3

The Modifiable Areal Unit Problem

3.1 Basic concepts

In a large proportion of analyses of spatial data, the data are spatially aggregated. Examples include remotely sensed images, where the pixels cover an area (rather than a point), and census area data where population counts are provided for zones. In the first case, the resolution of the imagery is necessarily finite. In the second case, practicality, and the need to product the anonymity of individuals, means that area data, rather than data at the individual level, are used in most applications. The analysis of area data is subject to a range of limitations, and this chapter discusses some associated problems, case studies and potential solutions.

3.2 Scale and zonation effects

In any context where spatial data are aggregations over some area, the results of an analysis are, in part, a function of the size and shape of the zones. The form of zones is usually arbitrary and they can be termed 'modifiable'. The modifiable areal unit problem (MAUP) comprises two parts:

1. The scale problem: relates to the *size* of the zones.
2. The zonation problem: relates to the *shape* of the zones.

Change in the size and shape of a zone impacts on any analyses based on zonal data. The scale problem is also relevant for raster data whereby analyses of raster grids are a function of the spatial resolution of the raster. The scale and zonation effects are illustrated in Figure 3.1.

Exploring Spatial Scale in Geography, First Edition. Christopher D. Lloyd.

Companion Website: www.wiley.com/go/lloyd/spatialscale

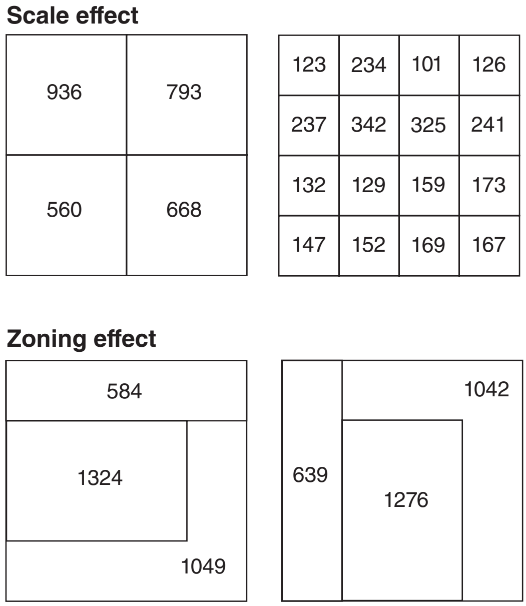

Figure 3.1 Scale and zoning effects.

For the counts shown in Figure 3.1, the mean will obviously increase as the original 4 × 4 cells are aggregated. When the counts are shown as rates (e.g. the counts aggregated for four original cells are divided by four), then the mean is similar for all of the four sets of zones (the same for the top two grids which have different scales but the same shaped zones) while the standard deviation differs. In general, a decreased standard deviation (or variance) would be expected as original units are aggregated.

In most applications zones are not defined objectively, and there are many sets of areal units that could be used to record information about the property of interest. It has been argued that any observed pattern in mapped areal data may be due as much to the zoning system as to the underlying distribution of the variable (Martin 1996).

As an example of alternative spatial aggregations, two maps of population counts in Belfast Urban Area (BUA), Northern Ireland, are shown below. Figure 3.2 shows population counts for 100 m cells while Figure 3.3 shows population counts for 1 km cells. The 100 m and 1 km cell counts are outputs from the 2001 Census of Population. Examination of Figure 3.2 shows that there is considerable variation in populations within each 1 × 1 km area. Therefore, Figure 3.3 contains much less than information than Figure 3.2 and the results of many forms of analyses based on each of the two datasets are likely to vary.

Figure 3.4 shows Catholics as a percentage of the population by Output Areas (OAs) in a part of Belfast, Northern Ireland, in 2001. OAs nest within

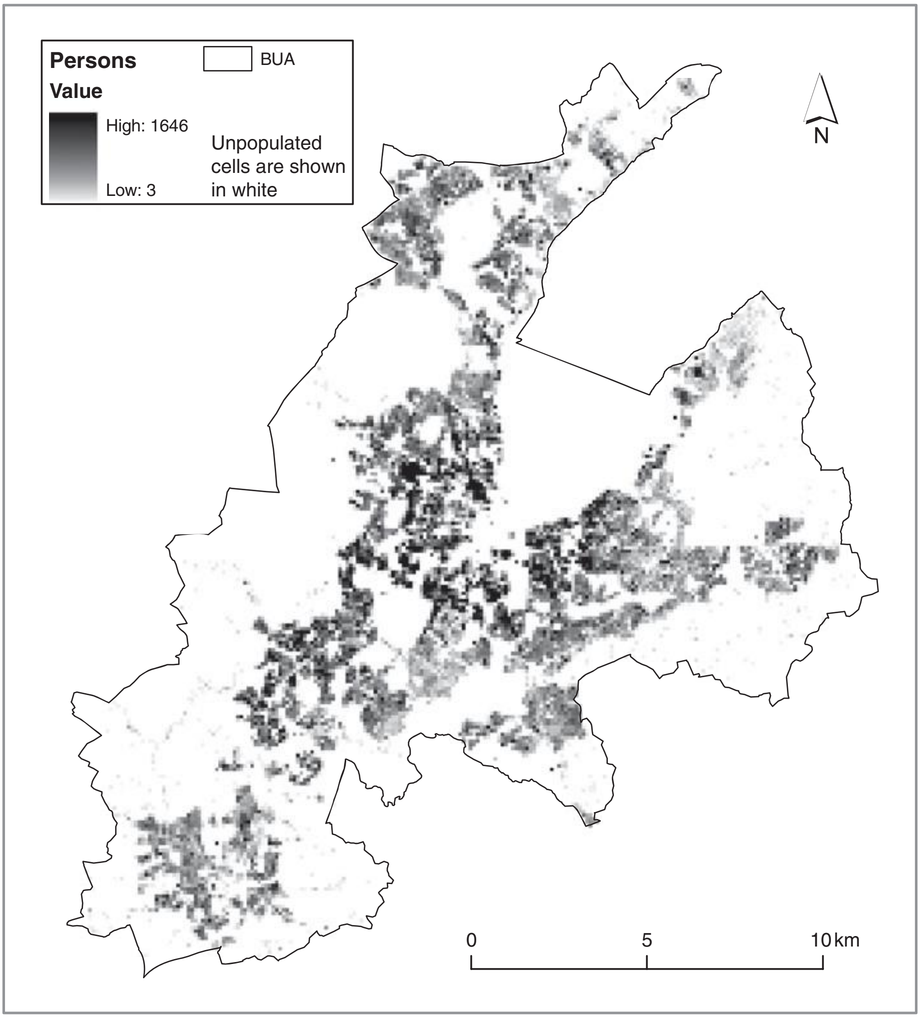

Figure 3.2 Belfast urban area (BUA): persons by 100-m grid cell. *Source:* 2001 Census: Northern Ireland Grid Square Data.

wards and the map shows OAs within one selected ward. Clearly, the percentages of Catholics vary markedly, and there is a distinctive north–south pattern. This suggests that any analysis based on OAs will return quite different results to any analysis based on wards. Section 3.4.1 outlines a case study based on the data for the whole of Northern Ireland.

Each level in a hierarchy of datasets has properties that may not be a simple sum of the disaggregated parts (Bian 1997). The spatial resolution of an image changes fundamental biophysical relationships (the ecological fallacy) (Lam 2004); this problem may exist in any application using aggregate data. Spatial

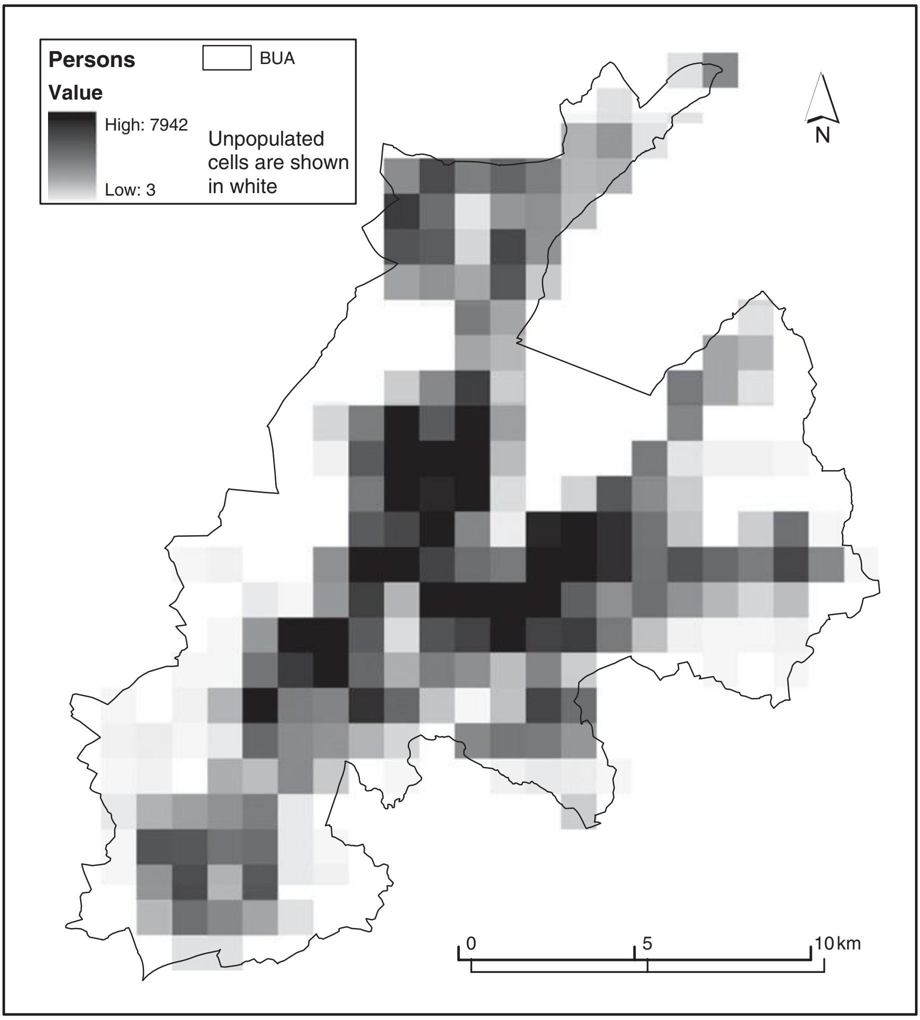

Figure 3.3 Belfast urban area (BUA): persons by 1-km grid cell. *Source:* 2001 Census: Northern Ireland Grid Square Data.

models are frequently scale dependent, and thus, models which are applied at one scale may not be appropriate at another scale (Bian 1997). The MAUP refers to the fact that areal units may be changed and that observed spatial variation may then alter.

3.3 The ecological fallacy

When the available data are spatially aggregate, users are at risk of committing the ecological fallacy (Robinson 1950). This refers to the fallacy of making

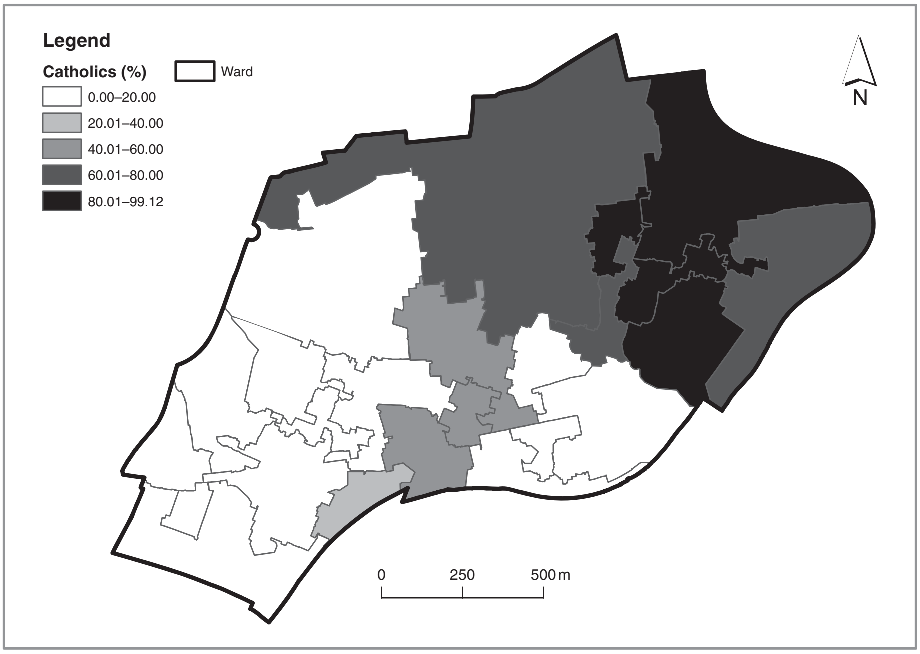

Figure 3.4 Catholics as a percentage of the population by Output Areas. *Source:* 2001 Census: Standard Area Statistics (Northern Ireland). Output Area Boundaries, Crown copyright 2003.

inferences about individuals from aggregate data. Thus, inferences made from aggregate data may be misleading and, as described above, such data are subject to the MAUP. The ecological fallacy is linked to the concept of cross-scale inference – the inferring of behaviour of a system represented at one spatial scale from its behaviour as observed at another, coarser, scale (Goodchild 2011). Thus, the ecological fallacy is the extreme form of cross-scale inference. Haining (2009) cites the case of an analysis of suicide rates in seventeenth century Germany – using aggregate data the rates were shown to be higher in areas with larger percentages of Catholics. However, this does not indicate that Catholics were more likely to commit suicide than Protestants, and an analysis based on individual-level data suggested that the opposite was the case. Models which have been developed to explore relationships between variables while accounting for the ecological fallacy as well as the atomistic fallacy are discussed in Section 5.7. The atomistic fallacy assumes that relationships observed at the individual level apply also at the group level. As Haining (2009) states, the risk of a young person becoming an offender may depend depend not only on individual (personal and household) characteristics but also on neighbourhood (however defined) effects and peer group effects. An exploration of the ecological fallacy using individual-level and aggregate Census data for England is provided by Tranmer and Steel (1998).

3.4 The MAUP and univariate statistics

Many studies in geography and related disciplines have considered the effect on summary statistics of changes in the form of zones or the spatial resolution of imagery. In a study concerned with the analysis of landscape pattern, Turner et al. (1989) assess the effect of changing resolution and study area extent on landscape characterisation using a variety of indices (relating to what are termed diversity, dominance and contagion). In that study (as in ecological contexts generally), the term grain is used to refer to resolution, while extent indicates the extent of the study area. Dungan et al. (2002) argue that resolution comprises more than grain alone, as resolution includes measurement scale of the attribute (z), and not just the size of the measurement cells.[1] Turner et al. (1989) found that land cover types which were found over large contiguous areas (cells with this type were clumped) disappeared slowly or were retained as the resolution was made coarser, while land cover types with more dispersed patterns were lost rapidly. As this study and many others suggest, the analysis of aggregated data is a function of

- the spatial structure of the data
- its relation to the zonal geography.

In a study assessing the impact in changes to the zonal system on measures of residential segregation, Wong (1997) demonstrated that changes in results were a function of spatial autocorrelation in the variable. Wong (2009) illustrates this principle using synthetic data. The main focus in the study by Wong (1997) was on the index of dissimilarity D (defined below). In the case of negative spatial autocorrelation, using different zonal systems may result in quite different values of D. For positive spatial autocorrelation, changing the forms of the zones may have little impact where the zones are smaller than the areas where positive spatial autocorrelation dominates (Shuttleworth et al. 2011).

Moran's I is a widely used measure of spatial autocorrelation (Moran 1950). Moran's I can obtained with:

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(y(\mathbf{s}_i) - \bar{y})(y(\mathbf{s}_j) - \bar{y})}{\left(\sum_{i=1}^{N}(y(\mathbf{s}_i) - \bar{y})^2\right)\left(\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}\right)} \tag{3.1}$$

where w_{ij} is the spatial proximity matrix and $\bar{y} = \sum_{i=1}^{N} y(\mathbf{s}_i)/N$; $w_{ij} = 0$ when location $i = j$. In this equation, y is used as, in the following chapter, z, refers to the deviation of y from its mean, thus $z(\mathbf{s}_i) = y(\mathbf{s}_i) - \bar{y}$.

[1] The example of Dungan et al. (2002) refers to the capacity to resolve lichen species whereby the radiometric sensitivity of a remotely sensed image must be great enough to distinguish lichen types. Equating *spatial* resolution with grain seems reasonable.

Binary connectivity is commonly used to define the weights w_{ij}. In this case, w_{ij} has a value of 1 if regions i and j are contiguous and 0 if they are not. For a contiguity matrix, in Equation 3.1 the number of zones is given by N and $\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$ is twice the number of adjacent zones. The weights are often row standardised (i.e. they are divided by the sum of the weights, and the new weights thus sum to one over all j relating to a unit s_i). Positive values of I indicate clustering of similar values, negative values of I indicate clustering of dissimilar values and values close to zero indicate zero spatial autocorrelation (a 'random' spatial pattern). Tests for spatial autocorrelation are described by Bailey and Gatrell (1995), Fotheringham et al. (2000) and Lloyd (2011) (see also Section 4.2).

Lloyd (2010) assessed changes in the measured spatial autocorrelation in a range of population variables in Northern Ireland in 2001, given change in the three different sets of spatial units used. These units were OAs ($N = 5022$), wards ($N = 582$) and 1-km grid squares (with 6071 grid cells containing the full range of variables,[2] and no cells for unpopulated areas). That study showed that the community background ('religion or religion brought up') variable was more structured at all spatial scales considered than any of a host of social, economic and demographic variables. Given queen contiguity (where zones sharing edges or vertices are neighbours), I for most variables was larger for OAs than for wards, but I for 1-km grid cell values was larger than the equivalent value for wards only in the case of the community background variable. This is a function of the high degree of spatial structuring of religion (and community background) in Northern Ireland. Section 4.2.2 presents the results from part of this analysis.

3.4.1 Case study: segregation in Northern Ireland

Shuttleworth et al. (2011) explored changes in measured residential segregation in Northern Ireland with changes in the zonal units used in the analysis. In this section, a similar analysis is presented based the index of dissimilarity D and the isolation index ${}_mP^*_m$. The data are counts of persons by religion in 2001 and the analysis assessed segregation in terms of the two main religious groups – Protestants and other Christians (denoted by n) and Catholics (denoted by m). As well as counts for standard 2001 zones, the individual-level data were allocated to the zones used in 1991.[3]

Figure 3.5 shows three sets of census zones for Northern Ireland, as used in 2001 – OAs (of which there are 5022), wards (these are collections of OAs

[2] Only total persons and households were reported for cells with less than 25 persons or 8 households; so, only cells exceeding those thresholds are included in this analysis.

[3] This was done by staff at the Northern Ireland Statistics and Research Agency; only the new aggregations, and not the individual-level data, were released.

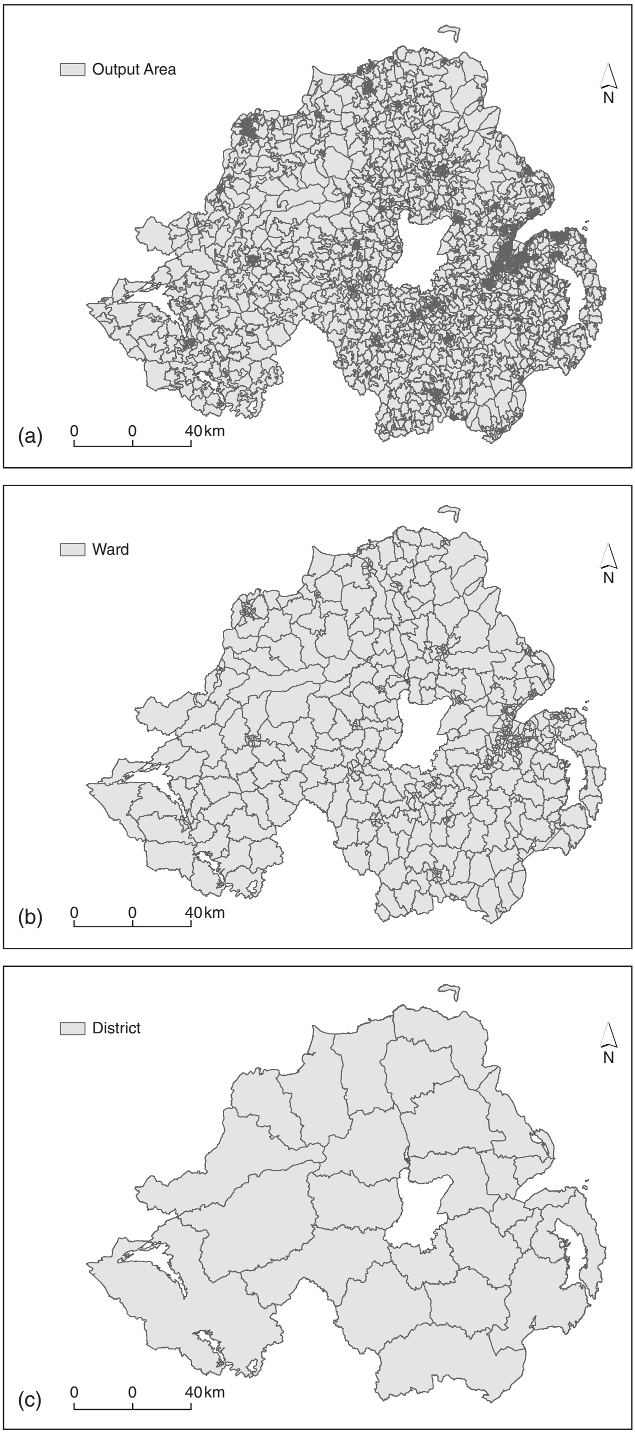

Figure 3.5 (a) Output Areas, (b) wards and (c) local government districts in Northern Ireland. *Source:* 2001 Census: Output Area Boundaries, Crown copyright 2003.

and number 582) and local government districts (LGDs; 26 in number). The present case study makes use of OAs and wards.

The index of dissimilarity D is given by:

$$D = 0.5 \times \sum_{i=1}^{N} \left(\left| \frac{N_{im}}{N_m} - \frac{N_{in}}{N_n} \right| \right) \tag{3.2}$$

where m is a population subgroup and n is another group, i is the index for the zones, N is the number of zones, N_{im} is the number of persons in group m in zone i and N_m is the number of persons in group m.

The isolation index ${}_mP^*_m$, for group m, is given by:

$${}_mP^*_m = \sum_{i=1}^{N} \left(\frac{N_{im}}{N_m} \right) \times \left(\frac{N_{im}}{N_i} \right) \tag{3.3}$$

where N_i is the number of persons in zone i and the other terms are as defined above.

Table 3.1 gives values of D, ${}_mP^*_m$ and ${}_nP^*_n$. The enumeration districts (EDs) nest within 1991 wards, while the OAs nest within the 2001 wards. The values of D are largest for OAs, with a similar value for EDs. D indicates unevenness in the population and it would take a value of zero if all zones had the same proportional share of members of the two groups; a value of 1 would be returned if all zones comprised members of only one of the two groups. So, the values of D for smaller zones suggest greater unevenness than the equivalent values for the two sets of wards. Smaller areas are more likely to be more homogeneous that larger areas; in this case smaller areas are likely to be dominated by either Catholics or Protestants and thus the values of D fall in line with expectation. The isolation index indicates interaction by members of one group with other members of that group. If it is close to 1, it suggests that members of a group are likely to mix with members of their own group and are unlikely to mix with members of the other group. Thus, the values of ${}_mP^*_m$ and ${}_nP^*_n$ also fit with expectation in that they are larger for smaller zones.

Another study which makes use of D in exploring scale effects is presented by Voas and Williamson (2000). The authors compared D for three different

Table 3.1 Index of dissimilarity (D) and index of isolation (${}_mP^*_m$, ${}_nP^*_n$,) by zone: religion in Northern Ireland for 2001.

Zone	D	${}_mP^*_m$	${}_nP^*_n$
ED 1991	0.690	0.769	0.796
OA	0.692	0.772	0.798
Ward 1991	0.618	0.723	0.755
Ward 2001	0.617	0.723	0.755

ED, enumeration district; OA, Output Area.

zonal systems (EDs, wards and districts) for multiple demographic and socio-economic variables across England and Wales. In that case, scale was conceptualised as relating to variation across and within each of the sets of zones.

3.4.2 Spatial approaches to segregation

D and ${}_mP^*_m$, as defined above, are aspatial statistics – they do not account for zones which neighbour given zones. Lloyd and Shuttleworth (2012) utilise geographically weighted indices in the analysis of segregation in Northern Ireland. Such approaches allow for more robust results in terms of sensitivity to changes in the size and shape of the zones – a summary is provided in the following section. In short, statistics derived using geographical weighting schemes (in segregation, as in other contexts) are likely to be robust to changes in the zones used if the geographical kernel size is large relative to the zone sizes. An alternative approach is to measure segregation at an individual (or household) level (Páez et al. 2012), although recent individual-level data are generally only available in secure environments.

3.5 Geographical weighting and the MAUP

In some contexts, the objective might be to compare statistics computed from data representing different time periods. Aggregating to a common set of zones is one possible solution (see Chapter 8). An alternative, which can be considered as conceptually similar, is to use some kind of geographical weighting (see Section 4.2) scheme whereby statistics are derived using a distance-decay weighting function. In this way, local variation is effectively smoothed. The transfer of values from one set of zones to another set of zones or to points is referred to as areal interpolation, and this is the subject of Chapter 8.

Figure 3.6 shows geographically weighted (where the weights were determined using Equation 4.2) versions of D and the isolation index (${}_mP^*_m$, given in the chart as P; see Section 3.4.1 for definitions) for OAs and 100-m grid cells (see Feitosa et al. (2007), Shuttleworth et al. (2011) for definitions of these measures and applications of them). The counts refer to persons by community background ('religion or religion brought up in'). As the bandwidth (which determines the size of the moving window) increases, the indices for the two alternative sets of zones tend to become more similar. The degree of similarity varies between the indices – D and ${}_nP^*_n$ are almost identical for the two sets of zones, while there are slightly larger differences between the two ${}_mP^*_m$ values. If the bandwidth is sufficiently large relative to the scale over which the variables are structured spatially, then the results are likely to be similar for different forms of zones. In this way the effects of the MAUP can

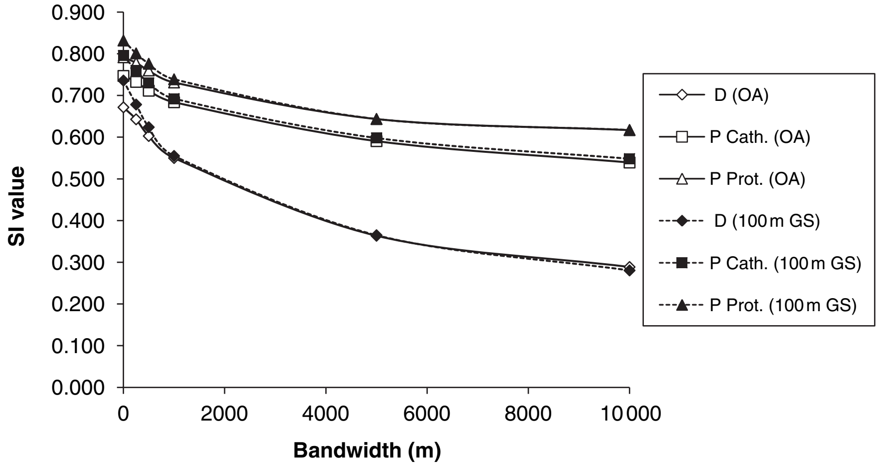

Figure 3.6 Spatial segregation indices by kernel bandwidth for OAs and 100-m grid cells. Cath., Catholic; Prot., Protestant; SI, segregation index; GS, grid square data.

be minimised. But, of course, the scale of the analysis is restricted in that fine scale features are likely to obscured by smoothing.

3.6 The MAUP and multivariate statistics

A key publication which deals with MAUP effects is the chapter by Openshaw and Taylor (1979). The authors compare regression results for several sets of aggregations of the same data source. Specifically, they explore the relationship between the percentage vote for republican candidates in the congressional election of 1968 against the percentage of the population over 60 years old recorded in the 1970 US Census in Iowa. The corresponding correlation coefficients are given in Table 3.2.

Differences between the results for the original 99 counties and the sets of six zones reflect the *scale* problem while differences between the results for

Table 3.2 Iowa county configurations and correlation coefficients.

Alternative combinations of counties	*r*
6 Republican-proposed congressional districts	0.4823
6 Democrat-proposed congressional districts	0.6274
6 Congressional districts	0.2651
6 Urban/rural regional types	0.8624
6 Functional regions	0.7128
99 Iowa counties	0.3466

**Source:* Slightly adapted from Openshaw and Taylor (1979). Reproduced with permission of Pion Ltd, London (www.pion.co.uk and www.envplan.com).

each of the the sets of six zones reflect the *zonation* problem. In many cases, scale effects are more marked than zonal effects. In this particular example, the set of 6 congressional districts corresponds to the smallest value of r, while the largest is for the 6 urban/rural regional types.

3.6.1 Case study: population variables in Northern Ireland

The effect of changing zonal systems is illustrated through a case study making use of four variables derived from Northern Ireland 2001 Census data for LGDs ($N = 26$) and wards ($N = 582$), as shown in Figure 3.5. The four variables related to community background ('religion or religion brought up in'), limiting long-term illness (LLTI), employment and qualifications. Percentages of people in particular groups were converted to log-ratios, as defined below.

1. ln(Cath/NonCath): natural log of Catholics by community background (%)/non-Catholics (%)
2. ln(LLTI/NotLLTI): natural log of persons with a LLTI (%)/persons with no LLTI (%)
3. ln(Unemploy/Employ): natural log of unemployed persons (%)/ employed persons (%)
4. ln(No qual/Qual): natural log of persons with no qualifications (%)/ persons with qualifications (%)

Raw percentages should not be analysed directly using standard statistical measures, and log-ratios are an appropriate solution. Alternative forms of log-ratios and justification for their use are discussed by Aitchison (1986) and Lloyd et al. (2012) (see Section 4.2.2 for more on this topic).

The correlation coefficients for the paired values are shown in Tables 3.3 (LGDs) and 3.4 (wards). Some coefficients have similar values for both sets of

Table 3.3 Correlation coefficients for LGDs.

Variable	ln(Cath/ NonCath)	ln(LLTI/ NotLLTI)	ln(Unemploy/ Employ)	ln(No qual/ Qual)
ln(Cath/NonCath)	1	0.49	0.68	0.60
ln(LLTI/NotLLTI)	0.49	1	0.69	0.51
ln(Unemploy/Employ)	0.68	0.69	1	0.53
ln(No qual/Qual)	0.60	0.51	0.53	1

*Cath, Catholic; NonCath, non-Catholic; LLTI, limiting long-term illness; Unemploy, unemployed; Employ, employed; qual, qualification.

Table 3.4 Correlation coefficients for wards.

Variable	ln(Cath/ NonCath)	ln(LLTI/ NotLLTI)	ln(Unemploy/ Employ)	ln(No qual/ Qual)
ln(Cath/NonCath)	1	0.13	0.49	0.14
ln(LLTI/NotLLTI)	0.13	1	0.69	0.71
ln(Unemploy/Employ)	0.49	0.69	1	0.66
ln(No qual/Qual)	0.14	0.71	0.66	1

*Cath, Catholic; NonCath, non-Catholic; LLTI, limiting long-term illness; Unemploy, unemployed; Employ, employed; qual, qualification.

zones (e.g. for ln(LLTI/NotLLTI) versus ln(Unemploy/Employ)), but they are very different in other cases (e.g. ln(Cath/NonCath) versus ln(No qual/Qual)). Values of r which relate ln(Cath/NonCath) to other variables have the largest differences. At ward level, ln(Cath/NonCath) is strongly positively autocorrelated, whereas the other variables are all much less strongly structured (Lloyd 2010). At LGD level the spatial structure is weak. Therefore, change in r values which relate to ln(Cath/NonCath) is not surprising as the spatial structure has changed with changes in the zones. These findings are, in general, in agreement with those of Flowerdew et al. (2001) who, using a case study based on simulated data, suggest that the MAUP will have an impact in analyses of correlation between variables when there is a spatial pattern in at least one of the variables.

3.7 Zone design

The zones used to report population data are often constructed subject to some criterion such as a minimum or maximum population count per zone. Many other criteria could be taken into account in designing zones. The optimal aggregation of zones or points into new areas is a common problem, although, in most applications, aggregation from points is not a possibility, as such data are generally not available. Openshaw (1977a) developed an automated zoning procedure, and this was extended to the case of spatial interaction (flow) data (Openshaw 1977b). Alvanides et al. (2001) provide a summary of a range of methods for zone design, and their particular focus is on aggregation of zones subject to contiguity constraints. The authors argue that zone design can be a useful analytical tool and they include a case study whereby different numbers of aggregations of zones are generated and visualised as part of an assessment of clustering in LLTI in north-east England.

Zones used, for example, in the output of census data have often been designed before the raw (individual level) data are available. An alternative approach is to use these raw data to design the zones. The OAs used to report data from the 2001 UK Census were generated using an automated zone design methodology. In this case, an intra-area correlation measure was used to maximise social homogeneity within Census OAs with the constraint that the total population and household numbers were above a predefined threshold and close to the target size (Martin et al. 2001).

3.8 Summary

This chapter has defined the MAUP and the ecological fallacy. The potential practical implications of the MAUP and some ways in which the effects can be explored were considered. Case studies relating to residential segregation and the relations between population variables were used to demonstrate how and why changing scales (and shapes of zones) might impact on the results of statistical analyses. The following chapter assesses some ways in which the spatial structure of one or more variables can be explored. Measures of spatial structure are, like any statistical measures, affected by change in the form of zones, as discussed in this chapter, and Section 9.2 links the material covered in this chapter with that covered in the next chapter.

3.9 Further reading

Introductions to the MAUP, along with examples, are provided by Openshaw (1984),[4] Arbia (1989), Green and Flowerdew (1996) and Wong (2009). Tobler (1989) provides an overview of the MAUP in a chapter which argues that spatial analysis methods should be independent of the spatial coordinates used – that is, the results should be what he terms frame independent. A recent review which is concerned with the MAUP and physical geography is provided by Dark and Bram (2007). Section 9.2 deals with the exploration of the MAUP using geostatistical methods.

References

Aitchison J (1986) *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Alvanides S, Openshaw S and Macgill J (2001) Zone design as a spatial analysis tool. In: *Modelling Scale in Geographical Information Science* (eds Tate NJ and Atkinson PM). John Wiley & Sons, Ltd, Chichester, pp. 141–157.

Arbia G (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Kluwer, Dordrecht, The Netherlands.

Bailey TC and Gatrell AC (1995) *Interactive Spatial Data Analysis*. Longman Scientific and Technical, Harlow, UK.

Bian L (1997) Multiscale nature of spatial data in scaling up environmental models. In: *Scale in Remote Sensing and GIS* (eds Quattrochi DA and Goodchild MF). CRC Press, Boca Raton, FL, pp. 13–26.

[4] This book is available at http://qmrg.org.uk/files/2008/11/38-maup-openshaw.pdf

Dark SJ and Bram D (2007) The modifiable areal unit problem (MAUP) in physical geography. *Progress in Physical Geography* **31**, 471–479.

Dungan JL, Perry JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin MJ, Jakomulska A, Miriti M and Rosenberg MS (2002) A balanced view of scale in spatial statistical analysis. *Ecography* **25**, 626–640.

Feitosa FF, Câmara G, Monteiro AMV, Koschitzki T and Silva MPS (2007) Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science* **21**, 299–323.

Flowerdew R, Geddes A and Green M (2001) Behaviour of regression models under random aggregation. In: *Modelling Scale in Geographical Information Science* (eds Tate NJ and Atkinson PM). John Wiley & Sons, Ltd, Chichester, pp. 89–104.

Fotheringham AS, Brunsdon C and Charlton M (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis*. SAGE Publications, London.

Goodchild MF (2011) Scale in GIS: an overview. *Geomorphology* **130**, 5–9.

Green M and Flowerdew R (1996) New evidence on the modifiable areal unit problem. In: *Spatial Analysis: Modelling in a GIS Environment* (eds Longley P and Batty M) GeoInformation International, Cambridge, pp. 41–54.

Haining R (2009) The special nature of spatial data. In: *The SAGE Handbook of Spatial Analysis* (eds Fotheringham AS and Rogerson PA). SAGE Publications, London, pp. 5–23.

Lam NSN (2004) Fractals and scale in environmental assessment and monitoring. In: *Scale and Geographic Inquiry: Nature, Society, and Method* (eds Sheppard E and McMaster RB). Blackwell Publishing, Malden, MA, pp. 23–40.

Lloyd CD (2010) Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: an application of geographically weighted spatial statistics. *International Journal of Geographical Information Science* **24**, 1193–1221.

Lloyd CD (2011) *Local Models for Spatial Analysis*, 2nd edn. CRC Press, Boca Raton, FL.

Lloyd CD and Shuttleworth IG (2012) Residential segregation in Northern Ireland in 2001: assessing the value of exploring spatial variations. *Environment and Planning A* **44**, 52–67.

Lloyd CD, Pawlowsky-Glahn V and Egozcue JJ (2012) Compositional data analysis in population studies. *Annals of the Association of American Geographers* **102** (6), 1251–1266.

Martin D (1996) *Geographic Information Systems: Socioeconomic Applications*, 2nd edn. Routledge, London.

Martin D, Nolan A and Tranmer M (2001) The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A* **33**, 1949–1962.

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23.

Openshaw S (1977a) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers, New Series* **2**, 459–472.

Openshaw S (1977b) Optimal zoning systems for spatial interaction models. *Environment and Planning A* **9**, 169–184.

Openshaw S (1984) The Modifiable Areal Unit Problem. *Concepts and Techniques in Modern Geography* **38**, Geobooks, Norwich, UK.

Openshaw S and Taylor PJ (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: *Statistical Applications in the Spatial Sciences* (ed. Wrigley N). Pion, London, pp. 127–144.

Páez A, Ruiz M, López F and Logan JR (2012) Measuring ethnic clustering and exposure with the *Q* statistic: an exploratory analysis of Irish, Germans, and Yankees in 1880 Newark. *Annals of the Association of American Geographers* **102**, 84–102.

Robinson WS (1950) Ecological correlations and the behavior of individuals. *Sociological Review* **15**, 351–357.

Shuttleworth IG, Lloyd CD and Martin DJ (2011) Exploring the implications of changing census output geographies for the measurement of residential segregation: the example of Northern Ireland 1991–2001. *Journal of the Royal Statistical Society: Series A* **174**, 1–16.

Tobler WR (1989) Frame independent spatial analysis In: *The Accuracy of Spatial Databases* (eds Goodchild M and Gopal S) Taylor and Francis, London, pp. 115–122.

Tranmer M and Steel D (1998) Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A* **30**, 817–831.

Turner MG, O'Neill RV, Gardner RH and Milne BT (1989) Effects of changing spatial scale on the analysis of landscape pattern. *Landscape Ecology* **3**, 153–162.

Voas D and Williamson P (2000) The scale of dissimilarity: concepts, measurement and an application to socio-economic variation across England and Wales. *Transactions of the Institute of British Geographers, New Series* **25**, 465–481.

Wong DWS (1997) Spatial dependency of segregation indices. *The Canadian Geographer* **41**, 128–136.

Wong D (2009) The modifiable areal unit problem (MAUP). In: *The SAGE Handbook of Spatial Analysis* (eds Fotheringham AS and Rogerson PA). SAGE Publications, London, pp. 105–123.