Executive summary

Spatial data are observations at specific locations or within specific regions – they contain information about locations and relative positions as well as measures of attributes. However, except for individual georeferenced records, there is no unique unit for spatial analysis. Areas over any continuous study region can be defined in a very large number of ways. In other words, the unit of spatial analysis is modifiable, unlike in time series where days, months and years give data a consistent structure. This is potentially problematic, since the results of quantitative analysis applied to such data will depend upon the specific geography employed.

The modifiable areal unit problem (MAUP) can be defined as ‘the sensitivity of analytical results to the definition of [area-based spatial] units for which data are collected’. It is a problem because changes in inferences can be - and more often than not are - associated with changes in geographical boundaries. Variations in areal units refer to their size and shape. Correspondingly, the MAUP has two components: the scale problem and the zoning problem. The scale problem refers to the variation in numerical results that is strictly due to the number of areal units used for the analysis of a given territory (i.e. the level of spatial resolution). The zoning problem arises because of the large number of possible ways of partitioning a study area into a given set of areal units, and the fact that how such a partition is configured may impact on the results of analysis.

This paper provides a summary of the academic literature on the MAUP and the main research themes that have been covered. It puts forward a research plan to investigate the impact of the problem at ONS using a large-scale simulation study based on individual-level data. The objective of the research is twofold:

a) To understand how scale and partitioning effects impact on the results of typical analysis scenarios at ONS.

b) To developing methods for quantifying the potential impact of MAUP or “MAUP risk factor” in given analysis scenarios.
Aim of paper

This paper provides background information on the Modifiable Areal Unit Problem and outlines a proposal for a research programme to investigate three issues:

a) Understanding how the MAUP impacts on ONS data.

b) Choosing sensible criteria to decide how best to aggregate unit-level data to areas for reporting and analysis

c) Developing diagnostic tests for sensitivity to geographical partitioning and scale effects.

Requested actions from the committee

The committee is asked to consider whether the proposed research programme is appropriate, and whether alternative areas of work or datasets should also be considered. We also request feedback on whether there are other areas of the statistical literature that are pertinent to this problem and should be investigated. Finally, no underlying theoretical framework for MAUP has been fully proposed. We are interested in the panel’s thoughts on how this problem fits with existing statistical theory, and whether there are particular theoretical properties that might be usefully applied here.

Main issues for discussion

QUESTION 1: Is there other important research that covers MAUP which we have omitted from this review? Can the panel advise us of other sources of information about the problem?

QUESTION 2: How does the MAUP sit with existing statistical theory? What statistical concepts underpin it?

QUESTION 3: Can the panel suggest other data themes that we might include in our analysis?

QUESTION 4: Do the panel think that our broad research plan is appropriate? Are there alternative approaches that we should consider instead?

QUESTION 5: At present, we propose to use real, unit-level data (such as census counts and house prices) as the base data for this project. Is this the right approach? Do the panel think there is merit in producing simulated datasets instead / as well? Can they advise us on how we might take this forward?
The Modifiable Areal Unit Problem (MAUP): Research Planning

1. Introduction

Spatial data are observations at specific locations or within specific regions – they contain information about locations and relative positions as well as measures of attributes. Three main types of spatial data can be identified: geostatistical, lattice and point pattern data. Geostatistical data consist of measurements taken at fixed locations (e.g. rainfall measured at weather stations). Lattice or area data contain observations for regions, whether defined by a regular grid or irregular ones (e.g. mortality ratio per ward). Point pattern data relate to situations where locations are of interest (e.g. household addresses). Of these, area data are the most common type of spatial data published by national statistical institutes.

However, except for individual geo-referenced records, there is no unique unit for spatial analysis. Areas over any continuous study region can be defined in a very large number of ways. In other words, the unit of spatial analysis is modifiable, unlike in time series where days, months and years give data a consistent structure. This is potentially problematic, since the results of quantitative analysis applied to such data depend upon the specific geography employed.

As long as the results of quantitative analysis across space are used simply to describe the relationship among variables, the dependence of these results on the specific boundaries used for aggregation is simply a fact that needs to be taken into account when interpreting them. The problem appears when the differences in parameters from quantitative analysis used to make inferences lead to different – at times contradictory – findings. These findings can be related to either the refutation of certain theoretical models or to the identification of specific policy implications.

Fotheringham, Brunsdon and Charlton (2000) identify the MAUP as a key challenge in spatial data analysis. Its consequences are present in univariate, bivariate and multivariate analyses and could potentially affect results obtained by all the users of area level data published by the Office. The implications of the MAUP affect potentially any area level data, whether direct measures or complex model-based estimates. Here are a few examples of situations where the MAUP is expected to make a difference.

1. The special case of the ecological fallacy is always present when Census area data are used to formulate and evaluate policies that address problems at individual level, such as deprivation. Also, it is recognised that a potential source of error in the analysis of Census data is ‘the arrangement of continuous space into defined regions for purposes of data reporting’ (Amrhein, 1995: 107).

2. The MAUP has an impact on indices derived from areal data, such as measures of segregation, which can change significantly as a result of using different geographical levels of analysis to derive composite measures (Taylor, Gorard and Fitz, 2003).

3. The choice of boundaries for reporting mortality ratios is not without consequences: when the areas are too small, the values estimated are unstable, while when the areas are too large, the values reported may be over-smoothed, i.e. meaningful variation may be lost (Nakaya, 2000).

More generally, the importance of area-level data to the allocation of resources for many key public services (schools, public transport etc.) is well known. The ONS publishes more and more data at various sets of boundaries and levels of spatial disaggregation (over 20 geographies in total, currently). Also, changes in subsets of boundaries for a given geography are very common. There is scope for both scale and zoning effects on the results of analyses based on our data, whether they are simple or intricate ones. The chance that users will obtain different – potentially contradictory - results on the same data, depending on the
Van Beurden and Douven (1999) draw attention to the fact that decision makers assume that the results of analysis applied to spatial data are generally valid and insensitive to any geographical aggregation. It is essential that we have a clear vision and understanding of the extent of the impact of the MAUP in our data and that we caution and guide users accordingly.

2. Background

The modifiable areal unit problem (MAUP) can be defined as ‘the sensitivity of analytical results to the definition of [spatial] units for which data are collected’ (Fotheringham and Wong, 1991: 1025). This is a problem because changes in inferences can be - and more often than not are - associated with changes in geographical boundaries. According to Openshaw (1984), this is ‘a major geographical problem with ramifications that need to be properly appreciated by geographers and all other interested in the analysis of spatial aggregated data’ (p.6); ‘the MAUP is endemic to all spatially aggregated data and will affect all methods of analysis based upon such data’ (p.25).

Variations in areal units refer to their size and shape. Correspondingly, the MAUP has two components: the scale problem and the zoning problem. The scale problem refers to the variation in numerical results that is strictly due to the number of areal units used for the analysis of a given territory (i.e. the level of spatial resolution).

The zoning problem refers to the differences in numerical results due to the way in which smaller areal units are regrouped into a given number of larger areal units, at a given scale. This follows from the existence of virtually infinite numbers of surface partitioning schemes, for any desired number of areas.

In practice, the two aspects – scale and zoning – interact. Both effects tend to increase in severity with decreasing number of zones and they are known to impact upon correlation coefficients size and variance and regression parameter estimates (Wrigley, 1995).

These two components are illustrated in the figure below, for both grid and irregular areas.

In each case, a given surface can be partitioned into four areas in more than one way. Two such possibilities are displayed along the horizontal dimension, labelled Aggregation/Zoning Effect. Also, the same surface can be partitioned into either few relatively large units or more relatively small ones. This is illustrated along the vertical dimension of the graph, labelled Scale Effect. Note that the zoning problem appears again when defining the boundaries of the areas at the level with the finer geographical resolution (e.g. the 9 small irregular areas).

---

1 Some authors call this the aggregation effect (e.g. Gotway and Young, 2002). This can lead to some confusion as some authors use “aggregation problem” to mean the MAUP, while other use “aggregation” to refer to zoning (see note 2).

2 Openshaw (1977, 1984) refers to this as the aggregation problem, where the aggregation is applied to the smallest territorial unit of analysis available, so-called “basic” spatial units. Amrhein (1995) calls it the zonation effect, while Gotway and Young (2002) name it the grouping effect.
Tobler (1989) indicates that, to some extent, the MAUP may be the outcome of spatial nonstationarity, i.e. of the absence of stability across space in the relationships analysed. Wong (1996) explains that the source of the MAUP lies in the alteration in spatial autocorrelation that comes with changes in the boundaries of the areas analysed: ‘[B]ecause (dis)similarity among areal units is not uniform over a large region in all directions, merging smaller areal units using different spatial partitioning schemes is the same as smoothing different combinations of spatial neighbours. Because different combinations have different degrees of similarity, the smoothed results also will differ from each other. Thus, different zoning schemes can render different results’ (p.88).

Gotway and Young (2002) identify twelve other concepts interlinked with the MAUP. Among these is the ecological fallacy, which arises when individual level characteristics and relationships are studied using area level data. The fallacy refers to drawing conclusions about individuals from area-level relationships that are only significant due to aggregation and not because of a real link. Robinson (1950) shows that correlations between two variables can be high at area level but may be very low at individual level. His conclusions are that area level correlations cannot be used as substitutes for individual correlations. Gotway and Young (2002) view the ecological fallacy as a special case of the MAUP, King (1997) argues the reverse.
Similarly, the MAUP is linked to the more general change of support problem in geostatistics, which refers to making inferences at one level different from the one of the data used for analysis (Cressie, 1993).

2.1 Key findings in the literature

The literature on the MAUP is over 70 years old, a large proportion of it having been produced around and after the 1970s. A literature survey by Gotway and Young (2002) indicates that the effects of the MAUP documented to date do not only refer to simple statistics such as variances and correlation coefficients, but also cover multivariate regression, Poisson regression, hierarchical random coefficients models, spatial interaction models and spatial autocorrelation statistics3.

Páez and Scott (2004) also give a brief overview of techniques known to be affected by the MAUP: correlation analysis, regression analysis, spatial interaction modelling, location-allocation modelling, discrete choice modelling and index construction from areal data. Due to the MAUP, results vary when the same technique is used to analyse the same spatial system, whenever data for areas delimited by different sets of boundaries are used. As they explain, ‘the outcome from a study using data reported for areal units of a specific zoning system is merely one manifestation from a range of possible outcomes’ (p.58).

Most of the research published up to now in this area is concerned with univariate and bivariate analysis, for which some theoretical foundations have been put, and concentrates on correlation analysis. Batty and Sikdar (1982) use spatial entropy to measure the interaction between origin zones and destination zones. They develop statistics that detect changes in spatial interpretation at different levels of aggregation. Arbia (1989) proposes a general framework to study the effects of transformations of spatial data on statistical analysis by considering the joint distribution and moments after aggregation in terms of the joint distribution and moments before aggregation.

2.1.1 The Scale Effect

Because it is easier to study, the scale effect seems to dominate much of the literature on the MAUP (Openshaw and Taylor, 1981). This effect usually translates into increasing correlations between variables as the size of the areas under analysis increases. This is a tendency recognised in the literature since Gehlke and Biehl (1934), who also find that no systematic effect is present in the case of random aggregations (as opposed to aggregations of neighbouring areas into larger ones). The presence of MAUP is usually associated with the aggregation of data across neighbouring areas and the alteration of the detected level of spatial autocorrelation (Wong, 1996).

Fotheringham and Wong (1991) explain the increase in the coefficient of correlation through the fact that aggregation into larger and larger area is equivalent to smoothing and therefore a reduction in variance for the variables of interest. This outcome is also affected by the presence of spatial dependence among the initial observations: positive spatial autocorrelation moderates the fall in variance from smoothing, while negative spatial autocorrelation exacerbates it4. Furthermore, spatial aggregation induces positive autocorrelation by design (Gotway and Young, 2002), so the change in variance is compounded further.

3 They review a number of variations on the theme of kriging, Kalman filters and Bayesian multi-scale modelling as solutions to analysing together data that is incompatible spatially. This is different from the aim of the analysis we propose below, which focuses more on the data than on the particular models used to analyse it.

4 This happens because in the presence of positive spatial correlation the values for neighbouring areas are more similar to each other than if they were random, while in the presence of negative spatial correlation they
For example, the fall in variance from spatial aggregation in the case of two variables $X$ and $Y$ amounts to a fall in the denominator of the ratio that gives the correlation coefficient and therefore in the values of this correlation:

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)}}. \quad (1)$$

To the extent that the covariance is relatively stable, this will lead to an increase in $r$.

The extent to which spatial autocorrelation and cross-correlation affect this particular outcome remains an empirical question. In a simulation context Bian and Butler (1999) find that, when the data are aggregated upwards using means or medians, the decrease in standard deviation is non-linear, a result which they attribute to spatial autocorrelation. Clark and Avery (1976) investigate the effect of proximity aggregation on the slope coefficient of a bivariate linear model. Their analysis indicates that the slope coefficient and the correlation coefficient increase at higher levels of proximity aggregation. They conclude that these results are related to the manner in which the covariation between the independent and dependent variables changes with increased aggregation and also to the way in which spatial autocorrelation is exhibited among the micro- and macro-level data.

Openshaw and Taylor (1981) report results from aggregations that produce correlations between -0.97 to +0.99 and ‘a level of fit which could be nearly perfect or incredibly poor’ (p.62). More recently, Openshaw (1994) reports correlations between -0.99 and 0.99, depending on the level of spatial aggregation. The link to gerrymandering, recognised in the political geography and political economy literature, is very clear. In another key early study, Yule and Kendal (1950) stress that estimated parameters are only valid for the specific set of boundaries used and that they measure not only the variation in the variables considered, but also the properties of the territorial divisions imposed to obtain area level measures. Robinson (1950) links the MAUP to the ecological fallacy for the first time.

The most prolific author on the MAUP, Openshaw (1984) concludes that ‘[Q]uite simply, different aggregations yield different results but without any systematic trends emerging that can be used for prediction or correction purposes’ (p.5). Challenging the view that the MAUP is fully unpredictable in its implications, Amrhein (1995) conducts a detailed study based on simulated point data aggregated to grids and concludes the following: (i) there are no effects on the mean from scale or zoning at any level of aggregation; (ii) there are no pronounced scale effects on variance, beyond those expected from the decrease in the number of observations; (iii) populations with high variances display stronger zoning effects than populations with low variances; (iv) regression coefficients exhibit scale effects that increase systematically as the number of zones falls; (v) standard deviations of regression coefficients show pronounced effects from zoning; (vi) Pearson’s correlation coefficient is affected by systematically increasing aggregation effects as the number of areas decreases (p.118). As his simulation uses only uniform and normal distributions, it is hard to know to what extent these findings generalise to data that do not follow, at all or closely enough, these profiles.

In terms of the distribution of values resulting from spatial aggregation (mean, median, standard deviation), Bian and Butler (1999) find, also using grid simulated data, that different aggregation methods alter the statistical and spatial characteristics of the data in different ways. They argue that, while all aggregation are more dissimilar; averaging over similar values changes the variance by less than does averaging over dissimilar ones.
methods lose details, some are better than others at retaining the statistical characteristics of the original data.

A number of ideas have been put forward in the multivariate context, too, but they remain exploratory and empirical rather than theoretical, given the complexities involved in obtaining analytical solutions. Openshaw (1984) argues that, since different variables can be affected differently by aggregation, ‘multivariate techniques based on correlations will tend to amplify the differences in results caused by the use of different zoning systems’ (p.14). Following an evaluation of the MAUP in a simulation framework in the context of multiple linear and logit regression models, Fotheringham and Wong (1991) conclude that this problem is ‘essentially unpredictable in its intensity and effects’ (p.1025) and warn against the unreliability of multivariate analysis undertaken with data from areal units. They conclude that ‘[I]t is clearly possible to find almost any desired result by aggregating the data in different ways’ (p.1041). Interestingly, they do not find any link between the level of spatial autocorrelation in a variable and its vulnerability to the MAUP. However, this could well be because the specific W matrix used to measure spatial autocorrelation (first order binary contiguity) was misspecified.

The MAUP remains an unsolved problem and there is no agreement in the literature over the precise scope of its implications and their predictability.

QUESTION 1: Is there other important research that covers MAUP which we have omitted from this review? Can the panel advise us of other sources of information about the problem?

2.2 Solutions proposed in the literature

Whilst some analysts dismiss the MAUP as an insoluble problem, many assume its absence, by taking the areas considered in the analysis as fixed or given. Most of those who recognise the validity of the problem, approach it empirically and propose a variety of solutions: using individual level data; optimal zoning; modelling with grouping variables; using local rather than global analysis tools (Fotheringham, Brunsdon and Charlton, 2000), applying correction techniques or focusing on rates of change rather than levels (Fotheringham, 1989).

The only way to have analysis of spatial data without the MAUP is by using individual level data. While being widely recognised (e.g. Fotheringham, Brunsdon and Charlton, 2000), this solution is of little practical relevance for most users of official statistics, due to confidentiality constraints5.

Presenting a set of results together with their sensitivity to the MAUP is a widely recommended but little followed practice. Reporting sensitivity of analytical results to scale and zoning effects has been done by several authors who used results for a large number of arbitrary regions produced by Thiessen polygons or using grids6.

Moellering and Tobler (1972) propose a technique that identifies and selects the appropriate set of boundaries on the basis of the principle that the level with most (statistically) significant variances is the one where spatial processes are ‘in action’. This solution, however, only deals with the scale effect of the MAUP7.

5 The Census samples of micro-data made available have limited use in spatial analysis, due to the coarse spatial scale that they cover (Fotheringham, Brunsdon and Charlton, 2002).
6 See Wong (1996); also, see Páez and Scott (2004) for a review.
7 Further limitations are discussed by Wong (1996).
2.2.1 Optimal Zoning

A well-documented solution is that of using data aggregated spatially at optimally-designed boundaries. In this sense, the scale and zoning aspects of the MAUP can be viewed as part of a problem of optimal design (Openshaw, 1977), where zones that optimise some objective function related to model performance, and accounting for constraints, can be identified. Several zoning criteria are reviewed in Openshaw (1984): equal area; equal population; equal density; compact zones; spatial entropy; zonal homogeneity; independent variable variation; relative variation; standard error of slope and optimal zoning. A comparative analysis of the effects of the varied criteria proposed show that, with the exception of optimal zoning, they produce different results which are as arbitrary as the criteria themselves (p.33). Nakaya (2000) proposes the use of the Akaike Information Criterion (AIC) for the selection of scale and zoning in a modelling environment that brings together cluster and regression analyses to produce adaptive and associative zoning. The application he considers is that of standardised mortality ratios in metropolitan Tokyo. Tagashira and Okabe (2002) derive and compare the variances of estimators for unit level and area level regression analyses (relative efficiency), accounting for scale and zoning effects. Under specific conditions, they explore the number of zones for which the variance for estimates from the area-level model is close to that from the individual level model. Also, for a given number of areas they obtain the zoning systems that correspond to the minimum and the maximum estimator variance.

Optimal zoning remains impractical as a generic solution, as it implies the use of analysis-specific boundaries. This means that either the analyst has access to unit-level data, which (s)he can aggregate to any boundaries desired, or that the data provider makes available aggregates at any conceivable set of boundaries. Furthermore, regardless of criterion, different variables may be optimally zoned to different sets of boundaries, adding complications to their being modelled together. It is unlikely that optimal zoning will lead to identical boundaries for different variables.

2.2.2 Modelling with Grouping Variables

In the context of ecological analysis, modelling with grouping variables has been proposed as a way of circumventing the MAUP, through the creation of a hierarchical model structure (Steel and Holt, 1996; Holt et al., 1996). The grouping variables are measured at individual level and are used to adjust the area level variance-covariance matrix and bring it closer to the unknown individual level variance-covariance matrix.

The availability of individual level grouping variables related to the area level process of interest remains an important constraint for the applicability of this solution. More important, however, is the limiting premise that ‘generally, individuals within the same area tend to be more alike than individuals in different areas’ (Holt et al., 1996: 245). This depends entirely on the definition of the area: while spatial proximity of similar values is well recognised in spatial analysis, their presence in the same area or in separate areas is exclusively the outcome of border design. In other words, it appears that the assumption of area homogeneity is not easy to defend. Also, the outcome is not really free of the MAUP, as the relationship between individuals and areas can change depending on the area definition used (i.e. zoning effects). Steel and Holt (1996) and Holt et al. (1996) take the areas as given and concentrate entirely on effects from aggregation of individual values to those for the given boundaries.

Using local analysis tools, such as geographically weighted regression (Fotheringham, Brunsdon and Charlton, 2002), may go some way towards limiting the global effects of the MAUP, but may introduce different distortions in the analysis, to do with the specific spatial structure used to identify the ‘local area’.
As far as the application of corrections is concerned, weights are amongst the most commonly proposed solution\(^8\), but they do not, in general, succeed in addressing both the scale and the zoning effects.

2.2.3 Towards a statistical framework for the MAUP

Despite the wide and varied coverage in the literature, the MAUP is far from solved. According to Wrigley (1995), two main hurdles remain: the absence of an adequate statistical framework for treating the MAUP and the absence of clearly formulated statistical models that can incorporate the empirical regularities identified by geographers working on the MAUP.

Openshaw and Taylor (1981) summarise attempts made to develop a statistical theory of the MAUP and dismiss the analogy between the scale problem and sample size on the one hand, and between the zoning problem and sampling error on the other, as unsatisfactory.

Often the MAUP is regarded not as a problem with the data, but as a consequence of using inappropriate statistics and models in conjunction with spatial data. For example, Fotheringham (1989) proposes shifting the focus of spatial analysis towards rates of change as a potential solution. They discuss substantive errors that may be associated with spatial data and spatial analysis and look at how these errors can be minimised by looking at estimating the rate of change in a variable caused by changes in the scale at which the variable is measured. King (1997) argues that the MAUP is not an empirical, but a theoretical problem and, as such, not difficult to solve: the solution lies in using scale-invariant statistics. However, while the use of fractals, for instance, may appeal to some academic users of spatial data, this solution is not particularly relevant when most analysts remain interested in simple statistics that are influenced by geography, such as correlation coefficients.

In a paper indirectly related to the MAUP, Coulson (1978) proposes the concept of “potential for variation” as a summary of the opportunity for unit data to differ from the mean value for the areal unit to which the units are aggregated, for the case when individual data values are unknown. He explores the assumption that phenomena measured – and therefore the data – are of equal value and evenly distributed within each areal unit (the choropleth assumption). The potential for variation is influenced by the size and the shape of the areas: the larger the area and the further away from a circular shape, the greater the potential for variation within that area. The size component can be measured by comparing all areas to the smallest area under analysis, while the shape component can be measured by comparing any given area to a circle of equal area. This approach hints at the idea of having some indicator of MAUP-risk associated with different geographies, an idea that we explore in more detail in the following section.

While a number of solutions have been proposed in relation to the MAUP, no universally accepted answer can be identified in the literature. Indeed, the practice of producing, reporting and analysing area level data seems, in general, to ignore the potentially serious implications of this problem.

\( ^8 \) See Gotway and Young (2002) for a review of this section of the literature.
3. Proposed analysis plan

The analysis of the nature and magnitude of the consequences of the MAUP has traditionally been based on large-scale numerical experiments. This approach is also the one we propose to follow in our analysis. Our objectives are twofold:

a) Through an experimental framework using appropriate data, to ascertain the magnitude of MAUP effects in a range of analysis scenarios. Our focus here will be in obtaining evidence that MAUP effects are potentially significant for ONS users and therefore worthy of further investigation.

b) To develop diagnostic methods for assessing the sensitivity of particular analysis scenarios to MAUP effects.

The analysis will cover five steps:

i. selecting individual geo-referenced cases for target variables or composite measures that can be computed from the Census;

ii. aggregating these individuals to various sets of boundaries;

iii. exploring the change in parameters from the analysis (univariate, bivariate, multivariate) of the selected variables;

iv. identifying a “potential for MAUP” or “MAUP-risk” factor;

v. exploring differences in the evolution over time of distributions of statistics computed at different geographical levels.

The details for each of these steps are as follows:

1. selecting individual geo-referenced variables available from the Census (Xs and Ys):

   o In the Office we are in the privileged position of having access to unit-level Census geo-referenced data, which is ideally suited for the study of the MAUP. While many other researchers have devoted time to the empirical analysis of this problem, they were constrained to using either simulated data or a small sample from the Census.

   o The choice of variables to be included in the analysis is constrained by the data available.

      a. The selection of specific variables could be informed by the variable selections in Small Area Estimation, for instance, since the MAUP is a problem with clear manifestations in a modelling context; this may also support on-going research work in the Small Area Estimation Centre.

      b. Also, building SMRs from individual level data, at a variety of boundary specifications, and studying the implications in terms of patterns and trends identifiable would have high relevance to understanding better the Office output.

      c. Illustrating the MAUP on area-level measures of segregation or diversity is another possibility.

      d. Area-level classifications are another possible target for study.

      e. An analysis similar to that conducted by Wong (1996) using nine US Census variables for the State of Connecticut could be designed.
2. **aggregating individual variables to various sets of boundaries:**

   o Separating the two types of MAUP effect is not easy. To date, researchers have looked at the scale effect using a fixed zoning system, or explored zoning for a given scale. This ceteris paribus assumption is not justifiable beyond the fact that it simplifies the analysis.

   o When studying scale effects, the aggregating upwards of area into larger ones becomes a zoning problem at a different spatial scale. Also, when studying zoning effects, there is no a priori reason why a certain scale is the right one for analysis; only by considering different zoning choices at a variety of scales is it possible to reveal the nature of the MAUP.

   o For this reason we propose an approach that produces a distribution of outcomes for one effect across a range of measures on the other effect, by repeating each step of the analysis “a large number of times.” Specifically:

     a. for **scale** effects: aggregate (a large number of times) individual level data and data from existing geographies to sets of fewer, larger, nested areas (e.g. from wards to districts and counties; from individual level to LSOA, MSOA and SOA);

        ▪ obtain a distribution for each parameter of interest;
        ▪ consider implications from the optimality properties of the SOA geography;

     b. for **zoning** effects: aggregate (a large number of times) individual data to a given number of areas (e.g. a large number of zoning simulations to obtain sets of areas with as many elements as the set of wards or SOA, but with different boundaries);

        ▪ obtain a distribution for each parameter of interest;
        ▪ consider implications from the optimality properties of the SOA geography.

3. **exploring the change in parameters from the analysis (univariate, bivariate, multivariate) of the selected variables:**

   o with changes in:

      i. geographical scale
      ii. zone system
      iii. scale and zoning;

   o through:

      ▪ analysing the distribution of parameters obtained for each of the many scale and zoning simulations applied;
      ▪ clarifying the implications from the optimality properties of the SOA geography;
      ▪ exploring patterns in error due to aggregation; this is defined as ‘the differences between the aggregated data and the data gathered or reported at the most disaggregated level’ (Wong, 1996: 104).

**QUESTION 3:** Can the panel suggest other data themes that we might include in our analysis?
4. **identifying a “potential for MAUP” or “MAUP-risk” factor:**
   - this would be a measure in the spirit of Coulson (1978);
   - as the MAUP is a global problem, referring to a set of areas rather than a specific area, this measure would be an overall indicator of risk of MAUP for a particular set of boundaries;
   - the risk of MAUP for a specific set of boundaries could be measured in comparison to a reference set of boundaries; a priori, SOAs offer a suitable candidate, as they have been designed in a way that can approximate optimal zoning for the purpose of certain analyses;
   - differences across variables will need to be identified and taken into account;
   - the differences in the spatial profiles of variables with different risk factors could be explored using spatial autocorrelation measures (Schabenberger, O. and C.A. Gotway, 2005) and map comparison indices (e.g. Van Beurden and Douven, 1999; Finn, 1993).

5. **exploring differences in the evolution over time of distributions of statistics computed at different geographical levels:**
   - the feasibility of this section may be restricted by data availability;
   - comparisons across space and time would offer novel insight into the nature of the MAUP;
   - parallel evolutions over time could identify suitable “correction” factors in interpreting the results.

This analysis, if successful, will produce a number of original contributions in both understanding the quality of outputs in the ONS and the literature on the MAUP:

- using a full individual level data set to study the effects of the MAUP on official statistics for the first time;
- producing distributions of outcomes from a large number of simulations, rather than a small number of point estimates;
- identifying a “potential for MAUP” or “MAUP-risk” measure;
- introducing the time dimension in the evaluation of effects from the MAUP in order to compare the dynamics of spatial relationships at various scales and for different zoning scenarios.

**QUESTION 4:** Do the panel think that this broad research plan is appropriate? Are there alternative approaches that we should consider instead?

**Real or simulated data?**

The issue of whether to use real or simulated data in any research programme is of key importance. On the one hand, the use of real data has the benefit that the specific features of key variables of interest become visible. The shortcoming associated is that the MAUP effects may be impossible to separate from all other possible effects, making it harder to formulate a solution. The use of simulated data has the benefit that, by
design, the analysis is capable of separating the effects that are of interest. On the other hand, simulated data will always lack the complexities of real data.

It is notoriously difficult to simulate meaningful spatial structures. This makes the study of MAUP effects with artificially constructed datasets less compelling. The usual structure deployed for generating such data for a variable $Z$ at location $x$ is:

$$Z(x) = m(x) + \varepsilon_1(x) + \varepsilon_2$$

where $m(x)$ is a general trend component, $\varepsilon_1$ is a spatially autocorrelated local random component and $\varepsilon_2$ is random noise. Some authors omit the first term in order to preserve spatial stationarity (Bian and Butler, 1999). The simulation produces a number of realisations of spatial data corresponding to chosen parameters of spatial autocorrelation for $\varepsilon_1$ and to the parameters specified for the distribution of $\varepsilon_2$, usually a normal one.

In our research we propose to use real, individual geo-referenced data in conjunction with a simulation approach to generate area boundaries. Our aim is to combine the strengths of real data and of simulations, and introduces the additional benefit of studying the problem starting from a full set of individual data rather than from data which have already been aggregated at least once into “basic spatial units”, as is the case with most studies to date (e.g. Openshaw, 1984).

**QUESTION 5:** At present, we propose to use real, unit-level data (such as census counts and house prices) as the base dataset for this project. Is this the right approach? Do the panel think there is merit in producing simulated datasets instead / as well? Can they advise us on how we might take this forward?

**References**


---

9 For this, Bian and Butler (1999) use kriging. The computationally intensive nature of such a data-generating process imposes severe constraints: they can produce a maximum of 512 areas (pixels). Other authors generate spatially autocorrelated data using matrices of spatial interaction, $W$. 

**NSMAC (11): Modifiable Areal Unit Problem**


Openshaw, S. 1984. The modifiable areal unit problem. CATMOG Concepts and Techniques in Modern Geography, 38


