Using GIS to Model the Diurnal Variation of Urban Population Distribution

Christopher NEEDHAM

The GI_Forum Program Committee accepted this paper as reviewed full paper.

Abstract

The spatial distribution of urban population varies widely throughout the day as people move from place to place. Information about the location of people at different times may assist emergency responders in the event of a disaster, although insufficient data has constrained modelling of the phenomenon. This paper explores the development of a model to simulate the diurnal variation of urban population distribution using the concept of temporal profiles to calculate population according to land use and time of day. A prototype software application is produced using GIS to enable the distribution of population to be simulated for different times. The model generates a more refined surface than existing models that use census data. Spatial analyses may then be used to calculate the estimated population within evacuation zones at different times of the day, week and year.

1 Introduction

Population distribution is ephemeral and its concentration in cities may vary by many orders of magnitude in a matter of hours. Modelling this distribution is highly complex and obtaining data at a suitable spatial and temporal granularity is challenging. As surveys with the explicit intention of mapping the daytime location of all people in urban areas are not routinely undertaken, population data for existing models have typically been derived from the decennial census, which is largely a record of population distribution at night as the main place of residence is recorded. If a disaster occurs during the daytime when the distribution of population is very different, then census data are insufficient for emergency responders to determine where people may be located. Methods used to acquire data for other geographical phenomena that also change rapidly, such as interpolating point samples to determine atmospheric temperature cannot be applied as population is fundamentally punctiform in nature, i.e. composed of individual people. Moreover, it is undesirable to record the location of people at such a scale due to the risk of identifying individuals. Modelling human activity can only be performed using population as an aggregate quantity of individual people and is therefore subject to generalisation.

The aim of this paper is to develop a mathematical model to calculate the diurnal variation of urban population distribution, assess potential data sources and devise a prototype software application to enable population distribution to be simulated for different times.

2 Review of Diurnal Population Modelling

The realisation that 'night-time' census data were of limited use in calculating daytime population change was made in the 1930's (WIRTH 1938; ZANDVLIET & DIJST 2005) although little research had been conducted until recently due to lack of alternative data (AL-SABHAN et al. 2003; SMITH 1989; TOBLER et al. 1997; WU et al. 2005). Census data are generally too coarse for modelling urban areas and omit large sections of the population (SAILOR & LU 2004; Smith et al. 2005) by assuming the daytime population of an area is merely the residential plus workplace populations, less resident workers to avoid double counting. This makes assumptions about population distribution at different times by age and economic status where activities such as travel or leisure are not accounted for. Current research is attempting to indicate the presence of people by deriving surrogate data from other sources, e.g. mobile telephone records, building footprints, land use and night-time light emissions (BRIGGS et al. 2007; MCPHERSON et al. 2006; MENNIS 2003; MOONEY & WALKER 2002; READES et al. 2007; SUTTON et al. 2001; XIONG 2004).

Population has primarily been represented in GIS by either the object view where areas are assigned a value or the field view where population density is defined as a continuously varying surface. Population density has long been visualised using choropleth maps. These have several shortcomings, specifically the use of arbitrary and irregularly shaped areas that have no real significance, an assumption of uniform population distribution within zones that masks internal variation, inappropriate class boundaries and the visual dominance of larger areas (LANGFORD 2005; MARTIN 1989; WU et al. 2005; YUAN et al. 1997). Dasymetric mapping seeks to refine population distribution to areas likely to be populated using a secondary variable (WRIGHT 1936). Choropleth maps may be transformed to surfaces using interpolation and represented using regular raster grids. These have advantages over choropleth maps by maintaining variation of population density over space (Martin 1996), although the field view breaks down when scale becomes so large that individuals are identified (LONGLEY et al. 2001 p. 69). Both representations are generalisations of reality as population does not vary continuously over space (BRACKEN 1994 p. 81; MENNIS 2003) unless when considered as a density (HOLT et al. 2004).

Recent research into daytime population modelling has been conducted by the Los Alamos and Oak Ridge National Laboratories (MCPHERSON et al. 2006; ORNL 2006). The model developed by Staffordshire University considers population based on activities and locations (MOONEY & WALKER 2002; SMITH et al. 2005). Research coordinated at the Massachusetts Institute of Technology may provide arguably the finest data source where mobile telephone location data are exploited to determine real time population densities (RATTI et al. 2006). Models can provide emergency planners with a valuable tool to determine the location of population at risk from hazards and enable evacuation plans to be devised. Although emergency responders find it easier to understand risk from maps than tabular data (HEINO & KAKKO 1998), there appears to be lack of demand for models from planners who do not see GIS as a high priority (THOMASSON & ELVES 2005). Responders are more concerned with general movement of people than high spatial precision and accuracy (ZERGER & SMITH 2003). SMITH et al. (2005) concede that maintaining database currency is problematic given processing times.

3 Building a Diurnal Model

The hypothesis presented infers locations are populated to various degrees throughout the day as people congregate to engage in activities related to the location's land use (WANG & CHENG 2001). The concept involves modelling population as an attribute of static locations rather than attempting to track a mobile population.

The basic components of the model are locations with population and land use properties, and time. Locations may be identified as discrete geographical features defined by land use e.g. parks, roads or buildings, represented in GIS respectively as areas, lines or points. The National Land Use Database (NLUD) provides a standard classification of land uses (ODPM 2006). Assuming population data sources count all people, modelling population based on land use inherently includes all sections of society. Population should be referenced to features for which population data are collected. Frequently these are point sources, e.g. OS Address-Point co-ordinates, addresses or postcodes which help overcome data referenced to artificial census areas that assume homogenous distribution of population.

The next stage of the modelling process involves defining the relationships between system components mathematically. The population of geographical features vary according to the time of day, week and year. By expressing time in this form, factors may be assigned to each hour, day and month to represent the relationship between cyclical time and population, for example:

$$x_t = ya_t$$

where x is the population of the feature at a particular hour t, a is the hour factor and y is the feature's 'usual' population. If there are half the usual number of people at the feature at 7pm (t) then the hour factor (a) would be 0.5. If the usual population (y) of the feature is 146, then the population (x) at 7pm would be:

$$x_t = 146 \text{ x } 0.5 = 73$$

The mathematical model consists of several elements that represent, at a particular time, the total population of a city comprised of many features such as schools, households or workplaces whose usual populations are adjusted according to their land use by factors that depend on the hour of the day, day of the week, month of the year and school terms. The model is expressed algebraically as:

$$x_t = \sum_{i=1}^{n} \left(y_{ij}\, a_{tj}\, b_{tj}\, c_{tj}\, d_{tj} \right)$$

where x_t is the total population of the city comprised of n features at time t, y_{ij} is the usual population of feature i that has land use j, a_{tj} is the hour factor of land use j at time t, b_{tj} is the weekday factor of land use j at time t, c_{tj} is the month factor of land use j at time t and d_{tj} is the school term time factor of land use j at time t.

The usual population of a feature depends on the data available for the feature type. The usual population of a household, for example, is the number of residents recorded by the census. It may be fully populated at night when residents are home but its population may fall during the day as the residents go to work or school. The usual population of other features may be defined by the average daily visitor population which is then apportioned throughout the hours of the day using temporal factors. The factors form ‘temporal profiles’ that represent the proportion of the usual population present throughout the time period. Figure 1 depicts a diurnal profile for road trips where the morning and afternoon peaks are clearly visible. Population distribution also varies widely throughout the week and is highly seasonal (SMITH et al. 2005). Where the temporal profile is unknown for a particular location or land use, use of surrogate data that indicate the presence of people at different times such as retail sales or hotel stays should be considered (SMITH 1989).

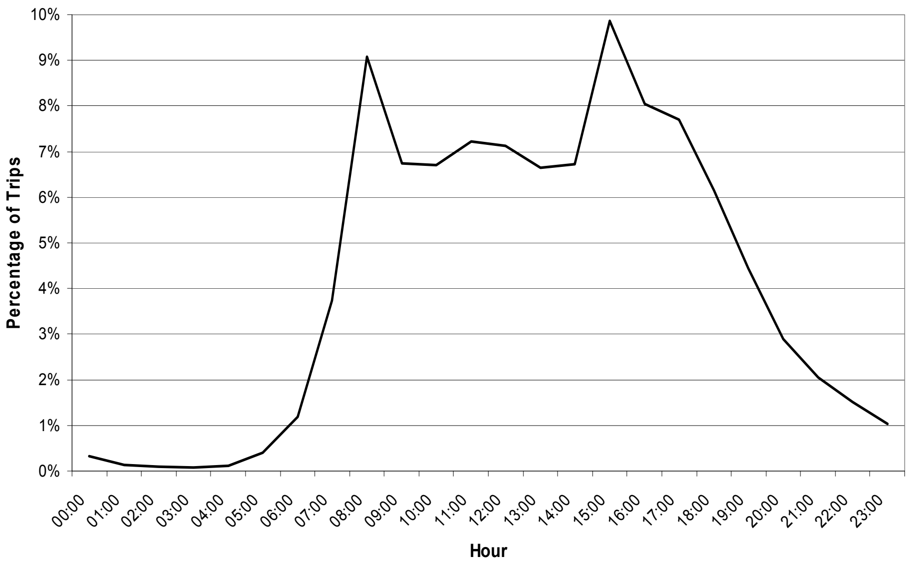

Fig. 1: Proportion of daily road trips by hour (2002-04) (DfT, personal email 20/2/07)

Temporal interpolation involves interpolating the population of a location between times as unsampled points may be interpolated across space. If the usual population of a location is taken as the daytime maximum and it is assumed unpopulated at night, then population between these times may be interpolated. However, a linear interpolation between the maximum and minimum could erroneously assign population to times when the location is unpopulated. In reality, the change in population between times may be more abrupt; a workplace may be unpopulated at 8am but fully populated by 9am as employees arrive.

Population data are collated from multiple sources as no single source is adequate (MCPHERSON & BROWN 2004; MOONEY & WALKER 2002; SMITH 1989; SMITH et al. 2005). Population data sources for each land use function are considered in turn. The smallest area for which residential population census data are disclosed in the UK is the Output Area, so the population of individual households cannot be determined precisely. The average Output Area household population is therefore assigned to individual households.

Workforce population datasets are collated by some local authorities but there are several problems with currency, completeness, consistency and referencing businesses with multiple buildings to a single address. The Inter-Departmental Business Register (IDBR) records employee counts and references premises by postal address. Population data for schools, colleges of higher education, hospitals, care homes, prisons and military bases may be derived from respective government departments and agencies that are responsible for these types of facilities. Monthly visitor numbers are available for some popular visitor attractions, but general tourism data are only available at county level (STARUK 2006). Tourist populations are apportioned between areas that tourists may visit such as parks or leisure centres. Data for retail populations are difficult to obtain. The method adopted in this study uses retail employment data as a surrogate for retail population on the basis that the number of staff employed is a reasonable indicator of level of custom. Stadia population data may be derived from average attendances available from the stadia in question. The approach taken to determine the road network population is to calculate the average vehicle density per length of road by dividing the average traffic flow per day for each class of road by the average vehicle speed. The population is then derived by multiplying the vehicle density by the average vehicle occupancy level, adjusted by the Department for Transport's (DfT) data on trips in progress by hour of the day, day of the week and month of the year. The same principle can be applied to population using the railway network. Transport terminus passenger data are available from the Office for Rail Regulation for railway stations and the DfT for airports and ports.

The primary locational data are address points from the Ordnance Survey's MasterMap Address Layer 2 that include NLUD land uses. Population values are then assigned to each address point. A problem occurs however when the population of an extensive geographical feature, such as a hospital or airport is assigned to a single point which would erroneously concentrate the population in one location and may result in it being falsely excluded from spatial analyses if the address point falls outside the buffer zone. The solution adopted is to divide the population proportionately between buildings based on estimates of gross floor space. Consequently, where buffer zones bisect a site, a proportion of the population is captured. For large features where precise population distribution is unknown, such as retail areas or visitor attractions, population is distributed evenly across the area to regularly spaced points. Linear features such as roads or railways are segmented and population assigned to points that represent the course of the feature. The point features are then converted to a raster surface and cells assigned the total population of the points in each cell. This introduces error into the model due to the loss in positional accuracy, although given the uncertainties of diurnal modelling in general, the error is considered minor on the proviso that a buffer equal to at least the minimum mapping unit is included in subsequent spatial analyses. The surface is displayed against a backdrop map which is important to put the results in context.

There will inevitably be a process of compromise from the ideal model to the achievable given the availability of data. A generic, repeatable model must be based on a supply of nationally consistent data, although additional local data if available may further refine the model. Even if such data are collected, they may not be disclosed for reasons of confidentiality or commercial sensitivity. Data can therefore only be used if they are disclosed. The aim of this study is not to devise theoretical methods to capture population data, but to model population based on actual data that are currently available. MOONEY & WALKER (2002) concede that the perfect solution is not achievable. To validate the model, it must generate results that can be tested against observed data. However, as data are not collected at the spatial and temporal resolution required, it can only be tested against results produced by other models. As the model generalises the distribution of population throughout the day, it must be stressed that the results are inevitably subject to uncertainty. Uncertainty is manifested as locational – not knowing precisely where people are located, temporal – not knowing precisely when areas are populated and quantitative – not knowing precisely how many people are at a location at a given time.

4 Results

The model was run for every hour of a day to assess the diurnal change in population distribution. A simulated evacuation zone around the city centre was then used to determine the extent of an evacuation at various times. To assess whether the results are more spatially and temporally refined than those produced using only census data, it is necessary to produce simulations for the case study area based on such data. Figure 2 shows a series of simulations at key times during an arbitrarily selected day, Tuesday 19th June 2007.

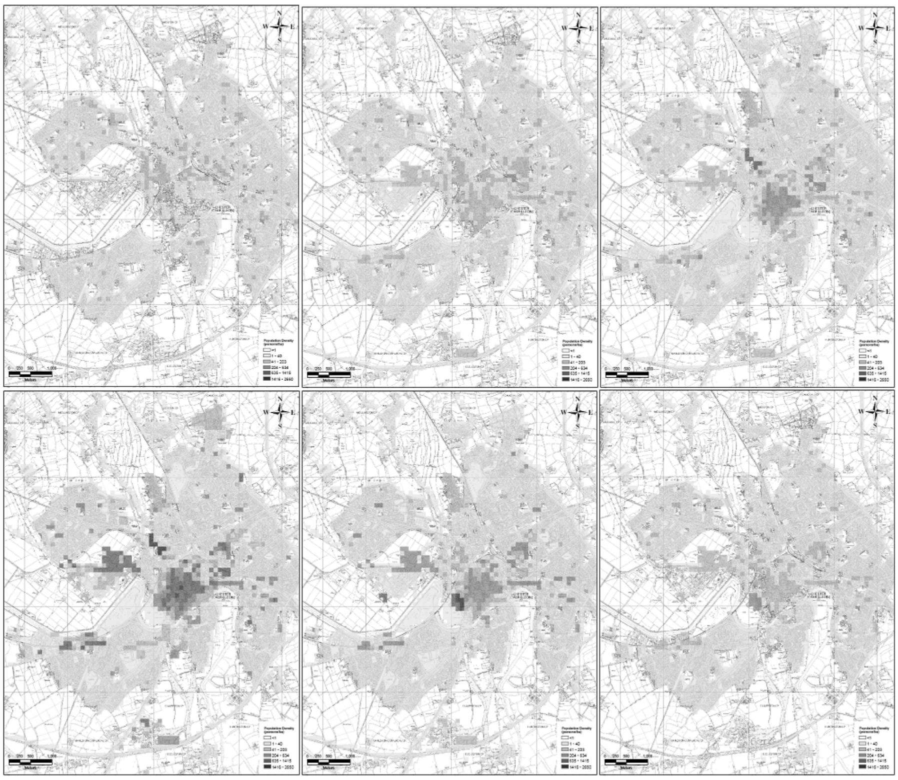

Fig. 2: Simulated population distribution at key times during June weekday (upper left: 4am; upper centre: 7am; upper right: 8am; lower left: 11am; lower centre: 6pm; lower right: 10pm)

There was found to be little change in population distribution between midnight and 6am, but a rapid change during the morning peak period as workers commute to work in the city centre and other places of employment. Population intensifies in the city centre and commercial/industrial areas during the day until the afternoon peak, after which the city centre begins to empty, continuing gradually during the evening.

A 1km radius evacuation zone around Chester city centre is used for the simulation. Figure 3 shows the predicted population within the evacuation zone throughout the day. Over thirty times as many people are present at the daytime peak than the night-time minimum. At the peak time of year for tourists, in August, the population within the evacuation zone is predicted to reach over 110,000. The usual resident population within the zone is approximately 6,700 people.

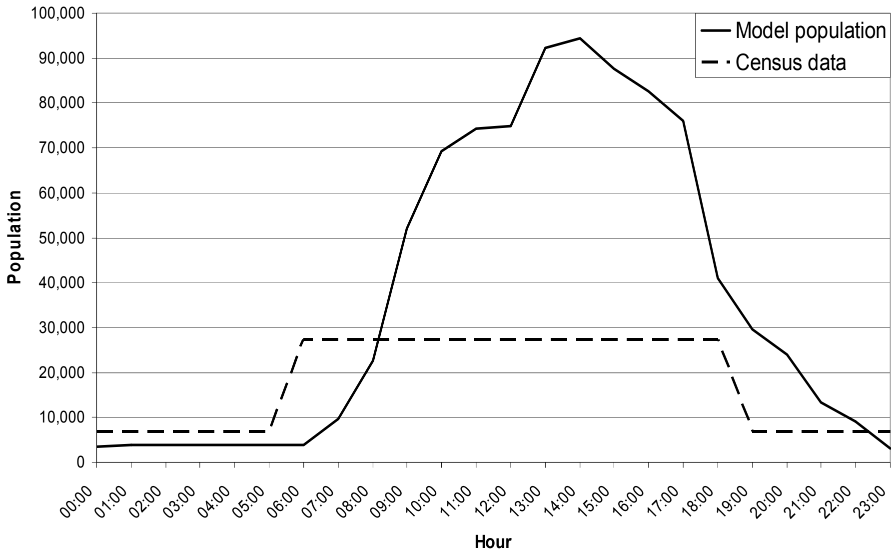

Fig. 3: Diurnal population within evacuation zone predicted by model and census data

The results were compared against output produced by existing models using readily available census data. The predicted population within the evacuation zone using such data is also shown in Figure 3. The daytime peak is less than one third of that predicted by the model in this study and no variation exists throughout daytime hours. The key differences in daytime population distribution produced by the existing model were that agricultural areas appeared populated as they were included in rural output areas that encompass large tracts of agricultural land, out-of-town business parks appeared populated at a much lower density as they were subsumed within large rural output areas, and areas that would be more densely populated with shoppers and tourists appear populated only by employees and daytime residential populations. Night-time simulations using existing models showed more similarities, but the city centre, retail and industrial areas appear to be populated as they were included in large output areas that contained some residential population.

Problems occur when the currency of the background map lags more recent population data; constructed buildings are not depicted, so apparently empty areas appear populated. Conversely, features on the background map may appear unpopulated because the source data are not current or erroneous. The quality of output is no better than the data quality as errors are propagated throughout the model.

5 Conclusions

The results confirm the hypothesis that population is concentrated in the city centre, industrial and commercial areas during the day and is relatively disbursed throughout residential areas at night. The model produces a more refined distribution of population when compared to the output of existing models, both spatially to more precise locations, and temporally to different hours of the day rather than binary day and night-times, and includes sections of the population omitted from other models. Testing the model against real world data would require the collection of data at a very high regular spatial and temporal resolution which would be prohibitively expensive, although current research using mobile telephone logs in the SENSEable City Laboratory at MIT may provide such data.

Recommendations to refine the model further include replacing the NLUD land use classification that has 41 types with the OS Base Function that has approximately 1700, the incorporation of additional profile types such as public holidays or major sporting/cultural events and replacing the monthly profile with a weekly profile.

References

AL-SABHAN, W., MULLIGAN, M. & BLACKBURN, G. A. (2003), A real-time hydrological for flood prediction using GIS and the WWW. Computers, Environment and Urban Systems, 27, pp. 9-32.

BRACKEN, I. (1994), Towards improved visualization of socioeconomic data. In Visualization in Geographical Information Systems, edited by HEARNSHAW & UNWIN, (Wiley, Chichester), pp. 76-84.

BRIGGS, D. J., GULLIVER, J., FECHT, D. & VIENNEAU, D. M. (2007), Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote Sensing of Environment, 108, pp. 451-466.

HEINO, P. & KAKKO, R. (1998), Risk assessment modelling and visualisation. Safety Science, 30, pp. 71-77.

HOLT, J. C., LO, C. & HODLER, T. (2004), Dasymetric estimation of population density and areal interpolation of census data. Cartography and Geographic Information Science, 31, pp. 103-121.

LANGFORD, M. (2005), Dasymetric mapping of industry and commerce across Wales, UK, using geospatial integration of national mapping agency products [online]. Accessed 28/7/06, http://www.acsm.net/sessions05/langford05.pdf.

LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J. & RHIND, D. W. (2001), Geographical Information Systems and Science. Chichester, Wiley.

MARTIN, D. (1989), Mapping Population Data from Zone Centroid Locations. Transactions of the Institute of British Geographers, 14 (1), pp. 90-97.

MARTIN, D. (1996), An assessment of surface and zonal models of population. International Journal of Geographical Information Systems, 10 (8), pp. 973-989.

MCPHERSON, T. N. & BROWN, M. J. (2004), Estimating Daytime and Nighttime Population Distributions in U.S. Cities for Emergency Response Activities. Symposium on Planning, Nowcasting, and Forecasting in the Urban Zone, January 10-16, Seattle [online]. Accessed 19/12/06, http://ams.confex.com/ams/84Annual/techprogram/paper_74017.htm.

MCPHERSON, T. N., RUSH, J. F., KHALSA, H., IVEY, A. & BROWN, M. J. (2006), A day-night population exchange model for better exposure and consequent management assessments [online]. Accessed 2/4/06, http://ams.confex.com/ams/pdfpapers/105209.pdf.

MENNIS, J. (2003), Generating Surface Models of Population Using Dasymetric Mapping. The Professional Geographer, 55(1), pp. 31-42.

MOONEY, J. & WALKER, G. (2002), The derivation and use of population data for major hazard accident modelling. Health & Safety Executive, HMSO, Norwich.

ODPM (2006), National Land Use Database: Land Use and Land Cover Classification [online]. Accessed 10/1/07, http://www.communities.gov.uk/pub/847/NationalLandUseDatabaseLandUseandLandCoverClassificationPDF2680Kb_id1163847.pdf.

ORNL (2006), LandScan Global Population Databases [online]. Accessed 3/4/06, http://computing.ornl.gov/cse_home/about/LandScan%20long.pdf.

RATTI, C., PULSELLI, R. M., WILLIAMS, S. & FRENCHMAN, D. (2006), Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis [online]. Accessed 13/4/09, http://senseable.mit.edu/papers/pdf/2006_Ratti_Pulselli_Williams_Frenchman_EPB.pdf.

READES, J., CALABRESE, F., SEVTSUK, A. & RATTI, C. (2007), Cellular Census: Explorations in Urban Data Collection. IEEE Pervasive Computing, 6(3), pp. 30-38.

SAILOR, D. J. & LU, L. (2004), A top-down methodology for developing diurnal and seasonal anthropogenic heating profiles for urban areas. Atmospheric Environment, 38 (17), pp. 2737-2748.

SMITH, G., ARNOT, C., FAIRBURN, J. & WALKER, G. (2005), A National Population Data Base for Major Accident Hazard Modelling. Health & Safety Executive, HMSO, Norwich.

SMITH, S. K. (1989), Toward a methodology for estimating temporary residents. Journal of the American Statistical Association, 84 (406), pp. 430-436.

STARUK (2006), Statistics on Tourism & Research website [online]. Accessed 27/9/06. http://www.staruk.org.uk/.

SUTTON, P., ROBERTS, D., ELVIDGE, C. & BAUGH, K. (2001), Census from Heaven: an estimate of the global human population using night-time satellite imagery. International Journal of Remote Sensing, 22 (16), pp. 3061-3076.

THOMASSON, E. & ELVES, S. (2005), Emergency Management: delivering cross boundary intelligence with GIS. Emergency Planning College, December 2, York [online]. Accessed 2/11/06, http://www.epcollege.gov.uk/upload/assets/www.epcollege.gov.uk/13.pdf.

TOBLER, W., DEICHMANN, U., GOTTSEGEN, J. & MALOY, K. (1997), World Population in a Grid of Spherical Quadrilaterals. International Journal of Population Geography, 3, pp. 203-225.

WALKER, G. P. & MOONEY, J. (1998), Spatially referenced population data for land use planning advice. Health & Safety Executive, HMSO, Norwich.

WANG, D. & CHENG, T. (2001), A spatio-temporal data model for activity-based transport demand modelling. International Journal of Geographical Information Science, 15 (6), pp. 561-585.

WIRTH, L. (1938), Urbanism as a way of life. American Journal of Sociology, 44 (1), pp. 1-24.

WRIGHT, J. K. (1936), A Method of Mapping Densities of Population: With Cape Cod as an Example. Geographical Review, 26 (1), pp. 103-110.

WU, S.-S., QIU, X. & WANG, L. (2005), Population Estimation Methods in GIS and Remote Sensing: A Review. GIScience and Remote Sensing, 42 (1), pp. 80-96.

XIONG, D. (2004), GIS and Remote Sensing Support for Evacuation Analysis. Geospatial Information Systems for Transportation Symposium, March 29-31, Rapid City [online]. Accessed 26/4/06, http://www.gis-t.org/files/5levz.pdf.

YUAN, Y., SMITH, R. M. & LIMP, W. F. (1997), Remodelling census population with spatial information from Landsat TM imagery. Computers, Environment and Urban Studies, 21 (3/4), pp. 245-258.

ZANDVLIET, R. & DIJST, M. (2005), The ebb and flow of temporary populations: the dimensions of spatial-temporal distributions of daytime visitors in The Netherlands. Urban Geography, 26 (4), pp. 353-364.

ZERGER, A. & SMITH, D. I. (2003), Impediments for using GIS for real-time disaster decision support. Computers, Environment and Urban Systems, 27, pp. 123-141.