Ann Reg Sci (1992) 26:67–78

The Annals of
Regional Science

Developments in areal interpolation methods and GIS

Robin Flowerdew[1] **and Mick Green**[2]

[1] North West Regional Research Laboratory
[2] Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YB, UK

Abstract. This paper is a review and extension of the authors' research project on areal interpolation. It is concerned with problems arising when a region is divided into different sets of zones for different purposes, and data available for one set of zones (source zones) are needed for a different set (target zones). Standard approaches are based on the assumption that source zone data are evenly distributed within each zone, but our approach allows additional information about the target zones to be taken into account so that more accurate target zone estimates can be derived. The method used is based on the EM algorithm. Most of the work reported so far (e.g. Flowerdew and Green 1989) has been concerned with count data whose distribution can be modelled using a Poisson assumption. Such data are frequently encountered in censuses and surveys. Other types of data are more appropriately regarded as having continuous distributions. This paper is primarily concerned with areal interpolation of normally distributed data. A method is developed suitable for such data and is applied to house price data for Preston, Lancashire, starting with mean house prices in 1990 for local government wards and estimating mean house prices for postcode sectors.

1. Introduction

A key theme in geographical information systems is the integration of different data sets, and one of the key problems in data integration is the diversity of areal units in use for different purposes. Frequently it is desirable to compare two or more sets of regional data which are available for different zonal systems. A common example concerns administrative zones, for which data are collected by national censuses, government departments and by the administrative units themselves, and postal zones, which are the easiest way to aggregate data for most commercial purposes, such as records of sales or client contacts. In order to compare data for administrative and postal zones, or any other pair of incompatible zonal systems, methods must be found for estimating what one set of data would

be like if it were available for the other zonal system. This problem is referred to as the areal interpolation problem (Goodchild and Lam 1980).

With colleagues at the North West Regional Research Laboratory, we have been engaged in a research project to develop improved methods for areal interpolation, in particular by using other available information to help improve our estimates. This paper develops and applies a method suitable for normally distributed data, and should be regarded as a companion to earlier work on areal interpolation for count data (Flowerdew and Green 1989, 1991; Flowerdew et al. 1991).

2. A review of the problem

Although some, perhaps most, researchers who have encountered the areal interpolation problem have given up, regarding it as an insuperable obstacle, there have been a number of attempts to overcome it. One approach, best exemplified by the work of Tobler (1979), has been to regard data for discrete zones as manifestations of an underlying continuous density surface. For data on population, for example, Tobler postulates the existence of a continuous population density surface over the entire study area, regarding the populations of specific zones as the integrals of that surface over the region defined by the zonal boundaries. This surface must have what Tobler calls the pycnophylactic property, in other words it must yield the actually observed data for the original set of zones. If a surface can be specified given data for one set of zones, it is possible to calculate the population for any other set of zones, however defined, by integrating the surface over the new zones. This is an effective procedure where it is sensible to regard the data being modelled as varying smoothly over space. For many types of data of social and economic interest, however, there are abrupt discontinuities in spatial distribution, making this approach and its variants less effective.

The main alternative approach is based on the assumption that data are likely to be uniformly distributed within the zones for which they are available. If we refer to the zonal system for which we have data as source zones, and the system for which we want to estimate values as target zones, then data for target zones can be estimated as a weighted average (or weighted sum) of data for the source zones with which they intersect. Weighting is in proportion to the area of the zones of intersection. This method will be referred to here as 'areal weighting'; a more precise account is given below.

In their account of the areal weighting method, Goodchild and Lam (1980) distinguish between extensive and intensive variables. If a zone is divided into a set of subzones, then a variable is described as extensive if its value for a zone is the sum of its values for the subzones; it is intensive if the value for the zone is a weighted average of the values for the subzones. Some variables, like relative relief, are neither extensive nor intensive. For an extensive variable, therefore,

$$Y_t = \sum_s Y_{st} ,$$

where t is a target zone which is divided into a set of zones of intersection st by the boundaries of the source zones. For an intensive variable,

$$Y_t = \sum_s \frac{Y_{st} A_{st}}{A_t} ,$$

where A_{st} is the area of intersection zone st.

The areal weighting method also requires a method of relating source zone values to intersection zone values. In the case of intensive variables, it would normally be assumed simply that:

$$Y_{st} = Y_s$$

but in the case of an extensive variable it is assumed that the share of the total value accruing to each intersection zone is proportional to the area of the intersection zone:

$$Y_{st} = \frac{Y_s A_{st}}{A_s} .$$

The areal weighting method is thus based on the assumption that the values of the variable of interest are evenly distributed within the source zone. It seems to be the best method when there is no information available to suggest the distribution might be uneven. In practice, however, there is often plenty of information available which would lead us to expect this distribution to be far from even. Population distribution, for example, is likely to be strongly influenced by slope and elevation, and more directly by the distribution of land use types. A method for areal interpolation is needed which goes beyond simple areal weighting to allow researchers to use their geographical knowledge for improving their estimates of likely target zone values.

In earlier papers (Flowerdew and Green 1989, 1991; Flowerdew et al. 1991) we have developed an approach to 'intelligent' areal interpolation in which other information is taken into account to give better estimates than are possible with areal interpolation. This information may be available for the target zones, or for a third set of zones which may be called control zones. Essentially the method works through establishing a regression relationship between the variable of interest and one or more ancillary variables (the 'other information' above). Once this relationship has been established, it can be used, along with area, to estimate values for the variable of interest for the target zones.

The regression relationship mentioned above must be estimated in a manner appropriate for the probability distribution the variable of interest is assumed to have. We have therefore been developing a series of methods appropriate for different distributions of this variable. Most of the early work was on extensive variables, especially count data for which the Poisson distribution is applicable (Flowerdew and Green 1989, 1991). More recently, work has been extended to deal with binomial and other distributions (Flowerdew et al. 1991; Green 1990). The approach taken can be regarded as a special case of the EM algorithm (Dempster et al. 1977). In this paper, we develop and apply an algorithm suitable for continuous variables. We assume that these variables are means of a set of observa-

tions in each source zone. In the example, these are the means of prices for houses on the market in each zone.

3. Algorithm for continuous variables

Consider a study region divided into source zones indexed by s. For source zone s, we have n_s observations on a continuous variable with mean y_s. We wish to interpolate these means onto a set of target zones indexed by t.

Consider the intersection of source zone s and target zone t. Let this have area A_{st} and assume that n_{st} of the observations fall in this intersection zone. In practice n_{st} will seldom be known but will, itself, have to be interpolated from n_s. In what follows we may assume that n_{st} has been obtained by areal weighting:

$$n_{st} = \frac{A_{st} n_s}{A_s} .$$

More sophisticated interpolation of the n_{st} would increase the efficiency of the interpolation of the continuous variable Y but this increase is likely to be small as quite large changes in n_{st} tend to produce only small changes in the interpolated values of Y. We will consider n_{st} as known.

Let y_{st} be the mean of the n_{st} values in the intersection zone st, and further assume that

$$y_{st} \sim N(\mu_{st}, \sigma^2/n_{st}) .$$

Now

$$y_s = \sum_t n_{st} y_{st}/n_s$$

and

$$\begin{bmatrix} y_{st} \\ y_s \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{st} \\ \mu_s \end{bmatrix}, \begin{bmatrix} \sigma^2/n_{st} & \sigma^2/n_s \\ \sigma^2/n_s & \sigma^2/n_s \end{bmatrix} \right) .$$

Clearly if the y_{st} were known we would obtain y_t, the mean for target zone t as:

$$y_t = \sum_s n_{st} y_{st}/n_t ,$$

where

$$n_t = \sum_s n_{st} .$$

The simplest method would take $y_{st} = y_s$ to give the areal weighting solution. Here we wish to allow the possibility of using ancillary information on the target zones to improve on the areal weighting method. Adopting the EM algorithm approach we would then have the following scheme:

E-step:

$$\hat{y}_{st} = E(y_{st}|y_s) = \mu_{st} + (y_s - \mu_s) \ ,$$

where

$$\mu_s = \sum_t n_{st}\mu_{st}/n_s \ .$$

Thus the pycnophylactic property is satisfied by adjusting the μ_{st} by adding a constant such that they have a weighted mean equal to the observed mean y_s.

M-step:
Treat $\hat{y}_{st}$ as a sample of independent observations with distribution:

$$\hat{y}_{st} \sim N(\mu_{st}, \sigma^2/n_{st})$$

and fit a model for the μ_{st}

$$\mu(s,t) = X\beta$$

by weighted least squares.

These steps are repeated until convergence and the final step is to obtain the interpolated values y_t as the weighted means of the $\hat{y}_{st}$ from the *E*-step, i.e.

$$y_t = \sum_s n_{st}\hat{y}_{st}/n_t \ .$$

In practice it is often found that it can take many iterations for this algorithm to converge. Thus although this approach is relatively simple it may not be computationally efficient. However, since we are dealing with linear models we can use a non-iterative scheme as follows. Since ancillary information is defined for target zones only we have

$$\mu_{st} = \mu_t \ , \quad \text{all } s$$

and

$$\mu(t) = X^{(t)}\beta \ ,$$

where $\mu^{(t)}$ is the vector of means for target zones and $X^{(t)}$ the corresponding design matrix of ancillary information.

Since

$$\mu_s = \sum_t p_{st}\mu_{st} \ ,$$

where

$$p_{st} = n_{st}/n_s$$

then

$$\mu^{(s)} = PX^{(t)}\beta \ ,$$

where

$$P = [p_{st}] \ .$$

Thus we can estimate β directly using data y_s and design matrix $X^{(s)} = PX^{(t)}$ by weighted least squares with weights n_s. The final step is to perform the E-step on the fitted values $\hat{\mu}_{st} = \hat{\mu}_t =$ computed from $X^{(t)}\hat{\beta}$ and form their weighted means to produce the interpolated values y_t.

4. Application

House price data were collected for the borough of Preston in Lancashire during January–March 1990 by sampling property advertisements in local newspapers. The sample used in this exercise included 759 properties, each of which was assigned to one of Preston's 19 wards on the basis of address. Sharoe Green ward contained 170 of these properties, Cadley contained 70, Greyfriars 69, and the others smaller numbers, with under 20 in Preston Rural East (3), Preston Rural

Table 1. Mean house prices in Preston wards (source zones), January–March 1990

Ward	Population	Area [ha]	Mean house price [£]	Number of cases
Preston Rural East	5827	5779	140667	3
Preston Rural West	9218	4597	145825	6
Greyfriars	6526	270	91206	69
Sharoe Green	9402	800	81632	170
Brookfield	7262	176	48176	21
Ribbleton	6993	542	54810	20
Deepdale	6466	124	47823	30
Central	4320	168	40700	6
Avenham	6015	149	48569	29
Fishwick	7381	322	42031	39
Park	7426	134	34762	25
Moorbrook	5269	92	38200	25
Cadley	6681	193	83986	70
Tulketh	6538	101	45262	58
Ingol	4564	158	60294	54
Larches	6801	177	52973	26
St Matthew's	6072	112	37671	26
Ashton	6105	276	63275	54
St John's	5551	102	35991	28

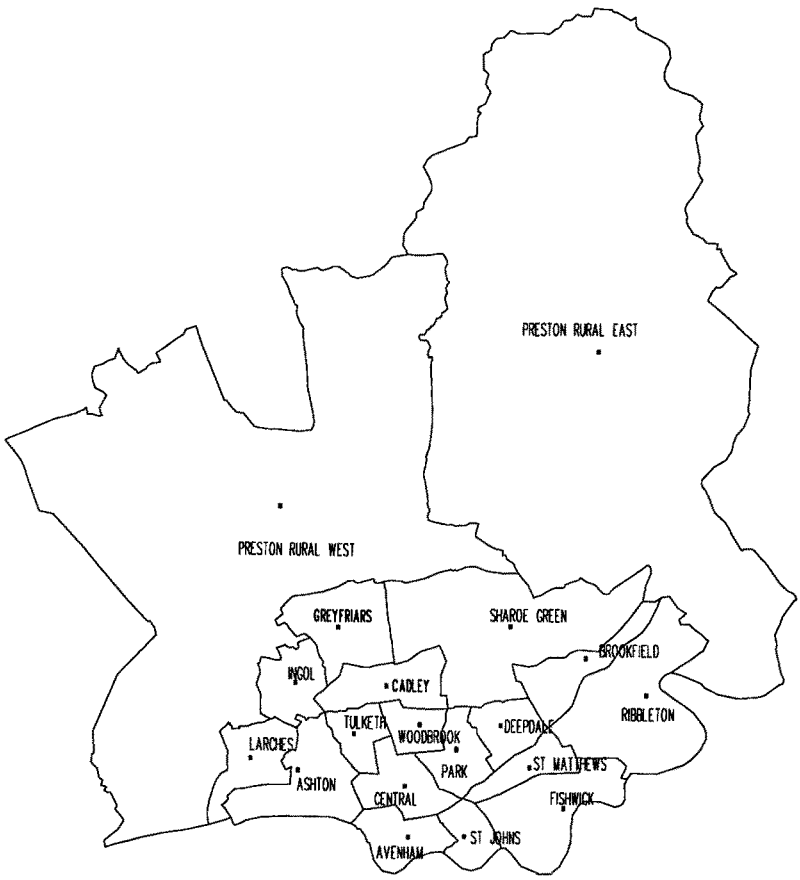

Fig. 1. Preston wards

West (6) and Central (6). The mean house prices for the wards ranged from £ 145,825 in Preston Rural West to £ 34,762 in Park. Table 1 shows the area, population and mean house price for the wards, and Fig. 1 shows their locations.

Wards were taken as source zones, with postcode sectors being the target zones. A set of 20 postcode sectors covers the borough of Preston, 6 of them also including land ouside the borough (Fig. 2). In some places, such as the River Ribble and some major roads, ward boundaries and postcode sector boundaries are identical, but in general they are completely independent. Ward and postcode sector coverages for Preston were overlaid using ARC/INFO to determine the zones of intersection and their area. There were many very small zones, some of which may have been due to differences in how the same line was digitised in each of the two coverages. The smallest ones were disregarded up to an area of 9 ha. This left 73 zones of intersection (Fig. 3).

Ancillary information about postcode sectors was taken from the June 1986 version of the Central Postcode Directory (CPD), the most recent version available to the British academic community through the ESRC Data Archive (Man-

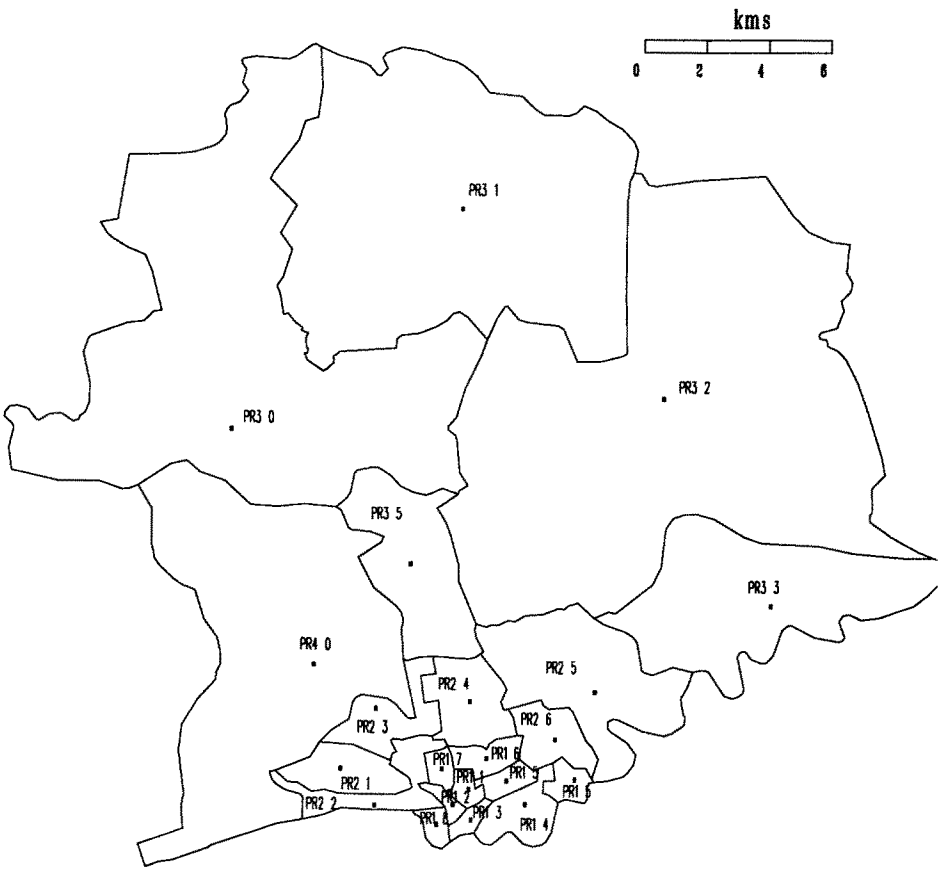

Fig. 2. Preston postcode sectors

chester Computing Centre 1990). This source lists all the unit postcodes, and gives a limited amount of information about them: the dates when they were introduced and terminated, and whether they are 'large user' or 'small user' postcodes. Most postcodes in the Preston area have been in existence since the establishment of the CPD in 1981 and are still in use. However, new postcodes are needed in areas of new residential development. Postcodes may go out of use in areas undergoing decline, or in areas where the Post Office has reorganised the system. Large user postcodes usually refer to single establishments which generate a large quantity of mail, such as commercial and major public institutions, while groups of private houses are small users.

The CPD can therefore generate a number of potential ancillary variables, including the number of individual postcodes in a postcode sector, the proportion of postcodes belonging to large users, the proportion of new postcodes and the proportion of obsolete postcodes. Each of these might be expected to reflect aspects of the geography of Preston. Although none of them are directly linked to house prices, they may be correlated with factors which do affect house prices.

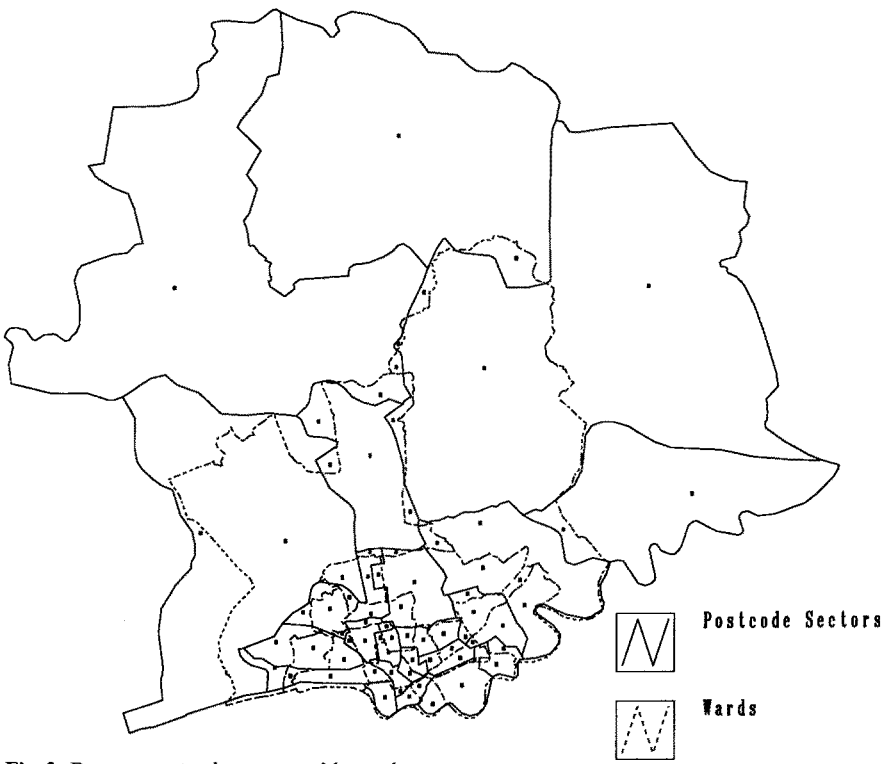

Fig. 3. Preston postcode sectors with wards

The most useful one may be the proportion of large user postcodes, which are likely to be most common in central or commercially developed areas. To some extent, these areas are likely to have less valuable housing stock, so a negative relationship to house prices is anticipated. For the purposes of the paper, this 'large user' variable was used as an ancillary variable in two forms. A binary variable was created separating those postcode sectors with more than 10% of postcodes being large users from those with under 10% large users; and a continuous variable was created representing the proportion of postcodes belonging to large users. The latter ranges from 0.03 in postcode sectors PR2 6 and PR3 0 to 0.60 in postcode sector PR1 2.

Two models were fitted, according to whether the ancillary variable was treated as binary or continuous. In all cases, the first stage was to estimate how many sample points were located in each intersection zone. This was done, as suggested above, by an areal weighting method. The relevant output from the models comprises the set of estimated house prices for the target zones, the relationship between house price and the ancillary variable, and the goodness of fit of the estimation procedure. The EM and direct estimation methods give identical answers, but the EM method takes longer to reach the solutions – in the binary case, 15 iterations were needed before target zone prices converged to the nearest pound.

5. Results

Table 2 shows the interpolated values for the 20 postcode sectors, together with the ancillary data. It can be seen that there are substantial differences between the areal weighting method and the estimates made using ancillary information. There are also major differences between the estimates reached according to the form of the ancillary variable. We do not have information about the true values for the target zones, so it is not possible to evaluate the success of the methods directly. However, the estimated values using the ancillary variable in its continuous form seem unrealistically low for sectors PR1 1, PR1 2 and PR1 3: there are few houses anywhere in England with prices as low as £ 20,000.

The goodness of fit statistics produced in fitting the models give further information for evaluating them. The house price data are assumed to be normally distributed; a model can therefore be evaluated in terms of the error sum of squares. The total sum of squares for house price values defined over the source zones is 335,600,000,000 (the numbers are very large because prices are expressed in pounds, rather than thousands of pounds, and because they are weighted by the number of houses in each ward).

When the large user postcode variable is incorporated in the model in binary form, the error sum of squares is reduced to 143,600,000,000. This yields a coefficient of determination (R^2) of 0.572. The model suggests that postcode sectors with high proportions of large user postcodes should have mean house prices of £ 44,056 and sectors with low proportions should have mean house prices of

Table 2. Interpolated house prices (£) for Preston postcode sectors (target zones)

Postcode sector	Area [ha]	% large users	Areal weighting	Binary variable	Continuous variable
PR1 1	82	44	37326	35513	14578
PR1 2	62	60	41896	44592	21816
PR1 3	68	57	43554	42489	20425
PR1 4	349	23	40284	38682	41322
PR1 5	258	17	43164	37784	43611
PR1 6	175	16	41216	38777	44093
PR1 7	87	13	40167	40762	51096
PR1 8	99	41	48569	48569	55580
PR2 1	413	4	93164	73882	69150
PR2 2	469	27	73465	44758	47295
PR2 3	453	5	91700	76745	76857
PR2 4	567	9	84618	96625	87439
PR2 5	1725	22	110793	52784	66205
PR2 6	364	3	54406	75946	67160
PR3 0	183	3	145013	147321	149497
PR3 1	223	9	140667	149021	140570
PR3 2	4437	6	140781	148923	144435
PR3 3	183	11	140667	106685	138045
PR3 5	1162	4	145640	147260	148283
PR4 0	2912	6	143509	97002	96692

£ 86,392. The estimated mean values in Table 2 are produced by adjusting these figures to meet the pycnophylactic constraint – in other words, ensuring that the observed source zone means are preserved.

It might perhaps have been expected that the crudeness of measurement of the large unit postcode variable would reduce the goodness of fit of the model, and hence that using the proportion of large user postcodes as a continuous variable would improve the model. In fact, however, this model had an error sum of squares of 229,300,000,000, with an R^2 value of only 0.317. The estimating equation involved a constant term of 84,560 and a coefficient of −126,260; in other words, estimated mean house price declined by £ 126,260 for a unit increase in the proportion of large user postcodes – more realistically, it declined by £ 1,263 for a unit increase in the percentage of large user postcodes. As noted above, this resulted in unrealistically low values for those postcode sectors with a high proportion of large user postcodes.

Examination of plots indicates that the relationship of house price to the ancillary variable levels off for higher values and use of a linear relationship can severely underestimate house price for zones with a high proportion of large user postcodes. This is confirmed by incorporating a quadratic term in the relationship which improves the fit ($R^2 = 0.365$) and lessens the underestimation problem. Using a logarithmic transformation gives a further small improvement. This highlights the general point that when using continuous ancillary variables careful choice of the form of the relationship may be necessary.

6. Conclusion

The poorer fit obtained when the ancillary variable was used in continuous rather than binary form is surprising, and suggests the need to experiment with other functional forms. It may also be worthwhile experimenting with other ancillary variables, either singly or in combination. Some other variables may be obtainable through the use of the CPD and digitised postcode sector boundaries, such as the density of small user postcodes, which might be expected to relate to house prices. In many practical applications, variables more directly related to house prices may be available, including of course things like client addresses, provided they are postcoded.

Nevertheless, the improvement in fit of the models discussed in this paper does suggest that taking advantage of ancillary data available for target zones can do much to improve areal interpolation. Without knowledge of mean house prices for postcode sectors, we are unable to quantify the improvement.

As stated in the introduction, this work is part of a more general project designed to develop better methods of areal interpolation. Unfortunately, different types of data require slightly different methods, but we hope that we have given some indication of how ancillary data can be used to improve areal interpolation for normally distributed data. Work is in progress to develop comparable methods where those described here are not applicable.

Acknowledgements. This research was supported by the British Economic and Social Research Council (grant R000231373). We are indebted to the North West Regional Research Laboratory for use of their GIS facilities, to John Denmead for use of advanced INGRES facilities he has developed, and to Evangelos Kehris and Isobel Naumann for research assistance. We also thank Susan Lucas for use of her data, collected as part of a postgraduate research project at the Department of Geography, Lancaster University.

References

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Flowerdew R, Green M (1989) Statistical methods for inference between incompatible zonal systems. In: Goodchild M, Gopal S (eds) Accuracy of spatial databases. Taylor & Francis, London, pp 239–247

Flowerdew R, Green M (1991) Data integration: statistical methods for transferring data between zonal systems. In: Masser I, Blakemore M (eds) Handling geographical information. Longman, London, pp 38–54

Flowerdew R, Green M, Kehris E (1991) Using areal interpolation methods in geographical information systems. Papers Reg Sci 70:303–315

Goodchild M, Lam NS-N (1980) Areal interpolation: a variant of the traditional spatial problem. Geo-Process 1:297–312

Green M (1990) Statistical models for areal interpolation. In: Harts J, Ottens HFL, Scholten HJ (eds) EGIS '90 Proceedings, vol 1. EGIS Foundation, Utrecht, pp 392–399

Manchester Computing Centre (1990) Post Office Central Postcode Directory (POSTZON file). CMS 628, Manchester Computing Centre

Tobler WR (1979) Smooth pycnophylactic interpolation for geographical regions. J Am Statistical Assoc 74:519–530