

Geography for the 2001 Census in England and Wales

1. Introduction

Census geography describes the subdivision of the country into geographical areas for the purposes of the census. There are two major requirements of a census geography system: firstly, that it should facilitate the organization and management of the census itself and secondly, that it should provide an appropriate framework for the publication of small area census statistics which meet users' needs. Examination of these two requirements reveals that the needs of enumeration are often very different from those of data publication. The 1991 Census used a single hierarchical system of geographical areas, the enumeration districts (EDs), for both data collection and publication and these were created manually, with census office staff drawing ED boundaries onto photocopied large scale Ordnance Survey (OS) maps according to a set of predefined design principles. For reference, a description of the 1991 geography system may be found in Clark and Thomas (1990), and more general 1991 procedures and outputs are covered in Dale and Marsh (1993). The period surrounding the 1991 Census was one of increasing interest in the use of geographical information systems (GIS) by census users, and ED boundary data were subsequently digitized by the ED-Line consortium and by GDC Ltd to create ED-Line and ED91 products respectively. Nevertheless the actual zone design process was entirely manual.

The 2001 Census represents a very different environment in which to undertake geography design, and a number of major innovations have taken place. These include the separation of collection (ED) and output geography; the incorporation of postcode geography into census output geography and the use of GIS and automated zone design procedures. These developments in geography have also occurred alongside major innovations in other aspects of census processing, in particular One Number Census (ONC) procedures. ONC refers to the use of a very large census coverage survey in order to estimate levels of under-enumeration and impute missing individuals and households into the census database before the creation of the published tables. Again, ONC details are not the purpose of this briefing, and readers requiring more detail are referred to *A guide to the One Number Census*, available from the Census organizations.

The purpose of this paper is to provide an overview of 2001 Census geography systems in England and Wales, concentrating primarily on the creation of output geography, to accompany the Census Geography Roadshows held in November and December 2001. It is important to note that there are significant differences between the systems being implemented by ONS in England and Wales and those which apply to Northern Ireland and Scotland. The Northern Ireland system is broadly similar to that described here, whereas the entire data infrastructure and OA design process in Scotland are different, and interested users are referred to the list of contacts at Annex A.

2.1 Enumeration district design: principles and practice

The principal consideration of ED design is to create geographical areas which facilitate efficient and accurate distribution and collection of census forms by enumerators, while attempting to equalize enumerator workload. This requires knowledge of the location of residential addresses, together with some understanding of the likely difficulty of enumeration. Factors particularly increasing enumerator workload include sub-divided properties such as bedsits; flats in which individual front doors are protected by entry-phone systems; residents who may not have English as their first language etc. Relevant physical considerations also include the density of housing: in rural areas, enumerators may have to travel long distances between addresses, reducing the number of households which can be enumerated within a standard workload. In planning for both 1991 and 2001 Censuses, ONS have attempted to trade-off the difficulty of enumeration resulting from such factors with the size of the ED. Further, it is necessary to place ED boundaries in such a way that enumerators have responsibility for areas which 'make sense' on the ground. This generally requires that EDs do not straddle major roads, rivers, railway lines or extensive areas of open space.

A further consideration has been to ensure that EDs nest within wards and parishes/communities (an important consideration in previous censuses when EDs were themselves used as output areas) as it is a requirement that aggregated census statistics provide exact counts for these areas. Parishes do not exist in all areas of England, being mainly present in rural areas. For census purposes, communities are the Welsh equivalent of parishes, representing a local level of administrative geography generally below that of the ward. A major consideration in any new census geography is one of incorporating all the changes that have taken place to these higher level boundaries. This combination of following higher-level areal boundaries while attempting to standardize workloads results in some highly irregular geographical sizes and shapes, as it is simply not possible to neatly reconcile these requirements that frequently work against one another.

Particular difficulties are faced when an area is undergoing significant residential redevelopment at the time of the census, and attempts must also be made to take into account any anticipated changes to the residential structure. This information has usually been sought from local authorities, with further changes necessary when enumerators first go into the field, and discover changes to housing which were not identified as part of the ED design process. The resulting EDs thus represent an attempt to reach the best 'trade-off' between the many competing design considerations.

In 1991 all these factors were taken into account by the ED planners working manually with paper maps, resulting in 116,919 EDs in England and Wales, plus 4,840 'special' EDs which were identified as without geographical areas, usually large communal establishments such as prisons or long-stay hospitals. Once enumeration was complete, a small number of EDs were found to have populations which fell below the confidentiality thresholds for publication of 50 persons and 16 households. These EDs were then identified

as 'restricted' EDs and all counts except the basic person and household totals were 'exported' to a nearby ED(s) to produce areas with above-threshold counts for which data could be published. The presence of restricted EDs within the ED geography is one of the inherent disadvantages of using a predesigned (collection) geography for data publication, as it is never possible to know with complete certainty which areas will turn out to be below the threshold once the data have been tabulated.

Another major user concern with the 1991 ED geography has been that it failed to take any account of postcode geography. Unit postcodes and address geographies have seen increasing use through the 1990s as the geographical referencing system of choice for many non-census applications, and the imprecise nature of the association between 1991 EDs and postcodes continues to be a frustration. This issue was widely discussed as early as the 1987 report of the government's Committee of Enquiry into the Handling of Geographic Information (Department of the Environment, 1987), but a major obstacle was the absence of any definitive (or digital) boundaries for the 1.7 million unit postcodes, the lowest level of the postcode geography, each typically containing around 14 addresses. The cost of postcode boundary creation as part of 1991 Census processing was considered prohibitive and indeed no conventionally digitized boundaries have been produced at the unit level to date. The case for closer integration of census and postcode geographies is summarized by Dugmore (1996), and the relationship between 1991 EDs and unit postcodes is recorded in the directory of EDs and postcodes (OPCS/GROS, 1992). The situation is very different in Scotland, where manually digitized unit postcode boundaries existed and were used in 1991 Census geography design. These postcode polygons have been continually maintained, making the creation of postcode-based census geographies a more obviously attractive and simpler exercise than elsewhere in the UK.

Other characteristics of 1991 EDs which make them less than ideal as data publication areas include the wide variations in population size which result from the attempt to standardize enumerator workloads, and their perceived lack of social homogeneity. Although a subsidiary design consideration was to change as few ED boundaries as possible, the combination of all the above considerations resulted in 68% of EDs being redefined between the 1981 and 1991 Censuses.

2.2 ED design for 2001

Design of EDs for 2001 has been implemented using a GIS-assisted methodology, in which ED planners at ONS have undertaken a design exercise very similar to the 1991 approach described above, but working within an entirely digital environment. The paper background mapping has been replaced by OS Land-Line and digital raster mapping, together with ADDRESS-POINT, which provides locations of all addresses with a spatial resolution of 0.1m (addresses being identified from the intersection of Royal Mail's Postcode Address File PAF and OS large scale mapping products).

ADDRESS-POINT also contains a certain amount of advance information on planned development, although generally with less precise locational referencing. The parish and ward geography to be respected is obtained from OS Boundary-Line, and the ED-Line representation of 1991 geography used as a starting point. The absence of any definitive unit postcode boundaries at the time of ED design meant that it was still not possible to incorporate postcode geography into the process.

ED planning using this methodology was undertaken from April 1999 until autumn 2000, resulting in a new digital ED database containing 116,897 EDs. ED design was completed in less time and by fewer planners than in 1991. Once the entire country was complete, certain areas were re-planned using the latest available ADDRESS-POINT data enhanced by additional addresses discovered during enumeration and processing, so as to ensure that the ED geography reflected the actual residential structure on census day as accurately as possible. An innovation resulting from this approach has been that it was possible to issue enumerators with appropriately scaled single-sheet ED maps (illustrated in Figure 1) and also to pre-print record books with those addresses known from ADDRESS-POINT which fell within the ED, (illustrated in Figure 2).

Figure 1: Sample enumerator's ED map

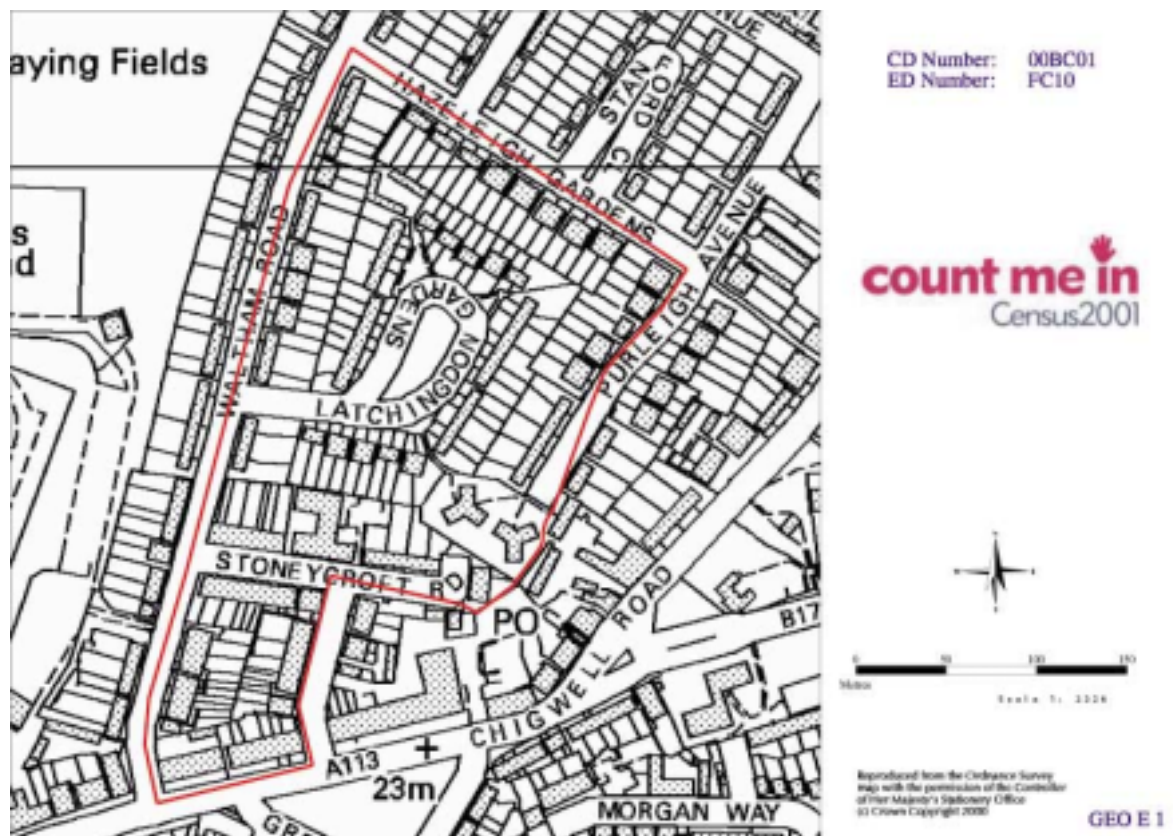


Figure 2: Sample page from enumerator's record book

Part 1

Form No.	Address (including Postcode)	Advance Record	NH	DEM	CE	Name of householder or CE Manager Person in charge	DEL	COLL REC'D	Dummy Form Information			Notes
									Count	No. of forms	Source	
Buildings	(Location or Description of Premises as required)					No. of persons in 1991						
A	B	C	D	E	F	G	H	I	J	K	L	M
0001	1 BRIDGE COTTAGES, FARNHAM ROAD, LISS, GU33 6LA			/					/	/		
0002	11 WINDBARN COTTAGES, FARNHAM ROAD, LISS, GU33 6LD			/					/	/		
0003	23 WIND COTTAGES, FARNHAM ROAD, LISS, GU33 6LA			/					/	/		
0004	7 HAMBARN COTTAGES, FARNHAM ROAD, LISS, GU33 6LB			/					/	/		
0005	5 BRIDGE COTTAGES, FARNHAM ROAD, LISS, GU33 6LA			/					/	/		
0006	31 WINDBARN COTTAGES, FARNHAM ROAD, LISS, GU33 6LD			/					/	/		
0007	4 BRIDGE COTTAGES, FARNHAM ROAD, LISS, GU33 6LA			/					/	/		
0008	4 HAMBARN COTTAGES, FARNHAM ROAD, LISS, GU33 6LB			/					/	/		
0009	5 BRIDGE COTTAGES, FARNHAM ROAD, LISS, GU33 6LA			/					/	/		
0010	6 BRIDGE COTTAGES, FARNHAM ROAD, LISS, GU33 6LA			/					/	/		
PAGE TOTALS				/					/	/		

SAMPLE

Page 1 of 14 P11402ENR/T

3. Output area design

Following on from prototype work undertaken at the time of the 1997 Census Test and reported in Martin (1997; 1998), the decision was taken to create separate geographies for data collection and output. As described in section 2.2 above, ED design proceeded within a GIS environment to create a digital representation of the ED geography from which enumerators' maps and address lists were generated. Although the cost and complexity of setting ED planners to go back over the entire country drawing up another separate geography according to output considerations are prohibitive, advances in computing power, digital data infrastructure and the use of zone design algorithms have made possible the creation of a separate output geography by automated means. Output geography design is based on the same GIS database that has already been created for ED design.

Reference to section 2.1 above reveals some of the problems inherent in the 1991 model, particularly the lack of integration with postcode geography and the pre-enumeration geography design which is therefore prone to the creation of restricted EDs and a relatively weak reflection of small area social homogeneity. The creation of a separate set of output areas (OAs) after 2001 enumeration opens up the possibility of addressing each of these problems to some degree, and of producing a new geographical division of the country designed explicitly for the purposes of census data publication.

3.1 Design principles

One of the first principles of any attempt to create a purpose-specific output geography is that the OAs should all be above the required population and household thresholds, thus avoiding the 1991 problem of restricted EDs with their consequent 'exporting' and 'importing' of population counts. A second significant issue is that it should be possible to construct OAs which are as far as possible constructed from whole unit postcodes, thus facilitating far better integration between geographical information referenced by census and postcode geographies. Additional, but nevertheless important, considerations then include the standardization of OA population sizes, maximization of internal social homogeneity and some control over the more irregular geographical shapes of the areas.

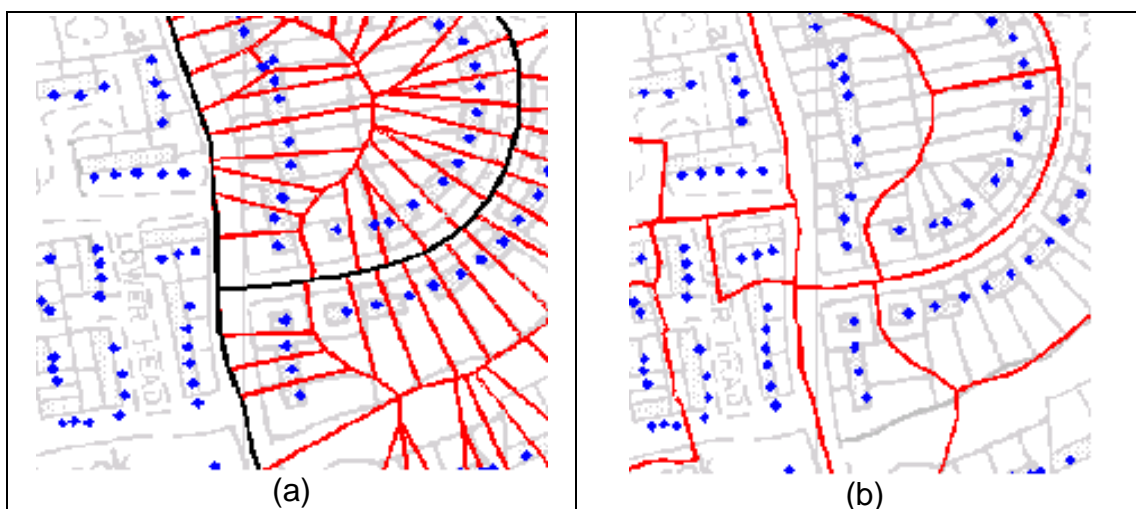
3.2 Unit postcode polygons

A prerequisite for the creation of OAs that respect unit postcodes is that there should be a national coverage of unit postcode boundaries. Throughout the intercensal period there has been discussion of this need within the UK geographic information community, but no such boundary set existed when it became necessary to choose and implement 2001 Census geography systems. Not only would any such boundary set inevitably be very large, but it will be subject to relatively heavy maintenance requirements due to the continual small-scale changes to the postcode system. Postcodes are created and used by Royal Mail for the prime purpose of speeding the sorting and delivery of mail, and there are therefore numerous minor complexities which have deterred any organization from creating a maintained set of postcode boundaries. Nevertheless, the advent of ADDRESS-POINT as a fully georeferenced, postcoded address list makes this problem too amenable to geographical computation rather than manual digitizing.

For the purposes of 2001 Census OA design, a complete set of unit postcode polygons is being created, using Thiessen polygons, ADDRESS-POINT and ancillary topographic data. Thiessen polygon creation is a standard GIS function, allowing the creation of space-filling polygons around a point dataset such that each polygon encloses the space which is closer to its own point than to any other. Creation of Thiessen polygons around address locations, as illustrated in the right hand side of Figure 3(a), producing a coverage of small polygons, each containing a single address. This polygon boundary set is then intersected with a series of topographic data layers such as principal roads and waterways, and also with the ward and parish boundaries which must be respected in the eventual output geography. The internal boundaries between any polygons sharing the same unit postcode in these intersected layers are dissolved, resulting in a set of synthetic unit postcode polygons, such as those shown in Figure 3(b). Situations exist in which unit postcodes are either spatially coincident (on separate floors of the same building) or overlapping, particularly in mixed residential/commercial neighbourhoods. In

these cases, more than one postcode is assigned into the same polygon, a situation referred to in 2001 geography terminology as 'stacked' postcodes. Following enumeration and imputation procedures, census data from individual forms are associated with addresses, and then aggregated for each postcode polygon. These polygons then form the basic building blocks for census OA creation. They must be configured in such a way that they nest exactly within parish or ward boundaries, and the external boundary that must be precisely matched is referred to here as a constraining polygon. A small proportion of unit postcodes straddle constraining polygon boundaries and must therefore be split for the purposes of census output, but these whole or split postcodes are here all termed postcode polygons.

Figure 3: (a) Thiessen polygons created around address locations; (b) Thiessen polygons merged to create unit postcode polygons



3.3 Automated zone design applied to census output areas

There are many different ways in which unit postcode polygons could be arranged within a given constraining polygon boundary, and the design problem is much the same as that of political districting, in which polygons must be created which represent a satisfactory trade-off between various considerations including population size and geographical shape. Several computational approaches to this kind of zoning problem exist in the academic literature, and 2001 OA design will make use of the algorithm suggested by Openshaw (1977) known as the automated zoning procedure (AZP). Application of this methodology to published census data has been demonstrated in the context of the 'reengineering' of 1991 UK Census outputs by using enumeration districts (EDs) as input building blocks, and assembling these into larger zones which may be made to display a variety of target characteristics by Openshaw and Rao (1995), in a paper which also summarizes some of the alternative algorithms available.

The AZP algorithm makes use of the contiguity information available from the GIS containing the unit postcode polygons. It begins by estimating the

approximate number of OAs that should fall within a constraining polygon, given an input population target size, and then randomly aggregating adjacent postcodes to form above-threshold OAs. A number of statistical measures for this initial configuration are computed for each of the selected design constraints. Overall distance from target population is measured by the sum of the squared differences between OA populations and the target population size. The measurement of shape and social homogeneity are discussed separately below. Consideration is then given to the swapping of postcode polygons between adjacent OAs in terms of their impact on these statistical measures.

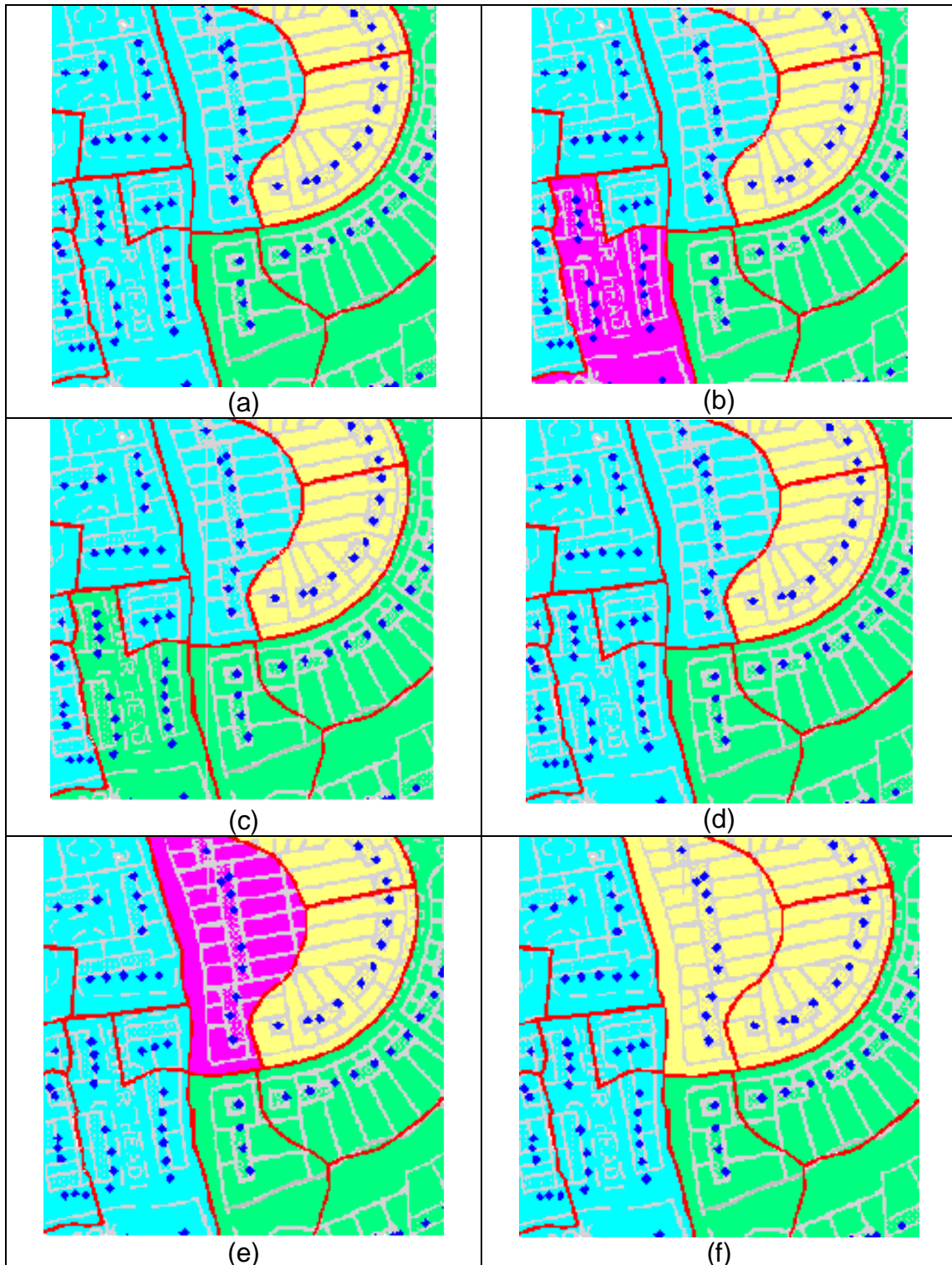
For example, regarding population size, an improving swap will be one that reduces the total squared difference from target size by bringing an above target and below target OA closer to the desired size. Any swaps which serve to improve the overall solution in this way are accepted and incorporated into the emerging OA geography, while any that cause deterioration in the objective criteria are rejected. This approach is illustrated in Figure 4. In Figure 4(a) an area is shown in which the postcode polygons have been grouped into three prototype OAs, indicated by the different shading. In Figure 4(b), one postcode polygon is selected for potential swapping into a neighbouring OA resulting in Figure 4(c). The overall quality of this configuration is assessed and found to be unsatisfactory, so the algorithm reverts to the original situation in Figure 4(d). Another postcode polygon is identified as a potential swap in Figure 4(e) and this time results in an overall improvement, leading to its retention within the current best solution, as shown in Figure 4(f). No swaps are permitted which would produce sub-threshold OAs or break the internal contiguity of an OA. Once all available combinations have been tested, the design process is recommenced using a different random starting configuration, and the overall best solution is chosen after a preselected number of iterations.

The resulting OA geography comprises complete unit postcodes, except where these are subdivided by constraining polygon boundaries. The final configuration represents a calculated trade-off of the various design objectives which are specified at the outset. While unable to provide a perfect solution to all the conflicting output requirements (an impossibility), the algorithm produces an approach to the identification of 'optimal' solutions which can be applied consistently across the entire country, and which represents significant improvements over the 1991 ED model in most respects. OAs produced in this way are above the specified population threshold, avoiding the need for restriction and importing/exporting of population counts, and are smaller and more homogeneous than 1991 EDs, making them a more flexible set of building blocks for census analysis.

An exception to the requirement to meet population thresholds occurs where the constraining polygons are themselves below the required thresholds, in which case there is no logical solution to the problem. 527 parishes fell below the population threshold of 50 persons in 1991. The thresholds to be used in 2001 have now been set at 100 persons and 40 households. (Based on the

estimates made during 2001 ED planning, 1,402 EDs fall below these thresholds.)

Figure 4: Illustrating the iterative swapping of postcode polygons between prototype output areas by the AZP algorithm



3.3.1 Measuring homogeneity

An area which has provoked considerable interest at consultation meetings with census users has been that of social homogeneity. This has assumed particular significance in the light of the use of Census OAs as building blocks for the release of sub-ward Neighbourhood Statistics proposed for 2003. A particular concern of many users has been that EDs from previous censuses have often obscured social divisions which would be of importance in the context of local service planning and resource allocation. In initial experiments with automated zone design for the 2001 census such as those reported in Martin (1997; 1998), a simple homogeneity measure was utilized, based on maximizing the uniformity of households falling within different categories. Tenure categories (owner occupied, privately rented, rented from local authority or housing association) within each OA were used in this early work. This measure helpfully introduced the concept of homogeneity but also has some weaknesses. If the dominant category (eg. owner-occupation) is the same across the entire area, then the algorithm simply attempts to maximize that category in all OAs. In many localities, there will be a fairly even spread of housing across several tenure categories, and the 'dominant' category may account for less than half of the total households. In the original ONS implementation, only tenure was used. What is really required is a more sophisticated measure that is able to take into account homogeneity among categories which may often never achieve absolute dominance within a given OA, and which provides for the meaningful combination of more than one multi-category variable.

Tranmer and Steel (1998) discuss the use of a statistic termed the intra-area correlation (IAC), which measures the similarity of values of variables within any area of interest. For example, if the variable is tenure, then the intra-area correlation measures the similarity of the values of this variable for each household in each OA. Although the theoretical maximum is 1.0, this will not be found in real-world census zones, and any value above about 0.05 implies a reasonable degree of homogeneity

IAC may be extended for use with variables that have K categories where K is greater than 2. For example tenure may have K=3 categories ('owner occupied', 'renting privately', 'renting from local authority or housing association'). For each category eg. 'owner occupied' a measure of homogeneity can be obtained using the intra-area correlation. This is calculated as:

$$\delta_k = \frac{\frac{1}{M-1} \sum_{g=1}^M N_g (P_{kg} - P_k)^2}{(\bar{N}^* - 1) P_k (1 - P_k)} - \frac{1}{(\bar{N}^* - 1)}$$

Where:

\bar{N}^* is the mean population size of the M areal units, with an adjustment to take into account variation in the population size of the areal units. Full details of the way in which this measure is derived are given in Tranmer and Steel (1998). In practice, \bar{N}^* is very close to \bar{N} .

N_g is the population size of areal unit g

P_K is the overall proportion of the population in category k, and

P_{kg} is the proportion in category k in areal unit g

The intra area correlation formula given here is, approximately, the ratio of the area level variance to the individual or household level variance, and this ratio is then divided by the mean area population size. Hence, this measure is relatively easy to calculate. Once we have calculated the intra area correlations, δ_k , for each category of the grouping variable, we can obtain an overall intra-area correlation measure, δ , which takes all categories of the grouping variable into account, using:

$$\delta = \frac{1}{K-1} \sum_{k=1}^K (1 - P_k) \delta_k$$

Maximising IAC may be thought of as configuring OAs in such a way that the largest proportion of the variation in the relevant variable(s) between different postcode building blocks occurs across OA boundaries, while the smallest proportion between postcodes inside the same OA. IAC provides a statistically valid measure that may be combined across different OA designs, and also compared with other zoning schemes with different target populations, such as EDs or alternative OA schemes. Previous work has identified dwelling type and tenure as the two variables that tend to experience the greatest degree of homogeneity – due to the structure of the built environment and its indirect reflection in property ownership patterns. The current proposal is to use four tenure categories and seven dwelling type categories which are then combined with equal weighting to produce the overall homogeneity measure. This combined IAC is compared in the light of each potential swap of postcode polygon between OAs. The categories are listed in Table 1. These are the two variables which other work on social homogeneity reports to play the greatest role in structuring neighbourhoods. Although it is tempting to include variables such as ethnic group, these have little discrimination in most areas – being present in non-standard mixtures in only a very small proportion of neighbourhoods nationally.

Table 1: Dwelling type and tenure categories used in intra-area correlation calculations

Dwelling type	Tenure
Owner-occupied	Detached
Rented privately	Semi-detached
LA/HA	Terraced
Other	Flat
	Part-house
	Commercial
	Non-permanent

3.3.2 Measuring shape

Initial work on the implementation of AZP for OA design used as a shape statistic the ratio of OA perimeter squared to area. Swaps were sought which minimized this ratio, the most compact shape of all (a circle) being that with the lowest perimeter in proportion to its area. Responses to consultation have shown that many census users would prefer OAs to respect discrete settlements where possible, rather than possess compact shapes. For this reason, an alternative shape control has been implemented which minimizes the distances between each postcode centroid and the mean of all postcode centroids in the OA. Postcode centroids represent the mean location of the addresses falling within a given postcode polygon, and are thus address-weighted spatial means. This approach has a beneficial effect in rural areas where the inclusion of a postcode polygon into a prototype OA is governed less by the compactness of the overall boundary shape which would result, than by the dispersion of the postcodes which it encompasses. In urban areas where population densities are higher and postcodes and OAs are relatively compact, the effect is less pronounced.

4. Matching 2001 output areas to other geographies

A perennial and fundamental problem with the census and all other administrative geographies is that of matching together data when the boundaries of the areal units are not coincident. This applies equally to the temporal comparison of data from two different censuses when the boundaries have changed, or the association of census data with some other areal units which have been independently constructed and which are not therefore assembled from the same elemental units.

Between the 1971 and 1981 Censuses, OPCS attempted to identify small areas whose external boundaries were unchanged for the Department of the Environment. This resulted in 48 300 new geographical areas called census tracts, each comprising aggregations of one or more 1971 and 1981 EDs that could be grouped to form areas with identical boundaries. These tracts were mainly in urban areas. Some small shifts in statutory boundaries did not result in any transfer of population, and some of these zones are therefore strictly

approximate rather than exact areas of comparability. Denham (1980) provides an illustration (p. 8) of both exact and approximate tract definitions with reference to a map of 1971 and 1981 ED and statutory boundaries. A further 10 700 parishes or communities were identified, primarily in rural areas, which had remained largely unchanged between the two censuses, although a small number of these contained large populations (over 10 000) and were subsequently subdivided to form smaller tracts.

Unfortunately, no equivalent exercise was undertaken in 1991, thus leaving census users without any directly comparable small area definitions between 1981 and 1991. Instead, a lookup table of 1991 EDs to 1981 wards was created by a team at the University of Newcastle (Atkins et al., 1993) using an approximate methodology based on a combination of existing lookup tables and GIS analysis of centroids and boundaries. A lookup table was produced which showed the estimated relationship between 1991 EDs and 1981 wards, based on the allocation of ED centroids from the small area statistics (SAS) into 1981 ward polygons, and also the creation of population-weighted ED centroids from the 1991 directory of EDs and postcodes. These alternative sources were subject to extensive checking and correction before a final allocation was produced. The decision to work back to 1981 wards was largely a reflection of the effort that had already been invested in the identification of 1971-1981 comparable areas, although it provided only a partial solution for 1991 users wanting to analyze change over time, or needing to use new 1991 ED-based geographies. One important consideration in relation to lookup tables is that, if mapped, they do not always produce spatially contiguous representations of one zonal system expressed in terms of another. For example, the 'best fitting' 1981 wards produced from 1991 EDs will not necessarily be single contiguous polygons.

Although a case can be made for using historical ED boundaries (eg. releasing 2001 data for 1991 boundaries so as to permit direct analysis of population change) or even regular grid squares (which are unchanging over time), there is no set of areal that meets all the conflicting demands of intercensal change, given the requirement to accommodate contemporary residential redevelopment and statutory boundary change.

In an academic context the usual approach to this problem has been to make use of areal interpolation— for which various algorithms exist, and the terminology of 'source' zones (for which data are held), 'target' zones (for which counts are required) and 'intermediate' zones (the intersection of source and target zones) is widely used. Most areal interpolation techniques assume that population is uniformly distributed within census areas, or at least within residential land uses, and that population counts can therefore be apportioned between overlapping zones in proportion to their geographical areas. Such an approach does not meet the needs of Census Offices that are required to produce exact or best-fitting census counts for a number of incompatible zonal systems. The address-based management of 2001 Census data would theoretically make possible the direct aggregation of address-level counts to any zonal system, but from the perspective of data output, the key problem is one of differencing (Duke-Williams and Rees,

1998). This refers to the possibility that if data are released for two very slightly different geographical areas, both of which are above the required population threshold, there is a possibility that the population revealed by subtracting one from the other will itself be sub-threshold and there is thus a risk of disclosure. Where the output geographical units are large, the probabilities associated with this risk are extremely small, but where for example two zones which are themselves small may differ by only a small amount, then the risk prevents the publication of precise aggregations for both zones. The only alternative is to implement some form of best-matching between small areas from the 1991 and 2001 censuses and to offer these to users for the purposes of intercensal comparison.

The AZP algorithm used for the creation of 2001 OAs offers one approach to the definition of such 'best matching tracts' because it is possible to set up the degree of match between two sets of zones as an objective function and to view the intersection of the two geographies to be matched (for example, 1991 EDs and 2001 OAs) as a set of building blocks. The advantage of such a procedure is that it produces computed best matches at a given scale of aggregation: a task which is even more challenging than ED design to accomplish by manual means. If 2001 OAs must be observed precisely, then only exact aggregations of OAs may be used in an attempt to achieve the best possible match (in population terms) with aggregations of 1991 EDs.

Prototype application of this approach to a set of prototype OAs for the City of Southampton is illustrated in Figures 5-7. Southampton comprises 417 EDs or 762 OAs within 15 wards. The move to unit postcodes as the basic building blocks means that there is little correspondence between detailed ED and OA boundaries, and Figure 5 shows the smallest areas which may be precisely aggregated from both geographies (exact tracts) – effectively representing the ward scale with a mean address count of 6 787, although there are some slivers in these prototype data which may be overcome in any production system. Choosing an intermediate zone size, approximately equivalent to that of the manually created tracts produced for 1981-1991 intercensal analysis (mean address count 4 524), results in a set of zones such as those displayed in Figure 6. These are assembled from whole 2001 OAs, and therefore could be aggregated directly from 2001 area statistics. At this scale 89% of the address locations fall within their best-matched tract. Due to the application of this process to whole OA data, users could potentially use such an approach to define their own tracts, or a set of standard 'best fit' tracts could be produced nationally. Further experiment and consultation is required in order to determine the size of tracts which would be of most interest to users. In general, quality of match decreases with size of zone, and improves to near-perfect at the ward or parish scale. For comparison, the same approach applied in a rural context achieves a far higher correspondence between the 1991 and 2001 geographies due to the mediating effect of parish boundaries which have acted as a constraint in both cases. In Pembrokeshire, 82 exact-matching 1991-2001 tracts can be created with a mean address count of 614.

A further application of these procedures is for the calculation of best matches between census and non-census geographies. Figure 7 illustrates the computed best match between prototype 2001 OAs and the 40 postcode sectors in Southampton (mean address count 2 375), which achieves a 90% match at the address level. For the corresponding postcode areas the address match is 97%, although these figures are to some degree dependent on which postcode sector boundaries are regarded as correct.

Figure 5: areas of exact match between 1991 EDs and prototype 2001 OAs within Southampton (scale bar: 2km)



Figure 6: Approximate aggregation of Southampton prototype OAs into tracts giving 89% population match with EDs (scale bar: 2km)



Figure 7: Best-matching aggregation of 2001 prototype OAs to postcode sectors in Southampton (scale bar: 2km)



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
- Clark, A. M. and Thomas, F. G. (1990) *The geography of the 1991 Census* *Population Trends* 60, 9-15
- Dale, A. and Marsh, C. (1993) *The 1991 Census User's Guide* HMSO, London
- Denham C. (1980) The geography of the census 1971 and 1981 *Population Trends* 19, 6-12
- Department of the Environment (1987) *Handling Geographic Information: The Report of the Committee of Enquiry Chaired by Lord Chorley* HMSO, London
- Dugmore, K. (1996) What do users want from the 2001 Census? In: *Looking towards the 2001 Census* OPCS Occasional Paper 46 OPCS, London 21-3
- 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* 12, 579-605
- 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 2, 459-72
- Openshaw S. and Rao L. (1995) Algorithms for reengineering 1991 Census geography *Environment and Planning A* 27, 425-46
- Martin, D. (1997) From enumeration districts to output areas: experiments in the automated creation of a census output geography, *Population Trends*, 88, 36-42
- Martin, D. (1998) 2001 Census output areas: from concept to prototype *Population Trends* 94, 19-24
- OPCS/GROS (1992) *ED/Postcode directory: Prospectus* 1991 Census User Guide 26 OPCS, Titchfield
- Tranmer, M. and Steel, D. G. (1998) Using census data to investigate the causes of the ecological fallacy *Environment and Planning A* 30, 817-31

Annex A

For further copies of this document or any feedback, please contact:

Census Customer Services
Office for National Statistics
Segensworth Road
Titchfield
Fareham
Hampshire PO15 5RR
Tel: 01329 813800
MINICOM 01329 813669 – for the hard of hearing
Fax: 01329 813587

Internet address: <http://www.statistics.gov.uk>
e-mail: census.customerservices@ons.gov.uk

This paper was prepared for ONS by:

Professor David Martin
Department of Geography
University of Southampton
Southampton
SO17 1BJ
Tel: 023 8059 3808
e-mail: D.J.Martin@soton.ac.uk
Internet address: <http://www.soton.ac.uk/~djm1/>

Contact details for the other Census organisations are:

Scotland

Customer Services Population Statistics Branch
General Register Office for Scotland
Ladywell House
Ladywell Road
Edinburgh EH12 7TF
Tel: 0131-314 4254
Fax: 0131-314 4344
Internet address: <http://www.open.gov.uk/gros/groshome.htm>
e-mail: customer@gro-scotland.gov.uk

Northern Ireland

Census Office for Northern Ireland
Northern Ireland Statistics and Research Agency
2 –14 Castle Street
Belfast BT1 1SA
Tel: 01232 526087
Fax: 01232 526949
Internet address: <http://www.nisra.gov.uk>
e-mail: census.nisra@dfpni.gov.uk