

Correlations between Census Dwelling Data and Remotely Sensed Data

Keping Chen

Natural Hazards Research Centre
School of Earth Sciences, Macquarie University, Sydney, Australia
Phone: +61 2 9850-8433 Fax: +61 2 9850-8428
Email: kchen@laurel.ocs.mq.edu.au

Presented at SIRC 98 – The 10th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
16-19 November 1998

ABSTRACT

With the increasingly available census data and remotely sensed data, to discuss their relationship is one of important issues in GIS data integration. This paper proposes a method to demonstrate the correlations between zonal census dwelling data and residential densities discriminated by RS classification. First, texture statistic (homogeneity) along with six TM bands (bands 1-5 and 7) is put together to classify residential density levels and it is shown that the homogeneity enhances classification accuracy. After classification, close correlations between residential densities and census dwelling data have been examined by using multiple linear regression. However, the accurate classification scenario does not methodically reveal higher correlations. It is concluded that data integration of zonal census data within the framework of RS-GIS is feasible, consequently the differentiation of residential densities could offer enormous opportunities to treat dwelling-related census data, such as identification of the scale and zonal effect of the modifiable area unit problem (MAUP), and topographical representation of zonal census data.

Keywords and phrases: census data; remote sensing; data integration; texture; classification; correlations.

INTRODUCTION

For the past a few years, census data and remotely sensed data are increasingly accessible and an extensive use of zone-based census data exists in RS-GIS context. For example, using census data as ancillary information for RS classification (Hutchinson, 1982; Sadler and Barnsley, 1990; Harris and Ventura, 1995; Mesev, *et al.* 1996; Vogelmann, *et al.* 1998), and using census data as separate variables in multi-variate analyses (Weber and Hirsch, 1992; Lo and Faber, 1997). However, little work has been done to identify the correlations between zone-based census data and discrete remotely sensed data. To analyze them in a statistical way, the key is the dis-aggregation of census data at an appropriate spatial scale which could be compatible with the raster-based data sets.

In terms of the dis-aggregation of census data, a few innovative methods have been devised during the last decade. They mainly focus on how to dis-aggregate population density data at a fine spatial resolution, such as the centroid-based surface model (Martin, 1989) and the dasymetric mapping approach (Langford, *et al.* 1991). Comprehensive comparisons between these two techniques are reviewed by Martin and Bracken (1993). They point out that the ability to re-model census data really offers an effective alternative to integrate socioeconomic data and physical environment data in a raster-based GIS environment. Particularly, the dasymetric approach demonstrates the transformation of socioeconomic data from the arbitrary zonal base to a physical settlement geography. There are significant geographical implications in treating census data in such way. It would be possible to establish correlations between detailed land divisions and zonal census data. Census data from hierarchical census boundaries can interactively link with remotely sensed data through different land classifications at the local/regional scale (Figure 1). For example, at the regional scale, Walker *et al.* (1998) discuss a practical method of dis-aggregating areal-based agricultural statistics with NOAA-AVHRR images in

the state of NSW, Australia. At the very local scale, two driving forces have come from the increasing use of airborne and spaceborne higher resolution images and smaller census districts with precise geo-referencing. For census dwelling data, based on a wide range of urban land uses and census dwelling counts at different boundary levels, the relationship between *different land uses/covers* and *census dwelling data* can therefore be constructed. Likewise, it is also feasible to build a relationship between *different residential densities* and *census dwelling data*, provided that the conventional RS is currently insufficient to identify urban land covers as discussed by Haack *et al.* (1987), Quarmby and Cushnie (1989), and the estimation of residential density is an alternative to distinguish urban surface from RS imagery. In terms of the integration approaches between RS land classifications and census data at different boundary levels, there exist “*many-to-many*” set relationships.

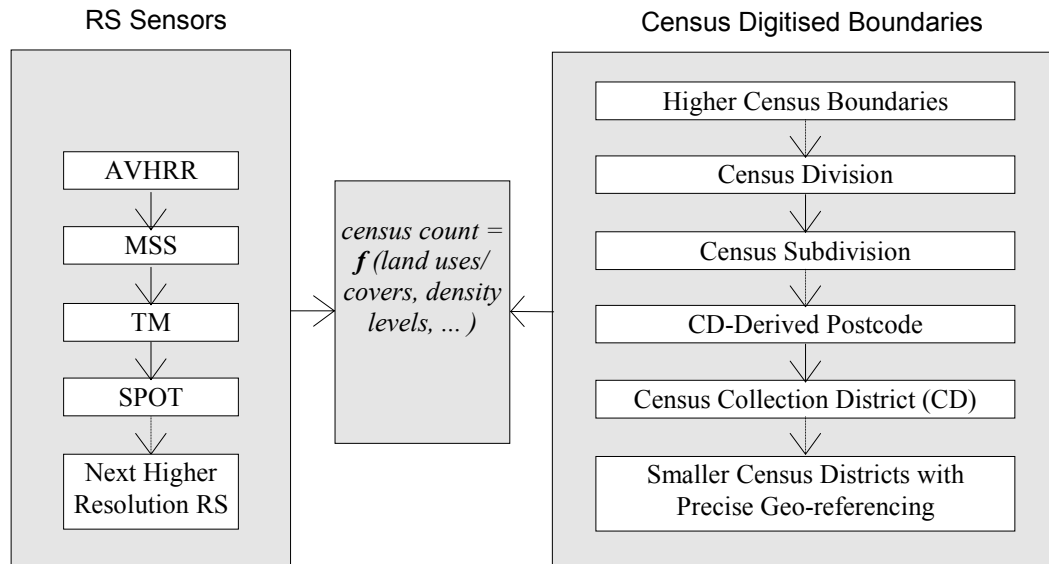


Figure 1. Linking hierarchical census data (e.g. Australia) and RS data

The purpose of this paper is to identify whether the correlations between areas of distinctive residential densities and census data at a certain boundary level exist. In the remainder of the paper, texture-based image classification scenarios to discriminate density levels will be detailed first, followed by the identification of correlations between residential densities and census dwelling data. Finally, topics relevant to the data model and spatial analysis of census data in a RS-GIS domain, and some geography-related implications will be discussed.

STUDY AREA AND DATA SOURCES

Hornsby Heights, located at approximately 20km north of the Sydney CBD, was selected as a study area (area 10.78 square kilometres and perimeter 18.53km). As a part of Sydney’s urban fringe, Hornsby Heights is a typical residential area surrounded by an extensive coverage of bush. The area covers 13 census collection districts (CDs) (Figure 2). A CD is the smallest areal unit for Australian census collection and each CD contains about 250 dwellings on average in urban area (ABS, 1993).

There were five primary data sources available for this research: (1) Landsat TM image captured on 9 November, 1991. A 150×175 section was clipped and pre-processed to an Australian Map Grid (AMG) coordinate system; (2) 1991 Census of Population and Housing data from the Australian Bureau of Statistics; (3) digital terrain data; (4) local cadastral maps (scale 1:8,000) between 1991/1992 (Hornsby Shire Council, 1992); and (5) local orthophoto maps (scale 1:4,000, 1986) and coloured aerial photography (scale 1:25,000, 1994). The timing of the census data, TM image and cadastral maps have almost matched, thus, it is expected that the data sets provide good representations of the study area during that period. The orthophoto map and aerial photo only serve as reference data for this study.

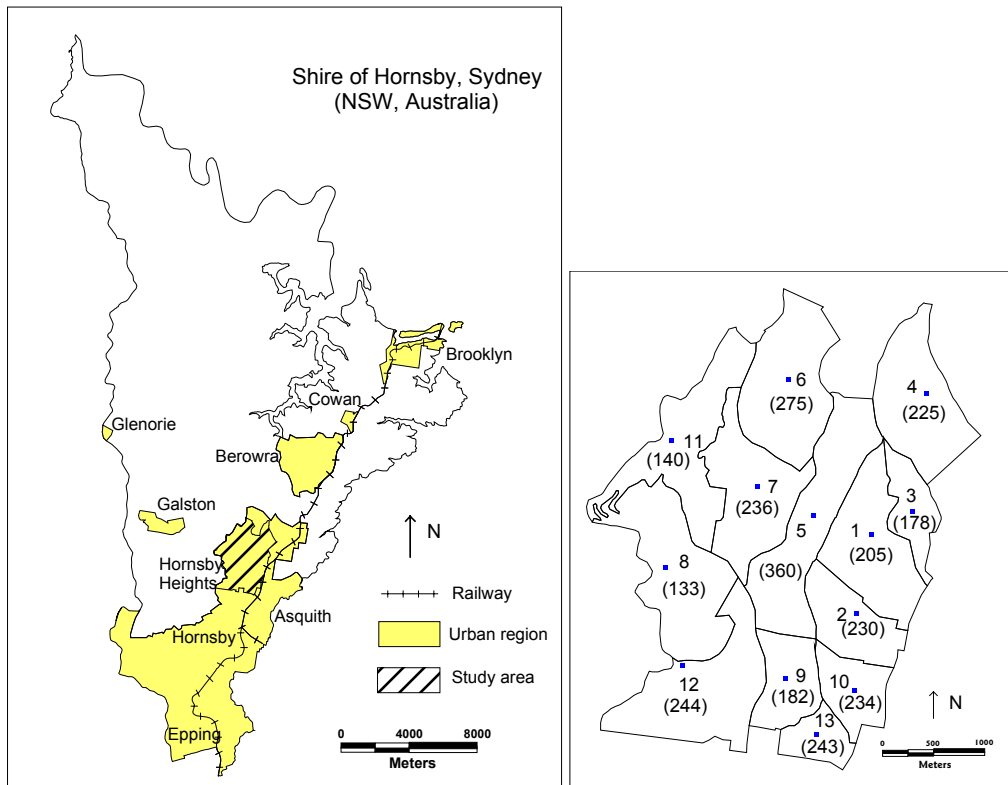


Figure 2. The study area and 13 CDs (dwelling number on each CD is included within parenthesis)

RESIDENTIAL DENSITY CLASSIFICATION

In terms of dwelling density estimation with remotely sensed data, Webster (1996) reports an empirical experiment in which residential density can be distinguished by using pattern-recognition techniques in the context of image processing and urban remote sensing. The approach profiles the distinctive forms of urban neighbourhoods and constructs textural signatures in order to demonstrate the correlation between textural form and land use density. Since different density areas should match with corresponding textural signatures, a number of texture statistics, such as edge contrast, street density, entropy, and homogeneity are selected. Besides, land use classification at the urban fringe using a mixture of spectral and textural features has been successful. For example, Jensen (1979) found texture data (such as contrast and high frequency filter) in conjunction with the traditional spectral bands could achieve accurate land cover classification at Levels II and III. During the past decade, a number of methods by using a combination of texture and spectral data to obtain satisfactory results have been tested, such as linear discriminate classification and neural networks (Franklin and Peddle, 1990; Kaminsky, 1997). The above review indicates that textures of an image can reflect heterogeneity, composition and pattern of urban land surface. Therefore, the method of incorporating physical textures of an image to reveal the underlying residential density associated with urban morphology is quite possible.

SELECTION OF TEXTURE STATISTICS

Choice of texture statistics and choice of scale (window size and pixel size) are important to characterise residential morphology mathematically. In this regard, familiarity with the study area and the data are necessary. Fourteen textural statistics were first developed by Haralick *et al.* (1973) and some of them, such as the angular second moment, the entropy and the inverse difference moment, were used to develop the texture-based image classification (Jensen, 1979; Peddle and Franklin, 1991; Kushwaha, *et al.* 1994). In this study, one texture discriminator was used - Homogeneity (H). Homogeneity is a commonly used statistic of the grey level co-occurrence matrix (GLCM). It is capable of discovering the degree of pixel adjacencies within the GLCM. Homogeneity can be measured by a grey level co-occurrence matrix $M(i, j)$ (Parker, 1997).

$$H = \sum_i \sum_j \frac{M(i, j)}{1 + |i - j|}$$

Since the image textural homogeneity is related to a region and not an individual pixel, moving windows with different width were used to produce appropriate discriminations. A pre-defined principle is that the moving window should result in the best differentiation between low and medium residential densities, and between dwellings and the neighbouring bushes. Two 5×5 and 7×7 moving windows were tested. Larger windows had very strong smoothing effects and therefore were not considered.

Homogeneity could be generated from each band using different window sizes, distances and orientations. A monochrome band was selected by undertaking a principle component transformation of the six bands of TM image (except TM band 6) which specifically covers the study area. The first three components account for 98.38% of the standardised variance, among which the first component carries 87.67% of the original variance and represents the biggest loading, 0.9822, on TM band 5, compared with the remaining bands. The first component also reveals the strong differentiation between bushland and residential. So it is anticipated that the middle infra-red band (TM band 5) has the greatest amount of information for the identification of different dwelling densities from surrounding bushland. By using TM band 5, homogeneity was calculated on the basis of Parker's (1997) TEXT program with a 5×5 window, one pixel distance and zero orientation (west-east direction). Figure 3 (right) displays the homogeneity of a 60×60 image section at the south part of the study area.

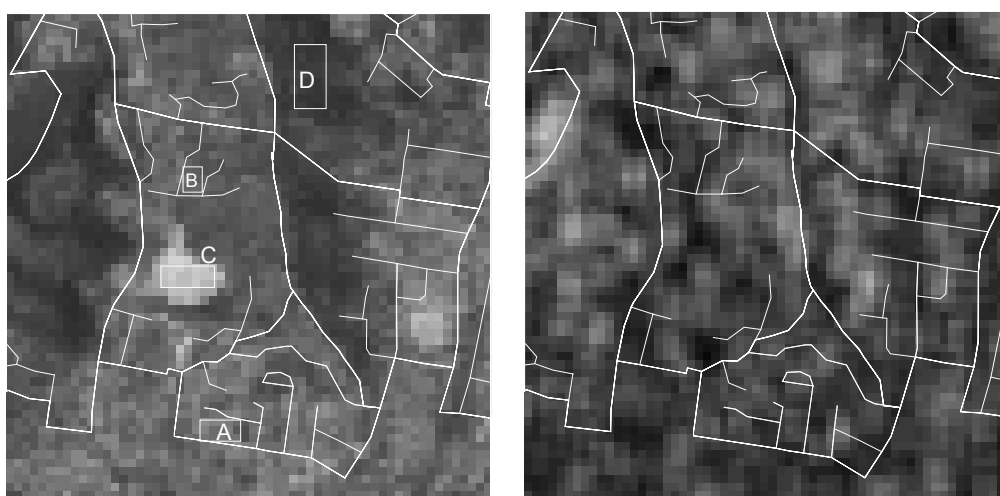


Figure 3. TM band 5 (left) and calculated homogeneity image (right) superimposed with local streets. Dark tones represent low values. A, B, C, and D are training areas: A - density Level I, B - density Level II, C - density Level III, and D - bushland (density levels are detailed at next section).

CLASSIFICATION WITH TEXTURE STATISTIC

The study area is basically a residential area and separate houses are the predominant category of dwellings (Table 1), therefore it is unnecessary to discriminate different structure types. Consequently only the relationship between residential densities and census dwelling data is examined in this study. Based on field observations and detailed interpretations of large scale cadastral maps, with the help of orthophoto maps and aerial photography, three levels of residential density were initially prescribed. Here, Level I primarily refers to a continuous group of dwellings with few overhanging trees, while Level II corresponds to dwelling units where each unit is largely scattered either by neighbouring bushland or adjacent to large gardens and parks. Level III includes school grounds, cricket ovals, bowling greens and tennis courts within the residential area. Thus, classifying the residential density based on textural neighbourhoods in this area is critical. According to the training sites, average dwelling number for each level are: Level I: 0.89 per pixel; Level II: 0.47 per pixel; and Level III: no dwelling on the pixel. Besides, forest, water, and road were also trained to make training signatures.

Table 1. Structure types (all private dwellings) of the study area

CDs (n=13)	Separate house	Semi-detached	Flat-apartment	Others*	Total
Total	2798	15	54	18	2885
Percentage	96.99%	0.52%	1.87%	0.62%	100%

* Others include caravans in or not in caravan parks, improvised home, house or flat attached to a shop, office, etc. A full description can be found in ABS (1993).

The classification approach used in this paper is the maximum likelihood (ML) classifier implemented by the IDRISI software (Eastman, 1997). Three classification scenarios were used (Table 2). Two methods of accuracy assessment are applied: (1) the first approach is somewhat general and is designed to assess the division between residential and non-residential areas. The whole residential area was made up of three density levels. The pre-defined residential area was delineated by using large scale cadastral maps, ancillary orthophoto maps and aerial photography. A stratified random sampling strategy was employed to generate 1,000 check points, 257 of which were located within the pre-defined residential area. All three classification treatments produced very good results (Table 2). It is shown that the inclusion of homogeneity (7×7 window) enhances the performance of the ML classifier by approximately 4.11 per cent compared with that of using TM 1-5 and 7 bands only. (2) Visual interpretations for three classified images show that density levels in the classified image with higher accuracy tend to be clumped while density Level I and Level II in low accuracy images tend to be fragmentary (Figure 4). Neighbouring pixels tend to be the same density category and an inclusion of homogeneity in this study facilitates this classification process. For objective appraisal purposes, the first-lag auto-correlation coefficient - Moran's *I* (King's Case) for the classified density images masked by the whole corresponding residential area was calculated individually (Table 2). The auto-correlation value expresses a propensity for data values to be similar to surrounding data values (Eastman, 1997). Therefore, it is reasonably conjectured that clumped density areas provide appropriate representations of underlying density levels.

Table 2. Accuracy assessment of three classification approaches

Classification Scenarios	Accuracy (%) - residential only	Moran's <i>I</i>
A1 - TM bands 1-5, 7	91.12	0.2632
A2 - TM bands 1-5, 7, Homogeneity (7×7)	95.23	0.3484
A3 - TM bands 1-5, 7, Homogeneity (5×5)	94.90	0.3128



Figure 4. Classified density levels from classification scenarios A2 (left) and A1 (right)
 Level I - dark grey, Level II - grey, and Level III - light grey.
 White area within 13 CDs represents non-residential.

CORRELATIONS BETWEEN CENSUS DATA AND DENSITY LEVELS USING MULTIPLE LINEAR REGRESSION

Having produced residential density levels by texture-based classifications, the correlation between the whole residential area and corresponding census counts of dwelling at each CD level was preliminarily tested and found to have no existing linear relationship. This proves that the residential distribution in the study area is not homogeneous. To test the correlations between zonal census dwelling counts and the underlying density levels, multiple linear regression was conducted using MATLAB statistical software. Pixel numbers (equivalent of area) of three density levels at each CD were selected as three independent variables, and the census count of each CD as dependent variable. Each regression proved to be valid after *F* test at a 0.005 level of confidence, and high correlations existed between them (Table 3). For example, in the classification scenario A1, 85.05% of

the variance of CD's dwelling counts is explained by the areas of three density levels. For each regression test, the percentage of residual against respective CD's dwelling count was plotted in Figure 5 (right).

Table 3. Multiple linear regression between three residential densities and census counts

Classification Scenarios	Apparent R^2	Mean Square (Regression)	Mean Square (Residual)	$F (n=13)$ $F_{0.005}(3,9)=8.72$	P
A1	0.8505	11810.49	692.16	17.0632	0.0005
A2	0.8213	11406.02	826.99	13.7923	0.0010
A3	0.8053	11183.30	901.23	12.4090	0.0015

From Table 3 and Figure 5, it can be generally concluded that the census dwelling counts have a close relationship with the underlying dwelling densities. Besides, density variabilities at each CD affect the total dwelling count with very low residual (about 10 per cent of the observation count), and density levels with higher classification accuracy do not reveal more significant regression results. Compared with the regression results between census counts and only the first two density levels, it is found that the R^2 decreases when the density Level III is not taken into account. Although the density Level III is not closely related to an absolute increase of dwelling number of each CD, it shows that the density Level III plays a role in the regression. High correlations can be attributed to the combined physical or possibly social importance of three density levels in the structure of built environment. Furthermore, the trend of residual percentage between neighbouring CDs largely displays a similar pattern (Figure 5, right) among three regression tests, and on the basis of this the compositions of different densities at each CD could be further identified.

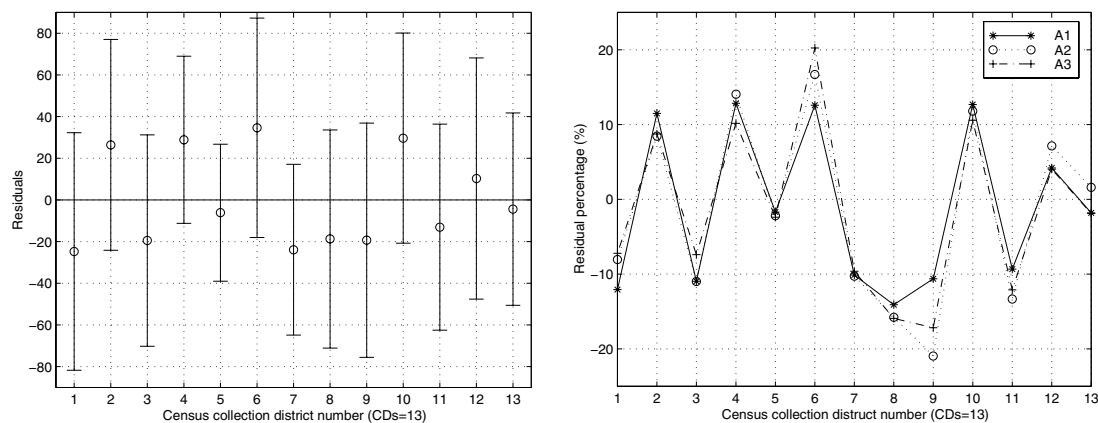


Figure 5. Regression residuals between three density levels and census dwelling counts residuals for A1 regression test (left) and residual percentage for three regression tests (right).

CONCLUSIONS AND DISCUSSION

This paper has examined the close relationship between census dwelling data and the underlying densities. Homogeneity was used to enhance classification accuracy, however, close correlations between residential densities and census dwelling counts do not correspond to high classification accuracy systematically. Census data and RS image will probably be very important data sources for human geography and physical geography in the future, and easily accessible data bases provide opportunities and impetuses to bridge the gap between them with compatible data structure and attributional and spatial analysis. In this paper, it is shown that dis-aggregation of census dwelling data through residential density classification using remotely sensed data is possible. Dis-aggregated census data could meet the needs of a wide range of integrated socioeconomic and environmental applications, such as integrated land uses and planning, natural hazards risk assessment, and environmental impact assessment. However, it should be noted that the census dis-aggregation approach will not substitute the traditional representation format. Rather, it will facilitate the fusion of making use of census and RS data and subsequently widen data sources and explore new spatial analyses in the next integrated GIS era. It can be expected that the dis-aggregation of census data by using RS will help solve a number of geography-related problems, such as:

- (1) Census data have long been represented as zone-based choropleth maps which lack underlying spatial

characteristics and an inherited problem is the MAUP (Openshaw, 1984). In the dis-aggregation of zonal census data, one of tasks is to assign values to suitable target zones from available source zones (here source zones often refer to census boundaries). Several methods have been discussed in the last decade. A piecewise approximation approach (Goodchild *et al.* 1993) by area weighting and area-related attributes between source and target zones assumes that source data are distributed uniformly and seems to be applicable to simple phenomena (i.e. population census). The important factor in spatial dis-aggregation as indicated by Flowerdew and Green (1991 and 1992) is local knowledge or control contributes. Image classification would delineate the rich tapestry of the underlying landscape (e.g. density levels) so as to provide multi-dimensional avenues for the spatial distribution of census data. By means of building the pattern-function correlations geometrically and statistically, the spatial and attributional allocations from source zones to target zones would become feasible, albeit a set of externally defined constraints (i. e. area, ground controls) in practice is very useful (Walker *et al.* 1998).

(2) The dasymetric method (Langford, *et al.* 1991), as an alternative representation to the choropleth format, assumes that the residential area is homogeneous. However, as previously mentioned, there is no linear relationship between the whole density area and census counts, nor between individual density level and census counts. Rather, the combination of different density levels exhibits close correlations with census data. Within the residential area, it is still possible to represent the subtleness of structure (e.g. categorical density levels). Effective representations of census data include at least two components: spatial boundaries or resolutions, and both qualitative contents and quantitative statistics. Therefore, any representational approximation will depend on both effective land divisions and associated socioeconomic dimensions.

(3) The regression equation based on an experimental area after fitting and checking for any inadequacies can be used to predict census counts at other urban fringes in a similar landscape siting and socioeconomic environment. Spatial, attributional and temporal changes could be identified. It is known that RS already offers a powerful tool to depict urban changes over time (Quarmby and Cushnie, 1989), however attributes of ground objects are generally ignored or disassociated from the spatial and temporal realm. Through the multiple regression test, it would be quite possible to predict the total dwelling number while the area of density levels changes. Unlike the example in this paper (the separate houses are the predominant dwelling structure), different dwelling structures may have different correlations with corresponding land categories.

(4) From the perspective of sociologists, residential segregation generally measures the dissimilarity - 'the degree to which two or more groups live separately from one another, in different parts of the urban environment' (Massey and Denton, 1988). However, for a very long time, a prominent problem of using a dissimilarity index to measure segregation based on the tract or block census data is the ignorance of spatial interaction between areal units. Geographers look at it spatially and capture the spatial components to the dissimilarity index. Thus segregation measures have been advanced over the years and include the distance-based approach (Morgan, 1982), the method of using the length of the common boundary of two areal units and the shape of the areal units (Wong, 1993), and the perimeter-based clustering index (Lee and Culhane, 1998). Those social segregation indices could be further explored by the understanding of the relationship between the residential patterns and human-oriented processes. As suggested, different residential densities, along with census distribution at a finer resolution could open a window to scrutinise the social segregation under spatial interaction with practical significance.

ACKNOWLEDGEMENTS

The author is grateful for Professor Russell Blong and Frank Siciliano of the Natural Hazards Research Centre, and Carol Jacobson of the School of Earth Sciences at Macquarie University for their encouragement, suggestion and support. Helpful comments of the paper reviewer are also appreciated. The census data and the TM image were provided by the Benfield Greig Australia P/L and the School of Earth Sciences, respectively.

REFERENCES

- Australian Bureau of Statistics (1993) *CDA91 Data Guide: 1991 Census of Population and Housing*, Canberra.
- Eastman R. (1997) *IDRISI for Windows User's Guide, Version 2.0* (Worcester, MA: Graduate School of Geography, Clark University).
- Flowerdew, R. and M. Green (1991) Data integration: statistical methods for transferring data between zonal systems, In *Handling Geographical Information: Methodology and Applications*, edited by I. Masser and M. Blakemore (London: Longman), pp. 38-54.
- Flowerdew, R. and M. Green (1992) Statistical methods for inference between incompatible zones systems. In *The Accuracy of Spatial Databases*, edited by M. F. Goodchild and S. Gopal (London: Taylor and Francis), pp. 239 - 247.
- Franklin, S.E. and D.R. Peddle (1990) Classification of SPOT HRV imagery and textural features. *International of Remote Sensing*, 11, pp. 551-556.
- Goodchild, M.F., L. Anselin and U. Deichmann (1993) A framework for the areal interpolation of socioeconomic data. *Environmental Planning A*, 25, pp. 383-397.
- Haack, B., N. Bryant and S. Adams (1987) Assessment of Landsat MSS data for urban and near-urban land cover digital classification. *Remote Sensing of Environment*, 21, pp. 201-213.
- Haralick, R.M., K. Shanmugam and I. Dinstein (1973) Texture features for image classification. *I.E.E.E. Transactions on Systems, Man, and Cybernetics*, 3, pp. 610-621.
- Harris, P.M. and S. Ventura (1995) The integration of geographic data with remotely sensed imagery to improve classification in an urban area. *Photogrammetric Engineering and Remote Sensing*, 61, pp. 993-998.
- Hornsby Shire Council (1992) *Hornsby Council land Information System - Cadastral Information*, Hornsby Council Central Library, Sydney, Australia.
- Hutchinson C.F. (1982) Techniques for combining Landsat and ancillary data for digital classification improvement. *Photogrammetric Engineering and Remote Sensing*, 48, pp. 123-130.
- Jensen J.R. (1979) Spectral and textural features to classify elusive land cover at the urban fringe. *Professional Geographer*, 31, pp. 400-409.
- Kaminsky, E.J., H. Barad and W. Brown (1997) Textural neural network and version space classifiers for remote sensing. *International of Remote Sensing*, 18, pp. 741-762.
- Kushwaha, S.P.S., S. Kuntz and G. Oesten (1994) Application of image texture in forest classification. *International Journal of Remote Sensing*, 15, pp. 2273-2284.
- Langford, M., D.J. Maguire and D.J. Unwin (1991) The areal interpolation problem: estimating population using remote sensing in a GIS framework. In *Handling Geographical Information: Methodology and Applications*, edited by I. Masser and M. Blakemore (London: Longman), pp. 55-77.
- Lee, C.M. and D.P. Culhane (1998) A perimeter-based clustering index for measuring spatial segregation: a cognitive GIS approach. *Environment and Planning B*, 25, pp. 327-343.
- Lo, C.P. and B.J. Faber (1997) Integration of Landsat Thematic Mapper and census data for quality of life assessment. *Remote Sensing of Environment*, 62, pp. 143-157.
- Martin, D. and I. Bracken (1993) The integration of socioeconomic and physical resource data for applied land management information systems. *Applied geography*, 13, pp. 45-53.
- Martin D. (1989) Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, 14, pp. 90-97.

- Massey, G.S. and N.A. Denton (1988) The dimension of residential segregation. *Social Forces*, 67, pp. 281-315.
- Mesev, V., P. Longley and M. Batty (1996) RS-GIS: spatial distributions from remote imagery. In *Spatial Analysis: Modelling in a GIS Environment*, edited by P. Longley and M. Batty (London: GeoInformation International).
- Morgan B.S. (1982) The properties of a distance-based segregation index. *Journal of Socioeconomic Planning Sciences*, 16, pp. 167-171.
- Openshaw S. (1984) *The Modifiable Area Unit Problem*, CATMOG 38, (Norwich: Geo Books).
- Parker J.R. (1997) *Algorithms for Image Processing and Computer Vision* (New York: John Wiley & Sons).
- Peddle, D.R. and S.E. Franklin (1991) Image texture processing and data integration for surface pattern discrimination. *Photogrammetric Engineering and Remote Sensing*, 57, pp. 413-420.
- Quarmby, N.A. and J.L. Cushnie (1989) Monitoring urban land cover changes at the urban fringe from SPOT HRV imagery in South-east England. *International Journal of Remote Sensing*, 10, pp. 953-963.
- Sadler, G.J. and M.J. Barnsley (1990) Use of population density to improve classification accuracy in remotely sensed images of urban areas. *EGIS (90)*, Amsterdam, The Netherlands.
- Volgelmann, J.E., T. Sohl and S.M. Howard (1998) Regional characterization of land cover using multiple sources of data. *Photogrammetric Engineering and Remote Sensing*, 64, pp. 45-57.
- Walker, P.A. and T. Mallawaarachchi (1998) Dis-aggregating agricultural statistics using NOAA-AVHRR NDVI. *Remote Sensing of Environment*, 63, pp. 112-125.
- Weber, C. and J. Hirsch (1992) Some urban measurement from SPOT data: urban life quality indices. *International Journal of Remote Sensing*, 13, pp. 3251-3261.
- Webster, C.J. (1996) Urban morphology fingerprints. *Environment and Planning B*, 23, pp. 279-297.
- Wong D.W.S. (1993) Spatial indices of segregation. *Urban Studies*, 30, pp. 559-572.