Environment and Planning A 2011, volume 43, pages 2399 – 2418

doi:10.1068/a43601

Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales

Samantha Cockings, Andrew Harfoot, David Martin, Duncan Hornby
Geography and Environment, University of Southampton, Southampton SO17 1BJ, England; e-mail: s.cockings@soton.ac.uk, ajph@geodata.soton.ac.uk, d.j.martin@soton.ac.uk, ddh@geodata.soton.ac.uk
Received 20 December 2010; in revised form 11 April 2011

Abstract. Automated zone-design methods are increasingly being used to create zoning systems for a range of purposes, such as the release of census statistics or the investigation of neighbourhood effects on health. Inevitably, the characteristics originally underpinning the design of a zoning system (eg, population size or homogeneity of the built environment) change through time. Rather than designing a completely new system every time substantive change occurs, or retaining an existing system which will become increasingly unfit for purpose, an alternative is to modify the existing system such that zones which still meet the design criteria are retained, but those which are no longer fit for purpose are split or merged. This paper defines the first generic methodology for the automated maintenance of existing zoning systems. Using bespoke, publicly available, software (AZTool), the methodology is employed to modify the 2001 Census output geographies within six local authority districts in England and Wales in order to make them suitable for the release of contemporary population-related data. Automated maintenance of an existing system is found to be a more iterative and constrained problem than designing a completely new system; design constraints frequently have to be relaxed and manual intervention is occasionally required. Nonetheless, existing zone-design techniques can be successfully adapted and implemented to automatically maintain an existing system. The findings of this paper are of direct relevance both to the Office for National Statistics in their design of the 2011 Census output geographies for England and Wales and to any other countries or organisations seeking to maintain an existing zoning system.

1 Introduction

For the purposes of this paper, a zoning system is defined as a set of areas used for collecting, reporting, mapping, or analysing data which are geographically referenced to the earth's surface. Some 'standard' zoning systems are defined nationally and used for many purposes: examples include those used for the release of census statistics, for the targeting and delivery of resources, or for the reporting of electoral votes. The design criteria of standard zoning systems (such as population size and placement of boundaries) are often defined by organisations such as statistical agencies or administrative authorities. Other, nonstandard, zoning systems are defined on an ad hoc basis, often for a specific study or application, and are generally used only for that purpose: the design criteria for such systems are usually defined by the individual or organisation carrying out the study.

Historically, most zoning systems were designed and created manually. Manual design enables humans to control the process and make decisions based on local knowledge or intuition, but such processes can lack objectivity and may be extremely time consuming and resource intensive [see, for example, Balinski et al (2010) on the design of electoral geographies in the UK]. Recent years have seen an increase in the use of automated techniques for creating zoning systems which are optimised to meet specific design criteria (eg, Cockings and Martin, 2005; Flowerdew et al, 2007; Haynes et al, 2007; Martin et al, 2001). Automated procedures offer more efficient, systematic, and objective methodologies for designing optimised zoning systems than

manual methods, although their success is still dependent on the extent to which it is possible to model real-world phenomena, whether it is feasible to parameterise the required design criteria, and the effectiveness of the zoning algorithm(s) employed.

All zoning systems face the challenge that the phenomena for which they were originally designed change through time: the quality of the zoning system with respect to those phenomena will therefore also inevitably change (usually degrade) through time. Some zones will remain fit for purpose, but others will no longer meet the required criteria. There are thus strong reasons to regularly update existing zoning systems in order to make them more accurately reflect contemporary data. By contrast, there is an on-going international desire for zoning systems to be stable and consistent through time. Such stability facilitates the comparison of statistics between and within countries through time (Martin et al, 2002), aids operational continuity, and serves to reinforce the sense of belonging associated with places. Historical zones may therefore sometimes persist even if they are no longer statistically optimal: for example, parishes (the lowest level of local government in England) have survived largely due to notions of neighbourhood identity and local representation.

When they need to update an existing system, most countries or designers have chosen either to completely redesign all zones within the system or to retain the entire system in its original form. Few have undertaken a process of what is termed "zone maintenance" by the Office for National Statistics (ONS): that is, the modification of an existing zoning system, such that some zones remain the same whilst others are modified to reflect changes in the underlying phenomena being measured. Scotland is unusual in this respect in that it has maintained its census geographies since 1981 by making modifications only in areas where there has been significant population change (Exeter et al, 2005). Where such maintenance processes exist, they are generally undertaken using manual or, at most, semiautomated procedures. While the use of automated zone-design techniques for creating entirely new zoning systems is arguably now well established, their potential usefulness for carrying out maintenance of an existing zoning system has not yet been explored. This paper addresses this gap by developing a generic methodology for automated zone maintenance and then demonstrating its application to the specific example of maintaining the 2001 Census output geographies for England and Wales.

The rest of this paper is organised as follows: section 2 briefly reviews existing automated zone-design techniques and their applications to date, identifying the pressing need for these techniques to be extended to enable automated maintenance of existing zoning systems; in section 3 we propose a generic methodology for the automated maintenance of existing zoning systems; in section 4 this generic methodology is applied to the empirical example in order to demonstrate how an existing zoning system can be maintained using automated techniques; finally, in section 5 we discuss the results of the empirical example, in terms of its implications both for the creation of the 2011 Census output geographies for England and Wales and for the application of automated maintenance procedures more generally.

2 Existing automated zone-design techniques and the need for automated maintenance procedures

Automated zone-design techniques have evolved partly to enable the efficient and objective creation of zoning systems for operational or research purposes and partly to explore phenomena related to the spatial analysis of data, such as the modifiable areal unit problem (Openshaw, 1984). Shortt (2009) provides a useful overview of the concepts, terminology, and methods involved in automated zone design (sometimes also termed 'regionalisation' or 'redistricting'). One of the most widely applied automated

zone-design algorithms is the automated zoning procedure (AZP), which was first developed by Openshaw (1977a; 1977b) and subsequently enhanced by Openshaw and Rao (1995), Alvanides (2000), and Alvanides et al (2002). The AZP algorithm works by iteratively combining and recombining sets of building blocks in order to create output zones which optimise a set of prespecified design criteria. Martin (2003) further developed the functionality of the AZP and his algorithm was subsequently used by ONS to create the 2001 Census output geographies for England and Wales (Harfoot et al, 2010; Martin et al, 2001). Other authors have also employed similar AZP-based algorithms for a range of purposes, including the development of standard geographies for the release of statistics and the creation of zoning systems for specific investigations (Cockings and Martin, 2005; Flowerdew et al, 2008; Grady and Enander, 2009; Haynes et al, 2007; 2008).

The vast majority of applications of automated zone-design techniques to date have had three characteristics in common: they have involved designing a completely new zoning system from scratch; all zones within the system have been created in one process; and all zones have been subject to the same design criteria. In some instances—for example in the creation of the 2001 Census output geographies for England and Wales (Martin et al, 2001)—the design process was undertaken from a completely blank canvas, with no preexisting building blocks or input zones. In others (eg, Flowerdew et al, 2008; Haynes et al, 2007; 2008), zoning systems have been created by taking an existing set of zones (such as census enumeration districts) and using these as building blocks which are then aggregated to create larger zones which optimise the required design criteria.

Recently, some authors (eg, Ang and Ralphs, 2008) have started to explore the use of automated zone-design techniques for creating 'refreshed' or updated geographies, but even these involve redesigning all zones within the system at once, with no attempt to preserve any of the existing zones which may actually still be fit for purpose. This means that, not only is any consistency of zones through time lost (thus reducing the ability to make comparisons), but also any existing data for the original zones must be transferred to the new boundaries.

An alternative approach is to try to maintain the existing zoning system such that any zones which no longer meet the design criteria are modified, but any existing zones which are fit for purpose are retained. There appear to have been no attempts to date to explore whether such a process of maintenance can be undertaken using automated techniques. There is therefore a need both to develop generic methods for carrying out automated maintenance procedures and to evaluate their usefulness for specific applications: this paper addresses both of these needs.

3 A generic methodology for the automated maintenance of existing zoning systems

'Maintenance' of an existing zoning system involves amending a subset of the system's zones, most likely via a combination of splitting, merging, or complete redesign of groups of the existing zones, to create a new set of maintained zones which are optimised according to specific design criteria. Figure 1(a) presents an example of a simplified zoning system which requires maintenance. The design criteria for the system are that all zones must be within-threshold (where the lower population threshold is 100 and the upper threshold 250) and as homogeneous (in population size) and as compact (in shape) as possible. The population within each of the four zones is shown. Zones A and B are both below the lower threshold (termed under-threshold), zone C is above the lower threshold and below the upper threshold (within-threshold) and zone D is above the upper threshold (over-threshold).

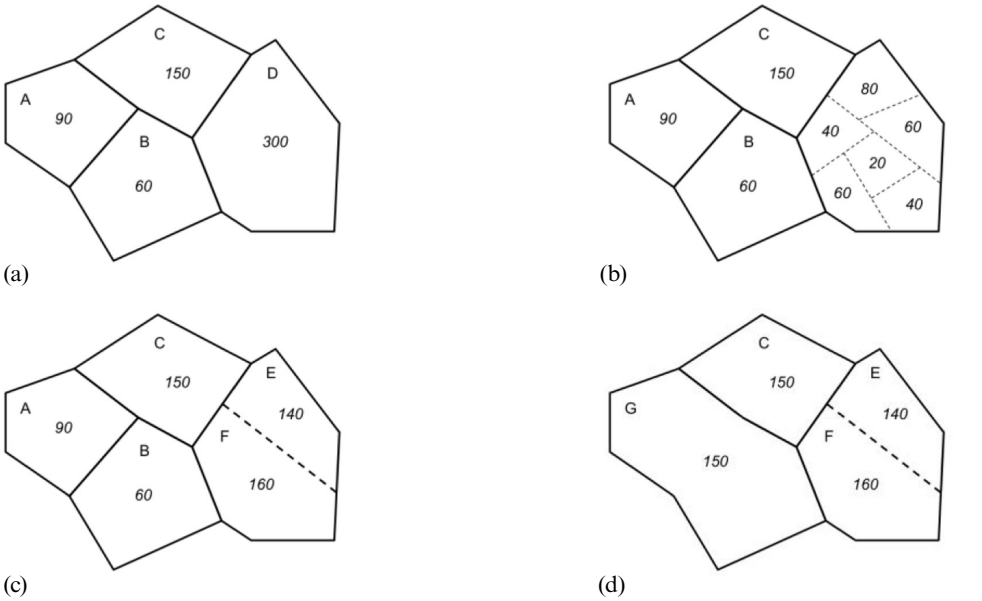


Figure 1. Simplified example of a zoning system requiring maintenance (lower threshold 100; upper threshold 250): (a) original (input) zones in zoning system; (b) building blocks for zone D which needs to be split; (c) split zone—input zone D has been split into two new output zones (E and F); (d) merged zones—input zones A and B have merged to create new output zone (G).

First, the input zones are separated into two groups: (i) over-threshold zones and (ii) within-threshold or under-threshold zones. The over-threshold zones need to be split: this requires a set of building blocks which are smaller than the input zones but which nest perfectly within them [as shown in figure 1(b)]. Using standard automated zone-design techniques, these building blocks can be aggregated in order to meet the design criteria. Each over-threshold zone is processed separately, which ensures that any aggregation takes place only within that zone, rather than across its boundaries with other zones (as this would reduce the uniqueness, and therefore utility, of look-ups between the original and maintained zones). This process results in two or more new 'maintained' zones which optimise the design criteria. Zone D is therefore split into zones E and F [figure 1(c)] as this particular solution creates two new within-threshold zones which are optimised for homogeneity and compactness.

Any under-threshold zones [such as zones A and B in figure 1(a)] need to be merged with one or more other zones. Under-threshold zones are only allowed to merge with other under-threshold zones or within-threshold zones; merging with an over-threshold zone and then splitting the resultant zone (eg, merging B with D and then splitting the resultant zone), or merging with any of the newly split over-threshold zones [B with F in figure 1(c)] is not desirable as this complicates any look-ups between the original and maintained zoning systems. The set of zones available for merging (usually sets of contiguous under-threshold and within-threshold zones) can be controlled via a list supplied to the program: in figure 1(c) this is a list of zones A, B, and C. The optimal solution in this case is to merge zones A and B, thus creating zone G [figure 1(d)]. Here, zone C, which was already within-threshold, remains unchanged after the maintenance procedures (although it might, if necessary, have been merged with one or both of zones A and B): stability is therefore retained wherever possible.

In some areas, there may be reasons why splitting or merging the existing zones is not desirable or does not produce the required results. In such cases a complete redesign of all zones may be deemed appropriate. This can be undertaken using the same standard aggregation algorithm as that used for splitting and merging, but this time supplying the program with building blocks for all of the original zones.

Figure 2 shows a system diagram of the generic automated maintenance methodology. In zoning systems which have a hierarchical structure—that is, where lower level sets of zones nest within one or more higher-level sets of zones (eg local within regional)—the process can be applied hierarchically. For example, maintenance can first be performed at the local level and the outputs from this process can then form the input zones for maintenance at the regional level. The order in which the maintenance process is carried out (eg, local to regional or regional to local) may influence the ability to successfully split, merge, or redesign zones, and can also affect the statistical and aesthetic characteristics of the resultant maintained geographies.

The requirements for maintaining an existing zoning system can be met using the same AZP-based algorithm as that used previously by a range of authors to design systems from new. The main differences between the two processes relate to how the algorithm is employed. In a maintenance situation it is applied to subsets of zones within the system, often at different levels of geography (eg, nested hierarchical), and frequently in an iterative process, rather than to all zones within the system, at one level, at once. The same basic aggregation algorithm is also employed in each of the splitting, merging, and redesign processes, but different sets of zones are supplied to the program in each case.

In a maintenance situation—because the problem space is more localised and the number of zones available for aggregation is smaller—there are usually fewer potential solutions than when designing from new. In some (possibly many) instances, it will not be possible for the algorithm to find a solution which meets the design criteria. For example, when attempting to split an over-threshold zone, the variable(s) being used for target or threshold constraints (such as population) may be unevenly distributed between the zone's constituent building blocks, thus preventing it from being split. Or, when attempting to merge an under-threshold zone with one or more neighbours, the input zone may be entirely surrounded by over-threshold zones, meaning that there are no neighbouring zones with which it can merge. In such cases it is possible to sequentially relax one or more constraint(s) to see if a solution can be found. If, after having relaxed all permitted constraints, some zones still do not meet the criteria, the only other option is manual intervention. This will usually require the relaxation of even more design constraints. At the end of this process, all the resulting zones are recombined to form the maintained zoning system. The new zoning system therefore comprises zones which are the same as in the original zoning system (ie, those that were already within-threshold and have not been used for merging with under-threshold zones, or those which were under-threshold or over-threshold but could not be resolved), zones which have been created by mergers, and those which are the result of splitting. In terms of commonality between the original and new zones, data can be directly compared for zones which have stayed the same in the two zoning systems, whereas zones resulting from mergers and splits will require look-ups to undertake comparative analyses: merged zones will require a simple aggregation of data, but zones resulting from over-threshold splits represent a new output geography and will therefore require some form of ancillary information (such as boundaries or weights) to enable the disaggregation of data.

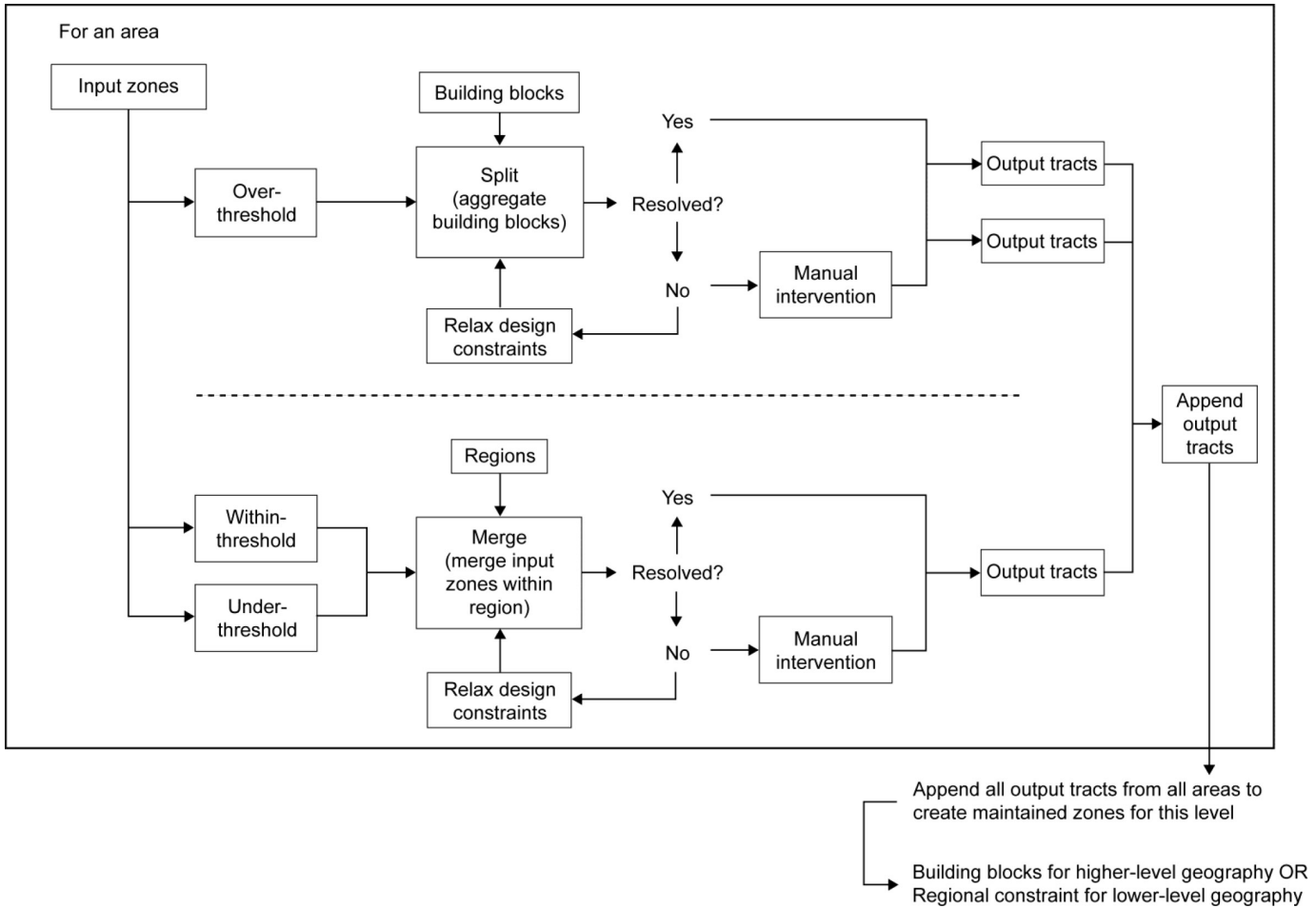


Figure 2. Generic automated maintenance method.

4 Empirical example: maintaining the 2001 Census output geographies for England and Wales

4.1 Background: 2001 Census output geographies and the need for maintenance

The 2001 Census output geographies for England and Wales were created by ONS using automated zone-design methods (Harfoot et al, 2010; Martin et al, 2001). First, Thiessen polygons were generated around the address points of households and communal establishments (CEs). These polygons were then constrained to fit within ward and parish boundaries and, where possible, aligned with geographical features such as roads. The boundaries of neighbouring address polygons within the same postcode were then dissolved to form a set of synthetic unit-postcode boundaries. The postcodes were aggregated into output areas (OAs) using a bespoke version of the AZP algorithm which optimised various design criteria, including minimum population and household thresholds, a target number of households per zone, socio-economic homogeneity (based on accommodation type and tenure) and spatial compactness of the zones. The OAs were subsequently aggregated into super output areas [lower-layer and middle-layer super output areas (LSOAs and MSOAs, respectively)] which have since been used for the release of a broad range of neighbourhood statistics (http://www.neighbourhood.statistics.gov.uk). Output geographies for Scotland and Northern Ireland were created via a separate, but similar, process (albeit with much lower thresholds in Scotland): these geographies are not considered here.

At the time of creation, the then National Statistician (Cook, 2004) stated that the output geographies should provide a stable buildings block base for the next twenty-five years. Martin (2006) noted that this desire for stability brings with it a need for the development of maintenance strategies to deal with inevitable population change. By the time of the next UK Census in 2011, in some areas of England and Wales, the 2001 output geographies will not be fit for the release of census data. Ralphs and Mitchell (2006) and Cockings et al (2009) have explored the level of population change since 2001. Cockings et al (2009) suggested that by 2005 only 0.89%, 0.37%, and 0.16% of OAs, LSOAs, and MSOAs, respectively, fell outside the relevant population thresholds. They concluded that, if current trends continue, the percentages of zones breaching the thresholds by 2011 are likely to be very low. However, whilst the total number of breaches might be low, these breaches are likely to be concentrated in specific areas because population and societal change tends to exhibit spatial clustering. In addition, due to problems with the 2001 address register (ONS, 2004), some areas (eg, Manchester and Westminster) are known to have output geographies which were not optimal for the release of 2001 data: there is therefore a case for completely redesigning at least some of the output geographies in these areas in 2011. In 2007 ONS conducted a user consultation on output geographies. This revealed a "strong user demand for stability in the small area geographies" but also a desire for the output geographies to "reflect 'reality' at the time" (ONS, 2007, page 3). As a result, the National Statistics' small area geography policy (ONS, 2007) is to retain a high degree of stability at both the OA and SOA levels, with an aim to limit change to a maximum of 5% of OAs nationally, to minimise changes at the LSOA level and to only make changes at the MSOA level in exceptional circumstances.

The aim of this empirical example therefore, is to evaluate automated methods for maintaining the 2001 Census output geographies such that existing fit-for-purpose zones are retained, but other zones are split, merged, or redesigned, as appropriate, in order to make them suitable for the publication of 2011 Census data.

4.2. Methods

4.2.1 *Selection of study areas and preparation of data*

Using mid-year estimates (MYE) provided by ONS and the Department for Environment, Food and Rural Affairs' (DEFRA) urban/rural classification (DEFRA, 2005) (see Cockings et al, 2009), six study areas were selected as being indicative of areas which will require maintenance in 2011. Table 1 shows the study areas and their characteristics.

Table 1. Study-area characteristics.

Local authority district (LAD)	Area type [a]	Population change [b]	Used by ONS [c]	Additional comments
Camden	major urban	high growth	test	
Isle of Anglesey	na	low growth	rehearsal	island; included as a control area
Lancaster	significant rural	mid growth	rehearsal	coastal
Liverpool	major urban	low decline	test	coastal
Manchester	major urban	mid growth	small-scale test	underenumeration problems in 2001
Southampton	large urban	low growth	local	coastal; local knowledge

[a] Based on DEFRA (2005) urban/rural classification for England. No similar classification available for Wales: Anglesey therefore does not have formal urban/rural type, but is rural.
[b] Population change between 2001 and 2006 mid-year estimates for LADs and between 2001 and 2005 for output areas, lower-layer super output areas, and middle-layer super output areas: low $\leqslant$ 5% change; mid = 5–10%; high $\geqslant$ 10%.
[c] Area used by ONS in 2007 Census Test (ONS, 2009), 2009 Census Rehearsal (ONS, 2010), or small-scale tests to support fieldwork (various years).

A contemporary (2007/08) household-level dataset was required for the study areas, containing the variables that will be used as design criteria for the 2011 output geographies (population count, accommodation type and tenure for each residential household, and population count for each CE). One of the difficulties with developing and testing methodologies for the census is that there are no readily available datasets which provide the small-area distribution of all people and households for England and Wales between censuses [see Martin (2010) for a discussion of the problems associated with candidate datasets]. A purpose-specific dataset was therefore constructed under secure-setting conditions at ONS Titchfield. Figure 3 summarises the data-creation process. 2001 Census household-level records for the study areas were matched to Ordnance Survey MasterMap Address Layer 2 (AL2) addresses for 2008, matching on Ordnance Survey Address-Point Reference address string or grid reference. Matched addresses were populated with their 2001 population, accommodation type, and tenure. The postcodes of large CEs (such as prisons and halls of residence) were identified using lists provided by ONS. Postcode-level MYEs for 2007 were used to allocate populations to unmatched addresses and to adjust the overall population totals at postcode, postcode sector, and local authority district (LAD) levels. Accommodation type for unmatched addresses was derived from a combination of building function/structure attributes from MasterMap and a bespoke building-type classification based on the topological relationships between neighbouring residential

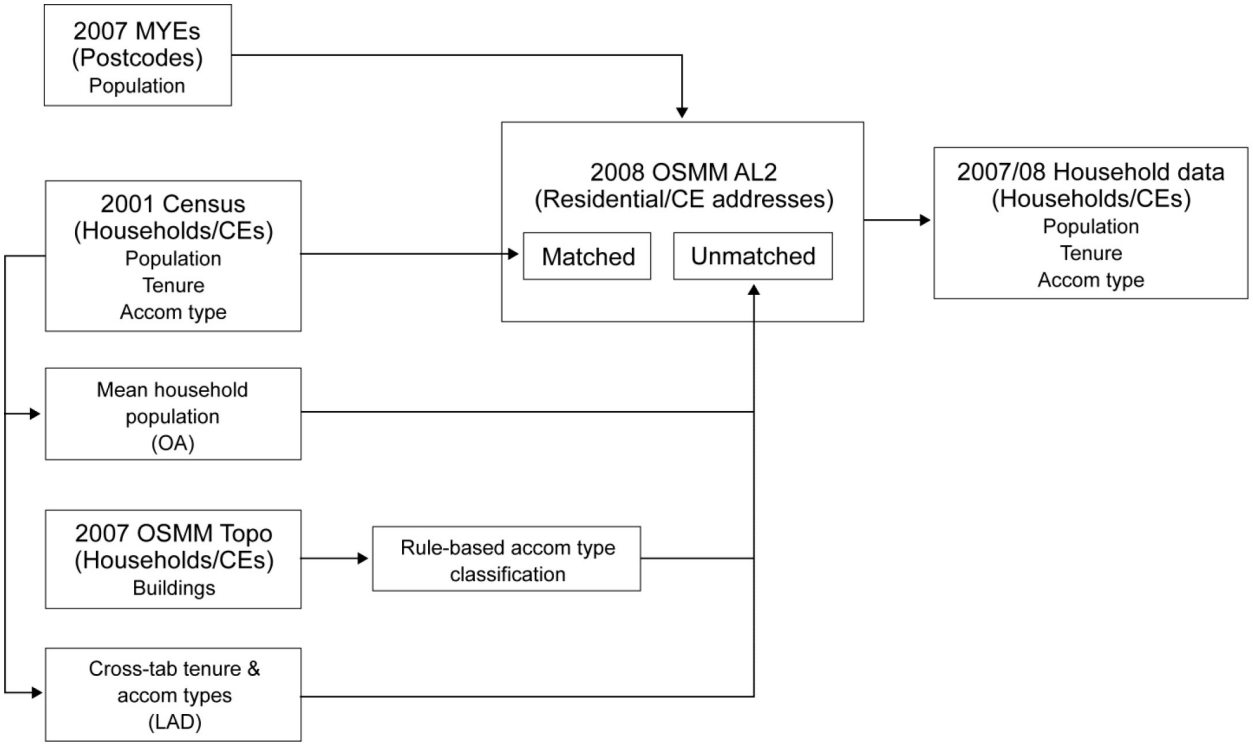


Figure 3. Methodology for creation of household-level data (AL2: Address Layer 2; CE: communal establishment; LAD: local authority district; MYE: Mid-year estimates; OA: output area; OSMM: Ordnance Survey MasterMap^{TM}).

buildings in the 2007 MasterMap Topography Layer. The proportional relationships between accommodation type and tenure in 2001 were calculated for each study area and tenure was allocated to unmatched residential addresses in the relevant proportions. This process thus created best-available estimates of population, tenure, and accommodation type for residential households and population counts for CEs in the six study areas for 2007/08 (hereafter termed 2007).

Postcode polygons (for use as building blocks when splitting over-threshold OAs) were created for each of the study areas, using similar methods to those employed in 2001 (Harfoot et al, 2010). Thiessen polygons were created around all residential and CE addresses, constrained to fall within the existing 2001 OA boundaries. Neighbouring address polygons with the same postcode were then merged to create a set of postcode polygons. The boundaries of these polygons were, where possible, aligned with roads (using the road centrelines of public roads from MasterMap Integrated Transport Network, 2007) and railways (from Meridian 2, 2008), with priority being given to dual carriageways, motorways, and railways.

4.2.2 *Identification of 2001 OAs, LSOAs, and MSOAs requiring maintenance*

ONS has recommended that the minimum population and household thresholds employed in 2001 are retained for 2011. In 2011, when the aim will be to identify and maintain zones which are no longer fit for purpose, it will also be necessary to consider upper thresholds. Table 2 defines the thresholds employed in this study, which were developed in consultation with ONS and are similar to those employed by Mitchell and Ralphs (2007). It is likely that similar thresholds will be employed in 2011. The household-level data for the study areas were aggregated to 2001 OAs, LSOAs, and MSOAs, and zones which had breached the lower or upper thresholds

Table 2. Population and household thresholds.

Geography [a]	Population thresholds [b]		Household thresholds [c]	
	lower	upper [d]	lower	upper [d]
OA	100	625	40	250
LSO	1 000	3 000	400	1 200
MSOA	5 000	15 000	2 000	6 000

[a] OA—output area; LSOA—lower-layer super output area; MSOA—middle-layer super output area.

[b] Population thresholds = household thresholds × 2.5 (equating approximately to average household size).

[c] Household threshold values from Mitchell and Ralphs (2007, table 1.1, page 4).

[d] No upper thresholds published in 2001 for OAs or LSOAs. Values are given by 2001 OAPS target mean × 2 (as in Ralphs and Mitchell, 2006). MSOAs did have a published upper threshold of 4000 households, but here = 6000 households (as in Mitchell and Ralphs, 2007) to be consistent with ratios used at other levels.

Table 3. Threshold breaches, 2007, all study areas combined, by output geography level.

Geography [a]	Total number of zones	Under-threshold	Within-threshold	Over-threshold
OA	4 988	43	4 836	109
LSOA	962	12	938	12
MSOA	200	1	198	1

[a] OA—output area; LSOA—lower-layer super output area; MSOA—middle-layer super output area.

Table 4. Statistical characteristics of output areas (OAs), lower-layer super output areas (LSOAs), and middle-layer super output areas (MSOAs) in 2001, 2007, and following maintenance, for all study areas combined.

	Count	Total population		Total households		Homogeneity		Mean shape score [b]
		mean	SD	mean	SD	tenure [a]	accommodation [a]	
OAs								
2001	4988	290.4	101.9	124.8	16.3	0.182	0.289	37.83
2007	4988	314.6	140.7	127.7	44.0	0.161	0.263	37.83
maintained	5074	309.3	128.6	125.5	29.5	0.162	0.264	37.79
LSOAs								
2001	962	1505.7	201.7	646.9	101.6	0.132	0.190	42.70
2007	962	1631.2	362.7	662.0	171.7	0.117	0.177	42.70
maintained	961	1632.9	321.1	662.7	132.3	0.117	0.177	42.74
MSOAs								
2001	200	7242.5	1078.9	3111.7	472.5	0.091	0.134	44.42
2007	200	7846.5	1465.0	3184.1	614.5	0.083	0.129	44.42
maintained	200	7846.3	1535.4	3184.1	588.3	0.084	0.128	44.60

[a] Intra-area correlation (see Martin et al, 2001; Tranmer and Steel, 1998).
[b] Perimeter^2/area (see Martin et al, 2001).

Figure 4. Output areas (OAs) breaching threshold(s) in an area of Liverpool, 2007. Crown copyright 2003. Crown copyright material is reproduced with the permission of the Controller of HMSO.

by 2007 were identified: these are shown in table 3. At all levels the majority of zones were still within-threshold in 2007. At the OA level the number of zones exceeding the upper threshold was 2.5 times the number falling below the lower threshold; within LSOAs and MSOAs, the numbers were much lower overall and the numbers of over-threshold and under-threshold zones were similar at each level. Figure 4 shows the OAs breaching the thresholds in an area of Liverpool: as can be seen, this area contains a number of both under-threshold and over-threshold OAs, but the majority of OAs remain within-threshold.

Table 4 summarises the statistical characteristics of the 2001 OAs, LSOAs, and MSOAs, together with the same statistics for the 2007 data within the 2001 geographies. This table clearly shows how the optimised distributions created in 2001 had degraded by 2007, with the means and standard deviations of population and household sizes having increased whilst the homogeneity of accommodation type and tenure within zones had decreased. Note that ONS is unlikely to use a decline in socioeconomic homogeneity as a reason for maintaining a zone in 2011,[(1)] although this may be of more concern to some users.

4.2.3 *Implementation and evaluation of automated maintenance procedures using AZTool*
Enhancements to our existing automated zone-design software (AZTool) were carried out to improve its functionality and performance for the specific challenges involved in maintenance procedures. The new version of AZTool (freely available at http://

[(1)] ONS has stated that it *may* redesign exceptional instances of OAs which were found to be socioeconomically heterogeneous in 2001 and which did not fit specified criteria for statistical zones, based on the results of the 2011 Census Outputs Geography consultation (http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-consultations/open-census-consultations/census-output-geography-consultation/index.html

Figure 5. Maintained output areas (OAs) in an area of Liverpool, 2007. Crown copyright 2003. Crown copyright material is reproduced with the permission of the Controller of HMSO.

www.geodata.soton.ac.uk/software/AZTool/) was employed to split or merge zones which had breached the thresholds, using the design criteria shown in table 5. Bottom-up (OA – LSOA – MSOA) and top-down (MSOA – LSOA – OA) approaches to the maintenance were implemented. Postcodes were used as the building blocks when splitting over-threshold OAs. The output zones from one maintained level of output geography went on to become the building blocks or regional constraints, as appropriate, for the next level to be maintained. Where solutions could not be found using all of the constraints, an iterative process of relaxing constraints and rerunning the procedures was undertaken. First, the minimum boundary length (MBL) constraint was relaxed; then the target tolerance; and finally both were relaxed together. Any zones for which solutions were not found after all constraints had been relaxed were left unresolved.

Table 5. Constraints and criteria employed in the maintenance procedures.

Constraint/criteria	Details	Weighting
Thresholds	As per table 2	na
Target (number of households)[a]	OA: 125; LSOA: 600; MSOA: 3 000	100
Homogeneity	Intra-area correlation scores for accommodation type and tenure	100
Shape	Perimeter^2/area	100
Minimum boundary length	10% of the total perimeter of the shared boundaries	na
Regional constraint	Respect higher level output geographies (eg, LSOA, MSOA)	na

[a] OA—output area; LSOA—lower-layer super output area; MSOA—middle-layer super output area.

Where identifiable, a reason for this nonresolution was recorded. The differences between the outputs from the bottom-up and top-down approaches were evaluated by comparing the statistical qualities of the maintained zoning systems produced by each approach.

4.3 Results and analysis

4.3.1 *Bottom-up versus top-down approach to maintenance*

The bottom-up and top-down approaches produced very similar results, other than when an under-threshold zone (eg, an OA) fell within an over-threshold higher level geography (eg, an LSOA). In this situation the order in which the maintenance was carried out influenced either the ability to fix the higher level geography (in the case of the bottom-up approach) or the ability to fix the lower level geography (in the top-down approach). There was only one such case in all of the six study areas. While it is impossible to predict the number of times that this situation may occur nationally in 2011, the study areas (other than Anglesey) were selected to be indicative of the type and scale of change likely to be seen in 2011. It is reasonable to assume, therefore, that there will not be many situations like this arising in 2011. Given that adherence to the lower thresholds is likely to be critical in 2011 (for statistical disclosure control reasons), a bottom-up approach is recommended as this ensures that the ability to merge under-threshold OAs is not reduced by any maintenance carried out previously on the higher level geographies. The disadvantage of adopting a bottom-up approach may be that a small number of higher level geographies (eg, LSOAs) remain over-threshold, but this is considered to be less critical. For conciseness, the rest of this paper presents only the results for the bottom-up approach and for all study areas combined: the full set of results, by study area, is available in Cockings and Harfoot (2010).

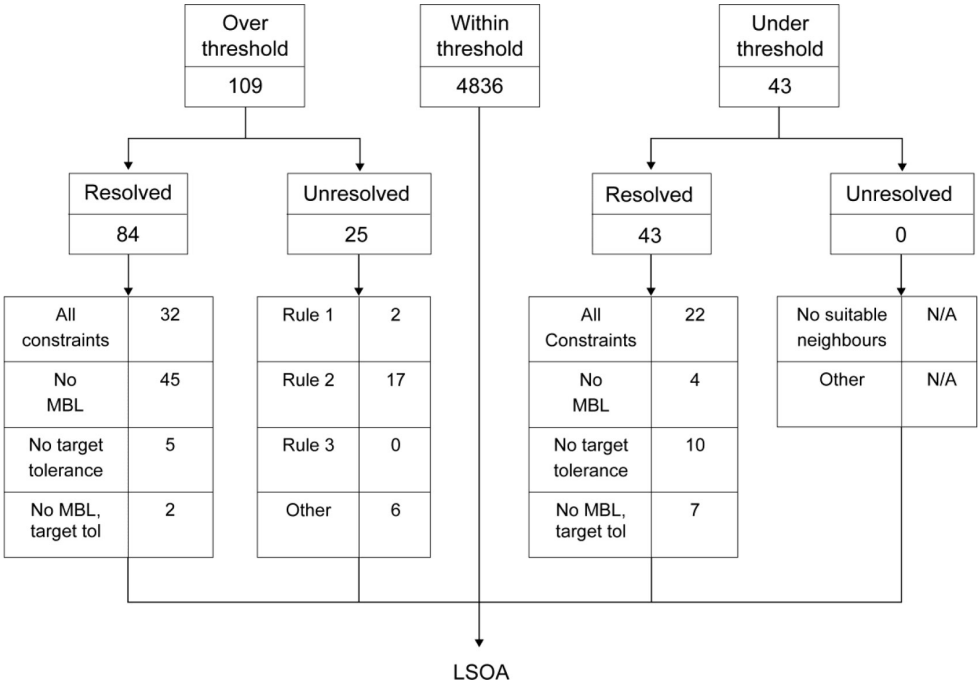
4.3.2 *Number of zones successfully maintained*

Figure 5 presents the maintained zones for the same area in Liverpool as that shown in figure 4. Over-threshold zones have now been split and under-threshold zones merged, so that all zones are now within-threshold.

Figure 6 details how many of the over-threshold and under-threshold 2001 OAs were resolved or not resolved, for all study areas combined, using the bottom-up approach. The schematic shows how many were resolved with all constraints in place and how many by sequentially relaxing first the MBL constraint, then the target tolerance, and finally both the MBL and target tolerance.[(2)] Where quantifiable, reasons for the nonresolution of zones are also shown. With all constraints in place, only 29% of over-threshold and 51% of under-threshold OAs could be resolved. Relaxing the MBL and/or the target tolerance constraints substantially increased the numbers resolved, taking the relevant percentages to 77% of over-threshold and 100% of under-threshold OAs. At the LSOA level, with all constraints in place, three out of twelve over-threshold zones were resolved and two out of twelve under-threshold zones. After relaxing both the MBL and target tolerance, six out of twelve over-threshold and ten out of twelve under-threshold LSOAs were successfully resolved. Of the two MSOAs requiring maintenance (one over-threshold and one under-threshold), both were resolved by relaxing the MBL and target tolerance constraints together. The results demonstrate that it was easier to resolve under-threshold areas (via mergers) than over-threshold areas (via splits).

At the OA level the main reason for nonresolution of over-threshold OAs (seventeen out of 109) was where at least one of the OA's constituent building blocks

[(2)] Relaxing the MBL tends to reduce the compactness of the maintained output geographies; relaxing the target tolerance potentially reduces the homogeneity of household size between zones.



Key
MBL: Minimum boundary length.

Rule 1: Area breached upper population threshold but has less than two times the lower household threshold (or vice versa) so cannot be split into within-threshold zones.

Rule 2: Area breached upper population and/or household threshold but one or more of its constituent building blocks also breached the same upper threshold so it cannot be split into within-threshold zones.

Rule 3: Area breached upper population and/or household threshold but distribution of population and/or households is overly concentrated within one building block, leaving insufficient population and/or households in other building blocks to create within-threshold zones.

Figure 6. Resolved and unresolved output areas (OAs).

[postcode(s)] had population and/or household counts which were themselves already greater than the OA-level upper threshold(s). This uneven spatial distribution of population/households between the building blocks prevented the OA from being split into two (or more) new within-threshold zones. By contrast, at the LSOA level, the main reasons were insufficient household counts to enable the zones to be split (three of the six unresolved LSOAs) or a specific geometric configuration of building blocks which prevented a solution being found (two out of six). A possible solution to the over-threshold building-block problem would be to subdivide the block(s) prior to carrying out the maintenance procedures. This would be similar to the manual intervention undertaken by ONS in 2001, when tower blocks with more than 250 households with the same grid reference were split (by postcode) and the grid reference(s) of the subblock(s) were moved to a nearby location: there were five such tower blocks within the study areas investigated here. In 2011 all under-threshold zones (especially OAs) will need to be resolved to ensure that statistical disclosure control requirements are met: manual intervention will therefore be required when such zones cannot be merged automatically. No such manual intervention was undertaken in this research. No upper thresholds were employed in 2001: ONS will need to consider how strictly these should be enforced in 2011. For example, thirty-four of the 109

over-threshold OAs and two of the twelve over-threshold LSOAs in the study areas would also have been over-threshold in 2001 had such a threshold existed: where such zones cannot be split by the automated procedures they could be allowed to remain over-threshold as they will not have exceeded the threshold(s) due to population/household change.

4.3.3 *Statistical qualities of the maintained geographies*

Table 4 presents the statistical qualities of the (bottom-up) maintained OAs. These can be compared directly with the statistics for 2001 and 2007 in the same table (already discussed in section 4.2.2). Note that the statistics for the maintained geographies include unresolved zones. As expected, the maintenance procedures were able to successfully move the OA-level means and standard deviations of total population and total households back towards their original (2001) values from their degraded 2007 values, but they were unable to improve significantly on the homogeneity of accommodation and tenure within zones. This is because the population/household thresholds and the target (number of households) have stronger influences on the final solution, especially when the number of building blocks is small. The shape scores for the postmaintenance OAs were actually very slightly better (ie, more compact) than the original 2001 OAs. This is mostly due to the fact that the maintenance procedures did not insist that split postcodes were placed within the same OA: in 2001 this acted as a significant constraint on the algorithm's ability to produce compact shapes. A slightly different shape score was also used in this research compared with 2001.[3]

Table 4 also presents the postmaintenance results for LSOAs and MSOAs. While there were improvements in the standard deviations of households and population at the LSOA level, there was little change in the population or household means or in homogeneity, and the shape score actually deteriorated slightly. At the MSOA level most of the statistics deteriorated, other than the standard deviation of households. Again, this is due to the very low number of zones involved in the maintenance processes: only twenty-four LSOAs and two MSOAs required maintenance and so, whilst the algorithm achieved its main aim (which was to ensure that all zones were within-threshold), not surprisingly, there was little scope to produce solutions which were statistically superior to those seen premaintenance.

5 Discussion

The empirical example reported here demonstrates that it is possible to adapt and apply the generic automated maintenance methodology developed in section 3 in order to maintain an existing set of zones which are no longer fit for purpose because the underlying data have changed. It has produced results which are specific to the 2011 Census for England and Wales as well as generic findings relevant to the wider application of such methods.

There are various limitations with the empirical example. A number of assumptions were made in linking the 2001 Census, MYEs, and AL2 addresses to create the household-level data. For example, 2001 households which matched to an AL2 address point were assumed to be unchanged in their population count, accommodation type, and tenure since 2001: there will clearly be cases where this is not true. Subdivisions of existing dwellings and dwellings which have been newly built should have been accurately identified and populated, but instances where the population count or

[3] In 2001 the shape score employed was the sum of the weighted squared differences between each OA's address-weighted centroid and the address-weighted centroids of its constituent postcode polygons; here we use perimeter^2/area, which tends to place more emphasis on the geometric properties of the zone.

tenure of an existing household have changed may have been missed. As is often the case, the data available for CEs were the least complete and least accurate (although the allocations for some large CEs will have been very accurate due to the provision of postcode lists by ONS, which enabled their unambiguous identification). The population counts were adjusted to match the MYEs at various geographical levels: the overall accuracy of the results is therefore reliant on their accuracy. Although the study areas were selected because they were areas undergoing population change, the number of zones requiring maintenance in each study area was still fairly low. If the number of zones requiring maintenance in 2011 turns out to be much higher, it is possible that there may be situations which were not encountered in the empirical example. However, the generic methodology and algorithm are both robust to large numbers of zones and other scenarios so there is no reason why they should not be able to cope with such situations. Overall, it is important to note that, whilst the household-level data may not be perfectly accurate in all areas, the main aim of the paper was to develop and test automated methods for maintaining existing zoning systems: in this respect, it was more important that the data and the study areas contained examples of the levels and types of change that the maintenance methods should be able to deal with, rather than them accurately representing the geography of population change in the study areas in 2007.

The maintenance process advocated here assumes that the building blocks employed to split over-threshold zones are available to the designer (ie, the person or organisation carrying out the automated maintenance). In the maintenance of standard geographies the designer is most likely to be a statistical organisation or data provider: there should therefore be no problem with accessing the required data. Likewise, this should not pose difficulties for researchers working with their own primary data. However, most individual users of standard geographies, such as researchers or local authorities using census data, are not able to gain access to such data and are therefore not able to modify existing standard geographies (unless they are operating under secure-setting conditions). The feasibility of using automated zone-design techniques to create user-defined geographies has been debated previously (Duke-Williams and Rees, 1998; Young et al, 2009). This paper shows that, technically, there is no reason why users should not use automated techniques to modify existing standard geographies to create their own flexible geographies; the limitations remain those related to statistical disclosure control and the potential for differencing.

The findings from the empirical example form a detailed evidence base upon which ONS can base decisions regarding the maintenance of the 2011 Census output geographies. A bottom-up approach to maintenance (ie, fixing OAs first, then LSOAs, then MSOAs) is shown to be preferable as it prioritises the need for all OAs to meet minimum population and household thresholds, which is critical for statistical disclosure requirements. An iterative maintenance process is proposed, whereby the procedures are first run with all constraints in place. Zones which cannot be resolved will then need to be reprocessed, sequentially relaxing specified constraints, until all zones are resolved or no more constraints can be relaxed. Some zones (eg, building blocks containing tower blocks) may require manual intervention prior to, or after, implementation of the automated maintenance process. Overall, the results suggest that it will be easier to resolve under-threshold zones than over-threshold zones. The software, methods, and approach developed here are being implemented and evaluated by ONS in preparation for processing of the 2011 Census results (ONS, 2011).

It is almost certain that the 2011 Census will be the last 'traditional' census in the UK (Martin, 2006). Some countries have already stopped undertaking a traditional census and have moved to a range of register-based or survey-based approaches

(Valente, 2010). Even if the census is replaced by another system of counting the population, existing census zoning systems will still need to be maintained or new ones created which enable the release of population-related statistics at an aggregate level which preserves the confidentiality of individuals, households, or organisations. This paper has demonstrated that, as well as being able to produce new zoning systems, automated zone-design techniques can be employed to maintain existing systems in an efficient, objective, and effective manner.

Many of the issues encountered in the empirical example are generic and directly relevant to other countries seeking to undertake a similar process of maintenance for census or any other zoning systems. This research has shown that in maintenance situations, just as when using automated zone-design methods to create new zoning systems, there are clear trade-offs between competing design criteria: for example, achieving a distribution tightly concentrated around the target value is often achieved at the expense of homogeneity of other variables. Despite this, the particular zone-design algorithm employed in this research (implemented using the AZTool software) usually managed to achieve a good compromise between the various zone-design criteria and constraints.

Unlike when designing a zoning system from new, maintenance of an existing system is a more cyclical process of running procedures, evaluating results, relaxing constraints, and repeating the procedures, until solutions have been found for all zones or all permitted constraints have been relaxed. In maintenance situations the solution space is much more tightly constrained. Constraints frequently have to be relaxed in order to enable solutions to be found. Having to respect a higher level geography constraint is particularly restrictive and often prevents solutions being found at all. Even when solutions are found, the statistical quality of these solutions is generally lower than that which could have been achieved had the system been designed from new. In general, more manual intervention is also required.

Automated maintenance procedures offer exciting methods for meeting other operational and research needs, such as the capability to redesign an existing zoning system where the design criteria themselves have changed. For example, within the UK there has been a recent submission to change the design criteria of parliamentary constituencies (Balinski et al, 2010). In this case an existing, predominantly manually defined, zoning system would need to be amended such that electorate size becomes homogeneous between (a reduced number of) parliamentary constituencies. One approach would be to split existing oversized constituencies (using some combination of wards, electoral areas, electoral divisions and/or polling districts as the building blocks) and to merge undersized ones (with other undersized or appropriately sized ones, constrained within LADs) whilst retaining existing appropriately sized constituencies where possible, in order to create the desired number of constituencies. Automated maintenance methods, as developed and applied in this paper, have the required capabilities to undertake such a task.

This paper has developed the first generic methodology for maintaining existing zoning systems using automated techniques and has demonstrated its application by maintaining the 2001 Census output geographies for six study areas in England and Wales. Whether updating a set of existing zones to reflect changes in the underlying data, or redesigning an existing set of zones because the design criteria have changed, the basic process of maintenance (ie, splitting, merging, or redesigning) is the same: this paper has demonstrated that automated zone-design methods can be successfully adapted and implemented in order to meet such needs.

Acknowledgements. This research was funded by ESRC Census Development Programme award RES-348-25-0019, carried out by the authors as Approved Researchers under secure conditions at ONS and guided by an Advisory Group chaired by ONS. The authors are grateful to ONS colleagues (particularly Andy Tait, Andy Bates, Steve King, and Brian Parry) for advice provided throughout the project. Views expressed in the paper are the authors' own. AZTool (available from http://www.geodata.soton.ac.uk/software/AZTool/) is copyright David Martin, Samantha Cockings, and University of Southampton. *MasterMap Address Layer 2* (March 2008): Crown Copyright Ordnance Survey; used under ONS PGA licence GD272183 2009. Household-level 2001 Census data, Postcode level mid-year population estimates and special population listings (mid-2007): access granted by ONS MicroData Release Panel. *MasterMap Integrated Transport Network Layer* (December 2007), *MasterMap Topography Layer* (December 2007), *Ordnance Survey Meridian 2* (October 2008): Crown Copyright/database right 2009; an Ordnance Survey/EDINA supplied service. *2001 Census Output Area, Lower Layer Super Output Area, Middle Layer Super Output Area, Local Authority District boundaries*: Crown copyright 2003; data provided through EDINA UKBORDERS with the support of the ESRC and JISC. *National Statistics Postcode Directory* (February 2008): Crown Copyright 2006; source: National Statistics/Ordnance Survey; data provided through EDINA UKBORDERS with the support of the ESRC and JISC. *Universities UK Student Residences List* (March 2009): obtained from http://www.universitiesuk.ac.uk

References

Alvanides S, 2000 *Zone Design Methods for Application in Human Geography* PhD thesis, School of Geography, University of Leeds

Alvanides S, Openshaw S, Rees P, 2002, "Designing your own geographies", in *The Census Data System* Eds P Rees, D Martin, P Williamson (John Wiley, Chichester, Sussex) pp 47 – 65

Ang L, Ralphs M, 2008, "Operations research for new geographies: zone design tools for census output geographies", Methodology Development Unit, Standards and Methods Group, Statistics New Zealand

Balinski M, Johnston R, McLean I, Young P, 2010 *Drawing a New Constituency Map for the United Kingdom: The Parliamentary Voting System and Constituencies Bill 2010* (The British Academy, London)

Cockings S, Harfoot A, 2010, "Census2011Geog: evaluation of automated maintenance procedures", School of Geography, University of Southamptonn, http://census2011geog.census.ac.uk

Cockings S, Martin D, 2005, "Zone design for environment and health studies using pre-aggregated data" *Social Science and Medicine* **60** 2729 – 2742

Cockings S, Harfoot A, Hornby D, 2009, "Towards 2011 output geographies: exploring the need for, and challenges involved in, maintenance of the 2001 output geographies" *Population Trends* **138** 38 – 49

Cook L, 2004, "The quality and qualities of population statistics, and the place of the census" *Area* **36** 111 – 123

DEFRA, 2005 *Defra Classification of Local Authority Districts and Unitary Authorities in England: A Technical Guide* Department for Environment, Food and Rural Affairs, http://www.defra.gov.uk/evidence/statistics/rural/rural-definition.htm

Duke-Williams O, 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

Exeter D, Boyle P, Feng Z, Flowerdew R, Schierloh N, 2005, "The creation of 'consistent areas through time' (CATTs) in Scotland, 1981 – 2001" *Population Trends* **119** 28 – 36

Flowerdew R, Feng Z, Manley D, 2007, "Constructing data zones for Scottish neighbourhood statistics" *Computers, Environment and Urban Systems* **31** 76 – 90

Flowerdew R, Manley D, Sabel C, 2008, "Neighbourhood effects on health: does it matter where you draw the boundaries?" *Social Science and Medicine* **66** 1241 – 1255

Grady S, Enander H, 2009, "Geographic analysis of low birthweight and infant mortality in Michigan using automated zone design methodology" *International Journal of Health Geographics* **8** 10

Harfoot A, Cockings S, Hornby D, 2010, "Technical summary: 2001 Output Area Production System (OAPS) methodology", School of Geography, University of Southampton, http://census2011geog.census.ac.uk

Haynes R, Daras K, Reading R, Jones A, 2007, "Modifiable neighbourhood units, zone design and residents' perceptions" *Health and Place* **13** 812 – 825

Haynes R, Jones A, Reading R, Daras K, Emond A, 2008, "Neighbourhood variations in child accidents and related child and maternal characteristics: does area definition make a difference?" *Health and Place* **14** 693 – 701

Martin D, 2003, "Extending the automated zoning procedure to reconcile incompatible zoning systems" *International Journal of Geographic Information Science* **17** 181 – 196

Martin D, 2006, "Last of the censuses? The future of small area population data" *Transactions of the Institute of British Geographers, New Series* **31** 1 6 – 18

Martin D, 2010, "Understanding the social geography of social undercount" *Environment and Planning A* **42** 2573 – 2770

Martin D, Nolan A, Tranmer M, 2001, "The application of zone design methodology to the 2001 UK Census" *Environment and Planning A* **33** 1949 – 1962

Martin D, Dorling D, Mitchell R, 2002, "Linking censuses through time: problems and solutions" *Area* **34** 82 – 91

Mitchell B, Ralphs M, 2007, "Developing maintenance rules for the neighbourhood statistics output geographies", Methodology Directorate, Office for National Statistics

ONS, Office for National Statistics

2004 *2001 Census: Manchester and Westminster Matching Studies Full Report* http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/local-authority-population-studies/index.html

2007 *National Statistics Small Area Geography Consultation 2007* http://www.ons.gov.uk/ons/about-ons/consultations/closed-consultations/2007/geography-policy-public-consultation/index.html

2009 *2007 Census Test: Summary Evaluation Report* http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-census-project/2007-test/index.html

2010 *2011 Census: Evaluation of the 2009 Rehearsal* http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/2009-census-rehearsal/index.html

2011 *2011 Census Output Geography (England and Wales) — Review and Consultation* http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-consultations/open-census-consultations/census-output-geography-consultation/2011-census-outputs-geography-consultation.doc

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

Openshaw S, 1977b, "Algorithm 3: a procedure to generate pseudo-random aggregations of N zones into M zones, where M is less than N" *Environment and Planning A* **9** 1423 – 1428

Openshaw S, 1984 *The Modifiable Areal Unit Problem* CATMOG 38 (Geo Books, Norwich)

Openshaw S, Rao L, 1995, "Algorithms for re-engineering 1991 Census geography" *Environment and Planning A* **27** 425 – 446

Ralphs M, Mitchell B, 2006, "Maintenance requirements for Super Output Area geographies: modelling changes from 2001 – 2006", Methodology Directorate, Office for National Statistics

Shortt N, 2009, "Regionalization/zoning systems", in *International Encyclopaedia of Human Geography* Eds R Kitchin, N Thrift (Elsevier, Oxford) pp 298 – 301

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

Valente P, 2010, "Census taking in Europe: how are populations counted in 2010?" *Population and Societies* **467** (May), http://www.ined.fr/en/publications/pop_soc/bdd/publication/1506/

Young C, Martin D, Skinner C, 2009, "Geographically intelligent disclosure control for flexible aggregation of census data" *International Journal of Geographical Information Science* **23** 457 – 482