.5 + 1.01; 1.02 (figure 3A) | 0.843 | 441 |

*See explanatory notes for table 2.
**Intercept + slope; slope when forced through origin.

regressions had small zero errors relative to the scale of the data, and the slope coefficients of both forced and unforced regressions were close to unity, indicating a lack of substantial bias of either type.

Notwithstanding these results for population densities, the ultimate criterion is the accuracy with which actual population counts can be estimated. Table 8 shows the results of multiplying the population density estimate for each CD by its area to generate an estimated CD population. These have been evaluated by regressing the ground reference data on the estimates, and by calculating the mean and median of the absolute values of the relative errors for each CD and estimates for the total population of the study area.

The more complex models 3 to 6 had median proportional errors within the range 17–24% overall, and 14–21% for the urban area. Mean proportional errors were somewhat higher, indicating a positive skew which is characteristic of absolute value distributions, and which may be exacerbated by the presence of a few outlying values (see for example figure 3A).

The results for the 122 urban CDs and those for the 16 rural CDs have also been tabulated separately in table 8, and here the inherent limitations of the methodology at low population densities become apparent. In most of the models, the low population densities for the 16 rural CDs were over-estimated by amounts which, though small in absolute terms, were large in relative terms. When these over-estimates were weighted by the large areas involved, the effect on population estimates was substantial. As a result, whilst the more complex models produced estimates of the total urban population which were accurate to within a few per cent, the total for the whole study area was in every case overestimated to a much greater degree. The much lower $R^2$ values in table 8 than in table 7 are also largely attributable to the effects of these extreme rural 'outliers'.

The opposite occurred with the first (untransformed) model, where, because of the curvilinearity of the relationship, the densities at the low end of the range were underestimated by the linear model to such an extent that a number of the estimates were actually negative. When combined with the large areal weightings of these low density CDs, this resulted in an estimated total population which was negative.

Negative estimates are always a potential problem when the lowest densities are small in comparison to the range of densities being estimated. Langford *et al.* (1991) described models which can lead to such estimates as 'logically flawed', whilst Lo (1995) took the more pragmatic view that a negative estimate can be interpreted as evidence of zero population. Webster (1996) also tabulated negative estimates but they were not discussed. Using a transformation such as the logarithmic makes it impossible to produce negative estimates, but this advantage has to be considered against other performance criteria. In the present instance the square root form, which generally produces better results than the logarithmic form, is potentially quite logically flawed in that negative results automatically back-transformed by squaring lead to positive estimates. In these circumstances, the author is inclined to take Lo's approach and reset any negative estimates to zero.

The level of accuracy achieved with the more complex models in tables 7 and 8 is rather better than the best of the models fitted to 47 suburbs of Harare by Webster (1996) using a similar methodology but a different suite of predictors. It is comparable to the accuracy of a 7-variable model Webster fitted to 65 Cardiff grid squares using yet another suite of 70 texture variables.

The 'mechanism' of these models—why these particular linear combinations of

Table 8. Summary of estimated census collection district populations based on TM data aggregated over census collection districts.

| | | Ballarat Statistical District ($n=138$) | | | | | | Ballarat urban area ($n=122$) | | | | | | Rural balance ($n=16$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model Type* | Class of predictors | Slope (forced) | $R^2$ | $s$ | Mean % error | Median % error | Est. total (% error)** | Slope (forced) | $R^2$ | $s$ | Mean % error | Median % error | Est. total (% error)** | Slope (forced) | $R^2$ | $s$ | Mean % error | Median % error | Est. total (% error)** |
| 1 | Band mean | −0.002 | 0.00 | 259 | 553.6 | 31.0 | −33 519 (−142.3%) | 0.57 | 0.26 | 225 | 83.2 | 28.5 | 88 178 (+26%) | −0.005 | 0.05 | 242 | 4141 | 2877 | −121 697 (−1458%) |
| 2 | Band mean | 0.12 | 0.00 | 259 | 185.9 | 32.9 | 151 019 (+91%) | 0.77 | 0.29 | 221 | 72.4 | 30.5 | 74 832 (+7%) | 0.06 | 0.06 | 241 | 1051 | 743 | 76 187 (+751%) |
| 3 | Mean, $(\text{mean})^2$ | 0.58 | 0.15 | 238 | 57.3 | 20.9 | 91 992 (+16%) | 0.78 | 0.32 | 216 | 33.5 | 20.0 | 74 811 (+7%) | 0.32 | 0.01 | 247 | 155.6 | 92.2 | 14 283 (+60%) |
| 4 | Ratios & difference to sum ratios | 0.66 | 0.10 | 246 | 47.6 | 24.1 | 89 094 (+13%) | 0.85 | 0.34 | 213 | 37.7 | 19.4 | 71 992 (+2%) | 0.25 | 0.01 | 248 | 207.4 | 97.1 | 20 000 (+120%) |
| 5 | Mean, std dev, variance, coefft of variation | 0.28 | 0.01 | 257 | 93.0 | 23.0 | 111 574 (+41%) | 0.97 | 0.49 | 186 | 36.5 | 21.0 | 68 805 (−2%) | 0.12 | 0.01 | 247 | 524 | 286 | 42 769 (+377%) |
| 6 | Mean & std dev of spectral transform'ns at pixel level | 0.58 | 0.18 | 235 | 39.4 | 17.4 | 89 849 (+14%) | 0.94 | 0.56 | 175 | 28.2 | 13.6 | 70 615 (+0.6%) | 0.24 | 0.18 | 225 | 124.2 | 57.6 | 19 234 (+115%) |

*The dependent variable for model type 1 was $D$, and for all other model types $\sqrt{D}$ (as shown in table 7).
**Ground reference populations are: BSD 79179; Urban 70222; Rural 8957.

particular spatially aggregated and averaged spectral characteristics should correlate highly with populations—is conjectural. It may be possible to relate the structure of the equations to reflectance properties of materials and combinations thereof, but one must always exercise extreme caution in placing interpretations on individual regression coefficients in a multivariate context, because of multicollinearity resulting from the correlations between the various predictor variables. There may be many other alternative combinations of variables that would estimate population density almost as well on this set of data, and perhaps better on a slightly different set of data. This problem, and the problem of capitalization on chance, are exacerbated by the relative paucity of aggregated observations. A total of 116 explanatory variables has been considered and assessed on the basis of just 138 observations.

There is also the issue of the ecological fallacy; the presence of a particular relationship between population and spatially aggregated and averaged spectral characteristics does not necessarily imply that a similar relationship holds at the level of individual pixels or individual dwellings.



## 5. External validation on the secondary study area

To investigate the robustness of the six models, the various remote sensing indicators discussed in §4 were calculated for each of the 225 CDs of the secondary image, and six sets of population density estimates were calculated using the equations in table 7. Table 9 shows some key indicators of comparative performance of these models on the primary and secondary images. In figure 3, plots of the first and the sixth model are compared for the primary and secondary study areas.

The $R^2$ values for the primary Ballarat area increased monotonically from 0.537 to 0.843; for the secondary Geelong area, the range was 0.453 to 0.741, with the value for model 6 being lower than that for model 5. Correspondingly, the residual standard deviation $s$ was in each case larger for Geelong than for Ballarat. These figures indicate a general degradation of performance when applying any of the estimating equations to the secondary image.

As well as reduced correlation, there was also evidence of bias, with consistent underestimation of population density in the secondary study area; all six slope coefficients were considerably higher than unity, indicating underestimation, with the ground reference values tending to be larger than the estimates.

Nevertheless, the plots for model 6 (figure 3) confirm that the relationship between $D$ and $\hat{D}$ remained linear, indicating that the underlying form of the link between population density and the particular linear combination of remote sensing characteristics chosen in Ballarat remained valid for the Geelong data. The reduction in the level of correlation and the associated broader spread of points on the Geelong plots was not unexpected. However the evidence of bias suggested some more systematic calibration problem requiring investigation.

As to the relative robustness of the 6 model types, the first two models, which were included only as benchmarks, and which were quite inadequate even on the primary data, were also the least robust with regard to the slope coefficient. At the other extreme, model 6, as well as producing the most accurate results for the primary area, also proved to be the most robust on most measures. The intermediate model types 3, 4 and 5 were consistent with regard to slope bias, though model 5 had much higher $R^2$ values.

Table 9 also shows that the population estimates for individual CDs were less accurate in the secondary area than in the primary area. The median percentage

Table 9. Comparison of six population density estimation models for primary and secondary study areas.

| | | Primary study area: Ballarat | | | | | | | Secondary study area: Geelong | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model type | Class of predictors | $D$ vs $\hat{D}$ coefft (forced) | $R_b^2$ | $s$ | Region median % error | Region total* % error | Urban median % error | Urban total* % error | $D$ vs $\hat{D}$ coefft (forced) | $R_b^2$ | $s$ | Region median % error | Region total* % error | Urban median % error | Urban total* % error |
| 1 | Band mean | 1.00 | 0.537 | 739 | 31.0 | – | 28.5 | +26 | 1.40 | 0.453 | 878 | 37.7 | +50 | 36.5 | −5 |
| 2 | Band mean | 1.05 | 0.557 | 739 | 32.9 | +91 | 30.5 | +7 | 1.52 | 0.448 | 882 | 40.7 | +22 | 39.5 | −20 |
| 3 | Mean, $(\text{mean})^2$ | 1.04 | 0.755 | 550 | 24.1 | +13 | 20.0 | +7 | 1.20 | 0.599 | 752 | 26.7 | +51 | 25.7 | +4 |
| 4 | Ratios & difference to sum ratios | 1.03 | 0.762 | 541 | 20.9 | +16 | 19.4 | +2 | 1.19 | 0.601 | 750 | 25.8 | +41 | 24.4 | −0 |
| 5 | Mean, std dev, variance, coefft of variation | 1.01 | 0.780 | 521 | 23.0 | +41 | 21.0 | −2 | 1.19 | 0.741 | 605 | 23.0 | +21 | 22.1 | −16 |
| 6 | Mean & std dev of spectral transformations at pixel level | 1.02 | 0.843 | 441 | 17.4 | +14 | 13.6 | +1 | 1.11 | 0.718 | 630 | 18.4 | +47 | 17.1 | −3 |

*Ground reference populations: Ballarat region 79179; Ballarat urban 70222; Geelong region 147910; Geelong urban 142250.

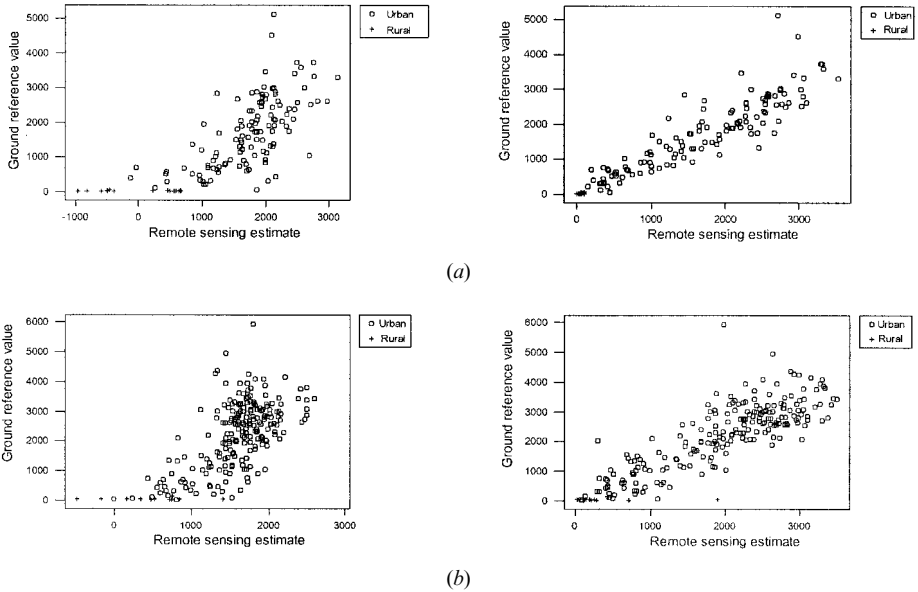


Figure 3. Population density estimates for census collection districts: ground reference versus remote sensing estimates from base and enhanced models. (*a*) Primary image: Ballarat. Population density model 1: band means. Population density model 6: pixel spectral transms. (*b*) Secondary image: Geelong. Population density model 1: band means. Population density model 6: pixel spectral transfms.

errors were higher in Geelong, both for the overall region and for the urban area, in all but two cases, both involving model 5. As in the case of Ballarat, the Geelong region totals were grossly overestimated in all cases. In Ballarat this was primarily due to the overestimation of the low densities in the large rural CDs. In the case of Geelong, this effect was further exacerbated by the presence of some large industrial sites in medium sized non-urban CDs (an oil refinery, an aluminium smelter, a car assembly plant, a cement works and a salt works), all of which were assigned large spurious populations by the remote sensing algorithms. Iisaka and Hegedus (1982) and Lo (1995) reported similar problems with a few non-residential or otherwise anomalous study units.

There were some apparent contradictions in the results. For example, whilst the slope coefficients for population density were much greater than unity for all models for the secondary area, the estimates of urban totals were in a number of cases comparable in accuracy with those for the primary area.

These results prompted a closer examination of the demographic characteristics of the two study areas, which are summarized in table 10.

There were a number of differences. GSD had little more than half the area of BSD, but almost double the population. A much larger proportion of GSD than of BSD was urban, but even within the urban areas, the average population density was 42% higher in Geelong than Ballarat.

The closest points of similarity were the occupancy ratio, which was almost identical at around 2.9 persons per dwelling for both regions and for both urban areas; and the percentage of non-separate dwellings, which again was almost constant at 19%. This suggested that the higher population densities in Geelong were not

associated with more multi-dwelling structures, but rather with smaller lot sizes, and houses which were either smaller, or closer together, or a combination of both.

Plots of the discrepancies ($\hat{D} - D$) against $D$ for models 4 and 6 (figure 4) show an attenuation or lack of sensitivity whereby the extremes of variation in density are not reflected in the remote sensing estimates. Higher densities tend to be underestimated and lower densities overestimated. The same problem was noted by Langford *et al.* (1991), and it is also apparent in the results of Iisaka and Hegedus (1982) and

Table 10. Demographic characteristics of primary and secondary study areas.

| Demographic characteristic | Primary study area: Ballarat | | Secondary study area: Geelong | | Ratio Geelong: Ballarat | |
|---|---|---|---|---|---|---|
| | Region | Urban | Region | Urban | Region | Urban |
| Area ($km^2$) | 613.89 | 74.85 | 351.9 | 107.15 | 0.57 | 1.43 |
| Population | 79 179 | 70 222 | 147 910 | 142 250 | 1.87 | 2.03 |
| Number of dwellings | 26 971 | 24 368 | 51 078 | 49 411 | 1.89 | 2.03 |
| Number of non-separate houses | 5154 | 4721 | 9989 | 9586 | 1.94 | 2.03 |
| Population density ($persons\,km^{-2}$) | 129.0 | 938.2 | 420.3 | 1327.6 | 3.26 | 1.42 |
| Dwelling density ($dwellings\,km^{-2}$) | 43.93 | 325.56 | 145.15 | 461.14 | 3.30 | 1.42 |
| Occupancy ratio (persons per dwelling) | 2.94 | 2.88 | 2.90 | 2.88 | 0.99 | 1.00 |
| % non-separate houses | 19.11 | 19.37 | 19.56 | 19.40 | 1.02 | 1.00 |



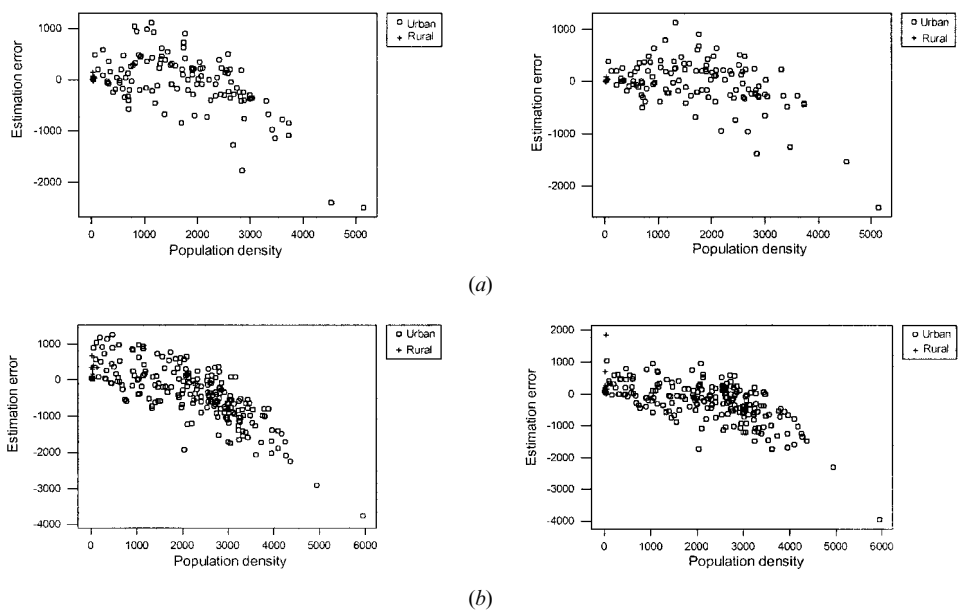


Figure 4. Estimation error versus population density. (*a*). Primary image Ballarat. Population density model 4: ratios of band means. Population density model 6: pixel spectral transfms. (*b*) Secondary image: Geelong. Population density model 4: ratios of band means. Population density model 6: pixel spectral transfms.

Webster (1996). The effect was exaggerated in the case of Geelong by the overall higher density than in Ballarat. As always, extrapolation beyond the range of the training data is dangerous.

The contradictory results alluded to in table 9 may be an artefact of the use of CDs as the unit of aggregation. CDs are designed to have approximately equal populations, and so CD area is inversely related to population and dwelling density. As a result, when population totals are calculated by multiplying the density estimate for each CD by its area and summing, there is perhaps a tendency for a slight overestimation of density over relatively large areas and more substantial underestimation of density over relatively small areas to counteract each other.

## 6. Summary

Starting from a very basic prediction model, substantial improvements were achieved in the estimation of CD population densities and populations on the basis of CD aggregates of remote sensing indicators. From the many models investigated, six representative models were chosen for validation on a second image. The first model was the simplest baseline model, with an untransformed dependent variable and four TM band means as predictors. The other five models were all formulated to predict the square root of the population density, and involved progressively more complex CD aggregate functions of the TM bands: basic means; squares of means, ratios of means, variation measures; and means and variation measures of selected pixel-by-pixel spectral transformations. The effective $R^2$ values of these models ranged from 0.54 to 0.84, corresponding to correlations of 0.73 to 0.92 between census and remote sensing estimates of population density.

For the primary study area on which the models were trained, the three most complex models produced median proportional errors for the population of individual CDs within the range 17–21%. Estimates of total population for the urban section of the primary study area were correct to within 3%, but total population for the low-density rural sections and hence for the whole study area were substantially overestimated by all models.

When applied to a secondary study area for purposes of validation, the performance of all the models was somewhat degraded on most criteria, with effective $R^2$ values ranging from 0.45 to 0.72 and median proportional errors within the range 18–26%. There was also evidence of estimation bias associated with the somewhat higher population densities in the secondary study area. This seems to be related to smaller lot sizes and a higher spatial concentration of separate houses, rather than to substantial differences in the distribution of other types of residential structure. Nevertheless, quite accurate estimates were obtained for the total population of the urban section of the secondary study area, though again the total populations of the low-density rural sections were substantially overestimated by all models, as was the case for the primary study area.

## 7. Conclusion

It has been demonstrated that a methodology based on aggregate measures of various remote sensing indicators can be ‘tuned’ to model moderate population densities moderately well, both in a training area and in another fairly similar area. Both the level of accuracy and the robustness to validation generally improved with increasing model complexity.

However, none of the models developed, nor those reported by other researchers,

performed very well at the extremes of population density. There was a consistent attenuation effect, with remote sensing estimates varying over a narrower range than the actual densities. In particular, low-density rural populations were substantially overestimated in all cases. It is concluded that the potential of this methodology is limited by heterogeneity of both land cover and population density within the individual CDs, and that further improvements are in principle unlikely using this approach. In particular, the sacrifice of detailed spatial information leaves no way to respond to the problem of over-estimation of population in large areas of low density.

Many of the estimates obtained for total urban population were quite accurate, and the best results for CD populations were obtained using spectral indicators calculated at the level of individual pixels. These two facts lend support to the notion that models formulated and fitted at a particular spatial level can produce relatively accurate and reliable population estimates for larger spatial aggregates, but not for spatial units at the same level of aggregation.

From both perspectives, it follows that to obtain more accurate estimates at the scale of CDs, more of the modelling process should be carried out at a lower level of spatial aggregation. The practical difficulty with this is that whilst remote sensing data is available at the higher resolution of individual pixels, ground reference population data is not available for areas smaller than CDs. An approach to modelling the population associated with individual pixels, which utilizes an EM (expectation-maximization) type of statistical algorithm to overcome the problem of combining two levels of data, has been developed by the author, and will be reported in a subsequent paper.



**Acknowledgments**

Ground reference population data was obtained from the Australian Bureau of Statistics. Digitized census collection district boundaries were obtained from the Australian Survey and Land Information Group.

**References**

Australian Bureau of Statistics, 1998, *Australian Standard Geographical Classification (ASGC) Cat. No. 1216.0* (Canberra: Australian Bureau of Statistics).

Ballarat and Western Victoria Regional Information Bureau, 1989, *A Social Atlas of the Central Highlands Region, 1986 Census Edition* (Ballarat: Ballarat and Western Victoria Regional Information Bureau).

Barnsley, M. J., and Barr, S. L., 1996, Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. *Photogrammetric Engineering and Remote Sensing*, **62**, 949–958.

Booz-Allen, and Hamilton, 2000, Master Address File/Topographically Integrated Geographic Encoding and Referencing System (MAF/TIGER) modernization study, for the Geography Division, United States Census Bureau, USA.

Chen, S., Zeng, S., and Xie, C., 2000, Remote sensing and GIS for urban growth analysis in China. *Photogrammetric Engineering and Remote Sensing*, **66**, 593–598.

Davreau, F., Barbary, O., Michel, A., and Lortic, B., 1989, *Area Sampling from Satellite Image for Socio-Demographic Surveys in Urban Environments* (Paris: Editions de l'ORSTOM).

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., and Worley, B. A., 2000, LandScan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, **66**, 849–857.

Fisher, P. F., and Langford, M., 1995, Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, **27**, 211–224.

Fisher, P. F., and Langford, M., 1996, Modelling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. *Professional Geographer*, **48**, 299–309.

Forster, B. C., 1980, Urban residential ground cover using Landsat digital data. *Photogrammetric Engineering and Remote Sensing*, **46**, 547–558.

Forster, B. C., 1981, *Some Measures of Urban Residential Quality from Landsat Multispectral Data (Unisurv S18)* (Sydney: University of New South Wales).

Forster, B. C., 1983, Some urban measurements from Landsat data. *Photogrammetric Engineering and Remote Sensing*, **49**, 1693–1707.

Forster, B. C., 1993, The coefficient of variation as a measure of urban spatial attributes, using SPOT HRV and Landsat TM data. *International Journal of Remote Sensing*, **14**, 2403–2409.

Haack, B., Craven, D., and Jampoler, S. M., 1996, GIS tracks Kathmandu valleys urban explosion. *GIS World*, **9**, 54–57.

Harvey, J. T., 1999, *Estimation of Population Using Satellite Imagery* (PhD Thesis, University of Ballarat).

Henderson, F. M., and Xia, Z. G., 1997, SAR applications in human settlement detection, population estimation and urban land use pattern analysis: a status report. *IEEE Transactions on Geoscience and Remote Sensing*, **35**, 79–85.

Iisaka, J., and Hegedus, E., 1982, Population estimation from Landsat imagery. *Remote Sensing of the Environment*, **12**, 259–272.

Imhoff, M. L., Lawrence, W. T., Stutzer, D. C., and Elvidge, C. D., 1997, A technique for using composite DMSP/OLS 'city lights' satellite data to map urban area. *Remote Sensing of Environment*, **61**, 361–370.

Jensen, J. R., and Toll, D. L., 1982, Detecting residential land-use development at the urban fringe. *Photogrammetric Engineering and Remote Sensing*, **48**, 629–643.

Kressler, F., and Steinnocher, K., 1996, Change detection in urban areas using satellite images and spectral mixture analysis. *International Archives of Photogrammetry and Remote Sensing*, **31**, 379–383.

Kwarteng, A. Y., and Chavez, P. S., 1998, Change detection study of Kuwait City and environs using multi-temporal Landsat Thematic Mappper data. *International Journal of Remote Sensing*, **19**, 1651–1662.

Langford, M., Maguire, D. J., and Unwin, D. J., 1991, The areal interpolation problem: estimating population using remote sensing within a GIS framework. In *Handling Geographical Information: Methodology and Potential Applications*, edited by I. Masser and M. Blakemore (London: Longman) pp. 55–77.

Lo, C. P., 1986, *Applied Remote Sensing* (Harlow: Longman).

Lo, C. P., 1995, Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach. *International Journal of Remote Sensing*, **16**, 17–34.

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

Martin, R. G., and Howarth, P. J., 1989, Change-detection accuracy assessment using spot multispectral imagery of the rural-urban fringe. *Remote Sensing of Environment*, **30**, 55–66.

Martin, R. G., 1989, Accuracy assessment of Landsat-based visual change detection methods applied to the rural-urban fringe. *Photogrammetric Engineering and Remote Sensing*, **55**, 209–215.

Masek, J. G., Lindsay, F. E., and Goward, S. N., 2000, Dynamics of urban growth in Washington DC metropolitan area, 1973-1996, from Landsat observations. *International Journal of Remote Sensing*, **21**, 3473–3486.

Mesev, V., 1998, Remote sensing of urban systems: hierarchical integration with GIS. *Computers, Environment and Urban Systems*, **21**, 175–187.

Myers, R. H., 1990, *Classical and Modern Regression with Applications. 2nd ed.* (Boston: Duxbury).

National Aeronautics and Space Administration, 1978, Application of Satellite Pictures to Census Operations. Bolivian Experience in Census-Taking of Population and Residences. Translation into English of Aplicaciones de Las Imagines de Satelite a Operaciones Censales. Experiencia Boliviana en El Censo de Poblacion Y Vivienda,

Rept. Inst. Nacl. De Estadistica, Min. De Planeamiento Y Coord., Rep. of Bolivia, La Paz 1977 p 1–14. Scitran: Santa Barbara. Report No. NASA-TM-75090.

OGROSKY, C. E., 1975, Population estimates from satellite imagery. *Photogrammetric Engineering and Remote Sensing*, **41**, 707–712.

OLERUNFEMI, J. F., 1986, Towards a philosophy of population census in Nigeria: remote sensing inputs. *Remote Sensing Yearbook 1986*, 117–125.

QUARMBY, N. A., and CUSHNIE, J. L., 1989, Monitoring urban land cover changes at the urban fringe from SPOT HRV imagery in south-east England. *International Journal of Remote Sensing*, **10**, 955–965.

STERN, M., 1984, Landsat data for population estimates – approaches to inter-censal counts in the rural Sudan. In *Remote Sensing from Satellites*, edited by W. D. Carter and E. T. Engman (New York: Pergamon), pp. 117–125.

SUTTON, P., ROBERTS, D., ELVIDGE, C., and MEIJ, H. (1997) A comparison of nighttime satellite imagery and population density for the continental United States. *Photogrammetric Engineering and Remote Sensing*, **63**, 1303–1313.

WEBSTER, C. J., 1996, Population and dwelling unit estimates from space. *Third World Planning Review*, **18**, 155–176.

WELCH, R., and ZUPKO, S., 1980, Urbanized area energy utilisation patterns from DMSP data. *Photogrammetric Engineering and Remote Sensing*, **46**, 201–207.

YEH, A. G., and LI, X., 1996, Urban growth management in the Pearl River delta: an integrated remote sensing and GIS approach. *The ITC Journal*, **1**, 77–86.

YEH, A. G., and LI, X., 2001, Measurement and monitoring of urban sprawl in a rapidly growing region using entropy. *Photogrammetric Engineering and Remote Sensing*, **67**, 83–90.

YUAN, Y., SMITH, R. M., and LIMP, W. F., 1997, Remodeling census population with spatial information from Landsat TM imagery. *Computing Environment and Urban Systems*, **21**, 245–258.