Environment and Planning A, 1991, volume 23, pages 1069–1075

# Techniques for modelling population-related raster databases

D Martin, I Bracken
Department of City and Regional Planning, University of Wales College of Cardiff, PO Box 906, Cardiff, CF1 3YN, Wales
Received 18 June 1990

**Abstract.** In this paper the refinement and application of a technique for the generation of surface models of population and related information are examined. With use of this technique the efficient generation of geographically extensive, high-resolution surfaces is described. The resulting database facilitates a range of improved spatial analyses. Some of these are more flexible means of accomplishing conventional tasks, such as the computation of incidence rates and the estimation of population for nonstandard areal units. Additionally, surface concepts are able to support innovative techniques, such as the identification and characterization of discrete settlements. Applications are described which demonstrate the range of possible analyses.

## Introduction

In a previous paper (Bracken and Martin, 1989) we introduced the underlying concepts of a general raster data model for population-related information. This showed that surface models are highly appropriate for the representation of socio-economic phenomena. These are in contrast to conventional (vector-based) approaches which, although widely used, suffer several weaknesses both at conceptual and at practical levels. These include the widely acknowledged modifiable areal unit problem (Openshaw, 1984), the difficulty of relating data from incompatible areal units (Flowerdew and Openshaw, 1987), the complexity of coordinate-based processing, and the more general problems of creating extensive vector databases. Examples of these general problems are the scale limitations of digitized boundaries, 'edge-matching' multiple-sheet map bases, and time and cost implications of digitization. However much care is taken in the construction of spatial databases, it is increasingly recognized that a degree of error is inevitable (Openshaw, 1989).

Since the original development of the surface modelling technique, a range of applications have been developed, which have demonstrated the important properties of the modelled database, and the wide variety of analytical operations which it is possible to perform. In this paper examples of these operations are described that will illustrate the use of this approach to more efficiently implement conventional forms of analysis. Equally important are the possibilities for novel forms of analysis, which cannot be performed with conventional data structures.

## Techniques for generation

The basic data-requirement of the surface-generation technique is a set of point locations to which counts of local population may be related. Examples of such points are census enumeration district (ED) centroids, unit postcode locations, or user-defined data points. No digitized boundary information is required. The underlying assumption of this technique is that the distribution of point locations is a summary of the distribution of the phenomena to be modelled. It must be noted that the locations of these summary points are not necessarily determined by precise methods, and a surface estimation algorithm is required which is appropriate to the 'fuzzy' nature of the input data.

The generation of such surface models has been more fully described elsewhere (Martin, 1989). In principle, the approach involves the redistribution of population (or population-derived, for example, unemployment) counts from the input data points into the cells of an output grid. A window is positioned over each data point in turn, the size of this window varying according to the local density of the points. This provides an estimate of the size of the areal unit represented by the current point. The count associated with the current point is then distributed into the cells falling within the window, according to weightings derived from a distance-decay function.

The general form of the assignment to each cell in the grid may be summarized as follows:

$$\hat{P}_i = \sum_{j=1}^{c} P_j W_{ij}, \tag{1}$$

where $\hat{P}_i$ is the estimated population of cell $i$, $P_j$ is the empirical population recorded at point $j$, $c$ is the total number of data points, and $W_{ij}$ is the unique weighting of cell $i$ with respect to point $j$. The practical implementation of this method to generate extensive surface databases requires a readily computed distance function which is appropriate to the nature of the data. A refinement of earlier work has been to use such a function (based on Cressman, 1959) which incorporates a finite window size, as shown by

$$W_{ij} = \begin{cases} \left( \dfrac{w_j^2 - d_{ij}^2}{w_j^2 + d_{ij}^2} \right)^{\alpha}, & \text{for } d_{ij} < w_j, \\ 0, & \text{for } d_{ij} \geqslant w_j \text{ with } \alpha > 1, \end{cases} \tag{2}$$

where $W_{ij}$ is the weighting associated with distance $d_{ij}$, and $w_j$ is the adjusted width of the window centred on point $j$. An increase in the value of $\alpha$ increases the weight given to cells close to the current point, thus controlling the form of the distribution. Although more sophisticated algorithms are available for density estimation (Silverman, 1986), more precise techniques are not considered appropriate given the spatial precision of these input data.

**Properties of the database**

The technique described is primarily a method for the generation of a population-related database. From experience with diverse applications of this approach, it is suggested that its most important features lie in the unique properties of the resulting socioeconomic models and in the analytical operations which they are able to support. These properties of the modelled database, which distinguishes it from conventional data structures, are considered here.

The cells of the raster database each contain a population estimate, and the grid thus represents a height matrix of density of the phenomenon represented by the count values. Many cells in this grid will have received a share of the counts from a number of different data points, but in a typical region the majority will remain unvisited and thus contain a population estimate of zero. In this way, the settlement geography is reconstructed in detail from the distribution of the point locations even within the zones used for data collection, and unpopulated regions are preserved. Discrete settlements are therefore readily identifiable from the pattern of populated cells, as illustrated in figure 1. A feature of the approach is a clearly understandable representation of the geography of the modelled variable. The total volume under the surface (the sum of values in every cell) will be the same as the sum of the

counts at the individual points, as these have simply been assigned to the surrounding cells of the grid. Moreover, the total population and all other counts for each discrete settlement will therefore also be correct, in terms of the input data.

For practical applications, it is important to understand the types of error which may be present in the modelled database. Errors may be of two, related types: first, 'locational' errors, which arise from the incorrect estimation of the spatial extent of populated areas; and, second, 'attribute' estimation errors, which arise from the assignment of incorrect population counts to individual cells.

In the context of the present work, locational errors can arise in two ways. A lack of detail in the input data may cause small settlements (within a single data-collection zone) to be missed. The overspecification or underspecification of window size will cause the population associated with a data point to be spread too far or not far enough. Errors in attribute values may arise from a lack of detail in the input data, or the specification of an inappropriate distance-decay function. Errors of this type will tend to be greatest in cells furthest from the data points, containing the lowest population estimates.

A general feature of raster databases is that error processes are more readily modelled than for zone-based data structures (Goodchild, 1989). As the approach models population distribution to a higher resolution than the input data, raw data are not available against which to evaluate errors in the modelled database. Nevertheless, locational errors can be assessed by comparing the populated cells of the model with digitized residential areas, and the population counts can be assessed by comparison with zones whose population is known. The extent of these errors has been examined in relation to models generated from UK census small-area statistics. A modelled population database with 200-metre cells was compared with digitized residential areas for a large region in South Wales, including Cardiff and Swansea. Of the eighty-thousand cells, 93.95% were correctly classified as populated or nonpopulated. If the distribution of error is mapped, it can be seen that these are primarily concentrated on the urban fringes and involve small population counts. In order to assess attribute errors, population estimates were obtained from the modelled database for sixty-one wards in the County of Avon, for which digitized boundaries were available. These estimated populations were compared with the actual census counts. The mean absolute error in individual ward populations was 5%, with a standard deviation of 4.7%, although the overall error in the modelled data is necessarily zero, as explained above.

**Figure 1.** A surface model of population density for the South West, showing all populated 200-metre cells (data source: 1981 Census).

Surface databases may be constructed for very extensive geographical regions, while maintaining a high resolution. The same database can therefore support analyses ranging from regional to local scales. This is illustrated by figure 2, which shows the population distribution for the City of Plymouth, extracted from the same data file as the South West region given in figure 1. This property of the raster structure can be contrasted with the representation of the same data by using vector means, which would involve a detailed boundary database, difficulties in handling the different resolutions required, and would necessitate complex coordinate processing. Despite the extensive coverage possible, population distribution is inherently sparse and databases held in this form can therefore be made compact by storing values for populated cells only. For example, in the $100 \times 100$ km (Ordnance Survey sheet TQ) region, which includes Greater London and a substantial part of the South East, less than 20% of cells contain any population count.

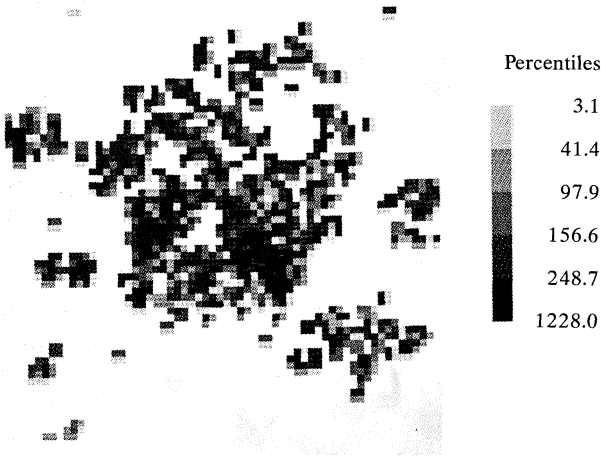


**Figure 2.** An enlarged portion of figure 1, showing population count for each cell by quintiles for the City of Plymouth (data source: 1981 Census).

**Techniques for analysis**

Current research has explored ways in which the modelled database can be exploited to perform traditional tasks more efficiently in spatial analysis, and to develop novel applications. The following examples illustrate some of these applications.

1 A frequent requirement in a policy context is the calculation and mapping of multivariate indices such as neighbourhood classifications or social stress indicators. Conventionally, these operations are performed on the attribute values of each zone, without reference to the explicitly spatial nature of the data. Indicator values are computed and mapped independently for individual data-collection zones, which are arbitrarily imposed on the underlying phenomena. In figure 3 a surface of percentage unemployment for the Greater London area is illustrated, which clearly retains the geography of the population and reveals local differences in unemployment incidence. Widely available cartographic modelling techniques (Berry, 1987; Tomlin, 1990) may be applied to these data models. Some of these techniques, such as neighbourhood-based functions, are particularly appropriate to the analysis of population-derived data.

2 Another common task, which is problematic if vector-based data are used, is the estimation of population or related counts for nonstandard areal units. An example

is the estimation of population for a community-nursing neighbourhood. A variety of options have been suggested for such estimation, but these require unrealistic assumptions of uniform population density, or ancillary data about settlement geography (Flowerdew and Green, 1989). A surface database provides a realistic model of both population distribution and geography, avoiding the need for such assumptions. Population estimates may be obtained for any ad hoc region by superimposing the boundary of that region on the surface. The high spatial resolution of the model described here is capable of supporting this type of estimation from conventional 'small areas' up to the regional scale.

3 Surface models provide an ideal base for the analysis of point-referenced 'event' data. An example is a postcoded mortality register, to which grid references can be assigned (Gatrell, 1989). Conventional approaches are hampered by several obstacles: the need to assign point-referenced events into zones of known population; the occurrence of small base populations in individual zones (Kennedy, 1989); and variations in the shape and area of zones. As a result, the derived incidence rates can be expressed in terms of the arbitrarily defined data collection zones only. Ideally, these rates should be continuous and related to the population density and geography. If the surface model is used, incidence rates may be calculated by a moving-window operation in which population and events in each window are counted and in which the rate is assigned to the central cell. In the calculation of a rate for each cell, adjustment of window size allows control over the base population and its areal extent.

4 In the examples above ways are described in which the raster surface models may be employed to improve three commonly required processes in spatial analysis. These are enabled by the reconstruction of detailed population density and distribution information. More significantly, these properties make possible powerful new analytical procedures, which can only be implemented in terms of the population surface concept. This final example draws attention to the implementation of one such technique.

One of the most striking features of the population model is its reconstruction of settlement geography. It is therefore possible to isolate discrete settlements by identifying groups of contiguous cells with nonzero population counts. This analytical procedure produces estimates of total population, area, and density for

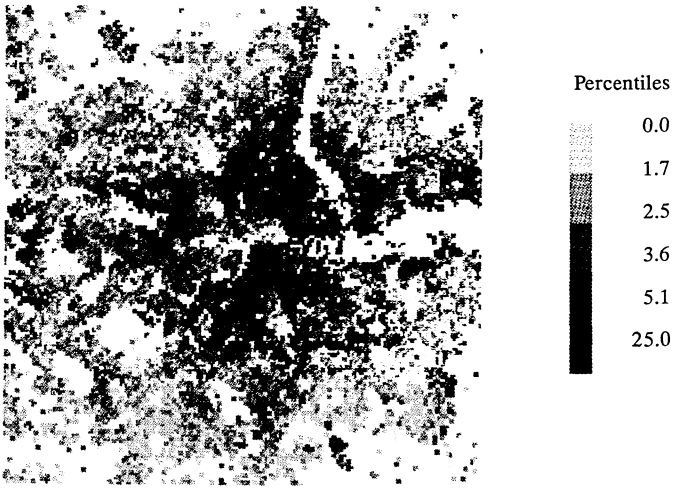


**Figure 3.** A surface model of percentage unemployment for the Greater London area by use of 200-metre cells (data source: 1981 Census).

distinct settlements. Contiguous populated areas identified in this way are entities in the modelled database, which correspond directly to the actual population which constitutes 'real' settlements such as Cardiff and Bristol. Thus, the towns and villages identifiable by eye in figure 1 are analytically separable in the data model. These entities are not separately identified in the input point data, and cannot be defined in terms of the arbitrary zones on which these data are based.

The idea of identifying regions of contiguous cells in this way may be extended to surfaces of other socioeconomic variables. For example, distinct neighbourhoods within settlements may be identified in terms of their unemployment or houseownership characteristics. This approach offers an alternative to conventional methods for neighbourhood classification and functional region definition, which is essentially 'data-derived'. An implementation of this technique has shown that the modelled database is suitable for settlement-based analysis, without the need for additional settlement-based data. Owing to the volume-preserving characteristic of the surface-construction technique, all attribute values relating to a discrete settlement will be correct. It may be desired to define as 'populated' only those cells whose population density falls above a certain threshold, or to define as 'contiguous' all populated cells which fall within a certain radius. These criteria are parameters of the settlement-identification algorithm, and are easily modified.

## Conclusion

In this paper, we have described refinements to our approach to the generation of surface models of population-related data. The application examples illustrate a range of different operations, made possible by the basic properties of this type of model. Three points should be noted. First, the surface model represents the underlying geography of the phenomenon independently of any zonal considerations. Second, because raster technology is used, these models can be stored and manipulated efficiently, avoiding many inherent complexities of large vector databases. Third, these models provide a valid basis for an extended range of fundamental spatial techniques.

In the context of current interest in geographic information systems (GIS), the techniques described here should not be seen merely as another 'GIS solution' to population mapping. Although the raster population databases are suitable for further manipulation in a GIS, the modelling of population and its characteristics presented here results from a fundamental reconsideration of the spatial nature of these phenomena. The unique properties of this form of representation offer a rich basis for analysis to those who are prepared to break away from conventional data structures. The principles identified have important implications for a wide range of policy and geodemographic applications which are currently severely restricted by the use of existing data models.

## References

Berry J K, 1987, "Fundamental operations in computer-assisted map analysis" *International Journal of Geographic Information Systems* **1** 119-136

Bracken I, Martin D, 1989, "The generation of spatial population distributions from census centroid data" *Environment and Planning A* **21** 537-543

Cressman G P, 1959, "An operational objective analysis system" *Monthly Weather Review* **87**(10) 367-374

Flowerdew R, Green M, 1989, "Statistical methods for inference between incompatible zonal systems", in *Accuracy of Spatial Databases* Eds M Goodchild, S Gopal (Taylor and Francis, London) pp 239-248

Flowerdew R, Openshaw S, 1987, "A review of the problems of transferring data from one set of areal units to another incompatible set", RR-4, Northern Regional Research Laboratory, University of Newcastle upon Tyne, Newcastle upon Tyne, England

Gatrell A C, 1989, "On the spatial representation and accuracy of address-based data in the United Kingdom" *International Journal of GIS* **3** 335-348

Goodchild M F, 1989, "Modeling error in objects and fields", in *Accuracy of Spatial Databases* Eds M Goodchild, S Gopal (Taylor and Francis, London) pp 107-114

Kennedy S, 1989, "The small number problem and the accuracy of spatial databases", in *Accuracy of Spatial Databases* Eds M Goodchild, S Gopal (Taylor and Francis, London) pp 187-196

Martin D, 1989, "Mapping population data from zone centroid locations" *Transactions of the Institute of British Geographers: New Series* **14**(1) 90-97

Openshaw S, 1984 *The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography Number 38* (Geo Books, Norwich)

Openshaw S, 1989, "Learning to live with error in spatial databases", in *Accuracy of Spatial Databases* Eds M Goodchild, S Gopal (Taylor and Francis, London) pp 263-275

Silverman B W, 1986 *Density Estimation for Statistics and Data Analysis* (Routledge, Chapman and Hall, Andover, Hants)

Tomlin C D, 1990 *Geographic Information Systems and Cartographic Modelling* (Prentice-Hall, Englewood Cliffs, NJ)