Progress report: 24-hour gridded population models

David Martin, Samantha Cockings and Samuel Leung

School of Geography, University of Southampton, Southampton, SO17 1BJ, UK

Email: D.J.Martin@soton.ac.uk

Paper presented at the European Forum for Geostatistics Conference 2010, Tallinn, Estonia, 5-7 October 2010

Introduction

In a paper presented at the 2009 EFGS Conference in The Hague (Martin et al., 2009) we set out the early stages of a UK research project focused on the production of time-specific gridded population models. The present paper reviews overall progress with this work and develops in more detail some aspects of the modelling principles and available data sources which make this possible. The paper seeks to demonstrate how readily accessible data are beginning to provide many of the key inputs required to make time-specific representations of population. Time-specific models have the potential to overcome many key weaknesses of current representations which are generally based on dated population statistics about population numbers at place of residence and which fail to reflect massive cyclical spatial redistributions of population. In many application areas, current analyses persist in using population maps with major deficiencies due to the absence of plausible alternative models. The remainder of this paper is structured as follows: the following section provides a brief review of the case for more sophisticated time-space population distributions. The third section reviews issues arising from the challenge of data source acquisition and integration and the final section indicates the progress made in the project to date with regard to addressing these challenges.

The need for better space-time population distributions

Conventional population mapping, whether based on regular grids or irregular areal units, relies on 'night-time' residential population counts. This is essentially due to the residential basis used by most censuses and population registration systems, whereby an individual's reported place of residence is taken to represent their geographical location for all purposes. Whether geographical referencing takes place at the address, postal code or small area level, the result is to create a geographical representation of population at residential addresses and for most purposes this therefore generates a map of population at their night time locations (when most population members are at home). One interpretation of this mapping in time-space is presented in Figure 1(a) in which the complete population remains associated with residential locations through all time periods, while the many other locations associated with human activity remain entirely unpopulated at all times. These distributions are entirely appropriate for many types of use, especially the planning of services and resource allocation which are essentially based on residence assumptions such as demand for primary school places or community-based health care services. There is nevertheless widespread demand for population maps which are more temporally appropriate for specific purposes and this can be separated into two elements. Firstly, there is demand for population maps which are as up to date as possible in chronological time: in this respect systems based on continuously-maintained population registration systems are far superior to those based on decennial census data, in which the conventional population map may represent the population distribution at night, up to 11 years ago (for example, with 2011-based small area data not replacing published 2001-based counts until late 2012). The second and more important aspect of temporal validity is demand for maps which realistically represent the distribution of population at specific times, reflecting the massive redistribution of population on daily, weekly, termly and seasonal timecycles through travel to work, school, college, business, leisure and vacation etc. These considerations are particularly pertinent to any application where population 'exposure' is a relevant consideration. This may be interpreted as direct exposure to hazard, such as an incident which occurs at a specific time and place – for example an explosion at an oil refinery requiring the mass evacuation of local populations; or assessment of the long-term exposure of population to hazards such as flooding or atmospheric pollution. A very different interpretation of the same process would be predicting the potential catchment population for a retail outlet, which will be massively affected by population movements during the day and, for the most part, unrelated to residential patterns, particularly in major commercial centres. There has been steady and growing interest in the development of population maps which attempt to adjust night-time populations to reflect 'day-time' distributions, particularly for the purposes of emergency planning.

Examples include Sleeter and Wood (2006), who propose a method aimed at providing more realistic daytime population estimates for emergency planning, as outlined above. They use US census data for small areas and transfer working populations out of residential areas during the daytime and redistribute these onto schools and workplace locations derived from a business directory, subdividing the population between school and working age groups. McPherson et al. (2006) present a national model for the US, this time based on a gridded model and again using the binary division of daytime and night-time. McPherson and Brown (2004) present static daytime and night-time models by allocating population to night-time and daytime residential locations, and to employment locations in the daytime model. McPherson et al. (2006) recognise the importance of time spent in travel and attempt to model population numbers in the transportation system. Their particular application example is that of planning for population transfer to hospital following a hypothetical airborne release of a hazardous substance in a densely populated area. In the UK, Smith and Fairburn (2008) provide a GIS database approach, which is not an explicit population model, but rather a geographical feature database with associated population capacities and annotation about temporal usage. In this approach, it is anticipated that the analyst will use the database to inform judgements about the population likely to be present in an area under any particular set of circumstances, for example following the commencement of a serious incident or in order to explore specific emergency planning scenarios.

Data source acquisition and integration

A key limitation to all published attempts to produce more sophisticated time-space representations of population distribution has been the relative paucity of non-residential population location data, and particularly the temporal component. Whereas official statistical sources provide good coverage of residential locations and to a certain extent cover workplace locations (either through census questions about place of employment or by linkage of population and employment registers, depending on country), there has until recently been very little formal recording (and even less publication) of population numbers associated with the numerous other types of population activity. This situation is changing rapidly with the explosion of relevant data on the web – most recently through deliberately constructed national data portals such as http://data.gov.uk and http://data.gov. Additional impetus to the exposure of these data has been the growth of the linked data community (Shadbolt et al., 2006) but there had already been a steadily growing tendency for (especially publicly-funded) organizations to use the web to publish service performance figures, which often take the form of user or visitor numbers. The time profiles of activities such as service opening hours, school term dates, etc. are also widely available on the web.

One approach to the estimation of time-specific population models is to work from remotely sensed data in order to measure population presence at different times, for example using night-time lights, and this is essentially the approach adopted by the Landscan USA project (Bhadhuri et al., 2007), a major initiative which recognises the importance of the temporal dimension in population representation. The approach used in our work has been to develop the population grid modelling method originally proposed for static residential populations by Martin (1989) and to extend this to operate on a database of activity locations at which the presence of population is described in both time and space, as outlined in Martin et al. (2009). Our conceptual model has much in common with that of Ahola et al. (2007) which sets out the relationships between geographical objects and their occupation at different times by different population sub-groups. The recent explosion in available data has begun to make such modelling a viable option over large areas. Our implementation treats all population locations as centroids (single x,y coordinate pairs) with an associated time profile for the presence of population, which may be applied separately to different population sub-groups (such as age groups). Gridded models are then assembled from this centroid database which use the specific values associated with a target time. A major focus of our work has therefore been on the assembly of a suitable dataset which could be interpreted in the terms of Figure 1(b) as representing the proportional shift of population over time between the spatial locations of a range of different activity types. The diagram shows only a single day, but weekly and longer time-cycles may readily be used.

Our data sources can be divided into three principal types: those relating to residential, non-residential and transportation locations of population. The decennial census and annual mid-year population estimates provide conventional residential population counts and a clear system of geographical referencing, although not always at the smallest areal unit level. We have used a national directory of georeferenced postal codes as the basis for allocation of these counts to exact locations and their population totals define the total population to be included in the model. Some categories of communal establishments, such as prisons and student halls of residence are available from independent sources and allow us to remove these sub-populations from the general residential base and directly assign exact counts and locations.

The second type of data are those relating to non-residential locations and in this category are locations of education, employment, health care, retail, leisure, etc. Recent data sources have begun to provide quite comprensive information on schools, colleges, universities and hospitals and the national statistical agency runs an annual survey of businesses which provides an employment dataset known as the Annual Business Inquiry (ABI) as a subscription service. The decennial census also provides information on places of employment, but these may be expected to become rapidly out of date as spatial patterns of employment will change much more quickly than residential patterns. Again, some of these such as ABI are not available for the lowest geographical units and some modelling is required. However, others such as educational and health care locations are identifiable at specific postal codes and can therefore be located with a high degree of precision. Further, most of these activity types can be assigned 'standard' time profiles based on survey data and published opening or service times. For example, working hours can be assigned to categories in the standard industrial classification based on the Quarterly Labour Force Survey (QFLS) while school opening times and university term-times are widely published on the web. For the purposes of a general model, single time profiles can be used for each activity type although we have adopted a data description model whereby unique time profiles can be assigned to individual centroids, if such data were available. Areas of influence, equating to catchment areas, can also be associated with these locations and these may be expressed as a variety of catchment area maps, travel-to-work distances or distance decay curves. Again, standard models may be applied to an entire category of centroids or individual data may be assigned where available. Census information contains a question on distance of travel to work which provides a useful starting point for the characterisation of standard areas of influence. A current area of deficiency is detailed counts for visitor numbers to residential and leisure facilities. These data are widely collected but considered to be commercially sensitive and our initial implementation therefore relies on modelled counts using employment numbers and reference data.

The third type of data are those relating to populations in the transport system. Here, most of the data available is a proxy for the actual population counts and tends to take the form of vehicle counts or timetable data associated with the movement of population through predefined networks (road, rail, etc.) Various data sources from Ordnance Survey, Britain's national mapping agency, have recently been published as open data, which allows us to identify key aspects of the transportation network and to associate these with typical traffic levels published by the Department for Transport. A very similar approach to UK transportation data has been adopted by Smith and Fairburn (2008).

Population 24/7 project progress

The Population 24/7 project has established a methodology for the redistribution of population from centroid locations into the cells of a regular geographical grid and described more fully elsewhere. This specific gridding algorithm could be replaced by alternatives while still retaining the integrity of the overall approach. The time-space implementation is based on the existence of an extensive database of spatial locations with associated time profiles indicating their population capacities. Centroids are divided into residential and non-residential locations and the transportation network is represented as a gridded model of population capacities generated at the resolution of the desired output model. For a specified target time, the time profile of each non-residential centroid in the study area is examined and a series of time-specific population estimates extracted. Populations are then redistributed from the residential centroids across the non-residential and transport locations so as to best meet the populations at each location. Again, a variety of algorithmic approaches would be possible, including complex spatial interaction modelling or microsimuation. In these initial implementations a simple distance-decay weighting is used whereby the total population summed across all residential locations is preserved within the model. The model is run separately on each age group identified in the data, allowing, pre-school, school, college, working and retired populations to be handled separately. Global adjustments may be made for visitors in or out of the study area. A substantial database of time-profiled centroid locations for England and Wales has now been prepared for the reference year 2006 and a range of tests and demonstrators prepared for the Solent region on the south coast of England. Figure 2(a-c) shows three time slices covering an area of approximately 80 x 40km at 200m resolution, including the cities of Southampton and Portsmouth.

The next phase of the work involves roll-out of these models to cover England and Wales (Scotland and Northern Ireland have some differently recorded data and are not covered in the current work) and consolidation and evaluation of the current models, using available data. The proposed methods are readily extensible in the light of more detailed data, providing that adjustments are made to any existing data sources to accommodate more specific sources, thereby avoiding double counting of population. Obvious areas for extension include modelling of non-road transportation from digital timetable data (Martin et al. 2008). An important area of future development will be to evaluate the possibility of drawing linked data dynamically from appropriate data services, which could in due course include the introduction of near-real time information on monitored population movements such as traffic flows. It seems probable that although the detailed algorithmic and data-specific details will continue to differ between countries and researchers (as between Landscan USA and the work described here), the most important conceptual issues will be consolidated and become more widely accepted, as the availability of more time-specific population data fosters new and more powerful application examples.

Acknowledgement and data sources

This research is supported by Economic and Social Research Council Award RES-062-23-0081. Employee data from the Annual Business Inquiry Service, National Online Manpower Information Service, licence NTC/ABI07-P3020. Office for National Statistics 2001 Census: Standard Area Statistics (England and Wales): ESRC Census Programme, Census Dissemination Unit, Mimas (University of Manchester). National Statistics Postcode Directory Data: Office for National Statistics, Postcode Directories: ESRC Census Programme, Census Geography Data Unit (UKBORDERS), EDINA (University of Edinburgh). Quarterly Labour Force Survey, Economic and Social Data Service, usage number 40023. Mastermap ITN layer: © Crown Copyright/database right 2009, an Ordnance Survey/EDINA supplied service.

References

Ahola, T., Virrantaus, K., Krisp, J. M. and Hunter, G. J. (2007) A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science* 21, 935-953

Bhaduri, N., Bright, E., Coleman, P. and Urban, M. L. (2007) LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *Geojournal* 69, 103–117

McPherson, T. N. and Brown, M.J. (2004) Estimating daytime and nighttime population distributions in U.S. cities for emergency response activities. Presented at Symposium on Planning, Nowcasting, and Forecasting in the Urban Zone, Seattle: American Meteorological Society http://ams.confex.com/ams/pdfpapers/74017.pdf

McPherson, T. N., Rush, J. F., Khalsa, H., Ivey, A. and Brown, M. J. (2006) A day-night population exchange model for better exposure and consequence management assessments. Presented at the Sixth Symposium on the Urban Environment, Atlanta: American Meteorological Society http://ams.confex.com/ams/pdfpapers/105209.pdf

Martin, D. (1989) Mapping population data from zone centroid locations Transactions of the Institute of British Geographers NS, 14, 90-97

Martin, D., Jordan, H. and Roderick, P. (2008) Taking the bus: incorporating public transport timetable data into health care accessibility modelling Environment and Planning A 40, 2510-2525

Martin, D., Cockings, S. and Leung, S. (2009) Population 24/7: building time-specific population grid models Paper presented at the European Forum for Geostatistics Conference, Statistics Netherlands, The Hague, 5-7 October

Shadbolt, N., Berners-Lee, T. and Hall, W. (2006) The Semantic Web Revisited. IEEE Intelligent Systems, 21 (3). pp. 96-101.

Sleeter, R. and Wood, N. (2006) Estimating daytime and nighttime population density for coastal communites in Oregon. Proceedings of the Annual Conference of the Urban and Regional Information Systems Association

Smith, G. and Fairburn, J. (2008) Updating *and improving the National Population Database to National Population Database 2*. Research Report 678, London: Health and Safety Executive

Figure 1(a): interpretation of population time-space in a conventional population map, with all population recorded at place of residence across all times.

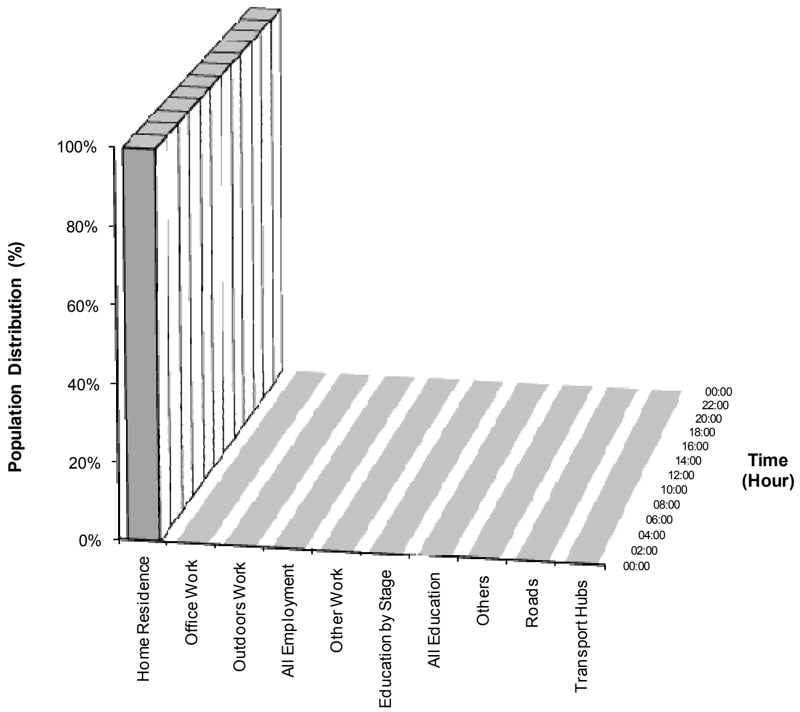


Figure 1(b): alternative interpretation of population time-space, whereby each non-residential population activity is associated with different time profiles.

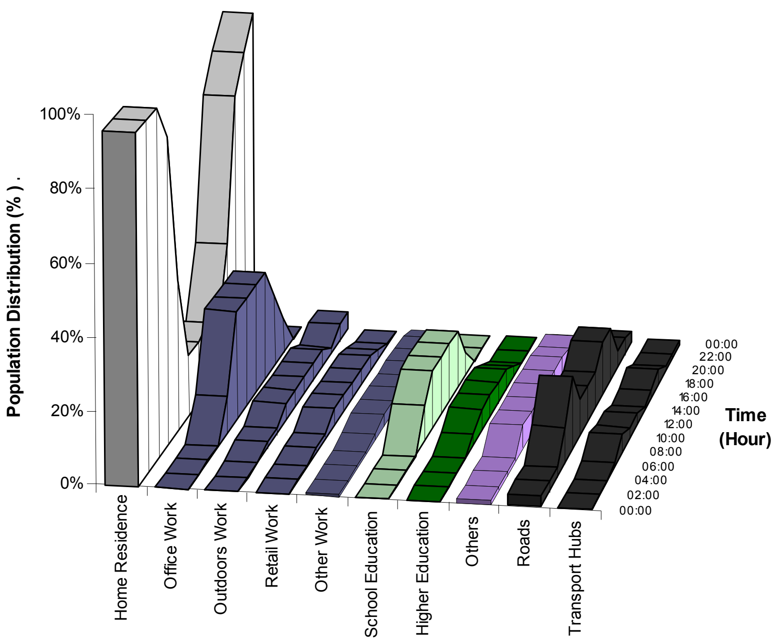
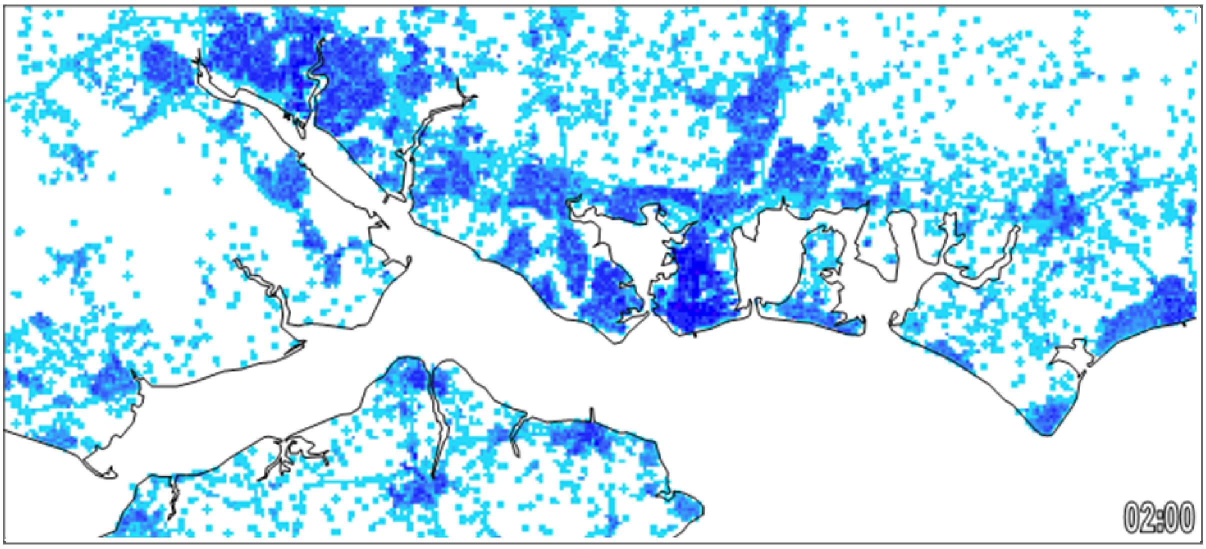
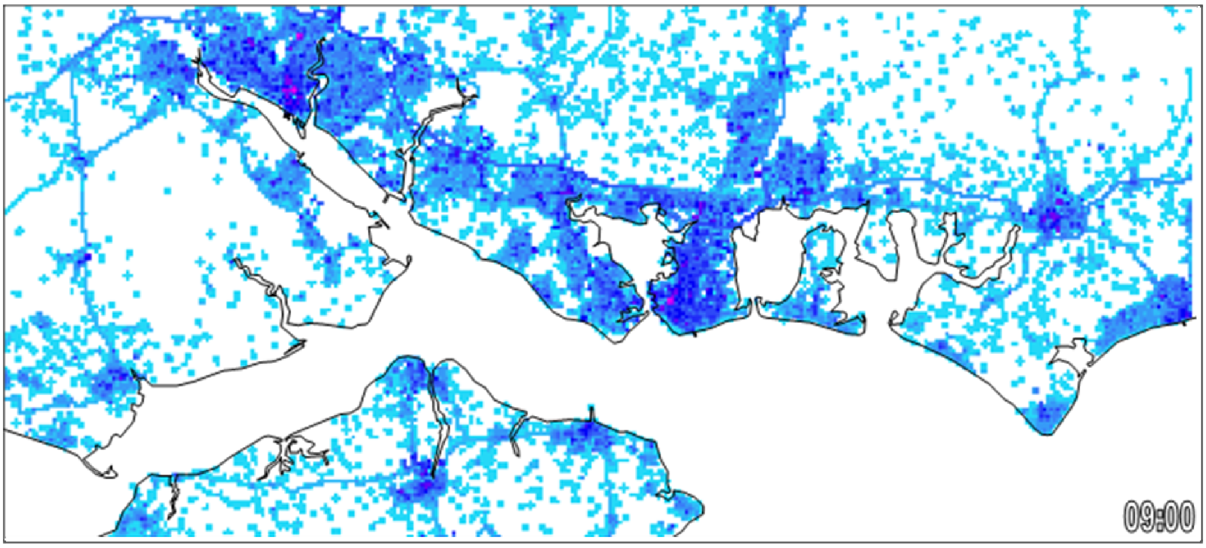
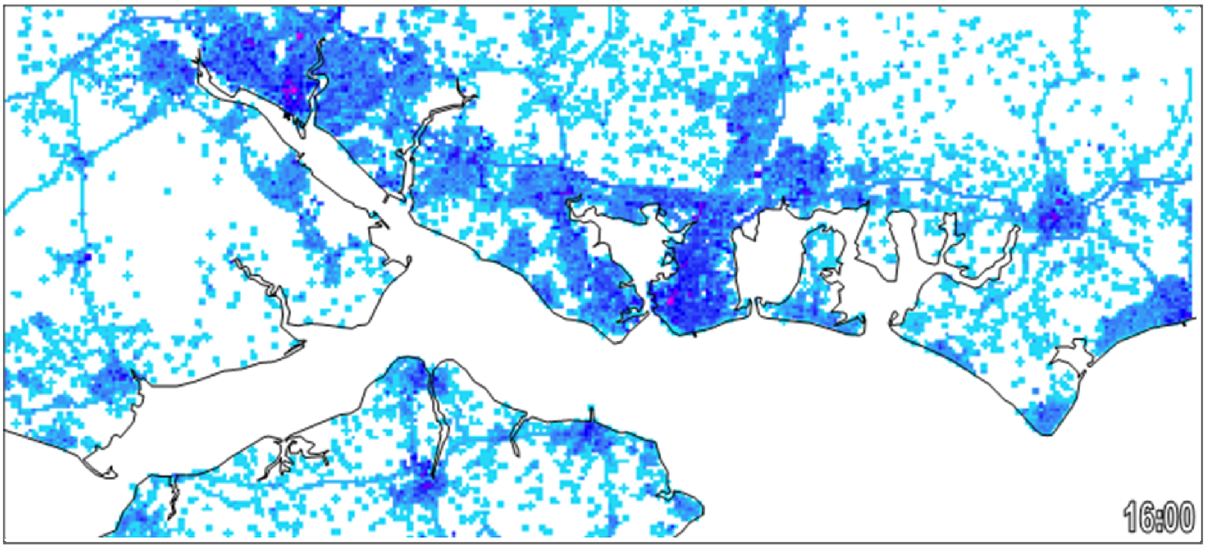


Figure 2(a-c): Weekday time sequence for the Solent Region, UK, showing temporal redistribution of population during the day (three sample time slices – 02:00 'night-time' and 09:00 and 18:00 subdivisions of 'day-time'






16:00