UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Geography and Environment

Automated Zone Design for the Spatial Representation of Population

by

Samantha Cockings

Thesis for the degree of Doctor of Philosophy

March 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Geography and Environment

Thesis for the degree of Doctor of Philosophy

AUTOMATED ZONE DESIGN FOR THE SPATIAL REPRESENTATION OF POPULATION

Samantha Cockings

Any mapping or analysis involving population data aggregated to geographical areas (zones) is subject to the modifiable areal unit problem (MAUP), namely that observed patterns are influenced by both the scale (size) and aggregation (boundary placement) of zones. This has important implications, not only for researchers undertaking analyses, but also for national statistical organisations needing to decide what zoning system(s) (sets of zones) to employ to release population data. This thesis explores, enhances and extends automated zone design techniques for the spatial representation of population. It addresses three key themes: the use of automated zone design to explore the modifiable areal unit problem; automated maintenance and adaptation of existing zoning systems; and the importance of building blocks in automated zone design. It shows that administrative geographies are not necessarily the most appropriate zones for exploring health and environment relationships and that automated zone design can be used to explore sensitivity of results to the MAUP. It demonstrates that automated procedures can be used to update existing zoning systems which have become unfit for purpose due to population change, and also to modify existing zones to make them suitable for representation of other phenomena such as workplace statistics. It provides evidence that building blocks are a crucial, but under-rated, component of the zone design process and concludes that all zone design should be based on sound theoretical reasoning and a clear conceptualisation of the phenomena and processes being represented. The thesis has had a demonstrable impact on the fields of environment and health, and official population statistics; its concepts and methods have been employed by a diverse range of researchers, as well as by the Office for National Statistics to create 2011 Census output areas and workplace zones for England and Wales.

Preface

I was appointed as Lecturer in Socio-Economic Applications of GIS in Geography and Environment at the University of Southampton in 2000. All of the research presented in this set of published works has been undertaken since that time and in that capacity.

The five papers that I am submitting for the degree of Doctor of Philosophy by Published Works are:

1. **Cockings S** and Martin D (2005) Zone design for environment and health studies using pre-aggregated data *Social Science & Medicine* 60(12) 2729-2742

2. **Cockings S**, Harfoot A and 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

3. **Cockings S**, Harfoot A, Martin D and Hornby D (2011) Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales *Environment and Planning A* 43(10) 2399-2418

4. Martin D, **Cockings S** and Harfoot A (2013) Development of a Geographical Framework for Census Workplace Data *Journal of the Royal Statistical Society: Series A* 176(2) 585-602

5. **Cockings S**, Harfoot A, Martin D and Hornby D (In Press) Getting the foundations right: spatial building blocks for official population statistics *Environment and Planning A*

Paper 1 was based on unfunded research, undertaken in collaboration with Prof David Martin: I conceived of the original concept, undertook the analysis and wrote the first draft of the paper. Prof Martin provided academic insight and software for the analysis (AZM). We jointly edited and approved the final manuscript for publication.

Papers 2, 3 and 5 were outputs from the ESRC-funded ‘Census2011Geog’ project (2008-2010), for which I was Principal Investigator. I developed the initial concept, managed the programme of research, provided the intellectual direction for the analysis and led the preparation and submission of the manuscripts. Andrew Harfoot was the Research Assistant employed on the project: under my guidance, he undertook the majority of the analysis under secure conditions at the Office for National Statistics (ONS) and prepared tables and figures for the papers. Duncan Hornby performed some of the early analysis as a Research Assistant. Prof Martin provided academic support and sat on the Project Advisory Group, which was chaired by ONS. All authors, except for Hornby, contributed to editing of the manuscript for final submission.

Paper 4 was derived from various research projects funded by the Department for Transport and the Office for National Statistics between 2009 and 2012. Prof Martin and myself were joint Principal Investigators on these projects. Together, we developed the concepts and methods and prepared the manuscript for publication. Andrew Harfoot undertook the analysis, at ONS, under our academic direction and contributed to preparation of the manuscript.

As a number of the papers submitted for this Doctorate are jointly co-authored with Prof Martin, a letter from him is provided in Appendix 1, corroborating our respective contributions to each.

Contents

ABSTRACT i

Preface i

Contents i

List of tables iii

List of figures v

DECLARATION OF AUTHORSHIP vii

Acknowledgements ix

1. Introduction to the Published Works 1

1.1 Introduction 3

1.2 Methodological and software development 15

1.3 Theme 1: The use of automated zone design to explore the modifiable areal unit problem 21

1.4 Theme 2: Automated maintenance and adaptation of existing zoning systems 27

1.5 Theme 3: The importance of building blocks in automated zone design 33

1.6 Conclusions 39

1.7 List of references 47

2. The Published Works 53

2.1 Paper 1: Zone design for environment and health studies using pre-aggregated data 55

2.2 Paper 2: Towards 2011 output geographies: Exploring the need for, and challenges involved in, maintenance of the 2001 output geographies .. 91

2.3 Paper 3: Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales 123

2.4 Paper 4: Development of a Geographical Framework for Census Workplace Data 163

2.5 Paper 5: Getting the foundations right: spatial building blocks for official population statistics 201

Appendices 239

Appendix 1 241

List of tables

Table 1 Constraints in AZTool 19

List of figures

Figure 1 Flowchart of AZTool algorithm.. 16

DECLARATION OF AUTHORSHIP

I, Samantha Cockings

declare that the thesis entitled

Automated Zone Design for the Spatial Representation of Population

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly while in candidature for a research degree as a staff member at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- a substantial part of this work has been published, during my employment as a staff member at the University of Southampton, as:

1. **Cockings S** and Martin D (2005) Zone design for environment and health studies using pre-aggregated data *Social Science & Medicine* 60(12) 2729-2742

2. **Cockings S**, Harfoot A and 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

3. **Cockings S**, Harfoot A, Martin D and Hornby D (2011) Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales *Environment and Planning A* 43(10) 2399-2418

4. Martin D, **Cockings S** and Harfoot A (2013) Development of a Geographical Framework for Census Workplace Data *Journal of the Royal Statistical Society: Series A* 176(2) 585-602

5. **Cockings S**, Harfoot A, Martin D and Hornby D (In Press) Getting the foundations right: spatial building blocks for official population statistics *Environment and Planning A*

Signed: …………………………………………………

Date: ………………………………………………………

Acknowledgements

My sincere thanks go to: Prof David Martin, for his unwavering enthusiasm, intellectual guidance and personal support throughout the time that it has taken to bring this thesis to fruition, and for his contribution to the papers herein; to Andy Harfoot, for his exceptional, high quality, work on the various census-related projects reported here; to the various colleagues at the Office for National Statistics (ONS), with whom much of this research has been jointly undertaken: there are too many to name individually, but in particular, Andy Tait, Ian Coady, Donna Viney and James Bayliss; and also to Alistair Calder and Martin Ralphs of ONS, with whom I conceived of the initial Census2011Geog project all those years ago! I can honestly say it has been a pleasure to work with you all, so thank you.

To Prof David Briggs, Prof Robin Flowerdew, Prof Robin Haynes and Prof Graham Moon, for being good-natured, intellectual, sounding boards at various times throughout the research.

To all of my friends at the University of Southampton and beyond, who have helped to make it enjoyable and put things in perspective when things weren't going quite so well!

Finally, my eternal thanks go to my parents, without whom none of this would have been possible: from their early support in getting me to University, through all the trials and tribulations along the way; I'll be forever grateful.

It has been a rocky road, with some significant challenges along the way, but I guess that is what makes you appreciate it all the more when you finally get there!

1. Introduction to the Published Works

1. Introduction to the Published Works

1.1 Introduction

The zones employed for mapping and analysing small area population statistics can have a potentially enormous impact on patterns observed. In particular, the size of zones and the placement of their boundaries can greatly influence inferences drawn from such data. This so-called ‘modifiable areal unit problem’ (MAUP) (Openshaw, 1984) means it is critical that data providers, analysts and policy-makers give explicit consideration to the design of any zoning systems (i.e. sets of zones) used for the release of official population statistics and to those zones employed for specific investigations. Automated zone design is the design of a zoning system using automated, computer-based, methods. There is a range of different algorithmic approaches to automated zone design: this thesis is concerned with that group of methods descended from the automatic zoning procedure (AZP), first introduced by Openshaw (1977a, 1977b). Extending key research in this field by Openshaw, Alvanides and Martin, the thesis explores the role that AZP-based approaches to automated zone design can play in designing zones for mapping and analysing population statistics. First, it illustrates the role that automated zone design can play in exploring the MAUP in studies involving small area population data, particularly within environment and health studies. Second, it considers the use of automated zone design for the output of official population statistics. While automated methods had previously been used to design zoning systems from new, they had never before been employed to update or adapt existing systems. This thesis presents novel methodologies for selectively updating and modifying existing zoning systems in two key situations: (i) where the underlying phenomena being mapped have changed substantially from when the zoning system was originally designed, thus making the zoning system no longer fit for purpose; and (ii) where a zoning system originally designed for representing one phenomenon is adapted to facilitate representation of a different one. Finally, the thesis presents the first systematic evaluation of the role that building blocks (the small, elemental, zones aggregated to create zoning systems) play in automated zone design, demonstrating how important they are to successful zone design.

The rest of this section discusses the key issues involved in representing population spatially. It then critically reviews approaches, algorithms and software used to automatically design zones for the spatial representation of population. Finally, it presents the aims of the thesis and its three key themes.

1.1.1 The spatial representation of population

> "The question is simply what objects at what scales do we wish to investigate?" (Openshaw and Taylor, 1979, p.143)

This thesis is concerned with spatial representation of population at the small area level or large (cartographic) scale. In most cases, the aim is to map or analyse the spatial distribution of population, to explore relationships between population and other factors such as economic, social or natural environmental features, or to explore changes in these patterns and relationships through time.

It is difficult to accurately map the location and characteristics of population. This is because population is the sum of individuals and these individuals are usually moving fairly continuously, albeit at varying spatial and temporal scales. The challenge then, is how to capture and map the location and characteristics of these mobile individuals. This is usually done via indirect geographic referencing, mostly via a residential address or some other relevant location (Martin, 1997). Decisions concerning what to map, and how, are often scale dependent: if we are concerned with mapping population at the national scale, daily movements to and from home to work are mostly inconsequential. However, if we are concerned with mapping at the small area level, for example to determine demand for local services or for emergency planning, such high precision may be vital.

Most small area population data in the UK has, to date, been based on place of residence. This is either because conceptually it makes sense for place of residence to be used as the geographic location of individuals e.g. because it is where they spend a large proportion of their time, or because it is the most pragmatic means by which to collect data about individuals and households e.g. via address-based surveys or censuses. For many applications, locations such as place of work or education may actually be more meaningful georeferences, but these are generally not yet widely employed in national datasets.

Population is most commonly represented spatially as points or areas. Although the decision as to which feature type to use is often determined by the scale of analysis, this is not always the case. For example, at a large scale, we might expect that points would be used to represent the locations of individuals (people or buildings) whereas areas might be better suited to the representation of geographical areas which are in some way distinct, such as neighbourhoods defined by type of housing or level of socio-economic deprivation. In practice, this is not always the case, and areas and points are often used interchangeably to represent features at both the same and different scales: census output areas (OAs), for example, may be represented either by points (population-weighted centroids) or areal boundaries. The choice often depends on the type of mapping or analytical operation being undertaken. For mapping densities of population, areas are usually more appropriate, but for linking population data to other data, such as via point-in-polygon procedures, points may be more useful.

This thesis is concerned with the representation of population as areas (hereafter termed zones), specifically at the local (small area) level. Mapping what is essentially a spatially discontinuous phenomenon in terms of zonal boundaries is challenging (Openshaw and Rao, 1995). As explained in Paper 1 (Cockings et al, 2005), researchers tend to use zones either because they have an explicit interest in area level effects e.g. neighbourhood deprivation, or simply because data are published at this level. In small area studies involving population, even where data are collected at the individual or household level, they are usually aggregated to zones by the national statistical organisation prior to release in order to prevent inadvertent disclosure of attributes pertaining to individuals or households. The design of such zones can greatly influence any patterns observed in the data. While others had previously demonstrated the problems associated with drawing statistical inference from spatially aggregated data (Blalock, 1964; Gehlke and Biehl, 1934; Kendall and Yule, 1950; Sawicki, 1973), Openshaw (1984) was the first to present a systematic review of the MAUP, demonstrating the wide range of correlation coefficients that could be achieved simply by aggregating the same set of data in a variety of different ways. Specifically, he identified the scale (size of zone) and aggregation (placement of boundaries at a given scale) effects. Openshaw (1977a, 1978) was also one of the first to recognise the opportunities presented by automated zone design techniques for both exploring the effect of the MAUP and also for attempting to 'control', or take advantage of, its existence.

1.1.2 Automated zone design: concepts, approaches and software

Zone design for the publication of population statistics is neither a novel concept nor a new practice. Parliamentary constituencies have been employed to present the results of elections in England since 1832 (Johnston, 2002) and census statistics have been available for geographical areas for many years, although 1971 was the first census to provide truly widely available small area data, in the form of aggregate statistics for enumeration districts (EDs). Historically, zone design is perhaps most synonymous with the practice of political districting (also known as redistricting, regionalisation, regional partitioning or, when carried out to achieve a desired political outcome, "gerrymandering") (Horn, 1995; Johnston, 2002; Rossiter and Johnston, 1981), but many other applications also exist, as evidenced in this thesis.

Zone design may be undertaken using manual or automated processes. Prior to the 2001 Census in England and Wales, the areas used to both collect and publish data (EDs) were designed by hand using paper maps. For the 2001 Census, Martin et al (2001) describe the transition from a manual to an automated process, which also paralleled the separation of collection zones (EDs) from output zones (OAs) (Martin, 1998a). Automated zone design may be formally defined as the design of a zoning system using automated, computer-based, methods. The output of an automated zone design process is termed a 'zoning system' – that is, a set of zones. Automated zone design allows us to tackle large zone design problems which would otherwise be too complex and time consuming using manual methods. They also facilitate the systematic application of design criteria across the entire zoning system. But, they are only feasible if measurable design criteria and decision rule sets can be defined and then implemented using appropriate data and computerised methods. Any trade-offs between competing criteria must be capable of resolution via an algorithmic approach, such as an objective function.

Various approaches to automated zone design exist: Shortt (2009) categorises these into clustering-based, hierarchical and rules-based methods. The research presented in this thesis is based on the automatic zoning procedure (AZP), originally developed and pioneered by Openshaw (1977a, 1977b)[1]. The AZP is a heuristic local boundary optimising algorithm designed to produce an optimal solution to the problem of grouping a set of small elemental units (hereafter referred to as building blocks, but termed 'basic spatial units' in Openshaw's original (1977a) manuscript) into a set of fewer, larger, output zones. It is a local algorithm because it does not search for the global best solution and is heuristic in the sense that it searches for an acceptable optimum solution, rather than the true, mathematical, optimum. Openshaw and Rao (1995, p.429) classify AZP as a 'mildly steepest descent' algorithm because it searches for the best single move possible at each step but only locally, rather than globally. A key difference between Openshaw's AZP approach and earlier redistricting methods is that AZP is inherently spatial and iterative. Previous attempts at solving this type of computational problem had involved mostly linear programming methods, which had not always taken full account of critical spatial measures (such as the contiguity relationships between building blocks, lengths of common boundaries, or measures of shape) and which had only been applied to small problems.

The AZP starts by randomly aggregating a set of building blocks into a set of larger zones in order to meet specified design criteria. This is known as the initial random aggregation (IRA) phase. These zones are then iteratively swapped with neighbours in order to optimise the value of an objective function. Openshaw (1978) outlines the use of a range of design criteria, including minimising the variation in population size, area or population density between zones, minimising the within-zone variation of variables and the use of a shape metric to maximise compactness of zones. A key requirement of any AZP-based procedure is the calculation of an adjacency or contiguity matrix, which records which zones share one or more common

[1] Note that in subsequent papers by other authors, AZP becomes referred to as the 'automated', rather than 'automatic', zoning procedure; this convention is therefore also adopted hereafter in this thesis, to ensure consistency with the literature.

boundaries. This contiguity information must be updated and re-evaluated for each (potential) swap in the iterative phase.

The original AZP algorithm was subsequently enhanced by Openshaw and his co-workers who investigated a range of alternative algorithms intended to improve its performance, in particular to overcome the occasional tendency for it to get stuck in local sub-optima. Openshaw and Rao (1995), for example, experimented with simulated annealing (SA). This permits swaps which do not initially result in an improvement to the objective function but which have a good probability of leading to a better solution in later iterations. The key barrier to simulated annealing algorithms at the time was that they required significant computing power to carry out the required number of searches.

Openshaw and Rao (1995) also proposed a 'tabu' searching algorithm: a procedure which implements a hierarchical series of preferred swaps – starting with the global best, through the local best which improves the current value of the objective function, to a local best even if it results in a deterioration. Once a move has been made, its reverse move is made 'tabu' for a specified number of iterations, thus precluding it from being used again and preventing cyclical behaviour. Two forms of the algorithm were trialled – a basic one, where the user defines the length of the tabu period, and a reactive one where the tabu period is data driven. Tabu searching was found to produce good results but was again considered computationally expensive at that time.

Although Openshaw pioneered significant methodological advances and demonstrated the huge potential of automated zone design techniques, he never truly managed to resolve the competing demands of a range of design criteria. Sadly, he also never reached the point of being able to test and demonstrate his methods on the complexities of a substantive real-world problem, although alternative approaches to Martin's AZP-based methods for the 2001 Census (discussed below) were proposed by Openshaw et al (1998). Rossiter and Johnston (1981) also demonstrated AZP's potential usefulness for applications such as political redistricting in the early 1980s.

In the late 1990s, increased amounts of digital spatial data, the growth of geographic information systems (GIS) and enhanced computing power facilitated the re-emergence of Openshaw's AZP algorithm. In collaboration with Openshaw, Alvanides took advantage of these developments to enhance the algorithm. He documented, implemented and tested more robust versions of the IRA, various objective functions and the simulated annealing and tabu search algorithms (Alvanides, 2000). He also demonstrated their application to some of the wide range of spatial problems which Openshaw had envisaged AZP could tackle (Alvanides et al, 2002; Alvanides et al, 2000; Openshaw and Alvanides, 1999).

Alvanides et al (2002) noted that one of the pervasive difficulties associated with implementing zone design algorithms for real-world applications was how to integrate an objective function with any constraints, given that they are usually at odds with one another, both conceptually and algorithmically. A key debate in the literature at the time was whether it was better to impose inequality constraints e.g. to ensure that all zones were above a specified population threshold, or to use equality constraints e.g. where the aim was to minimise the sum of the squared differences between each zone's population and the target population (Openshaw et al, 1998).

In the mid to late 1990s, Martin was experimenting with automated zone design methods for creating output zones for the 2001 Census. He developed a version of the AZP algorithm which employed a single weighted objective function to handle the competing design constraints required by a census zoning system (Martin, 1997). In contrast with Openshaw et al's (1998) view that it was better to select one of the design functions as the objective function and to treat all others as equality or inequality constraints, Martin demonstrated that it was possible to produce zones which were statistically optimised but still compact by using one or more of the constraints to set up the IRA and then trading the others off in the weighted objective function. He distinguished between 'hard' constraints (which must be met) and 'soft' constraints (which were traded off in the weighted objective function). Examples of hard constraints include minimum population thresholds or the requirement for all output zones to be contiguous. Soft constraints include minimising the between-zone variability in population size or maximising the compactness of zones. In Martin's weighted objective function, if all constraints were given equal weights, all contributed equally to the assessment of any potential solution. If constraints were weighted more heavily than others, they would be given greater prominence in this calculation. Critics such as Openshaw et al (1998) argued that variables with different distributions could not be effectively combined in this way. Martin's approach was pragmatic: in order to produce a solution which represented the best compromise for all of the given criteria, he accepted that the values for individual criteria would never be perfect. Wise et al (1997) adopted a similar method.

During research in subsequent years, including an ESRC-funded project (1999-2000), Martin and colleagues further enhanced the AZP algorithm for census purposes by improving its capacity to handle different criteria. They experimented with different shape constraints and introduced the use of intra-area correlation (IAC) as a measure of within-zone social homogeneity. This allowed multiple categories of more than one variable to be incorporated, thus providing a more sophisticated measure of homogeneity (Martin et al, 2001; Tranmer and Steel, 1998). Martin et al's methods were subsequently adopted by the Office for National Statistics (ONS) in order to design the 2001 Census OAs for England and Wales (Martin, 1998a; Martin, 2002; Martin et al, 2001). ONS is the organisation responsible for undertaking the decennial census in England and Wales.

The AZP algorithm was further developed by Martin (2003) into the automated zone matching (AZM) software, designed specifically for the purpose of matching two sets of geographically incompatible zones to create one common geography. In AZM, the fundamental AZP algorithm remained unaltered from Martin's previous work but a new intermediate layer was employed for undertaking the zone matching process and a new type of objective function (measuring population 'stress') was introduced.

The series of automated zone design methods described above was implemented by a number of key authors in bespoke pieces of software. The original AZP algorithm itself was written in FORTRAN 77 by Openshaw (1977a, 1977b) and was therefore GIS-independent, although separate spatial functions were required to construct the necessary contiguity information. ZDES was subsequently written with this purpose in mind, providing a GIS wrapper from within which the AZP algorithm was called and passed data, but which also offered various spatial routines (such as creation of the contiguity information), as well as command line and graphical user interface options. The original version of ZDES, written for UNIX ArcInfo, was produced by Openshaw and Rao as part of an ESRC-funded research project in the mid-1990s. Alvanides (2000) further developed ZDES by improving the functionality and usability of the software as well as encoding more robust versions of the simulated annealing and tabu search optimisation algorithms, resulting in ZDES3 and a later, even more user-friendly, version ZDES3b. Both of these implementations retained the AZP algorithm in FORTRAN and were written in ArcInfo. A further, non-GIS, Java-based Web version (ZD2K) was developed by Alavanides et al (2002). This implemented a restricted sub-set of the most commonly employed functions from ZDES3b, with the aim of opening up the functionality of automated zone design to a wider range of users.

Another key piece of software offering automated zone design functionality during the late 1990s/early 2000s was Wise, Haining and Ma's Spatial Analysis in a GIS Environment (SAGE) software (Haining et al, 1998; Wise et al, 1997, 2001). SAGE combined existing GIS functionality with bespoke spatial analytical functions to enable the user to undertake exploratory spatial data analysis. Two of its functions included calculation of a contiguity matrix and automated zone design (termed regionalisation or region building in this instance). Zones could be designed using three criteria: homogeneity (minimizing within-group variance of one or more attributes), equality (minimizing the difference between the total value of an attribute, such as population size, across regions) and geographical compactness (Haining et al, 1998). These criteria could be weighted via an objective function, similar to Martin's approach. SAGE showed much initial promise, particularly as it had a user-friendly GUI and was linked with other GIS functionality but, ultimately, its close-coupling with ArcInfo and the SUN/Unix environment limited its lifespan and use beyond GIS specialists, meaning that it was never really widely employed or tested on a substantive real-world problem.

Alongside this, Martin developed a FORTRAN version of AZP whilst experimenting with methods for designing the 2001 Census OAs (Martin, 1997, 1998b). This working prototype implemented Openshaw's AZP algorithm but with modifications to the IRA and the introduction of hard and soft constraints and a weighted objective function. A series of incremental enhancements followed, particularly to the measures employed in the weighted function (such as IAC) as Martin, ONS and colleagues experimented with different methods (Martin et al, 2001). A new FORTRAN version of the algorithm was produced as part of Martin's 1999-2000 ESRC-funded project. These procedures were subsequently adapted by ONS and embedded in their own Output Area Production System (OAPS), which was the system used to create the actual 2001 Census OAs for England and Wales. OAPS comprised a combination of Powerbuilder scripts, an ArcInfo/ORACLE database and Arc Macro Language (AML) routines.

Subsequently, Martin (2003) went on to develop the Automated Zone Matching (AZM) software, written in Visual Basic 6. This extended previous versions by including an intermediate processing layer and a new objective function for matching incompatible geographies. AZM is the software which this thesis enhances and extends.

1.1.3 Aims and key themes

Although automated zone design had been used to explore the MAUP in a range of experimental situations (Openshaw, 1984), prior to this thesis it had not been used to systematically explore its effects in environment and health studies (the focus of Paper 1). Automated zone design had also been used to successfully design a zoning system for the output of official statistics in England and Wales (for the 2001 Census) but this had then led to questions about how best to maintain such a system as population change occurs through time, thus making parts of the zoning system no longer fit for purpose. It was also becoming obvious that there was a need to define population zoning systems based on non-residential locations, such as workplaces, to make them more suitable for a range of mapping and analytical applications. Finally, queries about the role that building blocks play in the automated zone design process had been raised by a number of authors over many years. Ultimately, this series of questions led to the research presented in this thesis.

The overall aim of the thesis is to explore, enhance and extend the use of automated zone design methods for the spatial representation of population.

Specifically, the key themes explored are:

1) The use of automated zone design to explore the modifiable areal unit problem
2) Automated maintenance and adaptation of existing zoning systems
3) The importance of building blocks in automated zone design

In all instances, the aim is to develop generic approaches and procedures which, whilst providing a solution to the specific empirical problem being investigated, can also be applied to a diverse range of applications internationally.

1.2 Methodological and software development

This section describes the methodological and software developments undertaken as part of this thesis.

One of the key limitations of the AZM software (Martin, 2003) was that it could only deal with a restricted number of input zones (Flowerdew et al, 2007). From 2001, together with colleagues in Geography and Environment at the University of Southampton, the author led development of a new software package called AZTool. The original version of AZTool extracted the core AZP algorithm and associated zone design functions from AZM and enhanced their flexibility and performance. AZTool was initially written in Visual Basic 6 to take advantage of object-oriented constructs which removed limitations on the number of input zones which could be processed. The core functionality was similar to AZM, with the ability to specify a minimum threshold, target population, shape constraint and homogeneity (based on IAC). The IRA and weighted objective function were also as per AZM and the system had a simple graphical user interface.

During 2008-2010, AZTool was comprehensively redeveloped as part of the ESRC-funded Census2011Geog project led by the author (http://census2011geog.census.ac.uk/). The key aims were to extend its flexibility to allow it to be used for a wider range of applications (including 2011 Census zone design) and to improve some of the algorithmic issues associated with the IRA and iterative phases. The program was rewritten in VB .NET, and .XML was employed for parameter specification in order to maximise flexibility and portability. The core functionality of the current version of AZTool is described below.

As per AZM, the AZTool input file structures are based on ArcInfo formats (.pat and .aat), which are provided as text files by the user. Users may create their own input files or use the Shapefile Importer provided with AZTool. This facilitates data exchange with most major GIS packages. Figure 1 presents a high-level flowchart of the AZTool algorithm.

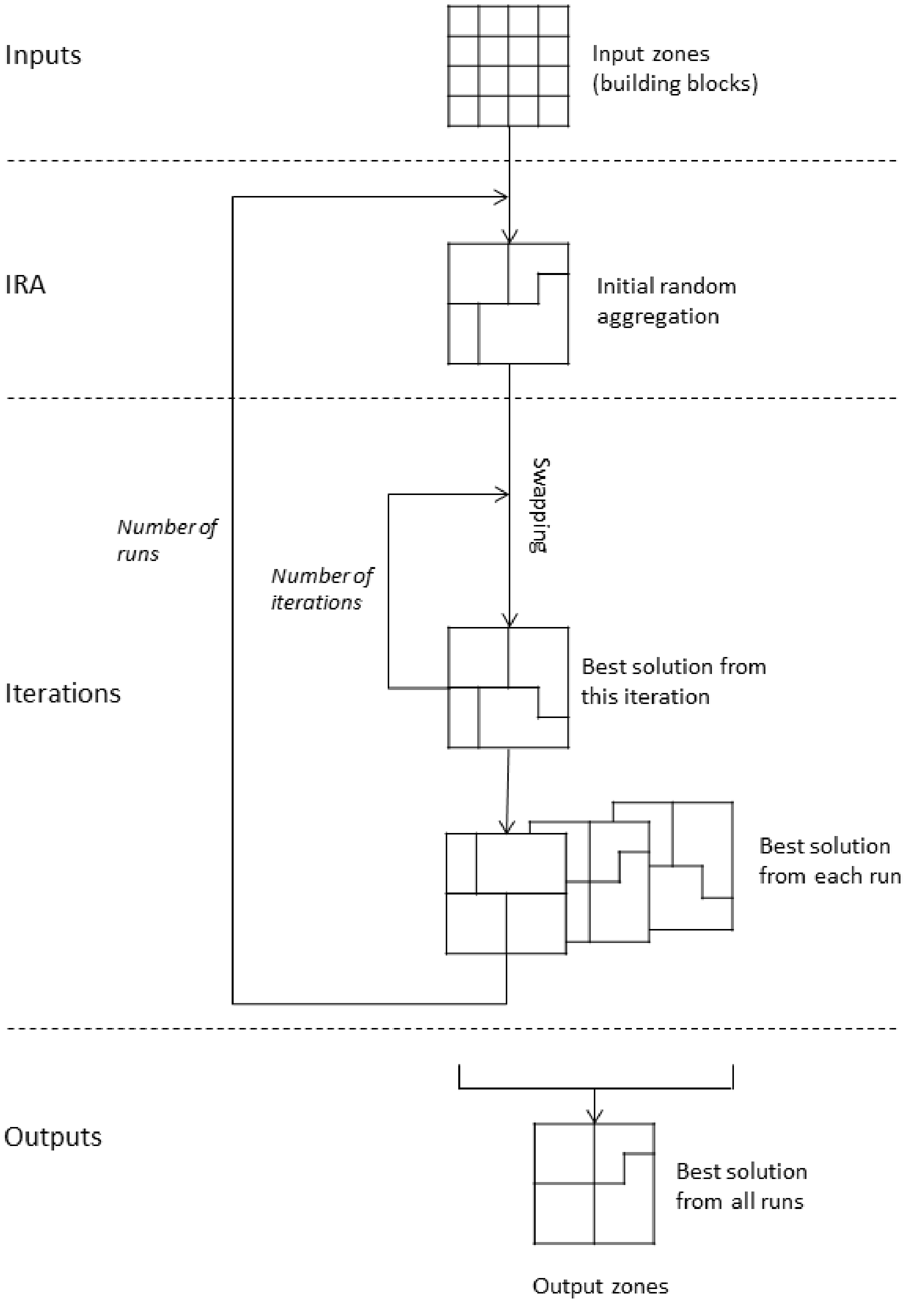


Figure 1 Flowchart of AZTool algorithm

There are four key phases to the algorithm: the input, IRA, iteration and output phases. A set of input zones (building blocks), the design criteria and various algorithm controls such as the number of iterations and runs, are provided by the user. A series of checks is then performed to construct the contiguity information required by the program. From a random starting point, the building blocks are then arranged into an IRA which satisfies the relevant design criteria and the value for the objective function is calculated for that IRA. The zones in the IRA are then iteratively swapped at random, with swaps only being retained if they result in an improvement to the value of the objective function. At the end of this swapping phase, the set of output zones with the best value so far for the objective function is stored. The swapping phase is then repeated (starting with this current best set of zones) for as many iterations as specified by the user in order to try to further improve the value of the objective function. When all iterations are complete, the best solution from all iterations is stored as the best solution from that run. The entire process is then restarted again, this time from a completely different random restart, which minimises the likelihood of the algorithm getting stuck in local sub-optima. At the end of each run, the best solution from that run is compared against the best solution from all previous runs and the current best solution is retained. The number of runs is specified by the user and is a matter of judgement derived from experience, observations of the results and factors such as the number of input/output zones and patterns in the data. Once all runs have been completed, the best solution for all runs is output as a lookup table containing each input zone and the output zone to which it has been allocated, together with various log files. This lookup table can be used to aggregate the input zones, thus creating a new set of output zones.

A key enhancement in the current version of AZTool is the algorithm used to generate the IRA: Alvanides, Openshaw and Martin had previously debated how best to set up the IRA and, in particular, how to determine the starting number of output zones. This is important because it is not possible to increase the number of zones during the iteration phase but only to reduce or retain it; the starting number of zones therefore essentially determines the scale of the final zoning system. Various approaches to this problem are possible. Two alternatives are currently provided in AZTool:

(i) a target tolerance: the value for the variable must fall within a specified (percentage) tolerance of the target value

(ii) best mean size: the number of output zones is constrained to give a mean size as close to the required target value as possible

Together with any associated minimum or maximum threshold values, these options provide the user with more control over the IRA than previously. Algorithmically, improvements have also been made to the iteration phase, with the introduction of an extra iteration loop within each run; this helps the algorithm to converge on an optimal solution more quickly. Greater flexibility is also provided with respect to hard and soft constraints, as shown in Table 1, which presents the set of constraints and criteria currently available within AZTool. Note that slightly different constraints are applicable during the IRA and iteration phases.

AZTool is written in a modular form to facilitate the addition of extra functions or measures as required and is freely available for download from http://www.geodata.soton.ac.uk/software/AZTool/. This stand-alone version of AZTool was subsequently packaged into a wrapper by ESRI (the providers of ArcGIS), forming the core of ONS' 2011 Census Output Area Maintenance Production System (OAMPS). It was this system that was used by ONS to produce the 2011 Census output geographies (including both the OA hierarchy and workplace zones, described later) for England and Wales. At the time of writing, AZTool has been downloaded and used by at least 25 research teams or organisations in 10 countries for a diverse range of applications.

In this thesis, Paper 1 (Cockings and Martin, 2005) employs AZM, whereas Papers 2 to 5 (Cockings et al, 2009; Cockings et al, 2011; Martin et al, 2013; Cockings et al, In Press) all employ AZTool.

Table 1 Constraints in AZTool

Constraint	Details	IRA[1] phase	Iteration phase
Hard constraints			
Threshold(s)	Minimum and/or maximum; as many variables as desired	Y	Y
Contiguity	Only allow mergers between geographically adjacent zones	Y	Y
'Bishops contiguity'	Allow/do not allow mergers between zones with Bishops contiguity (zones touching only by their corners and therefore with a shared boundary length of zero)	Y	Y
Minimum boundary length	Do not allow mergers between zones sharing a common boundary length less than the specified value (specified as % of total boundary length). Helps to prevent long, thin, zones forming	Y	Y
Region(s)	Do not allow zones to be merged across region boundaries (usually a higher level geography but could also be a categorical variable such as high-rise/low-rise housing)	Y	Y
Doughnuts	Allow/do not allow output zones to be created which are completely surrounded by only one other output zone	Y	Y
Soft constraints			
Target variable(s)	Minimise sum of squared deviations from target for all zones	Y	Y
Homogeneity	Maximise intra-area correlation, for as many variables and variable categories as desired	N	Y
Compactness	Minimise perimeter^2/area	N	Y

[1] Initial random aggregation

1.3 Theme 1: The use of automated zone design to explore the modifiable areal unit problem

1.3.1 Background

As noted above, there has been long-standing acknowledgement that researchers mapping or analysing population by areas (zones) should be aware of the MAUP (Openshaw, 1984). Despite this, many researchers (especially non-geographers) tend to treat geography as fixed, without giving explicit attention to the assumptions and implications associated with the zoning system employed. Openshaw (1977a) noted that usually only one of many alternative zonations is employed to analyse data and it is this that is the real 'problem' for real-world applications, rather than the MAUP itself. In his 1978 paper, Openshaw advocated the use of 'deliberate' zone design, arguing that such a practice should be based on sound theoretical reasoning.

1.3.2 Conceptual framework for use of areas

Prior to Paper 1 (Cockings and Martin, 2005), there had been little discussion of the impact that different zoning systems might have on environment and health investigations. Paper 1 discusses why and how zones should be used in environment and health studies, noting that they are usually employed either because the researcher has a specific interest in an area (or 'contextual') effect, such as an environmental factor or a policy intervention, or because data are not available at the individual level but are instead aggregated to areas. The paper proposes a conceptual framework for the spatial representation of processes and phenomena in environment and health studies and explains where zone design can aid with the practical application of such representation. Importantly, it argues that any such zone design should always be based on sound theoretical reasoning. This conceptual framework has since been adopted by authors such as Flowerdew et al (2008) and Schuurman et al (2008), who apply it in exploring neighbourhood effects on health, and the spatial epidemiology of trauma, respectively.

In common with Openshaw (1978), Paper 1 calls for deliberate zone design but goes further by outlining types of analyses where automated zone design may be useful in environment and health studies and by providing illustrative criteria for such situations. It is argued that whatever the reason for using zones for analysis, researchers and practitioners should, as a minimum, be aware that the scale and boundary placement for any zones can have a profound effect on patterns observed, and that, ideally, automated zone design methods should be employed to explore the sensitivity of results to such effects. This is a point picked up and later reemphasised by various authors such as Flowerdew et al (2008), Haynes et al (2007), Parenteau and Sawada (2011) and Riva et al (2009).

1.3.3 Scale and aggregation effects

The MAUP has two dimensions: scale and aggregation. As noted by Haynes et al (2007), Paper 1 (Cockings and Martin, 2005) was the first paper to systematically explore these dimensions in the context of environment and health studies. Grady and Enander (2009) consider it to be the seminal work in this field. Paper 1 explores the impact that scale and aggregation have on relationships between a causal factor (in this case, socio-economic deprivation) and a health outcome (limiting long term illness). It demonstrates that the scale effect is the stronger of the two, with higher correlations being seen for larger zones, but also provides evidence for a weak aggregation effect. Various authors have since reported similar effects (see, for example, Flowerdew et al (2008), Shuttleworth et al (2011)) while others have reported demonstrable aggregation effects (Grady and Enander, 2009). Further research by the author now suggests that the context in which the zoning system is employed is a critical factor, with relationships between the scale and geographical patterning of the phenomena, building blocks and output zones, in particular, affecting the impact of the MAUP on statistics. These theories are developed further in Section 1.5.

1.3.4 Standard and bespoke zoning systems

Zoning systems may be developed for standard or bespoke purposes, for example the design of zones for the output of national population statistics or the design of zones for a specific environmental health study. All zoning systems are ‘designed’ in that choices must be made about the size of the zones and where their boundaries should be placed. Even so-called ‘neutral’ zoning systems, such as grid squares, are designed in the sense that the size of the cells and the location of the grid's origin must be determined by a user, although the placement of cell boundaries thereafter does not involve any further design decisions besides cell size. As Haynes et al (2007, p.823) note:

> "Automated zone design has the semblance of objectivity, but each set of zones is the product of a set of criteria specified by the researcher. Far from defining a single optimal set of neighbourhoods, zone design opens the door to an infinite number of possibilities, all within the researcher's control."

The design of more complex zoning systems inevitably requires decisions to be made about the relative importance of design criteria. Frequently, design criteria conflict with one another, with the optimisation of one resulting in a sub-optimal solution for another.

One of the questions under consideration by the author in the late 1990s/early 2000s was the extent to which zoning systems which were optimised for a specific study might be better suited for such an analysis than standard zoning systems which had been designed with a multitude of users and uses in mind. At that time, most studies in England and Wales which required small area population data employed census zones (enumeration districts (EDs) or wards) as their units of analysis as there were no other publicly available alternatives.

Morphet (1993) and Openshaw and Rao (1995) argued that EDs are not an appropriate base for mapping and analysing population data. Building on this, Paper 1 (Cockings and Martin, 2005) demonstrates that the widely varying population sizes in 1991 Census wards (and, to a lesser extent, EDs) make them a less stable geographical base for statistical analysis than zoning systems designed specifically to minimise variation in population size between zones. Various authors have since cited our work and tested this hypothesis in a range of countries and applications, providing further weight to the assertion that administrative zones are not necessarily the most suitable units of analysis for exploring patterns of health: see, for example, Flowerdew et al (2008) – UK – limiting long term illness; Grady and Enander (2009) – USA – low birthweight and infant mortality; Jones et al (2010) – UK – physical activity amongst children; Parenteau and Sawada (2011) – Canada – respiratory health; Riva et al (2008, 2009) – Canada – active living potential and walking; Sabel et al (2012) – France – asthma. Flowerdew et al (2008) suggest that by using administrative zones we might actually be underestimating contextual (areal) effects on health.

Like Openshaw and Rao (1995), Paper 1 encourages investigators to consider designing bespoke zoning systems for their own analyses wherever possible, rather than assuming that administrative zones are suited for such purposes. It also provides a tool with which they can do this (at that time, AZM, now superseded by AZTool). Various authors have since employed AZM/AZTool for such purposes (Flowerdew et al (2008); Grady (2010); Grady and Enander (2009); Sabel et al (2012)).

Paper 1 has therefore had a demonstrable impact on the field of zone design in environment and health investigations and beyond, being cited by approximately 25 different research teams in 8 countries (Canada, Finland, France, Mexico, Northern Ireland, Portugal, UK and USA). It has also been referenced in reviews such as Curtis and Riva (2010 – writing on complexity theory and health), Rainham et al (2010 – time geographies, location technologies and spatial ecology in place and health research) and Shortt (2009 – regionalisation/zoning systems).

1.3.5 Appropriate building blocks

Paper 4 (Martin et al, 2013) further explores the MAUP by considering the suitability of zoning systems for different purposes. It explains why residential building blocks (which, prior to the 2011 Census, had been employed for the release of all census data in England and Wales) are inappropriate for mapping and analysing workplace-related statistics. Using automated zone design to create a zoning system based on workplace locations, it demonstrates how very different distributions are achieved in workplace and workforce statistics if we use these (workplace) zones rather than the traditional residence-based zones. This is an illustration of the very essence of the MAUP: the same attribute data published on different zoning systems provide different insights and can lead to very different results if used in analysis and decision-making. Data providers and analysts should therefore always aim to employ the building blocks and zoning system best suited to the specific analysis.

1.3.6 A modifiable spatiotemporal problem?

The MAUP has often been perceived and presented predominantly as a spatial problem – whereby the precise size and geographical placement of boundaries is vital – but, when dealing with population, there is also a significant temporal element which has, to date, been either completely ignored or not explicitly addressed. The spatial distribution of population is constantly changing, at virtually all scales: people travel to work, students move between term-time and vacation addresses, visitors enter and leave the country and special occasions such as festivals occur as ad hoc events. Even during night-time (arguably the least mobile time of day) there are sectors of the population, such as employees working night shifts or people undertaking leisure activities, who are not at their residential locations. Any superimposed zoning system therefore results in not just a modifiable spatial unit problem but also a modifiable spatiotemporal unit problem: by fixing zones in both space and time, we are inevitably influencing not just any spatial patterns observed but also any temporal or spatiotemporal ones.

Most zoning systems are designed to either represent the spatial distribution of population at a particular point in time, or to be representative for a period of time. The 2011 Census in England and Wales, for example, was designed to give a snapshot of the spatial distribution and composition of population on census night (27 March, 2011) but will also likely be used to represent the population for the period 2011-2021, albeit with periodic adjustments for events such as births, deaths and migration. In designing a zoning system which is optimised for a specific point in time but using it for a much longer period, we make assumptions and simplifications about the location of individuals. On various temporal scales, this (residential) location is simply not representative of the true location of the majority of individuals for much of their lives. As noted in Section 1.3.5, designing zoning systems which present alternative spatiotemporal representations (such as workplace zones) is one method for overcoming this modifiable spatiotemporal unit problem. A further approach is to update official zoning systems whenever feasible, to make them as accurate as possible. Papers 2 and 3 (Cockings et al, 2009, 2011) outline the challenges involved in undertaking such a procedure of ‘maintenance’. Paper 3 develops a generic methodology for selectively updating parts of a zoning system which are no longer appropriate for representing population at that time. In so doing, we are at least ensuring that zonal boundaries are optimised both in space and time, albeit for just one point in time. Such approaches are clearly limited, however, by frequency of data collection.

1.4 Theme 2: Automated maintenance and adaptation of existing zoning systems

1.4.1 Design of standard zoning systems from new

During the late 1990s, ONS was exploring automated methods for the design of output zones for the 2001 Census in England and Wales. The design of census boundaries clearly illustrates the challenges involved in reconciling the competing design requirements of a range of users. ONS' aim was to produce compact output geographies which had homogeneous population sizes but which also differentiated small areas in terms of their socio-economic characteristics and whose boundaries, wherever possible, followed recognisable geographical features on the ground. The boundaries also had to nest within existing administrative boundaries (wards and, where relevant, parishes). Martin (1997, 1998b) developed a prototype algorithm for the automated zone design process using synthetic postcodes as building blocks. Subsequently, ONS employed these methods to create the 2001 Census OAs for England and Wales. The same zone design code was then used to create the super output area (SOA) hierarchy (Lower-layer SOAs (LSOAs) and Middle-layer SOAs (MSOAs)), which became the key component for sharing of census and non-census data from various sources via the Neighbourhood Statistics service[2]. OAs were used as the building blocks for LSOAs, and then LSOAs for MSOAs. LSOAs were designed completely automatically, whereas MSOAs incorporated a consultation stage with key users and some manual changes were made to the automatically created zones, resulting in a more hybrid (automated/manual) product.

The 2001 Census output geographies represented a significant step forward in the creation of standard zoning systems in two respects: (i) the output zones used to publish data were different to those employed to collect data, and (ii) the output zones were more homogeneous in terms of population size and socio-economic characteristics than in any previous census, making them a more suitable statistical base for mapping and analysis. For the first time ever,

[2] http://www.neighbourhood.statistics.gov.uk

a set of purposely designed output zones had been created which were optimised for the population distribution at the time of the census.

1.4.2 Maintenance of existing zoning systems

Changes to the zones used to release census data in successive censuses present real difficulties for researchers seeking to analyse population through time. 56% of ED boundaries changed between the 1971 and 1981 censuses in England and Wales and 68% between 1981 and 1991 (Martin, 2003). The 2001 output geographies were intended to provide a stable geographical reporting framework which could then be retained for a number of years (possibly 25 years, according to the then National Statistician (Cook, 2004)); such stability should aid analysis of population trends through time and facilitate linkage with other datasets. The problem is, that the size, composition and distribution of population all change through time. Inevitably, zones which were optimised for a specific point in time become less fit-for-purpose through time. One challenge is predicting for how long a set of output zones is likely to remain useful. Paper 2 (Cockings et al, 2009) explored the extent to which the 2001 Census zones were likely to be appropriate for the release of 2011 Census data in England and Wales. Using mid-year population estimates and other ancillary datasets at various geographical levels, it estimated that by 2005/06 the vast majority of output zones (OAs, LSOAs and MSOAs) had not breached population and household thresholds which would be used by ONS for the 2011 Census. Assuming that these trends continued, it predicted that most of the 2001 OAs, SOAs and MSOAs would still be fit-for-purpose for the release of 2011 Census data. Nonetheless, because population change is usually spatially concentrated, the paper suggested that there would be specific areas where a higher proportion of zones would breach thresholds. Overall, the investigation showed that there had been more population growth than decline since 2001, resulting in more zones breaching upper thresholds than lower ones. In some cases, population change had been so great that there were nested breaches, where multiple layers of the geographic hierarchy had breached thresholds (e.g. at both OA and LSOA and, in some cases, even at MSOA levels). In some areas, patterns were complex, with upper threshold breaches adjacent to lower threshold ones. The paper concluded that, whilst in the majority of areas it should be possible to adopt a policy of no or minimal change in creating the output zones for 2011, in some areas it would be essential to update the zones to ensure accurate representation of the contemporary population distribution. A method was required, therefore, to facilitate this selective updating of some zones but retention of others. Such updates should ideally be carried out in a systematic, objective and efficient manner and ensure consistency with the design criteria used in creating the existing (automatically-generated) 2001 output zones. At the time, no such method existed internationally.

Paper 3 (Cockings et al, 2011) reports the development of a generic methodology for undertaking ‘maintenance’ of an existing zoning system. Maintenance in this context involves modification of a sub-set of existing zones, most likely via a combination of splitting, merging or completely re-designing, in order to create a set of newly ‘maintained’ zones, which are optimised according to specified design criteria. After setting out a generic methodology for such maintenance procedures, the paper implements the method, using it to adapt the 2001 Census output geographies for six study areas in England and Wales in order to make them fit for purpose for publication of 2011 Census data. Existing OAs, LSOAs or MSOAs, which have become too large or small in terms of population and/or household counts, are split or merged, as appropriate, using the AZTool software. The paper highlights that the process of maintaining an existing zoning system is more constrained (in algorithmic terms) than designing a system from new. Manual intervention and/or the relaxation of various design criteria are required in instances where the zone design algorithm is unable to split or merge zones automatically. Overall, the paper demonstrates that automated zone design techniques can be successfully employed to maintain existing zoning systems.

ONS went on to use AZTool and the methods developed in Paper 3 to design the actual 2011 Census output geographies (OAs, LSOAs and MSOAs) for England and Wales, which were released in October 2012[3]. For the first time in modern census history in England and Wales, changes to the boundaries of output zones were minimal, allowing direct comparison of data between censuses. The boundaries of only 2.6% of 2001 OAs changed between 2001 and 2011 (together with 2.5% of LSOAs and 2.1% of MSOAs) (ONS, 2012). These changes were mostly due to population change, but a small proportion were also due to local authority boundary change and modifications to improve socio-economic homogeneity in zones considered unfit for statistical purposes in 2001. As predicted in Paper 3 (Cockings et al, 2011), a small proportion of areas experienced a greater level of change to their OA boundaries: 22 out of 348 local authorities had 5% or more of their OAs changed, with City of London (27.8%), Tower Hamlets (10.8%) and Forest Heath (10.2%) experiencing the three highest levels. For the 97.4% of OAs which were unchanged between 2001 and 2011, direct comparisons can now be made. For the OAs split or merged, comparisons can be made by aggregating or merging data for the relevant 2011 or 2001 OAs respectively. The only OAs for which direct comparisons are not possible is the 0.1% which experienced complex changes to their boundaries.

It is too early to evaluate the success of the maintenance procedures and the 2011 census output zones, but the methods developed in Paper 3 do appear to have met ONS' aims and the output zones are receiving favourable initial responses from users. These output zones will now be utilised by an enormous range of users including central and local government, health authorities, industry and academia, until at least 2021. Unfortunately, full harmonisation of UK small area population statistics from the 2011 Census is not yet possible. Northern Ireland published statistics for their Small Areas (equivalent to OAs in England and Wales) at the end of February 2013 but

[3] See: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/new-developments-for-2011-census-results/2011-census-geography/modifications-of-output-areas/index.html and http://www.ons.gov.uk/ons/guide-method/geography/products/census/index.html

Scotland has yet to publish theirs. Northern Ireland undertook limited maintenance of their 2001 Census OAs, using mostly manual procedures, with 83% of their boundaries remaining the same and all modifications being mergers of existing OAs (NISRA, 2013). It is not yet clear what methods will be employed in Scotland. The generic methodology presented in Paper 3 and the AZTool software are both sufficiently flexible for use in other countries or other applications.

1.4.3 Adaptation of existing zoning systems for other purposes

As noted in Section 1.3, the use of one set of generic output zones for all applications can create problems for mapping and analysis. In Paper 4 (Martin et al, 2013), the fundamental flaw in using output zones which are based on residential locations for the publication of workplace statistics is illustrated: namely, that city centres and other high-density commercial areas contain numerous workplaces and workers, whereas suburban residential areas contain very few. The consequences of this are that in many cases data are suppressed, not published at all or are published with extremely large variances for key variables. For the 2001 Census for England and Wales, workplace population per OA ranged from 0 to 80,145 workers, with a mean of 134. This highly skewed distribution resulted in a number of important variables not being published due to disclosure concerns, including industry type, occupation and mode of travel to work (Martin et al, 2013).

Mapping and analysis is either impossible or severely limited by an inappropriate zoning system. One option for overcoming this problem is to design an entirely new zoning system based on more relevant building blocks (in this case, workplace locations); this would create a zoning system optimised for workplace statistics but would pose problems for integration with other datasets such as other (residence-based) census data. An alternative approach is to modify the existing residence-based zones by splitting or merging those which contain too few or too many workplaces and/or workers. This provides a compromise whereby spatial consistency with residential output zones is retained wherever possible, but a finer or coarser resolution is used in areas where it is most needed.

Implementing such an approach does present specific challenges related to the maintenance of statistical confidentiality, given that the spatial distribution and characteristics of workplaces and workers are very different to those seen for residents and households. Paper 4 (Martin et al, 2013) demonstrates how it is possible to take the generic automated maintenance methodology described in Paper 3 (Cockings et al, 2011) and employ it to adapt an existing zoning system to create a new system for a different purpose. A novel methodological framework for the creation of a set of output zones based on workplace locations (rather than residences) is described and the difficulties involved in protecting workers and workplaces from inadvertent disclosure via the publication of census data are discussed. Prototype 'workplace zones' (WZs) are produced for five study areas in England and Wales. Using the AZTool software, residential OAs containing too many workplaces and/or workers are split by aggregating a set of purposely created workplace postcode building blocks within the OA. OAs with too few workplaces/workers are merged with neighbouring under- or within-threshold OAs, while OAs with suitable numbers of workplaces/workers are retained. This produces a zoning system optimised for the release of workplace statistics but also aligned with the residential zoning system wherever possible.

The methods proposed in Paper 4 (Martin et al, 2013) were implemented by ONS to create the first ever set of output zones for England and Wales specifically designed for the publication of workplace statistics. The boundaries of these WZs were released by ONS in January 2013[4] but the release date for WZ data is not yet known. The WZs are already generating considerable interest amongst a diverse range of users, including central and local government, commercial organisations and transport planners (Department for Transport part-funded the research reported in Paper 4). To date, Scotland and Northern Ireland have not produced workplace zones as part of their 2011 outputs, so UK-wide harmonised workplace statistics are not yet feasible.

[4] http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-we-took-the-2011-census/how-we-planned-for-data-delivery/output-geography/index.html

1.5 Theme 3: The importance of building blocks in automated zone design

Some researchers (Manley et al, 2006; Openshaw, 1984; Openshaw and Rao, 1995 and the author) had hypothesised that the building blocks employed to produce zoning systems might significantly influence both the ability to meet the required design criteria and the magnitude of any scale and/or aggregation effects associated with such zoning systems. In fact, Openshaw (1984, p.21) went as far as to suggest that:

> "the aggregational variability in the range of possible results due to the choice of the first zoning system will exceed that of any subsequent re-aggregations of the data."

Most studies of MAUP (including Openshaw's seminal 1977a and 1978 works) had employed building blocks containing data which were already pre-aggregated to explore the effects of MAUP. This prevented them from exploring the impact that the design of such building blocks might be having on the zone design process. In 1984, Openshaw did attempt to quantify the effect of the first zonation, but his experiments were inconclusive.

From a more practical perspective, Martin (1998a, p.678) notes that:

> "The creation of any census output geography is fundamentally limited by the nature of the building blocks for which data may be aggregated and which can therefore be combined by the census offices to create output areas."

Martin (2000) also notes that the building block design stage of automated zone design has not been particularly prominent because in many cases the building blocks are already predetermined and analysts are therefore unable to select or design their own. He does comment, however, that:

> "a national digital data infrastructure and widespread GIS functionality will make these issues more important in the future" (Martin, 2000, p.110)

The research presented in Paper 5 (Cockings et al, In Press) arose as a direct response to these concerns and to observations from the author's own previous research. As predicted by Martin (2000), the importance of sound building block design was also gaining prominence amongst national statistical organisations as their digital spatial data infrastructures matured. The 2011 Census in England and Wales provided an ideal opportunity to test theories concerning the influence of building blocks on zone design, within the context of a large practical application of international importance.

The research presented by Cockings and Martin (2005) was a precursor to Paper 5 (Cockings et al, In Press). In it, the role of building blocks in the automated zone design process was explored by aggregating an individual (person) level dataset into zoning systems of a specified scale using various sets of building blocks (addresses, street blocks and grid squares). Intuitively, it might be expected that building blocks which are as small as possible and which represent the smallest features being modelled (in this case, addresses) would offer the greatest aggregational freedom and therefore produce the 'best' optimised geographies. This research suggested, however, that 'intermediate' building blocks which have some form of inherent internal homogeneity and which are aligned with recognisable geographical boundaries (such as roads or changes between different housing types) are more effective building blocks than individual address-based building blocks for producing zoning systems optimised for a range of criteria. The research also showed that such intermediate building blocks produce zones which are more homogeneous than 'neutral' building blocks such as grid squares.

Paper 5 (Cockings et al, In Press) develops this work further by undertaking a more rigorous evaluation of the role that building blocks play in the MAUP and in the design of zoning systems for official population statistics. It hypothesises that the statistical quality of output zones, the linking of data to other zoning systems and the maintenance of existing systems are all likely to be influenced by the scale and characteristics of the building blocks employed as the basis for the zoning system. Under secure conditions within ONS, it takes household data for six local authorities in England and Wales in 2007/08 and aggregates these into two sets of building blocks: postcodes and street blocks. At the time, following consultation with users, these were the two key sets of building blocks under consideration by ONS for utilisation in the 2011 Census maintenance procedures (ONS, 2007). The two sets of building blocks are used firstly to create an entirely new zoning system for the publication of census data and secondly to maintain (i.e. update via selective splitting or merging) an existing set of zones. Postcodes are shown to be more effective building blocks than street blocks, providing more uniform population and household sizes, greater precision for matching postcoded data to census data, and enabling more zones to be maintained. However, street blocks deliver more compact output zones and greater internal homogeneity of tenure and accommodation type.

Openshaw and Rao (1995) debated the virtues of using differing ratios of building blocks to output zones. Paper 5 provides evidence that the smaller sized postcodes offer greater aggregational freedom than the larger street blocks. This finding is contrary to that reported in Cockings and Martin (2005), which had suggested that smaller building blocks do not necessarily produce output zones of better statistical quality. In 1995, Openshaw and Rao had suggested that spatial interactions between the zoning system and spatial patterns in the data were important but poorly understood at that time. Paper 5 (Cockings et al, In Press) provides evidence that it may be the scale of the building blocks relative to the output zones and to the spatial patterning of the phenomena of interest that is critical. Shuttleworth et al (2011, p. 12) suggest:

> "If the phenomenon of interest is structured at a spatial scale that is larger than the areal units then the results are not sensitive to the changing scale."

In the case of postcodes, street blocks and 2011 Census OAs, postcodes appear to be at a smaller or similar scale to the spatial patterning of the phenomena being represented, namely population, whereas street blocks are either slightly too large or the placement of their boundaries does not capture the spatial variation in population quite so effectively. The findings from Cockings and Martin (2005), however, suggest that just as employing building blocks which are too large is a problem, using ones which are too small and have no real spatial structure can be equally as problematic. The address-based building blocks used in that paper were too small, making it difficult for the automated zone design procedures to pick out higher-level spatial patterning. Taken together, the series of papers by the author (Cockings et al, In Press), Flowerdew et al (2008), Manley et al (2006) and Shuttleworth et al (2011)) provides a more nuanced exploration of the role of building blocks in the MAUP than that presented by Openshaw (1984), Openshaw and Rao (1995) and Openshaw and Taylor (1979). Exploration of relationships between the scale of variation in zoning systems and spatial data and their impact on the MAUP remains an area of research interest for the author and others (see, for example, Flowerdew, 2011; Grady, 2010).

Paper 5 (Cockings et al, In Press) clearly demonstrates that there are two distinct, but closely related, stages to automated zone design: the building block design stage and the automated zone design (or optimisation) stage. For various reasons, including data unavailability, licensing issues or simply lack of awareness of its importance, insufficient effort has previously been focused on the building block design stage. Martin (1998b) and Martin et al (2001) note that the shape of output zones can be controlled at both the building block design and optimisation stages. Paper 5 clearly demonstrates that the influence of building block design extends much further than just to the shape of zones: it has a bearing on all elements of the zoning system and should therefore be granted just as much, if not more, attention as the automated zone design stage.

Paper 5 (Cockings et al, In Press) highlights the importance of investing in appropriately designed building blocks for official population statistics. Its findings and methods fed directly into ONS' policy and practice for the 2011 Census in England and Wales, with postcodes ultimately being employed as the building blocks for the automated maintenance procedures used to create the residential OAs. Paper 5 also presents the first review of international practice in building block design and identifies key conceptual and practical issues to be considered when designing building blocks. With many national statistical organisations seeking stable geographies to aid comparison of statistics through time, an appropriate set of building blocks is of vital importance. Even in a world beyond traditional census data gathering processes, output zones based on appropriate building blocks are essential for facilitating the release of maximal amounts of data and for enabling effective data mapping and analysis. In fact, suitable building blocks may become even more crucial in a future in which greater reliance is likely to be placed on linked administrative and survey data rather than conventional census enumeration (ONS 2013a, 2013b). Zonal boundaries will be vital components of such a data system, providing key linking mechanisms and the base units for any sampling and estimation.

The combination of appropriate building block design and automated zone design techniques offer many potential statistical and financial benefits. The Australian Bureau of Statistics (ABS) (Pers comm, ABS) has already demonstrated how automated zone design can be utilised to create improved building blocks for purposes beyond census design: they employed the AZTool software to design a set of new Base Frame Units (BFUs), which are now the basic building blocks used for geographical stratification of ABS' surveys. The new, improved, BFUs contain homogenous population sizes at a suitable scale and are nested within higher level geographies. Statistics New Zealand has also evaluated the use of AZTool for redesigning more equally sized census areas (Ang and Ralphs, 2008). Automated zone design therefore offers great potential for a wide range of applications, beyond those demonstrated in this thesis.

1.6 Conclusions

Overall, the set of papers presented here represents a significant step forward in the design of zones for the spatial representation of population. It raises awareness of the importance of MAUP in the mapping and analysis of population statistics at the small area level. It also develops and evaluates novel methods of automated zone design and applies these to zoning systems which are of fundamental importance to a range of disciplines and applications.

The thesis provides a conceptual framework for the use of zones in health and environment studies and presents the first systematic evaluation of the role of MAUP in such investigations. It presents evidence of strong scale and weak aggregation effects in the study of deprivation and limiting long-term illness and suggests that administrative boundaries are not necessarily the most suitable zoning systems for analysing such data.

It extends and enhances previous automated zone design techniques to allow automated maintenance of existing zoning systems to be undertaken. Both the updating of a system which has become unfit for purpose, and the modification of a system to make it suitable for modelling another phenomenon, are made possible by these methodological advances. It shows that maintaining an existing system is a more constrained process algorithmically than designing a zoning system from new.

The thesis also undertakes the first systematic review of international practice in the design of building blocks, providing conceptual and practical guidance for their creation and evaluating two key sets of building blocks (postcodes and street blocks) for the design and maintenance of census output zones in England and Wales. Postcodes are shown to be the more effective building blocks for the latter purpose. The thesis provides evidence that building blocks are a crucial component of the zone design process: getting the building blocks right provides more flexibility and accuracy for the zone design process, facilitates data linkage and enables effective maintenance of zones through time. It recommends that more attention is paid to the building block design stage, not only in studies undertaken by individual researchers but also by national statistical organisations in the design of standard zoning systems.

The outputs from the thesis have made a substantive contribution to the policies and methods employed by the Office for National Statistics (ONS) to create the standard residence-based output geographies from the 2011 Census in England and Wales, as well as proposing an innovative methodology which has been used by ONS to produce a completely new zoning system for the publication of census workplace statistics. These zoning systems will be employed for at least the next ten years by an enormous range of users. The approaches, methods and recommendations from the thesis are also of direct relevance to other countries and applications, as evidenced already by uptake by academic researchers and statistical organisations in other countries including Australia, Canada, Finland, France, Mexico, Portugal, New Zealand and the USA.

There are, of course, limitations to the methods and analyses presented in this thesis. Many of these are discussed in the individual papers, but some issues of generic importance are extracted here.

In employing zones to represent population it must be remembered that such zones are artificially imposed fixed representations of population in both space and time, whereas in reality population is fixed in neither space nor time. This thesis therefore argues that there is not only a modifiable areal unit problem but also a modifiable spatiotemporal unit problem. This should be borne in mind when interpreting statistics for any zoning system. A natural extension to the methods presented here would be to move towards more continuous representations of time and space, using automated zone design to create varying zoning systems which change more regularly through space and time: there are few remaining technical barriers to achieving this, the challenges instead lie in understanding the concepts and processes and in accessing suitable data to represent such phenomena at appropriate spatiotemporal resolutions. In addition, although not directly relevant in the case studies addressed in this thesis, some population-related phenomena and processes, such as catchment areas or social networks, may be better represented by overlapping or non-contiguous zones. This is an area for further research: it is not clear whether AZP-based methods can usefully contribute to the design of such zoning systems.

There will always be tensions between producing zoning systems which provide consistency and stability through time and those which accurately represent contemporary population distribution. In areas where population has changed since the design of the original zoning system, it is not feasible to achieve both; instead a compromise, such as that presented in Paper 3, may prove to be the most useful approach. If true stability through time is required, grid squares or some other regularly shaped zones which do not change through time, are probably the most useful units for analysis. The disadvantages of regularly shaped zones are that they rarely align to real-world geographical features and usually exhibit greater internal heterogeneity than other building blocks, such as postcodes, which tend to more accurately reflect social and economic phenomena. There are also tensions between designing zones which are aesthetically pleasing e.g. compactly shaped or aligned with real-world geographical features, and zones which are statistically optimised e.g. having equal population sizes or being internally socio-economically homogeneous. The boundaries selected and drawn by manual processes are often different to those produced by automated procedures, with humans favouring smoother shaped zones aligned to geographical features, and automated procedures tending to result in more irregularly shaped, but statistically optimised, boundaries. This thesis has demonstrated that building blocks have an important part to play in the process of zone design: creating building blocks which are aligned with geographical features wherever possible but which also satisfy relevant statistical criteria may therefore represent the best way forward in terms of reconciling these competing design requirements.

The AZP-based methods considered in this thesis are just one of many possible approaches to automated zone design. Others include graph theory and clustering methods. While there are differences in the methodological and algorithmic approaches of these methods, they are all united in their aim to design optimised zones for specific purposes. The key differences tend to be in how they measure and retain spatial relationships and whether they adopt a linear or iterative approach. The specific implementation of AZP employed here is just one of many developed by a range of authors. All of the implementations by Openshaw, Alvanides, Martin and the author are consistent with the original AZP algorithm proposed by Openshaw (1977a, 1977b), albeit with various modifications to the algorithm. It would be useful to compare the performance of different algorithmic approaches to automated zone design and different software implementations (such as AZTool, SAGE and ZDES) in terms of the statistical and aesthetic qualities of outputs as well as algorithmic qualities such as speed, but unfortunately many of these are no longer operational. AZTool itself could be readily enhanced by adding extra functionality to handle other types of constraints, alternative metrics (e.g. for measuring shape or homogeneity) or other approaches to the IRA.

While this thesis tackles a range of empirical problems from a variety of applications, it represents only a limited exploration of the full range of problems which might be addressed by using automated zone design techniques. Only limited long-term illness and deprivation were explored in Paper 1 (Cockings and Martin, 2005) and only for a limited number of scales and aggregations: this could readily be extended to other conditions and other scales/aggregations; some of this work has already been pursued by other researchers. Only two sets of building blocks were investigated in Paper 5 (Cockings et al, In Press): this needs to be extended to include grid squares and other types of building blocks employed internationally. The current version of AZTool lends itself much more readily to the repetition of large numbers of runs in order to produce estimates of the full distribution of scale and aggregation effects in the MAUP. This would be particularly useful for applications such as risk modelling or emergency planning. Finally, other applications of automated zone design in environment and health studies (Paper 1), official populations statistics (Papers 2 to 5) and other fields could be explored: good starting points for the range of potential applications can be found in Paper 1 (Cockings and Martin, 2005), Alvanides (2000) and Alvanides et al (2002).

There are no longer any technical reasons why it should not be possible for users to employ automated zone design techniques to design their own geographies from individual level data for specific applications. This was a potential application of automated zone design highlighted many years ago by Alvanides et al (2002). However, the reasons why such practices did not become commonplace at that time, namely statistical disclosure and differencing concerns (Duke-Williams and Rees, 1998), are still the reasons why this is unlikely to happen today, although a move away from traditional census enumeration processes may yet see radical developments in this respect. In countries where such concerns are not so important, automated zone design techniques offer immense possibilities for designing more effective and useful zones for a range of purposes. In the absence of individual level data, users can, of course, always aggregate pre-aggregated data to create their own zones, as we have seen in this thesis.

One of the key challenges associated with using automated zone design is the evaluation of its outputs. Although AZP-based methods produce an optimised zoning system, given that the algorithm is heuristic it is never possible to be certain whether the true (globally) optimal solution has been created because it is impossible to know what that is! Openshaw noted that every solution will tend to be unique and will tend to converge towards the optimum, but as we do not know what the optimum is, we cannot tell how far the solution is from it. This means that there will always be an unknown degree of error in any zoning system. The author's own experience suggests that the extent to which individual runs produce different optimal solutions depends on the scale and spatial patterning of the phenomena being represented compared to the scale and patterning of the building blocks and output zones. These characteristics tend to control the degrees of freedom in the aggregation: some datasets and zoning systems, or particular parts of a study area, will always tend to converge towards the same solution; others will generate very different zonations for every run. The designer therefore needs to explore the specific dataset and its characteristics prior to undertaking zone design. Rather than expecting to find the one 'optimal' solution, perhaps it is more realistic to aim for a zoning system which is statistically optimised with relation to the design criteria but which is also meaningful and useful to those who need it, as these are the primary grounds on which it will be evaluated.

This thesis provides evidence that the scale and spatial patterning of phenomena and processes being represented compared to the scale and boundary placement of building blocks and output zones are absolutely critical in determining the statistical qualities of the zoning systems and their sensitivity to the MAUP. These characteristics should be explored prior to, during and after the zone design process. This is an emerging area of research which warrants further investigation.

Another avenue for further research is the extent to which scale variability is important in zone design. The scale and spatial auto-correlation of population-related phenomena and processes vary across space and through time. It is therefore not logical to assume that one scale is appropriate for all phenomena and processes in all places at all times. Designing a zoning system at a specific spatial scale may therefore not be the most appropriate approach to representing population. 'Scale' most commonly refers to geographical size, although in the papers presented here, it relates to population size within zones. By minimising differences in variables such as population size and socio-economic homogeneity between zones, we should, in theory, be producing zoning systems at varying, but more appropriate, geographical scales. These types of effects require further research.

> "The MAUP is not so much an insoluble problem but rather a powerful analytical tool ideally suited for probing the structure of areal data sets" (Openshaw, 1984, p.38)

The MAUP cannot be ignored: it influences all aspects of zone design and the use of zones in population studies. Automated zone design offers tools to both explore the impact of the MAUP on analyses and produce zoning systems which are optimised for specific purposes. Ironically, one of the factors which may be limiting wider used of automated zone design may be its very versatility. The fact that it is possible to achieve virtually limitless zonations from any set of data may mean that it is perceived by some as a form of 'gerrymandering'. This was particularly the case, for example, with some of Openshaw's early experiments with maximising correlations, which were regarded as unscientific by epidemiologists. This thesis argues strongly that all zone design should be based on clear conceptualisations of the geographical phenomena or processes being represented. Equally, if relationships between variables are being investigated, clear hypotheses should underpin the specification of zone design criteria.

Further factors limiting widespread application of automated zone design techniques may be ignorance of its importance in mapping and analysis or lack of awareness of the existence of freely available software packages for its implementation. If the aim is to induce a genuine step change in the levels of sophistication with which spatial data are modelled, the principles and practices of automated zone design may need to become just as much a part of a mainstream quantitative education as the MAUP.

Openshaw (1984, p34) presents a five step methodology for 'solving' the MAUP. The author argues that this process can be adapted and simplified to three steps, but that all of these steps should *always* be employed when designing zoning systems:

i) Explicitly define the purpose of the study. Clearly conceptualise the objects, phenomena or processes that you are trying to represent. Be clear about what the zoning system will be used for. Identify the characteristics that you want the zoning system to exhibit.

ii) Select appropriate datasets which represent the objects, phenomena and processes defined in step 1. Define an objective function which accurately captures the complexities of any rules or decisions required to produce the zoning system.

iii) Run the algorithm. Evaluate the results, both in terms of whether the algorithm has produced a meaningful and statistically optimised zoning system, and what it tells you about the spatial patterns and geography of the study area.

This thesis concludes that automated zone design methods should be a vital component of both the provision and analysis of zonal data. Automated zone design provides methods and tools for exploring and minimising the impact of the MAUP on studies involving spatially aggregated data. 'Good' zone design allows for the publication of maximal amounts of data and provides statistically robust and appropriate geographical zones for mapping and analysis. 'Bad' zone design potentially limits the amount of data that can be published and may result in inappropriate inferences being drawn from zonal data. Automated zone design can also aid the analysis of population trends through time by maximising the consistency and stability of zones through time. Building blocks are a critical component of the zone design process and should be afforded greater attention when designing zones, either for standard or bespoke zoning systems. Automated zone design should always be underpinned by sound theoretical reasoning and a clear conceptualisation of the phenomena or processes being represented.

1.7 List of references

Alvanides S (2000) *Zone design methods for application in Human Geography* Unpublished PhD thesis, University of Leeds

Alvanides S, Openshaw S and Duke-Williams O (2000) Designing zoning systems for flow data In *GIS and Geocomputation: Innovations in GIS 7* Eds P Atkinson and D Martin (Taylor and Francis, London) pp 115-134

Alvanides S, Openshaw S and Rees P (2002) Designing your own geographies. In *The Census Data System* Eds P Rees, D Martin and P Williamson (J Wiley & Sons, Chichester) pp.47-65

Ang L and 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

Blalock H (1964) *Causal inferences in nonexperimental research* University of North Carolina Press, North Carolina

Cockings S and Martin D (2005) Automated zone design for environment and health studies using individual-level data *Proceedings of the 11th International Medical Geography Symposium* 9-10

Cook L (2004) The quality and qualities of population statistics, and the place of the census *Area* 36 111-123

Curtis S and Riva M (2010) Health geographies I: complexity theory and human health *Progress in Human Geography* 34(2) 215-223

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

Flowerdew R (2011) How serious is the Modifiable Areal Unit Problem for analysis of English Census data? *Population Trends* 145 106-118

Flowerdew R, Feng Z and Manley D (2007) Constructing data zones for Scottish Neighbourhood Statistics *Computers, Environment and Urban Systems* 31 76-90

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

Gehlke C and Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material *Journal of the American Statistical Association* Supplement 29 169–170

Grady S (2010) Racial residential segregation impact on low birth weight using improved neighbourhood boundary definitions *Spatial and Spatio-temporal Epidemiology* 1 239–249

Grady S and Enander H (2009) Geographic analysis of low birthweight and infant mortality in Michigan using automated zoning methodology *International Journal of Health Geographics* 8 10

Haining R, Wise S and Ma J (1998) Exploratory spatial data analysis in a geographic information system *Journal of the Royal Statistical Society: Series D* 47 457–469

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

Horn M (1995) Solution techniques for large regional partitioning problems *Geographical Analysis* 27(3) 230–248

Johnston R (2002) Manipulating maps and winning elections: measuring the impact of malapportionment and gerrymandering *Political Geography* 21 1–31

Jones A, van Sluijs E, Ness A, Haynes R and Riddoch C (2010) Physical activity in children: Does how we define neighbourhood matter? *Health & Place* 16 236–241

Kendall M and Yule G (1950) *An introduction to the theory of statistics* (Griffin, London)

Manley D, Flowerdew F and Steel D (2006) Scales, levels and processes: Studying spatial patterns of British census variables *Computer, Environment & Urban Systems* 30 143–160

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 (1998a) Optimizing census geography: the separation of collection and output geographies *International Journal of Geographical Information Science* 12(7) 673-685

Martin D (1998b) 2001 Census output areas: from concept to prototype *Population Trends* 94 19-24

Martin D (2000) Automated zone design in GIS. In *GIS and Geocomputation: Innovations in GIS 7* Eds P Atkinson and D Martin (Taylor and Francis, London) pp 103-113

Martin D (2002) Geography for the 2001 Census in England and Wales *Population Trends* 108 7-15

Martin (2003) Extending the automated zoning procedure to reconcile incompatible zoning systems *International Journal of Geographical Information Science* 17(2) 181-196

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

Morphet C (1993) The mapping of small area census data – a consideration of the role of enumeration district boundaries *Environment and Planning A* 25 267-278

Northern Ireland Statistics and Research Agency (2013) *Small Areas for Northern Ireland: a new statistical geography for the 2011 Census data* http://www.nisra.gov.uk/geography/home.htm

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

Office for National Statistics (2012) *Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011* http://www.ons.gov.uk/ons/guide-method/geography/products/census/report--changes-to-output-areas-and-super-output-areas-in-england-and-wales--2001-to-2011.pdf

Office for National Statistics (2013a) *Beyond 2011: Options Report* http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html

Office for National Statistics (2013b) *Beyond 2011: Options Explained 2* http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html

Openshaw S (1977a) A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling *Transactions of the Institute of British Geographers NS* 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 (1978) An empirical study of some zone-design criteria *Environment and Planning A* 10 781-794

Openshaw S (1984) *The Modifiable Areal Unit Problem* CATMOG 38 http://qmrg.org.uk/catmog/

Openshaw S and Alvanides S (1999) Applying geocomputation to the analysis of spatial distributions. In: *Geographical Information Systems: Principles, Techniques, Applications and Management: Volume 1* Eds P Longley, M Goodchild, D Maguire and D Rhind (J Wiley & Sons, Chichester) pp 267-282

Openshaw S, Alvanides S and Whalley S (1998) *Some further experiments with designing output areas for the 2001 UK census* Working Paper 9, School of Geography, University of Leeds http://www.geog.leeds.ac.uk/papers/98-9/

Openshaw S and Rao L (1995) Algorithms for reengineering 1991 Census geography *Environment and Planning A* 27 425-446

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

Parenteau M and Sawada M (2011) The modifiable area unit problem (MAUP) in the relationship between exposure to NO2 and respiratory health *International Journal of Health Geographics* 10 58

Rainham D, McDowell I, Krewski D and Sawada M (2010) Conceptualising the healthscape: contributions of time geography, location technologies and spatial ecology to place and health research *Social Science & Medicine* 70 668–676

Riva M, Apparicio P, Gauvin L and Brodeur J (2008) Establishing the soundness of administrative spatial units for operationalizing the active living potential of residential environments: an exemplar for designing optimal zones *International Journal of Health Geographics* 7 43

Riva M, Gauvin L, Apparicio P and Brodeur J (2009) Disintangling the relative influence of built and socioeconomic environments on walking: the contribution of areas homogenous along exposures of interest *Social Science & Medicine* 69 1296–1305

Rossiter D and Johnston R (1981) Program GROUP: the identification of all possible solutions to a constituency-delimitation problem *Environment and Planning A* 13 231–238

Sabel C, Kihal W, Bard D and Weber C (2012) Creation of synthetic homogeneous neighbourhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France *Social Science & Medicine* http://dx.doi.org/10.1016/j.socscimed.2012.11.018

Sawicki D (1973) Studies of aggregated areal data: problems of statistical inference *Land Economics* 49, 109–114

Schuurman N, Hameed S M, Fielder R, Bell N and Simons R (2008) The spatial epidemiology of trauma: the potential of geographic information science to organize data and reveal patterns of injury and services *Canadian Journal of Surgery* 51(5) 389–395

Shortt N (2009) Regionalization/Zoning Systems. In *International Encyclopaedia of Human Geography* Eds R Kitchin and N Thrift (Elsevier, Oxford) pp 298 – 301 http://www.sciencedirect.com/science/referenceworks/9780080449104

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

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

Wise S, Haining R and Ma J (1997) Regionalisation tools for the exploratory spatial analysis of health data. In *Recent Development in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Neurocomputing* Eds M Fischer and A Getis (Springer, Berlin) pp 83-100

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

2. The Published Works

2.1 Paper 1: Zone design for environment and health studies using pre-aggregated data

Cockings S and Martin D "Zone design for environment and health studies using pre-aggregated data"

The version included in this thesis is the author's version of a work that was accepted for publication in *Social Science & Medicine*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Social Science & Medicine*, 60(12) 2729-2742 2005 http://dx.doi.org/10.1016/j.socscimed.2004.11.005

Zone design for environment and health studies using pre-aggregated data

Samantha Cockings* and David Martin

School of Geography, University of Southampton, Southampton, SO17 1BJ, UK

* Corresponding author: Tel: +44 (0)23 8059 5519 Fax: +44 (0)23 8059 3295

E-mail address: s.cockings@soton.ac.uk (S. Cockings)

Abstract. Many environment and health studies employ geographical areas as the units of analysis, either through choice or necessity. The design of these areas can greatly influence any observed spatial relationships or patterns – an effect known as the modifiable areal unit problem. In this paper we identify the phenomena and processes which are typically measured in environment and health studies and present a conceptualisation for their representation as data objects in spatial analysis. We discuss the circumstances under which we find ourselves using areas for representation and outline the application of zone design techniques for the creation of such areas in environment and health studies. An empirical study of the relationship between deprivation and limiting long-term illness in the former county of Avon, UK, is employed to demonstrate the potential usefulness of zone design techniques for creating zones with stable estimates and for exploring the sensitivity of relationships to changes in the zoning system. In particular, we illustrate the inappropriateness of the 1991 Census enumeration district and ward zoning systems for such an analysis and conclude that automatically designed aggregations may be a more appropriate basis for analysis than any pre-existing zoning system.

Keywords: zone design, environment and health, modifiable areal unit problem, aggregated data, deprivation, limiting long-term illness

1. Introduction

Despite acknowledgement that health effects (morbidity or mortality) operate fundamentally at the level of the individual, much research concerned with environmental impacts on health involves the spatial analysis of areal data. This may be through necessity, where the individual level is inaccessible as a result of data or confidentiality constraints, or by choice, when there is an explicit research interest in area-level effects. ‘Environment’ in the context of this paper is taken to include the natural, social and built settings experienced by individuals. We here use the term ‘environment and health’ to encompass several research traditions variously described as spatial epidemiology, environmental epidemiology, geographical epidemiology and environmental health. While recognizing the different emphases associated with these approaches, we are here seeking to focus on their common concern with the handling of areal units and spatially aggregated data. Spatial aggregation is the aggregation of individual observations over geographically-defined units (e.g. census wards), as distinct from aggregation by non-geographic attributes (e.g. age, social class). This is undertaken for many reasons, for example, because the measurement of individual-level exposure to atmospheric pollution is usually limited to small numbers of participants over relatively short timescales; because disclosure control considerations may restrict or prevent access to individual-level data by researchers or because policy-oriented research may be constrained to those areal divisions through which services are delivered and policies implemented. There is evidence of the existence of area-level effects over and above those operating at the level of the individual (see for example Macintyre (1997); Macintyre, Ellaway & Cummins (2002); Mitchell, Gleave, Bartley, Wiggins & Joshi (2000); Pickett & Pearl (2001) and Shaw, Dorling & Mitchell (2002)). Such an explicit interest in area effects may lead to measurement or modelling at the area level by choice. Whichever the cause, the result is an analysis in which some or all of the input data have been aggregated over geographical zones. Despite the difficulties associated with conducting analysis using areally aggregated data, particularly the modifiable areal unit problem (MAUP) (Openshaw, 1984), very few environment and health studies give explicit consideration to the design of the zoning systems used.

From the earliest recognition of the MAUP, one suggested solution has been to control the design of the zoning system so as to create the most robust aggregation for the analysis to be undertaken (Openshaw, 1984). Automated zone design techniques have been developed for this purpose (Openshaw, 1977; Openshaw & Rao, 1995). However, their application has been limited, and rarely utilised in current research concerned with environment and health. This paper outlines the concepts underlying these techniques and seeks to illustrate their potential use in such studies. We are particularly concerned with issues associated with data already aggregated to pre-defined areal units. It is recognized that there are research contexts in which aggregation is performed from individual records, but that situation presents additional design challenges which are beyond the scope of this paper.

The rest of the paper comprises five sections. In section 2 we provide an overview of the spatially referenced phenomena and processes which are most frequently encountered in environment and health studies and suggest a framework for both the conceptualisation and measurement of these phenomena. In section 3 we review the methods available for zone design and explore their applicability in these types of research. In section 4 we present an empirical example for a county in the south west of England where zone design is employed to explore the sensitivity of the relationship between morbidity and deprivation to pre-defined and alternative zoning systems. In section 5 we discuss the results of this empirical study and draw conclusions for the application of zone design techniques in environment and health studies more generally.

2. Spatial representation in environment and health studies

This section is concerned with the types of spatial phenomena and processes encountered in environment and health studies and the ways in which these are typically represented as spatial data objects. In particular, we explore the way in which the representation of phenomena and processes as data objects often differ from their conceptualisation and we discuss the reasons why researchers frequently find themselves working with areally aggregated data.

2.1 Spatial phenomena and processes in environment and health studies

Fundamentally, environment and health research is concerned with describing and explaining the effects of the environment on health. Figure 1 presents a simplified conceptualisation of the phenomena and processes that we typically seek to measure and represent in such studies. A health "effect" experienced by an individual is likely to be the result of a combination of factors illustrated by the two lower layers in the figure.

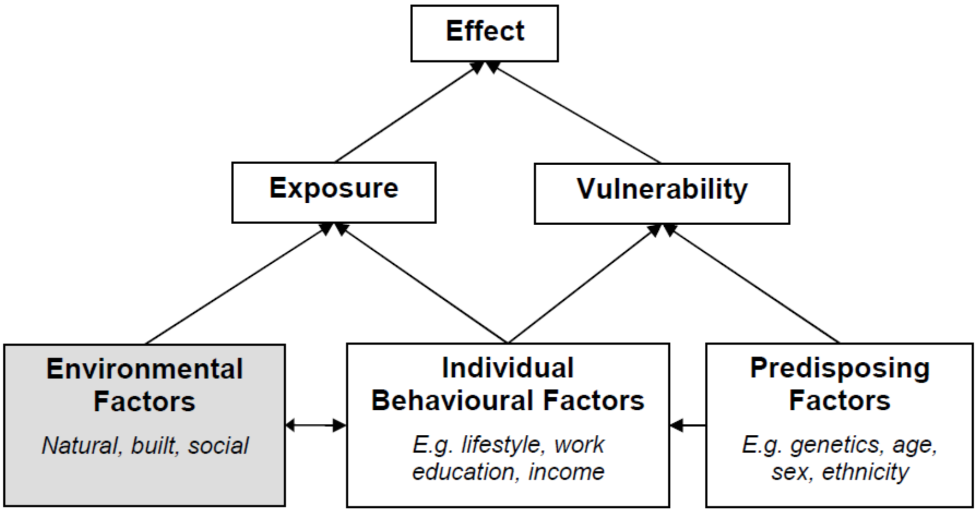

Effect
Exposure
Vulnerability
Environmental Factors
Natural, built, social
Individual Behavioural Factors
E.g. lifestyle, work education, income
Predisposing Factors
E.g. genetics, age, sex, ethnicity

Figure 1 Spatial phenomena and processes in environment and health studies

All individuals experience a range of risk factors which operate at the individual level and the conceptualisation of these is complex and frequently revisited (Andersen, 1995). Some, such as age, sex, birth weight, genetics and ethnicity, are beyond the control of the individual and may be termed "predisposing" factors in the sense used by Beaglehole, Bonita & Kjellström (1993, p. 74). Others, which may be termed "behavioural" or "lifestyle" factors (Gatrell, 2002, p. 113), are related to active decisions made by an individual. Examples of behavioural factors include smoking, diet and education.

We use the term "environmental factors” to refer to features of the natural, built and social environments, which are external to the individual. Examples of natural features include radon (Darby, Whitley, Silcocks, Thakrar, Green, Lomas et al., 1998) and volcanoes (Buist & Bernstein, 1986). Features of the built environment include landfill sites (Elliott, Briggs, Morris, de Hoogh, Hurt, Kold Jensen et al., 2001) and road traffic (Wilkinson, Elliott, Grundy, Shaddick, Thakrar, Walls et al., 1999). Social environmental factors include area-based resource allocation (Bentley, 2003) and neighbourhood disadvantage (Carstairs, 2000). While environmental factors are conceptualised as areally extensive, many, such as unemployment and atmospheric pollution, are continuously varying over geographical space and it is therefore difficult to place boundaries which have real meaning for representation of these underlying phenomena. Some phenomena which have genuine geographical boundaries, such as area-based health authority resource allocation, are in fact delivered to individuals through their interactions with the health care system. In this example a truly areal factor may be diluted by differences in individuals' utilisation of health care. A further example is that of water quality which, although conceptualised as an area-level effect of water supply zones, in reality varies at the point of delivery (the tap) for many reasons, including the water's residence time in the distribution system (Keegan, Whitaker, Nieuwenhuijsen, Toledano, Elliott, Fawell et al., 2001). Despite these representational difficulties, we suggest that environmental factors are not features of the individual and should still be conceptualised as areal phenomena. It is with these phenomena (indicated by the shaded box in Figure 1) that the rest of this paper is primarily concerned.

Although predisposing and behavioural risk factors fundamentally operate at the individual-level they are also influenced by area-level factors. For example, smoking behaviour is influenced by social context (Kleinschmidt, Hills & Elliott, 1995; Pickett, Wakschlag, Rathouz, Leventhal & Abrams, 2002; Frohlich, Potvin, Gauvin, & Chabot, 2002), educational achievement is related to social disadvantage (Conduit, Brookes, Bramley & Fletcher, 1996), and even predisposing factors such as birth weight are indirectly influenced by area-level effects because a mother's diet and health during pregnancy can be related to the area and social context in which she lives (Barker, 1998). Individuals with similar predisposing and/or behavioural factors often cluster together and this may in turn lead to area-level effects, which can influence an individual's health. Macintyre (1997) suggests a three-way classification of phenomena influencing people's health: compositional, contextual and collective. In a later paper, Macintyre, Ellaway & Cummins (2002) discuss the difficulties involved in conceptualising and measuring such features. They conclude that it is inappropriate to separate out collective features from contextual ones as they are so closely interlinked. The most appropriate way of dealing with collective and contextual features would appear to be an important area for further research, but is beyond the scope of the present study. Nevertheless, we suggest that both operate above the level of the truly individual factors and, in a similar vein to Macintyre et al. (2002), we therefore place both within our 'environmental factors' category in Figure 1.

An individual's activities and decisions through space and time influence the degree to which they experience behavioural risk factors and the extent to which they accumulate "exposure" to environmental factors. In this context, we define exposure as contact between an individual and an environmental risk factor. This exposure, together with the individual's vulnerability or susceptibility (which is strongly influenced by predisposing factors and their previous behaviour and exposure history) influences their likelihood of developing a particular health effect.

We recognise that the conceptualisation of environment and health relationships presented in Figure 1 does not adequately capture the complex interactions or feedback mechanisms which occur between factors. It also does not contain an explicit temporal dimension which is important when considering the lifecourse experiences of individuals (Kuh & Ben-Shlomo, 1997) or when attempting to assess exposure to environmental factors (Briggs, 2000). Nevertheless, we suggest that it is a useful summary of the ways in which these phenomena and processes are conceptualised in many published environment and health studies and it therefore has utility as a framework for our discussion of the spatial representation of environmental factors (highlighted in Figure 1).

2.2 Data representation in environment and health studies

Let us now consider how the phenomena or processes described above are represented as spatial data objects in analyses. We suggest that all the phenomena and processes in Figure 1, other than environmental factors, operate at the individual level and as such should ideally be represented by individual level data. It can thus be argued that environmental factors are the only phenomena which should be represented as areas. Yet, for many reasons, we frequently find ourselves also representing the individual level phenomena using areas. This is primarily due to the impracticality of measuring exposure and behavioural factors at the individual level, together with confidentiality or disclosure concerns which necessitate the aggregation of individual data prior to publication. Other reasons include the fact that many policy decisions are implemented at the area level and that the visualisation of results can be more effective for areas. Table 1 illustrates the situations in which we use areas. The key phenomena or processes are shown, together with ways of representing these – either as individual or areal-level data. For each phenomenon or process, we place a 'P' (phenomenon/process) in either the individual or area column depending on the level at which we believe it operates. We then place a 'D' (data) in either or both of the individual or area columns to indicate how we typically represent it as a data object in an analysis. Consider, for example, an assessment of the role of ethnicity in the development of leukaemia. Ethnicity is an individual predisposing risk factor so we would place a P in the individual/predisposing factors cell. But, in the UK, either for reasons of confidentiality or due to a lack of individual data, we frequently represent ethnicity as counts per area e.g. counts per census ward. In this example, our data representation of ethnicity would therefore be indicated by a D in the area/predisposing cell. The table illustrates how our representation of the phenomena or processes as data objects in an analysis can differ from how we might conceptualise them. The highlighted row and column in Table 1 correspond to the occasions when we use areas for an analysis: either we use them because we have an explicit interest in area-level environmental phenomena (the highlighted horizontal row); or because we choose to represent an individual level phenomenon at the area level or because data are only available at the area level (the highlighted vertical column). We have already noted the difficulty with placing collective features within these discussions. For some commentators, these features may more appropriately appear at * in Table 1, but our approach is to treat them as a feature of the social environment. It is essential to understand that in all cases where areal data are used, the design of the areal units, both in terms of scale and aggregation, will influence any observed relationships. Section 3 outlines the principles and applications of zone design and suggests ways in which it may be useful in studies of the environment and health.

Table 1 Spatial representation of processes and phenomena in environment and health studies and the role of zone design methods

	Individual		*Area*	
Effect	P	D	-	D
Exposure	P	D	-	D
Vulnerability	P	D	-	D
Predisposing factors	P	D	-	D
Individual Behavioural factors	P	D	* -	D
Environmental factors	-	-	P	D

P = processes/phenomena
D = data representation

3. Zone design

For the purposes of this paper, zone design refers to the placement of areal unit boundaries. In many contexts, the locations of these boundaries are determined by manual decision-making driven by organisational needs. For example, health care delivery authority boundaries in the UK represent a mixture of physical features such as coastlines or principal roads, existing political divisions such as local government areas and related service delivery areas such as those of community social services providers.

Zone design generally involves the application of a series of design principles to a set of elemental areal units, although the process is usually pragmatic and contested and is rarely formally defined in these terms. Designers frequently have in mind some elemental areal units (for example, blocks defined by street intersections or the smallest local political divisions) which they are reluctant to subdivide, and a series of loosely defined rules (for example the number of zones to be created or target population sizes which they are to encompass). In the UK context, the manual application of these principles results in zones which differ widely in characteristics such as denominator population size and composition but which are used as the basic units for publication of statistical data essential to the understanding of environmental impacts on health. 1991 Census enumeration districts (EDs) in the UK were designed in exactly this way (Clark

& Thomas, 1990) but, together with larger geographical units such as wards or local authority districts derived from the aggregation of EDs, have subsequently been used as the basis for many analyses in epidemiology and environmental health (see for example Elliott, Shaddick, Kleinschmidt, Jolley, Walls, Beresford et al., 1996; Haynes & Gale, 1999; Middleton, Gunnell, Frankel, Whitley & Dorling, 2003). Originally intended for the purposes of census enumeration, these EDs display wide variations in population size, geographical shape, area and social composition, and can be neither aggregated nor disaggregated neatly to any level of the postcode geography which has been extensively used for the georeferencing of health event data (including mortality data and hospital episode statistics). Further, some EDs proved to contain only small populations. In order to provide adequate disclosure protection, the counts for these EDs were suppressed and instead combined with those of neighbouring zones.

Automated zone design refers to the implementation of zone design procedures by automated means. Boundary placement is controlled by statistical design rules and computationally intensive procedures are employed to derive zoning solutions which are in some sense ‘optimal’ for a particular application. The task of assembling small geographical building blocks into larger regions so as to control population size has been of particular interest in political districting in the United States (see for example, Horn, 1995). Openshaw (1977) proposed a general purpose automated zoning procedure (AZP) based on the iterative recombination of building blocks into output areas so as to maximise the value of an objective function, and this methodology is further developed by Openshaw and Rao (1995). AZP has been developed and applied by Martin (1998) to the specific task of designing output areas (OAs) for the 2001 Census of population in England and Wales (Martin, Nolan & Tranmer, 2001). As yet, there are few published applications in fields beyond political districting and census output area design (Openshaw & Alvanides, 1999; Martin, 1998), but automated zone design would appear to offer a rich set of concepts and tools for applications, such as environment and health studies, where there is a desire or need to use areas for analysis.

We suggest that zone design methods may be useful for environment and health studies whenever we wish to use areas for the representation of phenomena or processes; essentially when the study involves any of the highlighted cells in Table 1. Table 2 presents a range of research aims, together with illustrative zone design criteria which might be potentially relevant in each case. This list is not intended to be exhaustive and raises many conceptual and methodological questions but it serves to illustrate the impact of zone design considerations on research design and implementation.

Table 2 Illustrative zone design criteria for different research aims

Research aim	*Criteria for zone design*
Hypothesis testing	Maximise internal homogeneity of risk and/or confounding factors
Assess non-stationarity	Maximise internal homogeneity of correlation between variables
Visualisation/exploratory analysis of spatial patterns of disease	Maximise internal homogeneity of disease rates/ratios or other relevant measures
Stability of estimates/power to detect relationships	Thresholds and/or targets for numerator and/or denominator and/or other relevant measures
Intervention (policy formulation and implementation)	Service delivery zones Zones representing groups at risk Zones representing neighbourhoods Zones of maximum effect/efficiency/ equity or other relevant measures
Disclosure control	Numerator and/or denominator threshold(s)

A potential application is the testing of hypotheses of deterministic or causal links between variables. In this context, zone design techniques may be used to create zones which maximise the internal homogeneity of the independent variable(s) within zones. For instance, if we hypothesise that living in a deprived area is a risk factor for the development of a limiting long-term illness (LLTI) we could aim to create zones which are as internally homogeneous as possible in terms of deprivation. If the hypothesis is true then we would expect the resultant correlation between the independent variable (deprivation) and the dependent variable (LLTI) for the newly aggregated zones to be strong. Essentially, this is a form of stratification at the design stage of the analysis. If there is also a possibility of confounding in the relationship, we might go further and design zones which aim not only for maximum internal homogeneity of the independent variable but also for a uniform dispersal of the confounding variable across the zones.

A closely related but conceptually different application of zone design techniques is to maximise the internal homogeneity of correlation between the hypothesised independent and dependent variables, as illustrated by Openshaw and Alvanides (1999). In this way, zone design methods could be used in an exploratory capacity, similar to the geographically weighted regression (GWR) methodologies developed by Fotheringham, Brunsdon and Charlton (2002), in order to provide information concerning how the strength of the hypothesised relationship varies over space. This is a potentially important use of zone design not least because it acknowledges the role of locality in environment and health relationships and is a way of exploring or confirming the non-stationarity of such relationships over space (Fotheringham, Brunsdon & Charlton, 2000). If non-stationarity exists, then it is likely that other variables are playing a part in the observed relationships. Identifying areas where the relationships are particularly strong or weak through the use of zone design may help to identify these variables. Note that this approach is different to that where the aim is to maximise the *global* correlation between two variables, which may also be useful in providing a summary of the relationship between the variables for the entire study area.

In many studies we wish to explore or to visualise the spatial distribution of health effects rather than to test a specific *a priori* hypothesis. In many of the applications proposed here, we are essentially using zone design as a form of data reduction where we aim to ensure parsimony (minimum number of units without loss of information) (McClelland & Kronmal, 2002; Morris & Munasinghe, 1993). In this context, zone design could be used to identify areas of high and low disease rates with a view to informing the delivery of health care or to aiding the generation of causal hypotheses. In such cases, zone design might be employed to derive sets of zones which maximise the internal homogeneity of the relevant measure of ill-health. Such exploratory applications can serve to enhance visualisation but, if used inappropriately, may be criticised for their post hoc development of hypotheses, akin to the Texas sharpshooter approach described by Rothman (1990). Zone design might also be used to improve the stability of estimates or to ensure a specified power to detect relationships. This is particularly relevant in investigations involving rare diseases, such as childhood leukaemia or congenital malformations, where small number effects may make interpretation difficult or meaningless. For example, one might design zones with a minimum (threshold) population or an ideal (target) population.

In the context of zone design for service delivery and policy making, the design criteria might include minimum and target populations, internal homogeneity of needs, or possibly zones of maximum efficiency, effect or equity of interventions. In more pragmatic terms, zone design techniques might be used to ensure that confidentiality thresholds are upheld when publishing health and health-related data, as has been the case with the 2001 Census OAs.

In summary, one of the major advantages of using zone design methods for environment and health studies may be as a means by which we can explore the sensitivity of observed relationships or patterns of ill-health to changes in the zoning system. In applying zone design methodologies we are forced to acknowledge the existence of the MAUP and we are able to explicitly and systematically explore its effects. In investigations where only pre-aggregated hierarchical area-level data are available, zone design methods also allow us to comment on the appropriateness or robustness of such zones for analysis. In the next section, we demonstrate the use of zone design techniques in an empirical example which explores the sensitivity of the relationship between morbidity and deprivation to pre-defined and alternative zoning systems using pre-aggregated data.

4. Empirical study: limiting long-term illness and deprivation in Avon

4.1 Background

Recent work (Martin, Brigham, Roderick, Barnett & Diamond, 2000; Barnett, Roderick, Martin, & Diamond, 2001; Barnett, Roderick, Martin, Diamond & Wrigley, 2002) has suggested that there is a strong relationship between morbidity and deprivation in the south west of England, but that the relationship is weaker in rural areas than in urban areas. This work used the administratively-defined hierarchy of census EDs, parishes, wards and local authority districts. There is a general requirement to standardise ward populations within a single district, but individual districts tend to be predominantly either rural or urban in nature, resulting in widely differing ward population sizes between urban and rural areas. For example the mean ward 1991 population within the

City of Bristol was 10,435, while in the neighbouring Wansdyke district it was 2,562. This structural feature of the administrative zoning system complicates interpretation of any spatial analysis in which there are potential urban-rural differences, as we are effectively using different geographical scales, under the overall description of 'ward-level' analysis. In many of the rural areas the calculation of rates is subject to small number problems. Haynes and Gale (2000) have explored the issue of aggregation in a case study in East Anglia. They have shown that by aggregating rural wards to create zones of approximately equal population size, the correlations between mortality and morbidity and unemployment (their chosen proxy for deprivation) in rural areas become more similar to those found in urban areas. However, the method of aggregation employed in this context did not require that the aggregated areas be adjacent to one another, obscuring the likely existence of area-level effects.

The present study aims primarily to examine the sensitivity of the relationship between morbidity and deprivation to the choice of zoning system. We are not concerned with questions of rurality *per se*, but rather with the way in which the observed relationships between the variables change as different zone designs are adopted. More specifically, we aim to demonstrate how zone design techniques might be used to create more stable estimates and to compare the robustness of the pre-defined census units with alternative zoning systems designed specifically for such an analysis.

4.2 Data

We have used the Townsend score (Townsend, Phillimore & Beattie, 1988) as a measure of deprivation which has been widely used in analyses of deprivation impacts on health, including the previously cited work on the south west of England. The Townsend score comprises of four variables: the proportion of people without a car, the proportion of households in overcrowded accommodation, the proportion of households not owner-occupying and the proportion of people unemployed, each based on total populations or households. We measure morbidity using self-reported LLTI obtained from the 1991 Census. In terms of our previous discussion, we are here using aggregated morbidity data (LLTI), aggregated predisposing factors (age and sex) and an area-level environmental factor (deprivation). This example is typical of investigations undertaken either because individual level data are not available for confidentiality reasons or because there is a need to use pre-defined administrative areas in order to inform policy decisions.

The analysis focuses on the former county of Avon (as defined at the time of the 1991 Census), a subset of the Barnett et al. (2001; 2002) study area. The total population of Avon is 928,423 and the county covers the full hierarchy from a large city (Bristol) through to small scattered rural settlements. Population totals and counts of persons experiencing LLTI for each of 10 age-sex bands covering the population aged 0-64 and constituent counts of the Townsend deprivation score were all retrieved at the ED level from the 1991 Census Small Area Statistics (SAS).

4.3 Methods

Our approach has been to take EDs as building blocks and to undertake repeated redesign of the Avon zoning system at different scales and aggregations, thus exploring both the scale and aggregation aspects of the MAUP (Openshaw, 1984) in relation to the deprivation-health relationship for this particular dataset. We are particularly interested to examine the way in which the observed relationship between the Townsend score and age-sex standardized LLTI rate changes as the zoning system is altered, and to understand the relative performance of the standard ED and ward aggregations within this spectrum. Meaningful use of age-sex standardized LLTI rates and Townsend scores in this context requires us to completely recompute these values for every new zone in each zoning system considered. Calculation of both measures involves standardisation and for this purpose we use England and Wales ED means and standard deviations throughout these experiments. In addition to the composite Townsend score, we examine the association between LLTI and the four Townsend components individually for each aggregation.

Zone design is undertaken using the AZM software described by Martin (2003). Although designed primarily to search for the best match between two incompatible systems of areal units, the software incorporates Openshaw's (1977) AZP and can thus be applied to more general zone design problems. Avon ED boundary data preparation in the ArcInfo GIS involved the removal of one uninhabited but genuine island polygon and several artificial unpopulated part-ED polygons caused by the intersection of administrative boundaries and the mean high water line. Arc attribute information was exported from ArcInfo and used in the iterative reaggregation of EDs to create output areas with our specified target characteristics. The output from each run takes the form of a constitution list, describing the membership of the newly-created output areas as a list of input EDs. This is sufficient to allow the generation of statistical data for each output area without the necessity for every configuration to be mapped and examined visually. For any zoning system of interest, the constitution list may be re-imported to ArcInfo and ED boundaries dissolved to create a map of the new geography. The term 'output areas' refers specifically to this analysis and should not be confused with output areas created for the publication of 2001 Census data.

AZM provides various constraints and values for the control of zone design. For this series of experiments a simple shape statistic (perimeter^2/area) is minimized, as is the sum of squared differences between actual output area counts of population aged 0-64 and the chosen target value. We shall refer to this count as Pop64 to avoid confusion with total population in the following discussion. A minimum threshold for Pop64 has been set at 90% of the target value in each case, and the best result after 50 random re-starts is used for analysis. Setting a threshold close to the target reduces the variation in acceptable zone sizes and thus aids in the production of alternative zoning systems at predetermined scales. The choice of these design constraints and target values is subjective, but serves to ensure consistency between our different zone design runs. The values were selected to achieve reasonably compact and uniformly-sized zones for analysis in a consistent and reproducible way. Note that there is no homogeneity design constraint specified in this analysis.

We have re-zoned the county with the target for Pop64 set at values from 250 to 4,500 in increments of 250, thus covering the range from ED to ward scales. The lowest of these targets cannot actually be achieved using EDs as building blocks, but the effect is to aggregate very small EDs to achieve values above the threshold, while preserving most of those that are already above the target. The mean size of the resulting output areas is thus greater than the target size. We also include the actual EDs and wards in our results for comparison. At each of four representative target sizes (500, 1,500, 2,500 and 4,000) we have also produced 10 different zoning systems by selecting different random starting configurations and setting the minimum threshold for Pop64 to 80% of the target value. This effectively allows us to explore the aggregation element of the MAUP by considering the sensitivity of the results to different zone designs at each chosen scale. For each new output area, the constituent ED data are aggregated, the LLTI and Townsend scores recalculated, and a range of diagnostic statistics examined in order to assess the variation in zone sizes and the effects on the deprivation-LLTI association.

4.4 Results

There are 1,970 EDs and 177 wards, displaying all the characteristic weaknesses of administrative zones as statistical reporting units. Their key characteristics in terms of total population and Pop64 are summarized in Table 3. The ED populations range from 0 to 1,416 with 16 zones having no recorded population (populations below 50 being suppressed for reasons of census confidentiality). The 177 wards all have non-zero population counts, but display a population range of 16,803. EDs and wards both present highly variable denominator population sizes, with systematic differences apparent between rural and urban areas and between the various local authority districts (Figure 2) reflecting administrative considerations as discussed in section 4.1 above. Figures 3 and 4 show Townsend scores and standardized LLTI ratios, both by wards. The general picture is of closely related distributions, albeit with differences in detail.

There is a strong urban-rural differential, with the greatest concentrations of both measures occurring in inner suburban areas.

Table 3 Population distribution summaries: Avon EDs and wards, 1991 Census

	EDs		*Wards*	
	Total pop	*Pop aged 0-64*	*Total pop*	*Pop aged 0-64*
Number of zones	1970	1970	177	177
Mean count	471	392	5245	4364
Minimum count	0	0	52	43
Maximum count	1416	1321	16855	14333
Standard deviation	143	135	3511	2947
Zero count zones	16	16	0	0

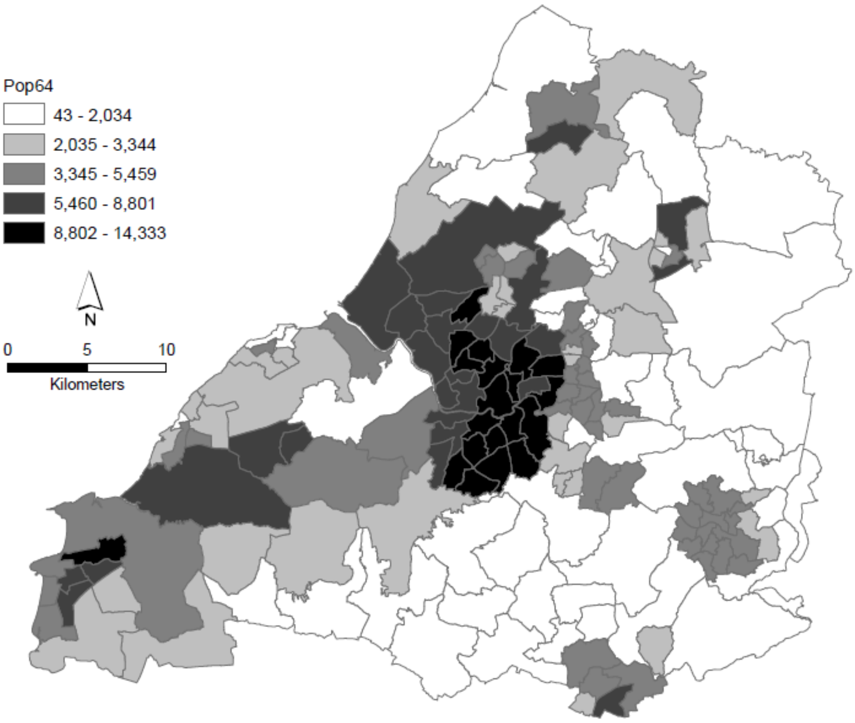


Figure 2 Population aged 0-64 (Pop64) in 1991 Census wards

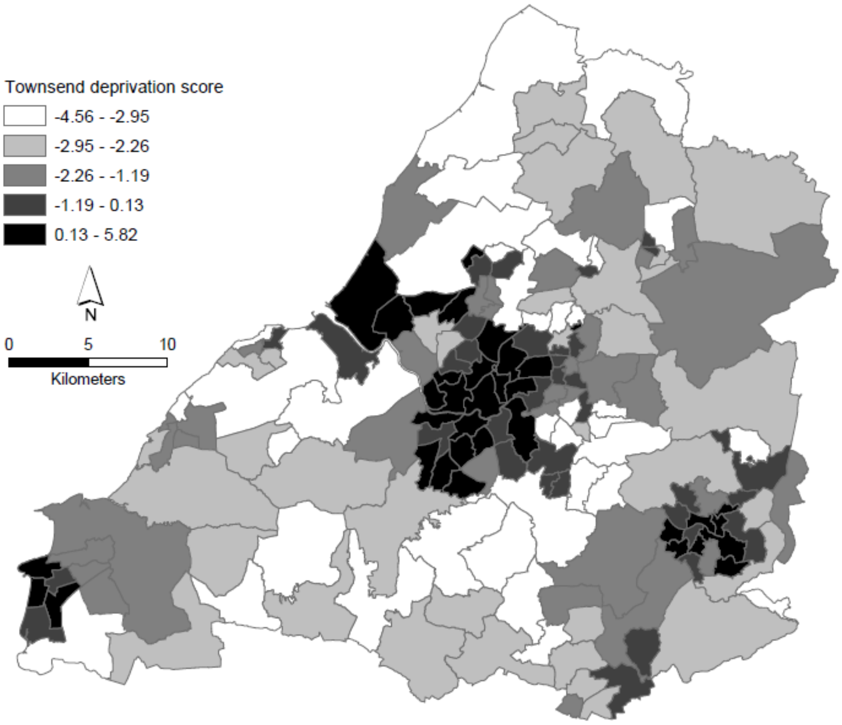


Figure 3 Townsend deprivation score in wards

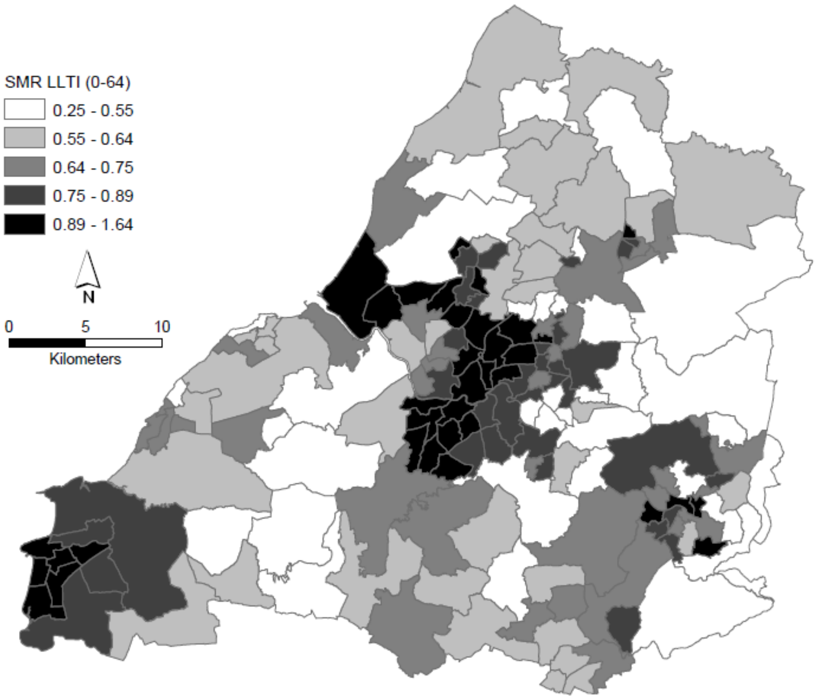


Figure 4 Standardised Morbidity Ratio (SMR) for LLTI 0-64 in wards

Figure 5 shows an illustrative zoning scheme from AZM with a mean Pop64 size of 4,291 and a Townsend-LLTI correlation of 0.883 (the target population size for this run was 3,750). This is closely equivalent to the ward 'scale' of analysis illustrated in Figure 2 (mean Pop64 = 4,364). Note the differences in the range of Pop64 values in the two figures, which should also be read alongside the histograms in Figure 6. The figures and histograms show that the Pop64 and LLTI 0-64 counts from the zone design run are much more narrowly distributed than those from the wards. These general distributional characteristics are present at all scales of zoning system produced by AZM. Further, the ED and ward zoning schemes display the greatest internal variations in zone sizes at the relevant scales, leading to instability in derived mapping and analysis: the range of Pop64 across Avon wards is 14,290 whereas the range in the nearest-matching AZM runs is only 3,000. All zero-value EDs are combined with neighbouring EDs by the zoning procedure resulting in distributions which are entirely above the Pop64 thresholds chosen.

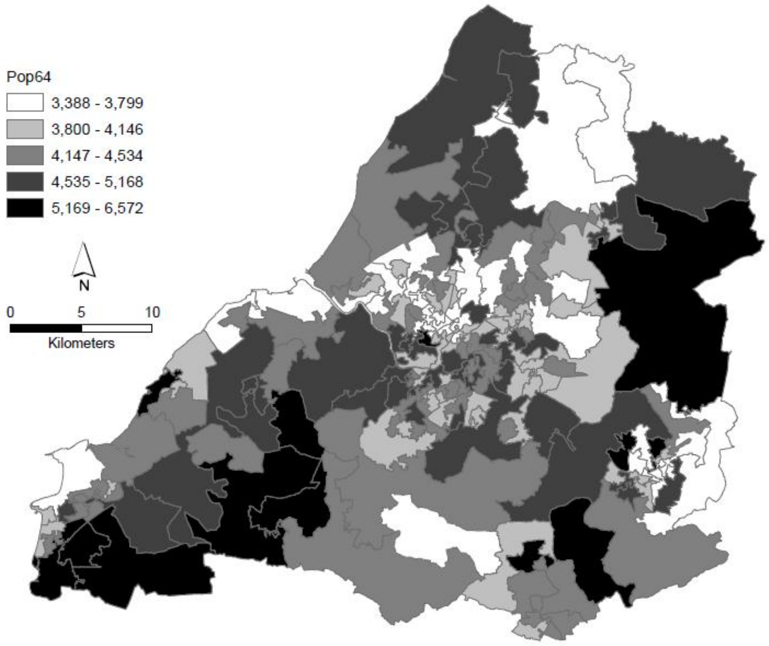


Figure 5 Pop64 in output areas from illustrative AZM zoning system (target Pop64 3750; mean Pop64 4,291)

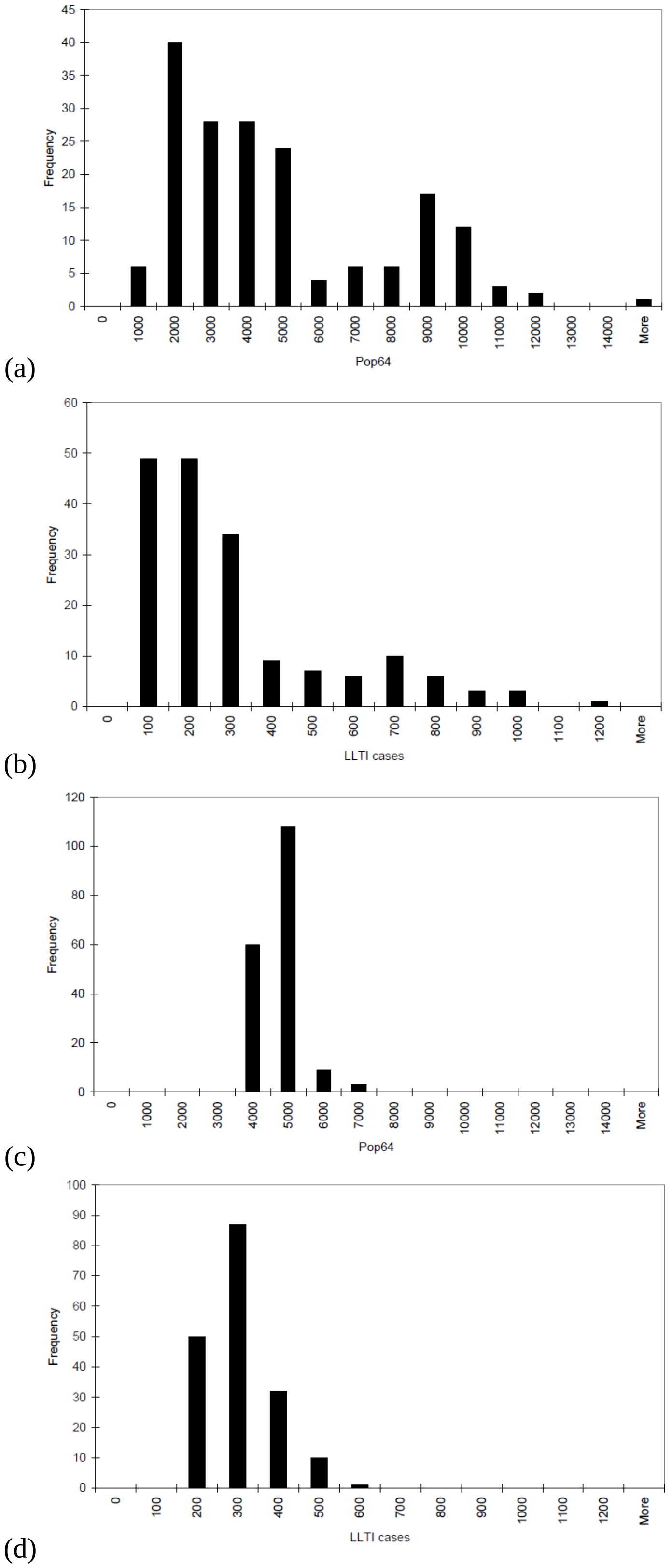


Figure 6 (a) Ward Pop64 distribution (b) Ward LLTI 0-64 distribution (c) Illustrative AZM zoning scheme Pop64 distribution (d) Illustrative AZM zoning scheme LLTI 0-64 distribution

The effect of these more stable distributional characteristics is to increase the observed correlations between LLTI and the composite Townsend score and also each of its individual components. These results are summarized in Figure 7 and it is apparent that the Pearson correlation coefficient strengthens steadily with increasing output area size (measured in terms of Pop64). The value of 0.724 for the correlation between LLTI and Townsend in the 1,540 non-zero EDs is the weakest association of all the zone designs, and the ward value of 0.861 is also below that of the AZM-derived aggregations at similar scales. Remarkably, the individual unemployment component is more strongly associated with LLTI at most scales within Avon than the composite Townsend score. It is important to note that this is despite the fact that in this experiment there was no attempt to design zones that were internally homogeneous in terms of deprivation (the zones were designed purely on population size and zone shape criteria), these increased correlations are therefore primarily a function of the more homogeneous zone population sizes rather than any deliberate efforts to maximise the correlations.

Even when the scale element of the zone design process is effectively fixed, simply by placing the boundaries differently at these scales we still see a notable degree of variation in the observed correlations. Figure 8 shows the results of repeated re-zonings at four scales. The variation in correlation between the composite Townsend deprivation score and LLTI at specific scales tends to be greatest for the larger (higher Pop64) zones, and appears to be smaller around the 1,500-2,000 Pop64 zone size. Further research is needed to explore whether this trend holds at other scales and also to see if the same trend applies for correlations between the components of Townsend and LLTI.

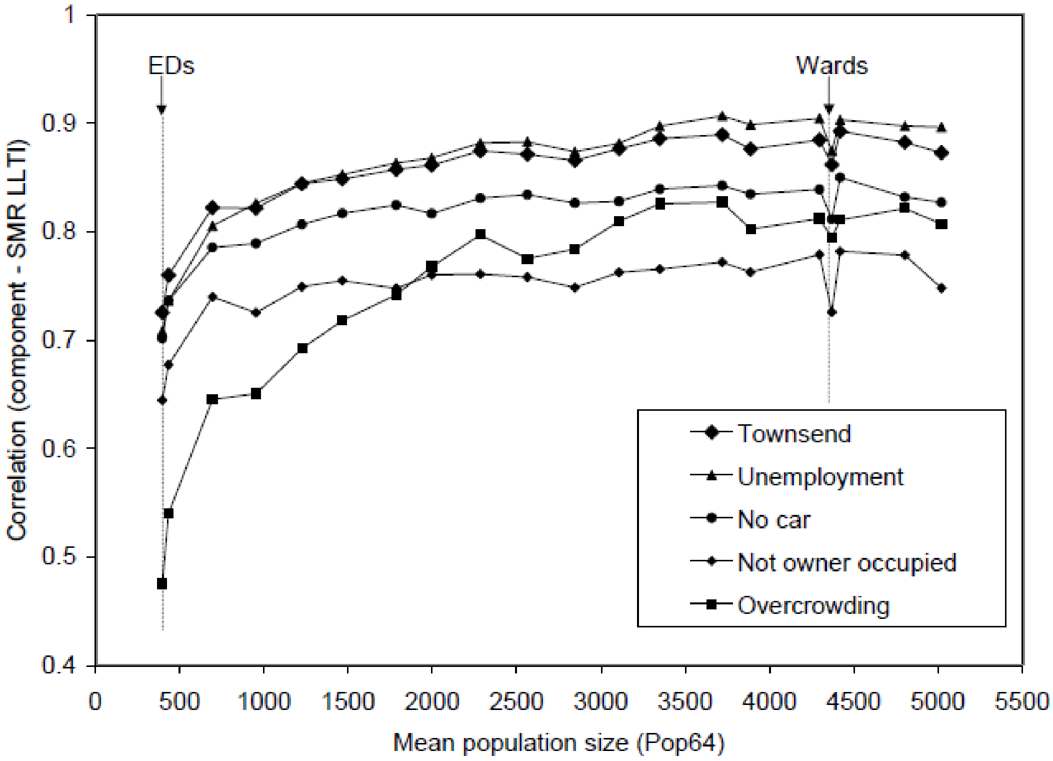


Figure 7 Correlation between Townsend deprivation score and Townsend components and SMR LLTI 0-64 by mean zone (Pop64) size

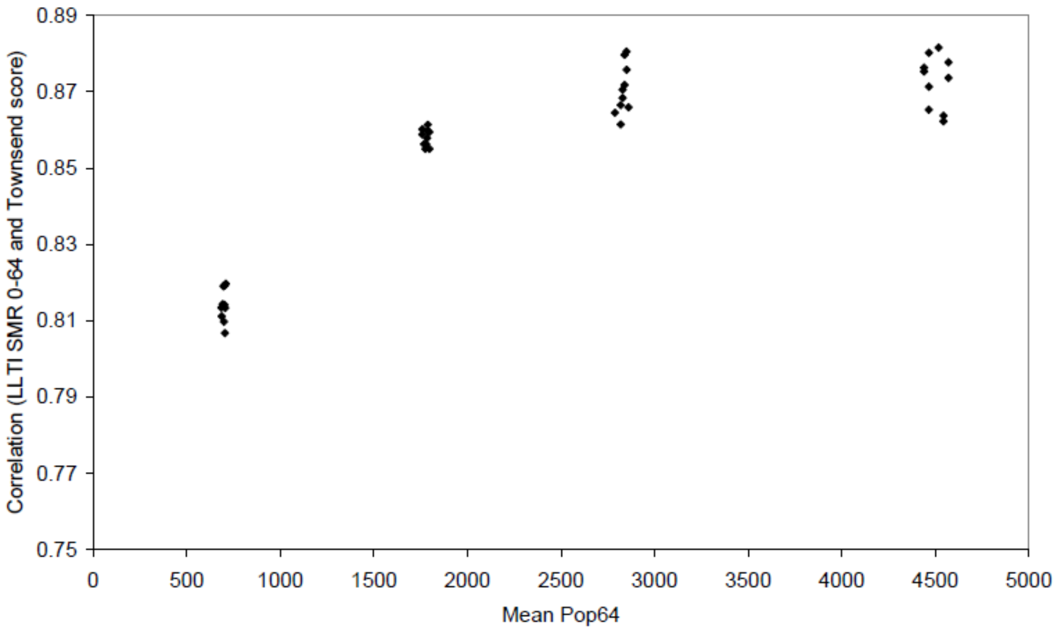


Figure 8 Variation in correlations between Townsend deprivation score and SMR LLTI 0-64 at specific mean zone (Pop64) sizes

5. Discussion

A simple descriptive analysis of the ED and ward population characteristics of Avon has served to illustrate the variability present within these widely used zoning systems that makes them highly problematic as the basis for population-based analyses involving rates and denominator populations. Neither zoning system could be considered optimal for the types of health and environment research of interest here. There is generally less variation between individual aggregations automatically produced than between the administrative zone designs and AZM zoning systems at the equivalent scales.

The pattern of increasing correlation between deprivation and morbidity as zone size increases is consistent with previous findings in the literature (Openshaw, 1984). The suggestion is that this is due to smoothing effects as the level of aggregation increases and this would appear to be supported by our results. The trend is remarkably consistent, and again serves to show that the ED and ward level aggregations are definite outliers in this distribution. Zone design tools in this context have enabled us to systematically explore this scale effect.

We are not suggesting that one specific zoning system is better than any other, but rather that researchers in environment and health should acknowledge the existence of the MAUP and be aware of its potential impact on analyses based on any single zoning system. This is especially relevant where the representation of the phenomena or processes being studied differs (for whatever reason) from the way in which they would ideally be conceptualised. The research design process should include alignment of concept and representation combined with exploration of the sensitivity of results to different zoning systems. Automated zone design techniques offer the ability to undertake such an analysis in a systematic and timely manner.

The empirical analysis presented here is of course based on a set of pre-existing building blocks - the 1991 Census EDs. Openshaw (1984) suggests that the relationships observed in any area-based analysis are likely to be heavily dependent on the first aggregation used to create the building blocks for subsequent zoning systems. Given that the 2001 Census OAs in England and Wales have been designed to be more homogeneous in terms of population size (as well as other socio-economic characteristics), in principle they may be more suitable building blocks for population-based environment and health studies than the 1991 EDs and their higher level aggregations. Despite this, it may be argued that whatever we do with the subsequent aggregations of building blocks, our results will always be dependent on how the initial zone design was undertaken. At present there is insufficient empirical evidence to assess the impact of building block design and choice of zone design criteria. There is also no obvious choice of statistic to measure the effectiveness of the output areas created. Investigation of building block design would require access to complete individual-level data for subsequent aggregation but, in the UK, such data are inaccessible to researchers.

Interpretation of statistical relationships based on any areally-aggregated data requires recognition of the ecological fallacy. Where researchers have access to individual level data, zone design techniques enable systematic exploration of the influence of aggregation on the relationships and patterns observed. In the UK, some authors (Tranmer & Steel, 1998) have used the census microdata samples to explore these effects. Future possibilities lie in the increasing availability of high quality geo-referenced and linked health and population data in countries such as Denmark, Sweden and Finland. If used in conjunction with zone design techniques, these data provide exciting opportunities both to design purpose-specific zoning systems for analyses and to explore the sensitivity of results to different zoning systems.

We conclude from these initial results that observed relationships between variables such as deprivation and morbidity are markedly affected by the choice of zoning system, with correlations strongly associated with the scale of aggregation. Even when the scale element is fixed, considerable additional variation arises from the aggregation aspect of the MAUP. In the context of the 1991 Avon data, the ward aggregation provides a particularly weak zoning system for the analysis of the relationship between deprivation and LLTI.

In more general terms, we suggest that there is a range of potentially important roles for automated zone design tools in environment and health studies, some of which we have discussed. In highlighting the profound impact that the design of areal units can have on observed patterns and relationships, they force us to acknowledge that areal units are neither neutral nor stable. Zone design tools offer the ability to explore the influence of pre-defined and alternative zoning systems and we suggest that in many cases, purpose-specific automatically designed aggregations may be a more appropriate basis for analysis than pre-existing zoning systems.

Acknowledgements

The work reported in this paper was partially supported by a University of Southampton Annual Research Grant (A2000/16). This work is based on data provided with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-LINE Consortium. The authors wish to thank Professor David Briggs for informal discussions which contributed to the development of this paper.

References

Andersen, R. (1995). Revisiting the Behavioral Model and Access to Medical Care: Does It Matter?. Journal of Health and Social Behavior, 36, 1-10.

Barker, D. (1998). *Mothers, babies and health in later life*. Edinburgh: Churchill Livingstone.

Barnett, S., Roderick, P., Martin, D., & Diamond, I. (2001). A multilevel analysis of the effects of rurality and social deprivation on premature limiting long term illness. *Journal of Epidemiology and Community Health*, *55*(1), 44-51.

Barnett, S., Roderick, P., Martin, D., Diamond, I., & Wrigley, H. (2002). Interrelations between three proxies of health care need at the small area level: an urban/rural comparison. *Journal of Epidemiology and Community Health*, *56*(10), 754-761.

Beaglehole, R., Bonita, R., & Kjellström, T. (1993). *Basic Epidemiology*. Geneva: World Health Organisation.

Bentley, C. (2003). *A South Yorkshire framework for NHS action to address health inequalities*. South Yorkshire Strategic Health Authority. Retrieved December 17, 2003, from http://www.southyorkshire.nhs.uk/resources/hiframework160703.pdf.

Boyle, P., Gatrell, A., & Duke-Williams, O. (1999). The effect on morbidity of variability in deprivation and population stability in England and Wales: an investigation at small-area level. *Social Science and Medicine*, *49*(6), 791-799.

Buist, A., & Bernstein, R. (1986). Health effects of volcanoes: an approach to evaluating the health effects of an environmental hazard. *American Journal of Public Health*, *76*(3), 1-2.

Briggs, D. (2001). Exposure assessment. In P. Elliott, J. Wakefield, N. Best & D. Briggs (Eds), *Spatial Epidemiology: Methods and Applications* (pp. 335-359). Oxford: Oxford University Press.

Carstairs, V. (2000). Socio-economic factors at areal level and their relationship with health. In P. Elliott, J. Wakefield, N. Best & D. Briggs (Eds), *Spatial Epidemiology: Methods and Applications* (pp. 51-67). Oxford: Oxford University Press.

Clark, A. M., & Thomas, F. G. (1990). The geography of the 1991 Census. *Population Trends*, *60*, 9-15.

Conduit, E., Brookes, R., Bramley, G., & Fletcher, C. (1996). The value of school locations. *British Educational Research Journal*, *22*(2), 199-206.

Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J., & Grundy, C. (1996). Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer*, *73*(5), 702-710.

Elliott, P., Briggs, D., Morris, S., de Hoogh, C., Hurt, C., Kold Jensen, T., Maitland, I., Richardson, S., Wakefield, J. & Jarup, L. (2001). Risk of adverse birth outcomes in populations living near landfill sites. *British Medical Journal, 323*(7309), 363-368.

Fotheringham, A., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

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

Frohlich, K., Potvin, L., Gauvin, L., & Chabot, P. (2002). Youth smoking initiation: disentangling context from composition. *Health & Place*, *8*(3), 155-166.

Gatrell, A. (2002). *Geographies of Health*. Oxford: Blackwell.

Hart, C., Ecob, R., & Davey Smith, G. (1997). People, places and coronary heart disease risk factors: a multilevel analysis of the Scottish Heart Health study archive. *Social Science and Medicine*, *45*(6), 893-902.

Haynes, R., & Gale, S. (1999). Mortality, long-term illness and deprivation in rural and metropolitan wards of England and Wales. *Health & Place*, *5*(4), 301-312.

Haynes, R., & Gale, S. (2000). Deprivation and poor health in rural areas: inequalities hidden by averages. *Health & Place*, *6*(4), 275-285.

Horn, M., 1995. Solution techniques for large regional partitioning problems. *Geographical Analysis, 27*(3), 230-248.

Keegan, T., Whitaker, H., Nieuwenhuijsen, M., Toledano, M., Elliott, P., Fawell, J., Wilkinson, M., & Best, N. (2001). Use of routinely collected data on trihalomethane in drinking water for epidemiological purposes. *Occupational & Environmental Medicine*, *58*(7), 447-452.

Kleinschmidt, I., Hills, M., & Elliott, P. (1995). Smoking behaviour can be predicted by neighbourhood deprivation measures. *Journal of Epidemiology & Community Health*, *49*(suppl 2), S72-7.

Kuh, D., & Ben-Shlomo, Y. (eds) (1997). *A Life Course Approach to Chronic Disease Epidemiology,* Oxford: Oxford University Press.

Macintyre, S. (1997). What are spatial effects and how can we measure them? In A. Dale (Ed.), *Exploiting national survey data: the role of locality and spatial effects* (pp.1-17). Manchester: Faculty of Economic and Social Studies.

Macintyre, S., Ellaway, A., & Cummins, S. (2002). Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science & Medicine, 55*(1), 125-139.

Martin, D. (1998). Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science*, *12*(7), 673-685.

Martin, D. (2003). Developing the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, *17*(2), 181-196.

Martin, D., Brigham, P., Roderick, P., Barnett, S., & Diamond, I. (2000). The (mis)representation of rural deprivation. *Environment and Planning A*, *32*(4), 735-751.

Martin, D., Nolan A., & Tranmer, M. (2001). The application of zone design methodology to the 2001 UK Census. *Environment and Planning A*, *33*(11), 1949-1962.

McClelland, R., & Kronmal, R. (2002). Regression-based variable clustering for data reduction. *Statistics in Medicine*, *21*(6), 921-941.

Middleton, N., Gunnell, D., Frankel, S, Whitley, E., & Dorling, D. (2003). Urban-rural differences in suicide trends in young adults: England and Wales, 1981-998. *Social Science & Medicine*, *57*(7), 1183-1194.

Mitchell, R., Gleave, S., Bartley, M., Wiggins, D., & Joshi, H. (2000). Do attitude and area influence health? A multilevel approach to health inequalities. *Health & Place*, *6*(2), 67-79.

Morris, R., & Munasinghe, R. (1993). Aggregation of existing geographic regions to diminish spurious variability of disease rates. *Statistics in Medicine*, *12*, 1915-1929.

Office for National Statistics (2003). Super Output Areas: a core geography for neighbourhood statistics: Consultation document. [Document]. URL http://www.neighbourhood.statistics.gov.uk/downloads/Proposal_SOA_200302.doc.

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, New Series*, *2*(4), 459-472.

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

Openshaw, S. and Alvanides, S. (1999). Applying geocomputation to the analysis of spatial distributions. In P. Longley, M. Goodchild, D. Maguire, & D. Rhind (Eds.), *Geographical Information Systems: Principles, Techniques, Applications and Management*, (pp. 267-282). Chichester: Wiley.

Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 Census geography. *Environment and Planning A*, *27*(3), 425-446.

Pickett, K., & Pearl, M. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology & Community Health, 55*(2), 111-122.

Pickett, K., Wakschlag, L., Rathouz, P., Leventhal, B, & Abrams, B. (2002). The working-class context of pregnancy smoking. *Health & Place, 8*(3), 167-175.

Rothman, K. (1990). A sobering thought for the cluster busters' conference. *American Journal of Epidemiology, 132* (Suppl.1), S6-13.

Shaw, M., Dorling, D., & Mitchell, R. (2002). *Health, Place and Society*. Harlow: Pearson Education Limited.

Shouls, S., Congdon, P., & Curtis, S. (1996). Modelling inequality in reported longterm illness in the UK: combining individual and area characteristics. *Journal of Epidemiology and Community Health, 50*(3), 366-376.

Townsend, P., Phillimore, P., & Beattie, A. (1988). *Health and deprivation: inequality and the North*. London: Croom Helm.

Tranmer, M., & Steel, D. (1998). Using Census data to investigate the causes of the ecological fallacy. *Environment and Planning A, 30*(5), 817-831.

Wilkinson, P., Elliott, P., Grundy, C., Shaddick, G., Thakrar, B., Walls, P. & Falconer, S, (1999). Case-control study of hospital admission with asthma in children aged 5-14 years: relation with road traffic in north west London. *Thorax*, 54(12), 1070-1074.

Wilkinson, R. (1992). Income distribution and life expectancy: a critical appraisal. *British Medical Journal, 311*, 1282-1285.

2.2 Paper 2: Towards 2011 output geographies: Exploring the need for, and challenges involved in, maintenance of the 2001 output geographies

Cockings S, Harfoot A and Hornby D “Towards 2011 output geographies: Exploring the need for, and challenges involved in, maintenance of the 2001 output geographies”

This is a post-peer-review, pre-copyedit version of an article published in *Population Trends*. The definitive publisher-authenticated version [Cockings S, Harfoot A and 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] is available online at: http://dx.doi.org/10.1057/pt.2009.46

Towards 2011 output geographies: Exploring the need for, and challenges involved in, maintenance of the 2001 output geographies

Samantha Cockings, Andrew Harfoot, Duncan Hornby,

University of Southampton

Abstract. This article describes and presents early results from the ESRC-funded Census 2011Geog project, which aims to develop and evaluate automated procedures to maintain (split, merge or re-design) the 2001 Census output geographies in order to create the 2011 output geographies for England and Wales. The article explores population change at the small area level between 2001 and 2005-06, and considers the extent to which the 2001 Census output geographies are likely to be appropriate for the release of 2011 Census data. It concludes that the vast majority of output geography areas are unlikely to have breached population thresholds by 2011, but that a small proportion of areas will require maintenance. The article finishes with a discussion of the key decisions that need to be made before the automated procedures can be implemented operationally.

Keywords: Output geographies, maintenance, automated zone design, population change

Introduction

The 2001 Census output geographies for England and Wales were designed to be an optimised representation of the population distribution and socio-economic characteristics at that time. By the next Census in 2011 there will have been changes in the size and composition of population in most areas. While recognising this, the National Statistics Small Area Geography Consultation in 2007[1] revealed strong user demand for output geography stability. The challenges involved in creating 2011 output geographies that maintain both a high degree of stability and also reflect real-world population changes are non-trivial. This article introduces the ESRC-funded Census 2011Geog project, which aims to develop automated methods for maintaining (splitting, merging or re-designing) the 2001 output geographies in order to create output geographies for 2011. The article presents preliminary results from the first stage of this project, exploring small-area population change in England and Wales between 2001 and 2005-06, and considering the extent to which the 2001 output geographies are likely to be appropriate for use in 2011. It also reviews the key decisions that must be made before the maintenance procedures can be implemented. Note that this article is not specifically concerned with whether or not to re-align the boundaries of the output geographies to real-world features, although this is a relevant and related issue.

Background: the 2001 output geographies

The smallest zones for which Census data were released in 2001 in England and Wales were output areas (OAs). These OAs were created using a process of automated zone design following the collection and processing of the household-level 2001 Census data.[2] [3] Automated zone design involves two key methodological stages.[4] First, a set of small building blocks is created. Second, these building blocks are iteratively aggregated into larger zones, with the aim of optimising an objective function based on pre-specified design criteria. The building blocks employed for the 2001 Census OAs were postcode polygons. Geographic Information Systems (GIS) were used to create small, space-filling, polygons around the addresses of households enumerated by the Census. Adjacent address polygons belonging to the same postcode were then merged to create postcode polygons. The boundaries of wards and parishes were then intersected to create a set of 'ward-parts' and the postcode polygon boundaries were constrained to nest within these, as well as being made to coincide with road centre lines where possible.

These synthetic postcode polygons were then aggregated to create OAs. All OAs had to exceed specified minimum population (100) and household (40) thresholds in order to protect individuals from inadvertent disclosure in the aggregate data. Note that no maximum thresholds were specified for the OA creation process. The OAs within each ward-part were then iteratively re-combined, using multiple random restarts, in order to identify the set of OAs which best optimised a set of design criteria. The criteria were: homogeneity of population size across OAs (aiming for a target mean of 125 households); internal homogeneity of accommodation type and tenure within OAs; and compactness of shape. The OAs from all ward-parts were then merged to form a national set of OAs. These OAs subsequently became the building blocks for sets of larger 'neighbourhood' geographies, namely the Lower Layer and Middle Layer Super Output Areas (LSOAs and MSOAs respectively). These LSOAs and MSOAs are now well established geographies for the release of neighbourhood statistics (www.neighbourhood.statistics.gov.uk). Similar zone design criteria to those used to generate the OAs were employed in the creation of the LSOAs and MSOAs, including minimum population thresholds of 1,000 and 5,000 respectively. Note that the MSOAs were the only output geography layer to have a published upper threshold (4,000 households). Importantly, the boundaries of all of these output geographies were made freely available for non-commercial use.

Population change since 2001

While the output geographies were optimised for certain population and socio-economic characteristics in 2001, changes in the population size and distribution since then are likely to mean that in some areas the 2001 output geographies will no longer be appropriate for representing the population or for maintaining confidentiality. The key drivers of population change since 2001 have included migration, an ageing population, people marrying later and higher divorce and separation rates. These factors have led to a reduction in mean household size and a consequent rise in the number of residential properties required, together with a greater demand for smaller properties. Residential development has primarily comprised the building of new properties (mainly on either green-field or brown-field sites) and the sub-division of existing properties. A minority of areas since 2001 have experienced population decline; where this has happened, it has mainly been due to internal outward migration. All of the above changes will not only have led to changes in the population size and distribution within the output geographies since 2001 but also, potentially, to changes in the homogeneity of the socio-economic characteristics of the areas.

In planning for the 2011 Census it is important to estimate how much change there has been since 2001, and to what extent the output geographies will have breached population thresholds by 2011. It is also important to understand the nature of these changes, especially in terms of the types of breaches and their geographical distribution.

This will enable the development and evaluation of methodologies that can take the 2001 output geographies and modify them, where appropriate, in order to create the 2011 output geographies, preferably using automated procedures.

The Census 2011Geog project

The Economic and Social Research Council (ESRC), via its Census Programme, is funding a collaborative research project (the 'Census 2011Geog' project, http://census2011geog.census.ac.uk) between the University of Southampton and the Office for National Statistics (ONS). The aim of the project is to create automated procedures for maintaining (i.e. splitting, merging or re-designing) the 2001 output geographies in order to create the 2011 output geographies. It also aims to investigate the implications of using different building blocks, such as postcodes and street blocks, for these maintenance procedures. The project, which builds on previous experiments by ONS [5], will deliver prototype software that can be tested by ONS following the Census Rehearsal in 2009. This can then adapted for operational use in the 2011 Census. It will also deliver an evidence base of the implications of using different building blocks and design criteria for the maintenance procedures. The first stage of the project has involved an exploration of the likely magnitude and geographical distribution of population change and consequent breaches in the output geographies. The results of this analysis form the basis of the findings presented here.

Use of ONS mid-year population estimates

Available ONS mid-year population estimates for 2001 to 2006 at local authority level and for 2001 to 2005 at OA, LSOA and MSOA levels were employed to investigate population change since 2001. Eighteen OAs were excluded from the OA-level analysis as they contained no postcodes and therefore did not receive any population via the postcode-best fit method employed to create the mid-year estimates at OA level. Note also that the MSOA counts include the Isles of Scilly pseudo-MSOA. The mid-year estimates of usual resident population include adjustments for births, deaths and migration.[6] In addition, the 2001 mid-year estimates included specific adjustments/corrections for under-enumeration at Census. This under-enumeration arose for many reasons[7], but particularly as a result of problems experienced with the address register in certain areas.[8] In assessing population change since the 2001 Census, it is therefore more appropriate to compare the 2005-06 mid-year estimates with the 2001 mid-year estimates, rather than with the actual 2001 Census counts, as this gives a more reliable estimate of population change. A further consideration when assessing population change over time is that some areas have high proportions of special populations that can be highly mobile, such as members of the armed forces. Changes in the geographical distribution of such populations can result in an apparent increase or decrease in an area's usually resident population since Census, even if the underlying non-special population is actually reasonably stable. This should be borne in mind when interpreting the results of the analyses presented here.

Population change at Local Authority level, 2001-2006

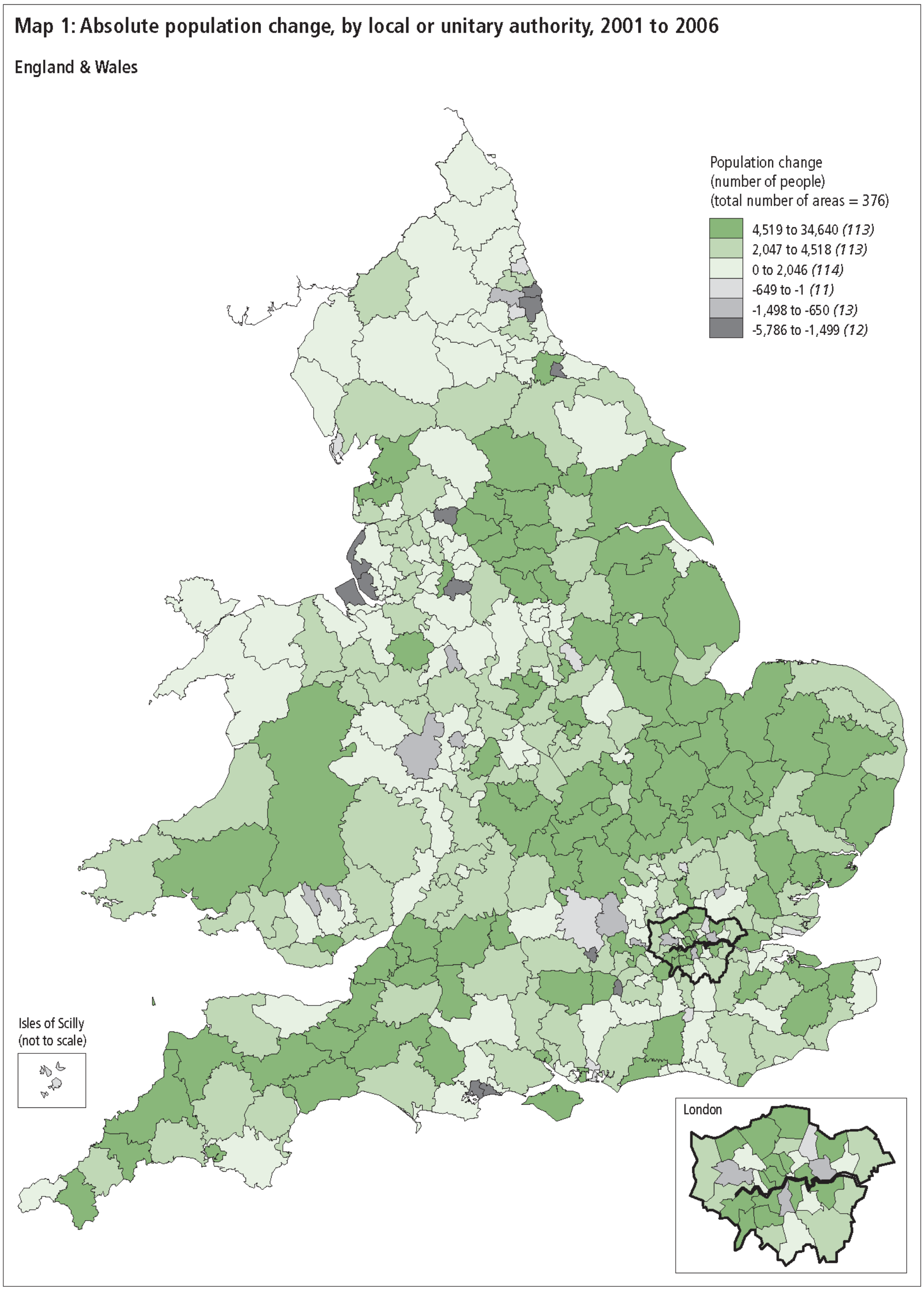
Map 1 shows absolute population change between the 2001 and 2006 mid-year estimates for local authorities in England and Wales. The average population change across all local authorities in England between 2001 and 2006 was a 2.9 per cent increase, representing an average growth in population of 3,710 per local authority. By contrast, the levels of population change over the same time period in Welsh unitary authorities were lower, with an overall average increase of 1.8 per cent (2,523 people).

These averages hide considerable geographical variation between local authorities of different area types. In order to explore this further, local authorities were classified by their Department of Environment, Food and Rural Affairs (DEFRA) urban/rural category[9] and then ranked within these categories according to their percentage increase or decrease in population between 2001 and 2006 (again using the mid-year estimates). **Table 1** and **Table 2** show the five English local authorities exhibiting the greatest population increases and decreases respectively, in each DEFRA urban/rural category. Note that for the most strongly rural categories ('significant rural', 'rural-50' and 'rural-80'), there were less than five local authorities experiencing a decrease in their population and so only those experiencing a decrease are shown. No similar urban/rural classification scheme was available for Welsh unitary authorities; instead population change for all Welsh unitary authorities is shown in **Table 3**.

It is clear from this analysis that a greater proportion of local authorities experienced population growth (90.4 per cent of English local authorities and 91.7 per cent of Welsh UAs) than population decline (9.6 per cent and 8.3 per cent for England and Wales respectively), and that the relative magnitude of the growth was greater than that of the decline. The levels of growth were reasonably similar across the various urban/rural categories: whilst the greatest increases were seen in the major urban areas (such as Westminster, Camden and Oxford), there was also significant growth in rural areas (for example, South Northamptonshire, Forest Heath and Rutland). By contrast, the levels of decline were less consistent: the greatest declines were seen in urban areas (such as Sefton, Middlesborough and Rushmoor), whereas very few rural areas (other than Bridgnorth) declined substantially.

Table 1. Local Authorities in England exhibiting highest population increases between mid-year estimates in 2001 and 2006, in each DEFRA urban/rural category

DEFRA Classification	Local Authority	Population		Change	
		Mid-2001	Mid-2006	Persons	Percent
1. Major Urban	Westminster	203,329	231,874	28,545	14.0
	Camden	202,567	227,453	24,886	12.3
	Kensington and Chelsea	162,199	178,021	15,822	9.8
	Manchester	422,915	451,984	29,069	6.9
	Tower Hamlets	201,090	212,804	11,714	5.8
2. Large Urban	Nottingham	268,939	286,378	17,439	6.5
	Bristol City of	390,049	410,487	20,438	5.2
	Portsmouth	188,043	196,379	8,336	4.4
	Wyre	105,800	110,371	4,571	4.3
	Southampton	219,539	228,635	9,096	4.1
3. Other Urban	Oxford	135,509	149,105	13,596	10.0
	Welwyn Hatfield	97,550	105,514	7,964	8.2
	Canterbury	135,381	146,181	10,800	8.0
	Exeter	111,180	119,606	8,426	7.6
	Cambridge	109,941	117,913	7,972	7.3
4. Significant Rural	South Derbyshire	81,738	89,779	8,041	9.8
	Colchester	156,016	170,846	14,830	9.5
	Ashford	103,024	111,177	8,153	7.9
	Kettering	82,304	87,858	5,554	6.7
	Lancaster	134,049	143,033	8,984	6.7
5. Rural-50	East Northamptonshire	76,835	83,954	7,119	9.3
	North Somerset	188,840	201,404	12,564	6.7
	Kerrier	92,634	98,008	5,374	5.8
	Tonbridge and Malling	107,771	113,937	6,166	5.7
	Braintree	132,482	139,688	7,206	5.4
6. Rural-80	South Northamptonshire	79,497	88,764	9,267	11.7
	Forest Heath	56,145	62,129	5,984	10.7
	Rutland	34,598	38,277	3,679	10.6
	North Kesteven	94,378	103,152	8,774	9.3
	Mid Bedfordshire	121,258	132,185	10,927	9.0


Map 1: Absolute population change, by local or unitary authority, 2001 to 2006

England & Wales

Source: Mid-year population estimates, Office for National Statistics

Map 1 Absolute population change, by local or unitary authority, 2001-2006, England and Wales

Table 2 Local Authorities in England exhibiting highest population decreases between mid-year estimates in 2001 and 2006, in each DEFRA urban/rural category

DEFRA Classification	Local Authority	Population		Change	
		Mid-2001	Mid-2006	Persons	Percent
1. Major Urban	Sefton	282,884	277,421	-5,463	-1.9
	Sunderland	284,601	280,593	-4,008	-1.4
	Stockport	284,557	280,619	-3,938	-1.4
	Liverpool	441,858	436,072	-5,786	-1.3
	South Tyneside	152,793	151,020	-1,773	-1.2
2. Large Urban	Middlesbrough	141,233	138,434	-2,799	-2.0
	Bournemouth	163,560	161,169	-2,391	-1.5
	Reading	144,684	142,756	-1,928	-1.3
	Wirral	315,004	311,210	-3,794	-1.2
	Poole	138,368	136,869	-1,499	-1.1
3. Other Urban	Rushmoor	90,892	88,744	-2,148	-2.4
	Burnley	89,521	87,979	-1,542	-1.7
	Harlow	78,799	78,065	-734	-0.9
	Slough	120,577	119,516	-1,061	-0.9
	Stevenage	79,794	79,307	-487	-0.6
4. Significant Rural	Wycombe	162,050	161,326	-724	-0.4
5. Rural-50	Blyth Valley	81,334	81,204	-130	-0.2
6. Rural-80	Bridgnorth	52,458	51,808	-650	-1.2
	Isles of Scilly	2,140	2,126	-14	-0.7
	South Oxfordshire	128,307	128,124	-183	-0.1

While the above analysis provides useful information about the general trends in population change since 2001, in planning for the 2011 Census it is more important to explore the extent to which the output geographies themselves have breached specified thresholds, as it is stability of the output geographies (particularly at the OA and LSOA levels) which forms the basis of ONS's small area geography policy and users' preferred requirements (ONS, 2007). The mid-year estimates for 2001 to 2005 were therefore employed to investigate this in more detail.

Table 3 Population change between 2001 and 2006 mid-year estimates for Welsh Unitary Authorities

Unitary Authority	Population		Change	
	Mid-2001	Mid-2006	Persons	Percent
Powys	126,398	131,141	4,743.0	3.8
Monmouthshire	84,984	87,882	2,898.0	3.4
The Vale of Glamorgan	119,277	123,275	3,998.0	3.4
Denbighshire	93,070	96,089	3,019.0	3.2
Carmarthenshire	172,845	178,043	5,198.0	3.0
Bridgend	128,735	132,584	3,849.0	3.0
Ceredigion	75,083	77,160	2,077.0	2.8
Pembrokeshire	114,199	117,280	3,081.0	2.7
Cardiff	310,088	317,523	7,435.0	2.4
Neath Port Talbot	134,380	137,052	2,672.0	2.0
Wrexham	128,540	130,990	2,450.0	1.9
Newport	137,642	140,125	2,483.0	1.8
Swansea	223,463	227,079	3,616.0	1.6
Isle of Anglesey	67,806	68,884	1,078.0	1.6
Conwy	109,674	111,273	1,599.0	1.5
Gwynedd	116,844	118,250	1,406.0	1.2
Caerphilly	169,546	171,349	1,803.0	1.1
Flintshire	148,629	150,077	1,448.0	1.0
Rhondda, Cynon, Taff	231,910	233,936	2,026.0	0.9
Torfaen	90,912	91,022	110.0	0.1
Blaenau Gwent	70,000	69,341	-659.0	-0.9
Merthyr Tydfil	56,207	55,530	-677.0	-1.2

Threshold breaches in the output geographies

The number of output geography areas breaching upper and lower thresholds by 2011 will be contingent upon the specific thresholds employed. ONS has not yet confirmed the thresholds to be employed in 2011, but it is likely that the levels will be similar to those used in 2001. Assuming this is the case, similar thresholds can be employed to explore the extent to which OAs, LSOAs and MSOAs had breached lower and upper thresholds by 2005. **Table 4** shows the population thresholds employed here, including our working definition of ‘upper thresholds’ which were not formally defined for Census purposes. It was not possible to explore household threshold breaches as inter-censal ONS mid-year estimates are not produced for households. Population thresholds were calculated by multiplying household thresholds by a factor of 2.5 (designed to approximate to mean household size).

Table 4 Population thresholds employed in the analysis

Geography	Lower thresholds		Upper thresholds	
	Households	Population	Households	Population
OA	40	100	250	625
LSOA	400	1000	1200	3000
MSOA	2000	5000	6000	15000

Notes:
1. Population thresholds obtained by multiplying household thresholds by a factor of 2.5 (equating approximately to average household size).
2. Household threshold values taken from Mitchell and Ralphs (2007), Table 1.1.
3. No upper thresholds were published in 2001 for OAs or LSOAs. The values used here are the target mean employed for the automated zone design process multiplied by 2 (as in Ralphs and Mitchell, (2006)). MSOAs did have a published upper threshold of 4000 households, but here we use the value of 6000 households (as in Mitchell and Ralphs, (2007)) as this is consistent with the values used at the other levels.

Number of breaches

Table 5 shows the numbers and percentages of OAs, LSOAs and MSOAs in England and Wales breaching and/or within threshold, in either or both of 2001 and 2005. The columns show the number (or percentage) of areas below, within or above threshold in 2005, whereas the rows show the number (or percentage) below, within or above threshold in 2001. The cells in the matrix are therefore cross-tabulations of these various combinations. The majority of OAs, LSOAs and MSOAs (98.9 per cent, 99.6 per cent and 99.8 per cent respectively) were within threshold in both 2001 and 2005. Even in areas with high population change, there were remarkably few OAs, LSOAs and MSOAs breaching the 2001 thresholds by 2005. Very few areas (<0.1 per cent of OAs; 0.1 per cent LSOAs; 0.1 per cent MSOAs) that were within threshold in 2001 had breached the lower threshold by 2005. Similarly low percentages (0.4 per cent of OAs; 0.2 per cent of LSOAs; zero MSOAs) were within threshold in 2001 but above threshold by 2005. 221 OAs (0.13 per cent) that were below threshold in 2005 were also below threshold in 2001. The only MSOA (in Cambridge) which was above threshold in 2005 was also above threshold in 2001. Only 1 OA had shifted from being below threshold in 2001 to above threshold in 2005.

Map 2 and **Map 3** provide examples of how the magnitude and distribution of these breaches can vary geographically in specific areas. Map 2 and Map 3 show absolute population change between the 2001 and 2005 mid-year estimates for Camden and Liverpool respectively. OAs that have breached lower or upper thresholds have a semi-circle or circle symbol within their boundary; those breaching a threshold in 2001 have a left-handed semi circle; those breaching in 2005 have a right-handed semi-circle; breaches below the lower threshold are shown in purple; breaches above the upper threshold are in green. It is clear from these figures that, overall, Camden has experienced more population growth whereas Liverpool has experienced more decline. While the majority of OAs within both local authorities have not yet breached thresholds, both areas have a number of OAs that have breached either the lower or upper thresholds in 2001 and/or 2005. This illustrates the complexity of designing automated procedures that will maintain stability in the majority of OAs but split, merge or re-design only those that require change.

Table 5 Number of OAs, LSOAs and MSOAs breaching population thresholds in mid-year estimates for 2001 and 2005 in England and Wales

Output Areas (OA)

(a) Counts

	2005 Below	2005 Within	2005 Above	2001 Totals
2001 Below	221	228	1	450
2001 Within	147	173,553	682	174,382
2001 Above	0	78	506	584
2005 Totals	368	173,859	1,189	175,416

(b) Per cent (of total)

	2005 Below	2005 Within	2005 Above	2001 Totals
2001 Below	0.13	0.13	0.00	0.26
2001 Within	0.08	98.94	0.39	99.41
2001 Above	0.00	0.04	0.29	0.33
2005 Totals	0.21	99.11	0.68	100.00

Lower Level Super Output Areas (LSOA)

(c) Counts

	2005 Below	2005 Within	2005 Above	2001 Totals
2001 Below	6	8	0	14
2001 Within	34	34,242	58	34,334
2001 Above	0	3	27	30
2005 Totals	40	34,253	85	34,378

(d) Per cent (of total)

	2005 Below	2005 Within	2005 Above	2001 Totals
2001 Below	0.02	0.02	0.00	0.04
2001 Within	0.10	99.60	0.17	99.87
2001 Above	0.00	0.01	0.08	0.09
2005 Totals	0.12	99.63	0.25	100.00

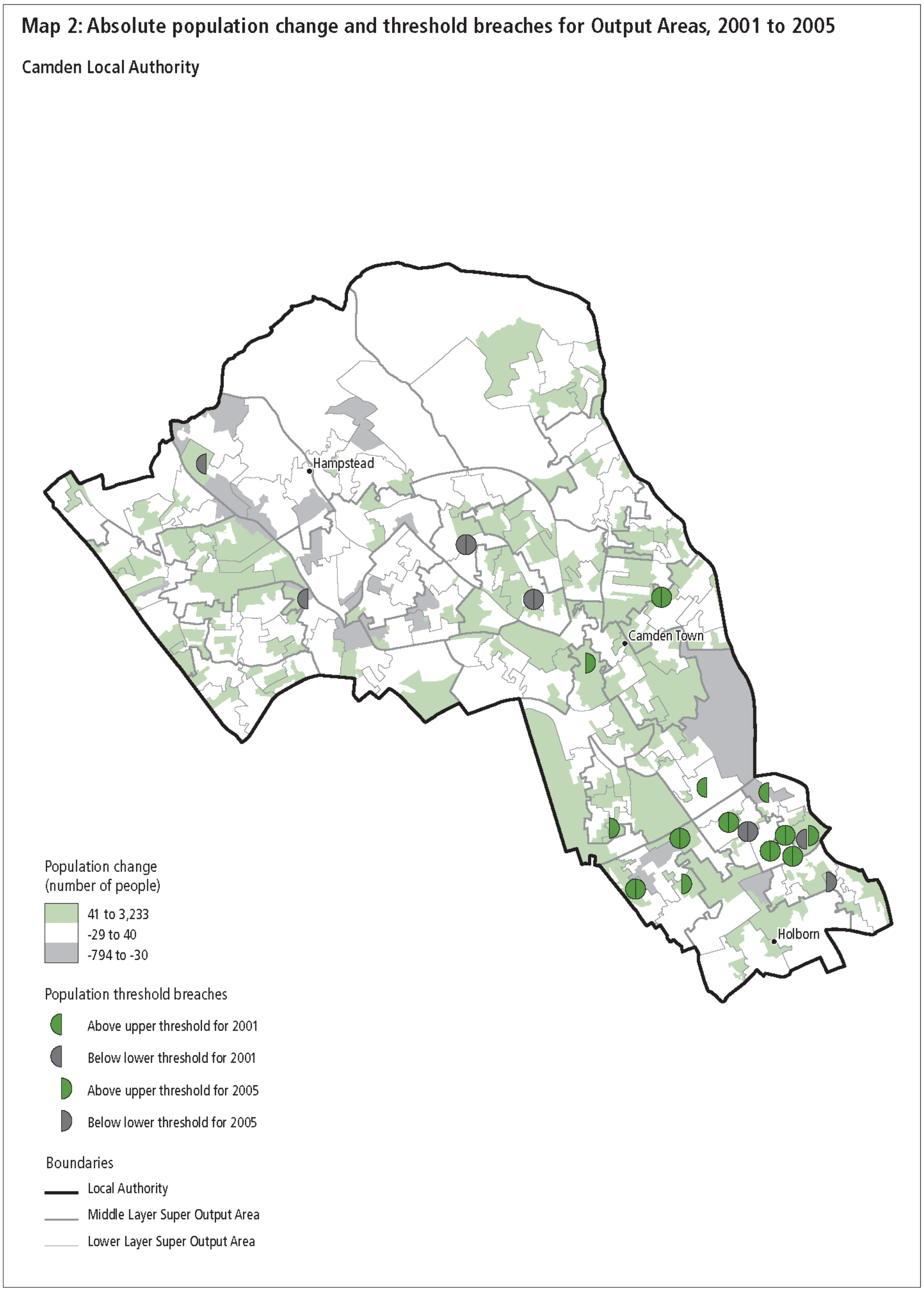
Middle Level Super Output Areas (MSOA)

(e) Counts

	2005 Below	2005 Within	2005 Above	2001 Totals
2001 Below	3	4	0	7
2001 Within	8	7,178	0	7,186
2001 Above	0	0	1	1
2005 Totals	11	7,182	1	7,194

(f) Per cent (of total)

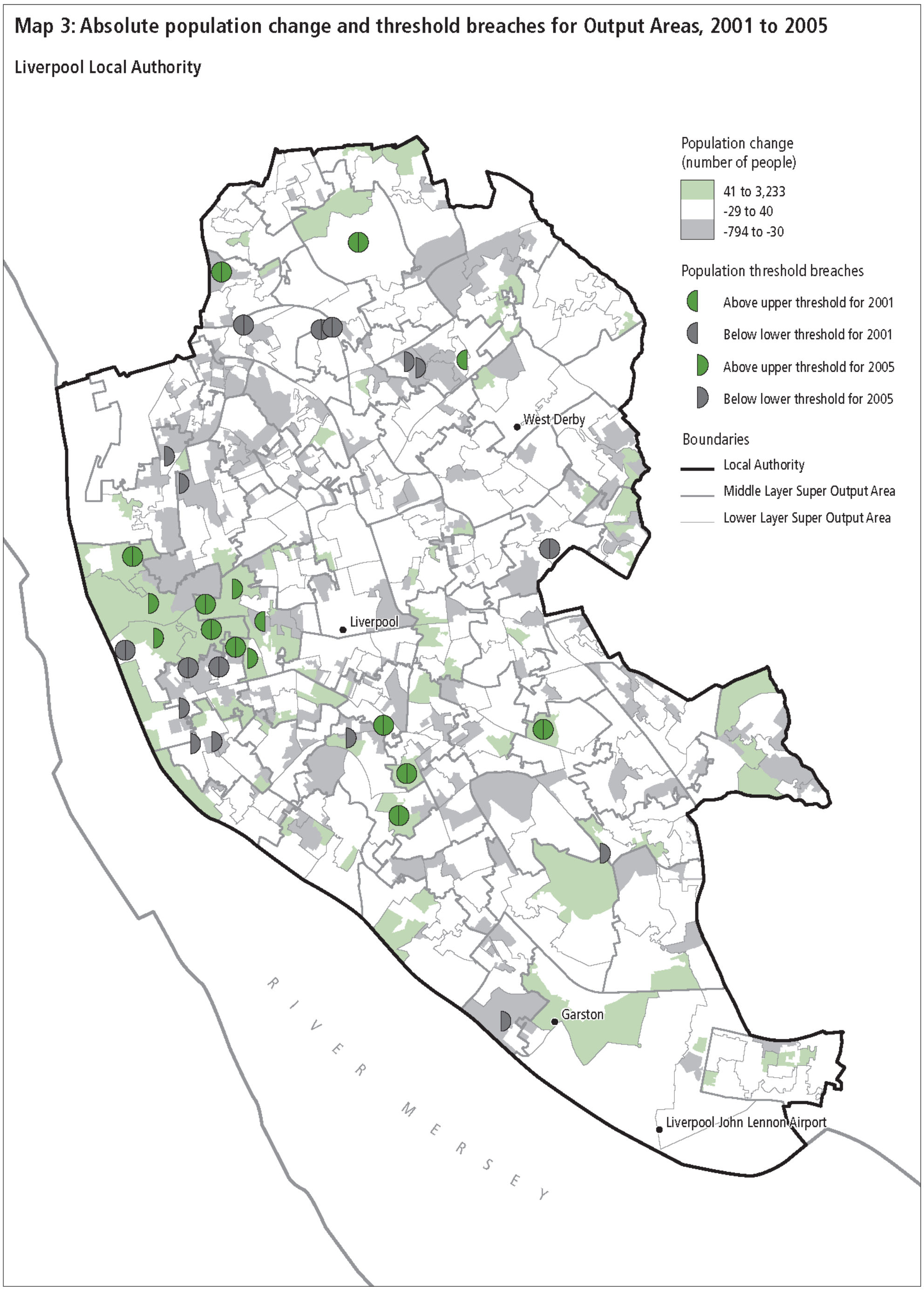
	2005 Below	2005 Within	2005 Above	2001 Totals
2001 Below	0.04	0.06	0.00	0.10
2001 Within	0.11	99.78	0.00	99.89
2001 Above	0.00	0.00	0.01	0.01
2005 Totals	0.15	99.84	0.01	100.00


Map 2: Absolute population change and threshold breaches for Output Areas, 2001 to 2005
Camden Local Authority
Hampstead
Camden Town
Holborn
Population change (number of people)
41 to 3,233
-29 to 40
-794 to -30
Population threshold breaches
Above upper threshold for 2001
Below lower threshold for 2001
Above upper threshold for 2005
Below lower threshold for 2005
Boundaries
Local Authority
Middle Layer Super Output Area
Lower Layer Super Output Area

Source: Mid-year population estimates, Office for National Statistics

Contains National Statistics data © Crown copyright and database right 2013
Contains Ordnance Survey data © Crown copyright and database right 2013

Map 2 Absolute population change and threshold breaches for Output Areas, 2001-2005, Camden Local Authority


Map 3: Absolute population change and threshold breaches for Output Areas, 2001 to 2005
Liverpool Local Authority
Population change (number of people)
41 to 3,233
-29 to 40
-794 to -30
Population threshold breaches
Above upper threshold for 2001
Below lower threshold for 2001
Above upper threshold for 2005
Below lower threshold for 2005
Boundaries
Local Authority
Middle Layer Super Output Area
Lower Layer Super Output Area
West Derby
Liverpool
Garston
Liverpool John Lennon Airport
RIVER MERSEY

Source: Mid-year population estimates, Office for National Statistics

Map 3 Absolute population change and threshold breaches for Output Areas, 2001-2005, Liverpool Local Authority

Change in number of breaches over time

Table 6 shows the numbers and percentages of OAs, LSOAs and MSOAs below, within and above threshold annually from 2001 to 2005 for England and Wales. **Table 7** explores the degree to which the output geography areas were approaching (within 5 or 10 per cent of) the upper and lower thresholds by 2005 and the extent to which areas that had already breached by 2005 had exceeded the thresholds.

Table 6 shows that at the OA level between 2001 and 2005 there was a steady annual increase in the numbers above threshold whereas the number of OAs below threshold remained fairly stable over time. Table 7 reveals that the percentage of OAs less than 5 per cent above the lower threshold was only 0.1 per cent, and that this percentage had decreased very slightly year-on-year. Approximately 0.1 per cent of OAs were also within 5 per cent of the upper threshold but, by contrast, this percentage had grown slightly year-on-year. Overall, the number of OAs that had breached, or were near to breaching, the upper thresholds appeared to be steadily increasing annually, presumably reflecting general growth in population. Interestingly, where OAs had breached, they tended to have breached by at least 10 per cent over the threshold, rather than having only just gone over threshold. The same is true of those breaching the lower threshold.

At the LSOA level, the trends of a small annual increase in the percentage of LSOAs breaching the lower threshold and a slightly larger increase in those breaching the upper threshold were similar to the trends observed amongst OAs. However, the degree to which non-breached LSOAs were approaching the thresholds, and the extent to which those already breaching had exceeded the thresholds, was different to the OAs. There was a smaller percentage of LSOAs near to the thresholds, with more being closer to the lower threshold than the upper threshold. Of those LSOAs that had already breached the lower threshold, most had done so by less than 5 per cent. Amongst those breaching the upper threshold, there appeared to be two distinct groups: those that had only just breached (by 5 per cent or less) and those that had breached more substantially (by greater than 10 per cent) – with the second group being more numerous.

There was an inconsistent, but very slight, increase in the number of MSOAs breaching the lower threshold between 2001 and 2005. Of those breaching the lower threshold, most had only just dipped beneath it. Approximately 2 per cent of MSOAs were less than 5 per cent above the lower threshold in 2001: by 2005 this percentage was only 1.3 per cent, reflecting the general growth in MSOA populations. Encouragingly, despite this growth, only a tiny percentage (approximately 0.04 per cent) were within 5 per cent of the upper threshold by 2005. As noted previously, only one MSOA had breached the upper threshold by 2005, but this MSOA was already above threshold in 2001. It is worth noting though that it had moved further above threshold between 2001 and 2005.

It is likely that the differences in the patterns observed at OA-level compared to those at LSOA and MSOA levels are due to the scale and size of the geographical units, and due to differences in the 2001 mid-year population distributions (in terms of how close the 2001 mean OA, LSOA and MSOA populations were to the thresholds initially). Overall, these findings should be reassuring for ONS and users who are hoping that it is possible to retain stability in 2001, especially at the higher output geography levels.

Table 6 Threshold breaches for OAs, LSOAs and MSOAs over time, 2001-2005 mid-year estimates, England and Wales

Output Areas (OA)

(a) Counts

Year	Zero people	Below threshold	Within threshold	Above threshold
2001	6	444	174,382	584
2002	8	307	174,465	636
2003	8	323	174,284	801
2004	4	339	174,055	1,018
2005	3	365	173,859	1,189

(b) Per cent (of total)

Year	Zero people	Below threshold	Within threshold	Above threshold
2001	<0.01	0.25	99.41	0.33
2002	<0.01	0.18	99.46	0.36
2003	<0.01	0.18	99.36	0.46
2004	<0.01	0.19	99.22	0.58
2005	<0.01	0.21	99.11	0.68

Lower Level Super Output Areas (LSOA)

(c) Counts

Year	Zero people	Below threshold	Within threshold	Above threshold
2001	0	14	34,334	30
2002	0	27	34,319	32
2003	0	25	34,307	46
2004	0	34	34,282	62
2005	0	40	34,253	85

(d) Per cent (of total)

Year	Zero people	Below threshold	Within threshold	Above threshold
2001	0.00	0.04	99.87	0.09
2002	0.00	0.08	99.83	0.09
2003	0.00	0.07	99.79	0.13
2004	0.00	0.10	99.72	0.18
2005	0.00	0.12	99.64	0.25

Middle Level Super Output Areas (MSOA)

(e) Counts

Year	Zero people	Below threshold	Within threshold	Above threshold
2001	0	7	7,186	1
2002	0	8	7,185	1
2003	0	10	7,183	1
2004	0	12	7,181	1
2005	0	11	7,182	1

(f) Per cent (of total)

Year	Zero people	Below threshold	Within threshold	Above threshold
2001	0.00	0.10	99.89	0.01
2002	0.00	0.11	99.88	0.01
2003	0.00	0.14	99.85	0.01
2004	0.00	0.17	99.82	0.01
2005	0.00	0.15	99.83	0.01

Table 7 Percentage of OAs, LSOAs and MSOAs near to thresholds, 2001-2005 mid-year estimates, England and Wales

(a) Output Areas (OAs) (n = 175,416)

Year	Below threshold			Around threshold			Above threshold		
	More than 10% below lower threshold	Between 5% to 10% below lower threshold	Less than 5% below lower threshold	Less than 5% above lower threshold	More than 5% above lower threshold and more than 5% below upper threshold	Less than 5% below upper threshold	Less than 5% above upper threshold	Between 5% to 10% above upper threshold	More than 10% above upper threshold
2001	0.12	0.05	0.08	0.09	99.24	0.07	0.05	0.05	0.23
2002	0.10	0.03	0.05	0.08	99.29	0.09	0.06	0.05	0.26
2003	0.11	0.04	0.04	0.08	99.16	0.11	0.07	0.06	0.32
2004	0.12	0.03	0.05	0.08	99.02	0.13	0.10	0.07	0.41
2005	0.12	0.03	0.05	0.08	98.87	0.16	0.10	0.08	0.49

(b) Lower Level Super Output Areas (LSOA) (n = 34,378)

Year	Below threshold			Around threshold			Above threshold		
	More than 10% below lower threshold	Between 5% to 10% below lower threshold	Less than 5% below lower threshold	Less than 5% above lower threshold	More than 5% above lower threshold and more than 5% below upper threshold	Less than 5% below upper threshold	Less than 5% above upper threshold	Between 5% to 10% above upper threshold	More than 10% above upper threshold
2001	0.00	0.00	0.04	0.45	99.39	0.03	0.02	0.02	0.05
2002	0.00	0.01	0.07	0.38	99.41	0.03	0.02	0.01	0.06
2003	0.01	0.01	0.05	0.39	99.35	0.05	0.03	0.03	0.07
2004	0.02	0.01	0.07	0.34	99.32	0.07	0.04	0.03	0.11
2005	0.02	0.02	0.08	0.32	99.23	0.09	0.06	0.04	0.15

(c) Middle Level Super Output Areas (MSOAs) (n = 7,194)

Year	Below threshold			Around threshold			Above threshold		
	More than 10% below lower threshold	Between 5% to 10% below lower threshold	Less than 5% below lower threshold	Less than 5% above lower threshold	More than 5% above lower threshold and more than 5% below upper threshold	Less than 5% below upper threshold	Less than 5% above upper threshold	Between 5% to 10% above upper threshold	More than 10% above upper threshold
2001	0.01	0.01	0.07	1.96	97.93	0.00	0.01	0.00	0.00
2002	0.01	0.01	0.08	1.65	98.22	0.00	0.00	0.01	0.00
2003	0.01	0.00	0.13	1.49	98.35	0.01	0.00	0.01	0.00
2004	0.01	0.00	0.15	1.38	98.40	0.04	0.00	0.01	0.00
2005	0.01	0.01	0.13	1.26	98.53	0.04	0.00	0.00	0.01

Nested breaches

It is also important to understand what types of breaches are occurring in order to ensure that the maintenance procedures will be able to deal with them. For example, the procedures required to deal with instances where an LSOA has breached a threshold but its constituent OAs have not, would almost certainly be different to those required where an LSOA has breached as a result of a high proportion of its constituent OAs breaching.

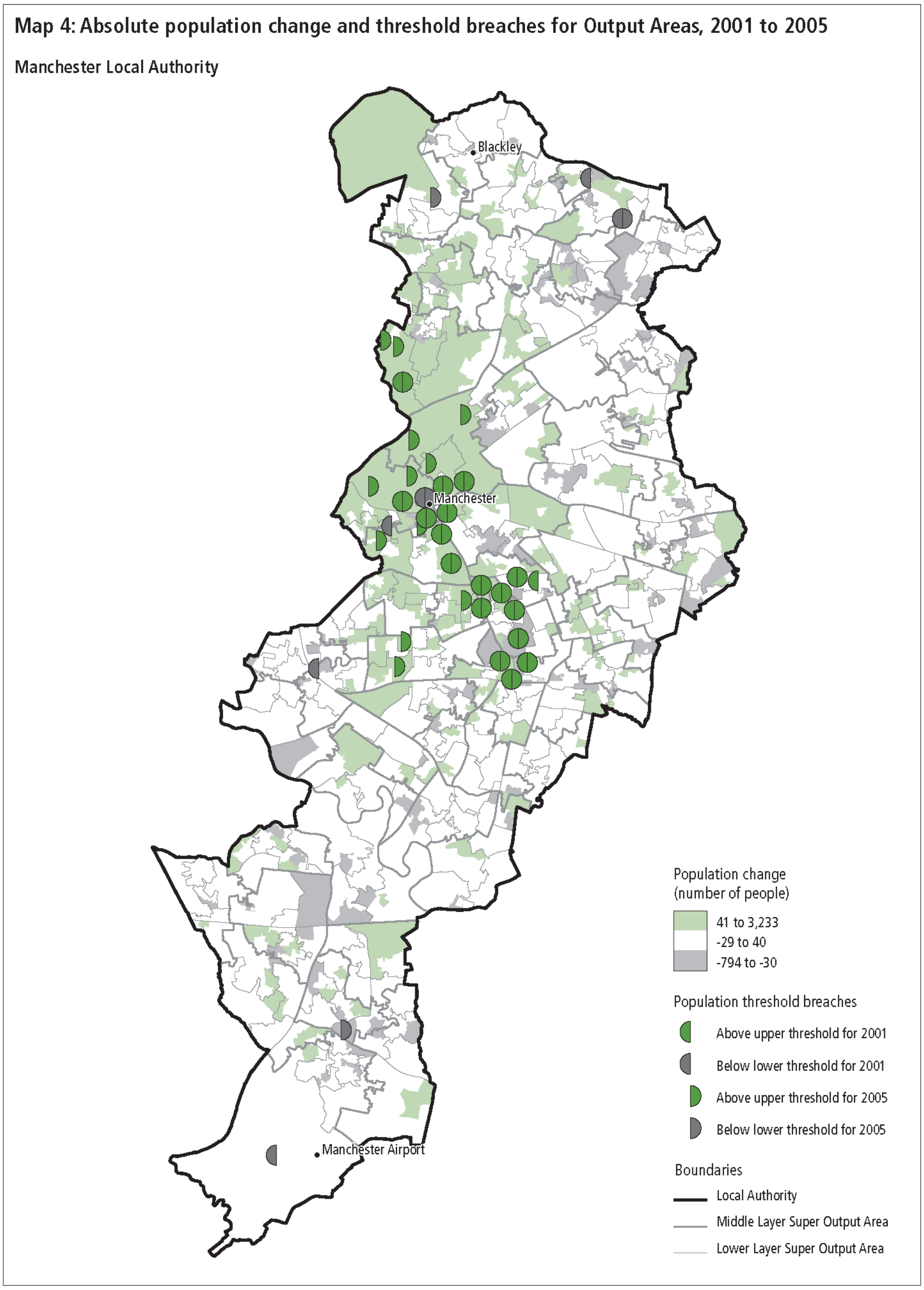
Of the 85 LSOAs above threshold by 2005, 83 contained above-threshold OAs within them. The percentages of OAs breached within an LSOA ranged from 75 per cent (three out of four) to 13 per cent (one out of eight), indicating that in a minority of areas, the LSOA breaches were not just due to one OA going significantly above threshold, but rather due to breaches across a number of OAs. The one MSOA that had gone above threshold by 2005 contained breaches at both the LSOA level (three out of 29), and at the OA level (two out of four OAs within the three above-threshold LSOAs).

In terms of below threshold breaches, only eight of the 40 LSOAs that had gone below threshold by 2005 also contained under-threshold OAs (ranging from 60 per cent (three out of five) to 20 per cent (one out of five). This suggests that the sub-threshold breaches have largely come about due to a general decrease in population across OAs within LSOAs rather than through significant decreases in specific OAs. Of the 11 MSOAs which were under threshold by 2005, only one contained any below-threshold areas within it (this one containing one out of four LSOAs breached but no OA level breaches).

Implications for maintenance of the 2001 output geographies

This analysis assumes that the ONS mid-year estimates provide an accurate picture of the rate and geographical distribution of population change since 2001. Any statistical or geographical bias in the mid-year estimates could significantly alter the levels and patterns observed. The number of breaches is of course dependent on the thresholds employed. It is possible that the factor (2.5) employed to calculate population thresholds from household thresholds over-estimates average household size and it has also been noted that average household size is decreasing over time. If this is the case, the number of threshold breaches reported here could under-estimate the scale of the problem. The magnitude of the breaches seen here is similar though to those previously reported by ONS[10], who employed different datasets and methodologies to explore potential output geography breaches by 2011. It is also not clear whether population will continue to change at the same rate and in the same geographical areas. For example, it is possible that some of the areas that have already undergone significant growth since 2001 may now become more stable. Growth may shift to other geographical areas, leading to new breaches in those areas, but this will be dependent on a number of factors such as trends in births, deaths, international and internal migration, economic prosperity and property development. Or growth may continue in the already breached areas, making the output geographies in these areas even more unsuitable. There are also uncertainties surrounding the extent to which the socio-economic homogeneity of the output geographies will have deteriorated by 2011: this article was unable to evaluate this because accurate contemporary tenure and accommodation type data were not available at the small area level, and indeed are only collected by the decennial Census.

If the trends presented here are accurate and do continue and if similar population thresholds are employed in 2011, it is likely that the majority of output geography areas will remain within threshold by 2011. The fact that population change tends to be strongly geographically clustered does mean though that in a minority of areas the output geographies are likely to be unsuitable for the release of 2011 Census data. In these areas, maintenance procedures that split, merge or completely re-design the existing geographies will be needed. Further, as a result of known issues related to the address register database used in 2001 and due to the complex and dynamic nature of the population in some areas (such as variations in the number of armed forces being stationed in some areas e.g. Rushmoor), some areas appear to have 2001 output geographies that are unlikely to have been optimal for the representation of population, even in 2001. For example, **Map 4** shows absolute population change and threshold breaches between 2001 and 2005 for Manchester UA. It is clear that a number of OAs had already breached the upper threshold by the time of the 2001 mid-year estimates, probably reflecting the fact that a large number of addresses were missing from the 2001 address register and were hence not used in the design of the output geographies. When corrections were made for these missing addresses, the size of the population in some areas will have increased, in some cases taking the OAs above threshold even by the time of the 2001 mid-year estimates. While these breaches arose for understandable, and often unavoidable reasons with respect to zone design, they do now present challenges for the maintenance of the geographies for 2011: should they be left as they are, maintained (that is split or merged), or completely re-designed?


Map 4: Absolute population change and threshold breaches for Output Areas, 2001 to 2005
Manchester Local Authority
Blackley
Manchester
Manchester Airport
Population change (number of people)
🟩 41 to 3,233
□ -29 to 40
▨ -794 to -30
Population threshold breaches
◖ Above upper threshold for 2001
◖ Below lower threshold for 2001
◗ Above upper threshold for 2005
◗ Below lower threshold for 2005
Boundaries
Local Authority
Middle Layer Super Output Area
Lower Layer Super Output Area

Source: Mid-year population estimates, Office for National Statistics

Contains National Statistics data © Crown copyright and database right 2013
Contains Ordnance Survey data © Crown copyright and database right 2013

Map 4 Absolute population change and threshold breaches for Output Areas, 2001-2005, Manchester Local Authority

Challenges involved in maintaining the 2001 output geographies

In 2007 ONS undertook a consultation on users' requirements for the 2011 Census small area output geographies. The consultation suggested that the majority of users would prefer to see the output geographies remaining stable rather than re-designing them completely for 2011. There were mixed views on the desirability of using postcodes as the building blocks for the 2011 geographies. Some users would prefer to see postcodes retained, while others would prefer the use of alternative building blocks such as street blocks. Some users argued for a better alignment of the output geography boundaries with real-world features. Some suggested that the OAs should better represent, and not split, 'neighbourhoods' or 'communities'. There was also some support for the resolution of known issues such as the address register problems experienced in Manchester and Westminster in 2001.

Population change since 2001, together with the requirements flagged by user consultation, present complex challenges for the design of the 2011 output geographies. Consequently, a series of important and potentially conflicting decisions need to be made concerning the processes, datasets and criteria employed to maintain 2001 output geographies:

- Perhaps foremost, is the decision as to whether geographic stability should be maintained at the OA, LSOA and MSOA levels where possible, or just at the LSOA and MSOA levels.
- For OAs which need to be split, what building blocks should be employed? The use of postcodes would allow datasets that are geo-referenced by postcode to be linked with census data. However, in some areas, the link between 2001 postcodes and OAs will have eroded during the inter-censal period making them less useful. Also, the use of postcodes for aggregation in the zone design process can be constraining in certain situations, for example, where postcodes are split across OAs or where vertical stacks (addresses with the same grid-reference but different postcodes) exist. Alternative building blocks, such as street blocks, have advantages in that they may appear to be more aligned to real-world features, but the ability to link via postcodes would be lost.

- A closely related issue is whether the existing and/or maintained output geographies should be aligned or re-aligned to real-world features. The output geographies were always intended to be synthetic statistical boundaries, albeit constrained where possible to some geographical features, such as roads. Alignment to more geographical features would possibly make the boundaries appear more 'real' but attempting to do so would be arguably futile: there is no agreed set of 'real world' features to align to; such alignment would introduce conflicts with other zone design goals, such as maintaining inter-censal stability; and doing so would raise significant boundary copyright issues, thereby potentially impairing ONS' ability to freely distribute the boundary data as an integral part of the Census outputs.
- Should the maintenance procedures ensure that the output geography boundaries nest within wards or local authority districts? This would be difficult given the regular changes in ward and LA boundaries and may lead to the need for a large number of changes to the output geography boundaries. It would also move away from ONS' stated policy of retaining stability of the output geographies over time.
- Should the same design criteria and values as in 2001 be employed? For example, should the same thresholds, targets and shape constraints be used?

In addition, the process of maintaining the 2001 output geographies should ideally be automated to enable the systematic, objective and efficient creation of the geographies for all of England and Wales in a timely manner following the collection of census data in 2011.

Conclusions

This article has explored the magnitude and geographical distribution of population change since the 2001 Census in the context of maintenance of the 2001 Census output geographies. Using mid-year estimates, it concludes that virtually all output geography areas had not breached upper or lower thresholds by 2005, and are unlikely to do so by 2011. Nonetheless, because population change is usually strongly geographically clustered, in some areas there have already been significant breaches of population thresholds; the output geographies in these areas and others are therefore likely to need maintenance in order to be suitable for the release of 2011 Census data. The challenges involved in carrying out this maintenance are non-trivial and this article identifies some of the key decisions that need to be taken before the maintenance procedures can be developed and implemented. The on-going ESRC-funded Census2011Geog project will develop prototype software for carrying out the automated maintenance procedures and will also evaluate the usefulness of different building blocks and maintenance methods. ONS will then need to evaluate these findings, make key policy decisions and then implement the procedures following the collection and collation of the 2011 Census data.

Key Findings

- The vast majority of Output Areas, Lower Layer and Middle Layer Super Output Areas had not breached specified population thresholds by 2005, and seem unlikely to do so by the 2011 Census
- Population change is strongly geographically clustered so in a minority of areas the 2001 output geographies will not be appropriate for the release of 2011 Census data
- Automated maintenance procedures that split, merge or re-design the 2001 output geographies are being developed by the ESRC-funded Census2011Geog project in collaboration with ONS

Acknowledgements

This research is supported by Economic and Social Research Council Award RES-348-25-0019. It uses mid-year population estimate data at OA, LSOA, MSOA and local authority levels provided by ONS, the local authorities' urban/rural classification from DEFRA and 2001 Census digitised boundary data from UKBORDERS which is Crown copyright and reproduced with the permission of the Controller of HMSO. The authors are grateful to Professor David Martin, Andy Tait, Andy Bates, Donna Viney and three anonymous reviewers for useful comments on earlier drafts, and to Nick Richardson for help with production of the maps.

References

[1] Office for National Statistics (2007) *National Statistics Small Area Geography Consultation 2007*. Available at: www.ons.gov.uk/about/consultations/closed-consultations/geography-policy-public-consultation/index.html

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

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

[4] Cockings, S and Martin, D (2005) Zone design for environment and health studies using pre-aggregated data *Social Science & Medicine* 60(12), 2729-2742.

[5] Mitchell, B and Ralphs, M (2007) *Developing maintenance rules for the Neighbourhood Statistics Output Geographies*, Unpublished research report. Methodology Directorate, Office for National Statistics.

[6] Office for National Statistics (2004a) *2004 Local Authority Studies: The Longitudinal Based Study Consequential Adjustment.* Available at: www.statistics.gov.uk/downloads/theme_population/LAStudy_LS_ConsequentialAdjustment.pdf

[7] Office for National Statistics (2004b) *2001 Census Local Authority population Studies: Full report.* Available at: www.statistics.gov.uk/downloads/theme_population/LAStudy_FullReport.pdf

[8] Office for National Statistics (2004c) *2001 Census: Manchester and Westminster Matching Studies Full Report.* Available at: www.statistics.gov.uk/downloads/theme_population/ManchesterandWestminster_FullReport.pdf

[9] Department of Environment, Food and Rural Affairs (2005) *Defra Classification of Local Authority Districts and Unitary Authorities in England: A Technical Guide.* Available at: www.defra.gov.uk/evidence/statistics/rural/rural-definition.htm

[10] Ralphs, M and Mitchell, B (2006) *Maintenance requirements for Super Output Area geographies: modelling changes from 2001-2006,* Unpublished research report. Methodology Directorate, Office for National Statistics.

4.1 Paper 3: Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales

Cockings S, Harfoot D, Martin D and Hornby D "Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales"

The version of this article included in this thesis is the author's own post-print copy. The definitive, peer-reviewed and edited version of this article is published in *Environment and Planning A*, 43(10) 2399-2418 2011 http://dx.doi.org/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, UK

* Corresponding author: Tel: +44 (0)23 8059 5519 Fax: +44 (0)23 8059 3295

E-mail: s.cockings@soton.ac.uk; ajph@geodata.soton.ac.uk; d.j.martin@soton.ac.uk; ddh@geodata.soton.ac.uk

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 (e.g. 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, non-standard, zoning systems are defined on an ad hoc basis, often for a specific study or application, and are generally only used 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 e.g Cockings and Martin, 2005; Flowerdew *et al.,* 2007; Haynes *et al.*, 2007 and 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 e.g. parishes (the lowest level of local government in England) have survived largely due to notions of neighbourhood identity and local representation.

When needing to update an existing system, most countries or designers have chosen either to completely re-design 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) i.e. 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, semi-automated 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 two 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; section three proposes a generic methodology for the automated maintenance of existing zoning systems; in section four 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, section 5 discusses the results of the empirical example, both in terms of its implications 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 (MAUP) (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; 1997b) and subsequently enhanced by Openshaw and Rao (1995), Alvanides (2000) and Alvanides, Openshaw and Rees (2002). The AZP algorithm works by iteratively combining and re-combining sets of building blocks in order to create output zones which optimise a set of pre-specified 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 pre-existing building blocks or input zones. In others (e.g. Haynes *et al.*, 2007, 2008; Flowerdew *et al.*, 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 e.g. 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 re-designing 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 to both 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 sub-set of the system’s zones, most likely via a combination of splitting, merging or complete re-design 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).

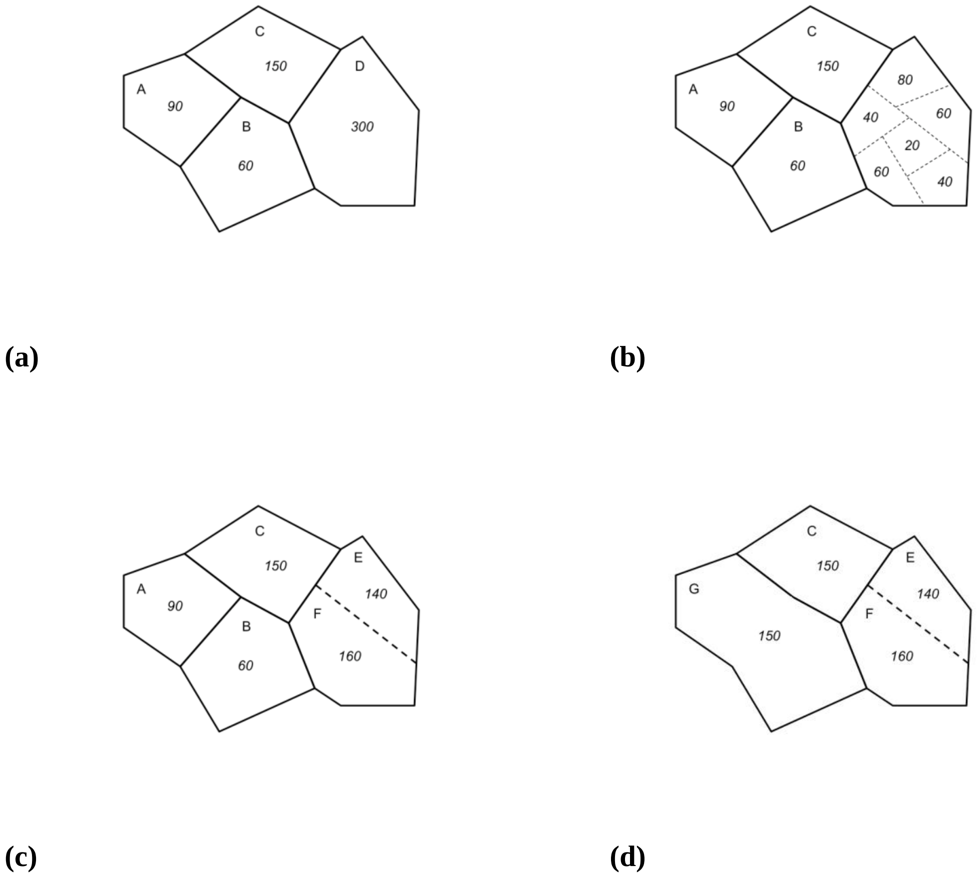


Figure 1 Simplified example of a zoning system requiring maintenance (lower threshold 100; upper threshold 250): (a) original (input) zones in zoning system; (b) building blocks for zone D which needs to be split; (c) split zone - input zone D has been split into two new output zones (E and F); (d) merged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

First, the input zones are separated into two groups: (i) over-threshold zones and (ii) within- 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 only takes place 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 (e.g. 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- 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 re-design 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 i.e. where lower-level sets of zones nest within one or more higher-level sets of zones (e.g. 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 (e.g. local to regional or regional to local) may influence the ability to successfully split, merge or re-design 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 sub-sets of zones within the system, often at different levels of geography (e.g. 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 re-design 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 (i.e. those that were already within-threshold and have not been used for merging with under-threshold zones, or those which were under- 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.

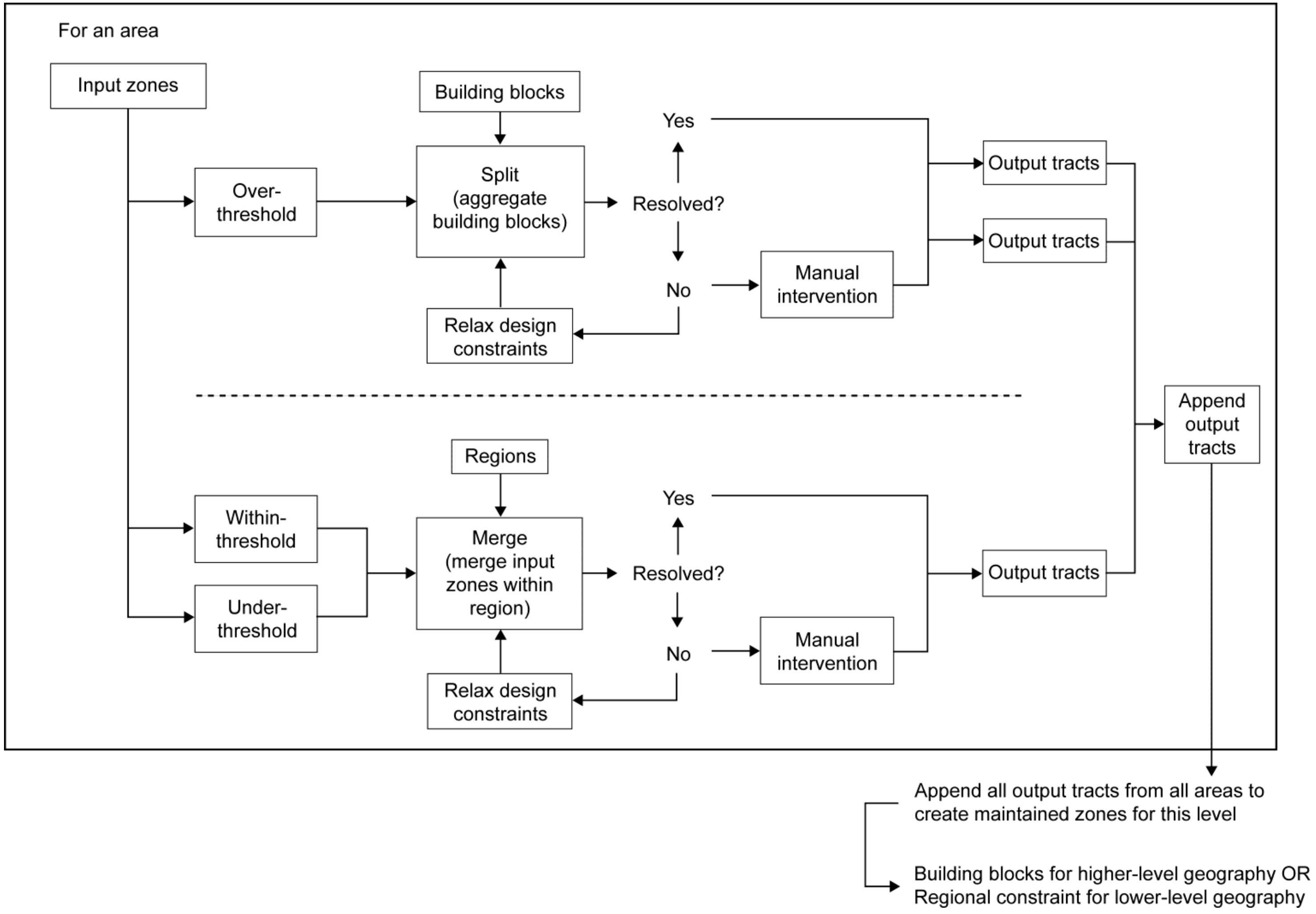


Figure 2 Generic automated maintenance method

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 building block base for the next 25 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 (e.g. Manchester, Westminster) are known to have output geographies which were not optimal for the release of 2001 data: there is therefore a case for completely re-designing 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, p3). 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 re-designed, as appropriate, in order to make them suitable for the publication of 2011 Census data.

4.2 *Methods*

4.1.1 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	N/A	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	Under-enumeration problems 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 to 2006 mid-year estimates for LADs and 2001-2005 for OAs, LSOAs and MSOAs: low = < 5% change; mid = 5-10%; high = > 10%

[c] Area used by ONS in 2007 Census Test (ONS, 2009), 2009 Census Rehearsal (ONS, 2010), or Small Scale Tests to support field work (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 (OSAPR), 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 residences) 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 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 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.

4.1.2 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 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- 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- and over-threshold OAs, but the majority of OAs remain within-threshold.

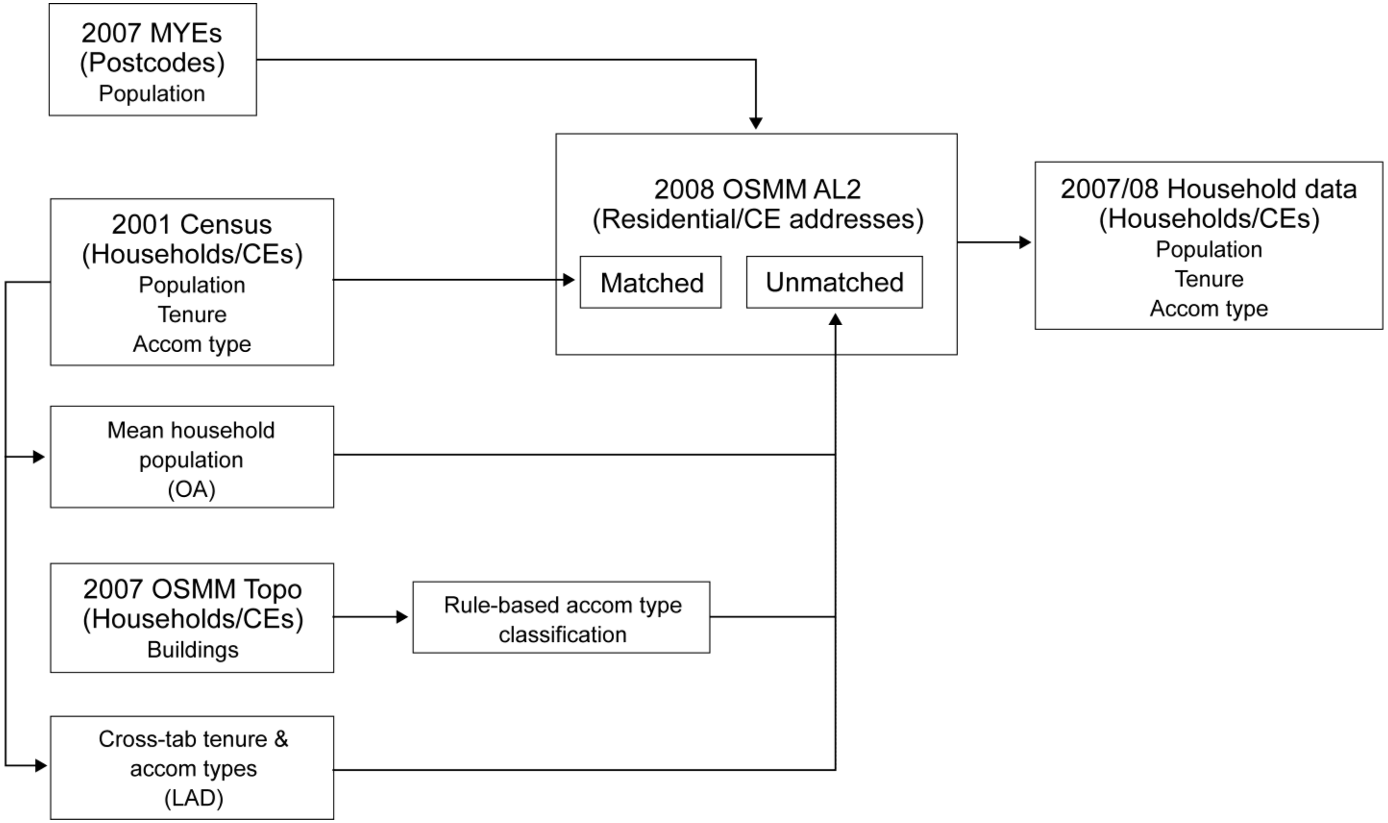


Figure 3 Methodology for creation of household-level data (AL2: Address Layer 2; CE: communal establishment; LAD: local authority district; MYE: Mid-year estimates; OA: output area; OSMM: Ordnance Survey MasterMap™).

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
LSOA	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)

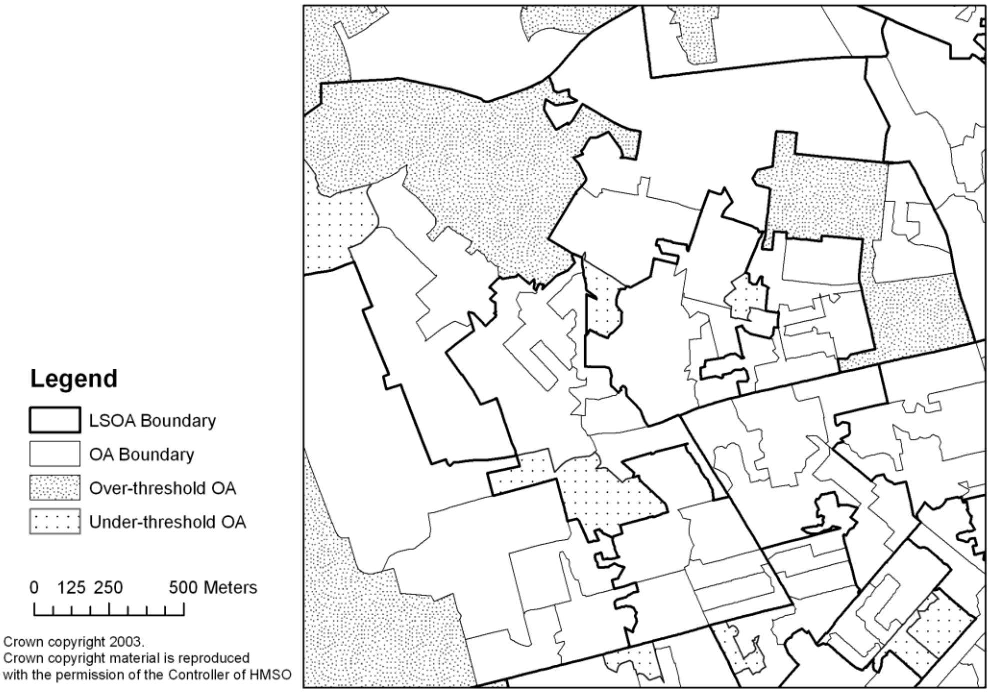
[c] Household threshold values from Mitchell and Ralphs (2007), Table 1.1, p.4

[d] No upper thresholds published in 2001 for OAs or LSOAs. Values = 2001 OAPS target mean * 2 (as in Ralphs and Mitchell, 2006). MSOAs did have 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*
OAs	4988	43	4836	109
LSOAs	962	12	938	12
MSOAs	200	1	198	1

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



Crown copyright 2003.
Crown copyright material is reproduced with the permission of the Controller of HMSO

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.

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 are unlikely to use a decline in socio-economic homogeneity as a reason for maintaining a zone in 2011[1], although this may be of more concern to some users.

4.1.3 Implementation and evaluation of automated maintenance procedures using AZTool

Enhancements to the authors' 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://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

[1] ONS have stated that they *may* re-design exceptional instances of OAs which were found to be socio-economically 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/census/2011-census/consultations/open-consultations/census-output-geography-consultation/index.html

geography went on to become the building blocks or regional constraints, as appropriate, for the next level to be maintained.

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 type* [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
2007/08	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
2007/08	200	7846.3	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).

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 (e.g. LSOA, MSOA)	Na

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

Where solutions could not be found using all of the constraints, an iterative process of relaxing constraints and re-running the procedures was undertaken. First, the minimum boundary length 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. Where identifiable, a reason for this non-resolution 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

4.1.4 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 (e.g. an OA) fell within an over-threshold higher-level geography (e.g. 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 e.g. 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).

4.1.5 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.

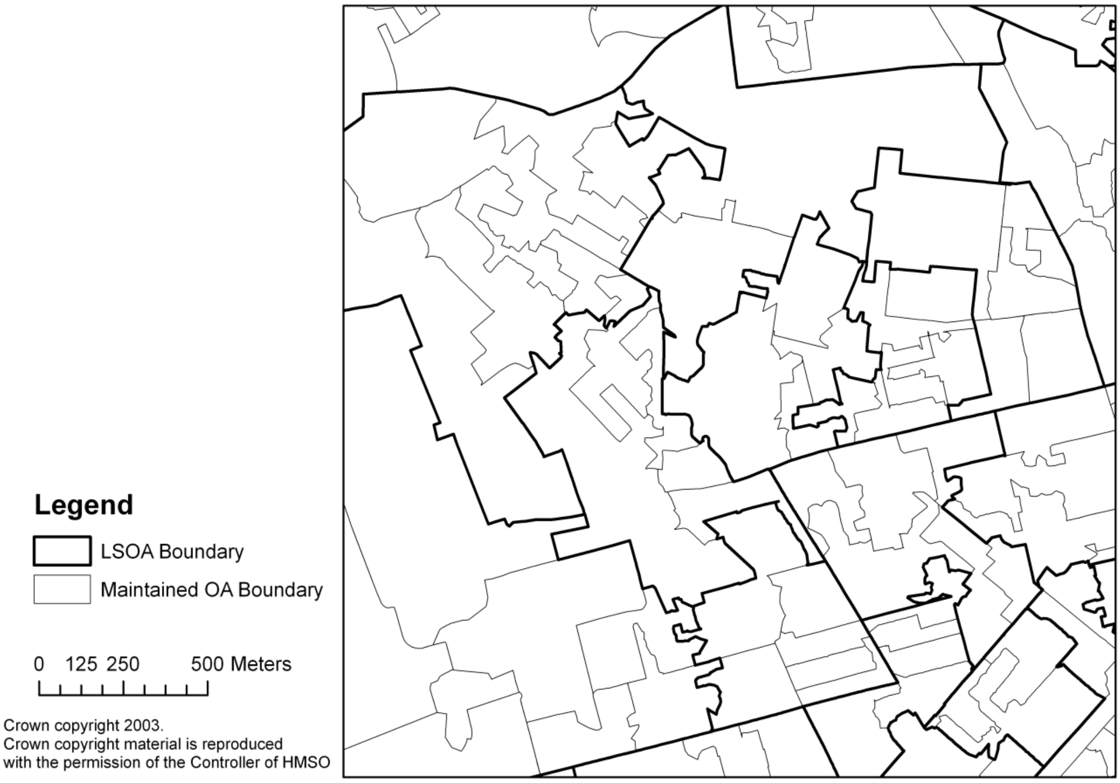
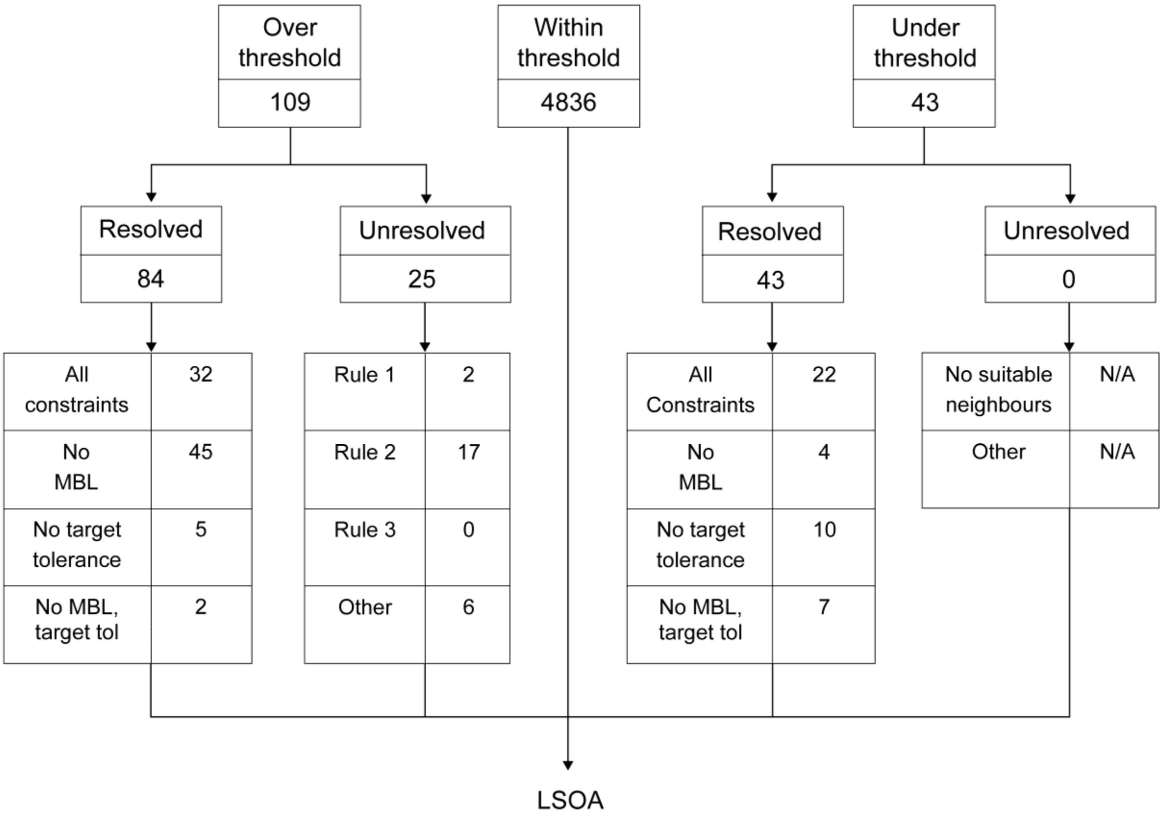

Legend
LSOA Boundary
Maintained OA Boundary
0 125 250 500 Meters
Crown copyright 2003.
Crown copyright material is reproduced
with the permission of the Controller of HMSO

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

Figure 6 details how many of the over- 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 minimum boundary length (MBL) constraint, then the target tolerance and finally both the minimum boundary length and target tolerance[2] . Where quantifiable, reasons for the non-resolution of zones are also shown.

[2] Relaxing the minimum boundary length 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)

With all constraints in place, only 29% of over-threshold and 51% of under-threshold OAs could be resolved. Relaxing the minimum boundary length 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, 3 out of 12 over-threshold zones were resolved and 2 out of 12 under-threshold zones. After relaxing both the MBL and target tolerance, 6 out of 12 over-threshold and 10 out of 12 under-threshold LSOAs were successfully resolved. Of the two MSOAs requiring maintenance (one over- 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 non-resolution of over-threshold OAs (17 out of 109) was where at least one of the OA's constituent building blocks (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 (3 of the 6 unresolved LSOAs) or a specific geometric configuration of building blocks which prevented a solution being found (2 out of 6). A possible solution to the over-threshold building block problem would be to sub-divide 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 sub-block(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, 34 of the 109 over-threshold OAs and 2 of the 12 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.

4.1.6 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 post-maintenance OAs were actually very slightly better (i.e. 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 to 2001[3] .

Table 4 also presents the post-maintenance 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 24 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 pre-maintenance.

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.

[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.

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. Sub-divisions of existing dwellings and dwellings which have been newly built should have been accurately identified and populated, but instances where the population count or 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 sub-input zone-level building blocks employed to split over-threshold zones are available to the designer (i.e. 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 a researcher 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 (i.e. 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 re-processed, sequentially relaxing specified constraints, until all zones are resolved or no more constraints can be relaxed. Some zones e.g. 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- 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 e.g. 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 re-design 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 over-sized constituencies (using some combination of wards, electoral areas, electoral divisions and/or polling districts as the building blocks) and to merge under-sized ones (with other under-sized 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 re-designing an existing set of zones because the design criteria have changed, the basic process of maintenance (i.e. splitting, merging or re-designing) 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 license 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 www.universitiesuk.ac.uk.

References

Alvanides S, 2000, “Zone design methods for application in human geography”, PhD Thesis, School of Geography, University of Leeds, Leeds

Alvanides S, Openshaw S, Rees P, 2002, “Designing your own geographies”, in *The Census Data System* Eds P Rees, D Martin, P Williamson (Chichester, Wiley) 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, Wellington, New Zealand, pp 40

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 Southampton, Southampton, http://census2011geog.census.ac.uk

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

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

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 & 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, 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* **31** 1 6 - 18

Martin, D, 2010, “Understanding the social geography of social undercount” *Environment and Planning A* **42** 2573 - 2770

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

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

Mitchell B, Ralphs M, 2007, “Developing maintenance rules for the Neighbourhood Statistics Output Geographies”, Methodology Directorate, Office for National Statistics.

ONS, 2004 *2001 Census: Manchester and Westminster Matching Studies Full Report,* Office for National Statistics, http://www.statistics.gov.uk/downloads/theme_population/ManchesterandWestminster_FullReport.pdf

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

ONS, 2009 *2007 Census Test: Summary Evaluation Report*, Office for National Statistics, http://www.ons.gov.uk/census/2011-census/2011-census-project/2007-test

ONS, 2010 *2011 Census: Evaluation of the 2009 Rehearsal*, Office for National Statistics, http://www.ons.gov.uk/census/2011-census/2009-census-rehearsal/index.html

ONS, 2011 *2011 Census – England and Wales Output Geography: Policy and Products*, Office for National Statistics, http://www.ons.gov.uk/census/2011-census/consultations/open-consultations/2011-geography-outputs-consultation/index.html

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 and N Thrift (Elsevier, Oxford) pp 298 - 301 http://www.sciencedirect.com/science/referenceworks/9780080449104

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 & Societies* **467** May 2010 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

4.2 Paper 4: Development of a Geographical Framework for Census Workplace Data

Martin D, **Cockings S** and Harfoot A "Development of a Geographical Framework for Census Workplace Data"

Development of a Geographical Framework for Census Workplace Data

David Martin[1] , Samantha Cockings and Andrew Harfoot

University of Southampton, UK

Summary. This paper addresses problems arising from the representation of workplace population data using geographical areas based on residential locations. This widespread international practice detrimentally affects publication of census workplace data. A novel solution is proposed for creation of new workplace zones using automated zone design techniques and applied to five prototype areas using England and Wales 2001 census microdata. Particular workplace-based disclosure control challenges are addressed and the characteristics of the proposed workplace zones reviewed. This approach offers important benefits for the international reporting of workplace data and is currently being incorporated into England and Wales 2011 census output plans.

Keywords. Output geography, census, workplaces, automated zone design, statistical disclosure control

[1] *Address for correspondence:* David Martin, Geography and Environment, University of Southampton, Southampton, SO17 1BJ, UK.
E-mail: D.J.Martin@soton.ac.uk

1. Introduction

Many contemporary national population statistics systems make extensive use of residence-based population definitions and predominantly publish data for geographical areas which have been designed based on the spatial distribution of places of residence. However, as Bhaduri (2008) observes, in modern societies the locations of residence and daytime activities are generally very different. Workplaces, in particular, provide a key alternative distribution of population, being a major focus of economic activity and generating some of the greatest regularly occurring concentrations of population. In general, commercial and employment centres have relatively few residents while most residential areas have few workplaces. The importance of workplace information is reflected in its inclusion in national data collection exercises such as the 2011 UK and 2006 Canadian censuses (Cabinet Office, 2008; Statistics Canada, 2006) and American Community Survey (US Census Bureau, 2009). Countries such as Finland, with population data systems based on administrative records, explicitly link residential and workplace locations of individuals (Statistics Finland, 2004).

While these official statistical systems include information about workers and workplaces, the resulting counts tend to be published for residentially-based areas, reflecting the primary concern with residents and households. National statistical organizations employ trade-offs between the amounts of geographical and statistical detail published, to protect the confidentiality of respondents (ONS, 2011a). The overall consequence is that many areas contain either too few businesses and workers for publication of even moderately detailed data, or large and diverse workplace populations with insufficient statistical disaggregation. This presents severe obstacles to those needing to analyse workplace statistics, employment structures, daytime population distributions or travel to work patterns. This is a generic problem requiring a novel solution.

In many countries, workplace-based counts are only available as a limited set of additional data for the standard residence-based areas (for example: Statistics New Zealand, 2006; Statistics Canada, 2008). Such data are not universally available for the smallest areas and display many of the weaknesses outlined above. By contrast, Australia and the USA offer an additional non-residential geography for the publication of workplace data, primarily developed to provide more appropriate destination zones for transportation modelling (Australian Bureau of Statistics, 2011; Federal Highway Administration, 2010). Neither is based on detailed workplace locations and only the Australian solution is explicitly based on an aggregation of workplace population. Both have been designed by regional transportation authorities, outside the main population statistics systems.

In this paper, we argue that there is a pressing need to develop a methodology for producing small areas for workplace data publication which truly reflect the geography of workplace locations. At the same time, these areas need to be fundamentally integrated with the national statistical organization’s hierarchy of data publication areas and to address all the associated disclosure control and end user requirements. We propose an innovative generic solution based on automated zone design and demonstrate its application in the context of planning 2011 census outputs in England

and Wales. Section 2 considers automated zone design applied to census data outputs and reviews the way in which the challenge of meeting census users' requirements for workplace data has been addressed internationally. Section 3 sets out a methodology for the creation of new workplace zones (WZs) and section 4 describes the implementation of a prototype based on five study areas in England and Wales. Results are presented in section 5 and a range of conclusions and recommendations are made in section 6.

2. Zone design for census workplace outputs

It is not the purpose of this paper to review automated zone design procedures in any detail as these have been covered extensively elsewhere (Martin, 1998; Martin et al., 2001). Rather, we consider how these techniques have been applied to the publication and analysis of census data, with particular attention to England and Wales, which provides the basis for our empirical study. Regionalisation algorithms have been applied to published census results for the creation of travel to work areas (TTWAs) (Coombes et al., 1986) or Metropolitan Statistical Areas (Adams et al., 1999) and these are well-established concepts in the UK and US respectively. There has been considerable interest in the use of automated zone design to produce census areas optimised for the analysis of flow data, including travel to work, for example by Alvanides et al. (2000) and Martinez et al. (2009). The objective in these cases is the demarcation of functionally meaningful labour market, transport analysis or metropolitan areas. By contrast, our primary concern is to provide national coverage of small geographical areas for workplace data which can be produced by a statistical organization prior to aggregation and publication.

Openshaw and Rao (1995) proposed the use of automated zone design methods to produce small areas with desired characteristics but again they were limited to re-engineering already-published census small areas. For the 2001 census in England and Wales, the Office for National Statistics (ONS) implemented a development of Openshaw's (1977) automated zoning procedure to assemble small building block zones into an 'optimal' aggregation for data publication, based on a set of design criteria. Essentially the same approach will be used to update these areas in 2011 (Cockings et al., 2011). Ang and Ralphs (2008) evaluate the same method as a potential means of generating a revised census geography in New Zealand.

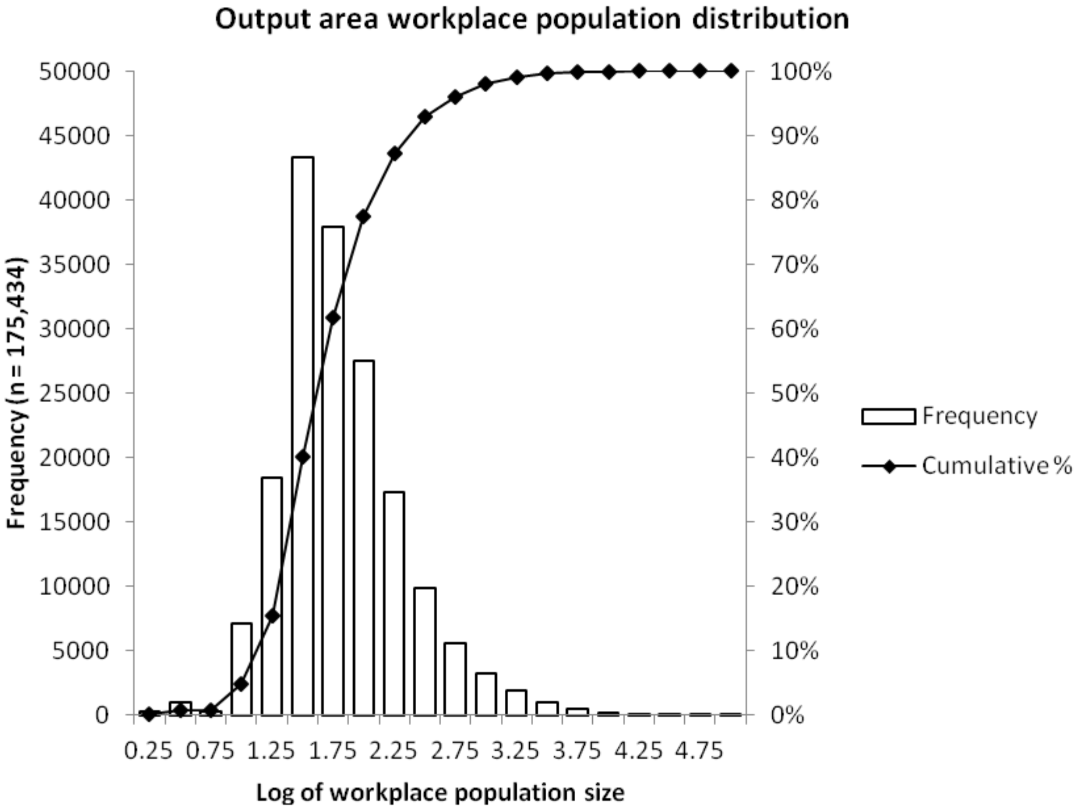
In common with much international practice, area-based statistics from the 2001 census in England and Wales were published for a hierarchy of geographical units, the smallest of which are termed output areas (OAs). The design of OAs was explicitly based on the geography of places of residence. The building blocks for 2001 OAs were whole or part unit postcodes, the smallest element of the UK postal referencing system, typically containing around 15 addresses. Some postcodes were split across higher-level administrative boundaries. As unit postcodes do not have definitive boundaries, polygons were first generated around address point locations and then merged to form polygons for postcode building blocks. Boundaries were clipped to follow a variety of additional geographical features as described in Martin (1998), although these data were limited by the requirement that OA boundaries would be distributed freely to census users. Postcode-based building blocks were aggregated to create OAs with resident populations above census confidentiality thresholds of 40 households and 100 persons, using a target size of 125 households, resulting in 175,434 OAs with a mean population of 297. Further design constraints were that OAs display geographical compactness (using a postcode centroid clustering measure) and social homogeneity (using an intra-area correlation measure based on tenure and accommodation type).

The method for creating 2001 OA geography was considered by ONS to be a significant success (ONS, 2004a), overcoming many of the deficiencies of earlier approaches and providing a structured mechanism for trading off the many (ultimately irreconcilable) user requirements. In particular, the consistent residential population sizes of OAs, with standard deviations of 71.65 persons and 16.86 households, replaced previous wide variations which had resulted in suppression of data for sub-threshold populations (in 1991, this applied to 3890 small areas for population and 4990 for household data). Following publication of 2001 results, the automated zone design methodology was re-applied to the OA layer in order to produce larger geographical units known as super output areas, which have become part of the standard geographical hierarchy for non-census neighbourhood statistics.

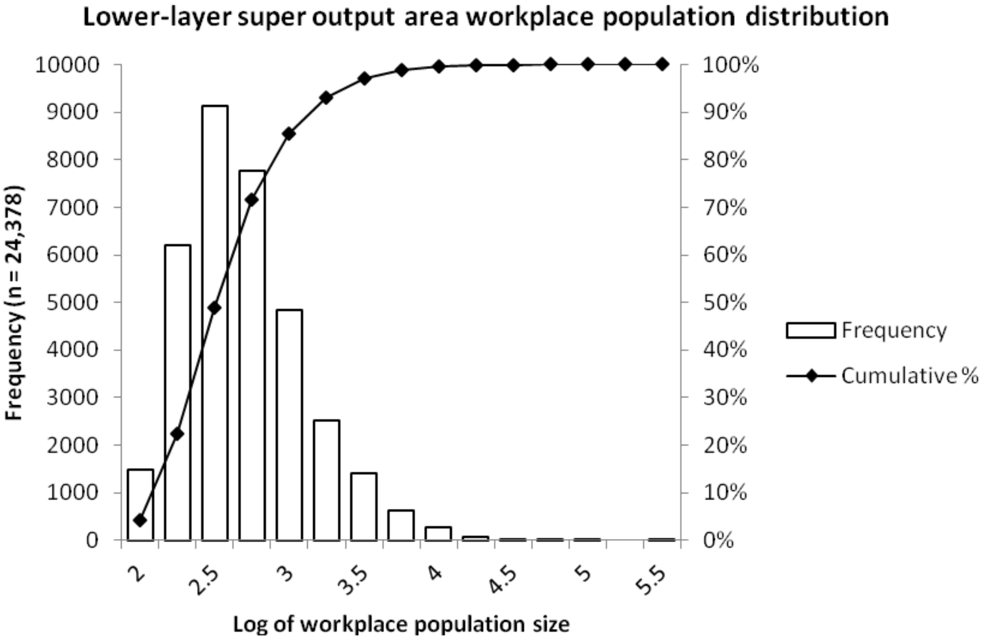
In common with other countries, the 2001 census in England and Wales included a question about usual place of work, which was subsequently utilised to produce workplace statistics and travel to work flows. Workplace postcodes were allocated to OAs using a point-in-polygon methodology. However, as in other countries, the completeness and accuracy of these reported workplace addresses varies widely and there is no definitive way of assigning exact locations to all workplaces. Statistical organizations edit and impute some records to allocate workers to workplace locations, whether based on census forms (ONS 2005; Statistics New Zealand, 2007) or derived from administrative lists (Statistics Finland, 2004). This is necessary, for example, where respondents fail to supply a workplace address or provide postal boxes or depot addresses which are not actual locations of work. It is also difficult to identify workplace locations within multi-site businesses or those with large physical extents. It is important to understand that although censuses ask about place of work, individual workplaces are not a key field in the census database, but rather they are immediately aggregated into (residentially-defined) geographical areas.

It is a core feature of UK census output geography design that there should be no vertically ‘stacked’ OAs, due to the difficulties this would present to users for mapping, linking and analysis, so addresses sharing the same map coordinates would necessarily be allocated into the same OA. The list of postcodes for which building block polygons were created was that generated from residential households and mixed-use dwelling spaces (e.g. apartment above a shop) on the census questionnaires. However, postcodes recorded only as workplace locations (e.g. office building or factory) were not used in OA creation and are not represented in the boundary set. These missing postcodes will often equate to large businesses or clusters of businesses on wholly non-residential sites. Users have noted that OA boundaries sometimes cut through non-residential buildings, which is a direct consequence of this process, which took no account of the physical structure of non-residential addresses.

The 2001 ONS implementation shows that automated zone design can deliver consistent residential population sizes but this still masks massive variations in workplace populations (i.e. the counts of people working in those areas). Fig. 1(a) shows the distribution of workplace populations for 2001 census OAs in England and Wales, ranging from 0 (various OAs) to 80,145 (00AAFE0001, City of London) with a mean of 134. The range is so great and the distribution so skewed that it is necessary to graph the log of workplace population sizes rather than their absolute values. Fig. 1(b) shows the same distribution for lower-layer super output areas (LSOAs), which ranges from 24 to 264,439 with a mean of 684. Due to post-tabulation modification of small cell values to protect confidentiality (Boyle and Dorling, 2004), the precise counts vary slightly between published census tables: these figures are based on Table UV75. There are two direct and important consequences of these highly skewed distributions. The first is that, after a review of disclosure risks, ONS decided not to produce all the planned workplace statistics, so as to avoid the possible identification of workplaces. At the OA level, only four univariate tables were published for workplace populations: age (UV75), National Statistics Socioeconomic Classification (UV76), approximated social grade (UV78) and distance travelled to work (UV80). Notably, some key variables such as industry, occupation and mode of travel to work were omitted. A second consequence is that users were left without any geographical subdivision of large working populations in major urban centres. Thus the residential geography not only limited the scope of the workplace data available, but is also an inadequate division of space for geographical representation of workplaces.



(a)



(b)

Fig. 1. Workplace population distribution by (a) output areas and (b) lower-layer super output areas, England and Wales, based on 2001 Census table UV75.

The US Federal Highway Administration (2010) has attempted a partial solution to the problem of widely varying workplace population sizes by the creation of Traffic Analysis Zones (TAZs) and Transportation Analysis Districts (TADs). TAZs are defined with reference to both resident and worker populations, setting a minimum workplace population of 600 and ideal resident populations of 1,200 and 22,000 for TAZs and TADs respectively. This maximises the publication of non-disclosive data but does not ensure similarity of workplace population sizes between zones. Being aggregations of census blocks, they do not usually cut across commercial premises, but neither are they explicitly delineated on the basis of workplace locations. Importantly, these are optionally defined by metropolitan and state transportation authorities and provide neither complete national coverage nor a nationally consistent set of dissemination areas. In Australia, Destination Zones (DZs) have been similarly defined by state/territory transportation authorities, but do provide national coverage (Australian Bureau of Statistics (ABS), 2011). They are used to code place of work data from the census and from 2011 will comprise aggregations of mesh blocks, the basic spatial units for the Australian Statistical Geography Standard (ABS, 2012). Neither the US or Australian approaches constitute a set of systematically designed units within the national statistics outputs system and both involve significant manual processing. In the UK, 2011 census user consultation (ONS, 2010) has demonstrated strong demand for geographical subdivision of very large OA workplace populations, a broader range of workplace-based statistics, more meaningful building block boundaries and the potential use of industrial structure as a homogeneity constraint in the zone design algorithm.

3. Workplace zones: a solution

We present a method for the creation of WZs for which more extensive and useful workplace statistics may be published than previously, while at the same time maintaining a strict hierarchical relationship with census OAs. The empirical work is based on England and Wales, although the approach would be readily transferable to other countries. The research has involved use of 2001 census microdata in order to develop a method applicable to the production of 2011-based census WZs. The OA geography for 2011 will represent a slight modification of that produced in 2001, reflecting population and local government changes (ONS, 2010). The production of WZs will therefore be undertaken after the finalisation of residential OAs.

To prevent the inadvertent disclosure of data about identifiable respondents arising from overlapping geographical areas (Duke-Williams and Rees, 1998), it is a basic requirement that WZs should be either exact subdivisions, aggregations or matches to the residence-based OAs. The methodology underlying the proposed creation of WZs is as follows. OAs must be analysed in order to determine those requiring aggregation or subdivision to meet specified thresholds (considered below) which permit the publication of a given range of output data. It is immediately apparent that threshold size and the range of data to be published are therefore highly interdependent. Many OAs will have workplace populations which already fall within the acceptable size range for WZs and will require no further processing. Merging some OAs would increase their combined workplace populations to the threshold size required of a WZ, and in these cases no new boundary data would be required. The key concern therefore relates to those OAs which require subdivision because they contain very large workplace populations.

The lack of detailed locations for individual businesses means that the postcode building blocks created during the 2001 census processing were not the most appropriate from which to generate WZs. For present purposes, in advance of 2011 data being available, we therefore created an entirely new set of building blocks comprising only the non-residential postcodes from 2001. This has the advantage of strongly reflecting the distribution of workplaces as represented by the locations of their postcodes. Our processing sequence is outlined in Fig. 2 which begins at top left with the existing residential OAs. This is a modification of the generic processing sequence proposed by Cockings *et al.* (2011) for the automated maintenance of existing zoning systems. OAs are firstly separated into (i) those requiring subdivision, and (ii) those which are either acceptably-sized already or which require merging. The same zone design algorithm is employed in both circumstances, but with different input building block sets in each case. Within those OAs requiring subdivision, shown in the top half of the figure, new polygons are created around all postcode centroids for which there are associated workplace populations. WZ design would then proceed by iteratively recombining the building blocks so as to produce a new set of sub-OA WZs, each of which meets any necessary design constraints. Threshold workplace population counts are the most important of these design constraints but others which could be used include compactness of geographical shape or homogeneity of some workplace-related statistics.

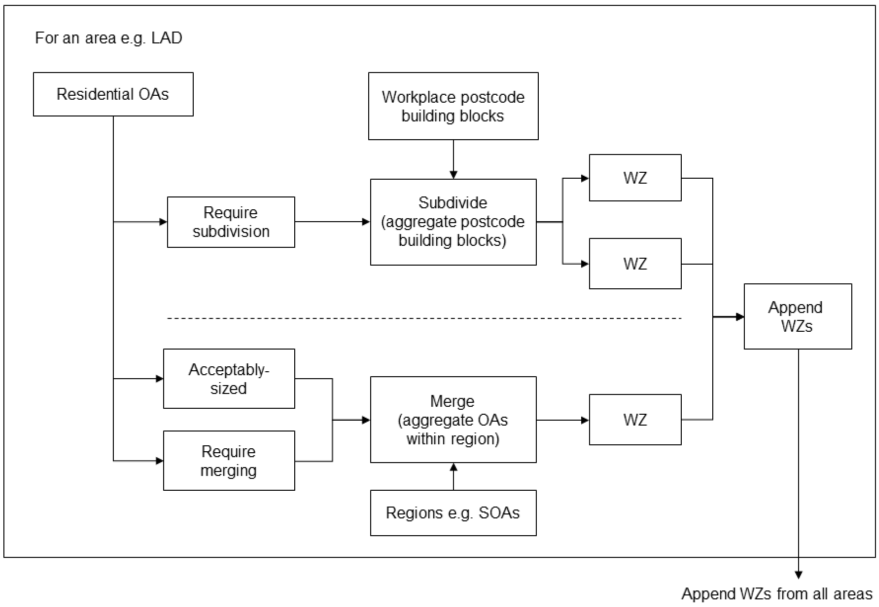

For an area e.g. LAD
Residential OAs
Workplace postcode building blocks
Require subdivision
Subdivide (aggregate postcode building blocks)
WZ
WZ
Append WZs
Acceptably-sized
Require merging
Merge (aggregate OAs within region)
WZ
Regions e.g. SOAs
Append WZs from all areas

Fig. 2. Proposed processing sequence for creation of workplace zones

OAs requiring merging are shown in the lower half of the figure. These would be aggregated with neighbouring sub-threshold or acceptably-sized OAs: the boundaries of these WZs would therefore be a sub-set of the existing OAs. OAs not breaching any thresholds require no further processing and their WZ boundaries would be the same as their OAs, unless they are required to be merged with an adjacent sub-threshold OA. For mergers, there is an additional option to constrain any aggregations within a higher level geographical unit, such as an LSOA. Once an entire set of boundaries has been produced (lower right in the figure), the required workplace statistics can be tabulated for the newly created WZs.

It has already been noted that statistical disclosure control considerations are central to the challenge of designing WZs. Disclosure concerns account for the relative lack of workplace data available for 2001 OAs and it is therefore essential that any alternative geographical units overcome these difficulties. Specific decisions for WZs must be part of broader ONS statistical disclosure control and census outputs policies. The relevant legislation is the Statistics and Registration Service Act 2007, which sets out an obligation to protect the confidentiality of all ‘personal’ information, explicitly stating (s. 39) that this extends to bodies corporate. Thus there is a concern not to publish data which reveal the workforce characteristics of an identifiable business. Businesses are not explicitly counted by the census and do not appear in the census database, hence ONS faces the intriguing problem of preserving the confidentiality of businesses which cannot themselves be identified with certainty from the collected data. While it is possible to treat workplace population as ‘equivalent’ to residents and set a simple threshold for the workplace population of a WZ, the treatment of businesses is more problematic. Businesses may be multi-site enterprises with or without remote or mobile workers and there is no reason to expect that employer names and workplace addresses provided on census forms will equate in any simple way with legal entities. It is clear that a threshold of 40 businesses (as used for households) could not be used, both because they cannot be identified with sufficient confidence and because most OAs contain far fewer than 40 businesses. Postcodes cannot be treated as simple proxies for businesses as there are known to be many-to-many relationships between them. Consideration has been given to the use of other government lists such as the Inter-Departmental Business Register (IDBR) (ONS, 2009) to obtain an independent estimate of the number of businesses in a given WZ and to use this, rather than the uncertain count arising from the census data themselves. However, it is not possible to match workplaces (as reported by census respondents) with the IDBR to the level required to confidently apply protection measures to the census data. The inherent ambiguity of business addresses and the difficulty of definitively listing these is internationally comparable (e.g. Statistics New Zealand, 2007).

A novel approach to the protection of business data has been devised for this research. The smallest unit to which workers can be aggregated in the census database is the postcode. Postcodes are often (but by no means always) allocated to individual businesses, particularly those which receive significant volumes of mail. Hence if workforce data were to be published at the individual postcode level, this would inadvertently generate data for some identifiable businesses. Instead of seeking to identify and then protect businesses explicitly, an alternative approach is to ensure that any postcode (which might equate to an identifiable business) is itself combined with an additional workplace population sufficiently large to obscure the exact characteristics of each postcode. Further, no WZ should contain less than a specified number of postcodes. Some very large workplaces will already greatly exceed any specified population threshold and consideration could be given to identifying and subdividing these prior to processing, although this option has not been implemented here.

In the following sections we demonstrate the implementation of this approach using automated zone design. Although setting of business-related thresholds is a critical issue to the overall task of WZ production, it is a policy decision for the particular national statistical organization involved and does not directly affect the proposed methodology, the chosen values being simply entered as program parameters.

4. Data and implementation

Five local authorities (LAs) were selected for this study: City of London (00AA), Tower Hamlets (00BG), Nottingham (00FY), Southampton (00MS) and Suffolk Coastal (42UG). This represented a substantial dataset requiring a microdata extract of 799,930 census records with subsequent analysis undertaken by approved researchers on a standalone PC in a secure setting on ONS premises. The production of a complete 'real' census geography boundary set is a major processing task for the national statistical organization. LAs were selected which had already been included in other census rehearsal, address testing or pilot projects and for which much valuable contextual information was therefore available. They cover a range of area types, workforce densities and configurations and were agreed in consultation with a user group. Microdata Release Panel clearance was obtained to extract individual-level records from the 2001 census database covering the entire workplace populations of these LAs (i.e. all persons that worked within those districts who were resident in England and Wales, some of whom would have been resident in another LA). The individual records were essential to facilitate re-aggregation to alternative geographical areas, many of which will be smaller than those for which data were published in 2001. Some of our effort has been devoted to overcoming artifacts of 2001 data processing which will no longer apply under 2011 census arrangements.

The extracts included workplace postcodes and OA codes, residential OA codes, employment status indicator, imputation indicators, occupation, industry, number of employees and some imputed workplace address data. The records covered those with a postcoded workplace, those working mainly from home and those with no fixed place of work. Those working mainly from home or with no fixed place of work had been assigned to their residential OAs by ONS but without a specific workplace postcode being allocated. These data correspond with the ONS 'workplace' population but not the ONS 'daytime' population definition (ONS, 2004b), the latter also including (for example) retired and unemployed persons. Using an ONS census form ID-to-postcode lookup table, the correct residential postcode was allocated to each record relating to a person who worked at home. No equivalent postcodes were available for persons with no fixed place of work: these were therefore randomly allocated to a postcode within the OA to which they had been assigned in the 2001 census outputs. This has been done for consistency with 2001 processing, reflecting the fact that these individuals were recorded as not working at home, but meeting the essential requirement that each workplace record carried a valid postcode within the correct OA. A summary is provided in Table 1. Although there are uncertainties associated with the georeferencing and imputation of some of these postcodes there is no additional evidence available within the census database on which to base any further evaluation or correction.

Table 1. Summary of workplace data characteristics for test local authorities (LAs)

LA code	LA Name	Total currently working	Of which		
			Imputed workplace data	Work mainly from home[1]	No fixed place of work[2]
00AA	City of London	312178	27610	432	124
00BG	Tower Hamlets	157162	19830	5658	3260
00FY	Nottingham	172274	19495	6785	3600
00MS	Southampton	111041	9958	6511	4609
42UG	Suffolk Coastal	48005	3870	5937	2487

[1] Allocated to residential postcode via lookup

[2] Allocated to random postcode within output area (OA) assigned by ONS

The next requirement was to create a set of postcode polygons for the workplace postcodes within OAs which required subdivision. The postcode polygon layer used to create the 2001 OAs was not helpful in this context due to the absence of polygons for non-residential postcodes and information about postcodes sharing the same map coordinates. A new set of polygons was therefore created for workplace postcodes, constrained within the residential 2001 OAs.

Although the software used by ONS to create the 2001 OAs has been archived, it is no longer functional due to changes to the IT environment. Implementation of 2011 OAs is now being taken forward using AZTool software (derived from AZM software introduced by Martin (2003), subsequently developed by Cockings *et al.* (2011) and freely available to download from http://www.geodata.soton.ac.uk/software/AZTool/). The creation of WZs from OAs is algorithmically analogous to the updating of 2001 OAs to 2011 OAs and can therefore be implemented using the same software.

For creation of 2011 OAs, ONS have identified residential population thresholds of 100, below which an area should be considered for merging and 625, above which it becomes a candidate for subdivision. The same thresholds can be applied to workplace population sizes to determine which OAs need merging or subdividing to create WZs, thereby treating workers as equivalent to residents. A target workplace population of 250 has been used for the analysis described below. Every postcode should be merged with a minimum of 100 additional workers (equivalent to the person threshold in the residence-based statistics) and there should be no less than three postcodes in a WZ. In keeping with the design of 2001 OAs described in Section 2, geographical compactness and homogeneity constraints were employed. Compactness was measured using the shape statistic perimeter squared divided by area (Maceachren, 1985) and homogeneity by intra-area correlation (Tranmer and Steel, 2001) of major industry types from the ONS Standard Industrial Classification (Hughes, 2008). Experiments were run with and without constraint to higher level geographical units. These values are consistent with current ONS disclosure control plans for 2011 (Cockings *et al.*, 2009; ONS, 2011b).

5. Results

This section presents the resulting WZs created for the five study areas in the sequence outlined in Fig. 2, using the data and constraints described above. Firstly, an illustrative example is provided for the Southampton study area. Secondly, the characteristics of WZs for all five study areas are presented. Finally, particular consideration is given to the statistical disclosure control issues, again with specific reference to Southampton.

Fig. 3. shows the Southampton census OAs which need to be merged or subdivided because their workplace populations are too small or too large. The sub-threshold OAs are generally in residential suburban locations across the city. Over-threshold OAs are located in the city centre (south-central), around the docks (south-west) and occasionally where large workplaces are found in suburban locations. The large OAs in the south-west encompass some areas of tidal water. Some of those requiring no action on the basis of their own workplace population sizes may still need to be merged with adjacent sub-threshold OAs.

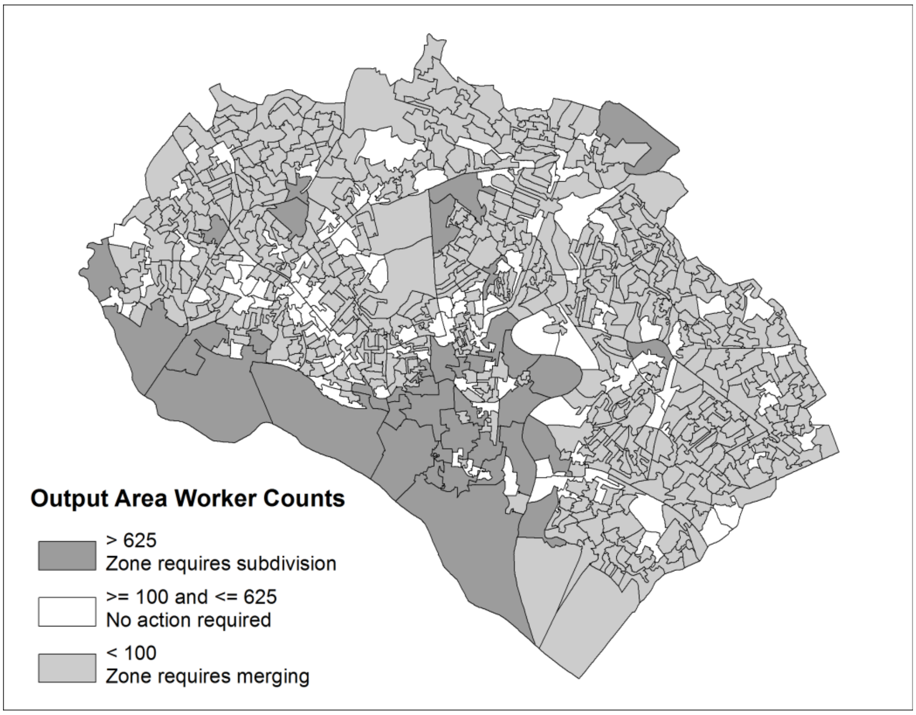


Fig. 3. 2001 output areas requiring merging or subdivision in the City of Southampton

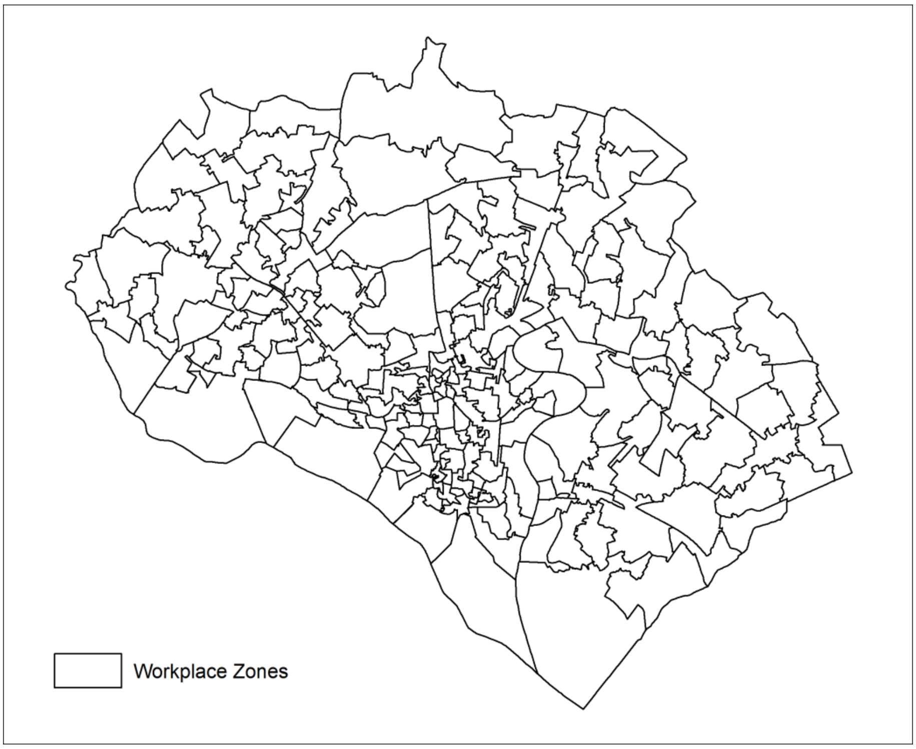


Fig. 4. Workplace zones for the City of Southampton

Fig. 4. shows WZs resulting from application of the complete zone design sequence. As intended, the overall effect is to merge OAs in suburban areas and to split OAs in commercial areas. Some difficulties arise with over-threshold OAs dominated by a large employer because there are no practical means of subdividing these into smaller WZs which would themselves be above the required thresholds. Those OAs which are already of an acceptable size and can be used directly as WZs generally cover suburban shopping streets and business districts.

Table 2 shows key characteristics of the workplace populations for all five study areas, aggregated by the original OAs. The columns show, from left to right, the number of geographical units, summary statistics of workplace population, the intra-area correlation of industry, for which higher values indicate greater homogeneity (only one value is calculated for a set of zones), the shape statistic (smaller values indicate greater compactness) and the numbers of units under, within and over the WZ threshold values. The statistics reveal the wide variation in OA workforce populations in all areas and particularly the enormous concentrations which appear in City of London and Tower Hamlets (which includes the Canary Wharf development in London's docklands). City of London also contains very few OAs due to its small residential population. Comparison with Fig. 1 indicates that these two study areas are extreme in the national context. All five areas include OAs with very small or zero workplace populations due to the geographical separation of residential and commercial premises. OAs across all areas exhibit low homogeneity of industry and broadly similar shape statistics. The maximum OA workplace population is slightly lower than that from published table UV75, most likely due to differences between our detailed processing of workers without clearly identifiable workplace locations and the procedures employed at the time of the 2001 census.

Table 3 presents comparable information to Table 2, this time for the input building block set, excluding the last three columns (thresholds are not relevant at this level). Across all LAs, there are 2.4 times as many building blocks as OAs, but for the City of London this rises to 41.5 due to the concentration of workplace population relative to residential population, generating very high numbers of postcodes. Even at this level, some building blocks still contain very large workplace populations due to the presence of large employers. As might be expected, the smaller building blocks are on average more homogeneous and more compact than the larger OAs.

Table 4 follows a similar structure to Tables 2 and 3, this time for the newly created WZs. The two rightmost columns show the number of cases in which the automated zone design software was not able to fully meet the specified criteria using the parameter values set. The first group of rows shows results for all WZs in all study areas, separately identifying those resulting from splits and mergers. The second set of rows shows the summary results for each of the five study areas.

Across all study areas, only 20 OAs could not be successfully subdivided or merged. These situations arose primarily where isolated sub-threshold OAs were surrounded by over-threshold OAs and therefore had no suitable neighbours with which to merge, or where postcode building blocks were dominated by isolated individual large businesses thereby preventing subdivision of the OA into acceptably-sized WZs.

Table 2. Summary of workplace population characteristics at output area (OA) level for all study areas

Area	Count	Workplace populations				Industry	Shape	Relative to thresholds		
		Min	Max	Mean	SD	IAC	Mean	Under	Within	Over
All study areas	2737	0	79956	292.27	2278.81	0.21	37.11	2078	518	141
City of London	36	8	79956	8653.86	16326.36	0.02	32.54	16	3	17
Tower Hamlets	627	0	24257	250.69	1322.58	0.13	34.03	445	146	36
Nottingham	929	0	16787	185.35	878.96	0.17	38.09	723	168	38
Southampton	730	3	7550	152.11	539.48	0.22	37.44	586	106	38
Suffolk Coastal	415	2	3768	115.62	288.04	0.24	39.41	308	95	12

Table 3. Summary of workplace population characteristics at building block level for all study areas

Area	Count	Workplace populations				Industry	Shape
		Min	Max	Mean	SD	IAC	Mean
All study areas	6520	0	11300	122.69	336.57	0.36	26.88
City of London	1495	1	3621	208.55	382.31	0.29	18.90
Tower Hamlets	1167	0	11300	134.49	502.95	0.22	27.70
Nottingham	1870	0	6646	92.08	252.41	0.33	28.63
Southampton	1389	1	5568	79.94	220.70	0.36	29.42
Suffolk Coastal	599	1	3581	80.11	189.54	0.27	33.85

Table 4. Summary of workplace population characteristics at workplace zone (WZ) level for all study areas

Area	Count	Workers				Industry	Shape	Failures	
		Min	Max	Mean	SD	IAC	Mean	Splits	Mergers
All study areas									
All WZs	1164	6	15771	687.23	946.98	0.27	30.72	12	8
- resulting from splits	645	200	15771	926.64	1182.62	0.29	24.85	n/a	n/a
- resulting from mergers	452	200	1340	386.30	162.08	0.10	38.90	n/a	n/a
By individual study area									
City of London	336	13	8557	930.43	957.39	0.19	21.25	0	4
Tower Hamlets	210	61	15771	743.32	1497.41	0.16	33.47	2	1
Nottingham	308	38	6716	559.05	660.72	0.23	32.38	2	2
Southampton	205	6	6132	541.64	606.24	0.25	36.57	6	1
Suffolk Coastal	105	121	3768	456.98	433.11	0.21	39.22	2	0

Table 4 clearly shows that the statistical qualities of the WZs are superior to those of the residential-based OAs for the publication of workplace statistics. The maximum and standard deviation of workplace populations have been substantially reduced, indicating much more uniformly-sized areas. In the tabulated solutions only eight WZs are left with workforces below 100. The improvements in the City of London are particularly notable, with a five-fold reduction in the maximum workplace population per WZ compared to per OA. In all areas, except the City of London (an eleven-fold decrease), the mean workplace population has increased: this is a direct consequence of the design requirement to produce acceptably-sized zones for publication of more detailed data. Even after such improvements, the City of London and Tower Hamlets still exhibit considerable heterogeneity in workplace population size due to the presence of large businesses, often in high-rise structures which are not subdivisible: these LAs represent extreme cases in the national distribution.

The WZs derived from subdivision of OAs have different statistical characteristics to those from mergers. The workplace populations of WZs created by subdivisions have much higher maximum, mean and standard deviation but exhibit more homogeneity of industry and compactness of shape (although mean compactness is strongly influenced by the City of London). In areas where merged WZs are created, the building blocks are generally larger but have lower workplace populations: it is therefore possible to produce more uniform sizes but more difficult to achieve compactness of shape. Overall, homogeneity levels are comparable to the original OAs although the intra-area correlation statistics are far lower for WZs resulting from mergers (grouping disparate workplaces) than from splits (increasing local homogeneity). These scores are scale-dependent, hence the results can be compared with others for the same area, but not directly with values for other study areas or the combined set. Some of the split OAs had been those with large geographical extents and sparse residential populations, hence there has been a considerable increase in geographical resolution in these areas. This would be particularly valuable for applications which require geographical locations of trip ends, for example modelling transportation demand, or for those employing area classifications, in which greater detail can be provided.

The risk of disclosure has been further reduced by improvement in the distribution of workplace postcodes per zone, with all but one WZs containing at least three postcodes and the mean number of workplace postcodes being reduced from 75 per OA to 19 per WZ. Again, there is a difference between the mean number of postcodes per WZ derived from subdivision (nine) compared to those from merging (35). The maximum number of postcodes per zone has also been reduced from 611 in one OA in City of London to 139 in a merged WZ in Suffolk Coastal.

User consultation (ONS, 2011b) has shown strong support for WZs to nest within larger geographical units for which non-census business statistics are published, such as LSOAs and middle-layer super output areas (MSOAs) (mean residential population sizes 1512 and 7248 respectively) . This constraint is only relevant when merging, where it influences the number and spatial adjacency of neighbouring zones available. We explored the impact of constraining the WZ creation process firstly to LSOAs and then to MSOAs, although full results are not shown here for conciseness. Constraining to LSOAs was found to be far too restrictive: 833 sub-threshold OAs could not be merged to meet the minimum workplace population threshold. By contrast, applying MSOAs as a constraint was much more successful. Compared to the completely unconstrained results presented in Table 4, only four additional OAs could not be merged because no suitable neighbours were available. The overall detriment to the statistical quality of the WZs was also only very slight.

Turning to the disclosure control implications, for Southampton only, a detailed analysis of the postcodes with the largest numbers of workers was undertaken. These might reasonably be expected to equate to the city's largest – and therefore most readily identifiable – businesses. Disclosure cannot be assessed strictly at the level of individual businesses because records are not matched into business-level data in the census database. The postcodes with the largest numbers of workers were individually assessed and a summary of the top 20, which contain most of the major patterns observed, is presented in Table 5. It is apparent from Table 5 that there are very few individual workplaces which account for the entire workforce of an OA. Where this does occur it relates primarily to the very largest workplaces, particularly if isolated in otherwise residential areas. The wide spread of persons working at home, at no fixed place of work and in small local businesses such as shops serves to obscure the exact population of large businesses in all but a very few cases. Only in the four cases whose postcode workforce count is indicated by a star does a single postcode account for within 100 workers of the total population of its OA. These are the situations in which information about the workplace population of the OA could potentially reveal most about a single workplace.

Table 5. Summary of 20 postcodes (PCs) with largest census workplace populations in Southampton

PC rank	PC workforce	Workforce of output area (OA) containing PC	Type of business to which postcode relates
1	5566	6536	One of four large sites (healthcare-related)
2	2900	3161	Single major employer (education campus)
3	1782	1986	Single major employer (suburban factory)
4	1598	1701	Single major employer (suburban offices)
5	1213*	1262	Single major employer (suburban factory)
6	1058	7663	One of several large employers (city centre)
7	1054*	1097	Single major employer
8	1037*	1120	Single major employer
9	999	1295	One dominant employer (city centre)
10	990	1847	One dominant employer (education)
11	917	7663	One of several large employers (city centre)
12	847	5163	One postcode for many small businesses (retail centre)
13	759	5163	One postcode for many small businesses (retail)
14	742*	760	Single major employer (healthcare-related)
15	730	2704	One dominant employer (docks-related)
16	634	2306	One of several large employers (city centre)
17	644	1714	One of several large employers (business park)
18	638	3333	One of several large employers (business park)
19	637	7663	One of several large employers (city centre)
20	634	673	One postcode for many businesses (city centre)

* Postcode accounts for within 100 workers of the total workforce for the OA and relates to a single large employer

Turning to the attributes associated with workplaces, our analysis shows that for each major employer represented by a unique single-business postcode, census responses contained a wide variety of different industry codes and the full range of possible responses for numbers of employees (from 0-9 to over 500). It would therefore not be possible to use either of these fields to directly identify individual businesses. Our interpretation is that employees tend to record the business of their employer in relation to their own job (e.g. ‘higher education’, ‘research and development’, ‘hospital activities’, ‘other human health activities’ all being recorded in large numbers at a teaching hospital) and similarly that they tend to interpret and report the number of employees in relation to their own unit rather than that of the entire enterprise. These are among the variables identified by the 2001 census quality report as having the highest level of non-response, with number of employees being the highest at 13.92%, particularly affected by respondents not being sure who to include in their answer (ONS, 2005).

6. Conclusions

This paper has highlighted particular problems which follow from the publication of workplace statistics through the medium of small geographical units based on residential location, currently a standard practice among many national statistical organizations. Some attempts have been made, for example in Australia and the USA, to produce alternative zones for workplace data but these are not at the smallest geographical scales, nor are they fully integrated with the population statistics systems. Taking the specific example of 2001 census OAs in England and Wales, the enormously skewed distribution of workplace locations can be seen as a major obstacle to important research and operational questions about workplace populations, as key information is restricted or its publication prevented. Underlying this issue is the fundamentally different spatial distribution of workplaces and residences. In response to this problem, an original solution has been proposed, using an automated zone design approach for the creation of WZs. These fit neatly alongside the existing hierarchy of residential OAs, merging or splitting areas to create a more uniform distribution of workforce sizes appropriate to the publication of more statistically detailed, non-disclosive data about workforces and places of work. Using confidential 2001 census records in a secure setting, it has been possible to create prototype WZs for five study areas, re-aggregate the original data to these new zones and examine their characteristics. The work has employed AZTool software written by the authors and available to other researchers.

As a result of the work described here, WZs and associated workplace population data will be published as 2011 census outputs (ONS, 2012). The final disclosure control policies to be applied have not yet been published. A key consideration in the design of WZs is the protection of individual workers and businesses from inadvertent disclosure in the published data. This is readily achieved for workers in much the same way as for residential populations by setting threshold values, but is far more challenging with regard to businesses which are not explicitly recorded in the census database. We have explored the relationship between individual businesses and geographical reporting units and propose a solution whereby postcodes are protected by the addition of further workers and postcodes to provide the required levels of uncertainty. The solution is readily implemented within the automated zone design approach. There will be richer record-level data available within the statistical organizations in 2011 than was the case in 2001 and more carefully designed processes to handle incomplete workplace addresses and workers without fixed workplaces, another challenge which appears to be internationally applicable.

The prototype areas display much improved statistical properties, with more uniform workforce sizes, less extreme values and compliance by design with the specified threshold values. It is possible to constrain the entire process within higher-level areal units such as MSOAs while still achieving acceptable results in terms of compactness, homogeneity and workplace population sizes. There are a small number of WZs which cannot be automatically resolved using the parameters evaluated here, either because no suitable neighbouring zones are available for merging or their constituent postcodes are inappropriately configured. These would need to be flagged for clerical intervention in order to permit specific mergers between areas or the relaxation of other design criteria to meet the absolute requirements of statistical disclosure control.

The creation of WZs does not predetermine or prevent any particular set of statistical tables being published but rather provides a more resilient workplace population distribution, permitting a wider range of tabulations than has been possible with previous geographical units based on resident counts. The production of WZs will not prevent the statistical organizations from producing alternative population counts for OAs such as 'daytime' populations, based on those working or judged to be remaining in an area during the working day. It will also be possible to generate flow data for travel to work based on WZs or to create travel to work areas and other functional geographies using WZs as the basic building blocks. A consideration, largely independent of the zone design methodology per se, is the extent to which geographical boundaries of WZs are constrained to follow real-world features such as roads and buildings.

Our approach to this internationally relevant challenge has the potential to facilitate considerably enhanced census workplace data. The methodology and tools are generic and serve to provide a further example of the practical application of automated zone design to a pervasive problem in spatial population data handling.

Acknowledgments

The authors gratefully acknowledge practical and financial support provided by the Office for National Statistics and Department for Transport and the helpful comments of two anonymous referees and the editors. All data access was provided directly by ONS. The analysis and views expressed here are solely those of the authors and are not intended to represent any aspect of data publication policy.

References

Adams, J. S., VanDrasek, B. J. and Phillips, E. G. (1999) Metropolitan area definition in the United States *Urban Geography* **20**, 695-726

Alvanides, S., Openshaw, S. and Duke-Williams, O. (2000) Designing zoning systems for flow data In *GIS and Geocomputation* (Eds P. Atkinson and D. Martin) pp. 115-134 London: Taylor and Francis

Ang, L. and Ralphs, M. (2008) *Operations research for new geographies: zone design tools for census output geographies.* Statistics New Zealand: Wellington.

Australian Bureau of Statistics (2011) Census Dictionary, 2011. Australian Bureau of Statistics, Canberra (Available from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2901.0Chapter29202011)

Australian Bureau of Statistics (2012) Australian Statistical Geography Standard (ASGS). Australian Bureau of Statistics, Canberra (Available from http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS))

Bhaduri, B. (2008) Population distribution during the day. In *Encyclopedia of GIS* (Eds S. Shekhar and H.Xiong) pp. 880-885 New York: Springer

Boyle, P. and Dorling, D. (2004) Editorial: the 2001 UK census: remarkable resource or bygone legacy of the 'pencil and paper era'? *Area* **36**, 101-110

Cabinet Office (2008) *Helping to shape Tomorrow: The 2011 Census of Population and Housing in England and Wales.* Office of Public Sector Information, London (Available from http://www.official-documents.gov.uk/document/cm75/7513/7513.pdf)

Cockings, S., Harfoot, A. and 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

Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2011) Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 census output geographies for England and Wales *Environment and Planning A* **43**, 2399-2418

Coombes, M., Green, A. E. and Openshaw, S. (1986) An efficient algorithm to generate official statistical reporting areas: the case of the 1984 travel-to-work areas revision in Britain *Journal of the Operational Research Society* **37**, 943-953

Duke-Williams, O. and Rees, P.H. (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

Federal Highway Administration (2010) TAZ Delineation Business Rules (March 2010). Federal Highway Administration, Washington DC (Available from http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/tazddbrules.cfm)

Hughes, J. C. (2008) SIC 2007: implementation in ONS *Economic and Labour Market Review* **2** (8), 41-44

Maceachren, A. M. (1985) Compactness of geographic shape: comparison and evaluation of measures *Geografiska Annaler B*, **67**, 53-67

Martin, D. (1998) 2001 Census output areas: from concept to prototype *Population Trends* **94**, 19-24

Martin, D. (2003) Extending the automated zoning procedure to reconcile incompatible zoning systems *International Journal of Geographical Information Science* **17**, 181-196

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

Martinez, L., Viegas, J. and Silva, E. (2009) A traffic analysis zone definition: a new methodology and algorithm *Transportation* **36**, 581-599

Office for National Statistics (2004a) Census geography evaluation report. Office for National Statistics, Titchfield (Available from http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/review-and-evaluation/evaluation-reports/geography/evaluation-report.pdf)

Office for National Statistics (2004b) Census 2001 Definitions. The Stationery Office, London (Available from http://www.ons.gov.uk/ons/guide-method/census/census-2001/data-and-products/data-and-product-catalogue/reports/definitions-volume/index.html)

Office for National Statistics (2005) Census 2001 quality report for England and Wales. Palgrave Macmillan, Basingstoke. (Available from http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/review-and-evaluation/evaluation-reports/quality-report/census-2001-quality-report.pdf)

Office for National Statistics (2009) About the IDBR. Office for National Statistics, London (Available from http://www.ons.gov.uk/ons/about-ons/who-we-are/services/idbr/about-the-idbr/index.html)

Office for National Statistics (2010) 2011 Census Output Geography Consultation (England and Wales) Report and Recommendations. Office for National Statistics, Titchfield (Available from http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-consultations/closed-census-consultations/2011-census-geography-outputs-consultation/report-and-recommendations.pdf)

Office for National Statistics (2011a) Beyond 2011: Consultation on User Requirements. Office for National Statistics, London (Available from http://www.ons.gov.uk/ons/about-ons/consultations/open-consultations/beyond-2011---public-consultation/beyond-2011-user-needs-consultation.doc)

Office for National Statistics (2011b) 2011 Census - England and Wales output geography: policy and products. Office for National Statistics, Titchfield (Available from http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-consultations/closed-census-consultations/2011-census-geography-outputs-consultation/2011-census-geography-outputs-consultation-document.doc)

Office for National Statistics (2012) The 2011 Census Prospectus (Available from http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/index.html)

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 – 446

Statistics Canada (2006) Journey to work reference guide, 2006 census. Statistics Canada, Ottawa (Available from http://www12.statcan.gc.ca/census-recensement/2006/ref/rp-guides/journey-to-work-deplacement-domicile-travail-eng.cfm)

Statistics Canada (2008a) 2006 Census Dictionary. Statistics Canada, Ottawa (Available from http://www12.statcan.ca/census-recensement/2006/ref/dict/index-eng.cfm)

Statistics Finland (2004) *Use of registers and administrative data sources for statistical purposes: best practices of Statistics Finland Handbook 45*. Statistics Finland, Helsinki (Available from http://tilastokeskus.fi/tup/julkaisut/kasikirjoja_45_en.pdf)

Statistics New Zealand (2006) Data Dictionary: 2006 Census of Population and Dwellings. Statistics New Zealand, Wellington (Available from http://www.stats.govt.nz/~/media/Statistics/Census/2006-reports/data-dictionary.pdf)

Statistics New Zealand (2007) 2006 Census information by variable: workplace address. Statistics New Zealand, Wellington (Available from http://www.stats.govt.nz/Census/about-2006-census/information-by-variable/workplace-address.aspx)

Tranmer, M. and Steel, D. (2001) Using local census data to investigate scale effects. In *Modelling scale in geographical information science* (Eds N. J. Tate and P. M. Atkinson) pp. 105-122 Chichester: Wiley

US Census Bureau (2009) *A compass for understanding and using American Community Survey Data: what researchers need to know*. US Government Printing Office, Washington DC (Available from http://www.census.gov/acs/www/Downloads/handbooks/ACSResearch.pdf)

4.3 Paper 5: Getting the foundations right: spatial building blocks for official population statistics

Cockings S, Harfoot A, Martin D and Hornby D (In Press) “Getting the foundations right: spatial building blocks for official population statistics”

The version of this article included in this thesis is the author’s own post-print copy. The definitive, peer-reviewed and edited version of this article is in Press for *Environment and Planning A*.

Getting the foundations right: spatial building blocks for official population statistics

Samantha Cockings*, Andrew Harfoot, David Martin, Duncan Hornby

Geography and Environment, University of Southampton, Southampton, SO17 1BJ, UK
* Corresponding author: Tel: +44 (0)23 8059 5519 Fax: +44 (0)23 8059 3295
E-mail: s.cockings@soton.ac.uk; ajph@geodata.soton.ac.uk; d.j.martin@soton.ac.uk; ddh@geodata.soton.ac.uk

Abstract. When publishing official population statistics, many national statistical organisations define small areas termed 'building blocks' which are then aggregated into larger 'output zones' for data release. While output zones are known to have enormous influence on spatial analysis, there has not been any systematic analysis of the effect of building blocks on characteristics of output zones. This paper evaluates current international practice in building block design, identifying key conceptual and practical issues. Using the example of six local authorities in England and Wales, it employs automated zone design to evaluate the influence of two sets of building blocks (street blocks and postcodes) on output zone characteristics. Household level census data, accessed under secure conditions, are used to evaluate the impact on both newly designed and maintained output zones. Postcodes are shown to be more effective building blocks than street blocks, providing more uniform population and household sizes, greater precision for matching postcoded data to census data, and enabling more zones to be maintained. However, street blocks deliver more compact output zones and greater internal homogeneity of tenure and accommodation type. The scale effect of the modifiable areal unit problem and the specific geographical patterning of variables are both shown to be important factors when designing building blocks. These findings have directly informed policies and processes for the 2011 census in England and Wales and provide useful conceptual and practical guidance for any national statistical organisation or analyst designing their own building blocks. The paper concludes that some aspects of international building block design practice could be more effectively harmonised but that such design should always be nationally-specific to incorporate locally-varying conceptual and practical issues. Further research should extend this analysis to other building block types, notably grid squares.

Keywords: Building blocks, basic spatial units, automated zone design, census, modifiable areal unit problem, output zones, output areas

1. Introduction

Many national statistical organisations define zoning systems for the release of aggregate population statistics. In particular, zones defined originally for census purposes are frequently used for the publication and integration of other data. These ‘output zones’ form the foundation for mapping and spatial analysis of an enormous range of social and economic data and consequently their design is of considerable significance. When designing zoning systems, many statistical organisations start by defining a set of small areas or ‘building blocks’. In some instances, these building blocks are themselves used for publication of data but, more frequently, they are aggregated into larger zones for full data release. Appropriately designed building blocks are essential if zoning systems are to have wide utility. This paper explores the effect that building block design has on the characteristics of output zones.

There has recently been renewed international interest in zone design, partly driven by the latest rounds of censuses worldwide (Martin, 2006; Valente, 2010) and a desire to define standard zoning systems for the publication of population data (Cook, 2004; ABS, 2011), but also by investigations into alternative models for producing national population statistics without a traditional census (House of Commons Treasury Committee, 2008). Motivations for creating such systems vary, but often include demands for greater value from public investments and requirements to collect data once but re-use it for different purposes (Cabinet Office, 2011; Cook, 2004). From research and planning perspectives, the use of standard zoning systems potentially aids integration with other datasets and stability through time if they are maintained.

Despite shared international interest in the design of zoning systems, there is considerable variation in the building blocks employed by national statistical organisations. The most common are: postal, or zip, codes (hereafter termed postcodes); blocks based on locations of streets and/or other topographical features; and grid squares. Some countries have recently invested significantly in creating sets of nationally-consistent building blocks (such as mesh blocks in Australia) and are hoping to use these as a stable basis for their zoning systems for many years to come (ABS, 2011).

Within the academic literature, the design of building blocks for the publication and analysis of population data has previously been couched in terms of 'basic spatial units'. Openshaw (1990) and Martin and Higgs (2004) debate the appropriateness of different basic spatial units for representing populations, while the Chorley Report (DoE, 1987, p.121) recommends addresses or postcodes as the "preferred bases for holding and/or releasing socioeconomic data" in the UK. Openshaw (1984) and others (such as Flowerdew, 2011; Holt et al, 1996; Manley et al, 2006; Openshaw and Taylor, 1979) demonstrate how the design of areal units (both in terms of scale and boundary placement) can influence spatial patterns – a phenomenon termed the modifiable areal unit problem (MAUP). Openshaw (1984, p21) also hypothesises that "the aggregational variability … due to the choice of the first zoning system will exceed that of any subsequent re-aggregations of the data", implying that the design of building blocks may be critical. And yet, the influence of building block design on output zones has not been investigated empirically before. Automated zone design (Cockings and Martin, 2005; Martin, 2003) provides a means to quickly and easily recombine zones according to specified criteria, thus enabling the systematic evaluation of alternative combinations, but it has not previously been utilised to explore the impact of building blocks on output zones.

This paper explores the impact of building block design on the statistical characteristics of output zones, on the ability to link other datasets to output zones, and on the revision of output zones over time. It employs automated zone design methods and 2007 household level data for a selection of local authorities in England and Wales, analysed under secure conditions within the national statistical organisation. The combination of a rarely available dataset and the novel extension of automated zone design techniques afford unique insights into a specific aspect of MAUP, namely the impact of the first aggregation of data. Such investigations are particularly timely given the current investment in standard zoning systems by national statistical organisations.

The rest of the paper is organised as follows: section 2 identifies the primary conceptual and practical considerations in building block design, with reference to current practice in different national contexts; section 3 describes the background, methods and data for the empirical evaluation; section 4 presents the results; finally, section 5 discusses the implications for international practice in building block design.

2. Current international practice in building block design

The ‘designer’ of a zoning system for official population statistics will usually be a national statistical organisation or similar body charged with the collection, collation and publication of data. Practical considerations often influence, and frequently outweigh, more conceptual aspects of building block design.

2.1 Basic conceptual considerations

The first requirement is a clear understanding of the purpose(s) for which building blocks will be used, including whether they will be used directly for data release or the basic spatial units from which output zones will be created. Designers need also to consider whether they will be used solely for the publication of population data or also for other datasets. For example, the ‘super output areas’ (SOAs) (ONS, 2012b) created from a building block base of postcodes in England and Wales were deliberately designed to integrate a wide range of data such as population, education, crime, health and quality of the physical environment (http://www.neighbourhood.statistics.gov.uk). Their postcode base provides a geographical reference common to many of these datasets.

Output zones are usually employed for purposes such as mapping and analysis. This can influence the design criteria for building blocks: zones with compact shapes whose boundaries follow recognisable features on the ground are often desirable for mapping purposes; whereas homogeneity of population size is often preferable for statistical analysis. Users inevitably tend to present a range of ultimately irreconcilable requirements which must be traded off by the designer.

“The question is simply what objects at what scales do we wish to investigate?” (Openshaw and Taylor, 1979, p143). Ideally, building blocks should directly represent, or be capable of aggregation to match, the spatial objects of interest. Objects frequently represented in national zoning systems are groups of people or areas of land which are in some way distinctive from their neighbours. Conceptually, these demand two distinct sets of building blocks. When modelling people, building blocks should represent the locations of people (often achieved through indirect georeferencing of residential properties) whereas for small areas, features of the built and social environment which make each location distinct (such as housing type or tenure) are more commonly employed. There is a significant literature concerning the definition of "neighbourhoods" (particularly in the fields of health geography and urban studies) which is of relevance here. This literature discusses the conceptual basis for defining neighbourhoods (e.g. Galster, 2001) and evaluates the impact that different zone designs can have on results (Flowerdew et al, 2008; Haynes et al, 2008). Importantly though, none of these studies address the issue of how building blocks should be designed. The design of building blocks for the output of national population statistics poses particular challenges: such zoning systems must meet essential statistical requirements, such as confidentiality thresholds, whilst at the same time satisfying the needs of a wide range of users with disparate needs. As such, they inevitably have to balance a number of competing criteria.

The appropriateness of different types of building blocks for representing specific objects varies by national context. For example, the UK postcode hierarchy is a coding system designed primarily for the delivery of mail, the smallest element of which is a 'unit postcode' (hereafter termed postcode) which typically contains, on average, 17 addresses. Postcodes are created, terminated or re-used in order to manage the workload of postal delivery workers as new addresses are constructed or old ones demolished. Coincidentally, due to residential clustering at the small area level, postcodes also tend to exhibit reasonably strong internal homogeneity of the socio-economic and built environments, making them extremely useful for analysis, linkage and representation of population data. By contrast, in countries such as the USA, Australia and New Zealand, postcodes are larger and more heterogeneous. In the USA, which exhibits gridded street patterns in many areas, blocks based on streets and other geographical features tend to more effectively capture the variation in housing type, socio-economic status of residents and characteristics of the built environment. Table 1 summarises the characteristics of three key building block types employed internationally.

Table 1 Characteristics of building blocks commonly employed internationally

Building block type	Country	Country-specific name	Scale/size (Year)	Method of creation	Design characteristics	Relationship to key output zones
Postcode	England & Wales	Unit postcode	Average 17 delivery points (2001)	Automated	Synthetic postcode polygons. Aggregations of address polygons. Aligned with topographical features where possible. Nested within administrative boundaries (electoral wards and civil parishes, where they exist).	Aggregate to census output zones (output areas, super output areas)
	Northern Ireland	Unit postcode	Average 17 delivery points (2001)	Automated	Synthetic postcode polygons. Aligned with topographical features where possible. Nested within administrative boundaries (electoral wards).	Aggregate to census output zones (output areas, super output areas)
	Scotland	Unit postcode	Average 15 delivery points (2001)	Manual	Digitised postcode polygons.	Aggregate to census output zones (output areas, postcode sectors, data zones)
Street block	Australia	Mesh block	Average 30-60 dwellings (2011)	Hybrid	Hierarchical design criteria: initial urban/rural split then uniformity of dwelling estimates and land-use key drivers. Based on cadastral boundaries. Aligned to 2011 Statistical Local Areas but this will not be maintained over time.	Aggregate to output zones in Australian Statistical Geography Standard
	New Zealand	Meshblock	Average 97 people (2006)	Manual	Boundaries follow cadastral boundaries, centre line of roads, rivers and other physical features.	Aggregate to output zones in New Zealand Standard Areas Classification
	USA	Census block	Average 28 people (2010)	Hybrid (mostly automated but some stakeholder input)	Boundaries of higher level geographic areas (e.g. counties, places, voting districts, census tracts, block groups, etc.) must form block boundaries; visible features (streets, roads, streams, and railroad tracks) usually incorporated, depending on pre-determined ranking system based on block size and boundary composition.	Always aggregate to higher level output zones due to method of creation

Grid squares	Denmark	National square grid (100m)	Average 6 households (2003)	Automated	100m grid squares covering whole country	Aggregate to larger standard grids or groups of cells meeting Statistics Denmark’s disclosure requirements
	Finland	Grid cells (250m)	Mean 16 people (2010)	Automated	250m grid squares covering whole country	Aggregate to 1km grid but not to other output zones (postal codes, municipal sub areas, municipalities)
	Northern Ireland	100m grid	Minimum 25 persons, 8 households (2001)	Automated	Since 2001, 100m grid squares available for whole country; previously, 100m for urban areas, 1km elsewhere	100m grids aggregate to 1km grid; 1km grid consistent since 1971. Neither are consistent with other census output zones (output areas) – see above

Sources: ABS (2011); CDU (2012); ONS (2012b); NISRA (2008); Sommer (2003); Statistics Finland (2012); Statistics New Zealand (1992, 2012); US Census Bureau (2010, 2011, 2012a, 2012b) and pers. comm.

Further conceptual considerations include whether building blocks should be space-filling (i.e. covering the entire land surface) and whether they should all be populated. This is often determined by their intended use. For example, if they will be employed for calculating densities it makes sense for them to be space-filling. All of the countries in Table 1 use space-filling building blocks. Most countries do not publish the full range of population data for their building blocks due to confidentiality requirements. Some countries, such as Australia and the UK, publish only key statistical data (usually counts of total population and households); others (e.g. New Zealand, Finland, USA) publish a broader, but still limited, range. In all cases, statistical disclosure control requirements are strictly enforced. Policies also vary in whether it is permissible to publish zero counts for building blocks. This may be driven by national legislation and attitudes towards privacy but also influenced by the spatial distribution of population. Countries with large expanses of unpopulated land tend to allow their building blocks to be unpopulated as this provides a more appropriate base for mapping and analysis. 47% of Finnish 250m^2 grid cells (Statistics Finland, 2012) and over 1 million US census blocks (out of a total 11.1 million) (pers. comm.) were unpopulated in 2010.

The scale effect of MAUP (Openshaw, 1984) implies that the size of potential building blocks (relative to the objects being represented and output zones) should be considered prior to their creation (Openshaw and Taylor, 1979). Shuttleworth et al (2011) suggest that as long as the phenomenon of interest is structured over larger areas than the output zones, results of analyses will not be sensitive to the scale of the zones. This suggests that building blocks should be smaller than the scale at which the phenomenon varies. As it is not possible for users to disaggregate building blocks, any spatial patterns not captured at the scale of the building blocks will not be identifiable. Whilst very small building blocks may therefore appear to be a prudent choice (Shuttleworth et al 2011), building blocks with some degree of structure of key variables at an appropriate scale may provide a more effective base. Such inherent structure provides a useful starting point for the zone design process and can lead to more homogeneous zones. Again, national context is important. Postcodes in the UK are relatively small units making them ideal building blocks relative to the larger output zones which are used to publish population data. In the USA, street blocks are more appropriate, although the size and characteristics of blocks varies significantly across the nation (US Census Bureau, 2012a). Table 1 provides an indication of the scale of building blocks employed in different countries, expressed in terms of average population, household or dwelling counts.

2.2 Practical considerations: opportunities and constraints

The way in which real-world features have historically been conceptualised and modelled in national geospatial infrastructures also influences which building blocks can be employed. For example, if population and other data are routinely georeferenced by street address ranges it may be attractive to base building blocks on these (as with mesh blocks in New Zealand). The UK has a long history of using postcodes whereas the USA has employed blocks in its census operations since the 1920s. As more countries move towards address-based georeferencing of population datasets, practical constraints on building block and output zone design should reduce as address points can be aggregated into virtually any building block.

In many cases, there are legal or administrative requirements for building blocks to nest within specific higher level zones (see Table 1). This is usually to enable exact statistics to be compiled for geographical units used in elections or public resource allocation. The 2001 postcode building blocks in England and Wales were nested within the administrative boundaries of electoral wards and (where relevant) civil parishes (Martin et al, 2001) while Australian mesh blocks are aligned with 2011 Statistical Local Areas (SLAs) (ABS, 2011). Such constraints can significantly affect the design of building blocks and often conflict with other requirements. An additional difficulty is that these administrative units change on a regular basis, making it particularly challenging to keep building block or output zone boundaries aligned with them. As a result, some countries have decided to prioritise stability through time over consistency with administrative units. Australia, for example, will not maintain the alignment of mesh blocks with SLAs beyond 2011 and in 2011 England and Wales removed the requirement for building blocks and output zones to be nested within electoral wards.

In some situations, ideal building blocks cannot be created due to lack of suitable input data or because licensing restrictions prevent their use. This was the case in England and Wales in 2001, where embedding commercially licensed data on topographic features into the postcode building blocks would have prevented unrestricted distribution of subsequently created output zone boundaries. As more countries embrace open data initiatives (Cabinet Office, 2011) such restrictions may recede.

Building blocks may be created using manual or automated methods, depending on feasibility of automation and availability of resources (digital data, skilled personnel, hardware and software). In reality, most statistical organisations employ a hybrid approach, with some processes being carried out by automated (usually GIS-based) procedures and the rest manually. However, the extent of manual intervention varies, as illustrated in Table 1. The creation of postcode boundaries for 2001 in England and

Wales was almost entirely automated, whereas the construction of mesh blocks in Australia involved more manual intervention. User consultation may also be incorporated (as in the USA), either *a priori* to determine the design criteria or *post hoc* to accept or suggest amendments to proposed building blocks. Similar procedures are usually required for the aggregation of building blocks into output zones, with a comparable range of approaches existing.

2.3 Changes through time: retention, redesign or maintenance

A generic problem with output zones and building blocks is that the spatial distribution of population changes over time, as do conceptual requirements and practical constraints. National statistical organisations must therefore decide whether to (i) *retain* the existing zones/building block, (ii) *redesign* completely new ones or (iii) attempt some hybrid approach of *maintenance* (where zones/building blocks which are still fit for purpose are retained but those which no longer meet requirements are redesigned). These options variously trade off comparability of statistical data over time against representation of the contemporary population distribution.

Automated zone design was employed to create output zones for the 2001 census in England and Wales (Martin et al, 2001; ONS, 2004). Entirely new boundaries were generated for the building blocks (postcodes) and these were then aggregated into output zones (output areas (OAs) and SOAs), with no attempt to retain zones from previous censuses. Ang and Ralphs (2008) explored how automated zone design might be used in New Zealand to create improved output zones. Rather than completely redesigning the entire zoning system, they demonstrated how existing, regularly maintained, building blocks (mesh blocks) might be re-aggregated to create output zones at different scales with more uniform population sizes. More recently, Cockings et al (2011) extended these methods to enable the hybrid maintenance approach described at (iii) above. This approach has now been implemented by the Office for National Statistics (ONS) (the national statistical organisation for England and Wales) to selectively update 2001 census output zones to reflect 2011 population distributions. In this process, new building blocks (postcodes) are created for zones which need to be split; these are then aggregated to create acceptably-sized output zones. The remainder of this paper explores how the design of building blocks influences the characteristics of output zones, whether they be redesigned or maintained.

3. Empirical investigation: building blocks for the 2011 census in England and Wales

3.1 Background

The OAs used for publication of 2001 census data in England and Wales were based on postcode building blocks. Address-based space-filling postcode polygons were aggregated such that OA population and household counts exceeded thresholds of 100 and 40 respectively, with a target household count of 125 (Martin et al, 2001). Since 2001 there has been population change in most areas of the UK, with a minority having experienced such substantial change that new OAs will be required for the 2011 census (Cockings et al, 2009). User consultation (ONS, 2007) indicated a preference that the 2001 zones be retained wherever possible in order to aid stability and consistency. ONS' policy is to undertake a process of selective maintenance, whereby OAs whose populations have grown or declined too much are split or merged while the rest are retained.

When consulted regarding which features any redesign should be based on (ONS, 2007), users expressed support for postcodes, “hard” physical boundaries, administrative zones and “neighbourhoods” but much less for grid squares. ONS’ preference is to retain postcodes as it provides consistency with 2001, but their boundaries often do not align with recognisable features on the ground and sometimes cut through footprints of buildings, due to their lack of formally defined boundaries and synthetic generation. Their ‘boundaries’ also change fairly regularly as new addresses are added or removed. In August 2012, 0.14% of current postcodes were terminated, 0.04% reused and 0.07% newly added (pers comm, ONS). The potential utility of alternative building blocks, particularly “street blocks”, has long been debated in the UK but never formally investigated. The JUG-T project (Cossey et al, 2005) did undertake a localised experiment in Manchester but there remains a need for a more systematic evaluation. The empirical example reported here thus explores the relative impact of two sets of building blocks (street blocks and postcodes) on the statistical characteristics of output zones (OAs), on the ability to link other datasets to OAs, and on the revision of OAs over time.

3.2 Study areas and data

Household-level counts of population, tenure (e.g. rented, owned outright) and accommodation type (e.g. detached, semi-detached) were extracted for 2007 for six local authorities in England and Wales (Camden, Isle of Anglesey, Lancaster, Liverpool, Manchester and Southampton). These areas were selected to be indicative of the types of areas experiencing population change since the 2001 census and also representative of different urban/rural types. Some were also areas used by ONS in census preparations or known to the authors. Full methodological details can be found in Cockings et al (2011).

3.3 Methods

Two sets of building blocks were created: street blocks and postcodes. As per traditional approaches to output zone creation in England and Wales, both were space-filling and contained at least one address.

Although various countries use some form of 'block' (e.g. census blocks in the USA and mesh blocks in Australia and New Zealand), there is no standard definition of the block as an entity and, as shown in Table 2, definitions vary. There are, nevertheless, clear areas of commonality, with most statistical organisations aiming to delineate areas broadly enclosed by streets and other recognisable geographical features. The US Census Bureau provides a useful discussion of the variation in census block characteristics across the USA (US Census Bureau, 2012a). For our purposes, the definition from the UK-based study by Cossey et al (2005, E/S-1) was employed: "an area of land surrounded by streets – or other major linear topographic features such as railway lines or water features". In keeping with this definition, centrelines of public roads (from OS MasterMap Integrated Transport Network, 2007) and railways (OS Meridian, 2008) were intersected with 2001 OA boundaries to produce street blocks, which were then intersected with the household data to identify blocks not containing any addresses. These empty blocks were merged with adjacent blocks containing addresses such that, wherever possible, their boundaries coincided with road centrelines or railways, with priority given to the retention of principal features such as dual carriageways, motorways and railways. Although conceptually desirable, it was not possible to incorporate water features as no suitable dataset was freely available at the required scale and level of generalisation.

Table 2 Definitions of "block" employed internationally

Country	Unit	Definition	Application context	Source
Australia	Mesh block	A mesh block "broadly identifies land use such as residential, commercial, agricultural and parks etc"	Basic unit for Australian Statistical Geography Standard	ABS (2011, p.15-16)
New Zealand	Meshblock	Meshblock boundaries were originally "identifiable on the ground … followed road centrelines, river courses or other prominent features … [there is now] an increasing tendency of aligning meshblock boundaries to legally defined cadastral boundaries."	Basic building block for New Zealand Standard Areas Classification	Statistics New Zealand (1992, p.13)
UK	Street block	"An area of land surrounded by streets or other major linear topographic features such as railway lines or water features"	Informal definition employed in specific research project	Cossey et al (2005, E/S-1)
USA	Census block	"A statistical area bounded by visible features, such as streets, roads, streams, and railroad tracks, and by nonvisible boundaries, such as selected property lines and city, township, school district, and county limits and short line-of-sight extensions of streets and roads"	Basic unit for US Census output zones	US Census Bureau (2011)

To create postcode building blocks, Thiessen polygons (constrained to fall within 2001 OAs) were first generated around each residential address and communal establishment. Neighbouring address polygons within the same postcode were then merged to produce postcode polygons. An automated process, similar to that used for street blocks, ensured that all postcodes contained at least one address and that their boundaries fell along roads or railways wherever possible. These methods are consistent with ONS' practices in the 2001 and 2011 censuses (Harfoot et al, 2010; ONS, 2012a).

The 2007 household data were aggregated to street blocks and postcodes. To explore the impact of building blocks on the redesign of output zones, the authors' AZTool software (http://www.geodata.soton.ac.uk/software/AZTool/) was employed to aggregate first the street blocks and then the postcodes into two new sets of output zones which met the criteria shown in Table 3. To evaluate the impact of building blocks on the maintenance of output zones, 2001 OAs which had breached the upper or lower thresholds shown in Table 3 by 2007 were split or merged using methods described in Cockings et al (2011). The process of splitting involved aggregation of building blocks within the OA to meet the required criteria. Where splitting or merging could not be carried out, even after the removal of permitted constraints, the original OAs were retained. Any 2001 OAs within the required thresholds were also retained without modification. All split, merged and retained zones were then joined together to create a new set of OAs. The whole process was applied first using street blocks and then again using postcodes as the input building blocks. All resultant sets of OAs were then evaluated in terms of their statistical characteristics and their utility for linking postcoded data with census data. In addition, for the maintained OAs, the impact of the building blocks on the ability to split over-threshold OAs was explored.

Table 3 Constraints and criteria employed in redesign and maintenance processes

Constraint/criteria	Details	Weighting
Minimum household threshold[1]	40	N/A
Maximum household threshold[2]	250	N/A
Minimum population threshold[3]	100	N/A
Maximum population threshold[3]	625	N/A
Target (number of households)	125	100
Target tolerance[4]	10%	N/A
Homogeneity	Intra-area correlation scores for accommodation type and tenure	100
Shape	Perimeter^2/Area	100
Minimum boundary length	10% of total perimeter of shared boundaries	N/A
Regional constraint[4]	Respect higher-level output geography (lower layer super output area)	N/A

[1] From Mitchell and Ralphs (2007), Table 1.1, p.4
[2] No maximum thresholds employed for output areas (OAs) in 2001. Values here = 2001OA target mean * 2 (as in Ralphs and Mitchell, 2006)
[3] Population thresholds = household thresholds * 2.5 (equating approximately to average household size)
[4] Only used during maintenance process, not redesign

4. Results

4.1 Statistical characteristics of building blocks

Table 4 summarises the statistical characteristics of the two sets of building blocks. The postcodes are much smaller than the street blocks, with almost twice as many postcodes as street blocks. Their small size was one of the key attractions for their use as building blocks in 2001. They also exhibit greater uniformity of population and household counts and greater internal homogeneity of accommodation type and tenure, measured using the intra-area correlation (IAC) score described in Martin et al (2001) and Tranmer and Steel (1998), with higher scores indicating greater homogeneity. Overall, they are also slightly more compact than the street blocks (measured by perimeter squared divided by area; lower scores represent greater compactness: see MacEachren, 1985).

4.2 Impact of building blocks on redesigned OAs

Table 5 summarises characteristics of the completely redesigned OAs resulting from aggregation of first the street blocks and then the postcodes. In both cases, similar numbers of OAs are produced (5086 from street blocks; 5094 from postcodes). When considering all study areas together, both sets of OAs have a mean household count very close to the target (125), indicating that zones were produced at the desired scale. However, the standard deviation of these household counts is lower for postcodes than street blocks. The standard deviation of total population per zone is similarly lower for postcodes than street blocks. This greater uniformity of household/population size for the postcode OAs results from their smaller scale, which affords greater combinatorial flexibility for meeting the required criteria.

Table 4 Statistical characteristics of postcode and street block building blocks

Building blocks	Count	Total population			Total households			Homogeneity[1]		Shape[2]
		Mean	Range[3]	SD	Mean	Range	SD	Tenure	Accommodation type	Mean
Street blocks	21627	72.6	0-42783	86.2	29.4	0-605	31.3	0.196	0.364	27.07
Postcodes	43211	36.3	0-1206	40.9	14.8	0-472	15.0	0.239	0.479	25.36

[1] Intra-area correlation (IAC) (see Martin et al, 2001 and Tranmer and Steel, 1998)

[2] Perimeter^2/area (see MacEachren, 1985)

[3] 424 street blocks and 1575 postcode building blocks have zero population

Table 5 Statistical characteristics of redesigned output areas (OAs) based on street block and postcode building blocks, by study area

Study area	Count	Total population		Total households		Homogeneity[1]		Shape[2]
		Mean	SD	Mean	SD	Tenure	Accommodation type	Mean
Street block OAs								
All study areas	5086	308.5	122.1	125.2	21.8	0.157	0.263	37.13
Camden	729	318.1	109.2	125.4	24.7	0.096	0.134	33.27
Isle of Anglesey	232	297.4	44.6	125.3	11.4	0.092	0.112	44.41
Lancaster	443	324.0	215.4	124.9	15.7	0.150	0.242	39.55
Liverpool	1469	296.5	85.5	125.1	20.3	0.139	0.225	39.23
Manchester	1480	309.6	130.1	125.2	25.4	0.152	0.171	35.03
Southampton	733	315.4	116.4	125.4	19.7	0.157	0.179	37.24
Postcode OAs								
All study areas	5094	308.1	107.6	125.0	11.9	0.151	0.252	47.88
Camden	732	316.8	104.5	124.9	19.0	0.094	0.134	37.06
Isle of Anglesey	232	297.4	42.2	125.3	6.5	0.087	0.116	51.29
Lancaster	443	324.0	158.7	124.9	7.6	0.138	0.219	49.02
Liverpool	1470	296.3	83.4	125.0	12.0	0.134	0.217	51.91
Manchester	1482	309.1	115.7	125.0	10.2	0.141	0.162	48.13
Southampton	735	314.6	110.3	125.1	9.1	0.152	0.167	48.34

[1] Intra-area correlation (IAC) (see Martin et al, 2001 and Tranmer and Steel, 1998)

[2] Perimeter^2/area (see MacEachren, 1985)

Conversely, OAs created from street blocks display greater internal homogeneity of tenure and accommodation type and also more compactness of shape. Overall, accommodation type exhibits more homogeneity than tenure. Given that the street blocks themselves were less homogenous and less compact than the postcodes, the fact that they produce more homogeneous and compact zones when aggregated suggests that they may be a more ‘natural’ areal unit for these variables. It is plausible that the scales over which tenure and accommodation type vary lie somewhere between street blocks and the derived OAs and that they vary geographically between and within the study areas.

Broadly speaking, variation in statistical characteristics between study areas is greater than that observed between the street block and postcode OAs within each study area. For example, Camden has the most compact OAs and Isle of Anglesey the greatest homogeneity of tenure and accommodation type, irrespective of the building block employed. This indicates that the specific spatial patterning and scale of variation in each geographical area are the key drivers behind the patterns observed, more so than the specific building blocks employed.

OA shapes tend to be least compact in the very rural areas (Isle of Anglesey) and most compact in the urban areas (Camden, Manchester). In general, it is easier to create zones with uniform household counts in the more rural areas than in the urban areas and with the smaller postcode building blocks than the street blocks, although this pattern is mediated to some extent by variations in the population distribution in specific areas. For example, Lancaster’s results are inconsistent, with a low standard deviation of household count but high standard deviation of population; this is due to the presence of large student halls of residence, which do not contribute to the household count, but add greatly to the population totals. Homogeneity of tenure and accommodation are not obviously driven by rurality: Camden and Isle of Anglesey are the least homogeneous in this respect, for both the street block and postcode OAs. Differences between and within urban and rural areas may be hidden by these aggregate statistics but it was not feasible to explore this in more detail given that only one of the study areas (Isle of Anglesey) was truly rural (and this was only included as a ‘control’ area, exhibiting very little population change). Nationally, very few rural areas exhibited significant population change between 2001 and 2011. To explore this fully, a contemporary sub-study area level urban/rural classification would be required but this was not available at the time of analysis.

The next series of investigations explores whether building blocks influence the precision with which postcoded data can be linked to census OAs. Figure 1 shows, for each study area, the percentage of postcodes whose constituent addresses match uniquely to one OA, first for the street block and then the postcode OAs. When building OAs from street blocks, 82% of postcodes have all of their addresses within one OA, compared to 91% when using postcodes. The latter is less than 100% because: OAs must nest within higher level administrative units (which sometimes splits postcodes); some postcodes comprise non-contiguous small areas; and postcodes change through time so the intended 1:1 match between postcodes and OAs in 2001 will inevitably have deteriorated over time. These results indicate the degree of certainty which would be lost if OAs were built from street blocks rather than the postcodes currently employed. The effect is most marked in Isle of Anglesey and Lancaster.

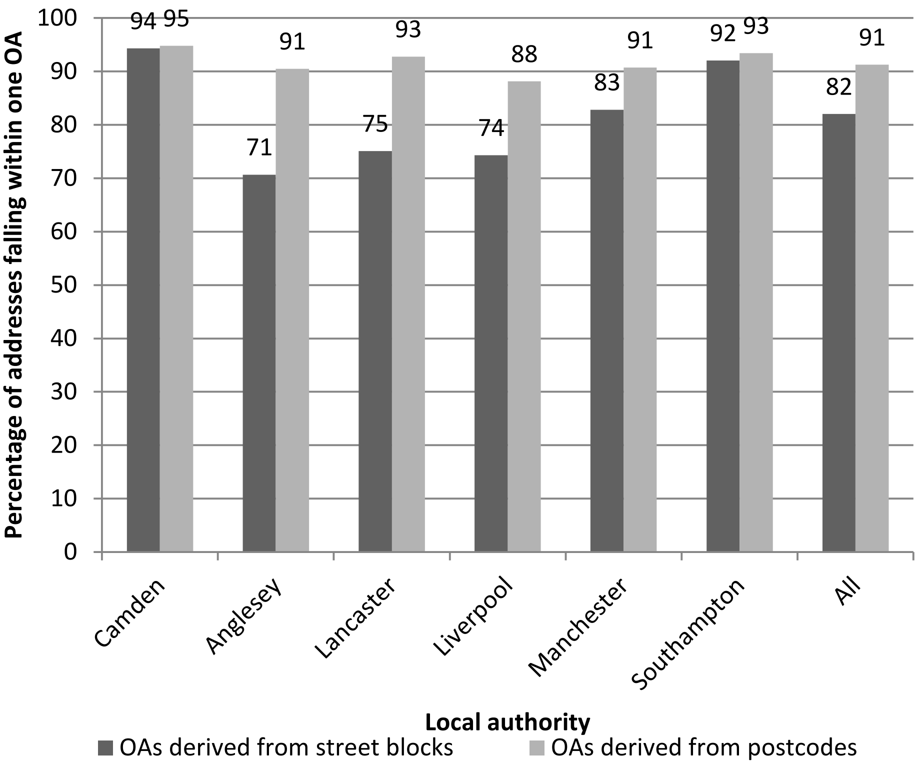


Figure 1 Percentage of postcodes whose constituent addresses fall within one output area (OA), for OAs derived from street blocks and from postcodes, by study area

4.3 Impact of building blocks on maintenance of output zones

The next set of experiments explores how the two sets of building blocks influence the maintenance of existing output zones. Using the postcodes enables more over-threshold OAs to be split (77%) than using the street blocks (54%). This superior performance is due both to the smaller size of postcodes and their geographical configuration within the types of areas experiencing population growth since the 2001 census. For example, postcodes tend to subdivide areas of high density population (such as apartments or student residences) more effectively than street blocks. Of the 46% of OAs which cannot be split using street blocks, half of these can be split using postcodes, but there are no situations where the reverse is true. 38% of street block OAs (compared to 16% for postcodes) cannot be split because at least one of their constituent blocks has population and/or household counts greater than the desired upper threshold(s). This limited granularity of street blocks at the very local scale in urban areas is a key constraint to their potential utility as building blocks. Postcodes, by contrast, are indirectly controlled for household/population size because they represent postal delivery workloads; the few places where they exceed upper thresholds occur where multiple postcodes are located on one geographical coordinate e.g. large residential buildings. New residential developments usually generate new postcodes, but are less likely to generate entire new street blocks.

Table 6 summarises the statistical characteristics of the maintained OAs created from street blocks and postcodes. As noted previously, these statistics include over-threshold OAs which cannot be split. For comparison, the results for the original 2001 OAs and the 2007 data aggregated to 2001 OAs are also included.

The OAs maintained using postcodes have lower standard deviations of population and household counts than those using street blocks and also have a mean household count closer to the target (125). Negligible differences are seen in homogeneity of tenure and accommodation type, but the street block-based OAs are slightly more compact than those based on postcodes. These findings are consistent with those observed for the redesigned OAs in section 4.1.2. Between 2001 and 2007, the OA-level means and standard deviations of population and household counts increased and homogeneity of tenure and accommodation type decreased, indicating that the statistical qualities of the 2001 OAs had started to decline. The two sets of maintained OAs demonstrate improved statistical characteristics over the 2007 results, but neither is able to return to the original 2001 values. This is to be expected given that the spatial distribution of population and households has changed since 2001 but the majority of OA boundaries have been retained.

Table 6 Statistical characteristics of maintained output areas (OAs) based on street block and postcode building blocks

OAs	Count	Total population		Total households		Homogeneity[1]		Shape[2]
		Mean	SD	Mean	SD	Tenure	Accommodation type	Mean
Maintained (street blocks)	5021	312.5	133.3	126.8	36.8	0.161	0.264	37.70
Maintained (postcodes)	5074	309.3	128.6	125.5	29.5	0.162	0.264	37.79
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

[1] Intra-area correlation (IAC) (see Martin et al, 2001 and Tranmer and Steel, 1998)

[2] Perimeter^2/area (see MacEachren, 1985)

In terms of implications for linking postcoded data, 94% of postcodes had all of their constituent addresses within one OA, for both the street block and postcode OAs. While this implies that the choice of building block has little impact in this respect, it should be remembered that the number of zones involved in the maintenance process here (152 OAs) is very small compared to the total number of zones in the six study areas (4988), thus the overall impact of changed zones is limited when measured in this way.

5. Discussion and conclusions

This paper has identified the key conceptual and practical requirements of spatial building blocks for official population statistics. It has reviewed current international practice and systematically evaluated the performance of two sets of building blocks.

Although the analysis is only for a sub-set of local authorities in England and Wales, it is based on the most up to date population estimates and address registers available and the areas selected are indicative of the types of areas which required maintenance in 2011. The results fed directly into ONS' policies and processes for the 2011 census and provide other national statistical organisations and analysts with evidence and guidance on the design of building blocks. Whilst the building blocks evaluated are two of the most commonly employed internationally, they are not the only alternatives. Further research should extend this comparative analysis to other types of building blocks, notably grid squares employed in Nordic countries.

The lack of a common international definition for a "street block" results in significant variation in practice. The development of such a standard would potentially harmonise international practice and facilitate comparison between and within countries.

Nevertheless, this paper argues that the specific conceptualisation and design of building blocks should be nationally-specific. The scale and characteristics of any building blocks should be carefully selected to match the intended purpose(s) and the objects being represented whilst also accounting for constraints related to existing geographic information infrastructures, legal and administrative requirements and available resources. The precise combination of these factors will inevitably vary by country.

These findings provide further evidence to support the contention that the scale effect of the MAUP is generally greater than the aggregation effect (Openshaw, 1984), with the smaller size of postcodes (relative to street blocks) proving to be the dominant feature in their effectiveness as building blocks. A number of authors have attempted to assess whether spatial auto-correlation, or spatial patterning, of variables is important in the MAUP: this paper concurs with Shuttleworth et al (2011) and Wong (1997) that both the spatial patterning and scale of zones relative to the objects being represented are critical in determining sensitivity to specific zonations. In their context, sensitivity is related to measures of segregation; here it is sensitivity of output zones to the building blocks employed. The findings also confirm claims by Openshaw (1984), Ang and Ralphs (2008) and Manley et al (2006) that the selection of the first zoning system is critical in determining the characteristics of any higher level aggregations. Variation due to this first aggregation is likely to be greater than any subsequent re-aggregations.

In the context of publishing population data for England and Wales, these results suggest that postcodes are more effective building blocks than street blocks, either for completely redesigning or selectively maintaining existing OAs. Postcodes result in more uniform population and household sizes than street blocks and provide greater certainty for geographical linkage between the many postcoded datasets and census areas. When maintaining OAs, postcodes also enable more (over-threshold) OAs to be split than street blocks. These findings provide strong support for ONS' policy of continuing to employ postcodes as the building block base for the 2011 census, a decision which also reflected users' preferences. Nonetheless, there are clear instances where street blocks outperform postcodes, most notably in producing OAs which are more compact and internally homogeneous in terms of tenure and accommodation type. The decision as to which to employ is dependent on the relative importance of the various design criteria. For the release of census data, adherence to confidentiality thresholds, publication of the maximum possible amount of data, and statistical stability generally take precedence over socio-economic homogeneity. This will not be the case for every application: any organisation or analyst designing building blocks and output zones should base their decisions on purpose-specific theoretical models tempered by practical considerations.

For England and Wales, the best compromise for a set of building blocks may lie in a combination of postcodes and street blocks. Arguably, the existing postcode polygons already represent such a hybrid solution, being both synthetic representations of postcodes but also aligned with key real-world features wherever possible. Recent experience in Australia suggests that it may be possible to produce an even more effective building block, which is at an appropriate scale and provides sufficient differentiation of social, economic and built characteristics of areas. To achieve this ideal would require a very clear reconceptualisation of the objects being represented and enhancement of various key datasets.

As countries such as the UK review the design of alternative, potentially post-census, official population systems, it is critical that explicit consideration be given to the conceptual and practical requirements for future building blocks. Irrespective of the methods used to collect or collate data in the future, stable and appropriate building blocks are essential. Building blocks must represent the best compromise of user requirements, be robust to change but easily updatable to reflect contemporary spatial population patterns, and provide an appropriate base for mapping and linkage. The impacts of building block design are extensive: fundamentally, all spatial analysis and research applications which employ small area data are dependent on the extent to which building blocks successfully capture spatial variation in the underlying objects being represented.

Acknowledgements

This research was supported by ESRC award RES-348-25-0019. The authors are grateful to colleagues from ABS, NISRA, ONS, Statistics Finland, Statistics New Zealand, UKBORDERS, University of Southampton and the US Census Bureau for material provided and comments on earlier drafts. They also wish to thank three anonymous referees who provided extremely useful comments on the manuscript. All views expressed in the paper are the authors' own. The research was carried out by them as Approved Researchers under secure conditions at ONS with data access granted by ONS' MicroData Release Panel.

References

ABS, Australian Bureau of Statistics

2011, *Australian Statistical Geography Standard (ASGS): Volume 1 – Main Structure and Greater Capital City Statistical Areas* Australian Bureau of Statistics, Canberra http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)

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

Cabinet Office, 2011, *Making Open Data Real: A Public Consultation* HM Government http://data.gov.uk/opendataconsultation

CDU, Census Dissemination Unit

2012, "NI 2001 Grid Data" http://cdu.mimas.ac.uk/2001/ni/grid/

Cockings S, Martin D, 2005, "Zone design for environment and health studies using pre-aggregated data" *Social Science & 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

Cockings S, Harfoot A, Martin D, Hornby D, 2011, "Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 census output geographies for England and Wales" *Environment and Planning A* **43** 2399-2418

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

Cossey R, Davies F, James R, Barr R, 2005, "The Joined-Up-Geography Testbed Project", Manchester Geomatics, Manchester

DoE, Department of the Environment

1987, "Handling Geographic Information: Report of the Committee of Enquiry chaired by Lord Chorley" (HMSO, London)

Flowerdew R, 2011, "How serious is the Modifiable Areal Unit Problem for analysis of English census data?" *Population Trends* **145** (Autumn) 1-13

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

Galster G, 2001, “On the nature of neighbourhood” *Urban Studies* **38** 2111-2124

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

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

Holt D, Steel D, Tranmer M, 1996, “Area homogeneity and the modifiable areal unit problem” *Geographical Systems* **3** 181 – 200

House of Commons Treasury Committee, 2008, “Counting the population”, Eleventh Report of Session 2007-2008, Volume 1, (The Stationery Office Limited, London)

MacEachren A, 1985, “Compactness of geographic shape: comparison and evaluation of measures” *Geografiska Annaler B* **67** 53-67

Manley D, Flowerdew R, Steel D, 2006, “Scales, levels and processes: Studying spatial patterns of British census variables” *Computers, Environment and Urban Systems* **30** 143 – 160

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* **31** 1 6 – 18

Martin D, Higgs G, 1997, "Population georeferencing in England and Wales: basic spatial units reconsidered" *Environment and Planning A* **29** 333 – 347

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

Mitchell B, Ralphs M, 2007, “Developing maintenance rules for the Neighbourhood Statistics Output Geographies”, Methodology Directorate, Office for National Statistics.

NISRA, Northern Ireland Statistics and Research Agency

2008, *NISRA Geography* http://www.nisra.gov.uk/geography/default.asp.htm

ONS, Office for National Statistics

2004, *Census geography evaluation report* http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/review-and-evaluation/evaluation-reports/geography/evaluation-report.pdf

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

2012a, *2011 Census Geography – Modifications of output areas* *http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/new-developments-for-2011-census-results/2011-census-geography/modifications-of-output-areas/index.html*

2012b, *National Statistics' – A Beginner's Guide to UK Geography: Super Output Areas (SOAs)* http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html

Openshaw S, 1984, "The Modifiable Areal Unit Problem" *CATMOG 38* http://qmrg.org.uk/catmog/

Openshaw S, 1990, "Spatial referencing for the user in the 1990s" *Mapping Awareness* **4** 24 – 29

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

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

Shuttleworth I, Lloyd C, Martin D, 2011, "Exploring the implications of changing census output geographies for the measurement of residential segregation: the example of Northern Ireland 1991-2001" *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174** 1 – 16

Sommer E, 2003, "Grid data from Denmark: Information bearing cells" 9th EC GI & GIS Workshop: ESDI Serving the User, La Coruña, Spain, 25-27 June.

Statistics Finland

2012, *Grid database* http://www.stat.fi/tup/ruututietokanta/index_en.html

Statistics New Zealand

1992, *New Zealand Standard Areas Classification Manual* (Statistics New Zealand, Wellington).

2012, *Meshblock dataset* *http://www.stats.govt.nz/Census/2006CensusHomePage/MeshblockDataset.aspx?tab=About*

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

US Census Bureau

2010, *Strength in Numbers: Your Guide to Census 2010 Redistricting Data from the US Census Bureau* http://www.census.gov/rdo/pdf/StrengthInNumbers2010.pdf

2011, *Geographic Terms and Concepts* http://www.census.gov/geo/www/2010census/gtc_10.html

2012a, *Geographic Areas Reference Manual: Chapter 11 – Census Blocks and Block Groups* http://www.census.gov/geo/www/garm.html

2012b, *Census Summary File 1* http://www.census.gov/prod/cen2010/doc/sf1.pdf

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

Wong D, 1997, "Spatial dependency of segregation indices" *Canadian Geographer* **41** 128-136

Appendices

Appendix 1

Confirmation of authorship

This note is provided in confirmation of the statement of paper authorship provided in the Preface to this thesis. I have been the candidate's line manager throughout the period and the research presented here builds on my own work initially conducted in the late 1990s. I have not at any stage formally acted as supervisor to this programme of research and it has been conducted with a degree of intellectual and practical independence far higher than that which would apply in any conventional postgraduate supervision arrangement. The initial paper (Paper 1) was undertaken using software which I had already written, whereas the remaining four papers are based on subsequent developments led by Samantha. Papers 2, 3, and 5 are direct outputs from Samantha's own single-investigator ESRC-funded research on which I served, with others, on a project advisory board. None of the papers included here are included in my own submission to the 2014 Research Excellence Framework.

DJMartin

Professor David Martin

Geography and Environment Academic Unit, University of Southampton