Developing the Automated Zoning Procedure to Reconcile Incompatible Zoning Systems

David Martin
Department of Geography, University of Southampton, Southampton, SO17 1BJ, UK
D.J.Martin@soton.ac.uk

Abstract. This paper concerns the problem of matching incompatible zonal geographies, for example in the context of comparing census outputs over time. The automated zoning procedure (AZP) proposed by Openshaw (1977) is reviewed, and a population stress statistic proposed which may be used in an AZP algorithm in order to maximise the match between two zonal geographies. An implementation of this approach is described, and illustrated by reference to UK Census data.

1. INTRODUCTION

A general problem in spatial analysis is that of reconciling data from two incompatible zonal systems. This is a particular issue, for example, when comparing the results of two censuses where there has been change in statistical zone boundaries during the intercensal period or when comparing published data for one zonal geography with a different set of application-specific zones. Previous approaches to this problem have included the production of lookup tables from manual records and address lists (Atkins et al., 1993); the development of areal interpolation techniques (Flowerdew and Green, 1991; 1992), and attempts to identify 'tracts' with constant boundaries within two incompatible systems of small zones (Morgan and Denham, 1982). This is a hard problem, in which the 'best' solution may actually represent a complex trade-off between competing constraints. This paper presents an automated zone matching (AZM) algorithm for the design of tracts, developed from Openshaw's (1977) automated zoning procedure (AZP). AZP is a computationally intensive procedure which seeks to optimize objectives such as minimum or target population sizes, zonal compactness or social homogeneity, by iteratively recombining a large set of building block polygons into a smaller set of output areas. AZM extends this approach by attempting to minimize the mismatch between two input zonal geographies as part of the zone design process.

AZP is being applied to the results of the 2001 Census in England and Wales in order to create Output Areas (OAs) which are entirely separate from the system of Enumeration Districts (EDs) used for data collection. This represents the first separation of collection and output geographies throughout the UK, and is described more fully in Martin (1998). The advantages of such a procedure include its ability to offer uniform OA coverage using an explicit design methodology, something, which has always been problematic when using manual approaches to census area design. Such an approach does not, however, overcome the perennial (in the UK case) problem of changing zonal boundaries between successive censuses, which is the focus of the work presented here.

The rest of this paper reviews the use of automated zone design procedures and relates them to the problem of matching two incompatible zonal geographies. The need for such matching is reviewed, and the unsuitability of areal interpolation approaches explained. A measure of population stress between two zonal systems is introduced and exact and approximate matching situations illustrated. These concepts are incorporated into a new zone design tool, and empirical applications are presented in urban and rural study areas in the UK. Conclusions are drawn with regard to the utility of zone design tools for matching incompatible geographies, and the situations in which such an approach might be applied.

2. AUTOMATED ZONE DESIGN

Data describing the socioeconomic characteristics of populations, most typically from censuses, are conventionally collected and mapped for zonal units. The use of areal aggregations stems from practical considerations (the organization of data collection); output requirements (provision of counts for established political and administrative divisions) and the need to protect the confidentiality of individual members of the population. A difficulty with most zonal geographies is that their relationship with the underlying population characteristics is undefined, resulting in the familiar modifiable areal unit problem (MAUP). Openshaw (1984) argues strongly that the most appropriate response to the MAUP is to design purpose-specific zonal systems. There is a long history of the modification of electoral geographies in order to achieve approximately equal political representation, and the automated partitioning of geographical space in this context is discussed by Horn (1995) and Mehrotra et al. (1998).

Openshaw's (1977) Automated Zoning Procedure (AZP) provides a means to automate the process of designing a zonal system in order to maximise the value of some objective function. The procedure is based on the iterative recombination of building block zones into output regions from an initial random aggregation (IRA) by examining the effect of swapping individual building blocks between output regions. Improving swaps are retained as part of the emerging solution and the IRA is thus refined to produce an 'optimal' boundary configuration, given a particular set of design constraints. Important aspects of the implementation of AZP-type algorithms are the methodology used for the construction of the IRA, and the method used for the weighting and combination of the different constraints (each measured in different units) into a single objective function. An essential part of the implementation of such zonal recombination is the maintenance of a contiguity matrix, allowing the identification of valid swaps between adjacent zones. Before the advent of topologically structured GIS, this was a major obstacle and the contiguity matrices for some early experiments were painstakingly compiled by hand.

Application of these zone design methodologies to published census data is demonstrated by Openshaw and Rao (1995), who consider the 'reengineering' of 1991 UK Census outputs by using enumeration districts (EDs) as input building blocks, and assembling larger zones which match a variety of objective criteria. Openshaw and Alvanides (1999) demonstrate similar applications using a national application with ward-level data. In both cases, the census user is encouraged to reaggregate the standard zones for which data are published to provide larger, purpose-specific zones designed according to clearly defined criteria. Openshaw and Rao (1995) discuss various search algorithms including the simple AZP, simulated annealing (SA) in which suboptimal swaps are permitted in the early stages of iteration, allowing the procedure to escape from potential local suboptima and a tabu search approach, in which recently tried swaps cannot be considered again until a certain number of iterations have elapsed. There has also been consideration of parallel implementations of AZP, although the partitioning of the problem into independent spatial sub-regions is problematic due to the fact that a boundary reconfiguration may have ramifications across the whole of the problem space. Software tools for the implementation of automated zone design include the SAGE package (Haining et al., 1998; 2001) and ZD2K (*http://www.ccg.leeds.ac.uk*), the latter designed with the specific objective of providing a tool for the reaggregation of 2001 Census output geography.

An automated zoning procedure has been adopted for the creation of 2001 Census OAs in England, Wales and Northern Ireland. Scotland has a rather different trajectory, and will be creating its own OAs using an alternative methodology, designed to maximise compatibility with those used there in 1991, based on their membership of higher level areal units. 2001 OAs will be designed by the Census Offices after census enumeration, coding and 'one number census' imputation are complete (ONS/GROS/NISRA, 1999), and represent a completely new subdivision of the country, separate from the ED-based collection geography which has also been used for data output in previous censuses. The use of EDs for output has a number of weaknesses, including wide variations in population size (some EDs being too small to appear separately in the published tables) and a poor match to the widely used postal geography. These difficulties may be significantly reduced by the use of AZP with the smallest divisions of the postal geography as building blocks and OAs being designed with an explicit target population objective and minimum population threshold. This approach will allow a larger number of smaller 2001 OAs than 1991 EDs to be produced, making them more amenable to use as building blocks in user-specific geographies. The use of automated zone design for 2001 Census processing is not the subject of this paper, having been discussed more fully in Martin (1997, 1998). However, the way in which these new 2001 OAs are used is very relevant here.

3. MATCHING ZONAL SYSTEMS

Census users are faced with many challenges, two of which are fundamentally the same, and relate to the matching of census output geography to other geographies. The first of these is the need to compare published data with those from a previous census in order to examine change over time, and the second is the need to aggregate data to other (usually larger) areal units which are important for research or policy purposes, but which cannot be assembled by neat aggregation of the census OAs. For example, UK ward boundaries are subject to continual revision and significant areas of the country will be redrawn between the 2001 Census reference date and the publication of the census statistics in early 2003. In both cases, the challenge is one of finding the 'best match' between two zonal geographies at a given scale of aggregation.

The usual approach to this problem in a research context has been to adopt some form of areal interpolation, and the terminology of 'source' zones (for which counts are currently held), 'target' zones (for which counts are required) and 'intermediate' zones (the intersection of source and target zones) has become widespread. Simple areal interpolation assumes that attribute values are uniformly distributed across the entire area of each zone, and may be redistributed from source zones to target zones in proportion to the areas of the intermediate zones, based on a simple intersection of the zone boundaries. This procedure is appropriate for ratios and percentages, and for when the zones are 'natural' areal units such as soil types or land use classes. However, the assumption of uniform population density is very often far from reality, and Flowerdew and Green (1991; 1992) suggest an enhanced approach which is able to take account of ancillary information in order to weight the interpolation process when count data such as population totals are to be estimated. An overview of interpolation methods for socioeconomic data is provided by Goodchild et al. (1993) and a comparison of methodologies by Fisher and Langford (1995). Others have considered implementations of such techniques within specific software environments (Bloom et al., 1996); the use of specific ancillary variables (Langford et al., 1991; Xie, 1995) and enhanced algorithms (Mugglin et al., 1999; 2000). Fundamentally, each of these is concerned with the statistical interpolation of data between zonal systems rather than the need for direct or approximate matching between zones.

The need to identify best matches between zonal geographies is of particular concern to those users who for some reason are unable or unwilling to engage in areal interpolation, and who are therefore forced to reaggregate. Specifically, it is a challenge for data providers such as Census Offices and government departments who are required to report population statistics for different zonal geographies but who must always be conscious of the risk of differencing if data are released for slightly differing zones. Although population thresholds may be applied to published data to prevent the inadvertent disclosure of information about individuals, the publication of above-threshold data for two slightly different geographies may result in the ability to produce data for sub-threshold intersections of these geographies (Duke-Williams and Rees, 1998). These organizations may have access to the actual counts for both source and target zones, but are unable to publish both sets due to differencing risk, and must therefore publish results for approximately matching geographies instead. Further, there may be many other situations in which a requirement to produce precise counts precludes the use of areal interpolation or weighted allocation from lookup tables. Lookup tables that include population counts in the overlaps between two sets of zones provide a very simple tool for reallocation of population, but do not incorporate adjacency information, and provide information only at the lowest level of aggregation.

In the UK, EDs are designed solely for census purposes and continual changes in higher level areal units and new residential development cause large numbers of EDs to be changed between successive censuses. In England and Wales, only 44% of ED boundaries remained unchanged between the 1971 and 1981 Censuses, and only 32% between 1981 and 1991 (again, the geography system in Scotland is rather different to that described here). The new smaller OAs to be used for 2001 offer many advantages in terms of increased aggregation flexibility, but will produce even higher levels of boundary change, with very few 2001 OAs likely to be coterminous with 1991 EDs. At the time of the 1981 Census, a major manual effort was undertaken to identify small areas whose external boundaries were unchanged, resulting in 48,300 census tracts (mainly in urban areas) comprising aggregations of one or more 1971 and 1981 EDs which could be grouped to form areas with identical boundaries. A further 10,700 parishes or communities (mainly in rural areas of England, and in Wales) which remained largely unchanged between the two censuses (Morgan and Denham, 1982). In 1991 no such exercise was undertaken, leaving census users with no directly comparable small areas between 1981 and 1991. Instead, a lookup table of 1991 EDs to 1981 wards was created by an approximate methodology based on a combination of existing lookup tables and GIS analysis of centroids and boundaries (Atkins et al., 1993). It is this general problem of devising tracts for comparison between zonal geographies which is the focus of the rest of this paper, hence the term 'automated zone matching' (AZM).

Development of the AZM concept requires some measure of the fit between two zonal systems containing population data, and it is proposed that this is most helpfully considered as the 'stress' between the two geographies which results from population misallocation due to approximate matching of source and target zones. Such a stress value may be expressed in terms of any attribute of the zonal data, such as geographical area, but the concern here is with population counts and we shall therefore refer only to population stress in the remainder of this discussion. Consider the example illustrated in Figure 1. When a zone from one (source) system is used to represent a zone from another (target), a perfect match would represent zero stress. As the match illustrated is approximate, both omission and commission errors may occur.

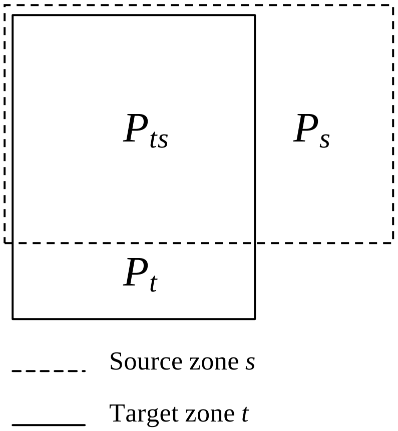

Figure 1: Approximate match between a source zone *s* and target zone *t*

In addition to the correctly matched population P_{ts}, some population P_t which really belongs in target zone *t* is omitted, while other population P_s which belongs in the source zone *s* but not the target zone is incorrectly included. The total population stress q_t for zone *t* may be viewed as some measure of the omission and commission errors as a proportion of the true population. Here, we use the sum of the squared omission and commission errors due from its approximation by zone *s*, divided by its true population, which is the sum of P_{ts} (the correctly matched part) and P_t (the omitted part). The errors are squared in this context in order to give greater weight to large misallocations that are highly unattractive from a zone design perspective.

$$q_t = \frac{P_t^{\ 2} + P_s^{\ 2}}{P_{ts} + P_t}$$

In order to get a summary measure Q_T of population stress across all target zones, we must sum over all T target zones and all S source zones (although many P_{ts} values will be zero, where the zones are disjoint).

$$Q_T = \sum_{t=1}^{T} \sum_{s=1}^{S} \frac{P_t^{\ 2} + P_s^{\ 2}}{P_{ts} + P_t}$$

Such a measure may be used as an objective function in an AZP-type algorithm, such that it is minimized during the iterative recombination stage. This is effectively an objective function for maximising the population match between two geographies. It will be reduced to zero if a perfect match is achieved between the two zonal geographies. In the aggregation of smaller to larger areal units within a perfect hierarchy, there is no population stress, as every member of the population may be correctly matched to a target zone on the basis of their source zone location. Some pairs of geographies are more similar than others. If minor boundary revisions are made to an administrative geography, but it retains the same number of zones, most of which are unchanged, the stress between the two systems will be small, with very few persons likely to have been reallocated between zones due to the revisions. If, however, the first boundary set is replaced with another that subdivides the space in an entirely different way, there is likely to be a high population stress, with great difficulty in approximating the populations of the new zones from the old. This stress will tend to be at its highest when the source populations are high in relation to the target populations, and will vary at differing levels of aggregation according to the precise relationship between the input geographies.

Haining et al. (2001) consider the provision of software tools for spatial analysis, and note that although there is a role for the incorporation of frequently-used functions within GIS software, specialist functions are probably best left to specialist software. In this context, the ability to extract topological information from GIS is a very important consideration. The work described in this paper has been implemented using a newly written AZM program, which embodies the original AZP and AZP-SA algorithms, combined with the necessary structures for boundary matching. AZM has been written in Visual Basic 6 and will run under Windows 95/NT 4 or later. As with Haining et al's (2001) SAGE software, it has been designed with polygon and arc attribute output from Arc/Info as the primary input, but will read the required topological information from any similarly formatted ASCII files. The program, including documentation, is available for download and experimentation from *http://www.soton.ac.uk/~djm1*.

While existing zone design tools effectively create an output geography from an input geography, AZM works with a third, intermediate, layer. Two input geographies are intersected in an external GIS and the attributes of the intersected polygons and arcs are supplied to the program to provide the input building block layer. It is thus necessary to maintain not only the contiguity matrices but the membership lists of each layer with respect to each other layer, making this task conceptually and computationally rather more demanding than a conventional AZP problem. These concepts are illustrated with reference to the following series of figures, all of which are based on the intersection of two input zonal geographies, A and B, illustrated in Figure 2.

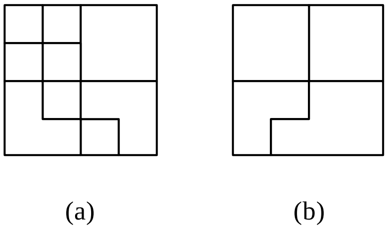

Figure 2: Two input zonal geographies, A and B

In each of the following four figures the first diagram (a) is the same, and shows the input zone layer resulting from the intersection of zonal geographies A and B. The second diagram (b) in each sequence shows the intermediate layer following aggregation analysis and this forms the input to iterative recombination using some objective function. The third diagram (c) represents a valid output layer from this processing. The details of the objective function are irrelevant here, the purpose being to illustrate the role of fixed and approximated geographies. If neither input geography is fixed, the intermediate layer is the same as the input layer, the program implements a conventional zone design procedure, treating all the input building blocks as separate entities (Figure 3), in this case producing three output zones which cut across both the input geographies.

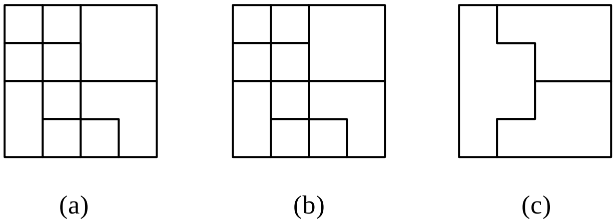

Figure 3: AZM with neither layer fixed

If one zonal geography is fixed (in this case B), then the input building blocks necessary to reconstruct each of the zones in that geography are aggregated to form intermediate building blocks (Figure 4) before recombination iteration, again resulting in three output zones, which respect B but cut across A.

If both zonal geographies are fixed, then a more complex aggregation analysis is required to find the smallest clusters of input building blocks which can be assembled from both input layers simultaneously without cutting across either (Figure 5). No iterative recombination is required if the target zone size is small in relation to the input zones.

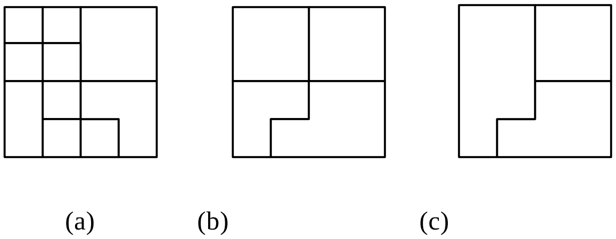

Figure 4: AZM with layer B fixed

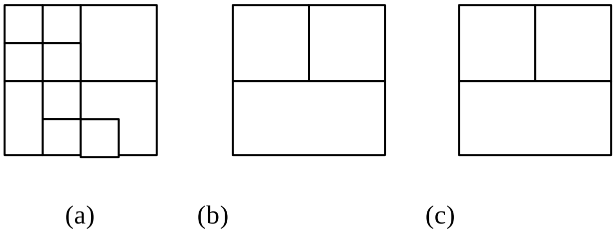

Figure 5: AZM with both layers fixed

A fourth scenario (Figure 6) is that in which one of the zonal geographies is fixed (in this case, A), and the design objective is to achieve the minimum population stress with geography B, using the stress measure introduced above. The two zones of geography B in the upper half of the map are reproduced exactly, whereas those in the lower half cannot be exactly reconstructed from geography A, and the result is therefore approximate.

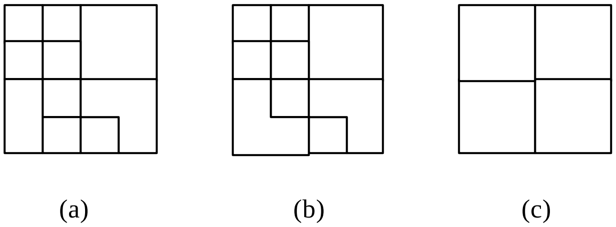

Figure 6: AZM with layer A fixed and layer B approximately matched

If the objective is to achieve a best possible match between two zonal geographies, and this has priority over all other constraints, an initial non-random configuration may be produced which begins to assign successive building blocks into tracts on the basis of the target zone to which they contribute the greatest proportion of their population. Using this simple allocation, similar to the use of a lookup table, the contiguity of the formative tracts must be monitored very carefully, as direct allocation will frequently result in non-contiguous tracts. This computed low-stress solution will thus attempt to create an output tract corresponding with each zone to be matched. Iteration may then be started from this configuration, which attempts to refine the solution in terms of any other design constraints that may be set. This has the effect of giving the algorithm a starting position that is strongly biased in favour of a good match between the two zonal systems. Alternatively, the population stress measure may be used in the same way as the other constraints in order to refine the output configuration from an initial pseudorandom aggregation.

4. APPLICATION

Two very different application areas have been selected for this study, to demonstrate the application of the zone matching approach in both urban and rural UK contexts. The urban example is the City of Southampton, and the rural the County of Pembrokeshire.

Southampton is a medium-sized city (1991 population 197,000) on the South Coast of England which at the time of the 1991 Census was divided into 417 enumeration districts (EDs) nested within 15 wards. The Pembrokeshire study area represents a remote rural region, with small settlements and extensive areas of sparsely populated land. Unlike the major metropolitan areas, rural areas in England and Wales have a further tier of local boundaries known as communities in Wales and parishes in England, which are to be respected in the construction of 2001 Census output areas (OAs). The study area is slightly less than the entire county, being that part for which the Office for National Statistics (ONS) has created prototype 2001 OAs. It contained 308 1991 EDs, although it is important to distinguish here between EDs/OAs and polygons, as the county includes a number of offshore islands both populated and unpopulated, which are represented by polygons in the dataset.

As part of preparation for the 2001 census, unit postcode polygons have been created around address locations by ONS for various test areas, including Southampton and Pembrokeshire. Thiessen polygons have been generated around each address, taking into account some additional topographic information, and boundaries dissolved between address polygons having a common postcode, in order to create a set of synthetic boundaries for the smallest units in the postal geography. These are based on 95,011 address locations in Southampton and 50,348 in Pembrokeshire. Using an implementation of the AZP algorithm, the postcodes have been grouped into 762 and 353 prototype OAs respectively, using a confidentiality threshold of 100 persons and 40 households, and a target population of 250. A simple tenure-based measure of homogeneity and square of perimeter divided by area constraints were applied in the production of these areas. Detailed 2001 OA design considerations are discussed more fully in Martin et al. (forthcoming). The OA boundaries resulting from this combination of constraints are indicative of what might be released to UK census users following publication of the 2001 Census outputs in the spring of 2003.

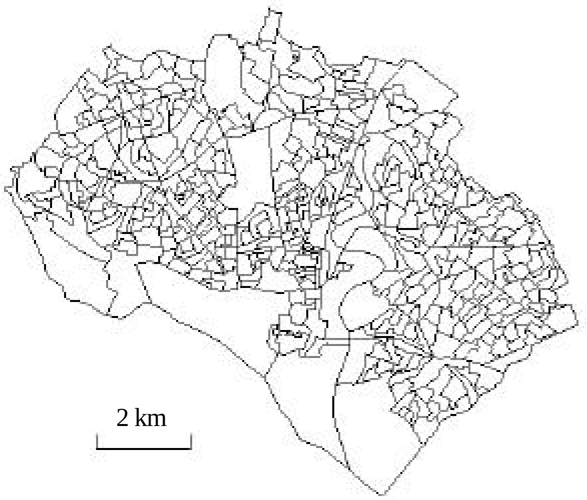

Figure 7: 1991 Census enumeration districts for the City of Southampton

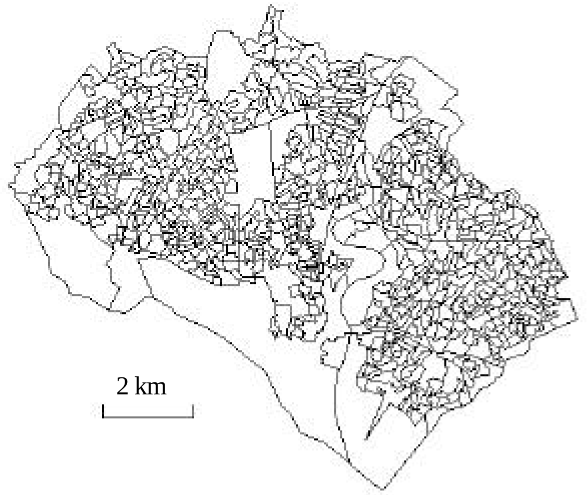

Figure 8: Prototype 2001 output areas for the City of Southampton

All GIS-based boundary manipulation has been performed within Arc/Info. The 1991 ED and 2001 OA boundaries have been intersected, sliver polygons merged with the neighbour with which they share the longest common boundary, and addresses counted within each intersection polygon. This results in intersection coverages containing 1771 polygons in Southampton and 1022 in Pembrokeshire, in which each polygon retains identifiers from both the ED and OA input layers. Polygon and arc attribute tables have been exported and used within the AZM program described above. In the following experiments 1999 address counts are used in lieu of actual 2001 population counts that are not yet known. Figures 7 and 8 show the 1991 EDs and 2001 OAs for Southampton, and Figures 9 and 10 show the corresponding EDs and OAs for the Pembrokeshire study area.

These datasets have been used to investigate a series of alternative zoning scenarios. In each of these the OA boundaries are fixed in order to reproduce the situation that will face users of the 2001 Census data when they are published. More experiments have been conducted with the Southampton data, as this represents a situation of considerably more complex boundary change between 1991 and 2001.

In the first scenario, both the 1991 ED and 2001 OA boundaries are fixed, and test areas are divided into tracts which can be exactly constructed from either of the input geographies: these are equivalent of the 1971-81 tracts described by Morgan and Denham (1982). This is the equivalent of the zoning problem illustrated in Figure 5. No iteration is required but the membership of each intersection zone must be examined, and a series of aggregations undertaken until tracts have been achieved which can be constructed by exact aggregation from either input geography.

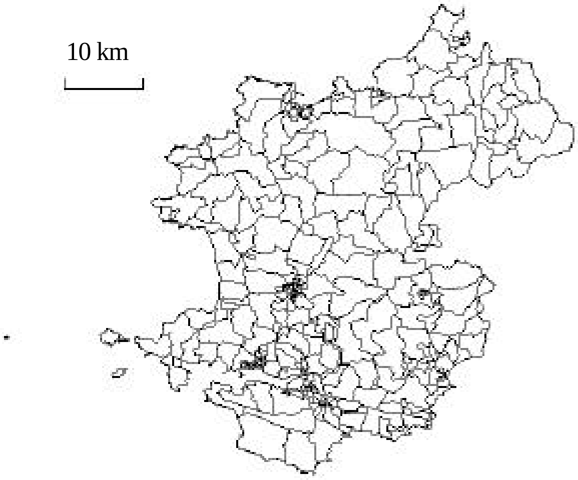

Figure 9: 1991 Census enumeration districts for the Pembrokeshire study area

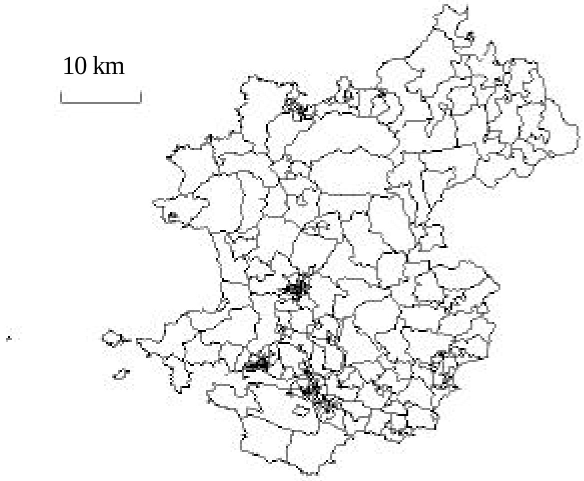

Figure 10: Prototype 2001 output areas for the Pembrokeshire study area

A series of further solutions are then sought at three different scales, corresponding approximately with the ward, the 1971-81 tract (the only empirical precedent for such an exercise in England and Wales) and the ED. In each case, the 2001 OA geography is fixed, but the solution may cut across the 1991 EDs, producing the equivalent problem to that introduced in Figure 6. At each scale the AZM algorithm is run firstly with only population and shape controls set, then with population stress added as a design constraint and finally with a computed low stress solution as the initial configuration. 250 iterations have been run in each case The scales of analysis have been set by taking the mean address count for Southampton wards (6333), 1971-81 tracts (4000) and EDs (228), and setting these as the target populations. A final option is to run the computed low stress solution without setting minimum or target populations and attempt to create a configuration in which there is one output tract corresponding to each zone of the approximated geography.

5. RESULTS

The results of the detailed experiments are shown in Tables 1 and 2. Table 1 shows the results of the full range of zoning scenarios for Southampton and Table 2 shows the results of only two experiments in Pembrokeshire. In the latter case it is much easier to achieve matches due to the higher degree of coincidence between ED and OA boundaries: this is due to the fact that in many cases both the ED and OA boundaries have actually been drawn to coincide with community boundaries. The first four table columns describe the zoning constraints, being a description of the scenario used, the minimum and target populations set and whether the shape constraint was applied. In the descriptions, A indicates that the population stress measure has been applied to maximize the approximation, and I indicates that the IRA has been replaced with a computed match configuration. The last six columns describe the characteristics of the output geography, namely the number of tracts created, their mean population and standard deviation, and the percentages of approximated EDs preserved without splitting, and of population correctly assigned into a tract which matches between the two geographies.

Table 1: Summary characteristics of Southampton tract geographies

Scenario	Min	Target	Shape	Tracts	Mean	St Dev	%EDs	% Pop
ED fixed	N/A	N/A	Off	14	6787	4024	100	100
Ward scale	5000	6333	On	15	6334	392	44	84
Ward scale A	5000	6333	On	15	6334	459	51	86
Ward scale A,I	5000	6333	On	15	6334	897	52	89
Tract scale	3500	4000	On	23	4131	314	41	83
Tract scale A	3500	4000	On	23	4131	384	39	84
Tract scale A,I	3500	4000	On	21	4524	599	47	89
ED scale	200	228	On	327	291	76	3	59
ED scale A	200	228	On	323	294	76	3	59
ED scale A,I	200	228	On	294	323	102	11	76
Smallest possible	N/A	N/A	Off	405	235	137	7	67

Table 2: Summary characteristics of Pembrokeshire tract geographies

Scenario	Min	Target	Shape	Tracts	Mean	St Dev	%EDs	% Pop
ED fixed	N/A	N/A	Off	82	614	539	100	100
Smallest possible	N/A	N/A	Off	233	216	129	30	77

The simplest task is the intersection of 2001 OAs and 1991 EDs with both geographies fixed, represented by the first row in each table. In this case, we are seeking the smallest set of output tracts that may be created precisely from both geographies. This processing is analytical and involves no iteration cycle, so the population and shape constraints are redundant. In Southampton, the large degree of boundary change from 1991 to 2001 leads to a subdivision of the city into 14 tracts with a mean address count of 6787, which are broadly equivalent to the 15 wards, although some boundary change and development causes two pairs of wards to be merged and one small ‘island’ zone to be formed. This solution is illustrated in Figure 11. The equivalent scenario in Pembrokeshire is shown as Figure 12. Here, the community boundaries are common to many 1991 EDs and 2001 OAs, allowing 82 tracts to be produced with a mean address count of 614. Due to this high degree of exact matching, the full range of alternative design scenarios is only applied to the Southampton test area.

Table 1 clearly reveals that poorer results are obtained as the scale of the output tracts is reduced, with great difficulty encountered when the target size is small in relation to the OA size. The final row in both tables 1 and 2 shows the results when no threshold or target are set and the shape constraint is not applied, but the initial configuration is simply computed so as to assign the population of each ED into the tract with which it has the greatest population overlap. In both the Southampton and Pembrokeshire contexts, there is still some loss of zones, with fewer tracts returned in each case compared to the corresponding number of EDs. Only 7% of EDs remain unsplit in Southampton and 30% in Pembrokeshire under this scenario.

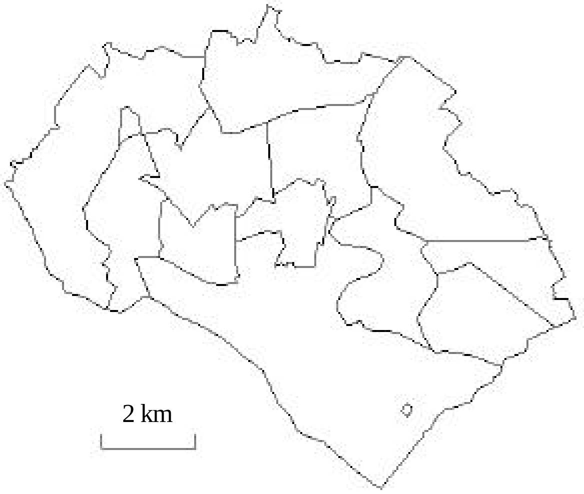

Figure 11: Exact aggregation of Southampton 1991 EDs and prototype 2001 OAs into tracts

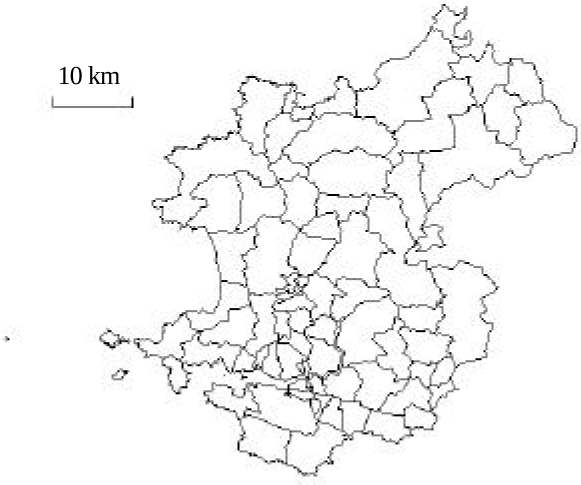

Figure 12: Exact aggregation of Pembrokeshire 1991 EDs and prototype 2001 OAs into tracts

It is apparent that at all levels of aggregation the proportion of population correctly preserved is much higher than that of EDs which can be preserved intact in the output solution. Including population stress in competition with the population and shape constraints generally results in only small overall improvements to the match, possibly because it is only one of four equally- weighted competing factors and therefore receives too low an emphasis in situations where matching is really the users' intention.

At each geographical scale of aggregation, the best matching results are obtained when a non-random initial configuration is used to maximize the match between the two input geographies. This approach effectively forces the algorithm to begin with the best-matching number of OAs to EDs, and the additional constraints are used only to make minor refinements to this configuration. At both the ward and tract scales, it is possible to achieve an 89% match in the populations – in other words a mean of 89% of the population of each 1991 ED is correctly carried through to the grouping of 2001 OAs to which it is assigned.

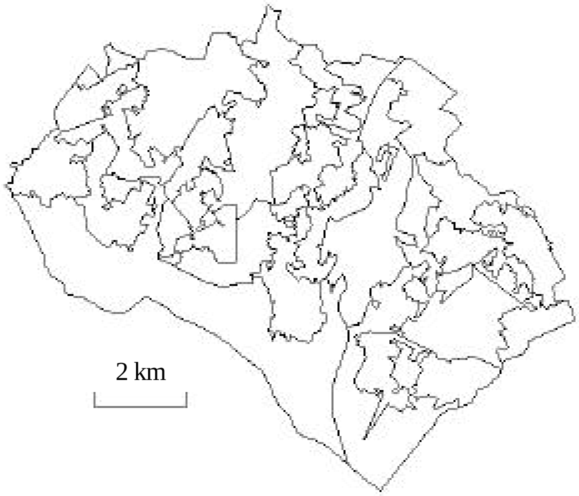

Figure 13: Approximate aggregation of Southampton OAs into tracts giving 89% population match with EDs

The 21 tract solution is presented as Figure 13, which illustrates the irregularity of shape which is necessary to achieve this level of matching, although this will have been increased by the relatively high population threshold in proportion to the target population. Even this approach does not guarantee a single OA to match every ED as some may be split between a number of OAs, none of which contributes the majority of its population to that ED, and others may be below the population thresholds set for this exercise. When an IRA is used, the algorithm proceeds to refine this by considering all the current constraints in parallel, and this has the effect of trading off target population size, shape and population stress. The most appropriate choice will depend on the importance to the user of achieving maximal matching between the geographies. There will be many, particularly urban, situations in England and Wales in which a high degree of match is only possible at the cost of significant irregularity of shape and/or broad variations in the population of the tracts. Generally, the degree of matching possible improves with target population size, making the approximation areas at the scale of the ward and above (either wards from a previous census, wards defined subsequent to the census, or non-census administrative areas) feasible with relatively low levels of approximation using this methodology.

6. CONCLUSION

This paper has reviewed the problem of matching population data between incompatible areal units, frequently encountered when census and administrative geographies change over time. The problem of maximising the goodness of fit of one zonal system to another has been characterised as one of minimizing population stress, and a stress statistics has been proposed. In certain situations, such as those in which both zonal geographies must be matched perfectly, the problem is one that may be solved analytically, but where it is necessary to approximate one geography by precise aggregations of another then the problem is more appropriately tackled with an iterative zoning procedure. Openshaw's (1977) AZP has been extended into an automated zone matching (AZM) algorithm and a program written for its implementation, which is available for other researchers to use.

A series of practical trials have been illustrated using the 1991 and 2001 Census geographies for Southampton and Pembrokeshire. These illustrate that (for UK-specific historical reasons) it will be much easier to achieve exact census tracts in rural areas where both 1991 and 2001 geographies have been drawn with reference to a common set of community and parish boundaries. In urban areas, the problem is complex, and there are no easy solutions or perfect matches but AZM offers an automated approach to the matching problem and allows an evaluation of the trade-off between resolution and precision in the creation of approximate tracts. The implementation described here will offer the possibility of very rapid identification of the smallest exactly matching zones, avoiding the significant manual effort or indirect calculations characteristic of previous attempts to devise intercensal tracts.

ACKNOWLEDGEMENT

The research described in this paper was undertaken while the author was a visiting researcher at Census Division, Office for National Statistics, Titchfield. The assistance of Steve King and members of the 2001 Census Output Production team at ONS are gratefully acknowledged. All views and opinions expressed in the paper remain the responsibility of the author. The maps presented here have been developed through the prototype 2001 Census Output Area Production System and are Crown Copyright.

REFERENCES

Atkins, D., Charlton, C., Dorling, D. and Wymer, C. (1993) *Connecting the 1981 and 1991 Censuses.* Research Report 93/9, NE.RRL, University of Newcastle: Newcastle-upon-Tyne

Bloom, L. M., Pedler, P. J. and Wragg, G. E. (1996) "Implementation of enhanced areal interpolation using MapInfo", *Computers and Geosciences*, vol 22, pp. 459-466.

Duke-Williams, O. and Rees, P. (1998) "Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure", *International Journal of Geographical Information Science*, vol. 12, pp. 579-605.

Fisher, P. F. and Langford, M. (1995) "Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation", *Environment and Planning A*, vol. 27, pp. 211-224.

Flowerdew, R. and Green, M. (1991) "Data integration: statistical methods for transferring data between zonal systems", in: Masser, I. and Blakemore, M. (Eds.) *Handling geographic information: methodology and potential applications*, Longman: Harlow, pp. 38-54

Flowerdew, R. and Green, M. (1992) "Developments in areal interpolation methods and GIS", *Annals of Regional Science*, vol. 26, pp. 67-78.

Goodchild, M. F., Anselin, L. and Deichmann, U. (1993) "A framework for the areal interpolation of socioeconomic data", *Environment and Planning A*, vol. 25, pp. 383-397.

Haining, R., Wise, S. and Ma, J. (1998) "Exploratory spatial data analysis in a geographic information system environment", *Journal of the Royal Statistical Society D*, vol. 47, pp. 457-469.

Haining, R., Wise, S. and Ma, J. (2001) "Providing spatial statistical data analysis functionality for the GIS user: the SAGE project", *International Journal of Geographical Information Science*, vol. 15, pp. 239-254.

Horn, M. E. T. (1995) "Solution techniques for large regional partitioning problems", *Geographical Analysis*, vol. 27, pp. 230-248.

Langford, M., Maguire, D. J. and Unwin, D. J. (1994) "The areal interpolation problem: estimating population using remote sensing in a GIS framework", in: Masser, I. and Blakemore, M. (eds) *Handling geographic information: methodology and potential applications* Longman: Harlow, pp. 55-77.

Martin, D. (1997) "From enumeration districts to output areas: experiments in the automated creation of a census output geography", *Population Trends*, vol. 88, pp. 36-42.

Martin, D. (1998) "Optimizing census geography: the separation of collection and output geographies", *International Journal of Geographical Information Science*, vol. 12, pp. 673-685.

Martin, D., Nolan, A. and Tranmer, M. (forthcoming) "The application of zone design methodology in the 2001 UK census", submitted to *Environment and Planning A*

Mehrotra, A., Johnson, E. L. and Nemhauser, G. L. (1998) "An optimization based heuristic for political districting", *Management Science*, vol. 44, pp. 1100-1114.

Morgan, C. and Denham, C. (1982) "Census small area statistics (SAS): measuring change and spatial variation" *Population Trends*, vol. 28, pp. 12-17.

Mugglin, A. S., Carlin, B. P. and Zhu, L. (1999) "Bayesian areal interpolation, estimation and smoothing: an inferential approach for geographic information systems", *Environment and Planning A*, vol. 31, pp. 1337-1352.

Mugglin, A. S., Carlin, B. P. and Gelfand, A. E. (2000) “Fully model-based approaches for spatially misaligned data”, *Journal of the American Statistical Association*, vol. 95, pp. 877-887.

ONS/GROS/NISRA (1999) *2001 Census: A guide to the one number census*. Accessed 31 May 2000 from: http://www.statistics.gov.uk/census2001/pdfs/onc.pdf

Openshaw, S. (1984) *The modifiable areal unit problem.* Concepts and Techniques in Modern Geography 38, Geo Books: Norwich

Openshaw, S. (1977) “A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling”, *Transactions of the Institute of British Geographers* NS, vol 2, pp. 459-72.

Openshaw, S. and Alvanides, S. (1999) “Applying geocomputation to the analysis of spatial distributions”, in: Longley, P. A., Goodchild, M. F., Maguire, D. J. and Rhind, D. W. (Eds.) *Geographical Information Systems: Principles, Techniques, Applications and Management.* Wiley: Chichester, vol. 1, pp. 267-282.

Openshaw, S. and Rao, L. (1995) “Algorithms for reengineering 1991 Census geography”, *Environment and Planning A*, vol. 27, pp. 425-446.

Tranmer, M. and Steel, D. G. (1998) “Using census data to investigate the causes of the ecological fallacy”, *Environment and Planning A*, vol. 30, pp. 817-831.

Xie, Y. C. (1995) “The overlaid network algorithms for areal interpolation problem”, *Computers, Environment and Urban Systems*, vol. 19, pp. 287-306.