# NEIGHBORHOOD PATTERN RECOGNITION FROM MAILING INFORMATION: LINKS WITH SATELLITE IMAGERY

Victor Mesev
School of Environmental Sciences
University of Ulster
Coleraine, BT52 1SA, Northern Ireland
tv.mesev@ulst.ac.uk.

## Introduction

Data from remote sensing and GIS have assumed pivotal roles in contemporary spatial analysis methodologies. Whatever the application, geo-coded data handled by both technologies have improved over the years not only in quality but also in the efficiency of how such data are processed. Improvements in quality and efficiency of data models inevitably result in “better” application results, which in turn provide the opportunity to challenge established theory (Longley, 2002). Within urban geographical theory, inroads in data quality, in terms of precision and disaggregation have recently provoked developments in system-wide models of urban form and function (Longley and Mesev, 2002). These, in turn, have fuelled rational urban planning, the delineation of store catchment areas, customer targeting (Harris and Longley, 2000), and land use change distributions. One of the most important urban monitoring applications has been the shift in the geographies of retail activity – the rise of Internet shopping and the growth of suburban “out of town” locations – which have diminished the attraction of traditional urban central cores. Along with similar centrifugal movements in both manufacturing and business sites urban systems in the developed world have experienced rapid land use adjustments without necessarily the scales of urban growth evident in many developing world cities. Such examples are increasingly monitored by data models generated from remote sensing and GIS.

The building of “data-rich” models has become the consensus within remote sensing and GIS research groups active in both the urban and natural environment domains. In urban planning, research has become sympathetic to the needs for precision and up to date maps of land use delineation, as well as accountable to the repercussions of government policy decisions on individual household behavior and interaction (Donnay, 1999). It is within this dynamic urban backdrop that this paper will seek to contribute towards the growing body of knowledge on the recognition of urban land use patterns as represented by remotely sensed images.

Traditionally, urban land use interpretation from remote sensing is problematic and highly dependent on the scale, generalization, and scope of the application (Forster, 1985; Mesev, 2003). Inherent spatial variability of urban land use and acute spectral heterogeneity between pixel values typically lead to low interpretation accuracy. Methodologies to improve accuracy have ranged from the manipulation of neighboring pixel values (textural and contextual measures, for example by Möller-

Jensen, 1990), to the spatial arrangement of classified pixels (for instance, using graph theory by Barnsley and Barr, 1997; or Pesaresi and Bianchin, 2001), and to the incorporation of information from beyond the spectral domain, usually during the classification process (for example, fuzzy sets, neural nets and Bayesian modifications by Mesev, 1998). On the whole, success in improving urban land use accuracy has been small to negligible, usually qualified by local site-specific and time-specific conditions.

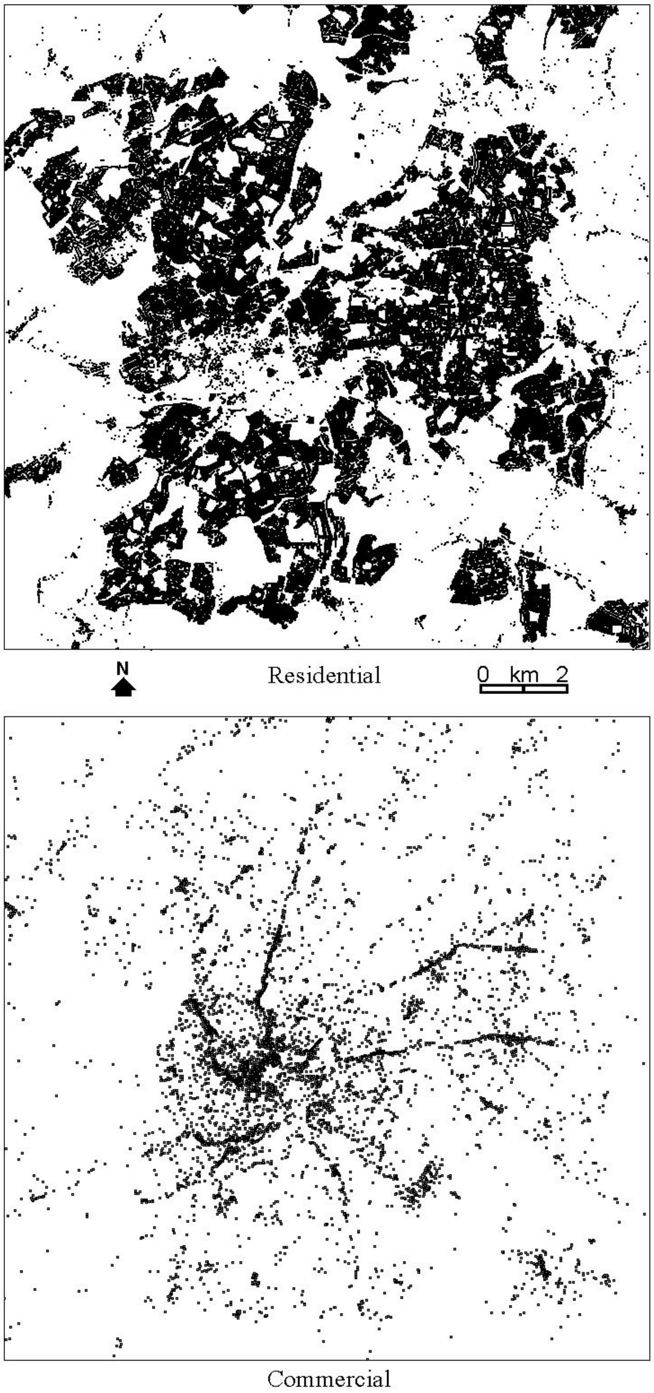

**Figure 1**: ADDRESS-POINT distributions for Bristol, UK

The common factor in most of these methodologies that restricts improvements in classification and pattern recognition accuracy is undoubtedly the inability to measure urban land use at a scale fine enough to identify individual building characteristics and hence infer human behavioral processes. If the objective is to delineate the maximum extent of human settlement then traditional approaches using coarse spatial resolution imagery and aggregated government statistics may suffice. However, such city-wide measures have as yet to convince planners and decision makers of their importance and as a consequence play only peripheral roles within local government policies (Donnay, 1999). If proponents of remote sensing and GIS want to rebuild the reputation of their data they need to seriously tackle the limitations of aggregated urban models and begin to embrace the challenges of disaggregated urban models. Although such models may be more demanding theoretically and technically they are nonetheless essential pragmatically.

The main objective of this paper is to outline a tentative agenda for building disaggregated models that infer urban land use distributions and therefore can be used to inform classified imagery of urban areas. The disaggregated models are based on digital postal records of every delivery address in a city; both residential and commercial properties (Figure 1). Knowing the spatial distribution of postal addresses introduces a number of key indicators of density (compactness versus sparseness) and arrangement (linearity versus randomness). These are measured using standard and linear readjusted nearest neighbor statistics. By establishing a relationship between image pixels and building distributions, the long-term research goal is to facilitate reliable and accurate spatial pattern recognition and multispectral classification methodologies to a level that renders resulting output irresistible to planners and policy makers (Donnay, 1999). Such work may even deflect criticism and restore flagging confidence in the applicability of urban remote sensing in the developed world (Mesev, 2003).

## 2 DISAGGREGATE DATA MODELS

Aggregate models are the standard vehicles for extraneous information commonly used in the augmentation of remotely sensed images representing urban land use (Chen, 2002). Typically, census records and other government directed statistics are aggregated into areal units for the sole purpose of preserving confidentiality. In the most recent UK Population Census, the finest level of aggregation is known as an output area (OA), which normally represents approximately 150 to 250 households. Within large, dense cities these aggregations are sometimes at a fine enough scale to adequately inform multispectral classifications (Mesev 1998; Mesev 2001), generate zonal-based dasymetric measurements (Langford, 2003; Lo, 2003), and even pixel-based population estimates (Geoghegan *et al*, 1998; Harvey, 2002). However, as with all aggregations of data, census tracts inextricably suffer from the ecological fallacy and the modifiable areal unit problem. Other than generalized city-wide applications, such aggregated zonal data have limited use for precision land use identification and therefore limited scope for informing accurate image classification of urban buildings.

| *Field* | *Format* | *Description* |
|---|---|---|
| Address key | I8 | Key identifying addresses |
| Building name | A50 | If number not available |
| Building number | I4 | Range 0–9999 |
| Change type | A1 | Insert; change; or delete |
| Change date | D6 | Date of last change to record |
| Department name | A60 | For organizations |
| Dependent locality | A35 | Subdivision of post town |
| Eastings (0.1m) | I8 | National grid (0.1m resolution) |
| Northings (0.1m) | 7 | National grid (0.1m resolution) |
| OSAPR | A18 | OS unique identifier |
| Physical status | I1 | E.g. planned |
| PO box number | A6 | PO box number |
| Postcode | A7 | Approx. 14 addresses |
| Positional quality | I1 | Accuracy of seed addresses |
| Post town | A30 | Name of post town |
| Royal Mail version | I8 | Date of last PAF update |

*Frc*

**Table 1**: ADDRESS-POINT attribute table

From the mid 1990s the Ordnance Survey of Great Britain began to compile a digital database of every one of the 25 million postal delivery points in Great Britain that has an address. The product is known as ADDRESS-POINT (Table 1) and the planimetric coordinates of this point-based dataset are claimed to be precise to within 0.1 meters (50 m in some rural areas) of the actual location of the building. It was created primarily using the Royal Mail's Postcode Address file (PAF) along with ground survey measurements, and is updated on a frequent interval (usually 3 months). The database represents one of the first attempts at disaggregating the geographical distributions of individual households and as such provides an unique opportunity to view the urban landscape as a surface of discrete entities rather than the traditional and administratively convenient partition by artificial zones representing aggregated and uniform values. It also offers, for the first time, the possibility of analyzing the spatial configuration and density of individual addresses within neighborhoods, which are typically hidden by zonal representations. All in all, the creation of ADDRESS-POINT is a major step forward in the pursuit of 'framework' data that encapsulate the desire for higher quality urban data not just for image pattern recognition improvement but also for all urban-based spatial data analyses.

In highlighting the merits of point-based disaggregated data it is also important to understand their limitations. For a start, point-based data are essentially dimensionless cartographic symbols, independent of the size and shape of the buildings they represent. The geographic co-ordinates of each point are positioned to correspond with the center of the building, and as such the only measurable parameters that can be derived are those related to the distribution of points indicating density (compactness or sparseness) and arrangement (linearity or randomness).

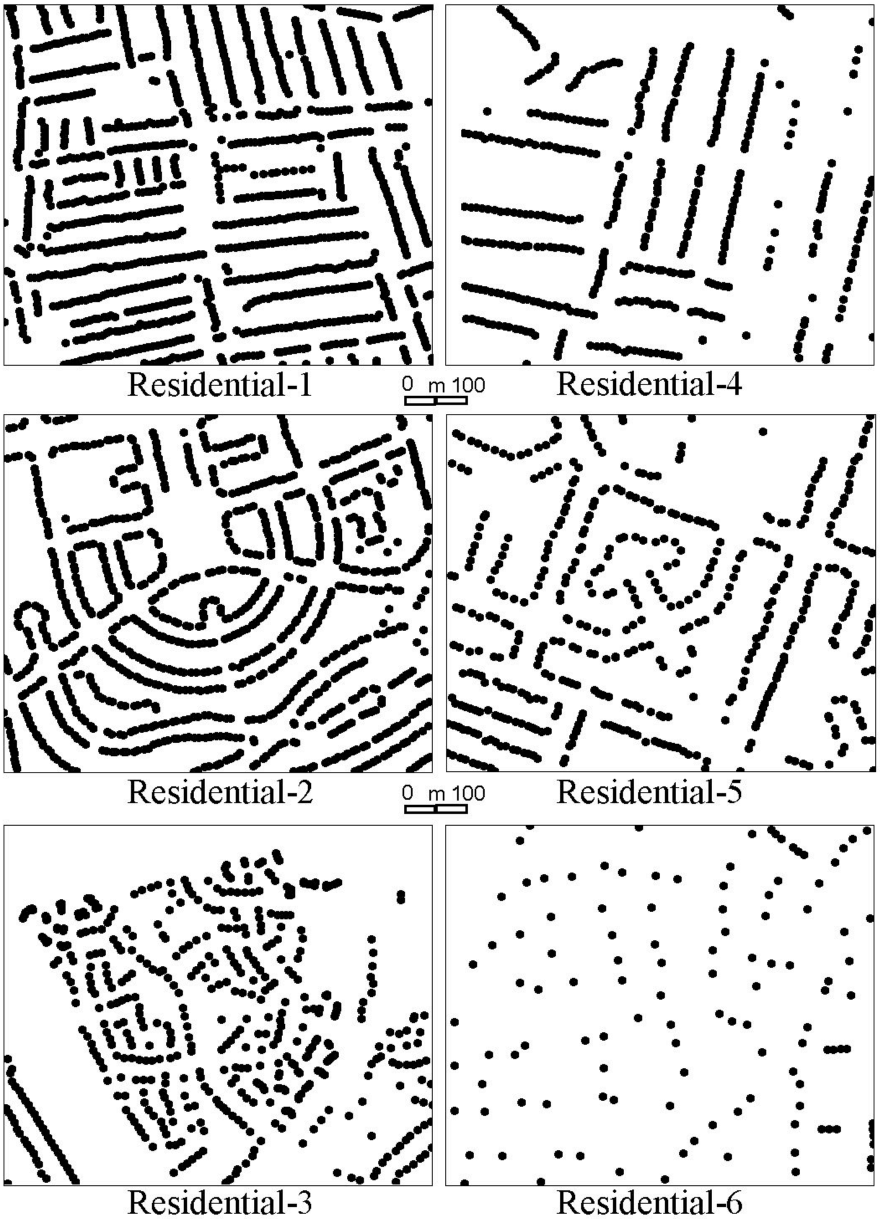

**Figure 2**: Residential ADDRESS-POINTS

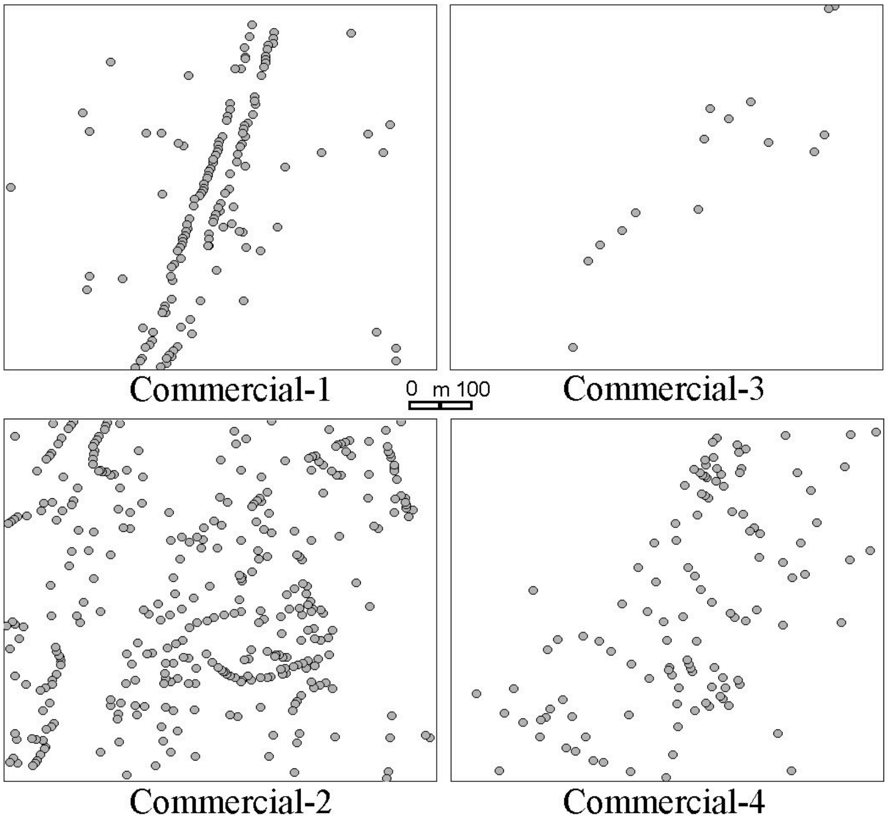

**Figure 3**: Commercial ADDRESS-POINTS

### Postal geography of Bristol, UK

The city of Bristol in southwestern England is large (population of approximately 375,000 at the time of the 2001 census) and dense (occupying around 200 km$^2$). It is an archetypal English city with high traffic congestion and a complex morphological structure. The mix of residential and commercial land use patterns is highly interlaced and only physically discernible by remote sensing in some parts of the central area and the more recent peripheral commercial estates. The spatial arrangement of residential street patterns is highly variable, ranging from dense inter-war linear patterns to modern curved geometrical layouts (Figure 2). Almost every address record is known by the Post Office to be in current use, although the database also contains other minor categories, such as properties under construction. Frequent maintenance updates by the Ordnance Survey ensure that the database is much more contemporary than the census, although it does not contain any of the socio-economic variables associated with the census. At first this may seem as a major disadvantage but not when placed into perspective that the only variables from the census to have had any impact on image classification are population and household categories. Moreover, even these standard census variables are difficult to accommodate into standard image classifications given the aggregate nature of census tract representations. Instead this paper will explore the possibility of calculating indices that characterize the spatial distributions of address points and how these indices can be used to infer land use from classified land cover.

**Spatial indicators of address point distributions**

The nearest neighbor technique is simple but ideal for expressing spatial distributions. It compares the *observed* average distance connecting neighboring points ($D_{OBS}$) and the *expected* distance among neighbors in a random distribution ($D_{RAN}$). The statistic is a straightforward ratio, where randomness is represented by parity; a clustering tendency has values towards 0; and perfect uniformity towards a theoretical value of 2.15. The nearest neighbor statistic $R$ is expressed as,

$$R = {}^{D_{OBS}}\!\big/_{D_{RAN}} \tag{1}$$

where $D_{OBS}$ is the total measured distance between neighboring points divided by the total number of points ($N$), and $D_{RAN}$ is calculated as,

$$D_{RAN} = \frac{1}{2\sqrt{N/A}} \tag{2}$$

where ($N/A$) is the density of points within area $A$. One of many strengths of the nearest neighbor statistic is the facility to compare spatial distributions on a continuous scale, especially when area ($A$) is constant.

| | *N* | *R* | *LN* | *LR* |
|---|---|---|---|---|
| Residential-1 | 1214 | 0.586 | 34 | 0.570 |
| Residential-2 | 1084 | 0.715 | 28 | 0.653 |
| Residential-3 | 503 | 0.910 | 16 | 0.841 |
| Residential-4 | 443 | 0.610 | 30 | 0.598 |
| Residential-5 | 494 | 0.747 | 21 | 0.717 |
| Residential-6 | 132 | 1.528 | 8 | 1.885 |
| Commercial-1 | 155 | 0.523 | 23 | 0.509 |
| Commercial-2 | 637 | 1.297 | 14 | 0.903 |
| Commercial-3 | 18 | 0.627 | 5 | 0.916 |
| Commercial-4 | 321 | 1.289 | 10 | 1.175 |

**Table 2**: Density and nearest neighbor statistics

Of the six different residential neighborhoods measured in the Bristol example, four are successfully identified as having strong clustering patterns (Residential 1, 2, 4 and 5) (Table 2). As expected, given the compact architecture of terraced housing in the UK, the most clustered neighborhoods are the linear patterns of Residential-1 and Residential-4. However, although their nearest neighbor values may be very similar, it is plainly apparent that Residential-1 exhibits a far denser concentration of address points. The same situation applies between Residential-2 (inner-city local government-owned estate) and Residential-5 (more affluent 1980s peripheral estate). Conversely, Residential-3 and Residential-4 have very similar densities yet somewhat dissimilar nearest neighbor values. The remaining neighborhood, Residential-6, is a highly affluent low density area of Bristol with large dwellings, and is the only one demonstrating a uniform tendency. Overall, what is clear is that if nearest neighbor and address point densities are taken together they are valuable measures for identifying and characterizing different residential types.

The same measurements can also be applied to commercial address points (Figure 3). A strong clustering pattern is, again, indicative of linear developments in Commercial-1, but the city center (Commercial-2) and peripheral estates (Commercial-4) exhibit definite signs of uniformity. Again, if both nearest neighbor and density values are taken in combination then unequivocal differences can be revealed. Both Commercial-1 and Commercial-3 (commercial development within residential areas), and Commercial-2 and Commercial-4 have similar nearest neighbor values but contrasting densities.

The conventional nearest neighbor statistic is effective for measuring clustering patterns but it lacks the ability to detect spatial arrangement. A variant of two-dimensional nearest neighbor analysis is the linear readjustment (*LR*) devised by Pinder and Witherick (1973). Instead of measuring all observed distances between neighboring points, $D_{OBS}$ is determined from a linear sequence (*L*) of consecutive points (*LN*) in all directions, whilst $D_{RAN}$ is,

$$LD_{RAN} = 0.5\left(\frac{L}{N-1}\right). \qquad (3)$$

Values for both *R* and *LR* are documented in Table 2. On the whole, they are very similar for both residential and commercial address point distributions. However, *LR* values are usually lower for linear patterns of address points (Residential-1 through 5, and Commercial-1, 2 and 4) and higher for inherently random or uniform distributions (Residential-6 and Commercial-3).

## IMAGE PATTERN RECOGNITION

Previous methodologies designed to improve the accuracy of image classifications representing the city of Bristol have had variable success (Mesev, 1998; Longley and Mesev, 2002). One of the more successful was centered on the use of a surface model to disaggregate census tracts to inform training samples as well as modify the prior probabilities of the classical maximum likelihood discriminant function. The surface model, which was built on a linear distance decay interpolation procedure, was essentially a dasymetric technique that sought to eliminate the non-residential areas of Bristol. Using an empirically-tested spatial resolution of 200m, the location of each surface cell corresponded to the location of the population-weighted centroid of each census tract. The disaggregation effect of the surface model was noticeable and it served as a vehicle for importing census ancillary information into the classification process by producing sharper estimates of the spatial distribution of property types (terraced, detached, semidetached, and apartments) than could be obtained from the image alone.

Although training sample selection using surface model cells was based on *disaggregated* census tracts, the disaggregation was still at a coarse scale (200m) and the information within the surface cell was an average number of households not individual dwellings. It was an improvement on zonal-based census surfaces but the surface cells were nonetheless limited in their usefulness. Address points, on the other hand, represent the entire distribution of individual dwelling units within a city, and as such are the ultimate in disaggregated surfaces. They convey valuable information on local spatial association – density and arrangement – information that is surprisingly overlooked in research on urban image classification, especially given the spatial nature of class distributions and the inherent limitations of spectral data.

Instead of informing image classification, this paper concentrates on pattern recognition, which is arguably more responsive to the extreme spatial heterogeneity of urban areas. Specifically, the objective is to explore the potential of measurements on the spatial structure of residential and commercial developments (from address points) to infer land use from classified land cover. Instead of classifying types of dwellings (detached, semidetached, terraced) address points are used to infer spatial structure on building density (determined by compactness and sparseness in dwelling spacing) and building arrangement (from linearity to randomness). In a way, density and arrangement are very similar to dwelling type, for instance low density and linearity would probably indicate detached housing; medium density and dwelling pairing are more likely of semidetached housing; and high density and linearity would point to terraced housing. However, the additional benefits derived from measurements in the level of linearity or randomness would further place density types within the history of the city’s development, with linear arrangements more characteristic of inner city inter-war and post-war architecture and randomness associated with 1980s and 1990s styles.

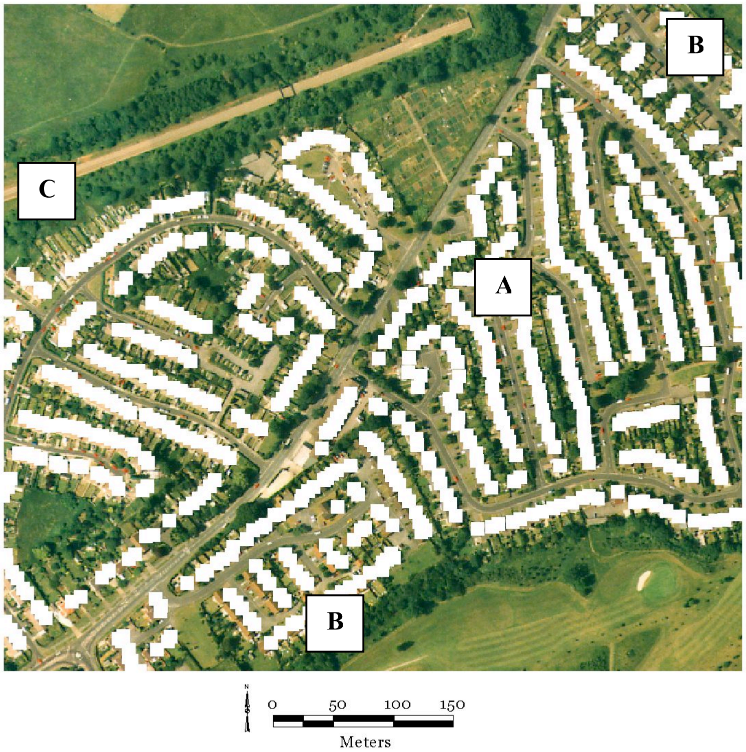

**Figure 4**: Pattern recognition of high spatial resolution imagery

The testing of image pattern recognition by unique address point characteristics is yet to be completed but preliminary results are very encouraging. Some early work includes the inference of residential land use spatial patterns from a classified digital aerial photograph of Bristol produced by Cities Revealed® at 15cm spatial resolution. Figure 4 represents the residential built land cover (shown by white pixels) for a subset of residential types in north Bristol. Using the spatial indices developed by density and nearest neighbor statistics, linear developments were identified in the right of the figure (labeled as “A”), more uniform patterns in the bottom-center and top right (“B”), and curved linear to the left (“C”). In each type of residential land use, density and nearest neighbor values were very close to the samples demonstrated by Figure 2. More testing and refinement is necessary before an automated pattern recognition system can be fully implemented and results evaluated. Nevertheless, there is considerable scope for the use of spatial attributes calculated by nearest neighbor techniques, as well as more sophisticated spatial metrics (Wu *et al*, 2000), in recognizing urban land use patterns. The importance of such work is especially relevant given the recent proliferation of very high spatial resolution imagery at 4m/1m and 2.4m/0.6m from the IKONOS and QUICKBIRD sensors respectively. Urban land cover Information at such spatial resolutions is discrete and highly identifiable. However, subsidiary information, in the form of ADDRESS-POINTS for example, is critical for converting urban land cover to urban land use with a reliable degree of consistency and accuracy.

## CONCLUSION

The OS ADDRESS-POINT product represents a type of disaggregated urban data set with tremendous opportunity for inferring land use from remote sensing. It can be used to help discern at least three geographies: the built environment, commercial, and residential, which are geographically exhaustive, regularly updated, and highly precise. On-going research has been presented on the possibilities for generating unique summaries of the various structural patterns of address points representing density and linear arrangement of residential and commercial buildings. These summaries have immense potential for inferring land use from land cover patterns classified from high spatial resolution remotely sensed data. Results so far, for the city of Bristol in southwest England, are most encouraging but further testing is crucial if an even closer relationship between imagery and postal information can be statistically established. In particular, research breakthroughs are needed in linking classified land cover with land use using not only nearest neighbor but also entropy maximization (Harvey, 2002); as well as the resolution of non-residential from residential land use patterns using targeted training sample selections; neighborhood differentiation using invariant fractal dimensions (Longley and Mesev, 2002), and urban growth using spatial metrics (Pesaresi, M. and Bianchin, 2001).

The methodology of characterizing address points for use in image pattern recognition is an effective means of integrating GIS data with remote sensing where the benefits of both are harmonized in the pursuit of greater accuracy (Mesev, 2004). A future research direction would be to establish direct relationships between socio-economic information within census tracts and the spatial distribution of address points. In this way, both attribute and spatial indicators would be readily available to inform multispectral classifications of urban areas. However, for the time being, research is focused on the spatial utility of address points and how spatial indices can be used to infer land use from high spatial resolution land cover data. Once residential and commercial characteristics from address points are established and comprehensively tested, a situation is envisaged where land cover patterns can be routinely categorized into a variety of types. A fully automated procedure is currently being built with many more address point configurations, which will allow consistent pattern recognitions both across settlements and through time. The advent of disaggregated models, such as address points, represents a major step forward in precision urban mapping, which is intuitively more realistic than the uniformity of traditional areal representations.

## ACKNOWLEDGEMENTS

ADDRESS-POINT is Crown Copyright (www.ordsvy.gov.uk). Cities Revealed is the copyright of the GeoInformation Group (www.crworld.co.uk). Preliminary work for this study was funded by the UK Natural Environment Research Council (NERC grant GST/02/2241). The author would like to thank three referees for their useful comments.

## REFERENCES

Barnsley, M.J. and Barr, S.L., 1997. A graph-based structural pattern recognition system to infer land use from fine spatial resolution land cover data*. Computers, Environment and Urban Systems*, 21, 209-225.

Chen, K., 2002. An approach to linking remotely sensed data and areal census data. *International Journal of Remote Sensing*, 23, pp. 37-48.

Donnay, J.-P., 1999. Use of remote sensing information in planning. In: Geertman, S. and Openshaw, S. (eds) *Geographical Information and Planning*. Springer, Berlin, pp. 242-260.

Forster, B.C., 1985. An examination of some problems and solutions in monitoring urban areas from satellite platforms. *International Journal of Remote Sensing*, 6, pp. 139-151.

Geoghegan, J., Pritchard, Jr L., Ogneva-Himmelberger, Y., Chowdbury, R.R., Sanderson, S. and Turner, B.L., 1998. “Socializing the pixel” and “pixelizing the social” in land use and land cover change. In: Liverman, D., Moran, E.F., Rindfuss, R.R. and Stern, P.C. (eds) People and Places: *Linking Remote Sensing and Social Science*. National Academy Press, Washington DC, pp. 51-69.

Harris, R.J. and Longley, P.A., 2000. New data and approaches for urban analysis: modelling residential densities. *Transactions in GIS*, 4, pp. 217-234.

Harvey, J.T., 2002. Estimating census district populations from satellite imagery: some approaches and limitations. *International Journal of Remote Sensing*, 23, pp. 2071-2095.

Langford, M., 2003. Refining methods for dasymetric mapping using satellite remote sensing. In: Mesev, V., 2003. *Remotely Sensed Cities*. Taylor & Francis, London, pp. 137-156.

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

Longley, P.A., 2002. Geographical information systems: will developments in urban remote sensing and GIS lead to ‘better’ urban geography? Progress in Human Geography, 26, pp. 231-239.

Longley, P.A. and Mesev, V., 2002. Measurement of density gradients and space-filling in urban systems. *Papers in Regional Science*, 81, pp. 1-28.

Mesev, V., 1998. The use of census data in urban image classification. *Photogrammetric Engineering and Remote Sensing*, 64, pp. 431-438.

Mesev, V., 2001. Modified maximum likelihood classifications of urban land use: spatial segmentation of prior probabilities. *Geocarto International*, 16(4), pp. 39-46.

Mesev, V., 2003. *Remotely Sensed Cities*. Taylor & Francis, London.

Mesev, V., 2004. *Integration of GIS and Remote Sensing*. Wiley, Chichester (forthcoming).

Möller-Jensen, L., 1990. Knowledge-based classification of an urban area using texture and context information in Landsat-TM imagery. *Photogrammetric Engineering and Remote Sensing*, 56, pp. 899-904.

Pesaresi, M. and Bianchin, A., 2001. Recognizing settlement structure using mathematical morphology and image texture. In: Donnay, J.-P., Barnsley, M.J. and Longley, P.A. (eds) *Remote Sensing and Urban Analysis*, Taylor & Francis, London, pp. 55-67.

Pinder, D.A. and Witherick, M.E., 1973. Nearest neighbor analysis of linear point patterns. *Tijdschift voor Economische an Sociale Geographie*, 64.

Wu, J., Jelinski, E.J., Luck, M. and Tueller, P.T., 2000. Multiscale analysis of landscape heterogeneity: scale variance and pattern metric. *Geographic Information Sciences*, 6, pp. 6-16.