An evaluation of error for piecewise areal interpolation: a research note

D. PULLAR
The University of Queensland, Brisbane 4072, Australia.
Internet: www.gpa.uq.edu.au\criss\tool\ludm

Areal aggregation over discrete zones, or piecewise polygonal representations, is a common problem in geographical analysis. Various solutions to the problem have been proposed, but in many cases their accuracy assessment is judged on arbitrary cases and no systematic evaluation of potential process errors. This research note analyses errors based upon patterns in the data, specifically using entropy to measure spatial segregation of mapped attributes. A general result is found that the mean relative error increases as interpolated areas have greater heterogeneity. This result was tested with synthetic data and actual census data. We also describe an implementation of piecewise areal aggregation using the EM algorithm in GIS.

1. Introduction

A common need in working with socio-economic information is transferring data between incompatible spatial units. It arises from data being reported for a set of collection zones but applications need the data for different spatial units. For instance a census collection zone has household counts for urban regions, but it is desired to know the break down of this variable for traffic analysis zones. The problem of transferring data between incompatible systems is known as areal interpolation (Lam, 1983). The set of spatial units with known variables is termed the source zone, and the other set of superimposed spatial units with unknown variables is termed the target zone. Transferring data is made possible by determining the area density for target zones, for example the number of households per hectare. Goodchild et al. (1993) describes a framework based upon the assumptions made in the areal interpolation process. One assumption is the *pycnophylactic* property which stipulates that the data value for source zones equals the sum of the constituent target areas multiplied by their estimated densities. The term was proposed by Tobler (1979) who also stipulated other conditions on the smoothness properties between zonal boundaries to model gradients of change for continuous surface representations. The main interest in this research note is to apply a piecewise approximation on the areal densities for source and target zones. Area weighted proportioning, a procedure commonly found in GIS, is one form of piecewise areal interpolation that assumes homogeneous density for source zones. For many problems dealing with socio-economic information this is an invalid assumption, and leads to the commonly described modifiable area unit problem (Openshaw and Taylor, 1981). Assuming variability in areal density for source zones; there are two assumptions that may be made on areal density for target zones. Both options make use of ancillary information for target zones. In the first option we have knowledge of the distribution of areal

densities for map classes. Dasymetric mapping uses this assumption where area density is distributed between binary classes with zero density and the remainder assigned according to a value class by area weighted proportioning. This may be further generalized to assign the variable according to predefined ratios between the classes (Mennis, 2003). A second option we believe is more common practice; the ancillary information uses a set of ancillary zones with constant area densities. These superimposed zones in combination with the source zones with known values define a set of conditioned equations that may be solved by various methods. The set of zones with ancillary information are referred to as control zones, and once their areal density is solved they may be trivially combined to compute values for any arbitrary set of target zones. See figure 1. We pursue this approach as it is very flexible (Goodchild et al., 1993) and may be applied to a large class of applications where controls zones are given by land use classes obtained from remote sensing classification.

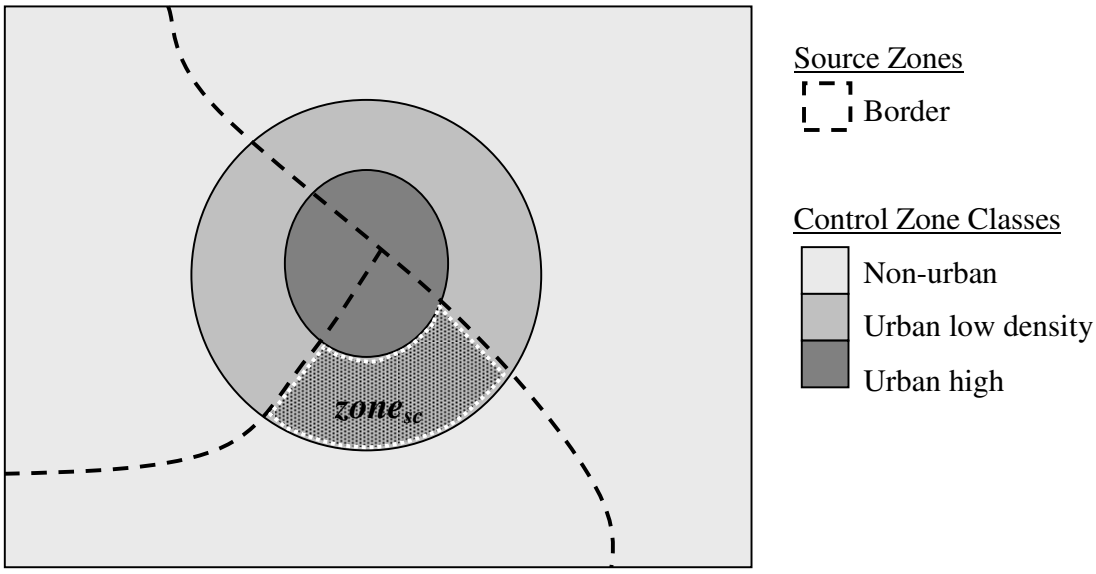


Figure 1. Source zones, control (target) zones and intersected sub-zones.

The choice of an areal interpolation technique should be guided by the assumptions underlying the approach and its accuracy. A number of studies have compared the results of areal interpolation to assess errors. The difficulty is that the error analysis only reports the consistency with which a procedure derives a result. Only in a few cases are independent geographical data available to assess the actual error in deriving areal density estimates for target zones. Although this provides just a single comparison, it is still a good reality check. More systematic approaches for error analysis require survey sampling or simulating zonal systems. Fisher and Langsford (1995) investigate the error distribution by randomly aggregating a set of elemental zones with known population to hypothetical zones. Other researchers have similarly generated synthetic zones for error analysis purposes (Gregory and Ell, 2005). These studies demonstrate that approaches which enforce the pycnophylactic property and make use of ancillary information give the most accurate results. Fisher and Langsford (1995) also make the observation that accuracy decreases as the number of target zones increase. Sadahiro (2000) investigated how the relative size of source zones to target zones affect errors; concluding that a

more accurate result is obtained if source zones are small, and hence are more likely to be nested in target zones. In this research note we explore a further characteristic for the spatial configuration of target zones superimposed on source zones. This follows from there being zero error when the source and target zones are equivalent, and error potentially increases for highly segregated spatial configurations. A statistic that may be used to measure the degree of segregation in maps is entropy. Entropy is used as a measure of the spatial distribution of mapped attributes in geostatistics (Journel and Deutsch, 1993) and in landscape ecology to indicate diversity (Forman, 1995). Can entropy provide additional insight into error analysis for areal interpolation?

The aim of this research note is to explore to what extent entropy characterizes errors. The next section describes the method of areal interpolation we used based upon Flowerdew and Green (1991), and its implementation as a geoprocessing tool in GIS. We then describe a systematic means of defining spatial configurations of source and target zones to assess error against a measure of spatial entropy. We compare errors from the EM algorithm and simple area weighted proportioning to show that as entropy increases so does potential error. A further example is given to disaggregate household data between to a zoning scheme with known data to provide a single comparison to a real world situation.

2. GIS Implementation of piecewise areal interpolation

The method used for areal interpolation was proposed by Flowerdew and Green (1991). The method uses the expectation-maximization (EM) algorithm (Dempster et al., 1977) which is an iterative method to estimate unknown parameters for the density given the data count we know for source zones. We choose this method as it is accurate and finds the best possible solution that satisfies the pycnophylactic property. In figure 1 we are given a set of source zones with counts y_s, and the goal is to obtain disaggregate values y_{st} for the intersected target zones given there are constant but unknown densities λ_c for these target zones. Referring to figure 1, a piecewise areal interpolation for an area is estimated as:

$$y_{sc} = \lambda_c \, a_{sc} \tag{1}$$

where y_{sc} is the data count (to be estimated) for the sub-zone given by source s and given control class c, λ_c is the area density for the target zones, and a_{sc} is the area of the sub-zone.

The pycnophylactic property enforces the condition that sub-zones belonging to a source zone sum to the given total count:

$$y_s = \sum_c \lambda_c \, a_{sc} \tag{2}$$

where y_s is the known data count for zone s over all constituent classes c. Combining equations (1) and (2) we apply estimates for λ_c to derive expected values for y_{sc} as:

$$\hat{y}_{sc} = \hat{\lambda}_c \, a_{sc} \, y_s \Big/ \sum_c \hat{\lambda}_c \, a_{sc} \qquad (3)$$

This is the expectation of the missing data (E step) in the EM algorithm (Dempster et al., 1977). The maximisation of likelihood over the complete data (M step) for updated values of the area density for each land class λ_c is given as:

$$\hat{\lambda}_c = \sum_s y_{sc} \Big/ \sum_s a_{sc} \qquad (4)$$

The EM algorithm iterates these steps in equations (3) and (4) until there are insignificant changes in estimated λ_c. Further details of the algorithm may be found in Flowerdew and Green (1991, 1994). A feature of the EM algorithm is that it finds the maximum likelihood estimates (Pickles, 1985) for the density parameters. We can think of the density estimates as having a distribution, and the aim is to identify the distribution parameters that most likely generated the data. Flowerdew and Green (1991, 1994) suggest a Poisson distribution with λ_c as the rate parameter, and y_s as a random variable that takes discrete values over the source areas. The Poisson distribution for a random variable y as a count per unit area is:

$$\Pr\{ y \mid \lambda_c \} = e^{-\lambda_c} \cdot \lambda_c{}^{y} \Big/ y! \qquad (5)$$

In this problem the given data are source zones and the probability needs to be calculated over these areas. For each source zone the Poison distribution for a random variable y_s and constituent the rate parameters λ_c where $c=1..k$ is:

$$\Pr\{ y_s \mid \lambda_{c=1,k} \} = e^{\sum_{c=1,k} -\lambda_{sc} a_{st}} \left(\sum_{c=1,k} \lambda_c a_{st}\right)^{y_s} \Big/ y_s! \qquad (6)$$

The likelihood of the data given different estimates for the density parameters is directly proportional to the probability in equation (6), so the likelihood may also be expressed as $L\{y_s \mid \lambda_c\}$. The likelihood of the set of n independent source counts is the product of the likelihoods of the individual sources, namely:

$$L\{ y_{s=1,n} \} = \prod_{s=1,n} L\{ y_s \mid \lambda_{c=1,k} \} \qquad (7)$$

Values for likelihoods may be very small, given that the denominator $y_s!$ in equation (6) may be very large, so the usual approach for computing equation (6) is as a logarithm of the likelihoods, called the log-likelihood. Figure 2 illustrates log-likelihood values for one density estimate from the initial estimate to the maximum-likelihood estimate.

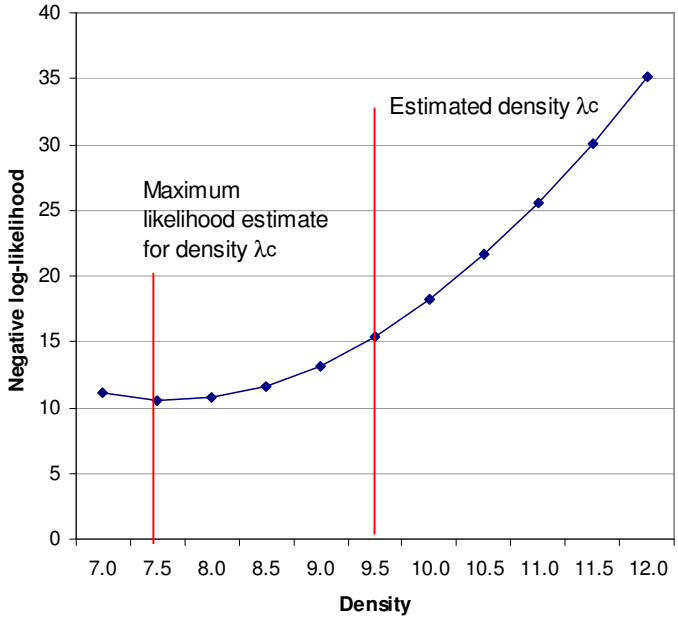


Figure 2. The maximum likelihood estimate for one density value. Note that the y-axis uses negative log-likelihhod so it appears as a minimum.

The method is implemented as a software tool within GIS. The interface for the tool requires a geospatial layer for the source zones with attributes for the zone identifier (subscript *s* in equation (2)) and known data value (y_s in equation (2)). A second geospatial layer is needed for the integrated source and control zones with attributes for their identifiers (subscript *sc* in equation (3)) and data value field to store the result data value (y_{sc} in equation (3)). This integrated layer may be computed in a GIS by a geometrical union of the source and land class zones. The union operation needs to produce unique source-zone areas (that may or may not be contiguous areas) and not to simplify the output polygons. This is because the index in equation (1) refers to a zone represented by a multi-part polygon composed of all areas that belong to source zone *s* and the land class zone *c*. It is also important to have an exact sequence in numbering source and control zone identifiers; this information is used in the GIS to properly associate features for source and control zones. The maximum log-likelihood value is displayed at each iteration during execution to provide feedback on the how the EM algorithm converges. It is difficult for a user to know what value to set for the convergence tolerance; therefore we allow the user to specify the number of iteration steps performed. By inspecting the output it is easy to gauge how convergence is progressing.

A tool for the EM algorithm to perform piecewise areal interpolation is available[1] from the author. The tool is implemented in ArcGIS with a dialogue interface following the description given above. The next section evaluates the error for piecewise areal interpolation.

[1] The tool may be downloaded from http://www.gpa.uq.edu.au/CRSSIS/tools/ludm/ (Lasted accessed Feb. 2007) with instructions and an example.

3. Error analysis

In this section we analyse the distribution of errors for different spatial configurations. By creating synthetic data we aim to find a general result. A method for characterizing the heterogeneity of spatial configurations and summarizing errors is discussed. A set of simple cases are created to explore errors.

We propose using spatial entropy as a measure of the spatial distribution of mapped attributes (Journel and Deutsch, 1993). Entropy is the measure of the disorder or randomness in a system. Shannon's diversity index is widely used to measure entropy for probabilistic and non-probabilistic distributions in spatial analysis (Shannon and Weaver 1949). Using the same notation for zones, Shannon's diversity index I has the form:

$$I_s = -\sum_c p_{sc} \ln p_{sc} \tag{8}$$

where p_{sc} is interpreted as the probability or area proportion of a discrete class for source zone s. I_s is a maximum when all classes occur as equal proportions within a source zone, i.e. $p_c = A_{sc} / A_s \ \forall_c$, and it is zero when a single class occupies a source zone, i.e. $p_c = 1$. An aggregate value for entropy is given as the average for I_s for all source zones, denoted as $\bar{I}$.

An error term is given by the difference between the known data value for integrated sub-zones and the estimated value from equation (3). To allow comparison among cases with differing data values it was decided to report error as the ratio of the absolute error divided by the true data value. This is the mean relative error (MRE) given as:

$$MRE_s = \sum_c \left(\left| \hat{y}_{sc} - y_{sc} \right| \Big/ \left| y_{sc} \right| \right) \tag{9}$$

A series of simple cases were created for three source zones and three land classes. Spatial configurations varied from the case where source and control zones matched exactly, to source and control zones crossed each other. See figure 2. The motivation to create these synthetic cases is that Shannon's diversity index will vary from zero to a maximum for these extremes, and the error term will also vary from zero to a potentially high value for uniformly overlapping classes, i.e. each source zone contains the same proportion of land classes and there is no way to infer their areal density.

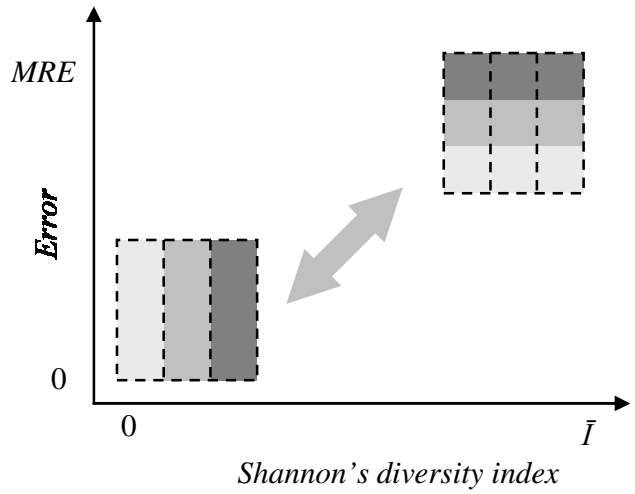


Figure 2. Error in estimating area density and spatial entropy for exactly overlapping (lower left) to for highly segregated spatial configurations (top right).

The results of analyzing a number of spatial configurations for three source zones and three control zones are shown in figure 3. The EM Algorithm clearly gives a more accurate result over the area weighted proportioning method for areal interpolation. Shannon's Diversity Index is a maximum when each land class occurs with equal proportions in each source zone; in which case the EM algorithm reduces to the same accuracy as area weighted proportioning. Shannon's Diversity Index does provide a reasonable predictor of expected errors but it is irregular. This is because the entropy measure is treated as additive, that is it is computed by averaging over all source zones, which violates the assumption that the sources of uncertainty are independent (Shannon and Weaver, 1949). This is the situation with the EM algorithm where you may have one control zone completely overlaping a source zone which provides a good partial solution to resolve areal density for that control class. It shows up in Figure 3 as cases with low MRE below the trend line. This is consistent with conclusions by Sadahiro (2000) with regard to the relative size of source zones to control zones affecting errors. In summary, we believe that entropy provides a good qualitative indicator of expected errors which is biased towards the worst case performance

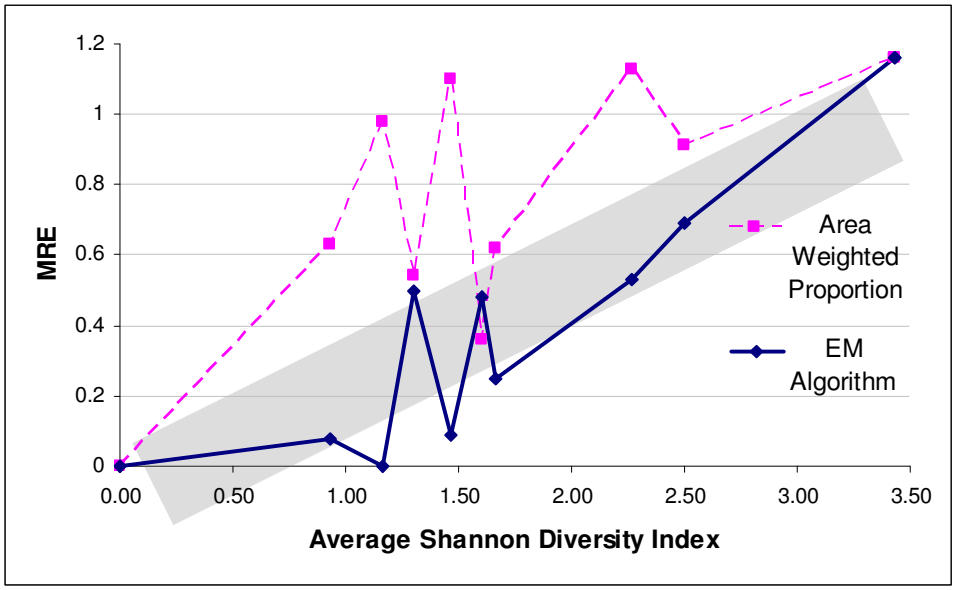
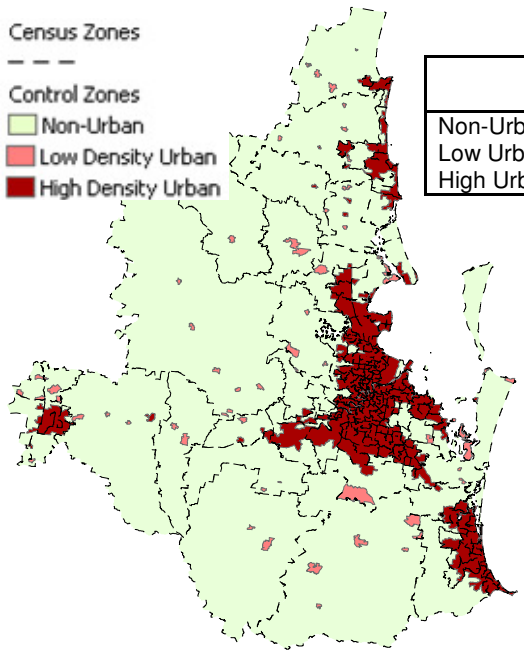


Figure 3. Root mean square error (RMSE) plotted against Shannon's Diversity Index Each point in the plot is a synthetic data case used to assess areal interpolation. Dark line shows error term for EM Algorithm, dashed line shows error term for Area Weighted Proportioning. The gray bar shows confidence limits for fitted linear trend line (r^2=0.78) for EM Algorithm.

4. Example

As a final test we apply this to a real problem. The Australian Bureau of Statistics (ABS) provides census data for statistical collection zones and urban areas. These are derived from aggregated household surveys which are reported for various zoning systems. We test a single case of error for population within the region around Brisbane (approximately 23,000 km^2 supporting a population of 2.5 million people) using 2001 census data. Source zones (298 in total) were obtained for statistical collection areas and the control zones for areas classed as high density urban, low density urban, and non-urban. A map of these zones is shown in Figure 4. The configuration of zones gives an average Shannon's Diversity Index of 0.75. Based upon the relationship in Figure 3 we expect the mean relative error (MRE) to vary between 0.1 – 0.4. In comparison the 2001 census data computes it is as 0.84. Although this error is outside the confidence limits for the simple cases it does incorporate variability in dealing with real data and is still within the same order of magnitude that was inferred. Therefore, it at least confirms that entropy can be used as a qualitative indicator of expected errors. It is concluded that the relationship between an entropy measure and error does provide useful insights.



	Estimated Population	Actual Population	Relative Error %
Non-Urban	120,817	218,543	44.7%
Low Urban	78,567	64,174	22.4%
High Urban	2,278,021	2,194,689	3.8%

Figure 4. Census zones overlaid on control zones. The table shows estimated populations from the EM algorithm using the ABS census population data and the three land classes as control zones. ABS also reports population for these control zones providing a true actual comparison.

5. Conclusion

Areal aggregation is a common problem in geographical analysis. The EM algorithm is relatively accurate and enforces the pycnophylactic condition. The approach infers areal density from the internal structure of patterns formed between known data values for source zones and a set of intersected control zones with regular, but unknown areal densities. The procedure is able to check the internal consistency of a result but not the statistical variation due to observation or process errors. This research note explores if process errors can be analysed based upon patterns in the data, specifically using entropy to measure spatial segregation of mapped attributes. A general result is found that the mean relative error increases as source zones show greater diversity in land classes. This was tested with synthetic data and actual census data. It gives a qualitative means to judge expected error terms based upon the spatial configuration of source and control zones. The research note also describes a way to implement piecewise areal aggregation using the EM algorithm in GIS. A software tool is available from the Internet site given with the author's details for ArcGIS, and implementation on other systems is planned for in the near future.

References

DEMPSTER, A., LAIRD, N. and RUBIN, D., 1977, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, pp.1–38.

FISHER, P.F. and LANGFORD, M., 1995, Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, 27, pp.211–224.

FLOWERDEW, R., GREEN, M. and KEHRIS, E., 1991, Using areal interpolation methods in geographical information systems. *Papers in Regional Science*, 70(3), pp.303-315.

FLOWERDEW, R. and GREEN, M., 1992, Developments in areal interpolation methods and GIS. *Annals of Regional Science*, 26(1), pp.67-78.

FORMAN, R.T., 1995, *Land mosaics: the ecology of landscapes and regions*. (Cambridge: Cambridge University Press).

JOURNEL, A.G., AND DEUTSCH, C.V., 1993, Entropy and Spatial Disorder. *Mathematical Geology*, 25, pp.329–355.

GOODCHILD, M. F. and LAM, N., 1980, Areal Interpolation - a Variant of the Traditional Spatial Problem. *Geo-Processing*, 1(3), pp.297-312.

GOODCHILD, M. F., ANSELIN, L., and DEICHMANN, U., 1993, A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25, pp.383–397.

GREGORY, I.N. and ELL, P.S., 2005, Error-sensitive historical GIS: Identifying areal interpolation errors in time-series data., *International Journal of Geographical Information Science*, 20(2), pp.135–152.

LAM, N., 1983, Spatial interpolation methods: a review. *American Cartographer*, 10, 129–149.

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

OPENSHAW, S. and P. TAYLOR P., 1981, The modifiable areal unit problem. In *Quantitative Geography: A British View*, N. Wrigley and R. Bennett (Eds.), pp.60-69. (London: Routledge & Kegan Paul).

PICKLES, A.R., 1985, An Introduction to Likelihood Analysis. Concepts and Techniques in Modern Geography 42 (Norwich: Geobooks).

SADAHIRO, Y., 1999, Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographical Systems*, 1, pp.323–346.

SADAHIRO, Y., 2000, Accuracy of count data transferred through the areal weighting interpolation method. *International Journal of Geographical Information Science*, 14, pp.25–50.

SHANNON, C.E. and WEAVER, W., 1949, The Mathematical Theory of Information. (Urbana : University of Illinois Press).

TOBLER, W.R., 1979, Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of American Statistical Association*, 74(367), pp.519 – 530