UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS & SOCIAL SCIENCES

School of Social Sciences

Division of Social Statistics

Methods of Geographical Perturbation for Disclosure Control

by

Caroline Jane Young

Thesis for the degree of Doctor of Philosophy

February 2008

Correction Sheet

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS & SOCIAL SCIENCES
SCHOOL OF SOCIAL SCIENCES
DIVISION OF SOCIAL STATISTICS

Doctor of Philosophy

METHODS OF GEOGRAPHICAL PERTURBATION FOR DISCLOSURE CONTROL

by Caroline Jane Young

Disclosure control methods are used to protect the confidentiality of individuals and households in aggregate census data. With growth in computational power, the disclosure control problem has been rapidly transformed. Increased analytical power has stimulated user demand for more detailed information for smaller geographic areas and to customized geographical boundaries. However, the possibility of allowing census users to create their own aggregates from census microdata, and for small areas, can lead to problems of disclosure by differencing. Traditionally, methods of statistical disclosure control have been aspatial in nature. This thesis describes a new framework of geographical perturbation methods designed to deal with the spatial nature of disclosure risk.
The research offers several new contributions, specifically;

(1) A framework of new geographical perturbation methods is defined, based on creating uncertainty around geographical location. Zone-independent methods are designed for protection in a flexible-tabulation scenario and to account for the spatial dimension of risk.

(2) Techniques for implementation of these methods are tested on a synthetic census dataset which show comparable risk-utility outcomes to RRS (an existing method used for the US and UK Censuses). The advantages and disadvantages of the proposed methods are discussed with regard to ease of implementation and flexibility of parameter values.

(3) One of these new methods; LDS, is then explored in more detail showing a significant improvement over RRS in terms of the risk-utility outcome. Risk reduction is illustrated in a geographical differencing scenario and distortion to utility explored in a spatial context of typical census users' analyses.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This research was supported by an NCRM-linked ESRC postgraduate studentship:- PTA-042-2004-00013. The work in this thesis would not have been possible without extensive use of census data disseminated through the ESRC/JISC Census Programme.

My sincere gratitude goes to both my supervisors, Chris and Dave, for their enthusiastic supervision during the past three years. I am extremely grateful for their much valued advice with research ideas, generosity with time, and guidance with papers and presentations. Coming from a statistical background, the opportunity to be involved with the geographical aspects of the research was particularly enjoyable. Throughout my doctoral work they have both encouraged me to develop independent thinking as well as other research skills, and have greatly assisted with my scientific writing; all of which will be invaluable in the future. It was a privilege to work with you both.

Thank you also to the staff in the Social Statistics department who encouraged me to do a PhD at Southampton. James Brown's advice has been especially beneficial in guiding this research. I would also like to acknowledge the computing staff that have solved many technical problems. Similarly I extend my gratitude to the administrative staff whose assistance has been much appreciated.

I am also grateful to the ONS and specifically the SDC team for the continued work on various projects linked to my PhD. This has allowed me to keep in mind a wider perspective on SDC issues. In addition, their thoughts relating to my research were always inspiring.

My time at Southampton will always be very special because of the friends I have made especially: Solange, Leo, Gloria, Amos, Mei, Sylke and Laura. Thank you particularly to Gail for your hospitality. Also, the POPFEST committee (Claire, Alex, Bernie, Guy and David) made the highlight of my third year! My personal thanks also go to my family and friends who have encouraged me during this period and to those specifically, who have given me support in some way: my parents, Tyrone, Louise and not forgetting Grandma. Thank you for giving me an outside perspective and listening to me talk about my work!

Abbreviations

AAD – Absolute Average Deviation

ED – Enumeration District

ESDA – Exploratory Spatial Data Analysis

FCSM – Federal Committee on Statistical Methodology

GIS – Geographical Information Systems

GOR – Government Office Region

GROS – General Register Office Scotland

IPF – Iterative Proportional Fitting

LAD – Local Authority District

LDS – Local Density Swapping

LISA – Local Indicators of Spatial Autocorrelation

LLTI – Limiting Long Term Illness

NeSS – Neighbourhood Statistics

NISRA – Northern Ireland Statistics and Research Agency

NSIs – National Statistical Institutes

OA – Output Area

ONS – Office for National Statistics

RAD – Relative Absolute Deviation

RRS – Random Record Swapping

SAS – Small Area Statistics

SAR – Sample of Anonymised Records

SDC – Statistical Disclosure Control

SOA – Super Output Area level

SRS – Simple Random Sampling

Chapter 1 Introduction

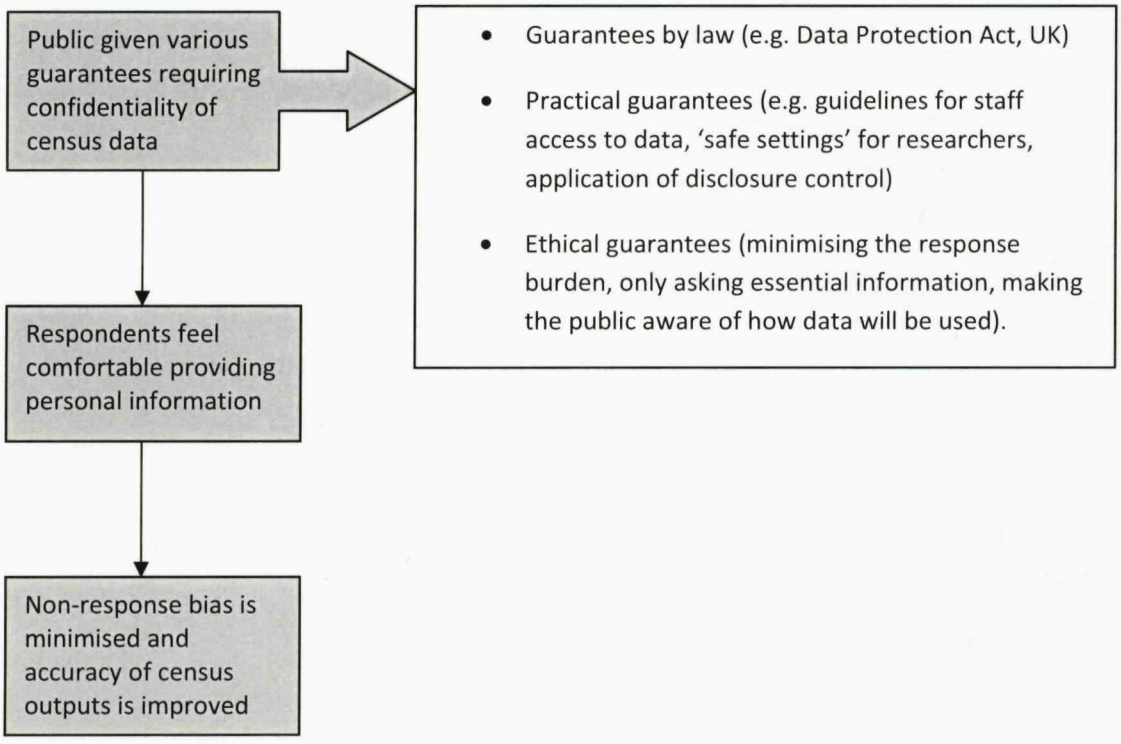
1.1 Motivation

This thesis is concerned with disclosure control of aggregated data produced from a census of population. Public release of census data is vital for researchers, businesses, policy makers, for government planning and resource allocation. The research addresses a problem faced by many National Statistical Institutes (NSIs), who would like to publish as much census data as possible to be able to meet user needs, but cannot do so without compromising the confidentiality of the respondents. The focus of this thesis will concentrate on the demand for NSIs to publish tables for small areas and to many differently defined geographies. Such demand pressures have led to discussion of the development of flexible tabulation systems in, for example, the UK, Australia and the US (Rhind et al. 1991, Zayatz 2003, Duke-Williams and Rees, 1998a). Any such system would allow users to create their own customised tables from unpublished individual records. This scenario presents a considerable disclosure risk. Protecting the confidentiality of census data by application of *statistical disclosure control* (SDC) methods is an integral part of the census process allowing use of protected data by researchers and policy makers across all sectors. SDC methods generally either

restrict or modify the detail released (Willenborg and De Waal, 2001). However little attention has been given in the literature to the spatial aspects of risk and the link between geography and disclosure arising from a flexible tabulation system. In this thesis, we attempt to incorporate 'space' into the development of new methodology to protect against these challenges.

The importance of the issue is magnified by reliance of official statistics on public trust in the safeguards employed. Keeping census data confidential is vital to obtain the benefits of censuses and to reduce biased conclusions arising from non-response. Figure 1.1 illustrates the need for disclosure control.

Figure 1.1: Obtaining the Benefits of the Census by Keeping the Data Confidential



NSIs are generally under strict legal obligations requiring confidentiality of census data. For example; the Privacy Act of 1974 for the US Census, the Census and Statistics Act (amended 2000) for the Australian Census and the 1920 Census Confidentiality Act and 1991 Data Protection Act in the UK. These laws generally have some requirement to the effect that no individual will be identified. The public are also provided with other assurances that their data are kept safe; NSIs will usually follow

certain principles in regard to any census activities. These include ethical principles such as making the public aware of how the data will be used as well as practical principles; who can access the data and how they are allowed to access it. A very important part of the census process is to apply disclosure control before the release of data. If disclosure control was not applied and a journalist, for example, was to make a sensationalist claim in the press revealing supposedly confidential information, it would damage the reputation of the NSI and it is likely that response rates would fall significantly. This is even more of an issue with the 2011 UK Census because more sensitive questions may be introduced; for example a new question on income. A high response rate is essential for the success of a census, to ensure accuracy of the data.

1.2 Context

In this thesis we will concentrate on disclosure issues in a UK context although we briefly review practices in other countries and in particular those which take a population census. The new methodology will be developed to be applicable to census data generally, not just in the UK, and possibly also to other high sample fraction data sources.

The national censuses in the UK are arguably the most widely used data source providing detailed information on individuals and their households. Whilst the full dataset is never released directly to the public, data from the census are often published as tables of counts down to neighbourhood level. If the raw data were to be published for neighbourhoods or other small areas, it may be possible to recognise data relating to individuals, especially when there is good background knowledge of the area. This is a *disclosure risk* since these individuals may be identified leading to potentially sensitive information being revealed. SDC methods are applied to disguise the data without destroying the relationships among the variables. User demand for data changes with each new census, in turn transforming and usually increasing the disclosure control challenges. New SDC methods need to be developed in response.

- User Demand for Small Area Data

Various sectors of society are users of the census and their demands for information shape the statistical outputs that must be provided. Over recent years, the needs of census users have evolved towards increasing levels of detail and for smaller areas. Small area data are crucial in understanding spatial variation. For example, government can locate the most deprived neighbourhoods by understanding the spatial distribution of poverty. Providing denominators for morbidity and mortality for small areas enables important studies in epidemiology. Information at this level can also be used to measure change in rural and urban migration on the micro-scale. Furthermore, geodemographic classifications or 'area profiles' such as ACORN* are widely used in businesses and marketing and are all based on data at the local level.

- User Demand for Flexible Tables

Different users want information for different geographical areas. For example, to analyse population change over time, small areas that match those used in previous censuses are essential. On the other hand local government requires data from small areas fitting exactly into current administrative units to be able to assign grants whilst businesses want data based on exact aggregations of unit postcodes to link to data in non-census databases. With each modern census, users have had more computing power and demanded more detail - thus increasing the capacity for linking records and potentially identifying individuals in aggregate datasets. The new Neighbourhood Statistics Service* (NeSS) for England & Wales is an example of data which are aggregated from multiple geographically referenced administrative records to a single set of output areas and super output areas. However in the future NeSS may be required to deliver counts for other non-standard geographies which are not neat aggregations of OAs. A method of disclosure control is needed to protect data sources such as NeSS regardless of the aggregation strategy adopted.

1.3 Potential for Disclosure

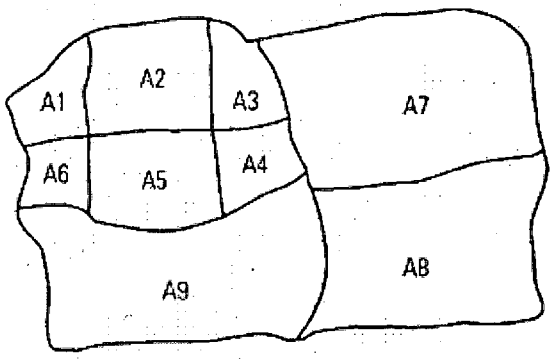
These developments in information technology coupled with increasing demand by the public for more detailed data for non-standard geographies, have generated rising levels of concern about confidentiality. Data released for very small areas for example, with fine variable breakdown can lead to disclosure. In table 1.1, the small cell counts are potential disclosure risks because it can be deduced that the only 20-21 year old in the area must claim benefits;

Table 1.1: Example table of potential disclosure risk

	...	16-17	18-19	20-21	22-23	24-25	...
Benefit claimed	...	2	2	1	3	2	...
Benefit not claimed	...	5	7	0	7	6	...

Moreover, appropriate disclosure control methods must be applied to ensure that different sources cannot be compared by geographical differencing. Geographical differencing occurs when information is published for two very similar geographies allowing the table relating to the smaller geography to be subtracted from the other to obtain information for a previously undefined area. This problem has been well described by Duke-Williams and Rees (1998a). They show an example of a Geography A defined by the following boundaries[1] as in figure 1.2(a):

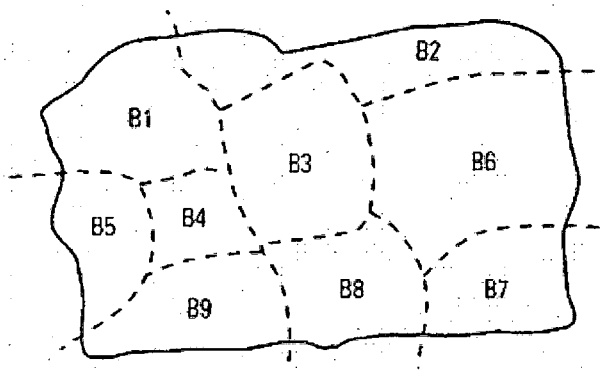
Figure 1.2(a): Geography A



[1] Reproduced from Duke-Williams and Rees (1998a)

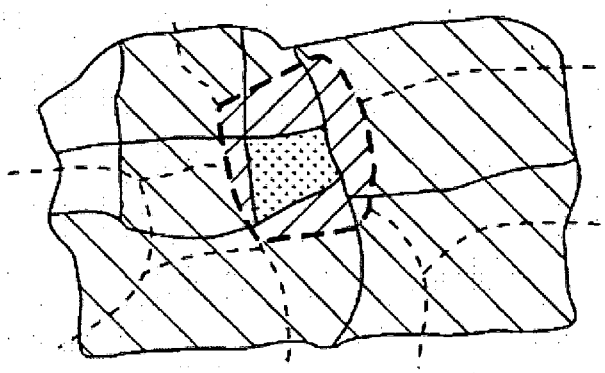
Suppose the existence of a Geography B in figure 1.2(b) defined by different boundaries (for the same base population):

Figure 1.2(b): Geography B



With modern GIS software, it is possible to overlay the two geographies (place one on top of the other) as in figure 1.2(c). Sometimes cases occur where one geographical area nests entirely within the other. The dotted polygon in the centre lies entirely within the outlined thick-dashed polygon, i.e. A4 lies within B3.

Figure 1.2(c): Overlay the two geographies



If tables were published for these nested areas then they could be 'differenced' to produce a new previously unknown table. For example suppose the two tables below show data published for the

two nested geographies. Table 1.2(b) (relating to the smaller dotted area) is nested within table 1.2(a) (relating to the larger dashed line area).

Table 1.2(a): Table relating to large boundary (Geographical Differencing)

	16-20	21-30	31-40	...
Benefit Claimed	10	16	19	...
Benefit Not Claimed	8	12	11	...

Table 1.2(b): Table relating to smaller boundary (Geographical Differencing)

	16-20	21-30	31-40	...
Benefit Claimed	9	16	19	...
Benefit Not Claimed	8	11	11	...

The corresponding cells in the two tables can be differenced to produce a new table relating to the differenced area which reveals previously unknown information. In fact table 1.2(c) shows cell counts of one which we later refer to as 'identity disclosure'. We can reveal that there is only one 16-20 year-old living in the differenced area, who must claim benefits.

Table 1.2(c): Subtracting tables 1.1 and 1.2 gives a new unpublished table

	16-20	21-30	31-40	...
Benefit Claimed	1	0	0	...
Benefit Not Claimed	0	1	0	...

This might occur for example when census outputs are published for two different sets of boundaries at the same point in time.

1.4 New Disclosure Control Methods

Disclosure control in the form of *geographical perturbation* can help protect against these risks by introducing uncertainty into the geographical location of households at the small area level. This can be done in such a way as to minimise damage to the data. The aim of this research will be to find an appropriate method (or methods) of geographical perturbation that will achieve the right balance in allowing more flexible outputs for users at the same time as minimising disclosure risk. The methods deal specifically with the spatial nature of the disclosure control problem of aggregated census data.

Random Record Swapping (RRS) is an existing approach applied to the UK Censuses in 2001 (Boyd and Vickers, 1999) and is one form of geographical perturbation. This involved moving records between Output Areas but within the same Local Authority District and was applied before the data were aggregated into tables. The method is dependent on the zonal geographies and thus can result in inconsistent perturbation over time due to boundary changes. This type of swapping also results in households being moved an arbitrary distance according to the sizes of the LADs and the population distributions within them. Simple relocation of records according to a distance from a probability distribution would result in excessively long moves in areas of high population density and have no effect in remotely populated regions. This thesis explores new approaches to geographical perturbation including methods that take into account the local spatial population distribution and ignore references to pre-existing geographical boundaries.

Throughout the text we refer to the risk-utility approach of Duncan et al. (2001) to assess the quality of SDC methodologies. This approach involves finding a balance between reducing the disclosure risk and preserving the utility of the data. As a general rule, the more the data are modified, the more information is lost. Ideally the disclosure risk should be reduced to a tolerable level while retaining the important patterns and trends in the data.

1.5 Outline of the Thesis

In summary, some techniques under a new framework of geographical perturbation methods will be described for disclosure control of census data. The risk-utility properties will be compared for the methods comparing against a benchmark RRS (replicating closely the approach of the UK Censuses 2001). The methods will be tested on a synthetic census population. An outline of the chapters is as follows:

Chapter 2: SDC for Census Tables

Chapter 2 describes in more detail the disclosure scenario and how disclosure might occur. This involves a review of the current literature with a particular focus on the risk from geographical differencing. Some current disclosure control methods are discussed and their advantages and disadvantages considered. All methods of disclosure control damage the data to a certain extent, thus

this should be balanced against the reduction in risk. The second half of chapter 2 focuses on the utility of the data, looking at typical analyses census users perform and finishing with some indicators of distortion after disclosure control has been applied.

Chapter 3: New Methods of Geographical Perturbation for Disclosure Control

Chapter 3 introduces a new framework of geographical perturbation methods for census data. These new methods have been developed specifically to deal with the spatial nature of the disclosure problem. Alternative ways of moving households instead of swapping are considered; displacement and rearrangement. New, zone-independent methods including Local Density Swapping (LDS) are described which ignore reference to geographical boundaries. RRS, the benchmark method, will be explained in detail. Ways to reduce the risk from geographical differencing and in small area data are hypothesised. These methods form the basis for the empirical work in chapter 5. At the end of this chapter, assessment measures for balancing risk-utility, in the context of frequency tables, are discussed.

Chapter 4: Building a Synthetic Population

Chapter 4 constructs a synthetic census population. These data will represent the Hampshire region. The objective will be to test new disclosure control methods on this synthetic dataset (or to particular areas within this region) so it needs to display all the complexities of real data. Microsimulation techniques are considered in some detail; methods which simulate based on true data such as samples of the real population. This chapter follows on by describing the creation of a synthetic dataset using spatial microsimulation based on 1991 census data. The fit of the synthetic population to the real Hampshire population is briefly reviewed.

Chapter 5: Empirical Assessment of Geographical Perturbation Methods

Chapter 5 presents some experimental work based on the methods proposed in chapter 3. The synthetic population is used for testing the new approaches which are assessed in terms of indicators of risk and utility. The analysis follows a sequential procedure, testing out different features of the methods (sampling fraction, distribution type, etc) to see which produce the best risk-utility outcome. Displacement is contrasted with swapping and results are shown at each stage of the experimental

process. The RRS is used as a benchmark for comparison. To end this chapter, the 'best' new method: LDS, is studied in more detail with reference to the geographical differencing problem.

Chapter 6: Impact on Complex Analysis Methods

In this penultimate chapter, the utility aspect of geographical perturbation is examined more thoroughly, comparing the two methods: RRS and LDS. Utility is assessed from a spatial perspective as it is the spatial relationships in the data which are distorted after geographical perturbation. We consider three different strategies to analysing the data; namely exploratory spatial data analysis, multivariate area classifications, and multilevel modelling.

Chapter 7 draws conclusions from the empirical work and indicates areas for further research.

Chapter 2 Statistical Disclosure Control for Census Tables

2.1 Introduction

This chapter reviews the current literature on SDC with reference to the geographical differencing and small area problem. In section 2.2 we set the scene for the problem providing a statistical framework and describing the terminology used. Disclosure scenarios are described as well as illustrations of disclosive tables. Section 2.3 discusses the specific issue of geographical differencing as distinct from differencing in general. We then digress briefly to consider disclosure arising from other types of spatial data but in non-census contexts such as in mapped health data for example (geoprivacy). This research is of relevance because the spatial confidentiality methods used may be applicable in a census context. Section 2.4 then examines SDC methods that might be implemented to protect against disclosure in a census context as well as geomasking methods used in the field of geoprivacy. All methods of SDC result in a loss of utility so ideally a disclosure control method should be applied that retains the usefulness of the data in the context of user needs. The remainder of the chapter then focuses on the utility of the data. Section 2.5 reviews and summarises the statistical techniques users employ to analyse census data.

2.2 What is Statistical Disclosure?

2.2.1 Framework and Notation

We consider tables based upon an underlying microdata file. A microdata file Z can be represented in the form of an $N \times M$ rectangular matrix:

$$Z = \begin{pmatrix} z_{11} & \cdots & \cdots & z_{1M} \\ z_{21} & \cdots & \cdots & z_{2M} \\ \cdots & z_{ij} & \cdots & \cdots \\ z_{N1} & \cdots & \cdots & z_{NM} \end{pmatrix}$$

where N denotes the number of units in some population $U = \{1,...,N\}$ and z_{ij} denotes the value for units $i \varepsilon U$ on variable j. The record for unit i is the $1 \times M$ vector $z_i = (z_{i1}, z_{i2}, ...z_{iM})$.

The form of the microdata set to be used in this research will consist of three types of variables; identifiers $I = \{I_1, I_2, ...\}$, attribute variables $A = \{A_1, A_2, ...\}$, and spatial variables $S = \{X, Y\}$.

$$z = \{I, A, S\}$$

Generally a microdata file is *anonymised* meaning that the direct identifiers *I* (such as household address) are omitted for confidentiality reasons. So we consider $z = \{A, S\}$. The spatial reference variables *S* refer to the spatial point location of unit *i* in the study region $\Re^2$ given by the geographic co-ordinates $(X_i, Y_i)\ \varepsilon\ \Re^2$. Individuals in the same household have the same spatial point location. *S* can also be considered a direct identifier since it pinpoints the exact location of a household. However *S* will be included in our microdata set since these variables are necessary to carry out the geographical perturbation methods.

The study region is divided into geographical zones (for example, wards or Enumeration Districts) denoted by $O_1, O_2, ... \subset \Re^2$ each defined by a particular boundary. Furthermore assume these

geographical zones are mutually exclusive $O_v \cap O_{v'} \neq \phi$ for $v \neq v'$. The spatial point data can be aggregated according to these different geographical zones which will be referred to generally as *output zones* throughout the thesis. A frequency table comprises count data relating to the number of households or individuals who possess the properties defining the cells. A frequency table may be derived from the microdata. Let z_i^* be an $1 \times R$ subvector of the original $1 \times M$ vector z_i for R selected variable values. Let c represent any $1 \times R$ vector of values taken by z_i^*. c represents a cell in a table formed by the cross-classification of the R variables. Then a frequency for the cell c in a table representing area O_v (representing geographical zone v) is defined as:

$$F_c = \sum_{i \in O_v} I\left(z_i^* = c\right) \tag{2.1}$$

In this thesis, we focus on the assessment of disclosure risk in frequency tables. F_c as defined in (2.1) will be used to refer to cells of the tables. F_c^o refers to a cell in the original (unprotected) table and F_c^p refers to a cell in the protected (disclosure-controlled) table. Given our interest in census data, we shall focus on the case where the units are people or households.

2.2.2 The Disclosure Scenario

The disclosure risk of a frequency table greatly depends on the motive of the intruder and the ways in which they will attempt a disclosure; referred to as the *disclosure scenario* by Willenborg and De Waal (1996). We define an intruder as the external user or users who attempt a disclosure as in Duncan and Lambert (1989). Intruders may group together as a coalition sharing their knowledge. To avoid confusion, other equivalent terminology: snooper, attacker, shall not be used.

The information that the intruder will try to disclose is called *sensitive information*. Lambert (1993) talks about data on diseases, debts, and credit ratings as typically constituting sensitive information. More general information on what constitutes sensitive data can be found in the Data Protection Act[2]

[2] Part 1 of the 1998 Data Protection Act has a section (2) describing what 'sensitive personal information' means. This includes ethnic origin, political opinions, religious beliefs and physical or mental health or condition.

and in the Caldicott report[3]. Variables that warrant greater protection because they might reveal sensitive information are called *sensitive variables*, e.g. variables on criminal behaviour. Deciding upon sensitive variables in practice is often subjective. For example the number of bathrooms in a household could be considered a sensitive variable by some. Also age could be considered sensitive if revealed to the exact year. Willenborg and De Waal (1996) define a sensitive variable as one where at least one of its possible values is sensitive.

Continuing with the disclosure scenario, we imagine that an intruder uses key variables to attempt a disclosure. *Key variables* form a subset of the attribute variables. These variables are not usually sensitive. Bethlehem et al. (1990) refer to key variables as those variables in the record that allow a person to identify a record, that is, to establish a one-to-one correspondence between the record and a specific individual. Well-known key variables are age, race, sex and occupation. Such variables are often *visible* and *traceable*. Visible refers to a characteristic that is easily seen. For example certain occupations; a doctor, a policeman, etc. Traceable refers to a characteristic that is easily traced such as the number of cars a household has, or marital status. The Federal Committee on Statistical Methodology (FCSM) (1994) refer to high visibility variables as that information available to others in the population which could be used with released data to uniquely identify someone. These variables require additional protection. Records which represent respondents with unique characteristics such as very unusual jobs (movie star) or very large income are very visible and represent a high risk.

In practice, there are various ways an intruder might try to discover information, encapsulated by Elliot and Dale (1999) for census and survey data. To understand the nature of the disclosure problem more fully, some scenarios of how disclosure might occur are presented in figure 2.1:

[3] The Caldicott Report (December 1997) was a review commissioned by the Chief Medical Officer *'owing to increasing concern about the ways in which patient information is being used in the NHS in England & Wales and the need to ensure that confidentiality is not undermined. Such concern was largely due to the development of information technology in the service, and its capacity to disseminate information about patients rapidly and extensively'*.

Figure 2.1: Scenarios describing how disclosure might occur

S1. Indirect disclosure from a database cross match – A private company wishes to enhance their external database and uses fields identical or recodable to the census to cross-match. If a record is unique, new information may be learnt from the non-matching variables on the census.

S2. Disclosing information for a specific target individual – there are a number of ways in which this type of disclosure may occur:

- Nosy neighbour: the intruder uses information about a single individual based on personal knowledge.
- An organisation may wish to enhance or verify information about a target individual, e.g. the Inland Revenue may want to search for income related information of an individual suspected of tax evasion
- Computer hackers wishing to steal another's identity
- Local search using information an estate agent might hold for example, accommodation type, number of rooms in the house, bathrooms, presence of central heating, etc to reveal further information about a household
- Self disclosure: a respondent complains to the media about privacy violation as they think they can see themselves in the data (also known as spontaneous recognition)

S3. Disclosing information for an arbitrary individual – in this scenario the intruder would be interested in the consequences of claiming that information can be disclosed, but not in the actual identity. For example a journalist interested in the political consequences of claiming that identification has been achieved.

S4. Disclosing information based on a specific group of individuals – this scenario may be an extension of S2 or S3 but specifically applies to a group of individuals which are selected because of their distinctive characteristics (e.g. come from a certain minority ethnic group) or because a great deal of matching information is held about them (e.g. belong to an occupation with a register of all members).

S5. Disclosure via reverse matching – the intruder starts with the census output searching for someone unique or distinctive and then tries to locate them in the real world.

Adapted from Elliot and Dale (1999)

These scenarios indicate that an intruder may not necessarily go about an attack by looking at an individual table on the off-chance that the intruder can find something out about someone they recognise. An automated method may be a more realistic approach where many tables are considered. We note that disclosure in these scenarios most commonly originates from the

occurrence of unique records (S1 - S3 and S5) or for small cell counts (S4). In section 2.2.4, we look at some examples of tables revealing disclosive information.

2.2.3 What constitutes Disclosure?

There is much discussion in the literature about definitions of disclosure and the different types. Keller-McNulty and Unger (1993) refer to disclosure as occurring under two conditions as a result of data release: (1) a specific entity (or entities) can be linked to one or more data objects. (2) confidential or sensitive information about this entity (or entities) is learned. The first part is commonly referred to as *identity disclosure*. An equivalent definition is given by Paass (1988) whose definition also alludes to identification of a respondent from a released file. Similarly, Lambert (1993) describes identity disclosure as occurring when a data subject is identified from a released file.

Learning sensitive information about the entity (the second part of Keller-McNulty's definition) represents another form of disclosure: - *attribute disclosure*, also proposed by Cox and Sande (1979). Cox and Sande (1979) state that if sufficiently accurate data are present for correct identification of a respondent and a good approximation of confidential data, and if it is possible to correctly associate that data with the respondent, then statistical disclosure has occurred. Lambert (1993) states attribute disclosure as occurring when sensitive information about a data subject is revealed through the released file.

Dalenius (1977) offers a third concept of disclosure relating to whether a microdata value can be determined more accurately than is possible without access to the release of data. This is a probabilistic concept commonly known as *inferential disclosure*. Duncan and Lambert (1989) also refer to inferential disclosure; when the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible.

2.2.4 Types of Disclosure

Three types of disclosure have been introduced: identity, attribute and inferential disclosure, which will now be considered in more detail and illustrated with examples. We only focus on frequency tables to concentrate on disclosure that would arise in a flexible tabulation scenario. The following tables are real examples taken from the census small area statistics website[4]. However these tables have been disclosure-protected by a technique called small cell adjustment meaning that the small cell values have been modified. Exact details of the method are confidential so we assume that *small cells* refers to ones or twos. We have probabilistically imputed back the ones and twos into the tables. Therefore these tables are not 100% accurate but give examples that are realistic. We assume the intruder has a set of key variables: information that is likely to be known about an individual in the public domain such as broad age-group, gender, possibly occupation, etc. These tables represent data from a complete census, that is (in theory[5]), all people in the area are included.

Identity disclosure occurs when a respondent can be identified from the released data. Table 2.1 shows an example of identity disclosure.

Table 2.1: Example of Identity Disclosure information (Family Type in Burlesdon and Old Netley)

	Age Grouping					
	16 to 17	18 to 19	20 to 22	23 to 24	25 to 29	Total
Has children	3	4	43	46	223	319
- lone parent	1	0	22	9	57	90
- married	2	0	6	26	104	137
- cohabiting couple	0	4	15	11	62	92
No children	181	141	196	134	343	995

This case of identity disclosure may not be of much concern to most statistical agencies as the intruder is unlikely to learn anything new about the data subject that they don't already know or can't

[4] The census small area statistics website (*www.census.ac.uk/casweb/*) disseminates census aggregate outputs for the 1971, 1981, 1991 and 2001 censuses. The census small area statistics provide detailed tables for small areas based on two or three variables.

[5] In reality, a census is likely to include response errors such resulting from under-count and item-imputation. These factors will contribute to reducing the disclosure risk for a real census.

easily find out. It may already be known to most people in the area that there is only one 16-17 year-old who is a lone parent with children and therefore the information considered non-sensitive. Table 2.2 shows a second example of identity disclosure. Here the table reveals that in this small area, only one pensioner lives in a caravan and she is a lone female. Referring back to the scenarios in figure 2.1, this may arise from self-disclosure, i.e. the woman spots herself in the table. Accommodation type may be used as a key variable whereas whether or not someone is a lone female (i.e. living alone) could be considered a sensitive variable. Whether or not the table is considered disclosive would be a subjective decision made by the statistical agency.

Table 2.2: Example of Identity Disclosure revealing potentially sensitive information (Accommodation for Pensioner Households)

	Lone Males			Lone Females			Total
	65-74	75-84	85+	65-74	75-84	85+	
House or Bungalow	19	21	9	105	78	23	255
Flat, Maisonette, Apartment	4	3	4	17	22	7	57
Caravan	0	0	0	0	1	0	1
Shared Accommodation	0	2	0	3	0	0	5
Total	23	26	13	125	101	30	318

Cell counts of one in tables such as these, will be referred to as *cell uniques:* $F_c = 1$, throughout this thesis. Cell uniques are the only units in the area (represented by the table) with that particular set of characteristics. The lone female in table 2.2 is *unique* in the area[6].

A further case of identity disclosure can occur with small cell counts such as values of 2 illustrated in bold in Table 2.2. One of the lone males aged 75-84 who lives in shared accommodation would be able to recognise himself in the table. Suppose he knew the other man who lived in shared accommodation, he would then learn that this other man is also aged 75-84 and lives alone.

[6] The term *unique* is more generally used in the context of microdata to describe records representing a high disclosure risk. See Bethlehem et al. (1990), Skinner et al. (1994) or Elliot (2000) for a discussion of uniques in microdata.

Attribute disclosure occurs when confidential information is revealed and can be attributed to an individual or a group as in scenario S4. Typically this happens when the marginal total is greater than one and most of the rows or columns contain zeros. This is of particular concern for frequency tables. In table 2.3 we see an example of attribute disclosure where it is revealed that all people in the area who stated their religion as Hindu have a long-term illness.

Table 2.3: Example of Attribute Disclosure

Religion	**Without LLTI**	**With LLTI**
Christian	3251	1004
Buddhist	8	0
Hindu	0	2
Jewish	4	1
Muslim	1	1
Sikh	0	0
Any other religion	13	3
No religion	556	86
Religion not stated	248	106
Total	4089	1205

A special form of attribute disclosure is within-group disclosure which occurs when a data subject who is a respondent in the table, reveals the confidential information and can attribute it to an individual or group. In table 2.3, the Jewish person living in the area represented, who does have a limiting long term illness can reveal that the other Jewish people in the area do not have a limiting long term illness. Within-group disclosure is generally not considered as important as other types of disclosure as it occurs less often.

The tables so far illustrate how disclosure commonly originates from small cell counts, zeros and in particular cell uniques. The third type of disclosure mentioned was inferential disclosure which results when the intruder infers new information about a respondent from the released data, even if no released record is associated with the respondent and the new information is inexact. In table 2.4, a fictitious example of inferential disclosure (not from the census website) is used where the information revealed may be very useful to a burglar! It would depend on whether these figures are absolute or relative. If the figures are relative to the national proportions, and the figures are higher than national proportions, it would be very disclosive information.

Table 2.4: Example of Inferential Disclosure

	Proportion with a computer	Proportion with a DVD player
Flat	0.16	0.24
Terraced	0.38	0.39
Semi-Detached	0.56	0.40
Detached	0.92	0.75

Cases of disclosure illustrated in the examples above are entirely feasible for England & Wales census data. 2001 Output Areas (OA) contained approximately only 300 people. This implies that firstly, as in tables 2.1 - 2.3, the cell counts from an OA are likely to be very low. Secondly it means that it is possible for an intruder to have a detailed knowledge of the area as it is so small. This makes identification easier. A potential risk of disclosure can easily translate into an actual disclosure.

In this thesis, all tables will have originated from a census rather than a sample. The FCSM discuss how disclosure risk is substantially reduced when a frequency table is based on a sample of data rather than the whole population. If an intruder possesses information about someone and is looking to find a specific individual, the chances may be that the individual is not even represented in the table (depending on the sample size). Secondly, unique cells containing a count of one need not represent respondents with unique characteristics in the population. There may be several other individuals in the population with the same characteristics that did not get chosen in the sample. Data based on a sample gains additional protection. However tables of census data or from samples with a high sampling fraction present a high disclosure risk.

2.3 Spatial Aspects of Disclosure

This section distinguishes disclosure from census data which are spatial in nature. Section 2.3.1 discusses disclosure by differencing, a term used in the literature to describe the disclosure arising from comparison of two or more different outputs. We then focus on geographical or spatial differencing which occurs when these two outputs relate to different geographical zones. Ways that geographical differencing can occur are illustrated with examples in section 2.3.2. A closely linked research field is the topic of geoprivacy which concerns the confidentiality of locational data such as

that typically represented in maps. This research field is briefly described in section 2.3.3, since the confidentiality protection approaches may have applicability to the objectives of this research.

2.3.1 What is Differencing?

Statistical agencies are obligated to protect the data they hold so that an intruder cannot obtain, directly or through inference, knowledge of the confidential data. Disclosure of the confidential data can occur through two means. *Direct disclosure* occurs with unauthorised access, for example, an intruder discovering a password to access the confidential data. However inadvertent (direct) disclosure, termed I.D.D by Fellegi (1972) occurs when the intruder uses legitimately accessible information to uncover confidential information. I.D.D can occur in the form of *residual disclosure* as defined by Fellegi (1972). Residual disclosure occurs inadvertently when two or more data tables, taken together, enable a user to identify information pertaining to individual respondents even though none of the data, taken by itself, is a direct disclosure. In the UK and other European statistical agencies, the term *disclosure by differencing* is more commonly used. The Confidentiality Guidelines specified on the ONS website use this term (e.g. ONS, 2006). Tables that are 'similar' are referred to as being at risk from disclosure by differencing producing sample uniques that could be population uniques. An example is a table containing 10 persons aged 16-20 being potentially disclosive if a similar table shows 9 persons aged 16-19, and hence 1 person aged 20.

In the literature, disclosure by differencing is referred to in different contexts and with different terminology. Statistics Canada and the US Census Bureau commonly use the terms *comparison* and *residual disclosure* in the context of comparing outputs protected by cell suppression (see section 2.4 on disclosure control methods), see for example Robertson (1993). Another example, from the Canada Customs and Revenue Agency (McElroy, 2003) refers to comparison disclosure as occurring when two or more sets of tabulations analyzed together make the identification of confidential information possible, even though none of these tabulations contains a direct disclosure; and residual disclosure as occurring when a blanked out cell in a tabulation can be deduced through the analysis of other component data and/or totals of the tabulation. However the Center for Economic Studies at the US Census Bureau instead often use the term *complementary disclosure* on their website, stating that tabulations can represent a significant risk of disclosure; by combining information from the

released table with other sources of information, it may be possible to infer information on an individual survey respondent.

There is also much discussion of differencing in the computing literature. It is an important concept for disclosure detection in statistical databases but is usually referred to as inferential disclosure. For example, Denning (1980) defines inferential disclosure as the deduction of confidential data by correlating declassified statistical summaries and prior information. An example is given, comparing the mean salary of two groups differing by only a single record which may reveal the salary of the individual whose record is in one group but not the other. Chowdhury et al. (1996) also refer to inferential disclosure when the intruder uses legitimately accessible information to infer confidential information.

In a broad sense, all the definitions have the same meaning in that the disclosure occurs from the comparison of two or more different outputs. The outputs are related in some way, be it by common margins, or the same respondents but different time periods to name two examples. However these definitions do not refer specifically to geographical (spatial) differencing. The term geographical differencing has been used to refer to disclosure specifically occurring when tables are published to different geographical boundaries. Duke-Williams and Rees (1998a; p.580, 1998b) have written extensively about this problem.

> *'... if data are published for more than one geography, then it may be possible to combine data to determine counts for very small areas, which may contain few people. Such cases may lead to a risk of information about individuals being disclosed.'*

For the remainder of this thesis, we refer only to the term **differencing** and take this in its broadest sense (comparison of any two outputs leading to disclosure). We refer to **geographical differencing** to mean only spatial differencing which occurs from the comparison of outputs to different geographical boundaries.

2.3.2 Spatial and Non-Spatial Differencing

In this section, illustrations of differencing are shown focusing on the case of geographical differencing in detail (other examples are shown to indicate what geographical differencing is not). The new methodology will be specifically designed to deal with the spatial nature of disclosure risk, aiming to reduce disclosure risk in geographically differenced slivers and for the small populations arising in small areas. However since the new geographical perturbation methods will be pre-tabular in nature, they should offer in addition, protection against other forms of differencing (because all outputs originate from the pre-confidentialised microdata; see the later section 2.4.2). Thus we present here an expanded discussion of forms of differencing. Tables of counts or frequencies are considered exclusively.

- Linked Tables (non-spatial differencing)

Linked tables are tables derived from the same base data. The tables are 'linked' as they share common attributes. Willenborg and De Waal (2001) discuss how it is possible that a table not containing any sensitive cells can be combined with the information of other non-sensitive tables and still disclose information about a respondent. The following is an example taken from their book:

Tables 2.5: Example of Differencing via Linked Tables

(a) Location of business and sex of self-employed shopkeepers

	Centre	Outskirts	Total
Male	19	6	25
Female	3	6	9
Total	22	12	34

(b) Financial position and sex of self-employed shopkeepers

	Weak	Strong	Total
Male	13	12	25
Female	3	6	9
Total	16	18	34

(c) Location of business and financial position of self-employed shopkeepers

	Weak	Strong	Total
Centre	7	15	22
Outskirts	9	3	12
Total	16	18	34

From these tables, it can be inferred by mathematical reasoning that the financial position of all male self-employed shopkeepers in the outskirts of the town is weak (see workings in Appendix A2.1). Linked tables do not pose too much of a problem when a fixed set of tables are released from a base dataset; all the tables to be released are known so can be treated in combination. However, when a non-fixed set of tables are released, it creates more problems in controlling confidentiality. It is not possible to modify previously released tables, and the agency would have to keep track of every table released and the users who requested them.

Much discussion of this type of disclosure can be found in computer-science related journals. As presented in Fellegi and Phillips (1974), computers have transformed the confidentiality problem. They have become powerful tools that allow users to analyse a variety of statistical information from separate or linked files, possibly collected over long periods of time and typically in a highly disaggregated form. Inferential disclosure, as this form of disclosure is commonly termed in computer science literature, can be defined using a structural approach with methods based on linear programming used to detect potentially disclosive tables created from a database. Statistical agencies, such as the US Census Bureau, which have research data centres or online databases where the user can submit requests for a defined table, commonly use such linear programming methods to monitor user requests. These programs detect users that are asking for many, very similar outputs.

- Special Tabulations (non-spatial differencing)

March and Norris (1987) describe a real example of how differencing from linked tables occurs in Canadian census data. There has been an increasing demand on statistical agencies for data on specific target groups such as ethnic minorities or certain occupational groups. March and Norris (1987) give an example from the Canadian Census of Population and Agriculture. A large majority of Canada's aboriginal population live on Indian Reserves that have been identified as census subdivisions in the standard census geographic hierarchy. Much data are available for these census subdivisions which are a basic tabulation unit. However there is a demand for data to be published concerning only the aboriginal population. There is a small non-aboriginal population living in the reserves and if data are released only for the aboriginal population, it could lead to disclosure by differencing (for the non-aboriginal population) when compared to data already available for census subdivisions.

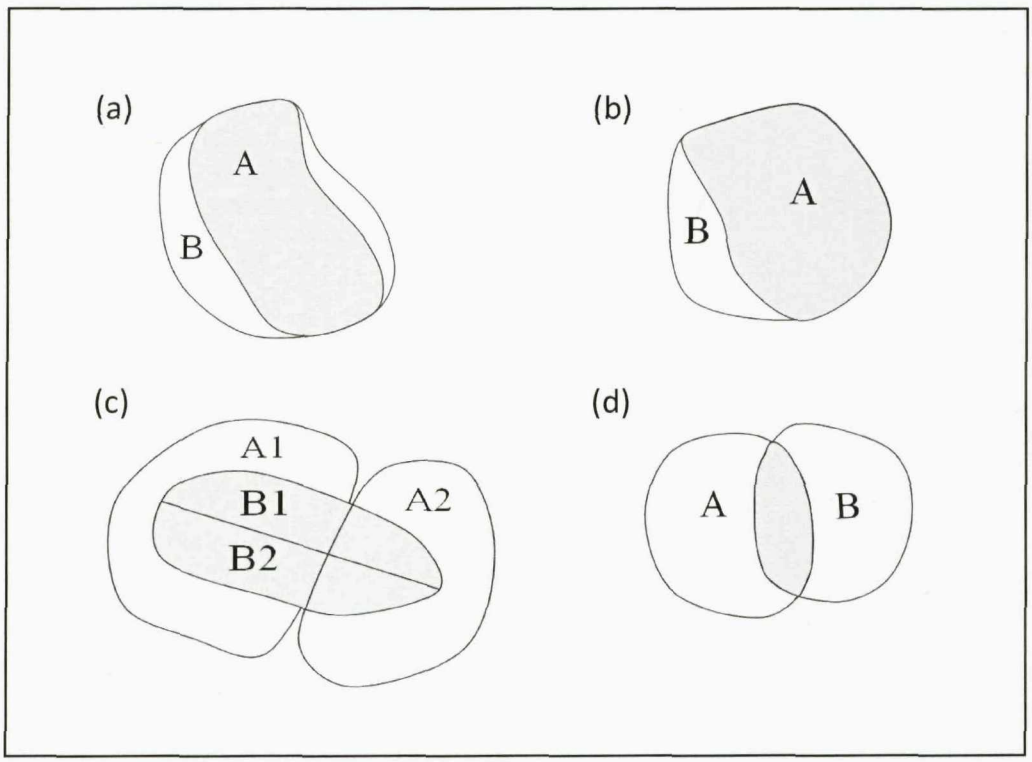
- Semi-linked tables (non-spatial differencing)

Willenborg and De Waal (1996) refer to semi-linked tables. These do not refer to tables that have been produced from the same microdata set, but from different ones that refer to (almost) the same population. In other words, the tables are not entirely independent. The common marginals of semi-linked tables at the population level are the same. A practical situation in which this occurs is when semi-linked tables are in longitudinal or panel surveys that yield information about particular subjects at different points in time. Algranati and Kadane (2004) provide an example of disclosure from semi-linked tables, extracting confidential information from public documents based on the 2000 Department of Justice Report on the federal use of the death penalty in the US (see Appendix A2.2).

- Geographical (Spatial) Differencing

Differencing can also occur when tables are produced for the same area but to different geographical boundaries. This is termed spatial or geographical differencing. Figure 2.2 shows how different output zones may overlap.

Figure 2.2: Ways that geographical output zones may overlap.



In the UK, data from the Census is published as aggregated data in tables to a variety of output zones including postcode sectors, wards and output areas. The structure of the output zones from the UK Census is such that one output zone may be wholly contained within another larger output zone (nesting) as in figure 2.2(a) or 2.2(b). For example some wards with 1998 definition published from the Census were larger than those in 2001. Cases of nesting illustrated can result in geographical differencing. The two tables for the output zones can be *differenced* to produce a *differenced area* sometimes known as a *sliver*. The table for the differenced area could contain very few individuals possibly below the confidentiality threshold as illustrated in the example in chapter 1 (table 1.2). This is especially so if the output zones are very similar in size, and are based on raw counts (as opposed to derived statistics which are difficult to untangle). Furthermore output zones could be aggregated such that in combination they can be differenced from a larger aggregate as in figure 2.2(c).

On the other hand, sections of two different output zones may overlap to produce an intersecting area as in figure 2.2(d) for example some 1991 postal sectors intersected with 1991 EDs. This scenario does not pose a disclosure risk. This is because there is no way of knowing which data units in the two tables belong to the intersecting area (the two tables cannot be differenced as with nesting - Duke-Williams and Rees, 1998b). Estimation techniques based on area or population shares possibly could be used but these methods are generally very inaccurate.

Fellegi (1972) discusses a mathematical theorem for detecting cases of disclosure by differencing in terms of sets of respondents corresponding to each published table. Each tabulation cell can be conceived as corresponding to a set of respondents. Publication sets may overlap to produce intersections and unions. When another set, not previously published, is deduced through manipulation, it also corresponds to a set of respondents. This set must be an intersection of some of the previously published datasets. Fellegi defines an intersection as the smallest mutually exclusive, non-overlapping set which can be created through the union of the publication sets. To deduce whether a [residual] disclosure would occur, he suggests that each intersection must be considered and examined to see 1) if it would be I.D.D. if the corresponding count or aggregate was published and 2) whether it can be isolated through a linear combination of the publication counts. Fellegi goes on to describes how any intersection can be represented as a linear combination of the total respondents in each of the sets and intersections. If [residual] disclosure is possible then the set totals can be manipulated arithmetically to give the exact number of respondents in the residual area. This

can be generalised so that an equation can be written for any intersection given any sets of publication tables. The fundamental part of Fellegi's definition of [residual] disclosure is knowing which respondents fall into each of the publication sets. His theorem can be used to show that if two output zones overlap but one does not nest entirely within the other as in example (d), then it cannot be determined which respondents belong to the intersection (from aggregate data). Thus unless output zones are nested, differencing cannot occur.

Geographical differencing is closely related to the problem of disclosure in small area outputs, because the differenced slivers are essentially tables relating to a small area. Both slivers and small area outputs have the potential to be highly disclosive because of the small populations within them. However it is not known where the slivers will occur unless the entire set of outputs to be released is known in advance. Methods of SDC such as iterative rounding and controlled rounding (see later section 2.4) have been developed in the literature which provide protection against differencing via linking by creating uncertainty around cell values. Rounding usually involves distorting all cell values in the tables. Random record swapping also provides some protection against differencing. However these methods have not been created with geographical differencing specifically in mind and often damage the data too greatly where it is not needed. We revisit these problems later in chapter 3.

2.3.3 Disclosure in Non-Census Spatial Data

A closely related confidentiality problem, arising from the availability of geographically referenced data, is termed geoprivacy and concerns the location of sensitive data at the disaggregate level. While not usually considered in the context of aggregate census data, there is a close link to the central focus of this thesis. *Geoprivacy* is an emerging area of research in the US and refers to an individual's right to deny disclosure of the location of one's home, workplace, daily activity or trips (Kwan et al., 2004).

Spatial analysis of georeferenced data allows geographic researchers to identify important patterns in the data at the disaggregate level, however it is important that the privacy of individuals is protected. Leitner and Curtis (2006) draw a distinction between statistical (attribute) and spatial (locational) confidentiality. Statistical confidentiality is associated with individual information, in GIS terms the equivalent of aspatial attributes, while spatial or locational confidentiality is concerned with the

placement of individual-level statistical information on a map. To date, relatively little has been written about methods to protect the point mapping of individual information. Geoprivacy is especially sensitive in studies of health and crime data. For example, law-enforcement agencies throughout the US provide crime maps (Leitner and Curtis, 2006), while point maps are often published representing cases of cancer or infectious diseases (for example, Zimmerman and Pavlik 2006, Armstrong et al. 1999). Leitner and Curtis (2006) note that an individual's residential location can be easily displayed, potentially leading to identification of the individual and disclosure of confidential information as inverse address matching technology can be used to reveal the street address and residents at a point location (Zimmerman and Pavlik, 2006). In the latter part of section 2.4, we consider some of the confidentiality methods (geomasking) used to protect this type of data, which may be applicable for protecting census data in a flexible tabulation scenario.

2.4 Disclosure Control: Statistical Methods and Procedures

NSIs must control the risk of disclosing sensitive information when releasing statistical outputs to the public. Disclosure control procedures can take two forms: (i) the Statistical Office can restrict and monitor researcher access to the data or (ii) methods of SDC can be applied to the data before its release. The second refers to techniques which either change and modify the data or restrict the detail released. Often both forms of disclosure control (i) and (ii) are implemented. This section will review some of the SDC methods and procedures used by NSIs to protect against disclosure. We will also comment on whether or not the methods offer protection against geographical differencing. Methods of geomasking for locational data will also be described. A summary of how statistical information is released from censuses, by NSIs in the UK and abroad, is first presented particularly focusing on small area data.

2.4.1 Release of Census Data

Before considering different SDC methods, we first discuss how census data are released and to what level of detail. Scandinavian countries such as Norway base their outputs mainly on registration data with only some data collected in the traditional way such as their Housing census. These countries tend to be more relaxed about disclosure risk. The Federal Statistical Office of Germany is another

example of a country that doesn't take a full census. Instead they have a continuous microcensus based on a random sample of 1% of all households. According to Giessing (2005), data from the census are only released to a strictly hierarchical geography so the differencing problem does not occur.

Countries which take a census of population include the UK, the US, Canada, Australia and New Zealand. Canada and the US have traditionally taken both long and short form censuses where the short form asks a limited number of questions on marital status, age and sex whereas the long form is given to much fewer households (one in five for Canada) and is much more detailed. In contrast, the UK, Australia and New Zealand give a full, detailed questionnaire to the entire population. Consequently the disclosure issues are most significant for these three NSIs.

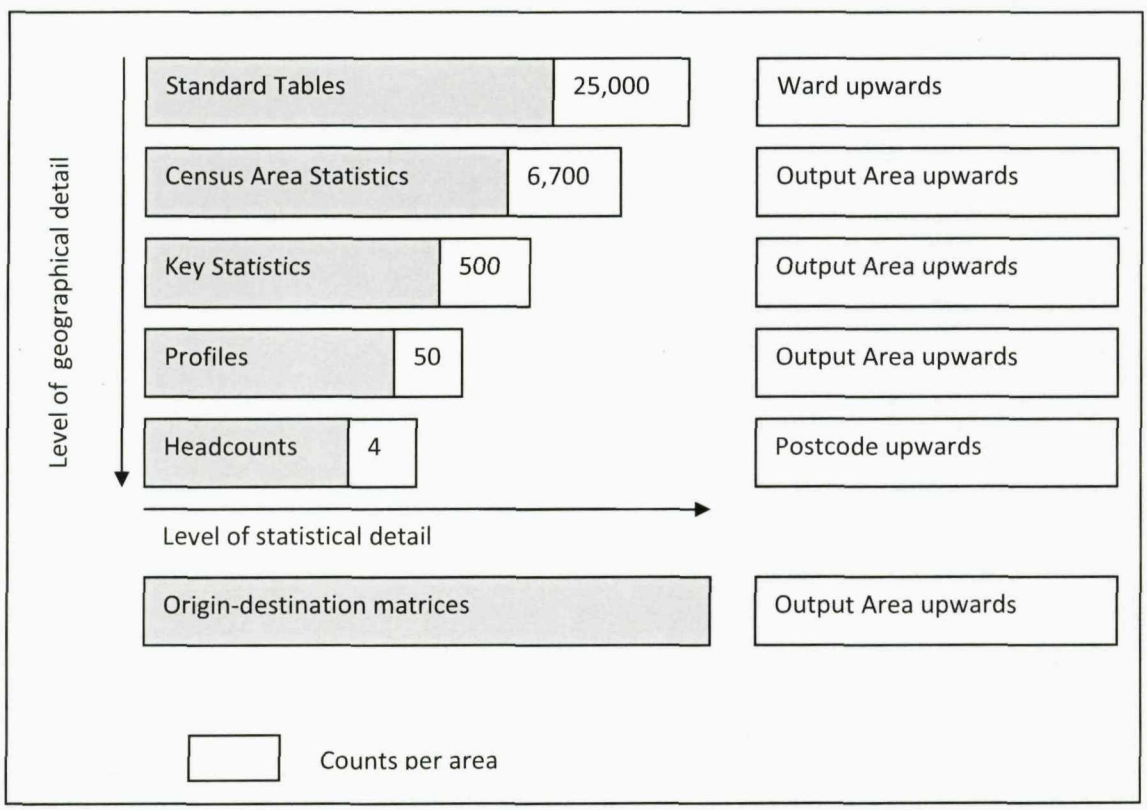
The most recent censuses in the UK were in 2001. In England & Wales the Census was carried out by the Office for National Statistics (ONS). Separate simultaneous censuses were conducted by the government census agencies in Scotland by the General Register Office Scotland (GROS) and in Northern Ireland by the Northern Ireland Statistics and Research Agency (NISRA). Methods of disclosure control applied to England & Wales and Northern Ireland census data were more severe than Scotland; GROS did not apply as much protection to their tabular output.

In countries which take a census, data are generally only released to a fixed set of geographies and more recently many NSIs have started to publish data down to small area level. The main output releases from the 2001 England & Wales Census were Key Statistics, Census Area Statistics and Standard Tables as shown in figure 2.3 (ONS, 2002). Key Statistics give an overview of the Census results for all geographies down to OA level. Census Area Statistics (CAS) are more detailed with counts generally presented as cross-tabulations and including simple univariate tables covering all Census topics available also down to OA level. At the top of the hierarchy there are Standard Tables which are more comprehensive than CAS with detailed cross-tabulations down to ward level. At the bottom, there are headcounts which are the number of people and households, followed by profiles which are standard 'templates' presented as percentages in a limited number of simple tables. The smaller outputs nest into the larger ones above. The origin-destination matrices provide statistics on migration and travel to work. Travel to work matrices show the flows of people travelling to work i.e. within the ward or within the district for example. These matrices can be affected quite badly by

disclosure control because the population can be sparse at OA level but another important reason is that these tables are linked (see section 2.3.2) to other census tables, for example to numbers of people in employment by geographical area.

It is the CAS which are potentially the most disclosive as they contain detailed information at the smallest level. They consist of approximately 30,000 cross tabulated counts for the constituent areas of Great Britain and Northern Ireland with information about households and individuals on age, gender, occupation, qualifications, ethnicity, social class, employment, family structure, amenities and tenure. The CAS are designed for analysis at small area levels such as neighbourhoods and are the equivalent of 1991 census Small Area Statistics. OAs are the smallest unit for which census data are published whereas EDs are used for collection only. OAs can be aggregated to certain geographical zones such as wards, local authorities and health authorities.

Figure 2.3: Standard Area Statistics (England & Wales).



To prevent geographical differencing, the current solution in the UK is to provide information for low population blocks (OAs) which are protected against disclosure, and assemble as necessary. If the user wants to compile statistics for areas which do not correspond to aggregations of the OAs then a synthetic estimate will have to be found. The quality of the estimates depends on how good the synthetic estimation model is; if the areas are not homogeneous then it is harder to produce a good estimate. Much work has been carried out on synthetic estimation for small areas (see Rao (2003) for example).

In general, most NSIs such as the US Census Bureau apply thresholds to determine the minimum size of areas as well as limiting the detail in classifications used in tables. Most relevant to this discussion are probably the most recent censuses as more small area data are being published with each new census. The Australian Bureau of Statistics (ABS) recently took a census in 2006. In the past, the lowest level of aggregation has been Census Collector District level, containing 250-300 households. However for the first time, data for finer geographic building blocks is being released called mesh blocks containing 30-50 households. The concept of using mesh blocks has only been feasible due to advances in GIS technology and improved access to digital topographic data (the mesh blocks will be independent of the census collection methodology). For the 2006 Census, only very limited information is being released for mesh blocks as a trial for full implementation with the 2011 Census[7]. In addition, a table builder service can be used for a fee, giving users remote access to the complete Census Unit Record file and allowing for the extraction and manipulation of an unlimited number of Census tables. A cell perturbation method of disclosure control is applied to the data during delivery that provides protection against geographical differencing. This method is described in section 2.4.2.

At Statistics New Zealand, mesh blocks are the smallest geographic unit (Census 2006) with roughly 100 households in each on average. Currently requested tables for a non-standard area are monitored to see if any cases of differencing occur. In the UK the smallest output level is Output Area (OA). In England & Wales, the minimum size is 40 households and 100 persons but recommended size of 125 households. In Northern Ireland, the sizes are roughly the same but in Scotland, the minimum

[7] Details can be found on the ABS website: *http://www.abs.gov.au/websitedbs/d3310114.nsf/Home/census*

size is 20 households, 50 persons but with target size of 50 households[8]. These thresholds are set dependent on the sparseness of the population in the output zones.

One way to prevent disclosure of confidential information is to control who has access to the data. This can be done by allowing researchers access to census data within a *safe setting*. Usually this would involve verifying researchers' jobs and research objectives as they do in Statistics Netherlands. There the researcher must sign a confidentiality pledge and are then trusted not to misuse the data. In addition, researchers are monitored in case they request anything unusual that might allow disclosure of confidential information (DeWolf, 2005). At Statistics Canada, researchers must take an Oath of Office to become a 'deemed employee of Statistics Canada' before obtaining entry to their research data centre. This oath has legal implications if violated.

Safe settings can also take the form of remote access or special licensing agreements and are useful in disseminating statistical data to a limited audience because disclosure control methods need not be so stringent and thus the data quality is not too badly damaged. However this would not be practical on a large scale so SDC methods play an important role in protecting confidentiality when releasing data to the general public.

Remote access refers to online access of protected microdata. Similar to the ABS table builder, the American Fact Finder has been developed in the US to allow access to frequency count data from the Census 2000. One part of the American Fact Finder is the Advance Query System which has been created such that tables are generated from the sample data and weighted up. Tables generated for more than one geographic area must pass through a filter and all tables must meet certain conditions such as the minimum population requirement to be permitted.

[8] Information taken from the UK Census geography website:
http://www.statistics.gov.uk/geography/census_geog.asp

2.4.2 SDC Methods

SDC methods are used to protect census data before release and can be categorised into *pre-tabulation* and *post-tabulation* methods. Alternatively methods can be classed into (a) methods which alter the data versus (b) methods that restrict the amount of detail in the information released. We will examine at the different classes of method here starting with post-tabulation methods. In addition, we discuss whether the methods offer protection against geographical differencing.

Post-tabulation Methods

Post-tabulation methods are applied to tables rather than to the underlying microdata. Post-tabulation methods in class (b) which restrict the amount of data released, include **cell suppression** and **recoding** (see Schulte Nordholt, 2001 for example). Cell suppression entails withholding certain values in a table. Recoding or Table-Redesign involves re-grouping continuous variables, for example age might be broad-banded into ten year age groups thus reducing detail (and usually increasing cell frequencies). Post-tabulation methods in class (a) which alter the table cell values, are known as cell perturbation. Both classes of methods are thought to offer limited protection against differencing in general because if the methods are not applied consistently they can be unpicked. For example if two tables have suppression applied but to a different set of cells, the tables could be compared and the original values discovered (differencing via linked tables, see section 2.3.2). Cell perturbation methods on the other hand typically introduce uncertainty around the true cell value. However if the same cell information is published many times with perturbation applied independently then it may be possible to narrow the uncertainty interval until the original value is obtained. We will now briefly review some cell perturbation methods.

Barnardisation was the post-tabulation method applied in the 1991 UK Censuses (up until 2001, only post-tabulation methods were applied). Barnardisation (Willenborg and De Waal, 1996) involves adding or subtracting a value of one randomly from some cells in the table. The idea is to create uncertainty around the true cell value. To give added protection, this method could be applied more than once. **Small Cell Adjustment** is a similar method applied in the 2001 England & Wales Census. Small cell adjustment modifies the small cells of the table only but for confidentiality reasons, the definition of a small cell count is not known. The small cell counts might be rounded down to 0 or up

to base 3 or base 5. Since only the small cell counts are affected, it would still be possible to geographically difference large cell counts (if such tables were published) resulting in tables containing previously unreleased small counts. When small cell adjustment is applied, totals and subtotals are calculated from adjusted data, thus ensuring consistency. However, the same totals appearing in different tables may be different.

Small cell adjustment can be thought of as a variant of **Rounding** (small cell adjustment involves rounding only the small cells). Rounding is often seen as a confidentiality measure for frequency tables, see Heldal (2003) for example. Random Rounding involves making a random decision as to whether the cell value will be rounded up or down and applies to all cells of the table. Random rounding was initially considered as an alternative method to small cell adjustment for the 2001 England & Wales Census but was disliked by users due to greater data distortion. Brown (2003) notes that although random rounding can give protection, in some cases the rounded figures can be unpicked. This may occur when areas from one boundary set are fully contained within areas from the other boundary set. The risk occurs partly because the smaller frequency might be rounded up and the larger frequency rounded down which the author refers to as 'contrary rounding'. An alternative is Fixed Rounding (or conventional rounding) to the nearest multiple of a base. This method prevents contrary rounding as described above, but it is less safe as the rounded table can be unpicked by exploiting the fact that rounded values could only have been a finite, narrow set of original values (see Armitage and Brown, 2003) for more details). A final possibility is Controlled Rounding proposed for three-dimensional (and higher) tables by Fischetti and Salazar (1998) which rounds all cell values in such a way that the table is additive and minimises damage to the data. Neither a formal assessment nor a theoretical study has been carried out to determine whether controlled rounding protects against disclosure by differencing, but practical examples give evidence to suggest that it does provide protection in the majority of cases.

The Canadian Census of Population has used an algorithm called Iterative Rounding to protect large frequency tables developed by Boudreau (2005). Every time a table is submitted, a differently rounded table is generated. This protects against differencing because of the random nature in which the cell values are rounded. It produces different rounded tables for tables based on similar overlapping areas and an intruder is unable to deduce any information because the pattern of the subtraction for the residual area will be random. At least 100 submissions would be required to get

precise intervals around cell values that lead to disclosure. Furthermore the Census confidentiality branch says that rounding works well for almost all cases of differencing overlapping regions. However it does not eliminate all cases of potential disclosure and suppression is still applied in some cases. A further disadvantage of this method is explaining to users that tables are different every time they are submitted.

A recent methodology designed for use with the 2006 ABS remote access Table Builder is **ABS cell perturbation** (Fraser and Wooton, 2005). This method works by assigning each record in the microdata a random number called a record key, which is kept highly confidential. When a table is produced from the microdata, the records composing each cell have their record keys summed together according to a special function to give a cell key. The perturbation added to the cell is then read from a look-up table which has the original cell value on its rows and possible cell key values on its columns. The look-up table is flexible and designed according to the specifications of the NSI but the cell perturbations are always dependent on the original cell size. This method typifies the problems encountered with post-tabulation methods in that it is very difficult to achieve both consistency and additivity. ABS cell perturbation ensures that the same cell values in different tables have the same perturbation added. However a further perturbation must be added to restore additivity which, even though it may be small, results in consistency being lost in some cases.

One advantage of post-tabulation methods such as rounding is their transparency; a user can see that a method of disclosure control has been applied. An element of the disclosure control problem is the perception of disclosure and whether the intruder believes an apparent disclosure is an actual disclosure. Thus rounding is very good for this purpose. Furthermore the damage to each cell frequency is often easily measurable: with rounding to base 3, the damage to each cell must be between minus two and plus two.

The disadvantages of post-tabulation methods are substantial - applying protection to each table is time-consuming and cumbersome. If flexible geographies are required for user-defined areas, this becomes more of a problem. Pre-tabulation methods eliminate this problem as well as the need to check every table for disclosure risk (unless the post-tabulation method inherently protects against differencing). Furthermore many post-tabulation methods create inconsistencies between tables at different levels of aggregation as the totals may not add up. If the totals do add up, then the same cell

in different tables is likely to be different, leading the possibility of narrowing the uncertainty interval. Post-tabulation methods in general are not well-suited to the demands of modern users who want flexible geographies. The main advantage of pre-tabulation methods is that they only need to be applied once and so the flexibility of outputs is maximised.

Pre-tabulation Methods

Pre-tabulation methods are applied to the microdata before it is aggregated into tables. Pre-tabulation methods commonly fall into class (b), that is they modify the data rather than restrict the detail released.

Imputation is a method used on census data for supplying the information for missing households as described in ONS (2001). However it can be seen as a form of disclosure control since there is uncertainty around the data values imputed; it is not known whether the values are true or false. The UK census in 2001 was referred to as the 'One Number Census' because a method of imputation was applied to adjust for under-enumeration. A Controlled Donor Imputation System (Steele et al., 2002) was used to impute individuals and households estimated to have been missed in the census. Donor households are selected at random from among households with the same weight where weights are based on the characteristics of existing households in the census database. The selected household is then used to supply all the variables to the imputed household. Imputation naturally offers some protection against disclosure (in addition to the donor imputation process for missing variables due to non-response[9]). In fact imputation has also been developed specifically as a disclosure control method. Over-imputation (Shlomo, 2005b) involves randomly deleting a percentage of selected records and having certain variables erased; standard imputation methods are then applied by selecting donors matching on control variables. Statistics Finland currently use a method of missing data imputation for protecting data relating to small areas according to Tammilehto-Luode (2001). In conjunction with the University of Jyvaskyla, Statistics Finland have developed a new improved method called 'Local Restricted Imputation'. This is supposed to do less damage than the previous method of missing data imputation and was developed in response for the demand for increased

[9]Further details at:
http://www.statistics.gov.uk/about/data/methodology/general_methodology/dataediting.asp

information at different levels of geography. Local Restricted Imputation developed by Markkula (2003) is a method that imputes new variable values for the original values in risk areas. The imputation is done locally only to a restricted set of units. The actual imputation can be done by imputing a mean, by random perturbation or by replacing with a sum-based value.

Imputation is generally not recommended as a disclosure control method as it can have the effect of attenuating the underlying relationships in the data. This is because imputation methods commonly impute values using the existing data thus making the data more homogeneous: and any biases in the data are exaggerated.

Record swapping is an alternative pre-tabular SDC method for microdata that involves swapping the values of the geographic variables for records that match on a key. Record swapping was first attempted on US census data. In the US a full census is conducted called the short form which asks a restricted number of questions (Zayatz, 2006) on sex, age, whether hispanic/non-hispanic, race, relationship to householder and tenure. The procedure for protecting the short form 100% data has been the confidentiality edit, otherwise known as record swapping. This involves swapping a small sample of households with data from other households that have identical characteristics on key variables but are from different geographic locations. Which households are swapped is not public information.

For the 1990 US Census, key variables for matching in the confidentiality edit were: number of people in the household of each race, whether hispanic/non hispanic, age group (<18, 18+), number of units in building, rent/ value of home and tenure. Furthermore a higher percentage of records were swapped in small blocks because those records pose a higher disclosure risk. The average block contained 34 people. Many tables were published at block level but some more detailed tables were published at block group level (average of 1,348 people).

In 2000 the confidentiality edit applied to the short form was targeted. This meant that only records that were unique on a set of key variables were swapped. The swapping rate of uniques was again higher in small blocks. In addition, a very large percentage of households containing a race category not anywhere else in the block were swapped and protection already provided by imputation was taken into account. Pairs of variables that were swapped were matched on a minimal set of variables.

Finally any table released had to have at least a minimum number of people of a given race in a given geographic area for the table to be released.

In addition, the US has carried out a long form survey which asks much more detailed questions but to only a sample (1 in 6) of people which in itself provides some protection against disclosure. The lowest level published was block group level. Swapping was also used as a protection method, consistent with the 100% procedure. However the plans for the 2010 census are to replace the long form with the American Community Survey (ACS)[10].

Special tabulations were generated from the swapped data files from the US census data. The confidentiality edit could not be relied on to give full protection so cell values were rounded. This meant that tables were no longer additive. Disclosure by differencing is also an issue in the US and the combination of disclosure control methods described were used to protect against potential risk.

In 2001, in contrast to the previous UK Censuses, a method of record swapping was also used based on the ideas in the US. The method of RRS is described in Boyd and Vickers (1999) and Shlomo (2005a). The idea was to select a percentage of records at random and to swap them with similar records in other Output Areas within the same Local Authority District (LAD). A random sample within strata defined by control variables was selected using a fixed swapping rate. The control variables (key) were hard-to-count index (associated with under-enumeration), household size, sex and broad age distribution of the household (0-25, 25-44, 45 and over). For each household, a paired household is found and all geographical variables are swapped. The key acts as a control so that the marginal distributions of these variables are not distorted. The precise percentage of records swapped was not released to the public. However in the paper by Boyd and Vickers (1999) they analyse distortion after swapping a hypothetical 1%, 3%, 5%, 10% and 20% of records.

Some recent research work described in a paper by Karr (2006) discusses the potential of a new record swapping method called 'doubly random swapping'. One problem often associated with RRS is that it attenuates the relationships between swapped and unswapped attributes. The idea of doubly

[10] The ACS is an ongoing statistical survey giving more current information and replacing the burden of a long form every ten years. It is planned to be fully implemented by 2010.

random swapping is to randomise both records and attributes that are swapped so that all relationships are attenuated uniformly, better for model selection and fitting.

A very different alternative pre-tabulation method is to create **synthetic data** based on the properties of the real census. Abowd and Woodcook (2004) have worked on multiple imputation to create synthetic data by estimating the joint probability distributions of all data using generalised linear models to model interdependencies between variables. A disadvantage of this approach is that the data are unlikely to be accurate for a full census dataset because of the complexities of the relationships which have to be modelled between the many census variables. Statistics New Zealand are currently using synthetic data called 'SURF for Schools[11]' which is based on their Income Survey. In this context synthetic data may be appropriate, since all the intricacies of a real census are not necessary.

Pre-tabulation methods are generally preferred for a variety of reasons. Swapping adds uncertainty to the data, for example if a cell value of one occurs in the perturbed table, then it is not known whether it is a genuine one or a swapped one. There isn't the potential to narrow the uncertainty interval around a cell because the tables all come from the same population base. Thus tables also add up. The major problem of obtaining both consistency (same cells in different tables taking different values) and additivity is avoided. Whether a pre-tabulation provides enough protection depends on the approach and level of perturbation to the microdata, e.g. taking into consideration the proportion of records swapped. This can be a downside of pre-tabulation methods because without knowing the set of tables to publish in advance, it can be difficult to ascertain whether sufficient protection has been provided. This contrasts with post-tabular methods where protection can be applied purposely to each cell. Often pre-tabular methods rely partly on the perceived uncertainty created. Moreover when a high level of pre-tabular perturbation has been applied, the distortion to utility can be high (for example correlations between variables can be damaged) and difficult for users to measure. An alternative is to apply a pre-tabulation method followed by a post-tabulation method as with the England & Wales Census 2001 (RRS and small cell adjustment). In this way, a balance may be more easily achieved between disclosure risk and data utility.

[11] *http://www.stats.govt.nz/schools-corner/secondary/teachers/surf-for-schools/default.htm*

Origin-destination tables (flows of migrants and journey-to-work) present a particular disclosure control challenge because these tables are linked to other census tables, for example flows on journeys to work relate to the workplace population. Application of disclosure control can lead to inconsistencies in linked tables and particularly with pre-tabular methods such as swapping which modifies the spatial relationships in the data. For example, a high number of records swapped could lead to inconsistencies such as a person who travels by bicycle to work over 60km. In the England & Wales 2001 Census, geographical variables such as workplace were not swapped. However this can lead to the possibility of differencing particularly at the OA level, via the linked tables, or at the very least inconsistent flows when comparing between the tables.

The origin-destination tables could be protected separately by an alternative method, or accessed within a safe-setting or under special licence. In this thesis we are concerned with the general problem of disclosure from small area data and flexibly aggregated outputs so little attention will be given to the special case of origin-destination tables. However as with any SDC method, it is important to consider all outputs as a whole if it is to be considered a realistic method to be applied to a census.

2.4.3 Geomasking for Spatial Data

In section 2.3.3 disclosure risk in non-census spatial data was considered. The conventional approach to preserving spatial confidentiality in locational data has been to adopt the same methodology as for census data, that is to aggregate records across populations large enough to ensure prevention of disclosure (Armstrong et al. 1999). However, aggregation damages the data, making research into causation with associated factors very difficult (Leitner and Curtis, 2006). Armstrong et al. (1999) introduced the term *geographical masking* (*geomasking*) for the modification of geographical coordinates to protect confidentiality. Methods include affine transformation and random perturbation. Affine transformations relocate each point by change of scale, rotation, flipping or some concatenation of these masks. Random perturbation or *jittering* involves adding noise to original locations. According to Armstrong et al. (1999), random perturbation is an effective geomasking technique, to some extent superior to affine and aggregation masks. Kwan et al. (2004) have assessed the spatial masks discussed in Armstrong et al. (1999), particularly levels of random perturbation in relationship to disclosure risk. Since mapped locations of disease or crime contain

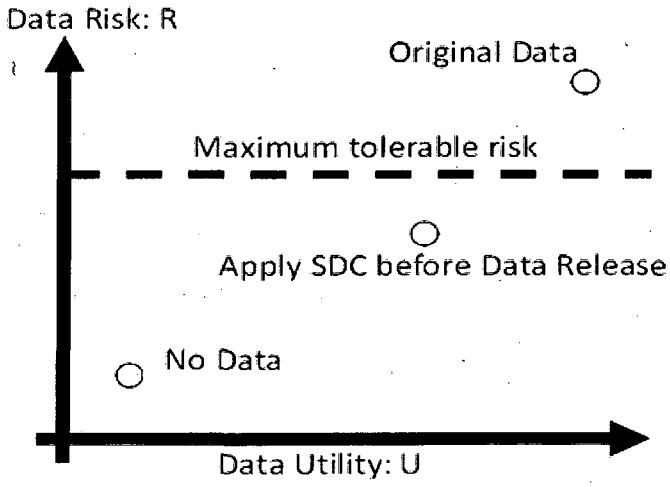
wide variation in population density, the amount of noise added to location can be allowed to vary with population density. The idea of encompassing population density into the disclosure risk model has also been discussed by VanWey et al. (2005) who simulated a sampling frame of public schools in the US. Their data contained the geographical location of each school with potentially sensitive attribute information. A solution was proposed whereby map symbol size was adjusted to cover multiple schools, providing locational uncertainty in proportion to a specified level of identification risk. For schools in large cities a much smaller point buffer was needed than in remote rural areas.

The research on geomasking methods again illustrates the special nature of the relationship between geography and statistical data. Gutmann and Stern (2007) focus on the challenges when precise spatial data are linked to confidential information. The report finds that several technical approaches for making data available while limiting risk have potential, but none is adequate on its own or in combination. In chapter 3, we review in detail some geomasking methods for potential application to census data. In particular the idea of incorporating population density into a census SDC method is considered.

2.5 Balancing Risk and Utility

All methods of SDC (or geomasks) damage the data to some extent. Thus the disclosure control dilemma involves finding a balance between disclosure risk and data utility. Duncan et al. (2001) summarised this neatly in a risk-utility framework as shown in figure 2.4. Release strategies are represented by points on the two-dimensional diagram.

Figure 2.4: The Risk-Utility Framework (Duncan et. al, 2001)



The dilemma is viewed as a decision problem where the optimal methods are determined by minimising the disclosure risk while maximising the utility of the data. R is a quantitative measure of disclosure risk and U is a quantitative measure of data utility. The original data has maximum utility and high disclosure risk and the aim of the SDC approach is to reduce this below some tolerable level. At the extreme this leads to no data being released at all. The optimal release will have enough SDC applied to be below the tolerable risk level but still retain data utility. This approach to SDC will be used throughout the thesis for comparing new methods. The remainder of this chapter considers the assessment of disclosure risk in protected data (section 2.5.1) and utility of the data (sections 2.5.2-2.5.4).

2.5.1 Practical Assessment of Disclosure Risk

Assessing disclosure risk in frequency tables is generally quite straightforward and following the principles of the earlier sections in this chapter, risk may arise from:

- Cell counts of one (cell uniques)
- All zeros in a row or column
- The majority of counts falling in just a few of the row or column cells
- Small counts in the margins

As a general rule, most statistical agencies do not publish cell counts of less than about five in tables from a census or high sampling fraction dataset. In a frequency table, the occurrence of small values is usually taken to present the possibility of a disclosure risk because of the risk of matching against other databases by an intruder. The FCSM (1994) mention that a cell is defined to be sensitive if the number of respondents is less than some specified threshold (well known as the threshold rule). Some agencies require at least five respondents, others require three. More specifically, the FCSM discuss the rules for limiting disclosure risk for frequency tables. As expected, these rules differ from agency to agency and from table to table. One approach is to not publish tables where the cell is equal to a marginal total (relating to attribute and identity disclosure). Another approach is to not publish cells which could allow a user to determine an individual's age within a five year interval, earnings within a specified interval or benefits within a specified interval[12] (relating to inferential disclosure).

The difficulty in assessing disclosure in frequency tables generally comes from the large numbers of tabular outputs to examine and the many ways in which they can be differenced. This would require use of some automated rules. In chapter 3, we discuss some new ideas on assessing disclosure risk after geographical perturbation has been applied, with a focus on cell uniques.

[12] This is called recoding and is described in section 2.4.2.

2.5.2 The Utility of Census Data

All SDC methods damage the data and distort statistical relationships. An appropriate disclosure control method should therefore minimise distortion of the data in the context of user needs. Measuring utility is often thought of as more complex than measuring risk, because the possibilities for user analysis are endless. The remainder of this chapter concentrates on the utility of census data, first looking at user demand for flexible outputs and then briefly reviewing the ways census users analyse the data. When developing and assessing the new SDC methods in chapters 3 and 5, the uses of the data need to be kept in mind, and in chapter 6 we consider the impact on utility explicitly in regard to some of these methods.

2.5.3 User Demand for Flexible Outputs

As Martin (2000) discusses, many census users have interests in geographical areas which cannot be neatly assembled from the statutory published areas. The ability to provide multiple geographies other than those provided, would be very helpful to many users. Four main geographies are most requested: local government require data from small areas fitting exactly into administrative units to be able to assign grants. The NHS and businesses want small area data based on exact aggregations of unit postcodes so they can link to data in non-census databases. Environmental data are also required for small areas that are regular spaces defined by the National Grid system of co-ordinates, convenient for use in simulation models. Lastly, to analyse population over time, small areas that match those used in previous censuses are essential. A flexible tabulation system is needed, not only to meet user needs as described but for revising the standard geographies which occur during the intercensus period, for example, new policy-related administrative areas may be developed by the Boundary Commission.

In general, users have expressed a strong desire for being able to develop their own tables from the Census or for being able to carry out their own statistical analysis using individual or household records. This demand has been met partly through release of the SARs but as mentioned, is only a sample of the full Census. Currently customised tables can be requested from the UK Censuses but this is a very slow process and can be expensive and this option has been used sparingly by academic

researchers. Various user consultations on the UK Censuses (both for 1991 and 2001) have stressed user preference for access to flexible geographies. Alternatively very small building blocks could be used, such as postcodes, to enable users to build their own customised outputs. This would follow the same goal as the ABS Census whereby the aim is to release very small mesh blocks as a building block. UK postcodes contain around 30 households on average.

Martin (2000) examines how digital boundaries became more widely used in the late 1980s so that by 1991 there was the first full national coverage of the UK right down to the smallest census areas. The concern of disclosure by geographical differencing for the UK Census is largely due to the explosion of Geographical Information Systems[13] (GIS) and use of digital boundaries. If coordinate information is available for areas, it is possible to use a GIS for each area to compare the two sets of boundaries. It is possible to 'overlay' the user specified output geography for the region of interest with the standard output geographies for the same region of interest using GIS. For the polygon fragments generated from the overlaid geographies, a disclosure risk assessment can be applied to identify the disclosive polygons. However this assessment is greatly dependent on the accuracy of the digitisation. If the digitisation is not accurate the GIS program will not pick up the fully contained ward and may pick up areas whose boundaries have not changed at all. The effect of inaccurate boundaries is that if two tables are released on both boundary sets (one being inaccurate), artificial differences may appear in the statistics. This is misleading since artificial cases of disclosure may appear. It is worth mentioning that one consequence is that errors in the digitisation process can inadvertently provide some kind of protection against disclosure.

2.5.4 How do Census Users Analyse Small Area Data?

In this final section we summarise how census users analyse the data focusing on information available at the small area level as this presents the greatest potential for disclosure. In chapter 6, we pick a selection of these methods to examine in detail to assess the new SDC methodology.

[13] A GIS is a computer system used to displaying and analyse geographically referenced information. It allows the user to combine spatial and aspatial information from different sources.

Descriptive Statistics – basic counts and rates

Often census users are interested in **descriptive statistics** reporting area profile information at the small area level. Several websites provide information in the form of counts and rates, some of these being commercial websites*. They provide statistics down to postcode level on property prices, crime along with other census related data on affluence (car and property ownership), age distributions and education level of residents. The Neighbourhood Statistics website provides information down to Super Output Area level using various government datasets including the census. Typical summary statistics presented include totals (e.g. numbers of motor vehicle offences), averages (e.g. average price of a detached house), percentages (e.g. percentages of people aged 0-19) and mode (e.g. most popular type of housing). Other descriptive statistics which users might be interested in are skewness of a variable, finding the median and inter-quartile range, or rankings within a variable (such as ranking of wards for percentage of elderly residents).

Measures of Social Homogeneity

Measures of social homogeneity are very useful in describing small areas. A simplistic approach to **measuring homogeneity** was used by Morphet (1993) who wanted to assess EDs to see if they were suitable for explaining patterns of socio-economic variation. In this study, an ED was regarded as homogeneous if 0%, 100% or 95% (almost homogeneous) of households in the ED have the characteristic (such as 100% of households being rented from local authority). A new measure of homogeneity was given by Martin (1998) based on maximising the uniformity of households falling within different categories of tenure within OAs. Moreover, Tranmer and Steel (1998) discuss the use of a statistic termed the intra-area correlation or IAC. This measures the similarity of values of variables within any area of interest. For example, if the variable is tenure, then the intra-area correlation measures the similarity of the values of this variable for each household in each OA. Any value above 0.05 implies a reasonable degree of homogeneity. Suppose tenure has three categories. For each category a measure of IAC can be calculated:

$$IAC_c = \frac{\frac{1}{V-1}\sum_{v=1}^{V} N_v (P_{cv} - P_c)^2}{\left(\overline{N}^* - 1\right) P_c (1 - P_c)} \qquad (2.3)$$

where: $\overline{N}^{*}$ is the (adjusted) mean population size of the V areal units,

N_v is the population size of areal unit v,

P_c is the overall proportion of the population in category c

and P_{cv} is the proportion in category c in areal unit v

This formula gives an approximate ratio of the area level variance to the household level variance divided by the mean population size. An overall measure of the IAC can then be calculated across all categories of tenure:

$$IAC = \frac{1}{C-1}\sum_{c=1}^{C}(1-P_c)IAC_c \qquad (2.4)$$

where C is the total number of categories in tenure (or the variable of interest). Essentially the *IAC* compares proportions of *tenure* categories with the national averages for each category, across all areas to arrive at a measure of homogeneity.

A measure of **spatial autocorrelation** can also be used to assess homogeneity in terms of how similar the value of an attribute in one location depends on the values of the attribute in nearby locations (Fotheringham et al, 2002, 1998). Spatial autocorrelation measures this dependency by examining the correlation between an area and its surrounding neighbours. If there is any systematic pattern in the spatial distribution of a variable, then it is said to be spatially autocorrelated. The impact of geographical perturbation on spatial autocorrelation is studied in detail in chapter 6.

Spatial Exploratory Approaches

A third approach to analysis of univariate data is to **map the spatial distributions** of variables, made possible using modern GIS software. Hirschfield and Bowers (1997) produced maps to show the spatial distribution of the 'have-not' geodemographic lifestyle classification. The local government of London investigated the concentration of late-night economies to identify hot-spots of late night activities. Barnett et al. (2002) created maps to compare the standardised residuals from a simple regression model plotted against urban, rural and fringe areas to see if there was any remaining unexplained geographical variation.

Spatial Rankings

Spatial rankings may be used to define relative positions of a geography e.g. wards with respect to a particular attribute (e.g. areas with the highest average income and areas with the lowest). Mapping of spatial ranks allows easy identification of high and low ranking areas in terms of poverty, income or health for example.

Spatial Processes

More complex analysis of spatial processes in a univariate context might involve **area interpolation** which is a process whereby data from one set of source polygons are redistributed onto a set of target polygons. Surface Modelling is a particular technique of area interpolation which models census populations and their characteristics as first outlined in Martin (1989) and Bracken and Martin (1989). The census variable is represented as a continuous mathematical function giving the appearance of a terrain with high values corresponding to peaks and low values corresponding to valleys. Population density can be modelled as a surface using population counts. Centroids are found for each ED and the population redistributed around the centroid. The extent of spread is determined by the local density of centroids through a distance decay function. Counts of individual and household population are needed for each ED. The centroids represent local summary locations for the distribution of the population. The surface modelling technique may be applied to any count data present at zone-centroid locations. For example, household counts have been used so that the output model is a household-density surface. Wu and Martin (2002) applied the surface modelling technique to the population total for each zone-centroid. Mesev (1998) spatially manipulated census data to determine the location of residential land use where image data were not available.

Another important aspect of a dataset to census users is how close events are to each other. For example, do people of certain age groups congregate in certain areas? These relationships can be examined by using semi-variograms and nearest-neighbour plots (see Birkin et al. 1995). These plots are based on the relationship between squared differences in value (squared to ignore direction) and geographic distance for pairs of observations in the data. Spatial autocorrelation is a further way of accounting for **spatial dependency** in a dataset (Fotheringham et al. (1996) and is explored further in chapter 6.

A very important use of census data distinct from the previously mentioned uses, is to **analyse change** between datasets from different time periods. Champion 's (1994) study on population change and migration in Britain used Small Area Statistics to identify trends within as well as across intercensal periods. This analysis was carried out for different types of area (such as urban or rural), at regional level and also for districts within regions. For example, the population change in rural districts was used to distinguish between districts of decline and districts of growth.

Scatterplots and Correlations

Univariate and bivariate methods of analysis are generally used as diagnostic tests to get a preliminary assessment of the data before continuing with more complex multivariate analyses. Spearman's rank correlation coefficient was used by Christie and Fone (2003) to assess the relationship between car ownership and seven other census variables. EDs were first stratified into deciles of population density within ONS urban/rural classifications and then **correlations** were calculated. Pearson's product-moment coefficient can be used when the variables are numerical and continuous. The correlation $\rho_{z_j z_k}$ between two random variables z_j and z_k for variables $z_j \neq z_k$ is:

$$\rho_{z_j z_k} = \frac{\sum z_j z_k - \frac{\sum z_j \sum z_k}{N}}{\sqrt{\left(\sum z_j^2 - \frac{\left(\sum z_j\right)^2}{N}\right)\left(\sum z_k^2 - \frac{\left(\sum z_k\right)^2}{N}\right)}} \qquad (2.5)$$

Another example is Walters et al. (2004) which used correlation coefficients to study interrelationships between both census and non-census variables, examining the association between smoking, depression, anxiety and population density.

Indices of Deprivation

Often census users employ techniques to condense multivariate data into univariate data (a single value is created from a set of observations from one entity). **Indices** are often considered as an appropriate way to do this representing a variable alongside analysis with other variables in the dataset and are one of the most common uses of census data. For example the correlation may be examined between the index of deprivation and smoking. Two examples of indices are the Townsend Index of Deprivation 1991 and the Index of Local Conditions (ILC)[14]. Different variables are used based on the Small Area Statistics from the Census. The statistical methodology to form these indices is generally to first transform the component variables, then to standardise them and finally add together to produce the overall score for each output zone.

Principal Components Analysis

Another common way of summarising census data is to use **principal components analysis** to describe the variability in the data as weighted components representing weighted combinations of the original variables. The first component will explain the most variation in the data, i.e. the longest axis in multi-dimensional space. Voas and Williamson (2001) used selected census variables in a principal components analysis to determine the extent to which certain underlying components might account for overall variation in the diversity of EDs in the UK. PCA relies on the covariance matrix calculated from the census variables. Given a set of variables denoted by $z_1, z_2, \ldots z_M$ the covariance $\sigma_{jk} = \mathrm{cov}(z_j z_k)$ of z_j and z_k is defined by:

$$\mathrm{cov}(z_j, z_k) = (z_j - \mu_j)(z_k - \mu_k) \qquad (2.6)$$

where μ_j and μ_k are the means of z_j and z_k respectively. The census variables $z_1, \ldots, z_M$ used by Voas and Williamson (2001) were in the form of proportions, e.g. proportion of single men in each ED, proportion of residents with a long-term-limiting-illness for each ED. Hirschfield and Bowers (1997) took various indicators of social cohesion, such as number of lone parent households, for

[14] Information about deprivation scores can be found at *http://census.ac.uk/cdu/Datasets/1991_Census_datasets/Area_Stats/Derived_data/Deprivation_scores/* Accessed July 2006

analysis using principal components. The weightings or 'factor loadings' were used to identify which variables contributed most to social disorganisation in small areas.

Geodemographic Classifications

Geodemographics or area classifications are used to classify and summarise small areas according to their inhabitants. People living in the same locality are likely to have the same lifestyle characteristics. Area classification refers to the classifying of areas into groups of similarity based on the characteristics of selected features within them (Everitt et al. 2001). Geodemographic classifications are based on the same idea and refer to information about population location typically at small scales such as postcodes. There are numerous examples such as Acorn from CACI[15] and Mosaic[16] from Experian. Wallace et al. (1995) describe area classifications as meeting a need for an indicator of socio-economic information contrasting the similarities and differences between areas. Geodemographics / area classifications have many uses including target marketing, consumer profiling, academic research and allocation of resources.

A traditional classification method as described in Debenham et al. (2003) starts with an $M \times V$ matrix describing the EDs (or areas) in the dataset where M is the number of variables and V is the number of EDs. An iterative relocation algorithm such as K-Means might be used to form the groupings. The method of K-Means Classification involves taking the V enumeration districts in the dataset and partitioning them into disjoint subsets so as to minimize a sum-of-squares criterion. The type of census variables used to create the classifications are again proportions such as number of households in an ED with greater than one person per room. We return to and explore area classifications more fully in chapter 6.

[15] Acorn is a geodemographic tool used to identify and understand the UK population and the demand for products and services *www.acorn.caci.co.uk*

[16] Mosaic groupings classify UK households into groups, types and segments *(http://www.business-strategies.co.uk/Content.asp?ArticleID=566)*

Distance Analysis

One way of representing a census dataset is to view each variable as an axis and so every output zone can be represented as a point in multi-dimensional space according to its value on those axes. Multivariate analyses in this context are based on **distance analysis**. Voas and Williamson (2001) used this approach to assess similarities and differences between sets of small areas. Various distance measures were employed to analyse differences between these points or distance from the 'norm'. This technique can be used to identify atypical areas. By first taking the mean of every variable, the mean for England & Wales could be calculated and located in multi-dimensional space. The distance of each area v from the norm is used as a characteristic of the area. In order to give the variables equal weight, they standardised them by transforming the variable values into Z scores.

$$zscore_j = \frac{z_{ij} - \bar{z}_j}{\sigma_j} \qquad (2.7)$$

where z_{ij} represents the value for variable j for household i. The square root of the sum of squares of these new values is the Euclidean distance of each output zone from the norm. Those areas with the highest figures where furthest from the norm. Voas and Williamson (2001) also examine within-class distances for selected geodemographic classes. For example, the distance between every pair of EDs within each GB profile class was averaged and weighted according to the number of EDs in the class to calculate the average-within class distance.

Multiple Regression

Regression analysis is often used at a small area level as a tool to describe how a mean response varies with the explanatory variables. Tickle et al. (2000) use a multiple linear regression model to explain the variability in the mean number of decayed, missing or filled teeth of 6-year old children for 30 districts in the North West Region of England. Deprivation related census variables (indices) at a small area level were used as independent variables in the model. The model was of the form:

$$y_i = \beta_0 + \beta_{1z_{i1}} + \beta_{2z_{i2}} + \ldots + \beta_{Mz_{iM}} + \varepsilon_i \qquad (2.8)$$

The explanatory variables $z_1, z_2 ... z_M$ represent census derived factors in the form of percentages, e.g. percentage of households with no car, percentage of children living in households with no earners. The dependent variable was the mean number of decayed, missing or filled teeth of 6-year olds for ward.

Multilevel Modelling

More complex modelling can be used to take into account the spatial nature of census data. Barnett et al. (2002) use a **multilevel model** to explore the impact of deprivation in explaining the spatial variation in health outcomes for example. Reijneveld (1998) used a multilevel logistic regression model to analyse the adverse effects of area deprivation over and above the effect due to individual socio-economic status. Multilevel models for census data will be explored in detail in chapter 6.

Geographically Weighted Regression

Geographically weighted regression (GWR) can also be used as an alternative to multilevel models. GWR results in a set of local parameter estimates for each relationship which can be mapped to produce a parameter surface across the study region. Fotheringham et al. (2002) describe many examples of GWR. There are many approaches to analysing census data, especially multivariate techniques, those described here provide a summary of the most popular methods.

2.6 Summary of Chapter

In this chapter, we have described what disclosure means and how it may occur with reference to disclosure scenarios and examples illustrated by small cell counts in tables. The disclosure by differencing problem was discussed including differencing via linked tables and in semi-linked data. Pre-tabular SDC methods by their nature provide some protection against differencing, unlike post-tabular methods. Geographical differencing is one type of differencing which is spatial in nature. A review of the literature examined the release of census data for different NSIs with particular reference to small area data and flexible outputs. We also considered the SDC methods that can be used to protect against disclosure and applicability to a flexible tabulation scenario. Ways of assessing risk in frequency tables were briefly discussed. Finally this chapter was ended with a summary of how

census users analyse data which is an important consideration; SDC methods must minimise disclosure risk but not at the expense of the utility of the data.

Chapter 3 New Geographical Perturbation Methods

3.1 Introduction

This chapter introduces some new disclosure control approaches developed in response to the problems discussed in chapters 1 and 2. The challenge is to find a methodology that reduces the disclosure risk in slivers created by geographical differencing and in general for small area data. Implementation and specifics of the new methods are described in chapter 5. Success can be measured by analysing the risk-utility outcome in comparison to RRS. RRS attempts to replicate the procedure used in the 2001 England & Wales Census as closely as possible given that full details of the method are confidential. Regardless of the aggregation strategy adopted, the aim of the new methodology is to offer protection against disclosure (greater than RRS) for a tolerable level of utility.

Section 3.2 briefly discusses the motivation for geographical perturbation methods. Such methods have a number of benefits but primarily the integrity of the attribute relationships is maintained. Section 3.3 then introduces a new framework of geographical perturbation methods. This framework

forms the foundation for this thesis and incorporates both existing methods as well as new methodology (sections 3.4 - 3.6). The potential advantages and disadvantages of different approaches of geographical perturbation for SDC are discussed in section 3.6. Different ways of implementing the new methodologies are explained in preparation for the empirical work in chapter 5.

Section 3.7 focuses on zone-independent methods of geographical perturbation developed with a flexible tabulation scenario in mind; in section 3.7.1 perturbation distance is sampled from a distribution, 3.7.2 considers perturbation in proportion to disclosure risk, 3.7.3 discusses implementation of these ideas and in section 3.7.4 local geographical perturbation is examined. In section 3.8, we look at varying either the sampling fraction and/or the selection of records to perturb, with the objective of balancing risk-utility. Finally section 3.10 deliberates measures for assessing disclosure risk. Indicators that are based on identification disclosure from output tables are discussed as well as more complex analysis to directly measure the risk from geographical differencing. A brief note on indicators of damage is provided in section 3.10.4.

3.2 Motivation for Geographical Perturbation

Chapter 2 discussed some post-tabulation methods which are thought to offer protection against differencing; for example, the ABS cell perturbation method (Fraser and Wooton, 2005) or controlled rounding (Bycroft, 2005). In general, post-tabulation methods have several disadvantages over pre-tabulation methods especially in a flexible tabulation scenario where the potential number of output tables may be unlimited. Achieving both consistency and additivity together can be a problem (for example in the ABS cell perturbation method), the former meaning that the same cell values in different tables do not always agree. Furthermore there is always the possibility of narrowing the uncertainty intervals around protected cells (particularly with rounding) in order to determine with some accuracy, the original count. The post-tabulation process of applying protection to each table individually may be time consuming and can require each table to be checked for potential disclosure risk. Pre-tabulation approaches may be considered more appropriate for protection in a flexible tabulation scenario because they only need to be applied once to the base population and thus any outputs produced would not need to be further checked for potential disclosure. By nature of the pre-tabulation process, it also ensures both consistency and additivity.

Pre-tabular SDC methods such as over-imputation involve modifying the values of the attribute variables *A*. However over-imputation may introduce hidden biases into the data. A developing area of research is to release synthetic data that have been generated from one or more population models. However this is not an approach that has been considered suitable for census data because of the loss of utility associated with a high-dimensional census dataset. An alternative pre-tabulation approach is to modify the spatial variables *S*, in other words to alter the geographic location of records. Methods that perturb the geographical location of households work well because they don't alter the integrity of the attribute relationships but introduce uncertainty into the interacting relationship between geography and the attribute data. Methods of this nature have some downsides as addressed in chapter 2, in particular the difficulty in determining the appropriate level of perturbation to achieve a balance between risk and utility. However they are particularly suited to addressing the disclosure risk from a spatial perspective.

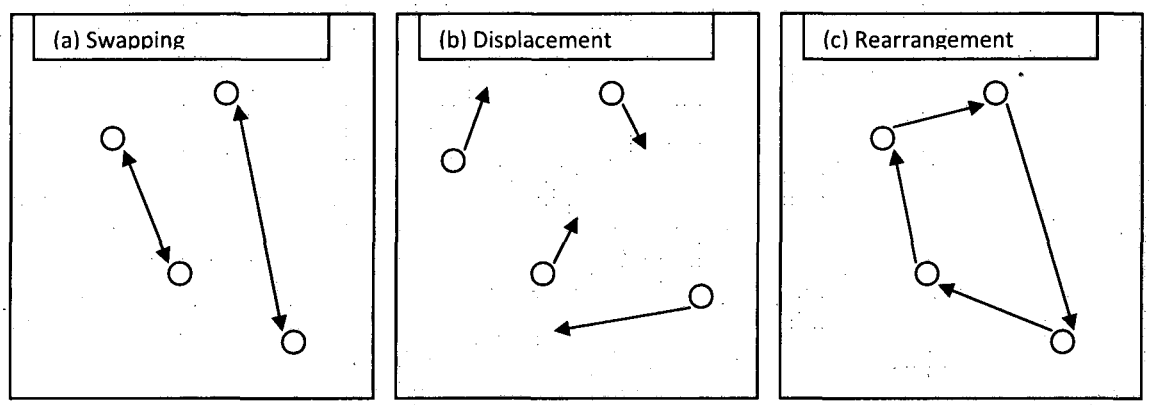
3.3 A New Framework of Geographical Perturbation Methods

The term **Geographical Perturbation** will be used to describe all methods which add uncertainty to either a subset or all of the point locations in the real space $\Re$ such that a grid reference with spatial variables $S=(X,Y)$ is modified to a new location $S'=(X',Y')$. Thus tables based on a particular output zone, produced from the perturbed microdata, may differ from those tables produced from the unperturbed microdata, as the set of households falling in the output zone may change after geographical perturbation. Throughout this thesis we refer to the perturbation rate as the total percentage of records perturbed. We refer to the households selected for perturbation as the sample; or in the case of swapping, the sample is half the size with the remainder being the swapping 'pairs'.

Geographical Perturbation can be achieved in three ways as defined in figure 3.1. The households in the census data can be swapped (existing approach), displaced or rearranged but essentially in all cases, noise is being added to the true location of the households. A 10% perturbation rate could be achieved by selecting a 5% sample of households and pairing each of these to produce a 10% swap in total (figure 3.1a). Or a 10% sample of households could be selected and each household moved a specified distance in unrestricted space (within the census region) as in figure 3.1 (b). Alternatively, a

household could be sampled from the population and a new location identified amongst the pre-existing household locations; i.e. 10% of households could be rearranged as shown in figure 3.1(c).

Figure 3.1: Geographical Perturbation of Households in Census Data



Geographical Perturbation methods can be further categorised into those methods which control movement of household location by output zones: they are zone-dependent, or those which are zone-independent. A zone-dependent method might involve swapping within LADs for example. A hypothetical idea suggested by Brown (2003) was to swap households around geographical boundaries to reduce the disclosure risk from geographical differencing. This would be very tricky to do in practice as the complete set of boundaries would need to be known in advance. Approaches which are zone-independent are particularly advantageous for protecting against geographical differencing as they ignore any reference to existing zone boundaries. Table 3.1 describes a framework for geographical perturbation which will be used as the foundation of the experiments for chapter 5 and is the basis for the discussion of new ideas in this chapter.

Table 3.1 Framework for Geographical Perturbation Methods for Census Data

	Swapping	Displacement	Rearrangement
Zone-dependent	e.g. RRS (section 3.4.1)	Possible new method	Possible new method
Zone-independent	Possible new method	Geomasking ideas applied to census data?	Possible new method

Zone-dependent methods have the disadvantage that the perturbation distance is controlled by the size and shape of the output zones which may be irregular. They are examined further in section 3.4.1. Geomasking methods applied in the literature to disaggregate data (e.g. maps) can be applied to census data and fall in the category of zone-independent displacement. Table 3.1 has some unfilled boxes including zone-independent swapping. These are previously unexplored methods to be considered. In particular zone-independent methods offer lots of flexibility and allow many different features of the method to be varied. Some of these options are now discussed fully. We note that the new methods described here are developed specifically for census data and the flexible tabulation scenario; however they have wide applicability for any area-based data publication and for other high sampling fraction data sources.

3.4 Swapping

Swapping involves pairing households and switching all the geographical variables between them. This technique preserves the global set of household locations and therefore ensures that the total households in each output zone remain unchanged. Zone-dependent swapping is a well-established SDC method that has been used for both UK and US Censuses. In the following section we describe RRS, the approach used in the UK and that will be used as the basis for comparison with all new geographical perturbation methods. New ways of swapping that are zone-independent may be considered for a flexible tabulation scenario and are explored in section 3.7 (- zone-independent methods).

3.4.1 RRS (Zone-dependent swapping)

RRS is a well-known approach for facilitating geographical perturbation. An illustrative example is shown in figure 3.2 where a pair of similar households is found and all the geographical variables swapped on the two households. Households 1 and 5 (identifiable by the *hnum* variable) have effectively had 'noise' added to their grid references so that they now lie within the same LAD but in different OAs. As can be seen in this example, the integrity of the attribute data are maintained since none of the attribute values within a record are modified. However the relationship between geography and the attribute variables is changed in order to protect confidentiality. It cannot be

known whether a household in a particular table actually represents a genuine household in that output zone. For example, an intruder may spot a count of one in a table for the category of 20-25 males unemployed for an OA in Southampton. The intruder does not know if this person actually exists in the area or is a 'false' person arising from swapping. A disclosure cannot be made with certainty. Record swapping provides additional protection on top of the inherent protection in raw census data due to imputed records since these records could also technically be "false" people/households to adjust for under-enumeration.

To keep data distortion to a minimum, similar households can be paired, as in figure 3.2. Similar households can be found by 'matching' on broad age-group, number of people in household, family type, economic status and household space type for example. In fact, the distributions of match variables with geography are also broadly maintained because these variable values are also held constant during the swap (see for example marital status in figure 3.2 and also approximately for econprim). By matching on broad-banded variables such as age, this information can be released to the census users to incorporate into their analyses (as the marginal distributions are maintained). A limited number of match variables can be chosen to preserve attribute-geography relationships (but not to the extent that the objective of introducing uncertainty is lost altogether!).

Figure 3.2: Fictitious Example of Geographical Perturbation using Swapping.

Census microdata before geographical perturbation (swapping)									
Hnum	**Marital Status**	**Age**	**Sex**	**Ppl in hh**	**Econprim**	**Household Spacetype**	**Grid Reference (X,Y)**	**Output Area**	**LAD**
1	Married	46	M	4	Full-time	Semi-det	4000,7186	24UDGT0001	24UD
1	Married	45	F	4	Full-time	Semi-det	4000,7186	24UDGT0001	24UD
1	Single	14	M	4	NA	Semi-det	4000,7186	24UDGT0001	24UD
1	Single	12	M	4	NA	Semi-det	4000,7186	24UDGT0001	24UD
2	Married	62	F	2	Full-time	Detached	4012,7112	24UDGT0002	24UD
2	Married	65	M	2	Retired	Detached	4012,7112	24UDGT0002	24UD
3	Widowed	78	F	1	Retired	Terraced	4032,7197	24UDGT0002	24UD
4	Divorced	50	M	1	Self-emp	Flat-resid	4025,7118	24UDGT0002	24UD
5	Married	31	M	4	Full-time	Semi-det	4039,7149	24UDGT0002	24UD
5	Married	31	F	4	Part-time	Semi-det	4039,7149	24UDGT0002	24UD
5	Single	1	F	4	NA	Semi-det	4039,7149	24UDGT0002	24UD
5	Single	8	M	4	NA	Semi-det	4039,7149	24UDGT0002	24UD
6	Single	24	M	2	Student	Flat-resid	4182,7544	24UCJFH0001	24UC
6	Single	23	M	2	Student	Flat-resid	4182,7544	24UCJFH0001	24UC
...	...	...	...	...	...	...	...	...	...

Census microdata after geographical perturbation (swapping)									
Hnum	**Marital Status**	**Age**	**Sex**	**Ppl in hh**	**Econprim**	**Household Spacetype**	**Grid Reference**	**Output Area**	**LAD**
1	Married	46	M	4	Full-time	Semi-det	**4039,7149**	**24UDGT0002**	**24UD**
1	Married	45	F	4	Full-time	Semi-det	**4039,7149**	**24UDGT0002**	**24UD**
1	Single	14	M	4	NA	Semi-det	**4039,7149**	**24UDGT0002**	**24UD**
1	Single	12	M	4	NA	Semi-det	**4039,7149**	**24UDGT0002**	**24UD**
2	Married	62	F	2	Full-time	Detached	4012,7112	24UDGT0002	24UD
2	Married	65	M	2	Retired	Detached	4012,7112	24UDGT0002	24UD
3	Widowed	78	F	1	Retired	Terraced	4032,7197	24UDGT0002	24UD
4	Divorced	50	M	1	Self-emp	Flat-resid	4025,7118	24UDGT0002	24UD
5	Married	31	M	4	Full-time	Semi-det	**4000,7186**	**24UDGT0001**	**24UD**
5	Married	31	F	4	Part-time	Semi-det	**4000,7186**	**24UDGT0001**	**24UD**
5	Single	1	F	4	NA	Semi-det	**4000,7186**	**24UDGT0001**	**24UD**
5	Single	8	M	4	NA	Semi-det	**4000,7186**	**24UDGT0001**	**24UD**
6	Single	24	M	2	Student	Flat-resid	4182,7544	24UCJFH0001	24UC
6	Single	23	M	2	Student	Flat-resid	4182,7544	24UCJFH0001	24UC
...	...	...	...	...	...	...	...	...	...

3.4.2 Implementing RRS

Figure 3.3 presents a methodology for implementing RRS which can be used to simulate the approach applied in the England & Wales Census 2001. The full details of the swap were confidential but an informed guess can be made as to what might have been done. It may be presumed that only a small percentage of records were swapped else the data distortion would be too great. It is also known that swapped records were not moved out of their LAD but were required to be moved between OAs (Boyd and Vickers, 1999).

Figure 3.3: A Methodology for Implementing a 10% RRS

1. Select 5% of records using SRS and flag them in the population
2. Give each record in the population a random number
3. Sort the population by LAD and then (within LAD) by the random number
4. Add a lag variable to the file where lag equals the previous record in the population
5. Remove all pairs of records which meet the following criteria and put into a new swapped file:
    - The donor (sampled) record has flag = 1 (in the sample)
    - The LAD of the donor record and the lagged record (potential recipient) is the same.
    - The ED of the donor record is different to the ED of the recipient record.
6. Take the reduced population and repeat steps 1 to 5 until 10% of records have been removed.
7. Swap the records in the saved file and replace into the population.

3.5 Displacement for Census Data

An alternative approach to swapping for geographical perturbation might be to instead add noise to household grid references using some kind of mathematical function, without having to pair households. The movement of households is unrestricted in the real space (defined by the census region), as in figure 3.1(b), and can be both zone-dependent (unrestricted within output zones) or zone-independent. Displacement is a commonly used method in the geoprivacy literature, for protecting the confidentiality of mapped points particularly for health data and involves adding noise by transforming the data (see chapter 2). Displacement has never been used before in the context of census data but it would be possible to apply the same ideas. Displacement may offer a number of advantages over swapping. To begin with, it allows more flexibility than swapping, as households do not have to be paired, and a function can be used to determine the movement of points according to some other rule. Displacement will be quicker and easier to implement than swapping as pairs of households do not need to be found. On the other hand, these benefits also mean that the marginal totals of tables will not be consistent as the global set of locations will be modified (so an output zone may end up with fewer or greater households than before perturbation). However this could also be controlled for in some way.

An obvious drawback with displacement is that perturbing a census household with unrestricted movement can be impractical if a household is moved into an uninhabitable or infeasible location such as a river. However if the aim is to publish census data at the tabular level (rather than perturbed microdata), the problem is reduced to preventing the perturbed households lying in an uninhabitable or infeasible output zone. In addition to this, displacement has to be carefully controlled in order to prevent households moving out of the census region altogether.

A key difference between displacement and swapping is that displacement does not involve match variables. There is the possibility for greater data distortion if households are moved randomly with no thought to the original patterns in the data. If small amounts of noise are added to household location, it would be arguable that match variables are unimportant (similar households are generally located near to each other). However moving households longer distances may distort utility more than swapping with match variables, because households are more likely to be moved to an area that

has different characteristics. A benefit of displacement over swapping may be that swapping has a greater potential to be 'undone' by an intruder who may trace the original location of pairs of households which don't 'fit' the location they have been moved to. For example, households which are not matched on hhsptype (detached, terraced, etc) in a major urban areas such as London, may be easily traceable, particularly if the swap involves one household from an area containing only detached households and the other predominantly high-rise council. Displaced households may be identified as not fitting with their area but it would harder for an intruder to work out where they came from. Such comparisons of displacement and swapping would have to be assessed empirically. Ways of incorporating match variables into displacement could be thought of, such as splitting the space into grids and using probabilities to define the likelihood of certain types of household being displaced.

3.5.1 Implementing Displacement with Census Data

The geographical location of household i can be represented by the spatial variable $S_i = (X_i, Y_i)$ where $i \in U$ and X_i and Y_i are real coordinates ($S_i \in \Re$). Displacement, no matter how it is carried out, can be represented in the form where noise is added to the true location of households in the form: $(X_i + \delta_{X_i}, Y_i + \delta_{Y_i})$ for all $i \in U$ where δ_{X_i} and δ_{Y_i} represent the noise added to X and Y. Displacement could be implemented in a number of ways:

(1) Sampling perturbation distance from a distribution

An obvious technique is to draw δ_{X_i} and δ_{Y_i} independently from distributions with specified means, and minimum distances to move (to give a certain level of protection) and maximum distances moved (so as not to distort the data too greatly). Alternatively the distance a household is moved could be sampled independently from the direction. The direction would take the form of a random angle θ sampled from a uniform with lower and upper bounds of $0°$ and $360°$ respectively. Given that the household would be moved in a straight line with specified angle $\vartheta°$, the perturbed location (X_i', Y_i') of household originally located at (X_i, Y_i) can be calculated using basic trigonometry for right angle triangles where $X' = X + d \times \cos(\vartheta)$ and $Y' = Y + d \times \sin(\vartheta)$. Another approach could be to add

noise to (X_i, Y_i) from a bivariate distribution as described in section 2.3 as a form of geomasking (Armstrong et al., 1999 and Kwan et al., 2004).

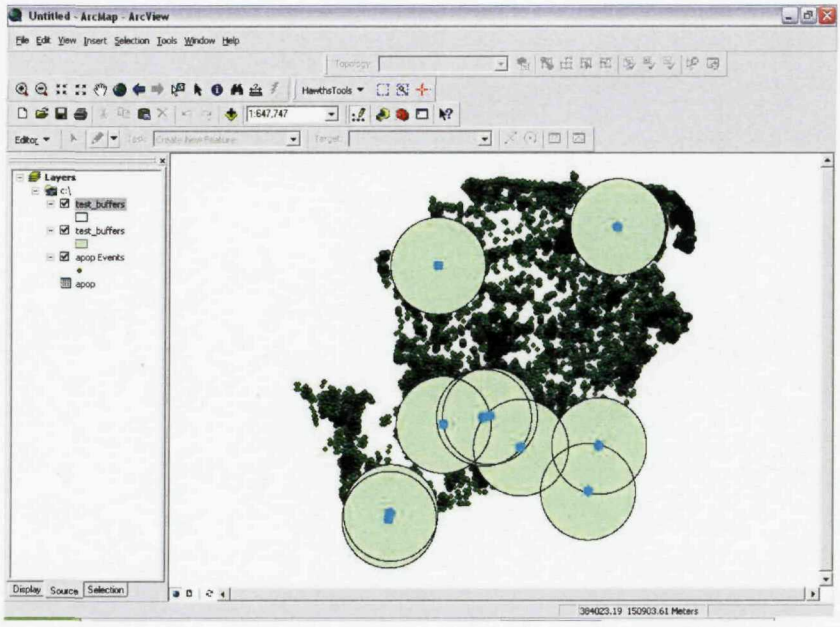
(2) Moving households in and out of output zones.

A zone-dependent displacement method could be designed. For every household moved into an output zone, another could be moved out to keep the overall balance of the total households in the zones. This could be implemented in the same way as RRS but without pairing (households are assigned a randomly selected location in another OA but not out of their LAD). This type of method would be more appropriate for a fixed set of output geographies rather than a flexible tabulation scenario.

(3) Using buffers to determine perturbation distance

A buffer is an area surrounding one or more points or perhaps an existing feature, as illustrated in figure 3.4. Buffers are commonly used in geography for several reasons; a 500m buffer might be created around a project site in order to locate which properties in the area might be affected, or a series of concentric ring buffers of variable widths might be created around a school to work out the number of pupils that fall within specified circular catchment areas. Another possible example is 'buffering' a road to find the number of census output zones that fall in the buffer.

Figure 3.4: Illustration of buffers in ArcGIS, created around a sample of points in Hampshire



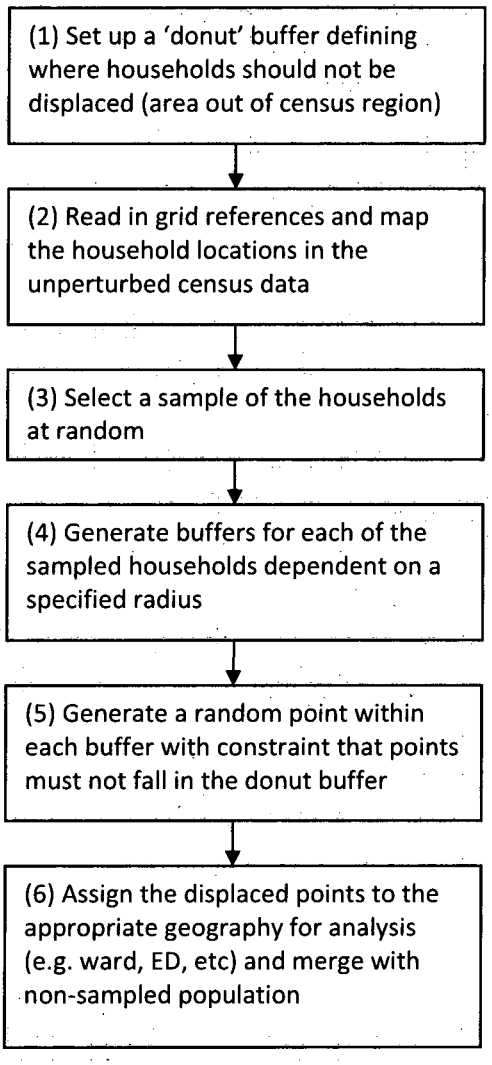
In the geoprivacy literature, techniques for protecting the confidentiality of mapped points have been discussed, for example to protect the location of people who represent cases of cancer or infectious diseases, (Zimmerman and Pavlik 2006, Armstrong et al. 1999). Stinchcomb (2004) describes how geomasking approaches can be implemented in a GIS environment; buffers were used for protecting the confidentiality of mapped health data to determine a random position within each buffer of where the point location should be moved. These buffers may also vary according to another variable such as density. These ideas can also be applied in a census context where the spatial points represent households instead of incidences of disease.

Implementing Displacement for Census Data using Buffers

We now describe a methodology in figure 3.5, for implementing displacement using Stinchcomb's technique. ArcGIS is very useful to prevent households from moving out of the census region as a buffer zone can be set up to define the area that households should not be moved to (this would be much more difficult to achieve in SAS). The methodology described is adapted from Stinchcomb's ideas for mapped points and applied to census data; the basic approach being to use (circular) buffers in ArcGIS to define where the households should and should not move. Note that geomasking

approaches such as that described in Stinchcomb are based on perturbing all point locations whereas in the case of census data, we perturb only a sample. The new methodology described relates to zone-independent displacement.

Figure 3.5: Procedure for carrying out Displacement with Census Data in ArcGIS (flow diagram)



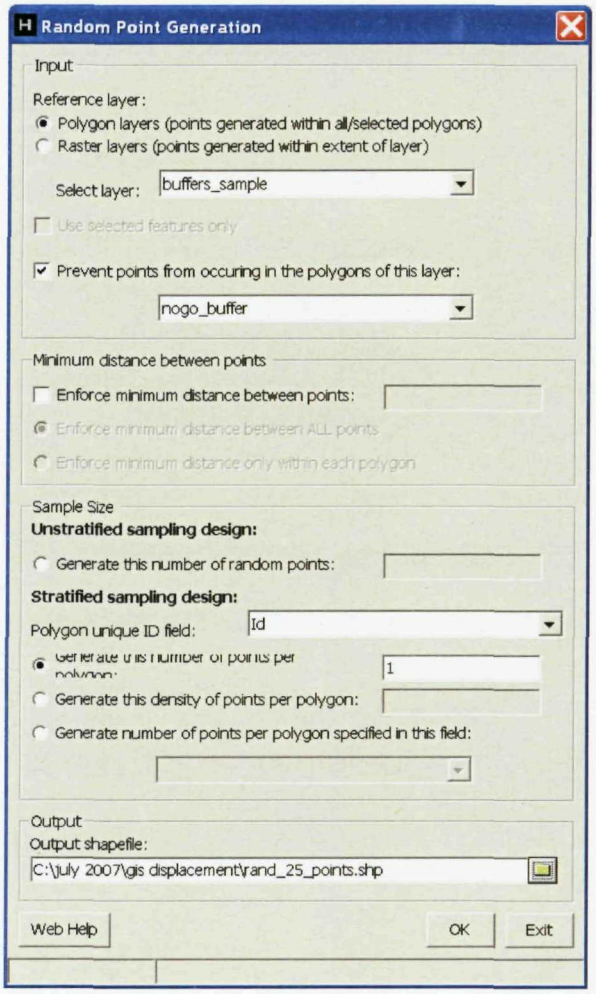
For step 1 in figure 3.5, the boundary of the census region can be downloaded from CASWEB and read into ArcGIS. A large polygon can be created in ArcGIS containing the census region area and the *Union* function used under *Analysis Tools* in ArcToolbox to merge the large polygon with the census

shapefile. The inner census region shape can be 'clipped' and deleted so that only the outer donut remains.

An extension called Hawth's Tools is available for ArcGIS[17] that performs a number of spatial analyses and functions. Using the *random selection* function under Hawth's Tools, a sample of random households can be selected from the household locations layer (step 3). The *create buffers* function under Hawth's Tools can be used to generate buffers around each selected point according to either a specified constant radius or according to a specified variable for step 4. The *random point generation* function, also under Hawth's Tools, can be used to generate one displaced point per sample buffer. There is an option to prevent points from occurring in the polygons of the donut buffer (see figure 3.6). Note that the radius of the buffers could be varied such that each sampled household has a different buffer size. It allows buffers to be created around selected features of a specified radius.

[17] Available free from www.spatialecology.com

Figure 3.6: Generating Displacements in ArcMap



3.6 Rearrangement and a Combination of Approaches

A rearrangement approach can be thought of as a method in between the extremes of displacement and swapping involving a one-to-one mapping of the selected objects in the dataset. This can be operationalised as a series of paths and swaps (cyclic permutations) as in figure 3.1(c). As with swapping, rearrangement is a more restrictive technique: households must be moved to an existing household location, but has the advantage over swapping that a one-to-one pairing of households does not need to be found. A subset of households to perturb could be identified. Then by using some measure of best fit, the households rearranged to maximise utility and minimise disclosure risk.

A simpler approach may begin with a household A, finding a suitable match B and moving household A to location of household B. Then a match C would be found for household B and then household B moved to location C, etc. This could be done in a series of paths or swaps. Table 3.2 concludes this section with a summary of the potential advantages and disadvantages of the three approaches to geographical perturbation (which may be both zone-independent or zone-dependent).

Table 3.2: Comparing the Potential Advantages and Disadvantages of a Displacement, Swapping or Rearrangement Approach

ADVANTAGES		
Displacement	Swapping	Rearrangement
• Quicker to implement • More difficult to find original location of perturbed household • Greater control and flexibility household movement • Solves problem of households which are difficult to pair (e.g. communal establishments, 12-person households)	• Match variables easily incorporated so utility should be better • Global set of household locations is unchanged	• Match variables easily incorporated so utility should be better preserved • Greater uncertainty introduced than with swapping (one-to-one pairing) • More flexible than swapping because households don't have to paired
DISADVANTAGES		
Displacement	Swapping	Rearrangement
• Match variables not found so utility may be worse • Method needs to be carefully implemented to avoid moving out of census region or to an infeasible output zone	• Easier than other methods to unpick households swapped to unlikely locations • Implementation is more difficult as paired households have to be found	• More complex to decide how households should be moved • Implementation is more difficult than displacement as a matching household has to be found

Alternatively a mixed approach could be taken. Particular types of records might be difficult to pair for swapping, e.g. large households and communal establishments. These records could be displaced while the remainder of the sampled households are swapped.

The ideas for displacement, rearrangement and swapping will be explored empirically in chapter 5.

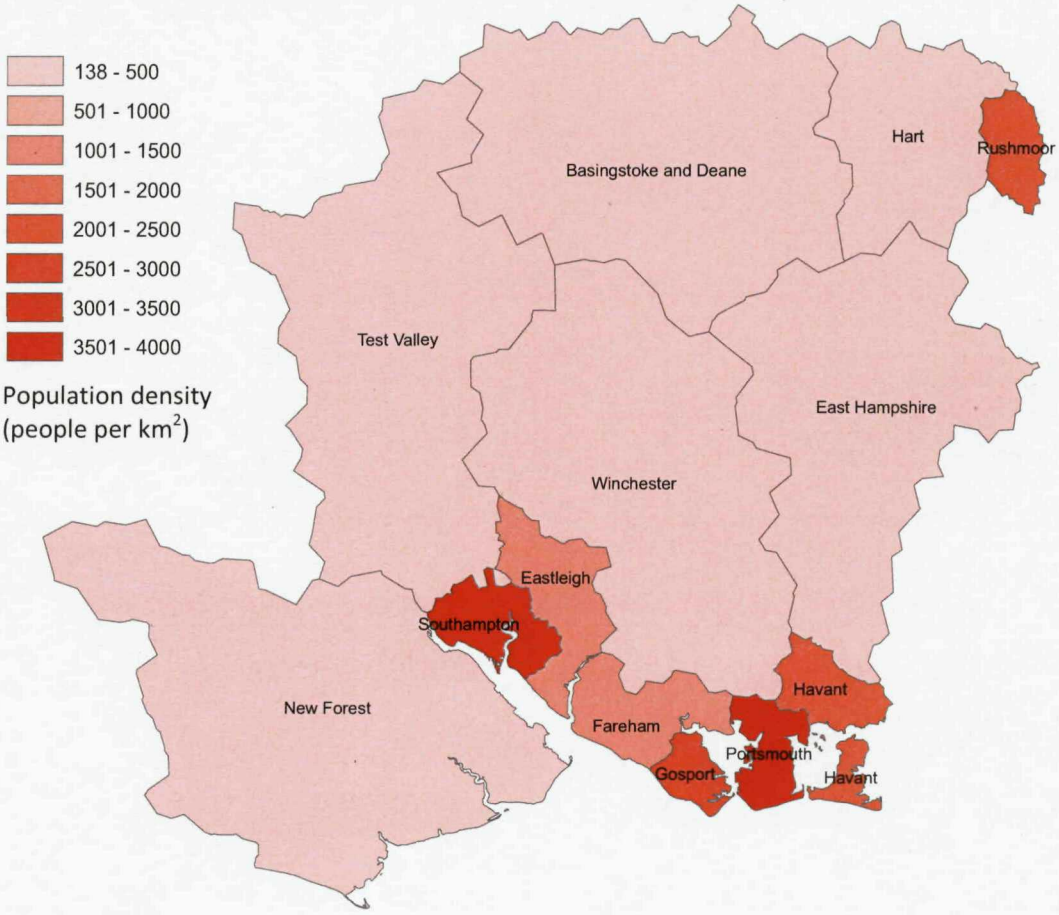
3.7 Implementation of Zone-independent methods

This section considers zone-independent methods of geographical perturbation; that is methods which ignore reference to any pre-existing output zones. These methods have been developed with a flexible tabulation scenario in mind. Swapping, displacement or rearrangements can be zone-independent. Section 3.7.1 considers sampling perturbation distance from a distribution (instead of being based on output zones). Zone-independent methods allow a lot of flexibility so we will also consider new ways to improve the risk-utility outcome from a spatial perspective in section 3.7.2 by perturbing households in proportion to spatial distribution of risk. Section 3.7.3 looks at implementing these approaches. In section 3.7.4, a new idea of local geographical perturbation is discussed with the specific objective of reducing the disclosure risk in small areas and geographically differenced slivers.

3.7.1 Perturbation Distances

The fact that RRS relies on pre-determined geographical boundaries can be a drawback of the method. In the US Census 2000, swapping was conducted between 'blocks' (Fienberg and McIntyre, 2004). In the England & Wales 2001 Census, households were moved between OAs but within their LAD (Boyd and Vickers, 1999). LADs can often be irregularly shaped with varying population distributions. Figure 3.7 shows the population density and boundaries of LADs in Hampshire.

Figure 3.7 Population Density and Geographical Boundaries of LADs in Hampshire



LADs such as Eastleigh and East Hampshire are either irregularly shaped or have very different shapes from one another; compare the long shape of Test Valley to Basingstoke and Deane. These shapes influence where the households are moved to if the method is zone-dependent. Moreover, Southampton and Portsmouth are similar in size to Rushmoor and Gosport but their population densities are almost twice the size; 3,786 people per km^2 and 4,159 people per km^2 compared to 2,027 people per km^2 and 2,832 people per km^2. Thus the perturbation distance ignores population density which may be considered an indicator of disclosure risk (see section 3.7.2).This can be avoided by taking a perturbation approach in continuous space (rather than zonal space) and generating the perturbation distance from a distribution. This could be applied to any geographical perturbation method (swapping, displacement or rearrangements) and such an approach would overcome the arbitrary distance moved by households according to output zones (in this example, LAD size).

The distribution can be chosen so that greater control can be gained over how households are moved. For example, the perturbation distance d could be sampled from an exponential distribution or a normal distribution. Generating from an exponential would mean a rapidly decreasing probability of d being a large perturbation distance but a high probability of being a small distance (defined by a small mean). A normal distribution would result in d being equally likely to take very large values, as very small values. The distribution could be truncated at maximum and minimum values to define a minimum distance moved. It is straightforward to implement such an approach with displacement as there is no restriction on where households can be moved. However swapping is more complex as there may not be a household to swap with at a distance d away. Instead d could be an approximate distance so that households are moved a distance $[d-\delta_d, d+\delta_d]$ where δ_d is large enough to define an interval encompassing matching households. A paired household would then be chosen from the interval.

3.7.2 Taking Account of Population Density

The relationship between geography and disclosure risk has been discussed in regard to smaller output zones with smaller populations presenting greater risk. However population density also is an important factor; in output zones which have a high population density, the risk is much smaller than in rural, sparsely populated zones. The perturbation distance could also take into account population density or household density. In fact similar ideas have also been discussed in the geoprivacy literature. Kwan et al. (2004) have assessed the spatial masks discussed in Armstrong et al. (1999), in particular the levels of random perturbation in relationship to disclosure risk. Since mapped locations of disease or crime spots often contain a wide variety of population densities, the amount of noise added to spatial location can be allowed to vary with population density. The idea of including population density in the disclosure risk model has also been discussed by VanWey et al. (2005) who simulated a sampling frame of public schools in the US. Their data contained the geographical location of each school with potentially sensitive attribute information. A solution was proposed whereby map symbol size was adjusted to cover multiple schools, providing locational uncertainty in proportion to a specified level of identification risk. For schools in large cities a much smaller point buffer was needed than in remote rural areas. This type of approach is also applicable to census data but first the underlying population density needs to be estimated.

Estimating Population Density

Kernel density estimation is often used when the underlying spatial distribution is unknown and can be used to estimate population density. Probabilities could be assigned to households dictating how far they should move based on local density. Since the true form of the density distribution is unknown, a non-parametric method can be used to estimate it given our microdata. The idea is to take geographical locations (co-ordinates) of households and produce a smoothed estimate of density. The following summarised from Simonoff (1996), Cressie (1991), Diggle (2003) and Gatrell et al. (1996) gives some background on kernel density estimation for a bivariate dataset

Assume the microdata set $\{(x_1,y_1),...,(x_N,y_N)\}$ is a random sample drawn from an unknown density function $f(x,y)$. A definition for the continuous function of $f(x,y)$ is:

$$f(x,y)=\frac{d}{dxdy}F(x,y)=\lim_{h\to 0}\left(\frac{F(x+h)-F(x-h)}{2h},\frac{F(y+h)-F(y-h)}{2h}\right) \tag{3.1}$$

However this formula assumes the values of x and y are continuous which is not the case with census microdata (which comprises a discrete set of household locations). A continuous function can be derived based on the random sample starting with an empirical estimator of the distribution function:

$$\hat{F}(x,y)=\frac{1}{N}\sum_{i=1}^{N}I((x_i,y_i)\le(x,y)) \tag{3.2}$$

where, I is an indicator function such that $I(A)=1$ if A is true and $I(A)=0$ otherwise. Replacing $I((x_i,y_i)\le(x,y))$ with a smooth approximation $G\left(\frac{x_i-x_j}{h},\frac{y_i-y_j}{h}\right)$ for $i\neq j$ then differentiating $\hat{F}(x,y)$ gives the estimated density $\hat{f}(x,y)$ which can be written as:

$$\hat{f}(x,y)=\frac{1}{Nh^2}\sum_{i\neq j=1}^{N}K\left(\frac{x_i-x_j}{h},\frac{y_i-y_j}{h}\right) \tag{3.3}$$

with kernel function $K[\cdot]$ and bandwidth h. The bandwidth determines the level of smoothing of the data. Assume $\int h^{-2}K(u)\,du=1$ and as a consequence $\int\int \hat{f}(x,y)\,dx\,dy=1$. Furthermore, the estimator is consistent at any point (x_i,y_i):

$$\hat{f}(x,y)=\frac{1}{Nh^2}\sum_{i\neq j=1}^{N}K\left(\frac{x_i-x_j}{h},\frac{y_i-y_j}{h}\right)\xrightarrow{N\to\infty}f(x,y) \tag{3.4}$$

This assumes that the bandwidth is the same for both x and y. Relaxing this assumption to have a vector of bandwidths $h=(h_x,h_y)^T$, the bivariate kernel density estimator then becomes:

$$\hat{f}(x,y)=\frac{1}{N}\cdot\frac{1}{h}\sum_{i\neq j=1}^{N}K\left(\frac{x_i-x_j}{h_x},\frac{y_i-y_j}{h_y}\right) \tag{3.5}$$

Note that the bandwidths can also be weighted so that in densely populated areas more points can be taken into account and vice versa.

A multiplicative kernel can be used for the multidimensional kernel $K(u)=K(u_x,u_y)$. This is the simplest form.

$$K(u)=K(u_x)\cdot K(u_y) \tag{3.6}$$

where $K(u)$ denotes a univariate kernel. The bivariate kernel estimator then becomes:

$$\hat{f}(x,y)=\frac{1}{N}\sum_{i\neq j=1}^{N}\left(\frac{1}{h_x}K\left(\frac{x_i-x_j}{h_x}\right)\cdot\frac{1}{h_y}K\left(\frac{y_i-y_j}{h_y}\right)\right) \tag{3.7}$$

The kernel function is a probability density function which is evaluated for every pair of units. A discrete version of the bivariate normal distribution is commonly used. Note that the kernel function often includes an indicator function such that values outside the local neighbourhood of the

(x_i, y_i) being evaluated are given zero weight. Another simpler kernel function that is often used is the Epanechnikov kernel which is (in the univariate case):

$$\frac{3}{4}\left(1-\left(\frac{x_i-x_j}{h}\right)^2\right) \text{ for } -1<\frac{x_i-x_j}{h}<1 \text{ and 0 outside that range} \tag{3.8}$$

Suppose this form of the kernel was used in our bivariate density estimator above, it would then become:

$$\begin{aligned}\hat{f}(x,y) &= \frac{1}{N}\cdot\frac{1}{h_x h_y}\sum_{i\neq j=1}^{N} K\left(\frac{x_i-x_j}{h_x}\right)K\left(\frac{y_i-y_j}{h_y}\right) \\ &= \frac{1}{N}\cdot\frac{1}{h_x h_y}\sum_{i=j=1}^{N}\frac{3}{4}\left\{1-\left(\frac{x_i-x_j}{h_x}\right)^2\right\}\cdot I\left(\left|\frac{x_i-x_j}{h_x}\right|\leq 1\right)\cdot\frac{3}{4}\left\{1-\left(\frac{y_i-y_j}{h_y}\right)^2\right\}\cdot I\left(\left|\frac{y_i-y_j}{h_y}\right|\leq 1\right)\end{aligned} \tag{3.9}$$

Note that a contribution to the sum for unit i is only given if (x_i, y_i) falls into both intervals $[x_i - h_x, x_i + h_x]$ and $[y_i - h_y, y_i + h_y]$. The univariate Epanechnikov kernel has been applied to the bivariate function above. Observations in a cube are then estimated around the point locations to estimate the density at those points. It is possible to use a multivariate Epanechnikov kernel: a *spherical kernel* as opposed to the *multiplicative kernel*. The result is that observations in a sphere around the point locations are included (rather than a cube).

Computational Tractability and a Gridded Approach

Suppose the sample consists of N distinct points. A direct application of formula 3.9 to the dataset that is unbounded (without the indicator function) would result in $O(Ng)$ operations determining the estimator at g grid points. With bounds, computer time can be reduced since the estimator is only calculated within a neighbourhood of size h. However with such a large dataset, although the improvement is likely to be significant, the algorithm would still be slow. Moreover, given that a large sample of the microdata may be needed to be perturbed to meet the required protection level, evaluating the density at a subset of points still may be computationally complex if the subset is large.

More efficient kernel smoothing algorithms can be defined by *binning* (gridding) the data (Wand, 1994). This is often done with variables in a univariate context (and amounts to rounding of the data or *left moving average,* but it could be applied in a two-dimensional context as well. The new dataset would be produced by moving the original data to the bin centres according to the binning rule. A simple scheme is to define equally spaced bins B of width and height β and to then move each data point (x_i, y_i) to the nearest bin centre. The bin count c_B is then:

$$c_B = \sum_{i=1}^{N} (x_i, y_i \in B(x^b, y^b)) \tag{3.10}$$

where bin B is defined by its bottom left-hand corner location for $x^b = (x_0 + w_B\beta)$ and $y^b = (y_0 + w_B\beta)$ where w_B is a weight indicating the location of the bin and (x_0, y_0) indicates the origin of the grid.

In effect, the data are compressed so that the new dataset is at most as large as the original dataset (if β is chosen such that each data point lies in a different bin). Binning in this case requires $O(N)$ arithmetic operations. A binned kernel density estimator can then be estimated given the grid size, the kernel type and the bandwidth. This would be a binned approximation to the ordinary kernel density estimator.

Nearest Neighbour Approach

Another way to take into account population density is to use nearest neighbours distances. Households could be swapped with their *nth* nearest neighbour. This will lead to households not moving very far in high density areas but in rural areas, the same *nth* household will be much farther away. However this approach is problematic because all the nearest neighbour distances between households would need to be calculated which, if there are N households, makes $(N-1)+(N-2)+\ldots$ or $\frac{(N^2 - N)}{2}$ calculations.

A solution is to combine the ideas: nearest neighbours, binning the data and using an interval to determine distance moved. The data can first be gridded/binned to a resolution that is suitable to perform the calculations (this is equivalent to the rasterization functions that can be done in a GIS). Then the household could be swapped with any neighbour from the interval $[n-\delta_n, n+\delta_n]$ where δ_n is some constant. Thus n is defined according to household space rather than Euclidean space. The interval of households permits choice over which household is selected so that the most similar household could be chosen if swapping is implemented. The use of gridding aids the swapping algorithm but individual locations are retained.

A further alternative is to estimate the population density of particular risky variables such as the elderly population and move households with respect to these densities. This could be implemented with any variables which are visible and traceable that may indicate presence of local uniques. The extension of population density to cover particularly risky variables is not considered in the empirical work of this thesis.

3.7.3 Implementation using a Grid Based Approach

The following methodology describes how zone-independent methods can be implemented using a grid based approach for computational efficiency. The methods are described with respect to a swapping approach.

Zone-independent Swapping ('distance swapping')

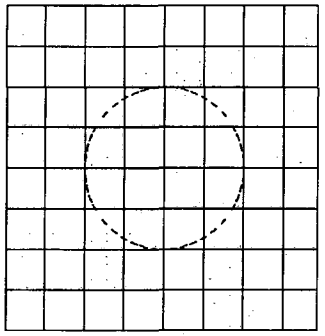
In this new method, swapping distances are sampled from a distribution rather than being determined by the shape and population distribution of pre-existing geographies, thus providing extra protection against geographical differencing. Each household in the sample is allocated a (straight line) distance *d* to move, drawn from a distance function with mean λ equal to that of the random swap. Suppose an exponential distribution were to be used, this ensures *d* cannot be negative and has a rapidly decreasing probability of taking a large value.

The decay function then takes the form:

$$f(d) = \lambda^{-1} e^{-d/\lambda}, \quad d \geq 0 \tag{3.11}$$

and this has the property that $\lambda = E(d)$. The distribution $f(d)$ may be truncated below by $\min(d_r)$ and above by $\max(d_r)$ which denote maximum and minimum distances moved between paired households of the random swap. The household should then be swapped with any other household distance *d* away. This would involve swapping with any household on the perimeter of a circle of radius *d*. For ease of computation, households are assigned to a 100m raster. A donor household can potentially be swapped with any recipient household that falls in a grid cell on the perimeter of the circle. We define the grid cells of the circle and exactly how this process is carried out using lattice points as described in the next paragraph.

Figure 3.8 Distance Swap: swapping with households on the perimeter of the circle



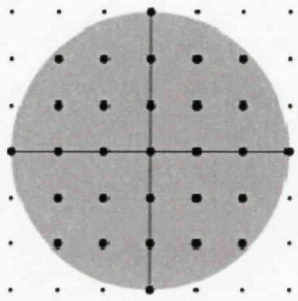
To implement this in practice it must be known which grid cells fall in a circle of radius *d*. Both *d* and the location of the centre of the circle can vary depending on where the donor household is and the distance sampled. A *template* of grid cells falling in circles of unit 1 radius, unit 2 radius, etc (where 1 unit relates to the 100m, the size of the grid cell) may be calculated from the centre of origin (0,0). The grid cell locations can then be rescaled according to the location of the donor household. A grid cell is defined as falling in the circle of radius *d* if its centre co-ordinates (X_g, Y_g) meet the criteria:

$$(X_g)^2 + (Y_g)^2 < d^2 \tag{3.12}$$

Gauss's circle problem (Hilbert and Cohn-Vossen, 1999) details how many *lattice points* fall in the circle as shown in figure 3.9 and is used to confirm the calculations. Five grid cells fall in a circle of radius unit 1, thirteen grid cells fall in a circle of radius unit 2, twenty-nine for radius 3, etc where 1 unit equates to the dimension of the grid cell.

Figure 3.9: Lattice Points falling in a circle

29 cells fall in a circle of radius 3 where a 'cell' is defined by its left hand corner (the lattice point)



The band of cells containing households for swapping are then identified by rescaling the template just described. Thus if a distance d is drawn from the distribution $f(d)$, the donor household can be swapped with any household in the interval

$\left(\left\lfloor \frac{d}{100} \right\rfloor, \left\lfloor \frac{d}{100} \right\rfloor + 1\right)$ of gridded cells (by differencing between the two rescaled templates).

Potentially this could result in a household moving out of the space or perhaps no households being present at the chosen swapping interval. In the rare case of this happening, a new distance can be generated (equivalent to rejection sampling).

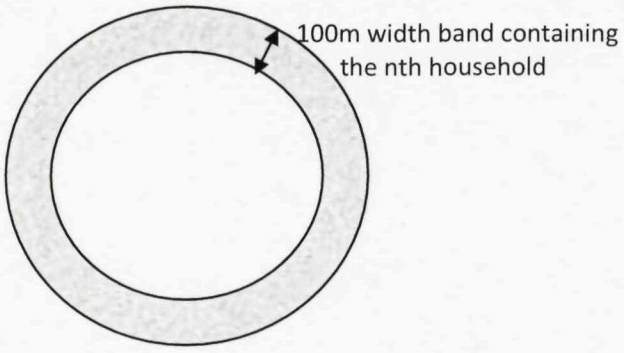
Zone-independent Swapping proportional to density ('density swapping')

This method is an extension of the previous, but this time taking into account the relationship between risk and population density. In areas where the population density is low, cells can be cell unique (because of small populations) up to relatively high levels of aggregation. In areas where the population density is high, the probability of obtaining cell uniques is very small because of larger

populations. Disclosiveness is generally related to the proximity of other uniques with similar characteristics. Thus cell uniques in low population density areas need to be moved farther to become non-disclosive. Household density will be used as a proxy for population density since the two are highly correlated and households are easier to deal with rather than multiplying by number of residents.

To implement a density swap, a 'household distance' n is generated randomly for each of the households in the initial sample, from an exponential distribution with a specified mean θ. This distance will determine the number of households, n, in the circle for which the initial household is at the centre and the matching household on the circumference. As before, households can be assigned to a 100m raster but this time a cellular approximation to a circular search is performed. The circular band corresponding to the number of households, n, is determined by counting the households in successive bands (using the scaled template in figure 3.8) until the cumulative count is greater or equal to n. The outer band of households of 100m width (obtained by differencing the scaled templates) contains the nth household. In other words, the household is swapped in an interval containing its nth nearest neighbour as in figure 3.10.

Figure 3.10: A household is swapped with another in the 100m band containing its nth neighbour



The probability density function of n might be exponential for example and is therefore given by:

$$f(n) = \theta^{-1} e^{-n/\theta}, \quad n \geq 0 \qquad (3.13)$$

so that $\theta = E(n)$.

3.7.4 Local Geographical Perturbation

Sampling perturbation distance from a distribution means the agency is faced with the decision of what parameter values to use. A new idea is to set the mean perturbation distance such that households are moved much shorter distances rather than the average distance between OAs in a LAD implied by RRS. This will be referred to as 'Local Geographical Perturbation' as the noise is targeted at the local level. The objective of the new methodologies is to reduce the number of risky records, particularly unique records which can lead to identification and possibly attribute disclosure through matching to other data sources. Uniques occur when there is only one record with a particular combination of variable characteristics, i.e. a cell count of one in a table representing an output zone. In terms of geography, disclosure risk can arise in two ways:

- 'Local' Uniques
- Special Uniques (Skinner et al., 1994)

Special Uniques are those which arise irrespective of geography definitions and will generally be unique in small aggregations (at lower levels) as well as in large aggregate areas (higher levels). For example; a 16-year old widow or an urban-farmer. These are people (or households) which are unique given the special combination of their attribute values. Thus geographical perturbation, being a spatial method, does not generally provide any protection for these records. Generally these people or households are rare and so a separate, additional method of disclosure control must be applied such as recoding or imputation to disguise the unusual combination of characteristics.

The remaining unique records can be defined as 'local uniques' and are generally unique because of geography. Most of these local uniques can be found in small output zones, at the OA or ED level for example, because of the small populations, thus are called 'local' uniques. The chances of another record with the same variable values is small at this level. At higher levels of geography, these uniques generally disappear due to larger populations. Moreover these records are potentially easier to identify at local levels because of the local knowledge that an intruder might have. When carrying out geographical perturbation, the mean perturbation distance can be adjusted so that uncertainty is added around household location at the local level where the majority of households are unique. This may be considered optimal, as moving households longer distances would result in larger damage to

the output zone tables, relative to the gain in protecting the extra households unique at higher levels of aggregation (as households located together are more likely to have similar characteristics). By targeting the perturbation around the postcode level or other small area, protection is given to small area data and slivers with some uncertainty carried through to the EDs and higher levels. However the damage to the data is likely to be much less than if the mean perturbation distance was to be larger. Setting the perturbation distance by targeting the level with the most uniques would depend on the number of variables in the table. For example, if on three variables most records are unique at the postcode level but not at the ED level then the mean perturbation distance could set to be equivalent to the mean distance between postcode centroids. These ideas are tested empirically in chapter 5. We note here that another advantage of local perturbation is that there is much less likely to be inconsistencies in the data; as records located near to each other are more likely to be similar in terms of general characteristics – Tobler's first Law (Tobler, 1970), although this obviously has to be balanced against reducing the disclosure risk.

Implementing Local Geographical Perturbation

An example of 'local' geographical perturbation combining the techniques local and zone-independent swapping also taking into account population density is a new method which we refer to as:- **Local Density Swapping**. Assuming for the moment most of the disclosure risk (in terms of numbers of uniques) is at postcode level, to find an appropriate *mean household distance* θ a small swap or rearrangement could be carried out between 10% of adjacent postcodes for example and the mean, maximum and minimum perturbation distance calculated between these swaps. This could be used as a starting point for determining θ and other parameter values at the local level, possibly adjusting later to obtain a suitable balance between risk and utility. The swap can then be implemented as before for the density swap.

3.8 Parameters which may be Varied

There are some other features of geographical perturbation that have not previously been mentioned but that can also be varied. Here we discuss varying the sampling fraction and the selection of records.

3.8.1 Varying the Sampling Fraction

If the mean perturbation distance is small then a larger proportion of records can be swapped to arrive at the same level of utility. In particular, the idea of 100% local geographical perturbation is a possibility where all households have a small amount of noise added at the local level. A 100% swap would have to be implemented slightly differently to a lower swap rate because matching pairs need to be found for all households in a 50% sample. Moreover the ordering of the households and the pairs available to match with may also be important (near the end of the swap, the set of possible matching pairs is very small) – this is considered more fully in chapter 5 with particular attention paid to utility since all records are perturbed. Figure 3.11 describes a method for implementing 100% local distance swapping.

Figure 3.11: A methodology for a 100% Local Distance Swap

1. Flag all households in the population with a value of zero.

2. Start iterations.

   3. Select first zero-flagged record in list (donor).

   4. Generate a random distance *d* from a distribution with parameters specified by random swap.

   5. Find cells in the circular band *d* cells away from household to swap using pre-calculated template.

   6. List candidate households in these cells and add to the list the donor household.

   7. Merge list with the census file to get match variable (head of household only) values.

   8. Score each candidate household in the list according to the number of variable values which equal that of the donor household.

   9. Sort the list of candidate households by ascending flag and then by descending scores. (Aim being to swap with already swapped households (flagged one) only as a last resort).

   10. Save into a new file; the donor household and candidate household at the top of the sorted list.

   11. Flag the two households swapped on this iteration with a value of one in the original population file.

12. Go to next iteration

Pairs of records to swap are identified and saved in a separate file. The households in this saved list can then be swapped in a separate program to create the new population. This may be done iteratively by swapping the first two households on the list, then the second pair on the list, and so on. When processing the list of households to swap, if a household has already been swapped once, then the new household at its original location should be swapped instead. It is important that each household in the population is marked with a flag of zero, becoming one after being swapped, so that households already swapped (flagged one) are limited from being swapped again. During the procedure, the number of zero-flagged households decreases, thus the choice of recipients for the donor decreases.

If density was to be taken into account (**100% density swapping**), a paired household would be found from the band of cells containing the *nth* household. In addition, the ordering of the swaps can be changed so that high density households are swapped first, therefore having the greatest choice of recipients to pair with (preferably with a household in the specified interval containing *n* households and preferably with another high density household). If low density households are swapped last, then they may end up being swapped with a household nearby (rather than a longer distance) if there are no alternative (unswapped) candidates.

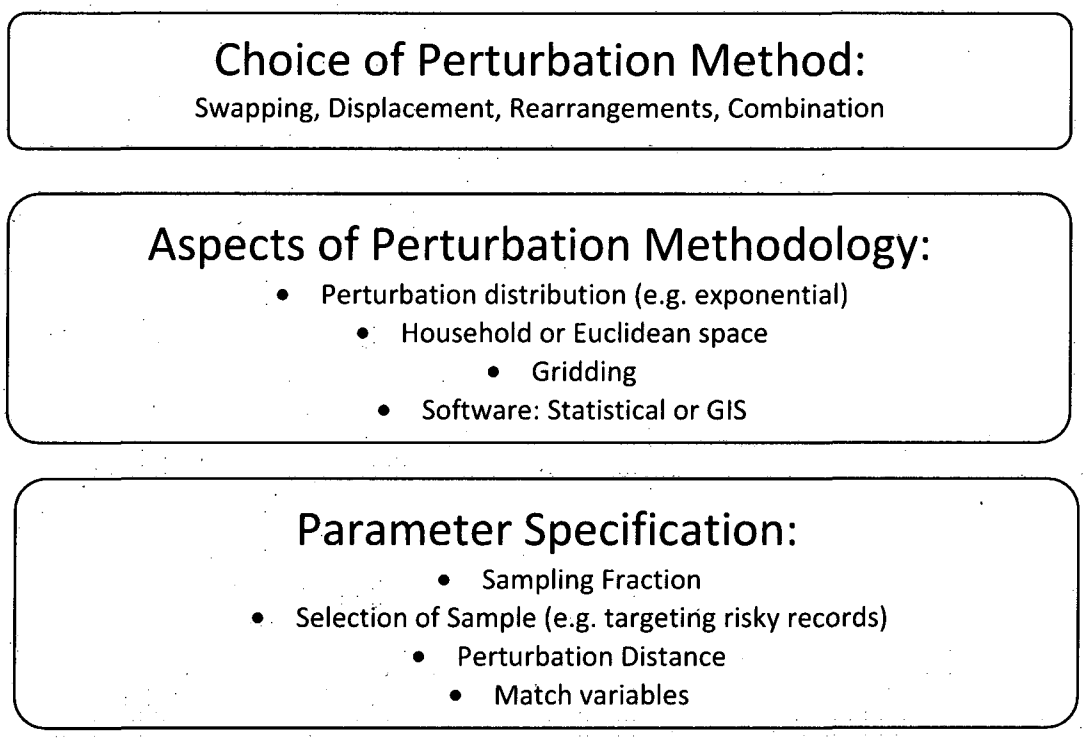
3.8.2 Selection of Records

Another factor to vary is which records should be selected for geographical perturbation. In the US, the method of swapping applied to their census data in 2000 was targeted, however such an approach has not been applied to UK census data. Targeted swapping means only risky records are swapped with the idea being to achieve the greatest reduction in disclosure risk by targeting the uncertainty to where it is needed most. Risky records are those that are likely to produce small cell counts such as records with variable values from minority ethnic populations or the very elderly. If a sample of records is to be perturbed in some way, be it by displacement, in continuous or zonal space, this sample could be from the subset of risky records only. What defines a risky record would have to be considered carefully. It is important to remember that at the small area level, many records are unique (section 3.7.4). Records could be assigned levels of risk with those representing a very high risk being perturbed much more than those records representing a lower risk.

3.9 Summary of Approaches to Perturbation

Figure 3.12 summarises the new methods described in this chapter (in particular the first two boxes). Zone-dependent methods are based on the households moving between pre-determined geographies. Zone-independent methods are based around a perturbation distribution and so more choices need to be made such as the type of distribution, setting the parameters and whether the data is gridded. As with traditional zone-dependent methods (RRS), the sampling fraction for swapping, displacement or rearrangements needs to be decided as well as how the sampled records are chosen.

Figure 3.12: New Methods for Geographical Perturbation



These ideas will all be put into practice in chapter 5.

3.10 Impact of Perturbation Methods on Risk and Utility

The aim when composing the new methodology has been to reduce disclosure risk in regard to the challenges described in the introduction. Disclosure protection can be measured by comparing disclosure risk before and after geographical perturbation.

3.10.1 Using Cell Uniques to Assess Risk

The focus will be on cell uniques, counts of one in the frequency tables, as an indicator of disclosure risk. Generally, statistical agencies are most concerned with releasing counts of ones because they can lead to identification disclosure. Identification disclosure can then lead to attribute disclosure through linking of different data sources. Measuring attribute disclosure directly (see chapter 2) can be complex, in particular the difficulty in determining what constitutes attribute disclosure. This could be defined as all zero cells in a row or column for example which equates to negative attribute disclosure; i.e. all persons of a certain age group not claiming benefits. Alternatively, the majority of cells being zero, an example might be a row in a table describing poor health where the only non-zeros fall in one or two religious groups. It would generally be expected that cases of attribute disclosure increase as cases of cell uniques increase (as they both originate from small cell counts) depending on the population size in the table. For simplicity, cases of identity disclosure will be considered only.

Assessment of cell uniques can be elaborated in the context of geographical perturbation. The cells in a frequency table T are defined by cross classifying a subset of the attribute variables in the vector A (see chapter 2). An arbitrary cell c in this table is defined by a combination of the categories of this subset of attribute variables. Let F_c denote the cell frequency in a specified zone, which is the number of units in the zone with the specified combination of values of the attribute variables. To be explicit about the effect of the perturbation process, let F_c^o denote the cell frequency before perturbation and F_c^p the cell frequency after perturbation. Cell counts of one in a published table after geographical perturbation could have arisen in one of the following ways:

(1) The cell count of one could relate to the same household in both the original and protected table, i.e. the household has not been geographically perturbed. $F_c^o = 1$ and $F_c^p = 1$

(2) The cell count of one in the protected table could have been a value of zero in the original table. In other words, a household with the characteristic has been perturbed into the area. $F_c^o = 0$ and $F_c^p = 1$

(3) The cell count of one in the protected table could have been a count of greater than one in the original table and as a result of *out-perturbation*, the number of households with the characteristic has reduced. $F_c^o > 1$ and $F_c^p = 1$

(4) The cell count of one could have been a cell count of one in the original table but relate to a different household. In other words, the *in-perturbed* household happens to have the same characteristic in this particular table. (It is unlikely that any two households are exactly the same on all 20+ census variables.) $F_c^o = 1$ and $F_c^p = 1$

Scenarios (2) and (4) will be called **False Uniques (FU)**. An intruder spotting a one in the table will not make a correct link to the household as it does not belong to the area. Scenario (3) will be called a **Disguised Unique (DU)** since an intruder has a 1/2 chance or less of making a correct link to the true household. The most undesirable scenario occurring in (1) when a cell count of one relates to the original household shall be referred to as a **True Unique (TU)**. Measuring disclosure risk in this way assumes the intruder has knowledge of a small area (perhaps their local area) and will attempt to use the variables in the table as key variables to try to identify someone. Table 3.3 summarises all the possible scenarios in which $F_c^p = 1$ can arise and assumes records cannot be perturbed more than once. 'A' refers to the original cell count F_c^o, 'B' represents all cell uniques in the protected table F_c^p and 'C', the number of matches in common (households that have not been geographically perturbed).

Table 3.3: Sources of Published Cell Uniques

This table applies for all cell uniques: $B = 1$

		Number of Matches in Common, C	
		0	1
Original Cell Count *A*	0	A household perturbed into the output zone FALSE UNIQUE	Not possible
	1	Unique perturbed out of the output zone and another unique perturbed in FALSE UNIQUE	Unperturbed unique (no in-perturbation or out-perturbation) TRUE UNIQUE
	2	Both households were out-perturbed and a new household in-perturbed FALSE UNIQUE	One of the two original households was out-perturbed and replaced with an in-perturbed household DISGUISED UNIQUE (1/2 chance of identifying a correct household)
	3	All three households were out-perturbed and a household in-perturbed FALSE UNIQUE	Two of the three original households were perturbed out of the output zone and replaced with one in-perturbed household DISGUISED UNIQUE (1/3 chance of identifying correct household)
	>3	FALSE UNIQUE	DISGUISED UNIQUE (less than 1/3 chance of identification)

Which of the scenarios are disclosive? Scenario (1) – True Uniques, would definitely be a disclosure as the intruder would make a correct link to the true household. Scenario (2) – a False Unique, would not be disclosive, the intruder would make an incorrect link. Scenarios (3) and (4) are more difficult to judge and may also be considered disclosive. In Scenario (4), the intruder may make a correct link for that particular table but an incorrect link if using information from other tables[18].

It is helpful to consider how other studies have evaluated protection after disclosure. In Steel and Zayatz (2003) protection was evaluated after targeted swapping by selecting three contrasting US states and examining two tables: one involving targeting criteria and one independent of targeting criteria. Their measure of disclosure risk involved looking at the households with unique combinations of characteristics (assigned a measure of disclosure risk from 1 to 4 depending on variables and geographic level). The measure of risk was the percentage of these uniques that were swapped and the number of households failing to find a matching partner. In other words they consider scenarios (1) and (3) to be disclosive from above.

Shlomo (2005a, 2005b) also discusses some disclosure risk measures after swapping including the probability that a record is perturbed. This measure is calculated as follows. Let z_i represent the record i. I is an indicator function having a value of 1 if true and 0 if false. C_1 is the set of cells with a value of 1, and C_2 the set of cells with a value of 2. $|C_1 \cup C_2|$ is the number of small cells with a 1 or a 2. The disclosure risk measure is then:

$$DR = \frac{\sum_{i \in |C_1 \cup C_2|} I(z_i perturbed)}{|C_1 \cup C_2|} \qquad (3.14)$$

Again this measure defines scenarios (1) and (3) only to be disclosive.

Despite the debate around which scenarios are most disclosive, the primary concern is with reducing the probability of finding true uniques (TU) since these cases relate directly to the true household. This is a conditional probability:

[18] for example, if the intruder used linked tables (based on information about the false household as in section 2.3.2 to make an attribute disclosure

Pr(TU | Observed Unique) = Pr(A = 1 , C = 1 | B = 1) (3.15)

A more conservative approach (scenario 3) would aim to reduce the probability of finding disguised uniques (DU) as well since there is still a chance a correct link can be made. Reducing the number of disguised uniques would then lead to the overall probability of making a correct link being:

$$\Pr(TU)+\frac{1}{F_c^o}\Pr(DU)=\Pr(A=1,C=1\mid B=1)+\frac{1}{2}\Pr(A=2,C=1\mid B=1)+\frac{1}{3}\Pr(A=3,C=1\mid B=1)+\ldots$$

(3.16)

These concepts of disclosure risk do not take into account the perception of disclosure. This refers to whether or not the intruder believes they have made a disclosure. Users of the data may perceive that the data are not confidential when counts of ones can be seen despite a note stating that SDC has been applied. In such a context, Pr(B = 1) would need to be reduced to 0, in other words, uniques would not appear in the table at all. This may be achieved by rounding small cell counts to base 3 (a.k.a small cell adjustment) for example as was carried out in the England & Wales Census, 2001 (Shlomo, 2006).

Furthermore false uniques as defined in table 3.3 might also be considered risky in this context particularly in the case where $F_c^p = 1$ and $F_c^o = 1$. If the cell takes the same value of one before and after perturbation, the intruder can still correctly deduce that there is only one person with the same set of characteristics in the area represented by the table, even though it relates to a different household. In fact if the two records swapped are identical then the original database remains unchanged after perturbation. In the actual census, this is unlikely on a full set of census variables. In the following experiments, false uniques will be presumed to be not very risky because even though the intruder would make a correct identity disclosure in that particular table, any other inferences would be incorrect (through linking of other tables) since we assume, as is highly likely with the actual census, that the two swapped households records would differ on the remaining census variables in some way.

3.10.2 Rules on Uniques Relating to Geography

When obtaining results for the numbers of uniques and true uniques after perturbation, it would help to know whether the results are plausible. This subsection defines some rules to help understand the possible outcomes in terms of numbers of uniques and true uniques when comparing hierarchically nested geographies (i.e. between ED and ward level but not between ED and postcode level since the latter are non-nested). In the following scenarios, α denotes a cell unique in the table $F_c = 1$, and the diagrams relate to sub-regions within a larger aggregate region.

Rule 1: There must be fewer or equal numbers of cell uniques in aggregations of nested geographic subregions.

Example: A region is split into nine subregions as shown in the table below.

α	α	
	α	
α	α	α

Rationale: If there are only 6 uniques in the subregions, then there cannot be greater than 6 uniques in the aggregation as a whole. Moreover, a unique at ED level must be unique at postcode level.

Rule 2a: There may be more *true uniques* at the aggregate level than there are true uniques in the nested subregions.

Example: The region is split into six subregions as shown in the tables below. There are 3 true uniques at aggregate level but only 1 true unique in the nested subregions. Let $\alpha_1, \alpha_2, \alpha_3$ represent uniques in three different cells of a table.

Before perturbation

α_1		α_3
	α_2	

After perturbation

	α_2	α_3
α_1		

Rationale: $\alpha_1, \alpha_2, \alpha_3$ are all true unique at the aggregate level (they belong to the same aggregate area) but only one is true unique at the subregional level – only α_3 belongs to the same area. α_1 and α_2 are false unique: they are uniques but don't belong to the same subregion after perturbation.

Rule 2b: There may be fewer true uniques at the aggregate level than there are in its nested subregions.

Example: Suppose the region is split into six subregions as shown in the diagram. In this case there are 3 true uniques at the subregion level but only 1 true unique at the aggregate level as a whole.

Before perturbation

α_1	α_1	α_2

After perturbation

α_1	α_1	α_2

Rationale: In this scenario, the uniques have been perturbed within their subregion so are still all true unique in their subregions after perturbation. In the aggregate as a whole, only α_2 is unique because there are two α_1.

3.10.3 Further Measures of Risk after Perturbation

As mentioned earlier, more complex measures of disclosure risk might take into account attribute disclosure looking at rows (or columns) where there is only one non-zero cell in both the margin and the row. Furthermore it is also important to assess the risk from geographical differencing if this is what the methods are intended to protect against. Geographical differencing can be studied using the

geographical package ArcGIS to assess similar geographical boundary sets and find those which intersect. Geographical differencing can occur by examining comparable geographies from different time periods (i.e. 1998 wards with 2001 wards) or by examining different geographies which are very similar in size. The disclosure risk in these differenced areas could then be assessed as above by treating the slivers as output zones. A simpler approach might be to analyse risk in the smallest geography available, say postcodes, which we would expect to be roughly the size of a 'sliver' - the differenced area. This would give a good indication of the risk from differencing (assuming the disclosure control method is not dependent on postcodes) and is a practical approach since a sliver may in theory, be located anywhere in the census region, when the location of all geographical boundaries is not known in advance. In chapter 5, the procedure for carrying out geographical differencing in ArcGIS is discussed.

3.10.4 Indicators of Damage

In this section we very briefly review some indicators of damage to the data after disclosure control has been applied or measures of *dis-utility*. These measures indicate a loss in the utility of the data and are to be used in our experimental work in chapter 5. A full and thorough discussion of utility is considered in the context of census user needs in chapter 6. The indicators in this section are primarily for comparison between methods in the empirical work.

Shlomo and Young (2006a and 2006b) summarise some measures which assess the quality of disclosure-controlled frequency tables. Simple measures include the Average Absolute Deviation (AAD) which looks at the absolute difference in cell values before and after perturbation.

$$AAD = \sum^{N_T} \frac{\left|F_c^p - F_c^o\right|}{N_T} \tag{3.17}$$

The Relative Absolute Distance (RAD) would also be useful to assess performance across different levels of geography, since ward level frequencies are likely to be larger than at postcode level.

$$RAD = \frac{1}{N_T}\sum^{N_T}\frac{\left|F_c^p - F_c^o\right|}{F_c^o} \quad for \quad F_c^o > 0 \quad else \quad RAD = 0 \tag{3.18}$$

The variance of the counts will also give an indication of whether an area has become more homogeneous:

$$V = \frac{V\left(F_c^p\right)}{V\left(F_c^o\right)} \tag{3.19}$$

where $V\left(F_c^p\right) = \frac{1}{N_T - 1}\sum^{N_T}\left(F_c^p - \overline{F_c^p}\right)^2$ and $V\left(F_c^o\right) = \frac{1}{N_T - 1}\sum^{N_T}\left(F_c^o - \overline{F_c^o}\right)^2$

Steel and Zayatz (2003) measure data quality by taking each cell in the table and finding the percentage of times the unswapped values are captured by the interval when it is placed around the corresponding swapped values. For example 95% of unswapped values might be within X of the swapped values. X could be a 95% confidence interval. They also looked at average change due to swapping in nonzero cells for different geographic levels, different variables and different size cells for three states.

These all measure distortion to tabular output after geographical perturbation has been applied. Another useful guideline to know in regard to utility of the data, would be the approximate percentage of cells that have changed value $\left(F_c^o \neq F_c^p\right)$ in a particular table, after geographical perturbation for a given sampling fraction.

The census data can be defined according to attributes $A_1, A_2, \ldots$ with associated frequencies $F(A_1), F(A_2), \ldots$. These frequencies are split according to how the census region is divided into geographies $O_1, O_2, \ldots$.

After swapping, the total frequencies $F(A_1), F(A_2), \ldots$ do not change. But the distribution of the counts across the geographies $O_1, O_2, \ldots$ change. Suppose we define the census data in terms of tenure, central heating and accommodation type. Then if for example, a 'detached rented with central

heating' household in O_1 is selected (the donor) and swapped with a recipient 'semi-rented without central heating' in O_2 the totals by geography change as follows:

- the total 'detached rented with central heating' in O_1 is minus one household
- the total 'detached rented with central heating' in O_2 is plus one household
- the total 'semi-rented without central heating' in O_2 is minus one household
- the total 'semi-rented without central heating' in O_1 is plus one household

Or for tables of accommodation type by household space:

- the total 'detached rented' in O_1 is minus one household
- the total 'detached rented' in O_2 is plus one household
- the total 'semi-rented' in O_2 is minus one household
- the total 'semi-rented' in O_1 is plus one household

Or for tables of household space by central heating:

- the total 'detached with central heating' in O_1 is minus one household
- the total 'detached with central heating' in O_2 is plus one household
- the total 'semi- without central heating' in O_2 is minus one household
- the total 'semi- without central heating' in O_1 is plus one household

In all tables, for one swap involving two households, four cell counts have changed value. Thus for a 10% swap, we would expect 20% of the cell counts at most to be affected. This is the worst case scenario, when households are not matched on the table defining characteristics and swaps are between two different geographies defining the tables for analysis. Every census attribute or attributes can be represented in this way. Therefore the probability that a cell count is *changed* in a particular table can be no more than 20% for a 10% swap. Similarly with displacement; one displacement affects two cell counts so for a 10% displacement, at most 20% of cell counts in the table would have changed value.

This is a different concept of risk to the false, disguised and true uniques described in table 3.3 because an unchanged cell count can relate to either a true unique or false unique (B = 1 and A = 1). Note that this looks at the utility perspective rather than risk. However it is a good guideline to give an idea of utility according to sampling fraction. Within this bound of the maximum probability of a

cell change in a table; the actual percentage may be much lower dependent on match variables, the distance moved, the number of geographic areas, etc. and this is what the empirical work is designed to assess; the best risk-utility outcome for different geographical perturbation methods (chapter 5).

Chapter 4 Building a Synthetic Population

4.1 Introduction

In chapter 3 some new methods of disclosure control were proposed to protect against geographical differencing. The next stage of the research will be to evaluate the proposed methods. Ideally the methods would be evaluated on the actual England & Wales census dataset. However such data are, of course, not available because of confidentiality reasons. However small samples of census data are released as well as small area tables defined by a limited number of variables but containing the whole population in that area. All releases from the census have disclosure control applied in some form. In this chapter a synthetic dataset is constructed based on combining these sources of census data. The aim will be to create a 100% dataset consisting of the full set of census variables. The census records have not been geocoded (assigned a one metre National Grid reference) but as disaggregate data are required; each unit will also be assigned geographic co-ordinates describing its point location.

This chapter begins by describing the characteristics the synthetic dataset should have. The data that are available from the UK Censuses is discussed in section 4.2. Section 4.3 looks at data available from the England & Wales Census. In 4.4 we review some approaches to creating synthetic census datasets

with particular attention on microsimulation methods (section 4.5). Spatial microsimulation is described in detail in section 4.5.1, a method that takes into account the relationship between census variables and geography. Section 4.5.2 concludes by describing the design of the synthetic dataset using a spatial microsimulation technique. Each household is assigned a geographic point location as described in section 4.6.

The microsimulation process involves a great deal of time and effort (writing the program, downloading and formatting the constraint tables, waiting for the program to run over many days) however it is the best way to obtain a realistic census-like dataset. Using a realistic census dataset is crucial to the final results in chapters 5 and 6 and to the interpretation of the usefulness of the methods.

4.2 Characteristics of the Synthetic Dataset

Household data are required for the purpose of assessing the geographical perturbation methodologies. The performance of the SDC methods will be evaluated by taking into account both the disclosure risk and the utility of the data. Thus the synthetic dataset should have the characteristics not only to carry out the SDC method but should also allow assessment of risk and utility. It should permit analyses of data that census users typically carry out (identified in chapter 2).

The synthetic population will need to consist of individuals in households in order to assess the effect of perturbation on both individual and household distributions. Variables such as tenure, number of people in household (household variables) and age, employment variables (individual variables) are all useful to census users (see section 2.5.4) so ideally the population should contain these variables. These distributions are linked so the effect of perturbation on joint distributions (e.g. age and tenure) should also be considered as well as the univariate distributions. Moreover to properly assess the effect on joint distributions, the individuals within households must be given some structure. Certain individual attributes, in particular age, sex and marital status cannot be assigned at random as this would result in strange households such as nine babies in a household or greater occurrence of unlikely households (five married people in one house).

Size of Dataset / Region of Study

The data will represent the entire population for the region (100% level data) as opposed to a sample. The size of the dataset should be large enough to incorporate both urban and rural regions to reflect a real population. The effect of geographical perturbation methods on statistical analyses that are spatial in nature should be considered as these relationships are likely to be distorted. For example the effect on multilevel models could be analysed or the change in patterns of spatial variation. This means the study region must be diverse. The county of Hampshire is a suitable candidate. Hampshire has a varying population distribution as illustrated in figure 3.7. It also includes the densely populated urban areas of Southampton and Portsmouth which have reasonably diverse populations in terms of wealth and family household types and also include a mix of students. Hampshire also contains some fairly rural areas such as the New Forest which is characterised by 'villages with wealthy commuters' (ACORN profile). However the dataset should not be so large that the experiments become too computationally intensive. If necessary, a smaller region such as the Basingstoke and Deane LAD could be analysed in isolation.

Spatial Distribution of Households

An important feature of the dataset is the distribution and concentration of the household locations. Geographical perturbation is a point based method so the households need to have geographical co-ordinates specifying point locations. The distribution of the households could be random or clustered. The points could be drawn from a specified distribution with some spatial correlation between points; for example independent points but correlated nearest neighbours. Alternatively the spatial locations could be predetermined from an existing dataset for that region. 'Address point' data are available from the Ordnance Survey* providing unique locations for every residential, business and public postal address in Great Britain. However this is only available for a fee so we can instead make use of a postcode file which provides details on the spatial location of postcode centroids (see section 4.6.2.). Noise could be added to the centroids to generate imaginary locations of households in the immediate area.

Ensuring the Data are Disclosive

In theory, if a realistic population is created, then the tables should automatically possess disclosive properties. Risky tables tend to arise when there are:

- Detailed tables with more than two variables
- Variables with skewed or non-uniform distributions such as age in single years or ethnicity
- Outliers in the data produced from rare combinations of variables (e.g. 16-year-old widow)
- Tables based on a sparse population, i.e. rural areas or the lowest level geographies e.g. enumeration district

Small area tables based on similar population frequencies as in the real population will naturally result in small cell counts. If necessary, outliers can be created in the data at higher aggregate levels by modifying the counts slightly. At a later stage, multiple tables may be generated from the synthetic microdata based on similar geographical boundaries so that the extent of differencing can be measured. Providing the households in the data have geographic co-ordinates, it would be possible to assign households to different sets of geographical boundaries using GIS software.

4.3 Data available from the Census

The ten-yearly Census is designed to collect information about every household and individual in England & Wales. A Census form is filled in by each household which contains questions on accommodation, relationship within the household as well as individual questions on demographics, cultural characteristics, health, qualifications, and questions on employment. The CCSR (Cathie Marsh Centre for Survey Research) website describes the microdata available from the Census. The Samples of Anonymised Records or SARs are samples drawn from 1991 and 2001 census data some of which are publicly available (can be downloaded from the internet) and have had identifying information removed to protect confidentiality (ONS, 2004a). There are two SARs; the Individual SAR and the Household SAR. These are samples of between one and five percent; the Individual SAR was a 2% sample in 1991, increased to 3% in 2001 but only released under licence. The Household SAR has been released as a 1% sample for both Censuses but was also licensed in 2001. The degree of geographical definition available is related to both sample size and reliability of estimates, as well as

confidentiality considerations. The minimum size of a geographical area in the 1991 Individual SAR was 120,000 population. This is large enough to be able to analyse small groups and sub-regions. The Individual file had the most detailed geography with most LADs separately identified (smaller LADs with under 120,000 population were grouped). In contrast, the 2001 SAR did not contain geography lower than Government Office Region (GOR). In the 1991 household file, the lowest level of geography released was Standard Regions with further subdivisions in the South East. In 2001, no geographic breakdown was made available for the Household SAR and the data are limited to England & Wales only (ONS, 2004b). Since the 2001 SAR was not yet publicly available when carrying out this research, the focus will be on 1991 census data. We note that the 2001 SAR would have been of limited use anyway due to lack of geographic detail.

The 1991 SARs cover the full range of Census topics including housing, education, health, transport, employment and ethnicity. In both files, households with more than twelve persons have been top-coded. The household SAR retains the structure of people in households and contains more detailed variables. Each record in the Household SAR also has an ONS classification attached to it called ONSCLASS. These classifications were calculated using the entire census dataset. Wards were classified into fourteen types based on the census variables. The ONSCLASS variable is very useful. A user of the Household SAR only knows the SAR region in which a household is located. However the ONSCLASS variable provides information about the characteristics of the ward the household is located in - so it tells the user something about the household at a much lower level geography. ONSCLASS can be used as a link to allocating households to wards within Hampshire rather than allocating at random. Because of confidentiality reasons, some extra ONSCLASS categories were created for potentially disclosive records. This is discussed again later on in this chapter.

The 1991 Small Area Statistics (SAS) are tables released from the census database which constitute the fully enumerated population for every Enumeration District which is typically around 200 households. The SAS are less detailed but are available at the highest geographical resolution (ED/OA). The data are arranged into tables defined by two or three variables. A wide range of variables is available to define the tables; there are 86 tables in total. The tables are either based on households or individuals (but not both together). A few of these small area tables contain only 10% sample data for hard-to-code variables (for example socio-economic group). Tables are only published for areas with populations of at least 50 people or 16 resident households. Furthermore,

cell counts are modified by the random addition of 0, -1 or +1 at the ED and ward level. This means that counts of the same population in two different tables may not necessarily be the same. Data for single EDs can be unreliable, especially for a variable where the count is low.

4.4 Approaches to Creating Synthetic Census Datasets

This section presents a brief review of the literature examining previous studies where synthetic census datasets have been created. Duke-Williams and Rees (1998a) carried out a study to assess the extent of the differencing problem for UK geographical boundaries. This involved creating a database of household locations along with information about the individuals who inhabit them. The Household (1%) Sample of Anonymised Records was used from the 1991 Census to form 2 million households and 5 million people living in Yorkshire and Humberside. This region represents 8.6% of the UK population and was used as it contains major cities and large, sparsely populated rural areas. The synthetic population was created using a set of ED boundaries, together with data published from the Small Area Statistics regarding the number of households in each ED. In each ED, a number of points were randomly selected, one for each household contained in the SAS. A household from the SAR was allocated at random to one of these points. The population had a realistic spatial distribution but no patterns in terms of socio-demographic characteristics because of the random allocation of households. For the purposes of our research it is very important that there is a realistic distribution between geography and the socio-demographic characteristics as one of the aims will be to see how geographical perturbation methods distort these relationships. However one advantage of the above method is that it solves the difficult problem of getting the correct correlation structure between persons in households by allocating households from a real dataset. If the correlation structure between individuals in households is ignored, then strange results may appear such as five babies in a household.

Another approach might be to use conditional distributions based on real census data as they did in the DACSEIS study (see Munnich and Josef, 2003). This project required a synthetic dataset as similar as possible to the German microcensus data to be used for simulating and evaluating variance estimation methods. In this simulation study, they considered only correlations within age and gender and assumed that other correlations would be implicitly generated by the influence of age and gender. Given a household of a particular size (e.g. 4 inhabitants) and in a particular region, they drew

households at random from a real dataset. They assigned the age and gender of the individuals in the household to the artificial household. The remaining variables were simulated by conditioning on age, gender, household size and region and modelling the multivariate frequency distributions.

In Chen and Keller-McNulty (1998), simulated datasets were used for demonstration of a new technique which estimates the risk in microdata. The data were simulated to illustrate a population of a large size containing 500,000 records and having six key variables. They mention that it is not clear exactly what a 'real' population dataset should look like as there are many relationships among the distributions of key variables that could exist. To have some structure in the data, they used a six-dimensional multivariate normal distribution with a covariance structure for 15 discrete scales of human response. The covariance structure among the first six variables was used in the simulation. It covered a range of correlation values from 0.0 to 0.8. Once the six-dimensional multivariate normal data were generated, they were discretised by dividing each marginal range into severally equally spaced intervals, one for each category. Cross-classifying the six variables created 15,047 cells with a positive count and 4,479 population uniques.

A second smaller simulated dataset was also created. Five variables were used to generate this dataset that were mutually independent with uniform marginal distributions. Therefore, the joint distribution was also uniform which is rarely the case in reality. There were 100,000 records and 65,640 non-empty cells. There were 40,217 population uniques. The problem with this type of approach which reduces the data to a multivariate normal or uniform distribution, is that it is likely to fail to capture all the complexities of a census dataset necessary for our purposes, in particular for assessing utility.

Voas and Williamson (2000) describe a method for creating synthetic data which builds on the approach taken by Duke-Williams and Rees (1998a). Instead of allocating households at random from the SAR, *spatial microsimulation* methods can be used to find the best fit of SAR households constrained by the ED characteristics. This makes the data more realistic. Microsimulation techniques have been developed in response to the lack of spatially disaggregate data available in official statistics. We describe microsimulation techniques in more detail in the next section.

4.5 Microsimulation

In this section we consider microsimulation techniques in more detail as this appears to be a satisfactory way of replicating the properties of a census dataset; giving individuals within households some structure and finding the best fit of households to small area characteristics. We first consider microsimulation methods in general and then explore methods of spatial microsimulation in section 4.5.1. A general definition of microsimulation was provided by Ballas et al. (1999; p.5):

> *'Microsimulation is a methodology aimed at building large-scale datasets on the attributes of individuals or households and on the attributes of individual firms or organisations and at analysing policy impacts on these micro-units.'*

Their definition refers to *'analysing policy impacts'*. Microsimulation was first conceptualised in the context of economics in Orcutt (1957) who noted that models of socio-economic systems had limited prediction usefulness and could only predict aggregates and not distributions of individuals or households needed for decision making. The new type of socio-economic system consisted of interacting units which receive inputs and generate outputs. The outputs are the results of a series of random drawings from discrete probability distributions which specify the probabilities associated with the possible outputs of the unit. The units are can be individuals, families, firms, etc. This idea was developed further and has been widely used in governments around the world for the analysis of redistributive policies and budget changes. Mertz (1991) noted that microsimulation models are especially well suited to estimate and analyse distributional impacts of policy change. However, today many types of microsimulation models are used with a wide range of specialities. All different types of entities can be constructed, including people.

Until recently microsimulation was aspatial, so the emphasis was on who is affected rather than where the individuals are located. Results were only available at the national level because existing models were constructed on top of sample survey data which did not go down to small geographic levels. Consequently it was not possible to use these models to predict spatial effects.

In response to a need for spatial models, spatial microsimulation has been developed which blends census and sample survey data to create synthetic small area populations. The creation of micro-level

databases to represent information on individuals in households has a number of benefits including efficient storage, flexibility of spatial and other aggregation, data linkage and the ability to update and forecast. In the UK, spatial microsimulation models have been developed by Leeds and Liverpool universities; see the examples of Williamson et al. (1998) and Ballas et al. (1999). In the example of Ballas et al. (1999), they modelled the impact of national social policies by using HES (Hospital Episodes Statistics) data and reweighting to detailed sociodemographic profiles from the Census for each district. Microsimulation models can be dynamic, projecting each individual in the simulated database into the future, or they can be static, applying to a specific time period only. Social science models tend to be static and either deterministic or stochastic in nature. Stochastic modelling is based on conditional probabilities that certain social conditions or processes will occur - for example, the likelihood that an 18 year old from a high income family will attend university. Wolf (2001) describes microsimulation as drawing a sample of realisations from a prespecified stochastic process. Microsimulation thus entails the generation of data (a set of realisations).

4.5.1 Static Spatial Microsimulation Models

As Ballas and Clarke (2001) note, there are three approaches to static spatial microsimulation modelling:

1) Probabilistically on the basis of random sampling and optimisation.

2) Deterministically without the use of random sampling.

3) Using synthetic probabilistic reconstruction models which involve the use of random sampling.

- *Synthetic Probabilistic Reconstruction*

Synthetic probabilistic reconstruction models were used before the SARs became publicly available and microlevel populations had to be generated by synthetic sampling from only SAS tables. As Birkin and Clarke (1995) point out, the power of representing data at the individual level becomes immediately apparent if we contemplate the list of characteristics from the UK Census. Consider the full list of twenty-four characteristics from the 1991 Census form. There are a huge number of

possible states: - even if there are only two categories per variable (as with gender), then there are still two to the power of twenty-four or sixteen million states or permutations. Microsimulation techniques aim to estimate this joint distribution by merging data from different sources considering three or four variables at a time. Clarke (1996) describes how there are two stages to this process. The first stage is to find conditional probabilities from available known data in order to reconstruct detailed micro-level populations. An example is given; assume we want to investigate the relationship between sex (SEX), age (AGE), educational qualifications (EQ), economic position (EP) and socio-economic group (SEG) for a given population group in area O_v.

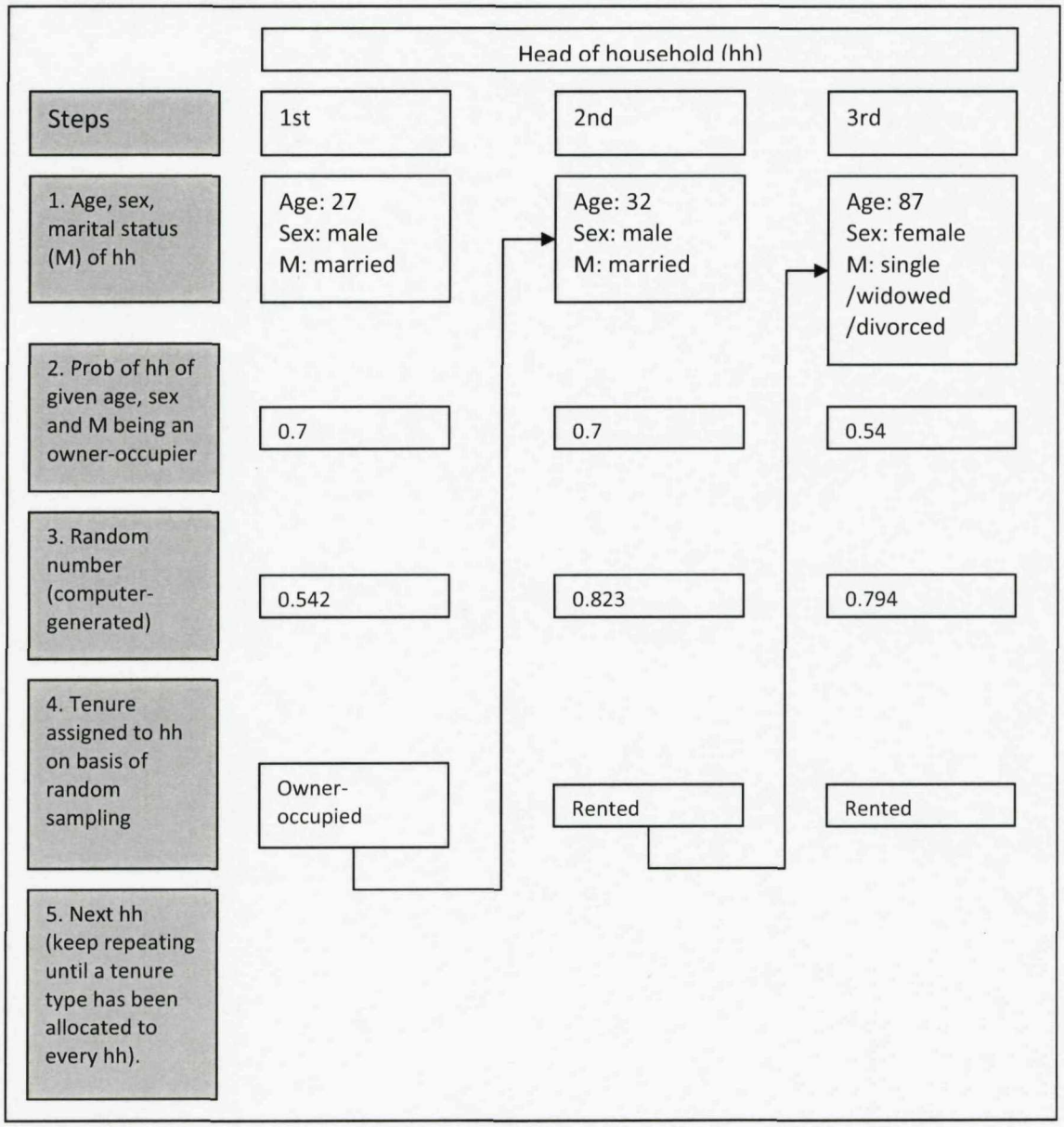
$$p(O_v, SEX, AGE, EQ, EP, SEG)$$

The task is to generate a single dataset containing the attribute set as above for the population of interest in the specific area O_v. These are the set of constraints or known probabilities from SAS tables:

$$p(O_v, SEX, AGE, EP)$$
$$p(O_v, EQ, SEX)$$
$$p(O_v, SEG, EP)$$

The aim is to generate a joint probability distribution for this attribute vector and in the second stage of the microsimulation procedure, synthetically create or extract a sample of individuals based on the distribution. The idea is to build up one attribute at a time, so that the probability of certain attributes is (conditionally) dependent on existing attributes. The difficulty is in selecting the order of the attribute dependencies.

Figure 4.1: An Example of the Microsimulation Process. Tenure Allocation Procedure (Reproduced from Clarke, 1996)



C

larke (1996) shows how this procedure can be employed for the creation of a micro-level population with the following characteristics: age, sex and marital status of the head of household as illustrated in figure 4.1. The method works probabilistically on the basis of random sampling for the creation of a micro-level population with the population characteristics: age, sex, marital status and household tenure. Suppose that age, sex and marital status of the head of household is available from the

-

census, it is then possible to estimate probabilities of household tenure. The first synthetic household has the following characteristics: male household head, age 27, married. The estimated probability that a household of this type would be owner-occupied is 70. The next step in the procedure is to generate a random number to see if the synthetic household gets allocated to the owner-occupier category. The random number in this example is 0.542 which falls within the 0.001 to 0.700 range needed to qualify as owner-occupied. The same procedure is then carried out sequentially for the tenure allocation of all synthetic households.

- *Probabilistically on the basis of random sampling and optimisation*

An alternative microsimulation technique is to reweight an existing micro dataset to fit a geographical area based on sampling and optimisation. Reweighting is a newer technique which makes use of the Census SARs. A crude way of doing this would be to reweight the 1% household sample by duplicating every household 100 times. A more sophisticated approach commonly used is based on the technique of Iterative Proportional Fitting (IPF). This works by adjusting a two-dimensional matrix iteratively until the row sums and column sums agree with row and column totals from alternative sources (Williamson et al., 1998).

A third variant of the reweighting approach is to view the households within the SAR as the parent population from which households can be drawn to recreate the population of an individual ED, as described by Williamson et al. (1998). This is a combinatorial optimisation approach which attempts to select a combination of households from the SAR that reproduces the characteristics of the chosen ED. Constraints on the combination of households chosen are provided by known tabulations of ED data from the Census SAS. The choice of SAS tables to use depends on the importance of the topics relevant to the study, the extent of correlation with other variables and computer resources - every additional table increases the number of iterations required to achieve a given level of fit.

In the next section, we consider the third variant – combinatorial optimisation in more detail. This is a simple concept which should produce a good approximation to the census data. Fit to the unconstrained distributions can easily be tested if required. Moreover it gives a real population of people in households unlike IPF which modifies existing tables. The main disadvantage of this approach compared to the other methods is that households are sampled from the SAR so if certain

types of households are not in the SAR, then they won't end up in the synthetic population; thus an exact fit may never achieved in some cases. However this is not important for our study; the population just needs to be realistic; not an exact fit.

4.5.2 A Combinatorial Optimisation Approach to Spatial Microsimulation

The idea behind this technique described by Williamson et al. (1998) is to allocate households based on their fit to the characteristics of each ED. Furthermore, it is possible to allocate other individual characteristics other than just age, sex and marital status. The method works as follows (with an example shown in figure 4.2):

STEP 1. Variables and corresponding SAS tables are selected to be used as constraints (figure 4.2a). Both tables defined by individual level variables and tables defined by household level variables can be used.

STEP 2. Using the Household SAR, n_v = 108 initial households are sampled at random. The number of households (n_v) sampled is determined by the known number of households in the enumeration district. Figure 4.2(b) shows a list of sampled households.

STEP 3. All n_v sampled households are tabulated based on the format of the SAS table(s) specified as constraints in step 1. See Figure 4.2(f).

STEP 4. The sampled data are assessed to see how well it fits the SAS table(s) according to some criterion of error. For example the *TAE* or Total Absolute Error is commonly used which looks at the sum of the absolute differences between the fitted and real SAS tables (N.B. this is the same as the AAD utility measure described earlier in 3.17). If the sampled data does not fit the SAS table well (the *TAE* is large), then a record is selected at random from the sample and swapped with a new record at random from the SAR. If the fit is better, the swap is kept, otherwise a different swap is tried.

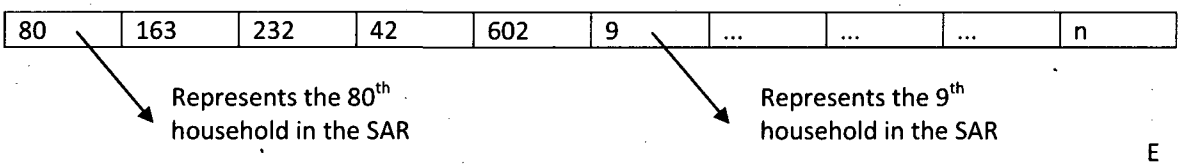
STEP 5. Steps 3 to 4 are repeated until the error is minimised.

Figure 4.2: The Microsimulation Procedure

Figure 4.2(a): Example SAS table required (household level variables)

		Tenure		
		Rented	Owned	Total
Number of cars	0	34	6	40
	1	6	14	20
	2	1	17	18
	3	2	23	25
	4+	2	3	5
	Total	45	63	**108**

Figure 4.2(b): Sampled Households from the SAR



E

ach household comprises a household record:

Figure 4.2(c): A Household Record from the SAR

Household number	...	Number of cars	...	...	Tenure	...	...
80	...	3	...	...	Owned	...	...

Figure 4.2(d): And one or more individual record(s)

Household number	...	Household Member	Age	Marital Status	...	Sex	...
80	...	1	40	Married	...	Male	...
80	...	2	38	Married	...	Female	...
80	...	3	11	Single	...	Female	...
80	...	4	8	Single	...	Female	...

One way of representing this household is as a tabulation of age by sex and marital status:

Figure 4.2(e): Representing sampled individuals in a table

	Male		Female	
Age	Single,w,d	Married	Single,w,d	Married
0-4				
5-9			1	
10-14			1	
...				
30-34				
35-39				1
40-44		1		
...				

Figure 4.2(f): Tabulating Sampled Households

1	...	...	...
2	...	...	...
3			
4			
5			
6			
7			
...			
...			
...			
...			

Select n_v households at random and tabulate into SAS table

	Single/w/d	Married	Total
<20	...	...	...
21-35	...	...	...
36-50	...	...	...
51-65	...	...	...
65+	...	...	...
Total	...	...	...

The procedure results in a sample of households 'best fitting' each ED. The characteristics of these households should be close to those of the ED for the constraining variables (depending on how far the error is reduced). The other household variables not used as constraints may not fit so well, depending on their degree of association with the constraining variables.

Ideally one would examine every possible selection of households that give the best fit to the SAS tables but the number of possible solutions is huge. Instead the fit can be measured in an iterative process. Various searching methods can be used such as hill-climbing, simulated annealing and genetic algorithms. Hill-climbing is the simplest method, starting from an initial set of households chosen randomly from the SAR and the effects of replacing one of the selected households with a fresh household from the SAR are considered. If the replacement improves the fit, the households are

swapped. If not, the swap is not made. This process is repeated many times, with the aim of gradually improving the fit between the actual data and the selected sample of SAR households.

As Williamson et al. (1998; p.794) noted,

> *'The main advantage of the hill climbing approach is the speed with which improvements in the initial combination being considered can be obtained. Hill-climbing is far less complex than synthetic population reconstruction plus far less subjective judgement involved.'*

When searching for the best solution given an extremely large number of candidates a decision must be made at some point to end the search. The optimum solution (perfect fit) is likely never to be found. Instead an acceptance threshold should be set that produces satisfactory results. Voas and Williamson (2000) provide a global measure of total absolute error across the SAS tables used as constraints. The disadvantage of any global measure is that one table might fit very well but another not so well. The suggested approach is to first fit the most unusual table then proceed as before in finding a global measure of fit (sequentially) so making sure the most unusual table is fitted first, then make adjustments to the subsequent SAS tables until they fit (whilst making sure no adverse changes to the first table).

Goodness of Fit Measures

Once the synthetic data have been created, a simple measure of goodness of fit to the real data is needed and thus whether the dataset is suitable for the empirical work in chapter 5. Some possible goodness of fit measures are listed here, with respect to the microsimulation procedure.

1. Visual Observation against Existing data: The difference between the unconstrained variables in actual small area tables can be compared with the synthetic small area tables. A further test can be performed to compare the tables such as Chi-Square measure of association.

2. Spatial mapping: In Ballas et al. (1999) five different microsimulation methodologies were used to assess the best approach to generating a model for the Leeds labour market including incomes, household wealth, taxation and welfare benefits. Both Small Area Statistics and data from the SAR

were used. They used a variety of measures to assess goodness of fit including mapping the TAE (total absolute error) by ED as well as mapping the generated spatial distributions of variables in each ED. The latter were compared with census proportions from the SAR.

3. Error per cell: Ballas et al. (1999) also examined error per cell in their study to assess a number of microsimulation methodologies and decide which is best.

4. Z statistic: Voas and Williamson (2000) used combinatorial techniques to produce estimates of the distribution of young urban professionals within the city of York. Williamson et al. (1998) also did a study estimating population microdata and assessed the suitability of the hill climbing algorithm. Both papers use the Z-statistic to assess goodness of fit. The Z-statistic gives a normal score for each table cell and is based on the difference between the relative size in that category in the synthetic and actual populations with an adjustment for counts of zero. The modified Z-statistic looks at relative rather than absolute frequencies and is:

$$Zstat_{ij} = \left(r_{ij} - p_{ij}\right) \Big/ \left[\frac{p_{ij}\left(1-p_{ij}\right)}{\sum_{ij} O_{ij}}\right]^{\frac{1}{2}} \qquad (4.1)$$

where $p_{ij} = \frac{O_{ij}}{\sum_{ij} O_{ij}}$ and $r_{ij} = \frac{E_{ij}}{\sum_{ij} O_{ij}}$

O_{ij} is the observed SAS count for row i in column j, and E_{ij} is the estimated count for row i in column j.

With the unmodified Z-statistic, the distribution of errors is assumed to be normally distributed and so the critical value at the 95% significance level is 1.96. Estimated table counts with associated modified z-scores of greater than or equal to 1.96 can then be flagged as being unacceptably dissimilar from the observed counts they are meant to fit (Williamson et al. 1998). A similar significance level may be assumed for the modified Z-statistic.

5. Departure from averages: This approach is also suggested by Williamson et al. (1998). The synthetic data can be assessed to see whether the output zones represent the geodemographic class they are drawn from or the extent of departure from the generated EDs with that of the national average can be examined.

4.6 Empirical work: Creation of the Synthetic Dataset

In this section the theory is put into practice by building a synthetic population for the Hampshire region. This involves two parts. Firstly Williamson et al.'s spatial microsimulation technique is used to create the attributes of individuals in households at ED level. Secondly spatial point locations are created for the households using a file describing postcode centre locations. As mentioned earlier, 1991 census datasets will be used to carry out the entire procedure because the 2001 SAR was not available at the time of writing the program. Moreover the Household SAR will be used to sample households rather than the Individual SAR (which only contains individuals) to get the correct structure of people in households (see section 4.2).

4.6.1 Part 1: Creating Attributes of Individuals within Households.

The first step is to decide which small area tables should be used as constraints. The more SAS tables there are and the higher their dimension, the longer the process will take. Only a limited number of variables should be chosen. Williamson et al. (1988) suggest at most nine or ten tables. A good selection of constraining tables needs to be selected in order that the correlations between the remaining variables are strong.

The next stage involves selecting an initial sample of households for each Enumeration District (ED). Voas and Williamson (2000) note the difficulty with the microsimulation process is fitting tables which are 'atypical' relative to the national distribution. These atypical EDs might have atypical values for a particular variable or have a few unusual people unrepresentative of the national distribution. If the initial sample is unrepresentative of the distribution in the SAS table, then many households will have to be resampled taking a long time for the error function to converge. One way around this may be to use stratified sampling to influence which households are drawn from the SAR. The stratification

variable must be available for (i) a lower level of geography (on the SAS) and (ii) present on the Household SAR. A suitable stratification variable mentioned earlier is ONSCLASS[19]. This variable is available on both the SAR and the SAS. Households will be sampled from the stratum which represents the same ONSCLASS classification as that of the ED. Although no experiments will be done to see whether this leads to better fitting and quicker results, it seems much more likely that the initial sample will be more representative of the ED and it should take fewer swaps before the error function reaches an acceptable threshold.

Table 4.1 gives further details on the various categories of ONSCLASS.

Table 4.1: Classification Types for ONSCLASS Variable

1991 Ward Classification	ONSCLASS grouping
Suburbia	1
Rural Areas	2
Rural Fringe	3
Industrial Areas	4
Middling Britain	5
Prosperous Areas	6
Inner City Estates	7
Established Owner-Occupiers	8
Transient Populations	9
Metropolitan Professionals	10
Deprived City Areas	11
Lower Status Owner Occupier	12
Mature Populations	13
Deprived Industrial Areas	14

One of the smallest ONSCLASS groups comprised only 1% of the SAR, therefore some ONSCLASS categories were more popular than others. This shouldn't be a problem as the smaller ONSCLASS groups should be represented by fewer EDs in the full census.

The microsimulation program will use replacement sampling as this avoids the loss of very unusual households from the sampling pool, however there will be a slight risk of introducing unintended dependencies amongst variables owing to duplication of households within an ED.

[19] Full details on ONSCLASS can be found in *http://www.ccsr.ac.uk/sars/publications/Areaclassifpap.pdf*

The 1991 SAS tables to be used have been disclosure-protected such that small cells have been modified by +1, 0 or -1 using the Barnardisation technique (applied twice in small areas). Since these adjustments are only minor and given that the microsimulated-SAS-tables are very unlikely to be a perfect fit anyway, we assume the tables represent the original values. Unfortunately the 12-person households in the SAR (our parent population) have had some variable values suppressed for disclosure control purposes. However there should still be plenty of uniques in the data. As for creating special uniques, an assessment could be made of the disclosiveness of the synthetic data and a small amount of special uniques made-up if necessary.

Finally Shipping EDs are present in the small area data. These are virtual zones without any physical area attached to accommodate persons on board ships. The shipping zones will be discounted.

Description of the Spatial Microsimulation SAS Program

The microsimulation program for simulating households was written in SAS. The following describes the microsimulation program step by step with a basic overview in Figure 4.3. The procedure is described in a general context, ending with specific details for implementation in Hampshire.

Before starting the microsimulation process, the appropriate data must be read into the program. Four small area tables at Enumeration District level were selected from the census small area statistics website site. Shipping EDs and EDs containing only communal establishments have been omitted. We will only be concerned with households in our perturbation methods and since communal establishments are not households we omit them.

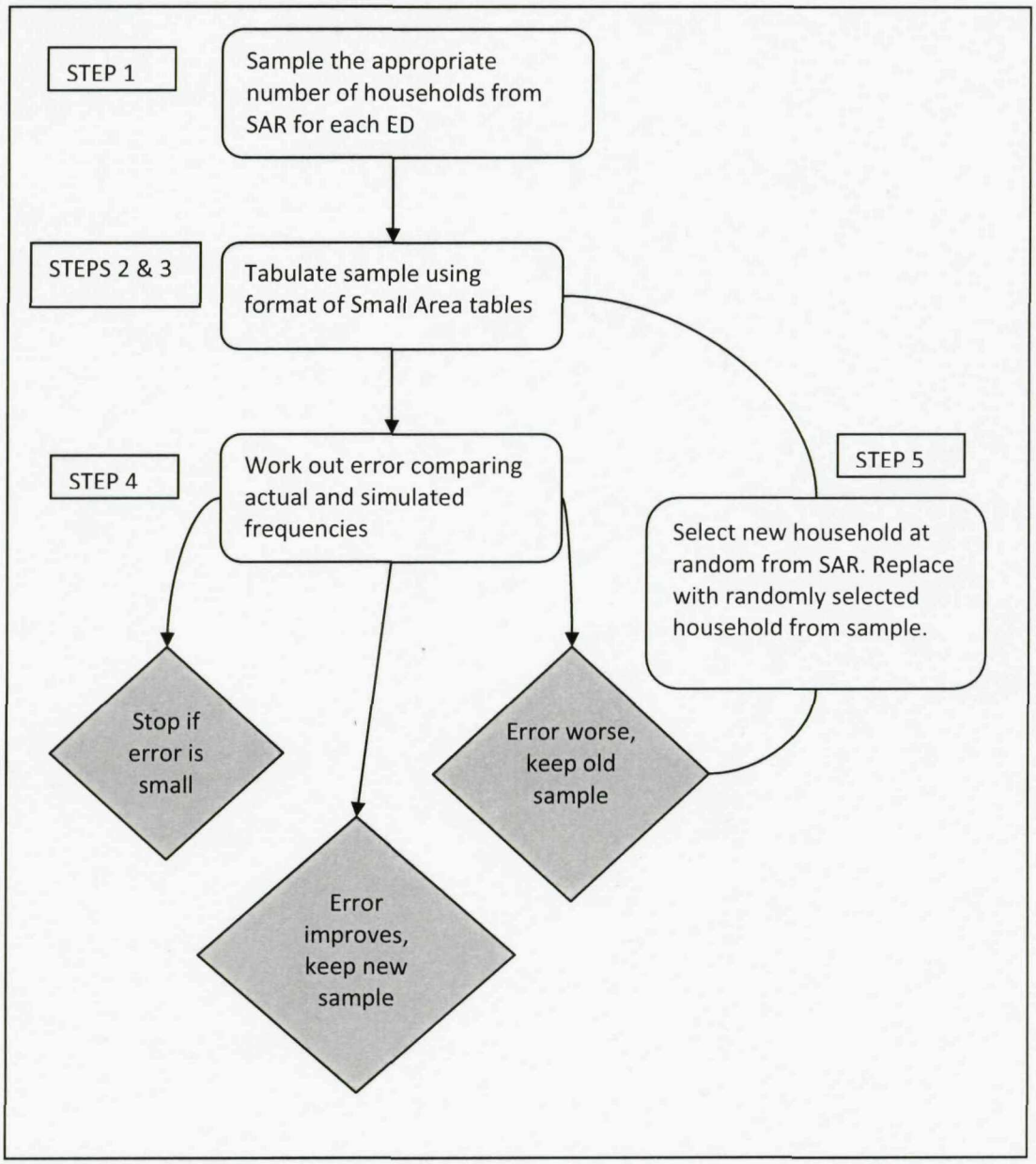
The presence of communal establishments leads to a slight inconsistency between the household and individual level tables. Whilst residents in communal establishments are represented in individual level tables, the same residents are not represented in household level tables. Tables with only private household residents are not supplied by ONS. Therefore simulated EDs consisting of a large proportion of communal establishments will have a greater number of individuals per household than in reality as the communal establishments are missing from the household tables. Since the simulated population is only an approximation, this inconsistency can be been ignored. Moreover the Sample of Anonymised Records (SAR) for households also omits communal establishments.

Earlier on a stratification variable was mentioned which is attached to each record in the SAR called ONSCLASS. Amendments were made to the ONSCLASS variable in the SAR to meet ONS confidentiality requirements. There is a potential disclosure risk when only a small number of wards have a particular value of ONSCLASS. Therefore ONS developed additional ONSCLASS codes based on probabilities of being in any two existing categories to create uncertainty thereby reducing this disclosure risk. However these additional codings do not directly relate to the 1991 ward classification. We use simple probabilities to replace these additional codings with one of the original fourteen. For example; a category '24' was created in Outer London which could represent a ward taking ONSCLASS '8' (63 wards) or '13' (3 wards). In this analysis households in the SAR with an ONSCLASS greater than '14' are; following the example, if they had ONSCLASS '24', reassigned either '8' with a probability of 63/66 or '13' with a probability of 3/66.

Finally a population base dataset needs to be imported to find EDs containing only communal establishments. An ED containing only communal establishments is defined as having at least one individual but containing zero households.

Once the data have been imported, the microsimulation process can begin. Figure 4.3 describes each step of the process.

Figure 4.3: Overview of the Microsimulation Process



- Step 1: Sampling initial households

An initial sample of households needs to be generated for each ED. The SAR is divided into strata according to the ONSCLASS variable. Households are sampled from the appropriate stratum corresponding to the ONSCLASS of the ward the ED falls in. The total number in the sample drawn is equal to the total number of households in the ED (specified by the small area tables at household level).

- Step 2: Define formats

In many cases, the coding of a variable in the SAR is different to the coding of the variable in the Small Area tables. For instance, the SAR contains five codes for marital status (single, married, divorced, remarried, widowed) whereas the Small Area tables have only two codes (single or married). This step therefore involves recoding the SAR variables in order to tabulate the sampled households.

- Step 3: Tabulate sampled households

Frequency tables should be produced for each ED based on the initial sample of households.

- Step 4: Comparison of simulated frequencies to actual frequencies

The true frequencies can then be compared with the simulated frequencies to see how close the fit is.

- Step 5: Iterative Fitting to True Table Frequencies

Finally new households can be selected for replacement and kept if the error is improved. This is an iterative process. We will now describe the measure of fit in more detail.

Selecting the Measure of Fit

At each iteration, a measure of fit must be used to compare the true ED tables with the simulated tables. Let O_c refer to the observed value and E_c refer to the simulated value in cells of the table. Figure 4.4 shows various options for calculating measure of fit.

Figure 4.4: Calculating Measure of Fit

Measure 1: The Standard TAE $TAE = \sum_{c=1}^{C} \lvert O_c - E_c \rvert$
Measure 2: TAE ignoring differences of less than 3. $TAE > 3 = \sum_{c=1}^{C} \max[\lvert O_c - E_c \rvert - 3, 0]$
Measure 3: "Chi-Square" type goodness of fit $\chi^2 = \sum_{c=1}^{C} \frac{(O_c - E_c)^2}{O_c}$

Measures 2 and 3 are two new measures. Measure 3, the chi-square, was devised with the idea being to make a poor fit in small cell counts more important. $TAE > 3$ was created as it seemed that a difference of one or two was not worth trying to improve; since the population should only approximate the true distribution. To decide on the most appropriate measure, experiments were run on a couple of sets of initial sample households. 4,000 iterations were run and the fit of the tables at the end analysed. Unexpectedly, the final tables fitted were very similar despite the fact that the chi-square measure was devised to focus on fitting smaller cells over larger ones. Exactly the same poorly fitting cells stood out under all three criteria. Also the tables under $TAE > 3$ and standard TAE showed little difference, despite expecting the standard TAE to give a 'closer' fit. The deciding factor in choosing to use the $TAE > 3$ was that it was easier to interpret than the chi-square and testing

showed it appeared to converge slightly quicker than the usual TAE. The final fitting procedure can then be summarised as follows in figure 4.5:

Figure 4.5: Microsimulation Fitting Procedure and Calculating TAE

- Set initial sampled households
- Set initial frequency tables
- Find overall $TAE > 3$

1. If $TAE > 3$ is smaller than threshold then STOP; else CONTINUE;

2. Drop a random household

3. Using Simple Random Sampling, select one replacement household from SAR of appropriate stratum.

4. Produce new tables with updated frequencies (including sub-totals)

5. Find the new $TAE > 3$ for each table and overall $TAE > 3$

6. If overall TAE is improved then keep replacement household otherwise drop newly selected household

7. Keep a log of the $TAE > 3$ at each iteration

8. Do the full set of iterations before going to next ED

Implementation for the Hampshire Region

Here we describe implementation of the procedure using the 1991 SAR and SAS tables. The four small area tables chosen to be fitted were:

- Table 02: age, marital status and sex of individuals (individual level) - *84 cells*
- Table 58: household space type and tenure of households (household level) - *35 cells*
- Table 08: primary economic position and age (individual level) – *99 cells*
- Table 86: tenure and socio-economic group of the head of household (household level) 10% sample grossed up to the 100% level – *76 cells*

Only four tables were used as constraints rather than the nine or ten used in other studies. This was to decrease computational time and for practical purposes but mainly because these four tables should be enough to capture the essential properties of the census population that are necessary for a realistic dataset. Variables were chosen in the tables above that would generally be strongly correlated with the remaining (non-selected) census variables. For example, the number of cars a household has is likely to be strongly correlated with the constraining variables tenure, occupation, and age. To simulate household level tables, only heads of household were sampled from the SAR. The entire SAR was used (not just those households in the Hampshire region), because the sample consists of only approx. 500,000 households altogether and a population of approx. 500,000 households needs to be simulated to represent Hampshire. Thus there is greater need to avoid duplicating households rather than having exact fitting results. Therefore the whole file was used. Households were omitted from the SAR if any of their records composed missing values for the above variables.

Regarding the number of iterations; since each ED took half an hour to fit the four tables (from preliminary experiments), and bearing in mind the resources available, this was set to 900 iterations. This implies that only 900 households combinations were tried. 900 iterations should have allowed for most EDs to converge (error stabilising after a fast initial improvement) and leaves a little longer for those EDs that are harder to fit. The $TAE > 3$ was then calculated for each of the four small area tables separately. The overall TAE was calculated by summing the four TAE scores. If this improved sufficiently then the simulation could stop. The error score of most EDs only declined down to about 300 at the least; however when averaged over all the cells in the table – this is still less than one count out per cell (on average). We next examine the goodness of fit more closely.

Goodness of Fit

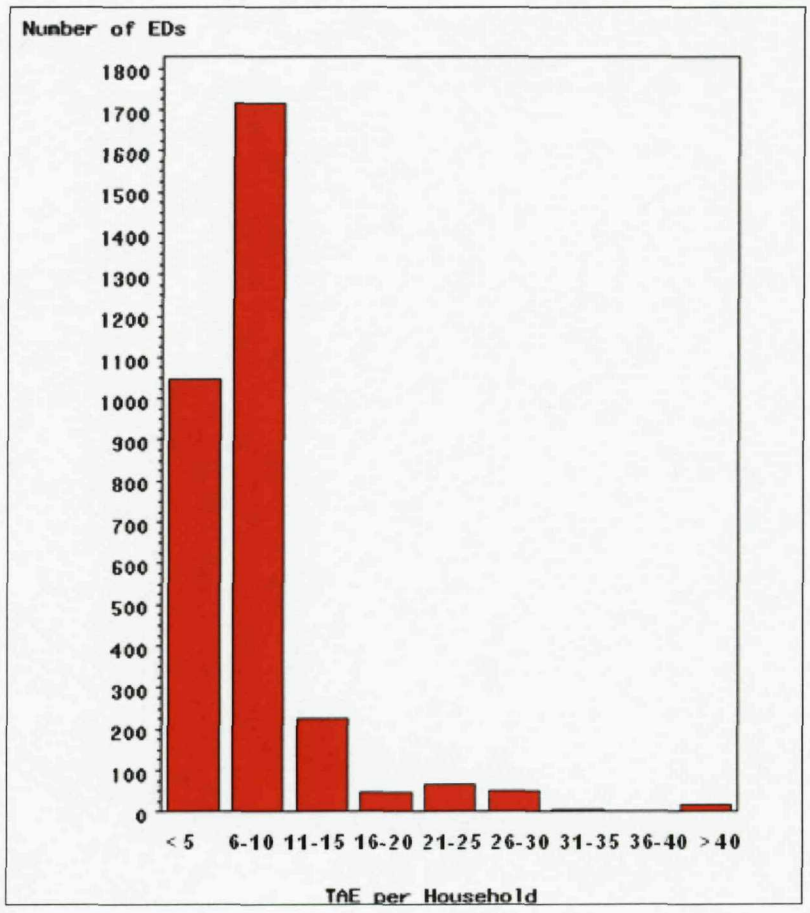
Now that the synthetic population has been created the goodness of fit must be assessed to check it is suitable for the empirical work. The fit does not need to be perfect but reasonable so that the experiments will be valid. Figure 4.7 displays a map showing the spatial distribution of error for Hampshire and its main urban areas, Basingstoke, Portsmouth, Winchester and Southampton. Firstly we want to check that the error is sufficiently small and secondly look for clusters where the error is very big - this may indicate that the area has unusual characteristics and the tables aren't fitted

properly. As shown in the map in figure 4.7, the tables seem to fit well in the west region of Hampshire with TAE scores per cell of less than one. The urban areas seem to show the worst fit however bearing in mind the number of cells over all four tables, the TAE is relatively small.

Figure 4.6 shows a frequency chart relating to the TAE divided by the number of households in each ED. The TAE provides a measure by ED, of the total difference in number of individuals/households over all cells in the constraint tables. Proportional to the total number of households in the ED the TAE seems quite small, being less than ten in the majority of cases.

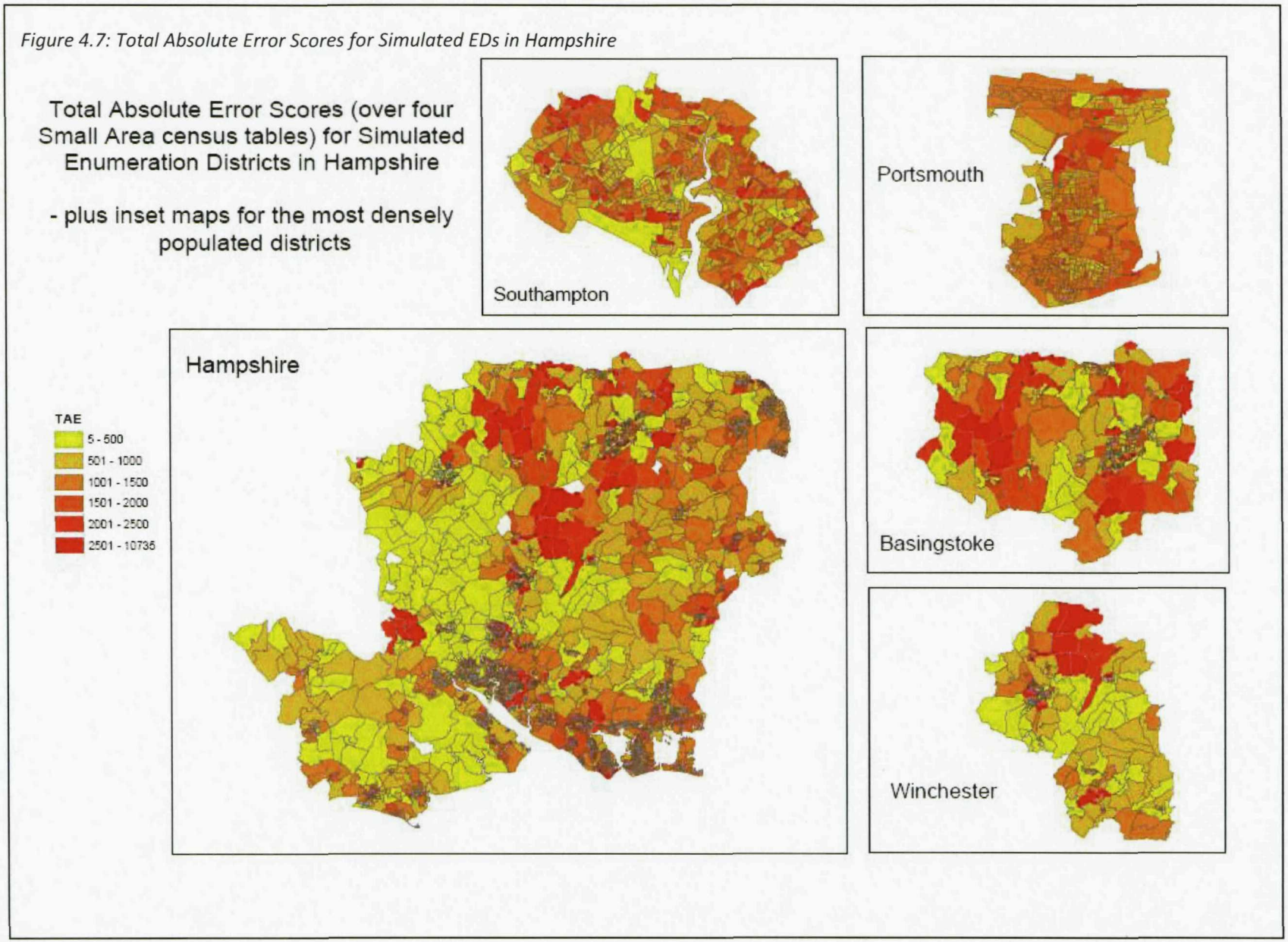
We conclude that the population demonstrates satisfactory fit for our purposes.

Figure 4.6: TAE by ED, divided by the number of households in each ED



Furthermore repetition of households in the synthetic dataset should be considered as it is possible that a household could be sampled from the SAR more than once (we note that the synthetic data consists of 595,174 households and the SAR 541,922 households). Tests showed that 70% of households were not unique in the dataset. 15% of households were represented three times (making 45% altogether) but less than 1% repeated more than four times. This would be a problem if the objective of the empirical work was to assess uniques in microdata since only 30% are unique records. However since the tests for disclosure will focus on tables defined by no more than about four variables, there should still be many small cell counts. Tests on the synthetic population showed that there were many uniques at the small area level with tables of two to four variables. This is because only 8% of duplicated households were located in the same LAD. Further perturbation could be applied to the households (such as adding +/-1 to the age of each duplicate person), however it was thought that 8% of households was too small to have a significant effect. When performing geographical perturbation methods, this result must be kept in mind as we may end up swapping some households which are identical thus biasing the results. In summary, we conclude that the synthetic population is sufficient for our purposes as any bias resulting from duplication will affect all geographical perturbation results equally and the main objective is to compare between methods, rather than examine their individual effect on risk and utility.

Figure 4.7: Total Absolute Error Scores for Simulated EDs in Hampshire



4.6.2 Part 2: Creating Geographical Point Locations for Households

It is essential that the households in the dataset have a spatial point location so that the geographical perturbation methods can be carried out. We use a GIS approach to assign point locations making use of a file from the census website called the PCED file* (Postcode to Enumeration District Directory). The PCED file provides a direct or virtual match between all the postcodes and enumeration districts in England & Wales. For every postcode there are details of which EDs that postcode falls into with grid references and total households at each postcode. Most importantly the file also gives postcode centre locations. The procedure for generating point locations is then as follows: (some simplified diagrams have been added to help explain)

1. The reference numbers (href) are read in, denoting the identification of each household in each of the fitted EDs (from the microsimulation procedure).

ED code	href
ED1	501
ED1	428
ED1	23
ED2	100
ED2	58

2. The number of households in each of the fitted EDs is recorded

ED	Number of households
1	39
2	81
3	298
...	...

3. The PCED file is then merged with the fitted ED codes. There is likely to be a discrepancy between the PCED household count and simulated household count as they are based on slightly different data.

ED	Postcode	Grid Reference	Households at postcode	Totals hhlds by ED (PCED file)	Total hhlds by ED (simulation)
1	RG247BT	5321018180	25	40	39
1	RG297BA	5322018130	15	40	39
2	RG297BA	5320018140	10	81	81
...	...	...	...	...	...

4. Where the postcode falls across more than one ED, the ED code for which the majority of the postcode falls in is kept.

ED	Postcode	Grid Reference	Households at postcode	Totals hhlds by ED (PCED file)	Total hhlds by ED (simulation)
1	RG247BT	5321018180	25	40	39
1	**RG297BA**	5322018130	15	40	39
~~**2**~~	~~**RG297BA**~~	~~5320018140~~	~~10~~	~~81~~	~~81~~
...	...	...	...	...	...

5. A separate file is made for each ED with postcodes, grid references and total households at each grid reference.

ED1		
Postcode	Grid Reference	Total Households from PCED file
RG297BT	5321018180	25
RG297BA	5322018130	15
...	...	...

6. Each record is replicated by the number of households at that postcode.

ED1		
Postcode	Grid Reference	Total Households from PCED file
RG297BT	5321018180	1
RG297BT	5321018180	2
...	...	...
RG297BT	5321018180	25
RG297BA	5321018180	1
RG297BA	5321018180	2
...	...	...
RG297BA	5321018180	15

7. The number of households in each ED from the PCED file is compared with the number of simulated households in each ED based on the small area tables.

ED code	PCED households	Simulated households
1	40	39
2	66	81
...	...	...

8. Where there are more simulated households than on the PCED file (over-generation) then the extra amount of needed is duplicated from existing households at random.

9. Where there are fewer simulated households than on the PCED file (under-generation) then a random subset of households is taken.

10. A *weak sort* is performed on the hhsptype variable (hhsptype: whether detached, semi-, terraced, flat etc) for each ED. A *traditional sort* might be performed by working through a list of items and comparing adjacent elements; if the first element is greater than the second in the list then swap them, and then do this for each pair of adjacent elements going down the list. A weak sort is a 'partial' sort, and has the effect of only partially sorting the list. The aim of this is to give a realistic distribution of household type across the ED since most detached households will be located together, most terraced together, and so on. But this won't be an exact pattern so we use a partial sort to get the appropriate effect.

ED1	
Household ref	Tenure category
23	5
2	5
109	4
43	3
63	4
73	3
12	2
74	4
95	2
25	3
16	1
19	1
82	2

11. The grid references are then sorted (obtained after step 9).

12. The grid references are attached to the household ref numbers (hnum) in the order of the weakly sorted hhsptype variable. (The grid references are already sorted in order of adjacency.) We then end up with a file containing households of different hhsptypes located near to each other within each ED.

13. A small amount of noise is added to the grid references so that each household has a unique grid reference (so that the point locations are plausible, would not be sensible to have more than one household in one place). Noise is generated from a uniform distribution (with sensible bounds) and added independently to the easting and northing[20] parts of the grid reference. The grid references are then compared to ensure that no two are the same. If they are, then more noise is added and the list compared again.

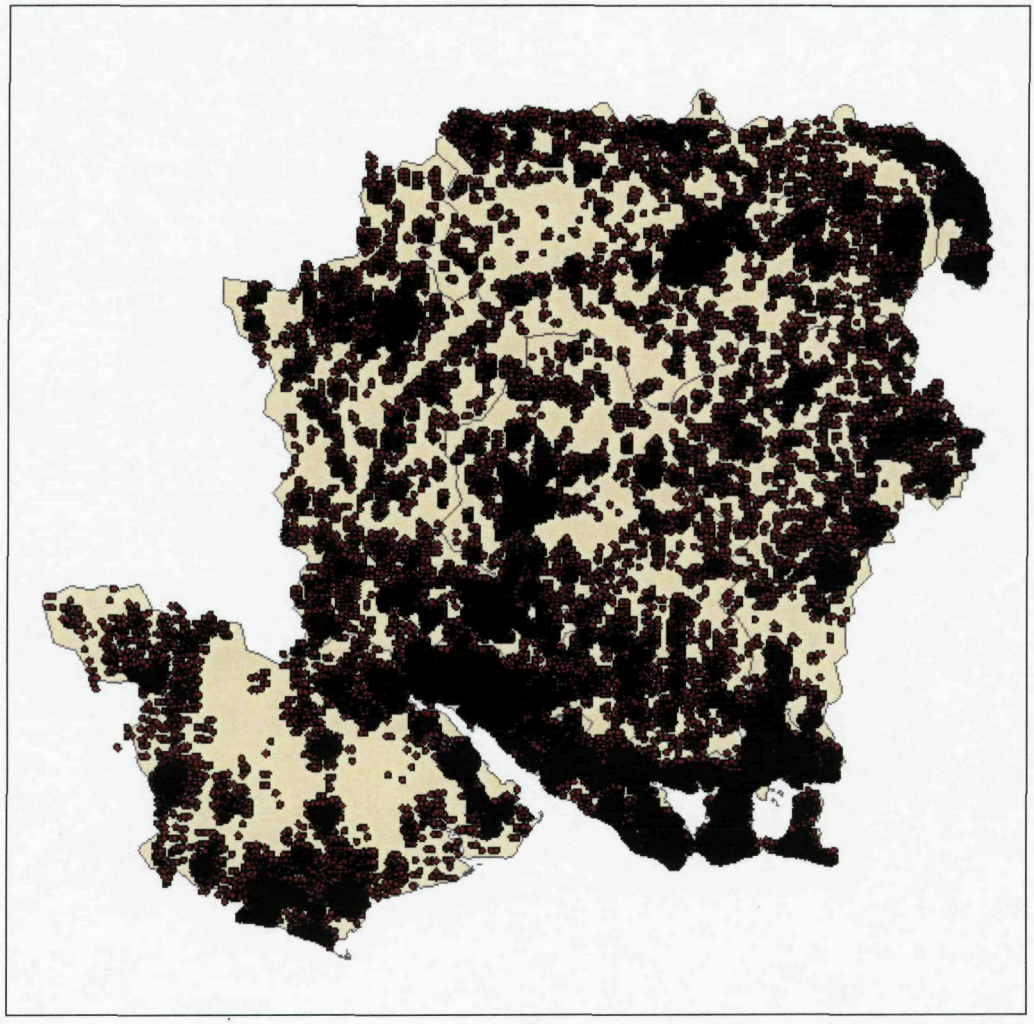
15. When all households have a unique grid reference, the data are imported into ArcGIS along with 1991 ED boundaries. In the GIS software, we are able to view point locations as well as the positioning of the ED boundaries.

16. The spatial join feature in ArcGIS is used to link households to the ED they lie in as some households will fall in a different ED after noise is added to their grid reference. The noise could have been constrained to retain households within their original ED, but this was thought to require more effort than deemed necessary, given that we require only a realistic population.

17. The household population has now been altered so that the number of households in each ED is slightly out in some cases (over-generation) plus a small minority of households have moved ED (noise added to grid references) however this is the simplest way to assign point locations. The population is still realistic and will be treated as the 'truth'. Figure 4.8 shows the map of simulated household locations for Hampshire.

[20] The grid numbers on the east-west or horizontal axis are called Eastings, and the grid numbers on the north-south or vertical axis are called Northings.

Figure 4.8: Map of Simulated Household Locations for Hampshire



After following this detailed procedure, a realistic micropopulation is obtained for the entire county with every household uniquely georeferenced and each with the full set of SAR variables, ready to act as the 'real' population.

4.6.3 Calculating Adjacent Postcodes

Although the population now consists of households with unique grid references, the households do not have a postcode variable attached. This was because in section 4.6.2 the grid references had noise added to make them unique thus the original postcode variable attached to the household became inapplicable. We can simply assign a new postcode variable to each household based on the Euclidean distance between household location and postcode centroids (the postcode locations are defined by Royal Mail as lists of sequential delivery points). In effect the household is assigned to the postcode it has the nearest straight line distance to. However this means that the postcodes

themselves have no defined shape other than the rough structure defined by the locations of households falling in them. Each record z_i must have a postcode attached to carry out swapping at the postcode level. In chapter 5, some of the experiments will be based on swapping households between adjacent postcodes hence the postcodes need to have some shape. In this last section of the chapter we consider how to calculate adjacent postcodes in preparation for the swapping experiments. Voroni diagrams will be used, as described in DeBerg et al. (2000), to define postcode shape. The Voroni diagram can then be transformed into a Delaunay triangulation to determine adjacent postcodes.

Voronoi diagram

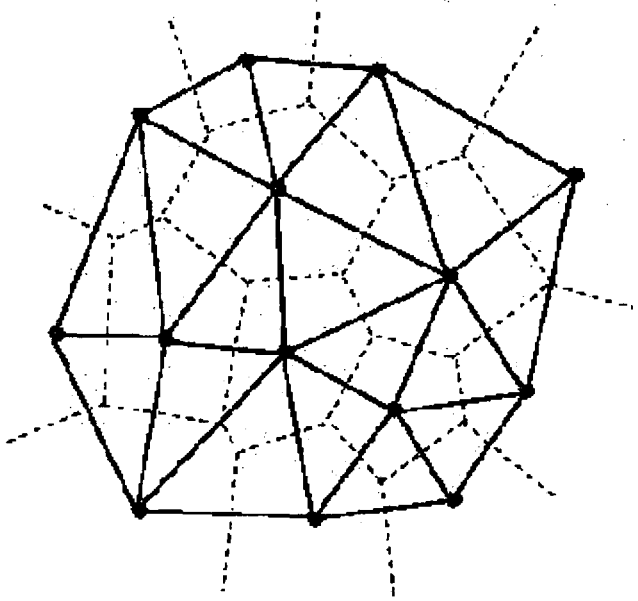
A Voronoi diagram is a geometric structure that represents proximity information about a set of distinct points $P=\{p_1, p_2, ..., p_N\}$ in the plane $\Re^2$. In the case of the synthetic data, these points relate to the postcode centroids.

The Voronoi diagram is the partition of $\Re^2$ into subregions known as Voronoi polygons (or Thiessen polygons) $Vo(p)$ where $p \varepsilon P$. There is one subregion $Vo(p)$ relating to each postcode. Each subregion is defined as the set of points which are closer to p than to any other points in P. In other words, the subregions $Vo(p)$ have the property that a point q lies in the subregion corresponding to a postcode p_k if and only if $dist(q,p_k) < dist(q,p_l)$ for each $p_l \in P$ with $l \neq k$. The distance is the Euclidean distance for the two points p and q. The Voroni polygons $Vo(p)$ can be thought of as defining the shape of the postcodes.

Delaunay Triangulation

The next step is to define the Delaunay Triangulation which will be used to identify adjacent postcodes. If a line is drawn between any two points p_k and p_l who share edges of a Voronoi subregions then a set of triangles is obtained, known as the Delaunay triangulation (shown in figure 4.6). Each intersection of Voronoi edges belongs to at least three Voronoi cells and is the centre of the circle through the generators of these three cells. These three vertices form a Delaunay triangle. A triangulation is a subdivision of an area into triangles (tetrahedrons). Generally, this triangulation is unique. The Delaunay triangulation produces a set of lines connecting each postcode point location to its natural neighbours. Figure 4.9 shows that each edge of a Voronoi subregion is the bisector of the connection of p to the corresponding neighbour cell.

Figure 4.9: Delaunay Triangulation



In R there is a package called *deldir* which can perform the above. The function computes the **Del**aunay triangulation and the **Dir**ichlet tesselation (called Voronoi diagram above) of a planar point set. The program requires the co-ordinates of the point set being triangulated (the postcode grid references). An option of the deldir package is a function called *tilelist* which returns a list describing the Dirichlet/Voroni tile $Vo(p)$ containing each point in the set being tessellated. Part of the R output from the deldir package also includes a matrix called *dirsgs* of which columns 5 and 6 give a list all of the postcodes (tiles) which are adjacent.

Finally we end this chapter having created a synthetic micropopulation for the Hampshire region. This population consists of individuals in households and their values on the full set of SAR variables. The characteristics of individuals in households mimic those of the real 1991 population. Furthermore the frequencies of households within each geographical region are very close to (1991) reality. The population of households are also georeferenced, that is, each household has a unique point location. Furthermore each household record has a postcode, ED and LAD attached describing the geography of the dataset. The postcodes are connected in real space but do not necessarily nest within EDs. EDs nest within wards within LADs.

Chapter 5 Empirical Assessment of Geographical Perturbation Methods

5.1 Introduction

This chapter presents an empirical study experimenting with some of the ideas proposed in chapter 3. The methods will be assessed in terms of risk and utility as defined by the measures formulated in section 3.10. To begin with, the focus will be on new ideas for improving the risk-utility outcome of swapping (as opposed to displacement or other rearrangement methods). The method of RRS carried out on the Census 2001 is used as a benchmark for assessing the new approaches. Various parameters for swapping will be studied including generating sampling distance from a distribution (section 5.5.1), targeting risky records (section 5.5.2), varying sample size (5.5.3) and the effect of distribution type. The objective is to find a new perturbation method that offers greater protection against geographical differencing and disclosure from small area data, for the same levels of utility as the RRS. Thus attention will be focused on the risk at the small area level.

Section 5.6 then goes on to study a displacement approach, where unlike swapping, the households have no restriction on where they are moved to, resulting in the global set of locations being modified from the original. Certain parameters for the displacement methods are set equivalent to their comparable swapping approaches as described in section 5.6.1. The pros and cons of displacement and swapping are also discussed.

The chapter ends with an examination of the best new approach identified via its risk-utility outcome. Further work compares this method against RRS to highlight the reduction in disclosure risk through a simulation of geographical differencing (section 5.7.2) and in simulated small area data (section 5.7.1).

5.2 Data Preparation

In chapter 4, a synthetic population of people in households was constructed. This dataset will be used to perform the experiments. The data have all the characteristics of a real census but for the sub-region of Hampshire (containing 595,174 households), with the variables coded as in the 1991 SAR. For ease of computation of some of the swapping methods, households are gridded by assigning to a 100m raster. 'Gridding' as described in chapter 3, involves assigning each household a cell location on a grid of 100m resolution (the original grid references have 1m resolution) and avoids the need for computationally intensive methods to estimate density for example, such as kernel density estimation. Local density can instead be estimated by a cellular approximation to a circular search which involves counting the number of households in the surrounding grid cells.

Swapping will be performed in the statistical package SAS, and displacement in the geographical package ArcGIS. Displacement involves the use of buffers and does not require the dataset to be gridded. The advantages of displacement in terms of implementation will be considered against swapping.

5.3 Assessing Risk and Utility

Following Duncan et al. (2001), the evaluation of the effectiveness of the geographical perturbation methods will be in a risk-utility framework, i.e. the performance of each of the methods studied in terms of both disclosure risk and the utility of the resulting outputs for analysis. Moreover, since the methods depend upon the specification of parameters, such as the proportion of records to swap, such choices will be examined to see how they affect risk and utility and the trade-off between the two. In order to set up this framework measures of risk and utility are introduced in sections 5.3.1 and 5.3.2.

5.3.1 Indicators of Risk

The aim in the following experiments will be to reduce the disclosure risk primarily by minimising the probability of finding true uniques as defined in equation 5.1 following the discussion in section 3.10. This measure of risk is dependent upon the output zones which may be EDs or wards, for example. To measure the risk arising from geographical differencing (see chapter 2), the disclosure risk for frequency tables for zones assumed to be equivalent in scale to a differenced 'sliver' is considered. The smallest output zones available are either postcodes or EDs and these will be used as an indicator of the risk in slivers followed by a more extensive analysis of geographical differencing in section 5.7. Let N_T denote the number of cells in table T and F_c^o and F_c^p represent a cell in the unprotected and protected tables respectively (as in the previous chapters). Let *match* = 1 if $F_c^o = F_c^p = 1$ and if the same unique household appears in the table before and after perturbation. The probability of finding a true unique in a postcode or ED can then be calculated as:

$$\Pr(TU) = \frac{\sum^{N_T} I\left(F_c^o = F_c^p = 1 \,\&\, match = 1\right)}{\sum^{N_T} I\left(F_c^p = 1\right)} \tag{5.1}$$

where the sums are over all the cells c in the table and I is an indicator function which equals 1 if true, 0 otherwise. A tolerable risk might be set below 50% at the very least, so that the odds are against an intruder making a correct link.

5.3.2 Indicators of Damage

Two indicators of damage will be used in the assessment process, as described in section 3.10.4: the AAD and the RAD:

$$AAD = \frac{\sum^{N_T} \left|F_c^p - F_c^o\right|}{N_T} \tag{5.2}$$

$$RAD = \frac{1}{N_T} \sum^{N_T} \frac{\left|F_c^p - F_c^o\right|}{F_c^o} \quad for \quad F_c^o > 0 \quad else \quad RAD = 0 \tag{5.3}$$

More complex analyses of utility are discussed later in chapter 6 including looking at the impact on multilevel models and geodemographic classifications (an expansion on the methods briefly summarised in chapter 2).

5.4 Outline of Experiments

Some new methods for geographical perturbation were discussed in chapter 3 which will be tested empirically; firstly swapping will be considered, looking at the two new methods distance swapping and density swapping. Both of these are zone-independent involving sampling distances to move from a distribution, but the distance swap will sample Euclidean distances whereas the density swap will sample 'household' distances. The latter approach is a way of taking into account household density; a spatial indicator of risk. Following on from this, other swapping parameters will be varied such as the size of the sample of households swapped, the choice of which records to swap and the effect of 'local' swapping (swapping short distances).

Secondly, a displacement method will be considered. This technique has no spatial restriction on where households are moved to and may offer advantages in terms of operational simplicity with a comparable R-U outcome in regards to swapping.

The aim of these experiments will be to find a disclosure control methodology that, compared to the existing method of RRS, achieves greater reductions in disclosure risk particularly in regards to geographical differencing of small area data, but does not damage the data too severely. The experiments will concentrate on the following geographies; postcodes which are the smallest (in terms of number of households), then OAs, followed by EDs, LSOAs and wards, as shown in table 5.1.

Table 5.1: Average number of households by geography in synthetic Hampshire dataset

Postcode	OA	ED	LSOA	Ward
31	109	220	535	2,619

The risk will be analysed predominantly for postcodes, being the smallest output zones, but also the general effect over different levels of geography compared. For displacement, the smallest geographies available for analysis are EDs because this method involves creating new household locations which do not have postcodes assigned (the postcode list represents a look-up table for the existing population only – see chapter 4).

Ideally, a geographical perturbation method that can be used as a sole pre-tabulation technique would be desirable but bearing in mind that a large number of records would need to be swapped to obtain minimal disclosure risk, this may be unachievable (or unacceptable to a statistical agency). In the England & Wales Census 2001, other post-tabulation methods were also applied to the aggregated tables.

To ensure initial comparability between the methods, they should equate to roughly the same levels of perturbation, i.e. certain parameters should be kept constant. After conducting RRS, the minimum, maximum and mean of the distance between swaps will be calculated and these parameters used to produce equivalent distance and density swaps. For the displacement methods, ensuring comparability is more complex because there are no paired swapping distances, only distances displaced. However the mean distance displaced can still be set to equal the mean swapping distance (for random displacement) and the median distance displaced set to be equivalent to the mean number of households between swaps (density displacement). This is explored further in section 5.6.1. Many results have been produced and reviewed but only the most relevant are reproduced in this thesis.

5.5 Varying the Parameters for Swapping

5.5.1 Using a distribution to generate swapping distances

Firstly, RRS (the benchmark) is compared with distance and density swapping. The latter two cases take their swapping distances from a distribution but the proportion of records swapped and all other parameters remain constant. At this stage, match variables will not be used. 10% (59,518) of households are swapped in each of the three cases (RRS, distance and density swaps)[21]. The same 5% initial sample of records selected by Simple Random Sampling (SRS) are perturbed in each case. Various statistics such as the minimum, maximum and mean will be kept the same across all three methods. Complete descriptions of the methodologies used can be found in chapter 3.

[21] Note that if 10% of records are swapped (a 10% swap) then an initial sample of 5% is first taken from the population and paired with 5% of the remaining records determined by the swapping algorithm.

(i) RRS (without match variables)

RRS attempts to simulate the approach applied in the England & Wales Census 2001. We presume that only a small percentage of records were swapped otherwise the data distortion would be too great. It is also known that swapped records were not moved out of their LAD but were required to be moved between OAs (Boyd and Vickers, 1999). The methodology is implemented as described in chapter 3: section 3.4.2. After performing the swap, the Euclidean distance between each pair of swapped records is calculated, having first converted the grid references into Easting and Northings (X,Y).

Table 5.2 Parameters kept constant from the 10% RRS
(in terms of Euclidean distance)

Minimum distance swapped $\min(d_r) = 5$m
Maximum distance swapped $\max(d_r) = 67{,}915$m
Mean distance swapped $\bar{d}_r = 1{,}354$m

(ii) Distance Swapping (without match variables)

In this case, distances are sampled from a distribution rather than being determined by the shape and population distribution of pre-existing geographies. A paired household is identified for swapping in an interval containing distance d (see section 3.5.3).To ensure consistency between RRS and the distance swap, certain distributional characteristics are kept constant. The mean λ is set to be $\bar{d}_r$ from table 5.2. The exponential distribution will be used so that most households are moved relatively short distances given the mean, and $f(d)$ is truncated below by $\min(d_r)$ and above by $\max(d_r)$ which denote maximum and minimum distances moved between paired households of the random swap.

(iii) Density Swapping (without match variables)

The idea behind the density swap is to take into account the relationship between risk and local population density. Household density will be used as a proxy for population density since the two are highly correlated and households are easier to deal with rather than multiplying by number of residents. To ensure comparability between the random and density swap, parameters from the random swap can be measured in terms of the number of households instead of in Euclidean distance. The number of households passed between pairs of swapped records in the random swap

is denoted n_r. A paired household is identified for swapping in an interval containing the n_rth households (see section 3.5.3).There are two ways of calculating n_r; (i) counting the number of households located in a straight line between the pair or (ii) counting the number of households within a circle defined by the donor household in the centre with radius extending to the recipient household on the perimeter. The second will give a more reliable estimate of local household density as it takes into account all households in the local area whereas in (i) households not on the straight line will be missed.

Table 5.3 Parameters kept constant from the 10% RRS
(in terms of household distance)

Minimum distance swapped	$\min(n_r) = 0$ households
Maximum distance swapped	$\max(n_r) = 218{,}472$ households
Mean distance swapped	$\bar{n}_r = 5{,}038$ households

As with the distance swap, a truncated (exponential) density function is used to generate the number of households passed between swapped households. Table 5.3 shows $\min(n_r) = 0$ because occasionally a household was swapped with its nearest neighbour and there are no other households between them (within the circle).

Initial analysis concentrates on disclosure risk (rather than utility) for the three swaps, since the primary aim is to minimise this with the secondary aim being to limit the damage that occurs as a side-effect. A table defined by ethnic group and limiting long term illness would generally produce many small counts being skewed towards the white without a limiting long term illness category. Tables of ethnicity by LLTI are constructed from the swapped populations for all postcodes (the smallest geography available) and the cells compared to the original (unprotected) tables. The probability of finding true uniques is assessed in table 5.4.

Table 5.4: Disclosure Risk after 10% RRS, Distance and Density Swaps at Postcode Level

Measure of risk (Hampshire study area)	Distance swapping (i)	Density swapping (ii)	Random swapping (iii)
Pr(TU)	0.92	0.91	0.94
Total uniques (in all postcodes)	11237	11242	11335

Although the distance and density swap perform slightly better than the random swap, the overall percentage of true uniques is very high. More than 90% of the time, the intruder would be able to

make a correct link to the record. This is a very high probability and swapping a 10% sample of records appears to offer very little disclosure protection and if this were to be the sample size used for the actual RRS, it explains why post-tabular protection would also have been needed.

5.5.2 Targeted Swapping at the Local Level

A higher reduction in disclosure risk may be achieved by targeting those records likely to be unique. In this section we return to zone-dependent methods for the time being using output zones to control the movement of households (the distinction between the three methods RRS, distance, density is ignored). 10% of records will be swapped using a targeted approach and will be conducted at the local level in order to try and reduce the risk further for small geographies. A baseline non-targeted method will be used for comparison.

Defining the 'Local' Level

Following the theory of chapter 3; section 3.7.4, the mean swapping distance needs to be decided upon to create noise around the greatest number of household locations without damaging the utility of the data too much. Figure 5.1 shows the number of unique households by spatial resolution (for the synthetic population). The spatial resolution is represented in terms of the mean number of households at that level of geography. The number of uniques is determined in tables of LLTI by ethnicity, at each level of geography. This table was chosen as it contains many small cells so it is easy to spot a pattern.

Figure 5.1: Number of Uniques by Spatial Resolution

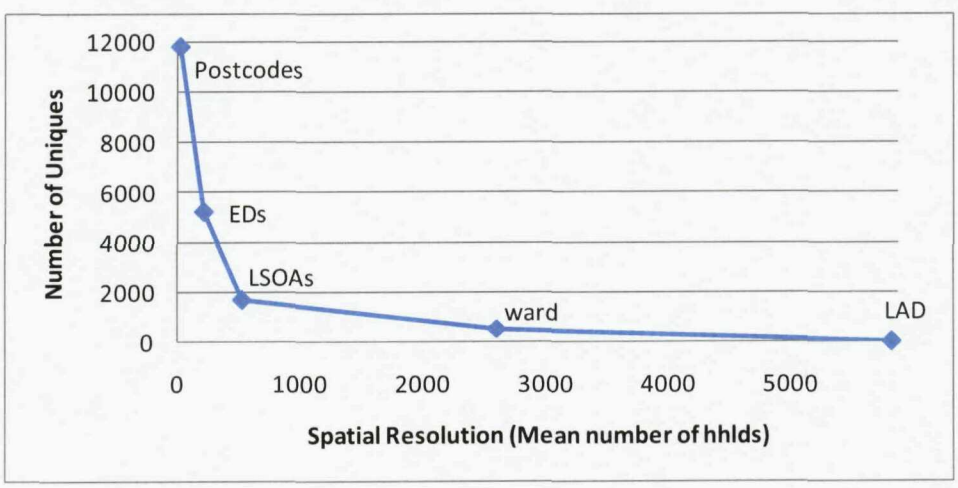


Figure 5.1 highlights the large number of uniques at the very smallest level of geography – postcodes; and more importantly, the number of uniques doesn't decrease linearly with spatial resolution. By targeting protection at postcode level, greater protection should be given to small area data and slivers without damaging the utility too severely at higher levels.

Defining Risky Records

A targeted swapping method was used to protect the US census data in 2000. The general approach can be replicated using the synthetic Hampshire data. The first stage was to identify records that are likely to be unique. Swapping can then take place amongst this subset of households. This subset can be identified by analysing tables at small area level and distinguishing those records which result in small cell counts. This was performed at ED level because at postcode level more than half of the households were unique. Only uniques were defined as risky households because counting larger values such as twos and threes would again produce too many risky households (and only 10% will be swapped). The tables analysed for identifying risky records were defined by variables that are the most visible and traceable to make up keys that an intruder would use to make an identity disclosure. Two keys were chosen as follows:

1st Key) Ethnic group, Age band, Marital Status and Sex. **2nd Key)** Family Type and Household Type.

The 1st key used a banded age variable grouping into ten years since it is more likely that an intruder would have an idea of rough age rather than to the exact year. Both ethnicity and marital status were recoded. Ethnicity became four groups: black, asian, white, other. Marital Status was reduced to either married or single. This was primarily to reduce the overall number of uniques. The 1st key resulted in 158,038 uniques. The 2nd key was made up of Family Type which gives the composition of the household (married couple with two children for example) and Household Type was recoded to give four groups: terraced, semi, detached, flat. The 2nd key resulted in 118,819 uniques. Both keys together resulted in 164,430 risky households out of a possible total of 595,518, since many of the cell uniques were unique on both keys. This is a very high percentage of uniques (around 27%), particularly as they are at ED level (rather than at postcode level). This implies that a high number of records might need to be swapped to reduce disclosure risk satisfactorily as tested in section 5.5.3. The targeted sample of records to swap were drawn from this subset of risky households.

Targeted Local Methods (all zone-dependent, without match variables)

Three targeted approaches were performed as well as a non-targeted approach at the local level to be used as a benchmark. A third type of geographical perturbation is introduced which is a rearrangement approach involving a household A moving to the location of a household B, then that household B is moved to the location of a household C, and so on. It seemed appropriate to test this perturbation method here since swapping between adjacent postcodes allows easy construction of rearrangements between one adjacent postcode and the next. The following four methods are all zone-dependent as the swapping is based around adjacent postcodes. Moreover no match variables are used at this point. In summary, the following methods were tested:

(iv) 5% uniques swapped with any other household in the data but in adjacent postcodes only (to make a 10% swap altogether)

(v) 10% uniques swapped amongst each other in adjacent postcodes only

(vi) 10% uniques rearranged in a series of paths and/or swaps (Rearrangement approach) through adjacent postcodes only

(vii) 10% swap of any household between adjacent postcodes only (non-targeted method used as a benchmark at the local level)

Different samples were used in all four experiments because the methods were performed as a sequential process (select a unique, find another unique in adjacent postcode to swap with, select next unique, and so on). We note that approaches (iv), (v) and (vi) involve risky records for swapping. As before, the probability of obtaining true uniques in tables of LLTI by ethnicity at postcode level is obtained, shown in table 5.5. To confirm the validity of the results, referring back to chapter 3: section 3.10.4, the total percentage of cell value changes $\left(F_c^o \neq F_c^p\right)$ can be at most 20% for a 10% swap. Alternatively this can be rephrased as at least 80% of cell values must be unchanged after a 10% swap. An unchanged cell in terms of uniques can be represented by a true unique (cell value of one unchanged) or a false unique (cell value of one where the swap has involved households with the same characteristics). Thus the sum of Pr(TU) and Pr(FU) must be no less than 80% which we see from table 5.5 is correct.

Table 5.5: Disclosure risk at postcode level, comparing four approaches of local, targeted swapping with RRS

Measure of Risk (Hampshire study area)	5% local unique swap (iv)	10% local unique swap (v)	10% local unique rearrangement (vi)	Local random swap (vii)	RRS (as per table 5.4)
Pr(TU)	0.78	0.74	0.73	0.84	0.94
Pr(FU)	0.22	0.17	0.22	0.14	0
Pr(DU)	0	0.09	0.05	0.02	0.06
Total uniques over postcodes	11832	11822	11818	11832	11335

Local swapping works well, reducing the percentage of true uniques at postcode level by around 10% compared to RRS. Local swapping creates additional uncertainty by increasing the number of disguised uniques (ones which relate to a different household) and false uniques (new cell counts of one). Moreover targeted selection further reduces the risk by up to another 10%. There doesn't seem to be a significant effect of using rearrangements as opposed to swapping (the difference could be due to natural variation since different samples of uniques were swapped). A sensible next step would be to perform local, targeted distance and density swaps since the above methods are not zone-independent (the swaps are based around adjacent postcodes) and thus may not offer much protection against geographical differencing.

5.5.3 100% Local Swapping

It is clear that swapping only 10% of records does not give sufficient protection unless further methods of disclosure control are applied (as with the England & Wales Census). The impact of swapping larger sampling fractions at the local level will next be assessed starting with the extreme of 100%. In the following experiments, 100% local swapping is considered; 100% distance swapping – (viii), 100% density swapping – (ix) and 100% random swapping – (x). Since the swapping will be at the local level, the noise applied to each individual location will be small, however measures of damage will be considered alongside the disclosure risk.

Swapping 100% of records is computer intensive so initial experiments are performed for the Basingstoke LAD only. Basingstoke does have some variation in population density to be suitable for testing the 100% local density swap. North Waltham ward has the lowest population density (10 households per km^2) and Westside has the highest population density (2,051 households per km^2). There is a cluster of high population density wards near the centre of Basingstoke.

Swapping 100% of the records means each record must be swapped at least once. Records are allowed to be swapped more than once to permit alternative swaps only when a suitable match cannot be found amongst the remaining unswapped records. The random, distance and density swaps are performed as a sequential process as described in section chapter 3: section 3.8.1 with slight variation on the methodology according to how the paired households are selected:

100% Random swap: A paired household is found amongst all candidate records in an adjacent postcode

100% Distance swap: A paired household is found amongst all candidate records in the circular band containing distance *d* away from the household to swap.

100% Density Swap: A paired household is found amongst all the households in the circular band containing the nth household (counted in cumulative circular bands).

Since the extreme of a 100% swap is likely to distort the data to a greater extent, this effect is minimised by matching pairs of households to be swapped on a set of key variables (as with the UK Census 2001). Potential recipient households are scored according to how well they resemble the donor household. A *hard match*[22] is implemented meaning that the variable values must be identical to give a score. Four key variables are used for this process: number of people in household, tenure, ethnicity of head of household, and family type. If a potential recipient household takes exactly the same values as the donor household on all four variables, then a score of 4 is given (score of 3 if only three variable values are the same, and so on). The key variables: tenure, ethnicity and number of people in household, were chosen as they are likely to correlate with other census variables so that matches will be of similar types of households in general. These three variables were also used on the matching process in either the UK and US Census swapping methods. Family type was chosen as a fourth variable as it is likely to reflect age, sex and marital status distributions.

An idea is also borrowed from the imputation method used to produce the UK One Number Census*. Imputation was applied when there was no answer on the Census form or the answer was invalid. The principle was to search for a single donor household to supply all the missing variables in a recipient household. Potential donor households were found based on a set of matching variables. In addition, potential donors were penalised if they had been used before. The same idea is implemented for 100% swapping, penalizing households if they have been swapped previously to

[22] An alternative to a hard match would be a soft match which would minimise a distance function based on the number of discrepancies.

discourage them from being swapped a second time. The 100% distance and 100% density swaps could then be carried out as before, with certain parameters kept constant (tables 5.6 and 5.7).

Table 5.6: Parameters kept constant from the 100% Local Random Swap
(in terms of Euclidean distance)

Minimum distance swapped $\min(d_r) = 19$m
Maximum distance swapped $\max(d_r) = 13{,}388$m
Mean distance swapped $\bar{d}_r = 1{,}137$m

Table 5.7: Parameters kept constant from the 100% Local Random Swap
(in terms of household distance)

Minimum distance swapped $\min(n_r) = 0$ households
Maximum distance swapped $\max(n_r) = 17{,}833$ households
Mean distance swapped $\bar{n}_r = 1{,}859$ households

In addition to the three swaps described above, the effects of ordering and distribution type on the local density swap are studied; methods (xi) and (xii). Both of these factors control which households are paired for swapping and may improve the R-U outcome.

- Method (xi) – Density Swapping using a Normal Distribution

The same parameters for a density swap are used as in table 5.7, but with the normal distribution as opposed to the exponential.

- Method (xii) – Sorted Density Swapping using an Exponential Distribution

The impact of sorting the households by density is considered (the rationale discussed in section 3.7.3). This means that donor households in the highest density areas are swapped first and thus have the greatest choice of recipient households, because there are fewer already swapped households (flagged with a one). High density households are defined as those where the household is in a higher density ED relative to the rest of the mean. The reason for this is to avoid high density households moving long distances, due to being at the bottom of the list with no nearby households in close proximity to swap with. Preference is also given to swapping households from similar density areas (if given a choice between similarly matched households). All high density households are swapped first (by sorting the list of households to be swapped by descending density). The low density households will be swapped last with other low density households (since the high density

households will have penalties attached so are less desirable for swapping). This means some low density households may be swapped shorter distances because of the shortage of unflagged households in the specified interval.

Utility and risk are assessed for the 100% swaps at postcode and ward levels. To assess distortion after swapping, as well as considering the AAD per cell, the RAD per cell (equation 5.3) will be calculated. At ward level the AAD will be larger as the noise is aggregated for larger populations but the RAD standardises for cell size. The means over all geographies are examined and split by high and low density areas. The two tables for which damage is assessed are:

Table A) age-band (20 years) by sex and marital status.

Table B) Ethnic group and tenure

Table A consists of independent variables, i.e. the variables were not used on the key for matching swapped pairs. Table B consists of match variables that were used to compose the key and therefore the cells in table B should suffer less distortion. Tables 5.8 to 5.10 present the results.

Table 5.8: Assessing Disclosure Risk after 100% swapping in tables of ethnicity by LLTI at postcode and ward level (non-targeted swaps with matching)

Measure of Risk: Pr(TU) (Basingstoke and Deane LAD study area)	100% local random swap (viii)	100% local distance swap(ix)	100% local density swap (Exponential)(x)		100% local sorted density swap (Normal)(xi)		100% local sorted density swap (Exponential)(xii)	
Postcode level	0.01	0.02	0 (High density areas)	0 (Low density areas)	0.16 (High density areas)	0.05 (Low density areas)	0 (High density areas)	0.01 (Low density areas)
Ward level	0.50	0.33	0.21	0.20	0.28	0.03	0.15	0.14

At postcode level, the risk is minimal as would be expected from a 100% local swap. All density swaps are performing well in comparison to the distance and random swaps. The fact that the percentage of true uniques is reduced at ward levels in all cases implies that many of the postcode uniques are also unique at ward level.

Table 5.9: Assessing Utility after 100% Swapping for tables of ethnic group by tenure (Match Variables) (non-targeted swaps with matching)

Utility: AAD and RAD (Basingstoke and Deane LAD study area)	100% local random swap (viii)	100% local distance swap (ix)	100% local density swap (Exponential) (x)	100% sorted local density swap (Normal) (xi)	100% sorted local density swap (Exponential) (xii)
Postcode Level					
Low density (AAD)	0.3	0.2	0.4	0.17	0.33
High density (AAD)	0.5	0.3	0.8	0.24	0.41
Low density (RAD)	0.73	0.63	0.79	0.66	0.65
High density (RAD)	0.87	0.72	1.18	0.63	0.70
Ward Level					
Low density (AAD)	2.5	3.6	11.3	7.6	4.42
High density (AAD)	4.0	4.4	12.7	7.04	5.51
Low density (RAD)	0.21	0.27	0.40	0.59	0.47
High density (RAD)	0.23	0.39	0.54	0.50	0.46

Table 5.9 shows distortion for tables defined by the match variables is generally small in terms of AAD. The local density swaps result in greater distortion but the unsorted local density swap is particularly bad at ward level. RAD can sometimes be large because of the emphasis on small values where a small change (1 becoming a 2) can represent a high RAD.

Table 5.10: Assessing Utility after 100% Swapping for tables of ageband by marital status by sex (Independent Variables) (non-targeted swaps with matching)

Utility: AAD and RAD (Basingtoke and Deane LAD study area)	100% local random swap (viii)	100% local distance swap (ix)	100% local density swap (Exponential) (x)	100% sorted local density swap (Normal) (xi)	100% sorted local density swap (Exponential) (xii)
Postcode Level					
Low density (AAD)	1.3	1.2	1.5	1.09	1.12
High density (AAD)	2.6	2.1	3.7	1.93	2.93
Low density (RAD)	0.73	0.63	0.79	0.66	0.65
High density (RAD)	0.87	0.72	1.18	0.63	0.70
Ward Level					
Low density (AAD)	15.6	23.5	54.6	37.7	35.1
High density (AAD)	19.3	34.9	55.4	34.9	12.9
Low density (RAD)	0.13	0.18	0.37	0.30	0.20
High density (RAD)	0.14	0.23	0.38	0.19	0.18

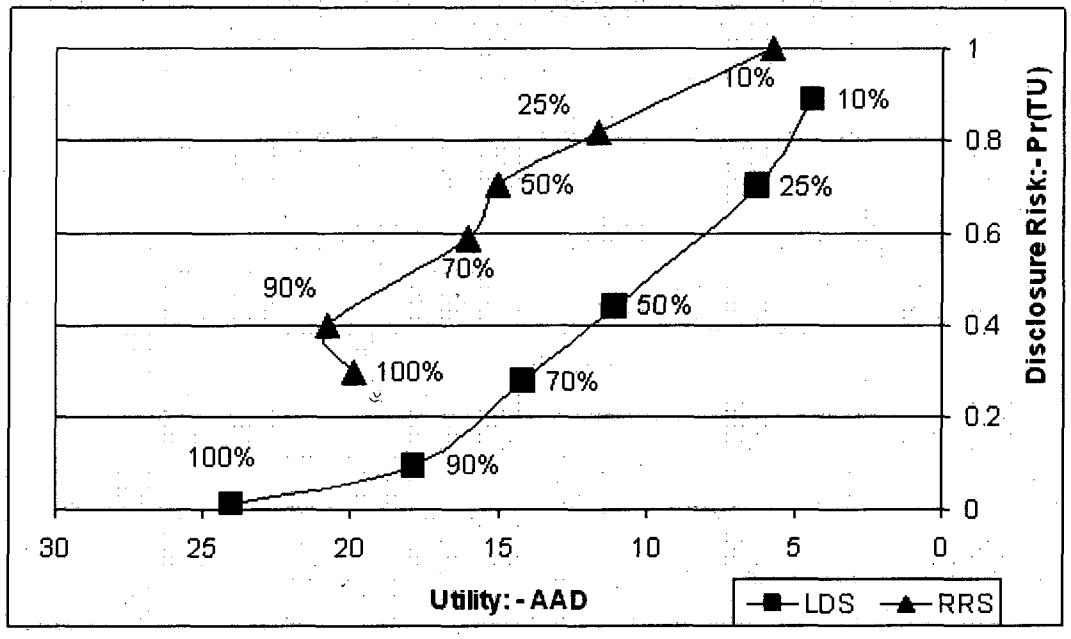
Table 5.10 shows that local sorted density swap is again performing less well in terms of utility in low density areas compared to the local random swap, because the households are moved farther in low density areas. This explains why the risk is relatively lower at ward level in table 5.8 because the households are moving longer distances in some rural cases, between wards. The sorted swap (xii) produces much better results in terms of better utility compared to the other density swaps. The results of the RAD show that there is an average RAD of around 60-90% in the cell count at postcode level, falling to 20-50% at ward level. This seems quite high, particularly for larger cell counts. However this measure needs to be interpreted with caution since it results in very high values for small cell counts and may include 100 differences where 1 has become a 2, etc.

5.5.4 Local Density Swapping (LDS)

In this section, the effect of sampling fraction is analysed on (sorted) LDS using an exponential distribution. This is an attempt to find a balance between risk and utility between the two extremes of 10% and 100% swapping. Later in this section, the effect of varying perturbation distance is also considered on the change in risk-utility, as well as the behaviour of LDS over various levels of geography. Perturbation distance is initially set to $\bar{n}_r$ from table 5.7. LDS will be compared to RRS. RRS will be performed as described for method (i) (and is not a local swap). As before, only a sub-region of the synthetic population is used relating to the Basingstoke and Deane LAD because the simulations are computationally intensive. We note some interesting statistics in Appendix A5.1 regarding 10%, 25% and 50% RRS and LDS swaps looking at how many households actually move ED or ward. It illustrates that although the LDS has a mean set at the 'local' level, many households are actually moved long (Euclidean) distances because of the household density effect.

In this section we first consider the effect of sampling fraction. Initial samples are drawn which when paired with matched households make total swapped samples of 10%, 25%, 50%, 70%, 90%, 100%. Each sample was independent meaning that the 10% RRS initial sample was different to the 10% LDS initial sample, so we may expect there to be some sampling variation. Figure 5.2 shows the R-U outcomes measuring risk at the lowest level available (postcodes) with utility measured at ward level representing a popular scale for analytical use. Both LDS and RRS are non-targeted but with matching.

Figure 5.2: Comparing LDS and RRS:Effect of varying sample size on the Risk-Utility Outcome[23]



The graph shows how LDS improves upon RRS across all sampling fractions. Thus, for a given utility (a vertical line on the graph), LDS always has a lower disclosure risk at the small area level than RRS. Conversely, for a given level of risk (a horizontal line on the graph), LDS always achieves greater utility at ward level than RRS. Suppose a statistical agency wanted to ensure disclosure risk was below 0.5; following figure 5.2, they would need to swap approximately 70% of the records to achieve this through RRS but around 50% of the records would need to be swapped if LDS was used. Moreover if 50% of records were swapped with LDS, higher utility would still be obtained at ward level than if 70% of the records were swapped with RRS. We examine the pattern of variation with risk and utility at other geographical levels in figure 5.3 .

As a side note here, we see that the effect of using different samples does not obscure the general interpretation of the data. Allowing for sampling variation, we might expect the pattern to be approximately linear with a 50% swap producing twice as much damage and reduction in risk as a 25% swap. The pattern in figure 5.2 shows this is roughly the case with some amount of variation, for example the 70% RRS doesn't quite produce as much damage as would be expected, if comparing against the 50% RRS and 90% RRS. Sampling variation has been discussed in the context of RRS in a study by Boyd and Stokes (1999). This concluded that the sampling variation was small (see Appendix A5.2). In conclusion, sampling variation is not considered to have large enough an

[23] Note that the risk for 100% LDS corresponds to the postcode risk in table 5.8. Utility for 100% LDS corresponds to the utility in table 5.10 for wards. 10% RRS and LDS are replicated for the Basingstoke and Deane LAD, not for the whole of Hampshire and thus produce slightly different results to table 5.5.

impact to distort the interpretation of the LDS and RRS so will be overlooked throughout this chapter.

The sampling fraction is one parameter of the LDS method that can be changed. Another is θ, the mean perturbation distance. This distance is measured in terms of number of households and thus doubling the perturbation distance does not mean the households are moved twice as far in Euclidean space. The relationship between the area of the circular band and the radius of the circle containing n households is not linear and this needs to be taken into account when selecting an appropriate perturbation distance. Table 5.11 shows how doubling θ has limited effect on the risk-utility outcome for a 10% swap. We show the AAD here as opposed to the RAD because we want to compare the impact for different values of θ in each column, rather than compare between different levels of geography in each row (where the magnitude of the deviation is relative to the zone size). However results not shown here indicated that within each level of geography, the RAD shows the same patterns as the AAD.

Table 5.11: R-U Outcome over Varying Mean Perturbation Distance for LDS (risk with utility in brackets)

Risk- Pr(TU) and in brackets: Utility - AAD	Mean perturbation distance, θ	Postcode	ED	Ward
	200	0.90 (0.24)	0.97 (0.78)	0.97 (2.46)
	1000	0.89 (0.25)	0.93 (1.03)	0.96 (3.27)
	3000	0.88 (0.25)	0.90 (1.15)	0.91 (3.98)
	5000	0.88 (0.25)	0.90 (1.13)	0.93 (3.78)
	7000	0.88 (0.25)	0.91 (1.16)	0.84 (4.35)

Further experiments showed that a small sample size of 10% would require setting $\theta \geq 10{,}000$ households to reduce disclosure risk by a significant amount (less than 50%) with an average distortion of 5 per cell at ward level. On the other hand, with a sampling fraction of 70%, to reduce disclosure risk below 0.5 at ward level, $\theta = 2{,}000$ households would be appropriate but the distortion per cell would be 15.

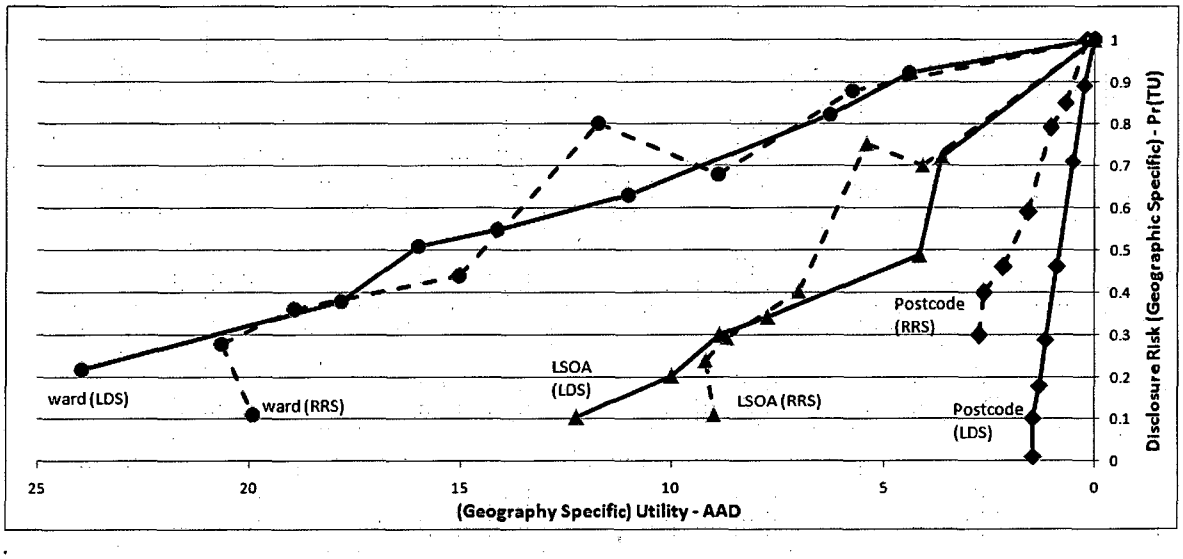
Finally we explore the pattern of risk-utility over different output scales. First table 5.12 shows how the number of true uniques decreases as the size of the geography increases for a 10% sampling fraction.

Table 5.12: Comparing RRS and LDS: Number of True Uniques per 1,000 population at risk (non-targeted swaps with matching)

	Postcode	ED	Ward
10% RRS	7.55	3.13	0.30
10% LDS	6.79	2.92	0.27

Figure 5.3 then shows the general pattern in terms of risk-utility over different output scales. These results include a completely independent geography derived from the 2001 census: Lower Super Output Areas. LSOAs are larger than EDs but smaller than wards. Risk here is measured in terms of the probability of being a true unique for the respective geography (i.e. postcode, LSOA or ward). Utility is the AAD for the respective geography. The figure shows a definite scale effect. As the zone size increases, the utility worsens in terms of AAD, with wards having the greatest average cell deviation and postcodes having the smallest average cell deviation: the larger zones of course have larger populations. However, the most important effect observable in Figure 5.3 concerns the disclosure risk at postcode level. LDS results in better utility (lower AAD) and lower risk than RRS for equivalent sampling fractions (0%, 10%, 25%, 50%, 70%, 90%, 100%) as indicated by the positioning of the lines. However, it is difficult to detect any difference between the methods at the higher levels of geography; partly because of the more unpredictable effect of RRS. Similar patterns were picked up for OAs and EDs (not much difference between the two methods) but are not included in the graph for clarity.

Figure 5.3: Comparing the Risk-Utility Outcome of LDS and RRS over different levels of geography



Note that the R-U results in this graph are for the Basingstoke and Deane LAD only (not for the whole of Hampshire) and thus are slightly different to the table 5.5. However 100% LDS corresponds to the results at postcode and ward level in table 5.8 (utility) and table 5.10 (independent variables).

5.5.5 Conclusions for Swapping (Varying Parameters)

The disclosure control procedures applied to the England & Wales Census 2001 consisted of RRS on the microdata plus small cell adjustment for tabular outputs. Assuming the swapping rate was 10% as in the experiments, this would lead to 80-90% of the small cell counts being true uniques at the small area level. Thus the emphasis in the earlier experiments in this chapter has been to reduce disclosure risk, minimising the data distortion as a side-effect. The targeted and local methods of swapping analysed in section 5.5.2 showed improvements in reducing disclosure risk.

The experiments involving swapping at the 100% level significantly reduced the disclosure risk in terms of percentages of true uniques in postcode level tables; in particular the sorted density swap being the most effective. The probability that a cell count of one can be linked directly to the true household was less than 1%. However it is important to consider the utility of the data. For any of the methods described to be suitable alternatives to the RRS, at the minimum, the same level of utility would need to be achieved. Sorted LDS proved to be a good alternative with better risk-utility outcomes but the AAD could still be up to 35 per cell which is a lot of distortion compared to rounding to base 5 for example. Alternatively the sampling fraction could be set equivalent to the utility obtained with RRS and still the disclosure risk would be lower at the small area level. In addition, targeted, sorted LDS was not studied in much detail but should further reduce the risk according to the results of the experiments. Targeted, sorted LDS is not explored empirically in this thesis as the results will be specific to the set of key variables used to determine which records are risky. Thus the records for swapping were selected at random in these initial experiments. However this may be an area for further work.

5.6 A Displacement Approach

Displacement adds noise to household location but has no restriction on where households can move to, unlike swapping where households are paired (see chapter 3: section 3.5). In this section, two new displacement approaches are implemented: random displacement and density displacement. The density approach will use variable buffers according to the household density whereas for the random approach, buffers will have a constant radius. The randomly sampled households will then be 'displaced' somewhere within their buffers (each sampled point having its

own buffer). Displacement will be implemented on the synthetic population representing the county of Hampshire using ArcGIS[24].

For density displacement, a variable needs to be defined providing the radii of the variable buffers. First the household density was calculated; one way to do this could have been to use kernel density estimation but for computational efficiency a ratio is found using the number of households and the size of the OA the household falls in. After adding an area field to the census Output Area shape file (using the appropriate VBA code), household locations were *joined*. The *normalisation* feature under Properties was then used to create household density. Therefore each household in the same OA has attached the same household density.

After following the displacement procedure in ArcGIS as described in section 3.5.1, the new population files can be created by deleting the ids in the sample from the original data and then adding the new households (with id, displaced grid reference, re-attached geographies) onto the non-sampled data.

5.6.1 Comparing Displacement with Swapping

The radii of the buffers for displacement can be set appropriately for comparison with swapping. Whereas swapping involves distances between paired households, displacement involves distances between an original location and new displaced location. These distances can be set equivalent with some calculation. In other words the radius R needs to be found for each of the buffers (circles). The two displacement approaches we consider will be non-targeted.

Setting equivalent parameters: random swapping and random displacement

Random displacement involves a constant buffer size. An obvious approach for setting equivalent parameters, is to set the radius of the buffer such that the *mean distance displaced* is equal to the *mean distance between the swaps of RRS* (see table 5.2). We can derive this based on the fact that:

$$\text{Area} = \int_0^R 2.\pi.r \; dr \quad = \pi.R^2 \qquad (5.4)$$

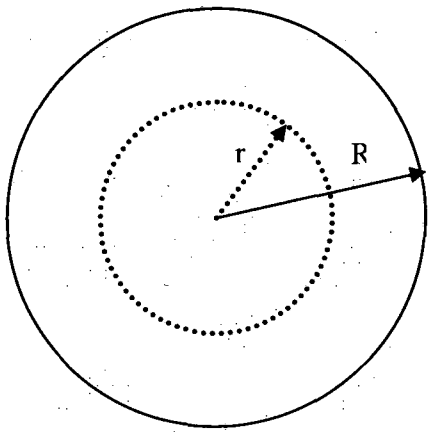
In other words the total number of points in a circle can be represented as each possible radii

[24] ArcInfo workstation was needed for displacement in order to get full functionality from the package.

multiplied by its circumference (the weight).

Consider therefore the buffer as a circle with radius R. The distance displaced can take any value between 0 and R, call this distance r. The number of times r occurs depends on the circumference of the circle defined by the length of the radius r as in figure 5.11.

Figure 5.11: Displacements r in the Buffer with Radius R.



Long displacements have a greater likelihood of occurring. Thus we can weight each possible displacement with its relative frequency (circumference), and then divide by the total number of points (the frequency over all r which equates to the area) to get the mean as in formula 5.5.

$$E(r) = \frac{\int_0^R r.(2.\pi.r)\ dr}{\int_0^R 2.\pi.r\ dr} = \frac{\left[\frac{2.\pi.r^3}{3}\right]_0^R}{\left[\frac{2.\pi.r^2}{2}\right]_0^R} = \frac{2}{3}R \qquad (5.5)$$

Thus the constant R should be set to $\frac{3.\bar{d}_r}{2}$ where $\bar{d}_r$ is the mean from the random swap in table 5.2.

Setting equivalent parameters: density swapping and density displacement

Density displacement involves a variable buffer size. For the density swap, the average number of households between swaps was found (and not Euclidean distance). This was done by counting in a circle, with the initial household at the centre and the paired household on the circumference (in effect nearest neighbour distance)). Counting a specified number of households away in order of increasing distance (points in a circle) is similar to the process involved in finding a median.

Therefore one approach for setting equivalent parameters is to set the *median* displacement distance equivalent to the *average number of households passed* from the density swap by taking into account the household density.

We can derive this based on the fact that the median cuts a probability distribution in half.

$$\int_0^r f(r)\,dr = \int_r^1 f(r)\,dr = 1/2 \qquad (5.6)$$

- The area of the buffer for displacement = *Area* with radius *R*.
- The median over all points (distance to centre) in a circle corresponds to radius *R'* for the circle with area *Area'* = *Area/2* (because by definition the median cuts the probability distribution in half)
- The circle *Area'* with radius *R'* (the median) should cover $\bar{n}_r$ households (average number of households passed in density swap from table 5.3).
- The household density *h* is known (can be calculated from the dataset (in terms of households per m^2)
- Thus *Area'* = $\frac{\bar{n}_r}{h} m^2$
- Then the radius *R* of the buffer can be found using the fact that $Area = \pi R^2 (= 2.Area')$ so:

$$R = \sqrt{\frac{Area}{\pi}} = \sqrt{\frac{2.Area'}{\pi}} = \sqrt{\frac{2.(\bar{n}_r / h)}{\pi}} \text{ = radius of buffer (variable depending on } h) \qquad (5.7)$$

Displacement is quicker to perform than swapping (once set up) because households don't have to be matched and paired, nor does a cellular search have to be performed. However for a population of 595,174 households, it still takes a number of hours to process and generate each buffer. For a sample size of 50%, the memory was not large enough (on a standard desktop PC) and ArcGIS would crash. Thus there are only results for the smaller sample sizes here (10% and 25%).

In table 5.13, disclosure risk is assessed at ED level as this is the smallest geography available to compare with the swapping results. The displaced population cannot be attached to postcodes in ArcGIS as the postcode polygons were created artificially and have no real boundaries: they form only a look-up table / list. The risk-utility results for 10% RRS at ED level are obtained from the swapped population derived earlier in this chapter (section 5.5.1) and are shown in brackets

Table 5.13: Risk-Utility Outcome comparing Displacement Methods to (non-targeted, non-matching) Swapping

ED Level (Hampshire study area)	Risk Measure: Pr(TU)	Utility Measure: AAD
Random displacement 10% (Random swap 10% non-matching)	0.87 (0.93)	0.7513 (1.39)
Random displacement 25%	0.69	1.5821
Density displacement 10% (Density swap 10% non-matching)	0.86 (0.86)	1.3475 (1.29)
Density displacement 25%	0.71	2.8958

Apart from the AAD result marked in red, the results of displacement are very similar to that of swapping as we would expect, since the methods were set with equivalent means/medians to be comparable. The results of the AAD for random displacement in table 5.13 however are surprising. Table 5.14 indicates the percentage of points that are displaced out of the census region (without using the donut buffer).

Table 5.14 Points displaced out of census region

	Random Displacement 10%	Random Displacement 25%	Density Displacement 10%	Density Displacement 25%
	= 24.0%	= 24.1%	= 6.8%	= 6.9%

Table 5.14 may help to explain why random displacement resulted in such high utility (low AAD); if 24% of the points are displaced out of the census region and a new point generated until it lies within the buffer (falling in the census region) then the distribution of displacements is not actually represented by whole circle buffers.

The results from tables 5.13 and 5.14 imply that for density displacement there were a high proportion of small buffers and a few very large buffers (because of the non-uniform household density distribution) compared to the constant buffer size for random displacement. Only 7% are displaced out of the census region because the buffers are much smaller in the densely populated areas (Portsmouth, Southampton, etc) which mostly lie at the boundary of the census region.

A positive outcome from displacement is that the results appear to be generally consistent indicating that the variation in this approach is small for different samples. This is shown by the near doubling of points displaced out of the census region from 10% to 25% (table 5.14) and the near doubling of the risk and utility figures from 10% to 25% (table 5.13) despite independent samples being used. A summary of the pros and cons of displacement compared to swapping is given in table 5.15.

Table 5.15: Advantages and Disadvantages of the Displacement and Swapping Methods

Displacement	Swapping
Advantages o Less computer intensive than swapping o Flexibility in parameter specification o Appears to be more consistent than swapping (for different sample sizes) Disadvantages o Difficult for ArcGIS to handle large sample sizes (runs out of memory) o More difficult to control data distortion o Difficult to define perturbation distribution near boundaries using buffers	Advantages o Match variables can be used to minimise data distortion o Flexibility in parameter specification o Can pair households of similar density to minimise distortion Disadvantages o Computer-intensive to search for paired households o Density swap particularly computer intensive in conducting the cellular search o Less consistency in outcome

The displacement method will not be explored any further because of the problems discussed; large sample sizes cannot be handled in ArcGIS and the risk-utility outcome does not show any advantages over swapping. Given the advantages of displacement as indicated in table 5.15, one

possibility for its use is for hard-to-pair units such as communal establishments and 12-person plus households. These records could be displaced as a simple alternative to swapping.

5.7 Disclosure Risk from Geographical Differencing and Small Area data

In this final section, we concentrate on the method of LDS because this has been shown to have a more favourable R-U outcome compared to RRS. In particular LDS has resulted in reduced disclosure risk at the small area level (postcodes) as shown in section 5.5.4 but provides comparable utility to RRS. This was the initial objective of the thesis as set out in chapter 1. The final two sections of this chapter analyse two small area geographies that might be simulated by an intruder to confirm whether LDS does indeed perform better than RRS.

5.7.1 Disclosure Risk in Small Area data

The LDS method is now explored further to analyse the disclosure risk in small area data that might be generated from published output. An intruder may obtain a published geography and use other available information to isolate small slivers; for example by requesting two sets of outputs for different geographies; e.g. OAs and OAs sub-divided by road features. In this section, we create small areas using the 2001 (England & Wales) Census OAs and split them by roads in Hampshire to create small polygons. It may be possible to identify people in these new geographical zones which we will examine (however in reality an intruder might geographically-difference these from the larger OAs). The data can be generated using ArcGIS.

OA boundaries

First the 2001 Census OA boundaries are read into ArcGIS; these boundaries were downloaded from the UKBORDERS[25] website.

Road data

The Ordnance Survey road file was downloaded from the Digimap[26] website in *ntf* format. Ntf files can be converted into ArcGIS format by saving them within a personal geodatabase using Map

[25] Output Area boundaries are available from http://edina.ac.uk/ukborders/

Manager 9. The roads have pre-determined codes which were imported once in ArcGIS. Lines with codes 3000 to 3004 represent roads (motorways: 3000, A-roads: 3001, B-roads: 3002 and minor roads: 3004) whereas all other lines are railways, coast or district boundaries, etc.

Merging the OAs with Roads

Merging roads with OAs can cause problems in ArcGIS unless the roads run all the way through an OA and do not intersect with other roads within the OA. To avoid these problems, the OA polygons were first converted to a line file to merge with the road (line) file. This was done by copying the OA shape file and pasting it, whilst in edit mode, and setting task to *create new feature*. Following this, a new shape file was created from the merged file representing the small slivers (or split OAs). This can be performed in ArcCatalog, within the personal geodatabase. A *new feature class from lines file* was created (under *properties*) and the new 'split OAs' shape file generated.

The disclosure risk and data utility is assessed for the small slivers by joining the swapped populations to the new geography and comparing with the households in the same slivers for the original population.

Table 5.16: R-U Outcome for LDS and RRS in Split OAs

	Risk - Pr(TU)	Utility - AAD (Age/sex/marital status)
RRS 10%	0.92	0.72
RRS 25%	0.85	0.81
RRS 50%	0.77	0.93
LDS 10%	0.85	0.67
LDS 25%	0.78	0.69
LDS 50%	0.72	0.74

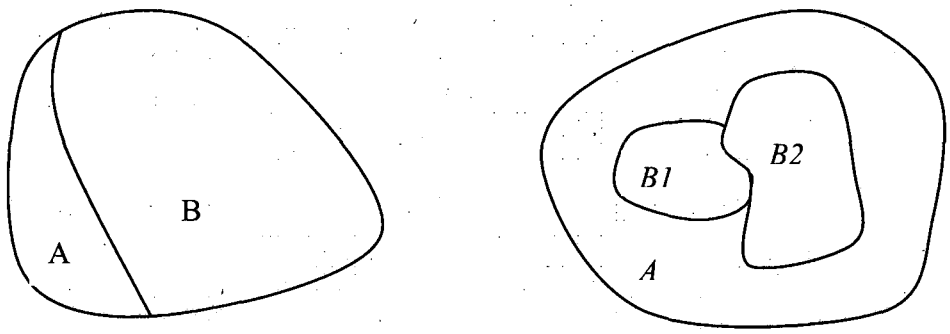
The results highlight how LDS performs much better than RRS for small areas with lower risk and smaller damage. This confirms the previous results from section 5.5.4. However the disclosure risk is still over 70% for both methods which is high (the chances are a unique in the table does represent the actual household). As found before, a very high percentage of records would need to be swapped if the method were to be applied as a sole protection method, or alternatively LDS would need to be applied in conjunction with another SDC method, in order to give sufficient protection against disclosure (lower than 50%).

[26] Road data for Hampshire is available at http://edina.ac.uk/digimap/

5.7.2 Disclosure Risk from Geographical Differencing

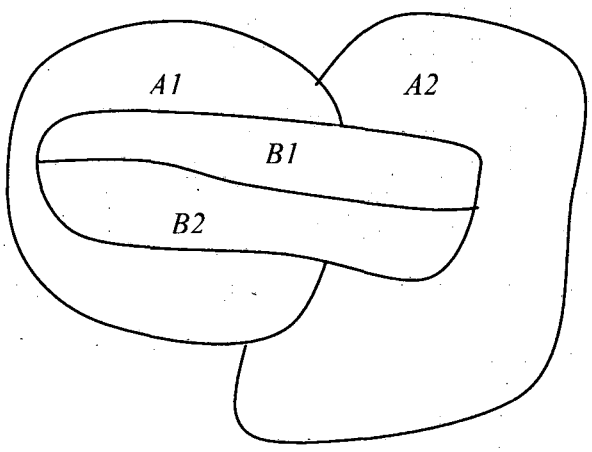
Two independent geographies of similar size are compared using ArcMap. Using the *join* function, zones of one geography which wholly nest within the other are identified. New tables are then created for the geographically differenced areas by subtracting the data for zone(s) B from zone A as in Figure 5.4.

Figure 5.4: Geographically Differencing Zones using 'Join' in ArcMap



A second way to geographically difference zones is via aggregation as depicted in figure 5.5. However this is not possible using the *join* function in ArcMap and requires more complex analysis.

Figure 5.5: Zones which cannot be geographically differenced using 'Join' in ArcMap



For simplicity, geographically differenced slivers created only as in figure 5.4 are studied in this thesis. To analyse the impact of LDS and RRS, the differenced tables are created for the populations corresponding to different sampling sizes such that the fractions 10%, 25% and 50% represent the total number of records swapped. These 'protected differenced tables' are then compared to the 'unswapped differenced tables'. The focus here is to produce enough slivers to allow the effect of

LDS and RRS to be compared rather than to carry out a complete analysis of geographical differencing in Hampshire.

Differencing 2001 OAs from 1991 EDs

Using ArcGIS, 43 OAs were found to nest entirely within EDs; OAs generally being the smaller geography. The variables *econprim* and *famtype* are used to generate tables of employment status (employed, self-employed, student, sick, etc) by family type (married, cohabiting, children, etc). Two sets of these tables were produced, one for the OAs and for the EDs. The differenced tables (ED – OA cells) were analysed to see which disclosure method provided the most protection. Since many of the differenced areas were very large (containing 200-300 households in the synthetic data), the ten smallest differenced tables were studied only (see table 5.17), where the OA nested in a large part of the ED. The appendix section A5.3 provides the full information on the number of households in all of the differenced OAs and EDs.

Table 5.17: Smallest Differenced Areas (OAs from EDs) in terms of number of households

Differenced Area ID	Households in 1991 ED	Households in 2001 OA	Households in Differenced Area
1	195	113	82
2	197	90	107
3	200	77	123
4	199	69	130
5	215	83	132
6	254	117	137
7	227	88	139
8	225	82	143
9	208	59	149
10	247	53	194

Table 5.18 presents the risk in terms of small cells in the differenced areas which are unprotected after disclosure control by LDS or RRS. By unprotected, it is meant that the cell values are unchanged after geographical perturbation. Small cells are defined as those with a count smaller than six.

Table 5.18: Percentage of Disclosive Cells in the Differenced Tables (OAs from EDs), for LDS and RRS

	LDS 10%	LDS 25%	LDS 50%	RRS 10%	RRS 25%	RRS 50%
Disclosive cells (unchanged small cells) in differenced tables	67%	57%	51%	87%	69%	62%

The risk in table 5.18 is a lot higher for RRS. For both methods, it decreases as the sampling fraction increases. Since only ten EDs were examined (there are 3,167 1991 EDs altogether) this small sub-sample may not be representative of differenced tables in the entire population but does confirm previous results. They are probably the only ten that matter, however. There appears to be a pattern in table 5.18 such that the difference in percentage of disclosive cells becomes less pronounced between larger sample sizes. However this pattern does not appear to hold for the previous results in section 5.4 so has not been explored further.

Differencing 1991 EDs from 1981 EDs

11 of the 1981 EDs nested wholly in 1991 EDs and could be geographically differenced. As before, the same tables of employment status by family type were created and the protected differenced tables compared to the unswapped differenced tables. The number of households in the differenced areas was much greater in this case because EDs are generally much larger than OAs and 1981 EDs in particular contained many more households.

Table 5.19: Smallest Differenced Areas (1991 EDs from 1981 EDs) in terms of number of households

Differenced Area ID	Households in 1991 ED	Households in 1981 ED	Households in Differenced Area
1	235 (Basing)	957	722
2	177 (3 EDs aggregated in Basing)	911	734
3	374 (2 EDs aggregated in Farleigh Wallop)	1527	1153
4	86 (Church Crookham)	302	216
5	97 (Hawley)	616	519
6	86 (Waterloo)	1179	1093
7	180 (Forest west)	503	323
8	160 (Forest west)	457	297
9	124 (Marchwood)	573	449
10	150 (St.Johns)	299	149

Table 5.20 presents the risk as before for small cells of the differenced table.

Table 5.20: Percentage of Disclosive Cells in the Differenced Tables (1991 EDs from 1981 EDs), comparing LDS with RRS

	LDS 10%	LDS 25%	LDS 50%	RRS 10%	RRS 25%	RRS 50%
Disclosive cells (unchanged small cell counts) in differenced tables	50%	39%	36%	67%	47%	40%

Because the differenced areas are much larger this time, fewer small cells are unprotected. LDS still offers more protection. However overall the risk is still high (over 36%) for both methods so other forms of protection would need to be considered. The difference between the two methods is less pronounced than before.

5.8 Discussion

Throughout this chapter, the emphasis has been to develop a new methodology that improves on RRS. As a benchmark for comparison, we have considered a 10% sample of records for RRS. We have explored the variation of a number of factors for geographical perturbation: matching versus non-matching, zonal versus non-zonal, different sampling fractions, different perturbation types (swapping, rearrangement and displacement) and examined the effect of density in determining perturbation distance. Whether zone-dependent or zone-independent methods are used depends on the practical needs of the statistical agency. It may be important that marginal distributions are unchanged at a high level geography such as LADs, in which case zone-dependent methods may be more appropriate. In the case of flexible aggregation, zone-independent methods are clearly more relevant to reduce the risk from geographical differencing. In terms of the risk-utility outcomes, distance and random swapping were very similar.

There appeared to be little difference between a rearrangements, displacement or swapping approach although this was not explored in great detail. An important advantage of swapping over displacement is that match variables can be used which significantly improves the utility of the data. For displacement to be a viable method, utility would need to be controlled more effectively in a comparable way.

The level of uncertainty created via geographical perturbation is directly related to the proportion of records perturbed so that swapping only 10% of records meant that the probability of a true unique

was around 90%. This is very high and raised questions as to the validity of using this type of approach as a sole SDC method. Moreover the experiments showed many risky records in the data in terms of uniques; in tables of three basic demographic variables, around 27% of the data were unique at ED level. Targeting risky records produced a significant improvement in lowering disclosure risk, while matching helped to retain the utility of the data. Risk-utility outcomes followed a consistent pattern for the different levels of geography; the larger the geography, the worse the utility in terms of AAD for a given swapping level, but the lower the risk.

This lead to the investigation of local swapping since the smallest areas had the highest risk in terms of proportions of uniques. Local swapping did as expected and reduced risk in the lowest levels of geography compared to RRS. Moreover density swapping helped to reduce risk farther in rural areas by moving households longer distances. Compared to RRS, the utility overall was similar as the worsening of utility in rural areas resulting from LDS was balanced against improved utility in urban areas (where households were moved relatively shorter distances). Choosing the right distribution was also important with the normal distribution not being very effective whereas the exponential distribution produced much better results. Overall, LDS demonstrated significant improvements compared to RRS at the small area level.

Putting the results together for a 10% sample using the new methods, the following table (5.21) can be obtained focusing on risk at the small area level and utility at the higher aggregate level. Note that the starred methods involved match variables whereas the remaining methods did not use match variables.

Table 5.21 Risk-Utility Analysis Comparing RRS to New Methods (10% swaps)

10% swaps	Risk (in terms of True Uniques)		Utility (in terms of AAD)	
	Postcode	ED	ED	Ward
RRS* (benchmark)	0.99	0.93	1.39	5.75
RRS	0.94	0.83	1.11	6.66
Distance swap	0.92	0.87	1.25	5.01
Density swap	0.91	0.86	1.29	5.64
Sorted LDS*	0.89	0.92	0.97	4.40
Random Displacement		0.87	0.75	3.88
Density Displacement		0.86	1.35	6.27

Apart from random displacement which produced unexpected results in terms of very good utility due to the reasons discussed earlier, the risk-utility outcomes for the methods are all very similar for

a 10% swap. In particular, the density and distance swap, as well as density displacement, all produce very similar results compared to RRS (non-matching methods). This result is interesting, particularly in regard to displacement; it means that the other zone-independent approaches, which may have advantages in being easier to implement for example, could be applied.

Moreover, sorted LDS compares favourably to the benchmark RRS (both using match variables). Both the risk and utility are consistently lower at the smaller levels of geography. This difference has been evident throughout the chapter. A simulation of geographical differencing and risk from small area data also highlighted this benefit.

Existing work by Duke-Williams and Rees (1998a) found that the occurrence of small populations resulting from geographical differencing was uncommon; differencing 1991 EDs from postal sectors produced no slivers (or haloes) with less than 48 households or 200 persons. Moreover differencing regular grid cells of varying size (1km, 5km and 10km) showed very few differenced areas falling below the threshold. However as the examples have just illustrated; splitting OAs by roads or grid cells, differencing between geographies of different time periods and of different levels in the hierarchy; there are many ways in which small areas can be created if the data were made available. If a flexible tabulation system were to be implemented, many different geographies could be compared. Moreover if areas smaller than OAs were to be published, the potential for disclosure would further increase. In these instances, a method such as LDS clearly shows benefits in offering additional protection over the traditional RRS as demonstrated in the results.

Throughout this chapter, there has been a strong emphasis on risk-utility with utility measured in terms of AAD or RAD. The improvement in utility through LDS as compared to RRS, means that larger sampling fractions could be considered; greater than 10%. In chapter 6, the LDS method will be explored further comparing with RRS in terms of more complex utility measures that census users employ, and for different sampling fractions. Since the operation of LDS is different to RRS, it may impact on the data in different ways that cannot be accounted for, and that are likely to be averaged out by the AAD measure. These impacts will be spatial in nature because it is the geography variables which are modified rather than interactions between the attribute variables (this is one advantage of geographical perturbation over other SDC methods). Spatial analyses will therefore be the focus of the next chapter; studying the changes before and after perturbation.

Chapter 6 Impact of Geographical Perturbation on Complex Analysis Methods

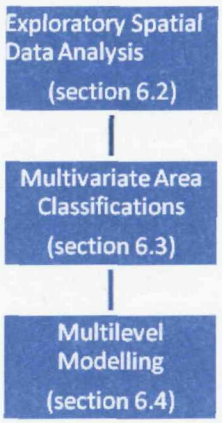
6.1 Introduction

In chapter 5, various geographical perturbation methods were examined and assessed empirically with a focus on disclosure risk. The final chapter of this thesis concentrates on the utility of the data after geographical perturbation. In particular, LDS is examined in more detail comparing against the benchmark RRS. LDS showed a noticeable effect, compared to the benchmark, at the small area level producing lower levels of disclosure risk for the same sampling fraction. Utility was assessed in terms of AAD but this metric may mask underlying effects not picked up by averaging. Studies have shown (Shlomo and Young, 2006a and 2006b) that a range of utility measures is often necessary to differentiate between the effects of different SDC methods. Since there are many approaches to measuring utility, the best strategy would be to consider the analyses census users carry out (as detailed in chapter 2) with regard to those which will be most affected by geographical perturbation. In this chapter we revisit some of the methods in chapter 2 in much more detail.

Swapping doesn't change the relative locations of households but distorts the relationship between the attribute variables and the geography variables. Record *i* with values $(M, A_i, (X_i, Y_i))$, where M are the match variables, when swapped with record *j* with values $(M, A_j, (X_j, Y_j))$, becomes $(M, A_i, (X_j, Y_j))$. Thus the relationship between the match and attribute variables is unchanged,

similarly the relationship between the geography and match variables is unchanged. For example, the relationship between tenure and occupation would be expected to be the same after swapping. In fact this is one advantage of geographical perturbation methods over other SDC methods. Rounding for example, distorts the interrelationships between the attribute variables, often artificially increasing correlations (Shlomo, 2005a). In this chapter, utility after geographical perturbation is considered in a spatial context. Figure 6.1 indicates some statistical methods a census user might perform specifically in a spatial context.

Figure 6.1: Statistical Methods for Analysing Census Data



The three sections of this chapter are devoted to analysis of the impact of geographical perturbation on each of these methods. They have been chosen to represent the diverse kinds of analyses that census users typically carry out. In addition, these methods have been chosen because of the extensive literature that describes application of these approaches to census data. Exploratory spatial data analysis (ESDA) encompasses methods used to detect spatial patterns in the data, to formulate hypotheses and to identify local and global trend patterns. In chapter 2 for example, studies on homogeneity within an area of interest were carried out by Morphet (1993), Tranmer and Steel (1998) and Martin (1998). Another example of ESDA is that of Hirschfield and Bowers (1991) where maps were produced to show the spatial distribution of those with a certain geodemographic classification. Different approaches to ESDA are summarised in section 6.2 and the impact on change in spatial rank and spatial autocorrelation studied. These kinds of analyses can typically be performed using ArcGIS.

Following this, section 6.3 uses area (or geodemographic) classifications to describe and summarise the data. Examples from chapter 2 where census data have been used in this context include the classifications of Acorn from CACI and Mosaic from Experian. The fit of the classifications to the RRS and LDS perturbed data will be considered. Secondly the difference in outcome in terms of which

output zones belong to which classification type will be examined after creating the classification from the swapped populations from scratch, as opposed to using the unperturbed data.

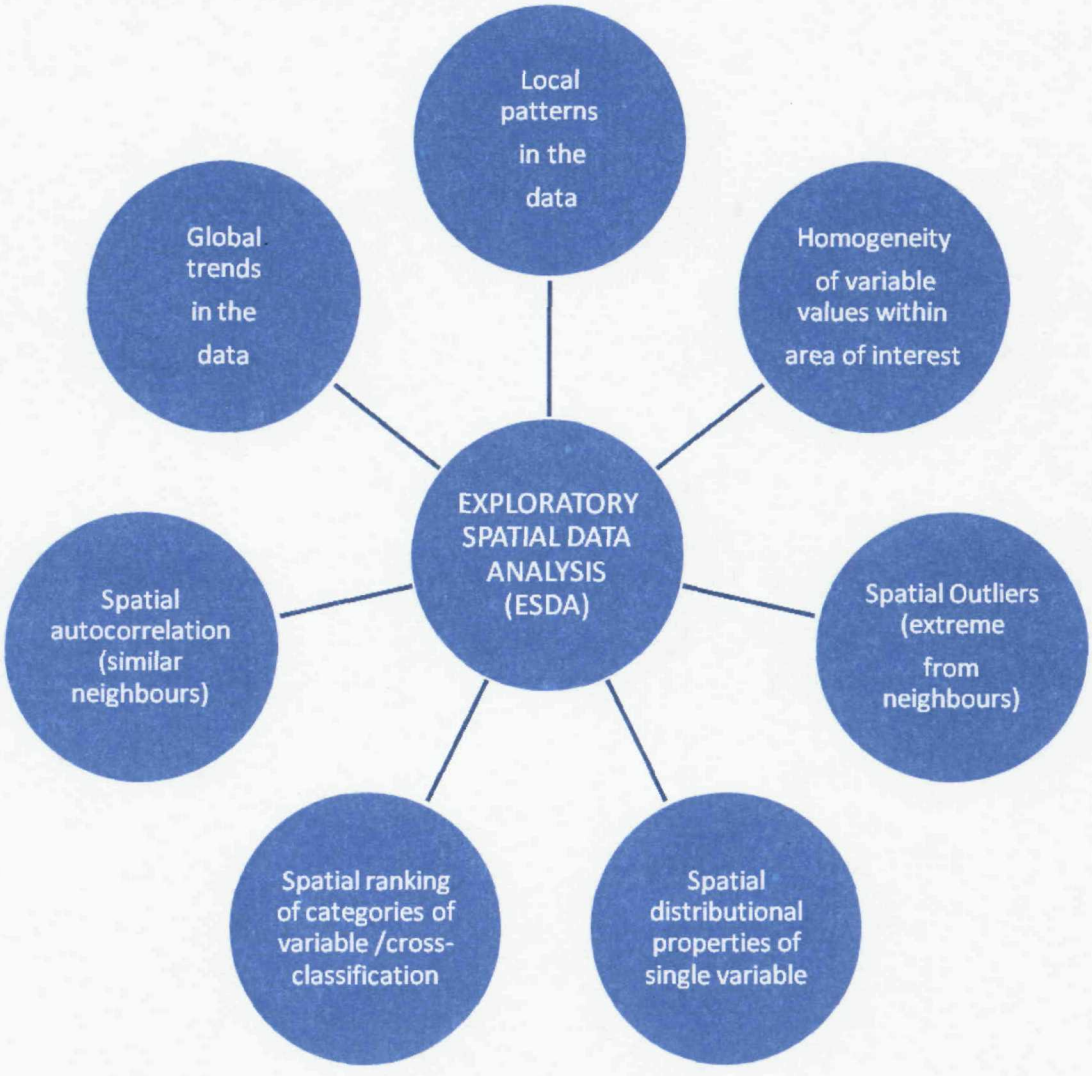
Finally section 6.4 looks at multilevel modelling of the data where EDs represent the nested geographies within wards and the area classification will be used to form a level 2 predictor. The aim will be to observe any changes in the parameter estimates and variation of the swapped models as compared to the unperturbed model. In chapter 2, two examples were given relating to census data; Barnett et al. (2002), studying the impact of deprivation in explaining the spatial variation in health outcomes, and Reijneveld (1998), looking at the adverse effects of area deprivation over and above the effect due to individual socio-economic status.

Ideally the distortion in the data should be kept to a minimum at all levels of output, but it is primarily the higher levels of geography e.g. large output zones such as wards that are most important in terms of accuracy for census users when making important policy decisions and drawing significant conclusions. In this chapter utility will be given less attention at the lower levels of geography. To continue the theme of the previous chapters, this work will be based on 1991 census data and definitions.

6.2 Impact on ESDA

Figure 6.2 identifies some properties of the data that may be found using ESDA. ESDA might involve identifying both local and global trends and patterns in the data. The techniques can also be used to determine spatial outliers and spatial distributional properties.

Figure 6.2: Properties of the Data Identified using ESDA



The impact of geographical perturbation on two distinct ESDA techniques will be explored; the spatial ranking of attributes in the data and secondly the impact on spatial autocorrelation. The second test will be studied from the perspective of identifying both local and global patterns in the data.

6.2.1 Impact on Spatial Rankings

A spatial ranking is an ESDA method used to define relative positions of a geography (e.g. wards) with respect to a particular attribute. This test considers the effect of swapping on the changes in overall spatial pattern, such as would alter the shading classes on a choropleth map; that is, the changes to rank order rather than changes in scale. The former is most likely to be distorted by swapping. The test is performed as follows:

(i) Each zone is measured according to the attribute under investigation; e.g. percentage unemployment

(ii) The zones are sorted by this attribute

(iii) The zones are grouped (to allow for small movements between ranks)

(iv) Each zone is ranked by group

(v) The rank group for each zone is compared before and after swapping to look for large movements between ranks

The procedure will be carried out on wards and LSOAs for two different attributes; (1) percentage unemployment and (2) the percentage of male head of households, aged 35-50, in a professional job with a first degree or higher[27]. These attributes were chosen because, as with any large mixed urban/rural area, they are likely to vary over space. (2) is a cross-classification of the variables and will show the extent to which interactions of the variables are distorted by geography. The wards are split into groups of five and LSOAs into deciles to allow for small movements between rankings. The results showed that changes in rankings were symmetrical about zero and approximately normally distributed. Thus the results are presented as absolute percentage change (in rank group) showing the median of the distribution and the maximum. Swaps of 25% and 80% for both LDS and RRS are compared in order to show the effect for contrasting sampling fractions.

[27] First degree or higher equates to the (1991) census variable qualevel taking values of 2 or 3, and socclass = 1 equates to professional job

Table 6.1: Changes in Rank Group for LDS compared to RRS

		RRS25	LDS25	RRS80	LDS80
Single attribute Wards	Median	1	0	1	0
	Maximum	3	2	4	3
	Proportion no change	13/34	22/34	10/34	22/34
Single attribute LSOAs	Median	2	1	2	2
	Maximum	7	9	9	20
	Proportion no change	29/103	36/103	11/103	17/103
Cross-classified attributes Wards	Median	0	0	0.5	0
	Maximum	1	1	4	2
	Proportion no change	15/34	21/34	6/34	11/34
Cross-classified attributes LSOAs	Median	0	0	0	0
	Maximum	1	1	1	1
	Proportion no change	91/104	97/104	89/104	93/104

Table 6.1 indicates that for the cross-classification, there is little change in spatial ranking with the median around zero for all cases. Many wards or LSOAs do not change rank group. However the single attribute shows lots more change, particularly at LSOA level with LSOAs moving one or two rank groups away. However this may partly be due to the definition of the size of the rank groups, since there were 104 rank groups for the LSOAs altogether and therefore change may be more easily picked up than in wards. A clear difference can be seen between RRS and LDS. The proportion of zones showing no change is higher with LDS than RRS, and at ward level the proportion of zones showing no change with LDS is almost twice that of RRS.

6.2.2 Impact on Spatial Autocorrelation

Another technique used for ESDA is to study the effect on spatial autocorrelation. Spatial dependency is the extent to which the value of an attribute in one location depends on the values of the attribute in nearby locations (Fotheringham et al, 2002, 1998). Spatial autocorrelation measures this dependency by examining the correlation between an area and its surrounding neighbours. If there is any systematic pattern in the spatial distribution of a variable, then it is said to be spatially autocorrelated. A random pattern would exhibit no spatial autocorrelation. Swapping is likely to distort any patterns of spatial autocorrelation, particularly when swapping over large distances, which is likely to make the data become more homogeneous and pockets of households exhibiting unusual characteristics would tend to become more like the region as a whole. If it is known that a variable (or set of variables) exhibit spatial dependency, this relationship can be exploited to assess the effect of the two swapping methods.

Typically a single measure of spatial autocorrelation is calculated which describes an overall degree of spatial dependency across the whole dataset (assessing the global pattern). However this can often mask the true pattern. When spatial data are distributed such that high values are located near to other high values and low values near to other low values, then the data are said to exhibit positive spatial autocorrelation. On the other hand, data with high values close to data with low values (or vice versa) exhibit negative spatial autocorrelation. If both negative and positive spatial autocorrelation are present, this is not picked up by a global measure. In this case, local measures of spatial dependency must also be examined. Local measures of spatial autocorrelation allow spatial variations in the spatial arrangement of data to be examined. In this section spatial autocorrelation will be assessed for the two attributes (1) percentage unemployed and (2) male head of households aged 35-50 in a professional job with a first degree (the same attributes used for the test on spatial rankings). The results for the swapped populations will be compared against the unperturbed data.

The global measure of spatial autocorrelation to be used is the Moran's *I*. This assesses spatial dependency in a particular attribute for the whole region: (Moran, 1950)

$$I = \frac{m\sum_u \sum_v w_{u,v}(Z_u - \bar{Z})(Z_v - \bar{Z})}{\left(\sum_u \sum_v w_{u,v}\right)\sum_u (Z_u - \bar{Z})^2} \qquad (6.1)$$

where m is the number of zones,

Z_u is the percentage in a particular category of a variable or a cross-classification of variables A, for zone u,

$\bar{Z}$ is the mean of the percentages across all zones,

w_{uv} is an element of a contiguity matrix, taking the value 1 if zone u is a neighbour of zone v and 0 otherwise.

The Moran's *I* is similar to Pearson's correlation coefficient where the numerator is a covariance and denominator is the sample variance. Also it can take values between -1 and 1 and the strength of the correlation is reflected in higher values of *I*. The weights reflect geographic proximity and define the local neighbourhood. Values of Moran's *I* larger than 0 indicate positive spatial autocorrelation; values smaller than 0 indicate negative spatial autocorrelation.

Spatial autocorrelation at a local level will be measured using the LISA statistic (Local Indicators of Spatial Association – see Anselin, 1995). This will indicate clusters of significant spatial autocorrelation and is computed as:

$$I_u = \frac{\sum_u \sum_v w_{u,v}(z_u - \bar{z})(z_v - \bar{z})}{\sum_u (z_u - \bar{z})^2} \qquad (6.2)$$

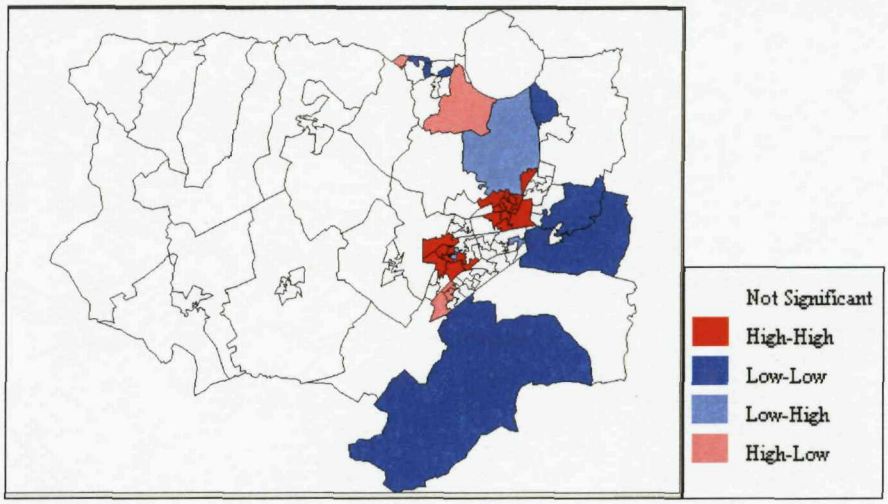
Spatial maps can be produced showing the value of I_u for each zone. In the LISA maps which follow high-high and low-low relate to incidences of positive spatial autocorrelation whereas high-low and low-high relate to incidences of negative spatial autocorrelation. The Moran's *I* and the LISA maps were computed in GeoDa[28] and relate to the Basingstoke and Deane local authority. This local authority was chosen as it was small enough to allow the relationships to be studied in detail which would have been difficult with the whole of Hampshire.

Spatial Autocorrelation for (1) Percentage Unemployed

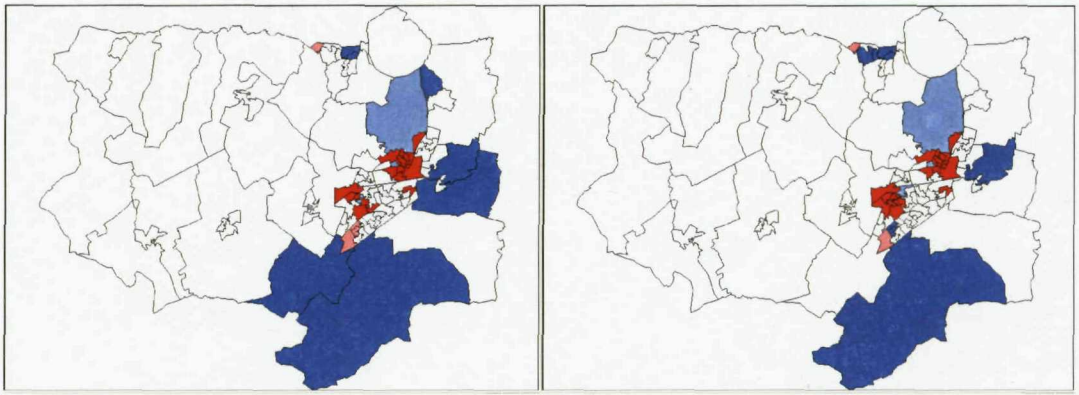
The LISA maps in figure 6.3 were computed for LSOAs in Basingstoke and Deane for the original data and for the swapped populations (for samples of 25% and 80%). The weights matrix needs to be selected in GeoDa and the nearest neighbours was used. In the following maps, the eight nearest neighbours were selected. This means that there is some randomness in which 'neighbours' are selected (if there are more than eight nearest neighbours). For this reason, an identical weights matrix was used in every case. LSOAs were chosen as the size of geography to analyse, containing enough households to show accurate and interesting patterns of spatial autocorrelation. There are 104 LSOAs altogether but only 34 wards which are too few to discern any spatial patterns with nearest neighbours.

[28] GeoDa is used to implement techniques for ESDA on lattice data (points and polygons) and was downloaded from https://www.geoda.uiuc.edu/

Figure 6.3: LISA Maps showing Spatial Autocorrelation in LSOAs for Percentage Unemployed.

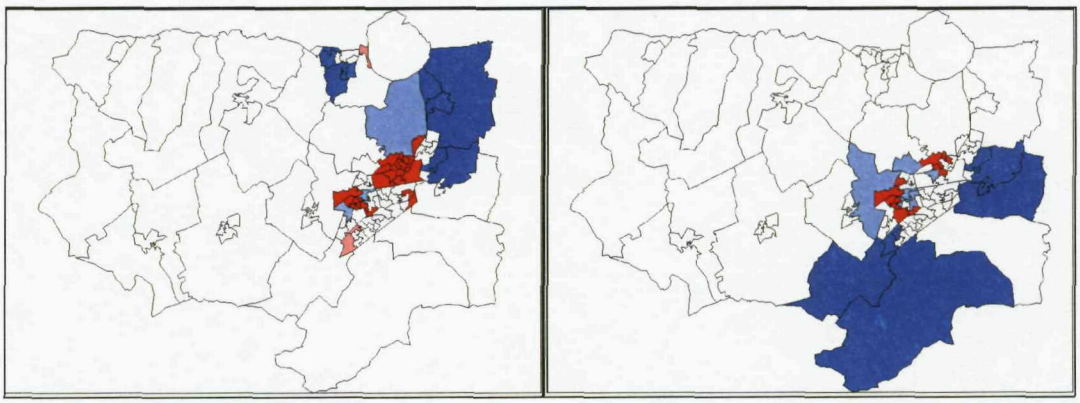


Unperturbed (Original) Data



After 25% LDS

After 25% RRS



After 80% LDS

After 80% RRS

The LSOAs (in the unperturbed data) show evidence of negative spatial autocorrelation north of the centre of Basingstoke but positive spatial autocorrelation in central Basingstoke and southern surrounding LSOAs. The darker shades show areas of positive spatial autocorrelation whereas the

lighter shades show areas of negative spatial autocorrelation. This pattern seems to be mostly retained even after a 25% sample swap for both LDS and RRS methods. However at the 80% sample level, the original pattern is starting to disappear. It is interesting to note that the plots display fewer significant areas of spatial autocorrelation as the sampling fraction increases suggesting that the data are possibly becoming more homogeneous as expected.

Table 6.2: Moran's I at LSOA level for Single Attribute

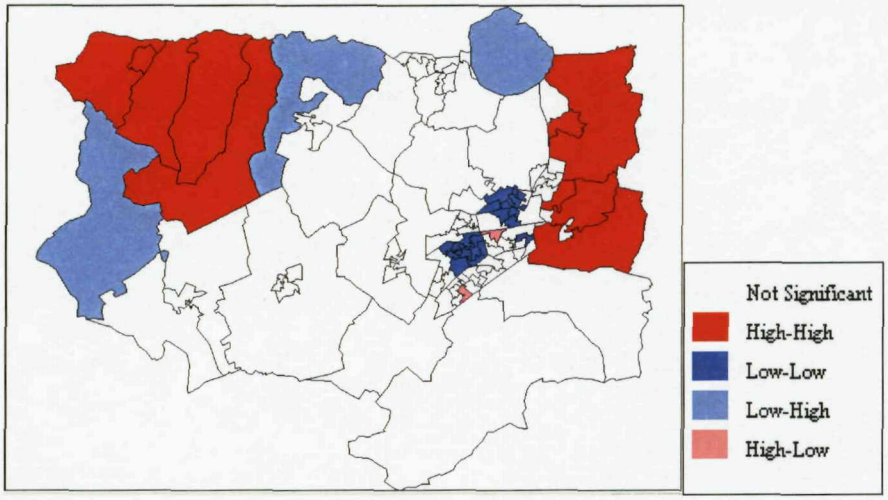
Original data	0.3491	
	25% sample	80% sample
RRS	0.2984	0.2341
LDS	0.3093	0.3338

The Moran's *I* gives a global indicator of spatial autocorrelation relating to the data as a whole (formula 6.1). Table 6.2 shows the global Moran's *I* for the 25% and 80% samples. It is perhaps misleading in this case because both positive and negative correlation are present. However the Moran's *I* is always around 0.3 for the swapped populations possibly indicating the data aren't too different from the unperturbed after LDS and RRS.

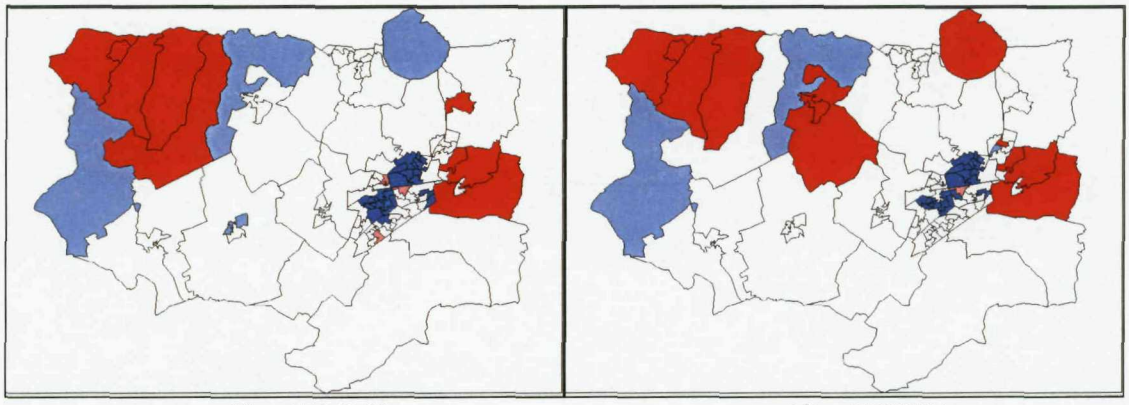
Spatial autocorrelation for a Cross-classification of Variables

The spatial autocorrelation can also be assessed in a cross-classification of the variables to distinguish any different effects of swapping on the interrelationships between the variables. The cross-classified attribute was the percentage of male head of households, aged 35-50, with a first degree or higher in a professional job, as shown in figure 6.4.

Figure 6.4: LISA Maps showing Spatial Autocorrelation in LSOAs for Cross-Classified Attribute.

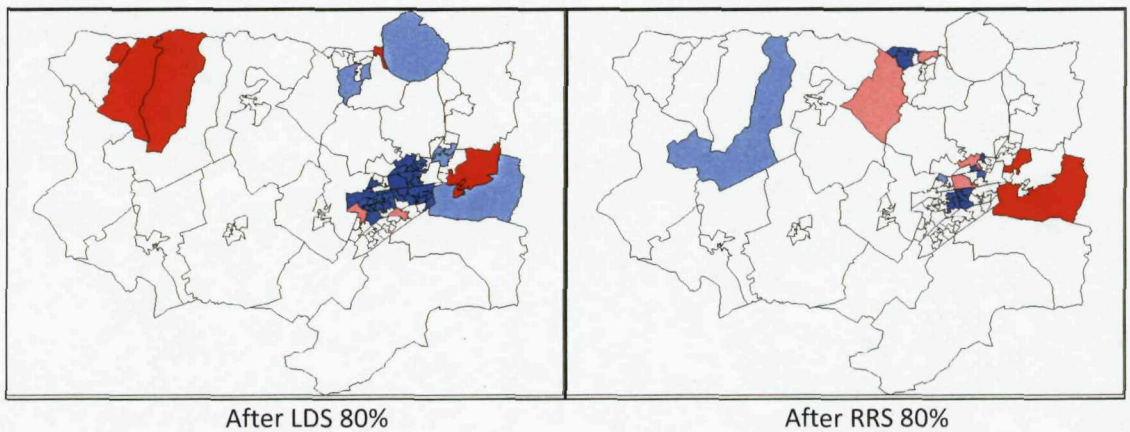


Unperturbed (Original) Data



After LDS 25%

After RRS 25%



After LDS 80%

After RRS 80%

As might be expected, the maps show evidence of a reverse pattern to proportions of unemployed meaning that areas of high unemployment typically have low proportions of male head of households in a professional job with a degree (central and surrounding areas of Basingstoke). As before these maps show that most of the pattern is retained at the 25% swap rate, perhaps more so

for LDS. On the other hand at the 80% swap rate, the large number of swaps mean that although some parts of the pattern is retained with LDS, with RRS many of the significant LSOAs are showing incorrect directions of spatial autocorrelation.

In fact this observation is also indicated in the Moran's *I* shown in table 6.3. At 25% the spatial correlation is roughly the same as the original data for both RRS and LDS. At 80%, with the RRS there is almost no correlation in the data. This seems to make sense since RRS swaps over much longer distances so has the potential to damage the data more than LDS which generally swaps over shorter distances. This contrasts with the results of the spatial rankings however, which indicated greater disturbance to the single attribute rather than the cross-classified attribute.

Table 6.3: Moran's I at LSOA for cross-classified attribute

riginal	0.2640	
	25% sample	80% sample
RRS	0.2564	0.0367
LDS	0.2747	0.2336

6.2.3 Conclusions: Impact on ESDA

In summary, it is difficult to use the spatial rankings to compare the effects of LDS and RRS across different levels of geography, since this depends on the size of the rank groups. However there was a clear observable pattern of LDS showing less disturbance than RRS, sometimes around twice the LSOAs or wards moved into a different rank group with RRS than LDS. The spatial autocorrelation showed an obvious pattern across different sampling fractions, with the 80% swaps showing a large amount of damage in the LISA maps. This was particularly true for the RRS swap at the 80% level for the cross-classified attribute; the Moran's *I* dropped significantly whereas with LDS some elements of the correlation remained.

In conclusion, the work in this section provides some evidence to suggest that LDS performs better than RRS and this pattern is particularly evident for larger sampling fractions. Further analysis will indicate whether this carries through to more complex methods.

6.3 Impact on Area Classifications

After ESDA, a statistical procedure commonly performed on census data is to try to summarise the statistical relationships. A typical spatial procedure is to use area / geodemographic classifications. Area classifications are commonly used with census data to capture and condense information about clusters in the population. The approach works by grouping together similar areas based on their values on a set of chosen variables. Geographical perturbation is likely to impact on area classifications since the relationship between the attribute and geography variables becomes distorted. After perturbation, the classifications that applied to the original data may no longer apply to the perturbed data; the groups may not be so distinct or new, different groupings may emerge. The effect of geographical perturbation on a classification of census data will be assessed in two ways:

1) A classification of wards (and EDs) is first created based on the unperturbed synthetic data for the Hampshire county. The classification is then applied to the swapped populations (LDS and RRS) and the fit assessed in terms of homogeneity within the clusters, measured by analysis of the distance to cluster seed.

2) New classifications are derived entirely from the swapped populations. The classification is then compared to that of the unperturbed data. A further possibility is to use the cluster means from the unperturbed data as the starting values for the cluster seeds at the first iteration on the swapped data (this shall be done for EDs).

6.3.1 General Area Classification Methodology

Area classification refers to the classifying of areas into groups of similarity based on the characteristics of selected features within them (Everitt et al. 2001). Geodemographic classifications are based on the same idea and refer to information about population location typically at small scales such as postcodes. There are a number of commercial and well-known area classifications such as ACORN and MOSAIC. These are often seen as a data reduction tool to offer a simplified description of the census database. Wallace et al. (1995) describe area classifications as meeting a need for an indicator of socio-economic information contrasting the similarities and differences between areas. Geodemographics / area classifications have many uses including target marketing, consumer profiling, academic research and allocation of resources.

Openshaw and Wymer (1995) provides a basis for classifying data which involves several steps. The first being to elucidate the purpose of the classification so that the appropriate census variables and areas can be selected. The next stage might involve transforming the data to remove outliers and to reduce skew. The variables should not be too highly correlated (one approach to this is to derive a new set of variables using principal components analysis). However assuming they are not, the data may be standardised to allow use of a cluster similarity measure that is a simple distance-based one. Otherwise the classification may be influenced by certain variables. The clusters can be obtained via 'cluster analysis' using some measure of similarity (or dissimilarity) such as within cluster sum of squares, and there are many ways of doing this. The researcher must select the number of clusters based on judgement requiring expertise in the discipline and knowledge of the purpose of the investigation. The results can then be mapped and the resulting classification displayed. An important step in area classification is to validate the results; determining whether there is significant structure within the clusters and whether the result would be the same on re-run.

The process of clustering is a well-known technique (Everitt et al, 2001) and involves two major decisions; choosing a measure of association / proximity measure and secondly the clustering method. Hierarchical agglomerative methods work by joining each object separately (a top-down approach) creating a cluster hierarchy. De-agglomerative methods work in the opposite way with all objects starting in one large cluster which is then split into smaller clusters. K-Means is a different and popular approach where the number of clusters is pre-determined. The K-Means clustering algorithm assigns each point in the dataset to the cluster whose average value on a set of variables is nearest to it by some distance measure (usually Euclidean) on that set. The algorithm computes these assignments iteratively, until reassigning points and recomputing averages (over all points in a cluster) produces no changes. The clusters may be summarised by their respective centroids (average of the cluster members' coordinates) in that space. Cluster analysis involves choice of a similarity or dissimilarity measure and is dependent on the type of variables; whether continuous, categorical or discrete. Euclidean distance is commonly used for continuous variables.

6.3.2 Creating an Area Classification for the Synthetic Data

The classification methodology used for the purposes of the experiment will closely follow the application of Vickers et al. (2007) where full details are described for creation of a national classification of census output areas. This classification was derived entirely from 2001 Census data which means the process can be replicated closely with the synthetic data. The Vickers' classification

was derived for the whole of the UK whereas the synthetic population only applies to the Hampshire region. Therefore the number of clusters and levels of hierarchy will be much smaller in this case. The K-Means method to be used will be performed in SAS rather than the SPSS approach used by Vickers et al. (2007).

- Step 1: Variable selection

Vickers et al. (2007) discuss in detail the variable selection process which involves studying correlations between the variables. Highly correlated variables are undesirable as the information would then be repeated in the clustering process. Variables that are correlated due to causality (one being a property of the other such as percentage of flats and percentage of households being the lowest level above the ground) should be eliminated as well as variables where the presence of one might indicate the presence of another (e.g. religion and ethnicity). In this analysis, a similar set of forty-one variables will be used (see appendix A6.1). However the work done by Vickers et al. (2007) was performed on 2001 census data whereas the synthetic data are based on the 1991 data, thus variable definitions vary slightly.

- Step 2: Standardisation

The variables should be approximately normally distributed to be able to differentiate between output zones with different values in the cluster analysis. This presents a problem with census data which is typically skewed:- the majority of the data are at the lower end of the scale. As in Vickers et al. (2007), the variables were first transformed by taking logs, which does not destroy the integrity of the data (unlike rankings for example), but reduces the skewness of the distributions.

Transforming the data by taking logs

$z_j^* = \log_{10}(z_j + 1)$ where z_j refers to variable j $(j = 1,\ldots 41)$

The variables were then standardised in order to ensure that each variable had the same weighting and to avoid the problems caused by outliers. Vickers et al. (2007) define two techniques for standardisation:

Range Standardisation

$zs_j^* = \frac{z_j^* - z_{\min}^*}{z_{\max}^* - z_{\min}^*}$ where $z_{\min}^*$ is the minimum and $z_{\max}^*$ the maximum of the transformed z_j^*

This idea was developed by Wallace and Denham (1996) and does not work well with outliers. An alternative is:

Inter-decile Range Standardisation

$zs_j^* = \frac{z_i^* - z_{p50}^*}{z_{p90}^* - z_{p10}^*}$ where z_{p10}^*, z_{p50}^* and z_{p90}^* are the tenth, fiftieth (median) and ninetieth percentiles respectively of the transformed z_j^*

The synthetic data were found to have the same problems as in Vickers et al. (2007), and so clustering was performed after applying both the inter-decile range standardisation and range standardisation techniques. Inter-decile range standardisation was most appropriate in this case for creating comparable variable distributions (of similar scales and magnitudes).

- Step 3: Clustering via K-Means

Cluster analysis is used to place each geographic zone into a group according to the key characteristics of the people who live there. The clusters should have roughly equal numbers of zones in them. In Vickers et al. (2007), K-Means was used in SPSS, an algorithm which minimises the within cluster variability in an iterative process. The number of clusters in the dataset must be pre-specified. K-Means can also be carried out in SAS using *proc fastclus*. First a random seed is chosen from the dataset for each cluster, at the next iteration the cluster seeds are reset to be the median of the clusters. The clusters are recalculated and clusters seeds reset, iteratively, until the maximum relative change in the cluster seeds is less than 0.0001. The relative change is calculated as the difference between the old and new seed values (in terms of Euclidean distance), divided by the mean absolute deviation from the cluster seeds in the current iteration.

- Step 4: Selecting the Number of Clusters

To produce a hierarchical classification, the K-Means algorithm can be run on the dataset to produce the top-tier clusters. The original dataset is then split into separate datasets according to the number of clusters (representing the higher level of the hierarchy) and the K-Means algorithm run on each dataset separately to create the next tier of clusters. This can be done a number of times to create the desired levels of classification (a top-down approach).

6.3.3 Fit of Swapped Data to a Classification of Wards

Following this process for the unperturbed synthetic data and using wards as the zones to be clustered, four main clusters were identified. The statistics for determining the clusters can be found in appendix A6.2. Appendix A6.3 then shows the radial plots for the sub-classifications. The description of the clusters is summarized in table 6.4. The description of the clusters has been given based on their distinctive characteristics but this does not need to be entirely accurate. The aim is to assess whether the same clusters appear (have the same distinctive characteristics) in the swapped and unperturbed populations; the 'type' of cluster does not matter. However the cluster detail can be used to validate the success of the synthetic data in replicating the Hampshire population.

Table 6.4: Description of the Clusters from a Ward Classification of the Unperturbed Data

Summary Description of Cluster	Cluster Number	Full Description
Single city dwellers	*Cluster 1*	*Urban, people living alone, Mid age group (25-44), service industry, above average minority ethnic groups.*
Affluent Older families	*Cluster 2a (urban, wealthy older families)*	*Urban, detached housing, two car households, older age group (45-64), families with non-dependent children.*
	Cluster 2b (rural, wealthy older families)	*Rural, two cars, detached housing, part-time workers, older age group (45-64), high SIR, families with non-dependent children.*
Families with young children	*Cluster 3a (traditional families with young children)*	*Rural, two car detached households, families with children (5-14), looking after home, older age groups (45-64).*
	Cluster 3b (rural families with young children)	*Rural, renting privately, families with predominately babies and young children.*
	Cluster 3c (suburban families with young children)	*Suburban, two car detached households with children and families with non-dependent children.*
Urban terraced blue collar	*Cluster 4*	*Urban, all age groups but not elderly, terraced, families with children and babies, many people per household, routine occupations, higher than average unemployment.*

Figure 6.5: Mapping the Unperturbed Ward Classification

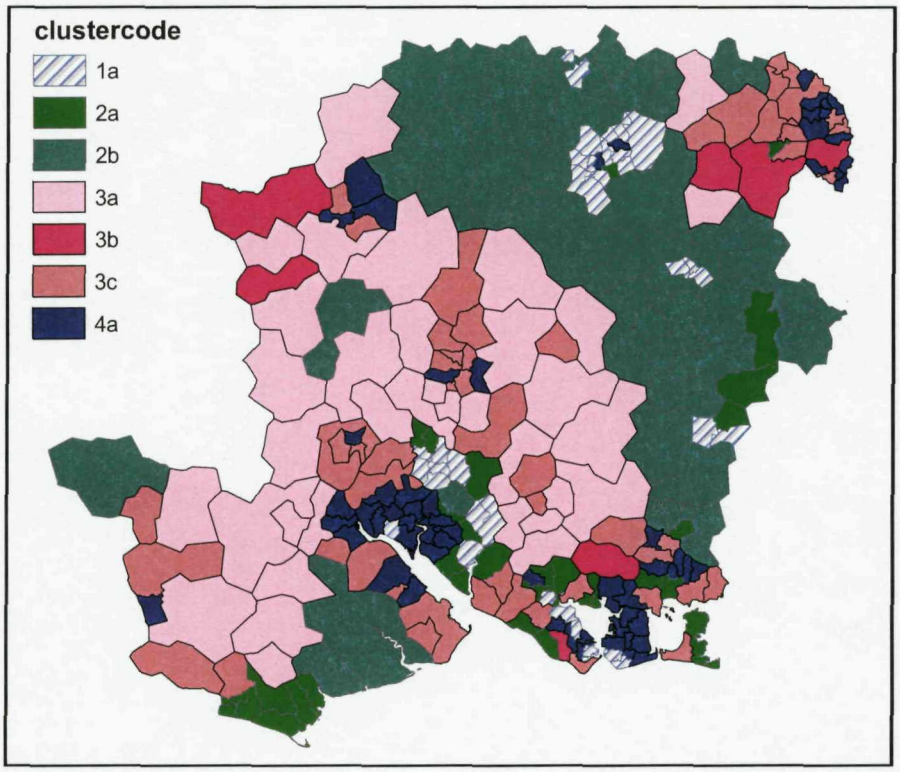


Figure 6.5 maps the clusters of the unperturbed (synthetic) ward classification. Urban areas such as Southampton and Portsmouth and central parts of Basingstoke are classified as 4a and 1a (urban areas) as expected. Suburban areas (cluster 3c) surround the urban areas for example around Southampton. Rural areas (clusters 2b and 3b) are also where there are expected, such as the New Forest area and the Test Valley. In that respect, the synthetic data seems to be a good representative of the true 1991 Hampshire census data. The wards in the synthetic data can be defined according to whether they are urban, suburban or rural areas as shown in table 6.5. This information will be used later in section 6.4 (for spatial modelling).

Table 6.5: Number of wards classified by urban/rural/suburban

Classification Type	*Number of Wards*
Urban	153
Suburban	53
Rural	61

The fit of the swapped populations to the unperturbed ward classification can now be assessed using various statistics. The Root Mean Square Deviation (RMSD) is a measure of the average distance between zones in the cluster for a particular cluster '*clus*' where η is the Euclidean distance between N_{clus} pairs of zones (within the cluster):

$$RMSD = \sqrt{\frac{1}{N_{clus}} \sum_{i=1}^{i=N_{clus}} \eta_i^2} \qquad (6.3)$$

The maximum distance from cluster seed to observation or the mean distance to seed can also be calculated. Results are presented for 10% and 25% RRS and LDS in table 6.6.

Table 6.6: Fit of the Swapped Populations to the Clusters in the Unperturbed Ward Classification

	Original	RRS 10%	LDS 10%	RRS 25%	LDS 25%
Cluster 1a (freq = 36)					
RMSD	0.1369	0.1345	0.1331	0.1355	0.1363
Max distance to seed	1.6714	1.7022	1.7393	1.8081	1.7274
Mean distance	0.7985	0.7780	0.7706	0.7767	0.7930
Cluster 2a (freq = 44)					
RMSD	0.1190	0.1167	0.1174	0.1215	0.1217
Max distance to seed	1.3846	1.4159	1.4280	1.5295	1.4502
Mean distance	0.7184	0.7023	0.7062	0.7267	0.7337
Cluster 2b (freq = 31)					
RMSD	0.1388	0.1354	0.1353	0.1318	0.1355
Max distance to seed	1.2995	1.3436	1.3647	1.2962	1.3028
Mean distance	0.8549	0.8317	0.8136	0.8118	0.8348
Cluster 3a (freq = 35)					
RMSD	0.1039	0.1081	0.1078	0.1123	0.1092
Max distance to seed	1.2128	1.2362	1.2690	1.2740	1.2668
Mean distance	0.6164	0.6439	0.6388	0.6692	0.6466
Cluster 3b (freq = 7)					
RMSD	0.1025	0.1060	0.1048	0.1046	0.1064
Max distance to seed	0.9097	0.9215	0.9411	0.9010	0.9374
Mean distance	0.5678	0.5903	0.5794	0.5851	0.5873
Cluster 3c (freq = 44)					
RMSD	0.0968	0.0983	0.0981	0.0998	0.0987
Max distance to seed	1.3711	1.3985	1.3893	1.4879	1.4258
Mean distance	0.5835	0.5931	0.5924	0.5966	0.5938
Cluster 4 (freq = 61)					
RMSD	0.1306	0.1305	0.1296	0.1305	0.1302
Max distance to seed	1.7083	1.7300	1.7213	1.7447	1.7514
Mean distance	0.7804	0.7784	0.7726	0.7735	0.7741

At both the 10% and 25% level, both LDS and RRS preserve the fit to the clusters with the *RMSD*, the maximum distance to seed and the mean distance to seed barely showing any changes from the

unperturbed values. This is in contrast to section 6.2.2 which showed the data (not the clusters) was becoming *more* homogeneous with a larger sampling fraction.

In conclusion, aside from the impact of sampling fraction, this test for utility of the data shows virtually no difference between LDS and RRS suggesting that both methods have little impact on the clusters at the ward level. This might be justifiable because wards are large aggregate areas containing approximately 2,500 households in each. This is a positive result from both methods; distortions to the data are minimal.

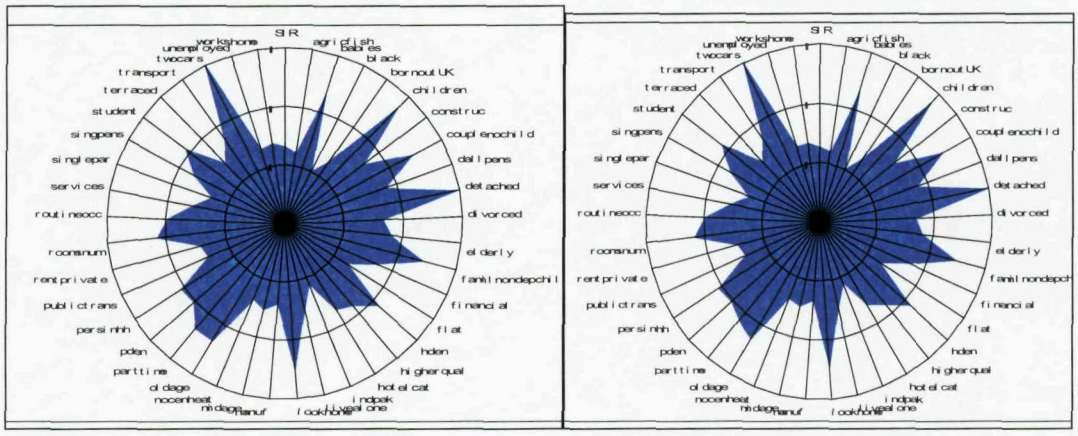
6.3.4 Creating a Ward Classification from the Swapped Data from Scratch

An alternative approach is to create the area classifications from scratch from the perturbed data and compare the outcome to the classification created from the perturbed data. The clustering procedure is run on the swapped data (RRS and LDS) and clusters identified as in the same way as before (no reference is made to the unperturbed data or classification).

Creating a Ward Classification for 10% LDS

The radial plots for the ward classification created from scratch from the 10% LDS population are shown in Appendix A6.4. The clusters were almost identical and could be directly matched to one of the four original clusters in the unperturbed classification. For example; compare cluster 3 original, to a cluster from LDS 10% in figure 6.6.

Figure 6.6: Radial Plots for the Ward Classification fitted from scratch to 10% LDS Data



Cluster 3: unperturbed data

A cluster from the 10% LDS classification

Moreover only seven out of the 267 wards were assigned to a different cluster comparing between the populations (after matching up the four clusters which were almost identical). This is indicated in the mapping of the classification in figure 6.8 which when compared to the unperturbed mapped classification in figure 6.7, shows hardly any change. In figure 6.8 we assume the four clusters created by 10% LDS are assumed to be identical to those found in the unperturbed population.

Figure 6.7: Mapping of the Unperturbed Ward Classification (four main clusters)

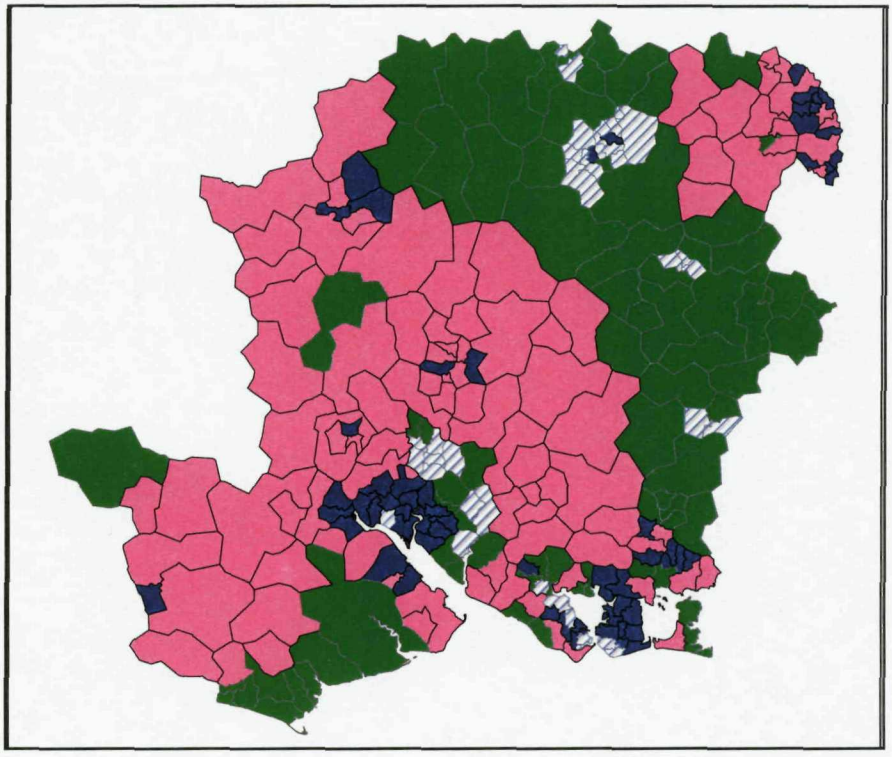
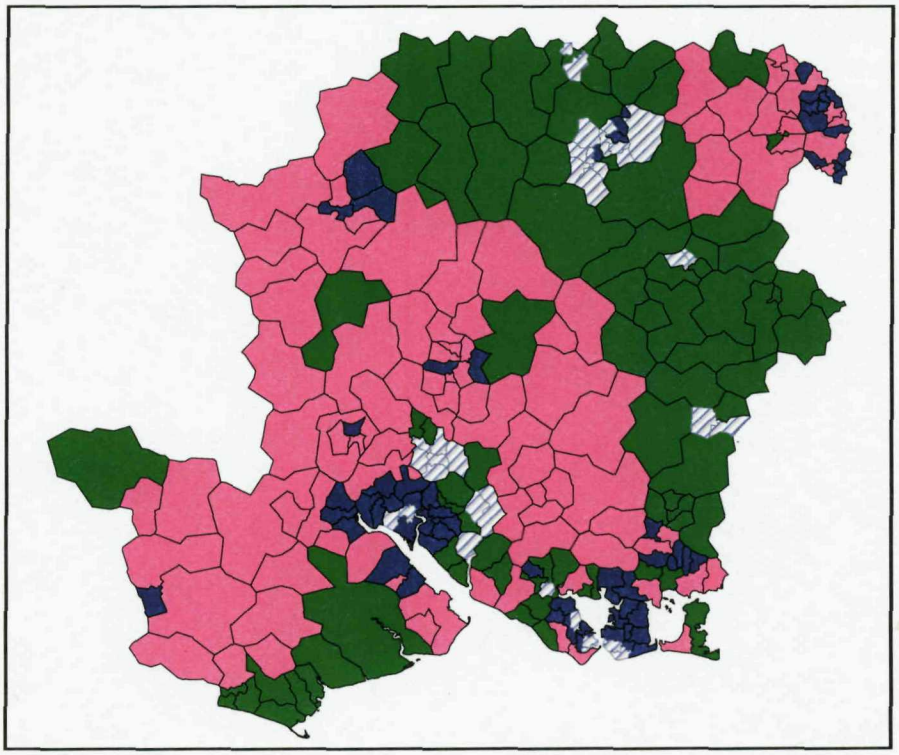


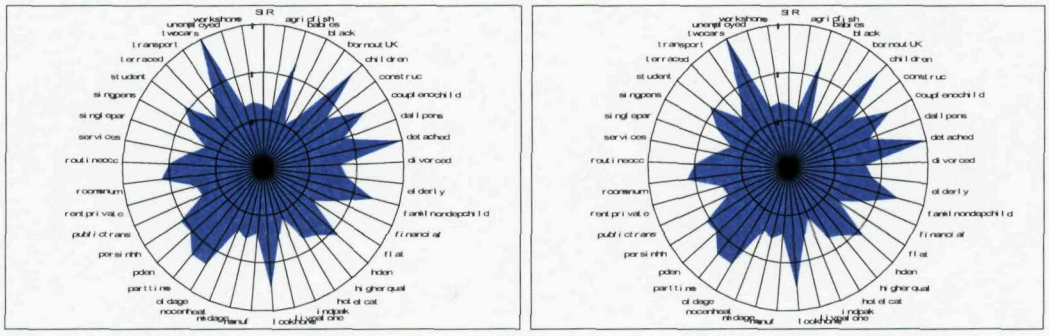
Figure 6.8: Mapping the LDS 10% classification from Scratch (four main clusters)



Creating a Ward Classification for 10% RRS

Again, the radial plots showed almost identical clusters which could be directly matched to the unperturbed classification. Compare in figure 6.9 the cluster from RRS 10% to cluster 3 from the unperturbed classification.

Figure 6.9: Radial Plots for the Ward Classification fitted from scratch to 10% RRS Data



Cluster 3: unperturbed data

A cluster from the 10% RRS classification

However, after matching up the clusters to the original classification, this time there was a lot of movement of wards between different classifications. Specifically almost half (121/267) were in the wrong cluster. This is reflected in the mapping of the RRS classification in figure 6.10. On the whole, the general pattern is the same, but at the smaller scale, it is easy to spot areas where the clusters have changed classification.

Figure 6.7 (repeated): Mapping of the Unperturbed Ward Classification (four main clusters)

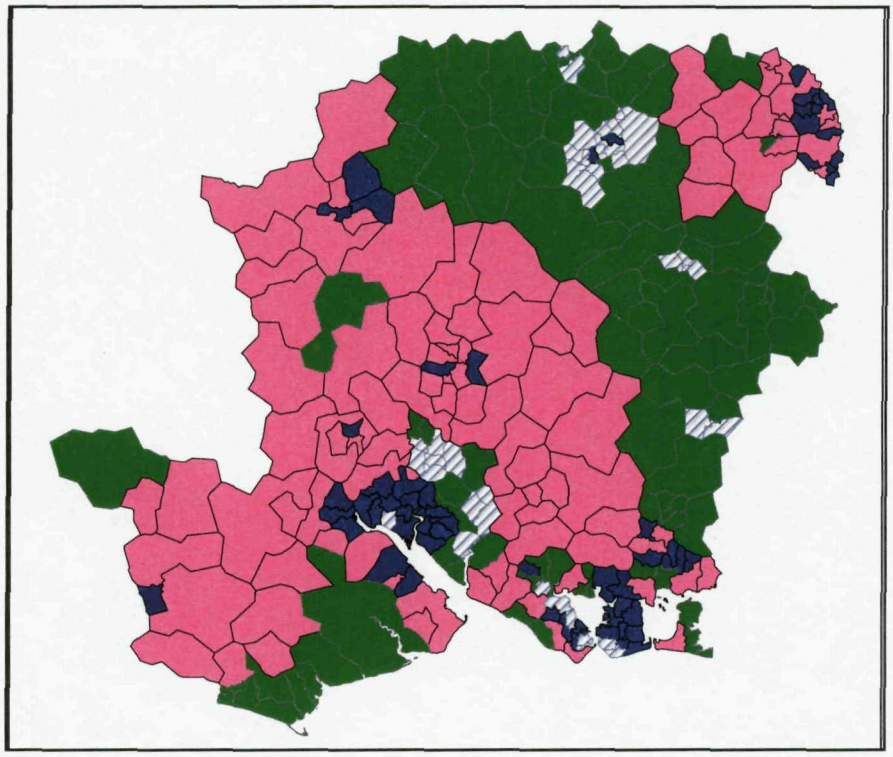


Figure 6.10: Mapping the RRS 10% classification from Scratch (four main clusters)

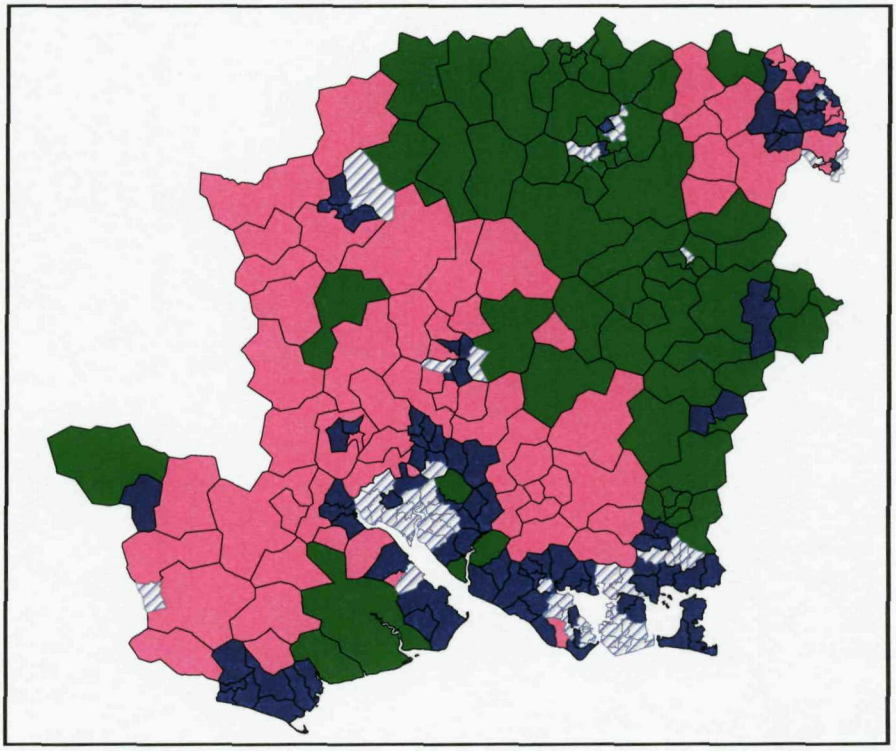


Table 6.7 displays the *RMSD* for each of the four main clusters found from creating the ward classifications from the unperturbed, 10% LDS and RRS 10% data, as well as the number of wards in each cluster.

Table 6.7: RMSD & Cluster Frequency of Ward-level Clusters created from 10% LDS and RRS swaps

Cluster ID	Unperturbed	LDS 10% Matched clusters	RRS 10% Matched clusters
1	0.1369 (36 wards)	0.1370 (34 wards)	0.1527 (65 wards)
2	0.1473 (75 wards)	0.1454 (79 wards)	0.1231 (65 wards)
3	0.1299 (95 wards)	0.1316 (92 wards)	0.1251 (86 wards)
4	0.1306 (61 wards)	0.1279 (62 wards)	0.1554 (51 wards)

Although the *RMSD* for the swapped populations is similar for both LDS and RRS, the number of wards in each cluster is quite different for RRS for cluster 1 in particular.

Creating a Ward Classification for 25% LDS and RRS

With a swapping rate of 25%, the RRS clusters were unrecognizable compared to the unperturbed. After LDS, cluster 4 is still evident but with some distortion. The other clusters from LDS are distorted and do not 'fit' to any of the unperturbed clusters. The urban classification (cluster 4) is probably still intact because the records in these areas are moved shorter distances by nature of the LDS methodology.

6.3.5 Fit of Swapped Data to a Classification of EDs

As EDs are smaller in size with an average of 220 households compared to an average of 2,619 in wards, the effect of swapping may be more pronounced. EDs in the unperturbed population for Hampshire were classified into two main clusters. The first of these was small containing only 46 EDs. The other cluster was large containing the majority of the EDs. Six sub-clusters were evident within this larger cluster. The clusters assigned to the EDs in the unperturbed population were then assigned to the same EDs in the swapped populations and the measures of homogeneity assessed (the *RMSD* and the maximum distance to cluster centre) as before with the wards. In this case,

sampling fractions of 10%, 25% and 50% are studied. Results are shown in table 6.8 (RMSD) and 6.9 (maximum distance).

Table 6.8: RMSD for each cluster when fitting LDS and RRS data to the Unperturbed ED Classification

	Unperturbed	LDS10	LDS25	LDS50	RRS10	RRS25	RRS50
1	0.1919	0.1928	0.1940	0.1981	0.1918	0.1952	0.2184
2	0.1207	0.1187	0.1203	0.1258	0.1187	0.1227	0.1267
2a	0.0899	0.0889	0.0910	0.0928	0.0887	0.0946	0.1005
2b	0.0909	0.0912	0.0899	0.0910	0.0923	0.0969	0.1027
2c	0.1165	0.1171	0.1151	0.1213	0.1175	0.1244	0.1265
2d	0.0963	0.0933	0.0945	0.1160	0.0933	0.0976	0.1031
2e	0.1156	0.1120	0.1110	0.1249	0.1128	0.1132	0.1178
2f	0.0884	0.0892	0.0895	0.0926	0.0900	0.0952	0.1016
Average	**0.1138**	**0.1129**	**0.1132**	**0.1203**	**0.1131**	**0.1175**	**0.1247**

Table 6.9: Maximum distance to cluster seed, for each cluster when fitting LDS and RRS data to the Unperturbed ED Classification

	Unperturbed	LDS10	LDS25	LDS50	RRS10	RRS25	RRS50
1	2.2173	2.2329	2.2305	2.2583	2.2555	2.2932	3.3007
2	1.7223	1.7179	1.7381	1.8656	1.7462	1.7322	1.8586
2a	1.2608	1.2567	1.2747	1.6326	1.2417	1.1607	1.3069
2b	1.4330	1.4412	1.4740	1.4805	1.4555	1.3061	1.5928
2c	1.3439	1.3607	1.4174	1.6482	1.3705	1.4554	1.5429
2d	1.2536	1.3127	1.2350	1.8251	1.2016	1.2046	1.3422
2e	1.3550	1.4128	1.6577	1.8260	1.4197	1.4438	1.6050
2f	1.4272	1.4541	1.4289	1.4743	1.4555	1.5267	1.8321
Average	**1.5016**	**1.5236**	**1.5570**	**1.7513**	**1.5183**	**1.5153**	**1.7977**

Tables 6.8 and 6.9 show little difference between the swapped and the unperturbed ED classification, for both RRS and LDS. Both the *RMSD* and maximum distance are almost unchanged. The only noticeable effect is of sampling fraction which tends to make the clusters less homogeneous as it becomes larger (both the *RMSD* and maximum distance increase). Thus, despite EDs being smaller in size, swapping still has little impact on the fitted clusters, and for both methods.

6.3.6 Creating an ED Classification from the Swapped Data from Scratch

Upon repeating the same analysis in section 6.3.4 at ED level (creating an ED classification from scratch for the swapped data), it was very difficult to identify similar clusters to the unperturbed data. Instead, the EDs in the swapped populations were classified according to the cluster centres from the unperturbed data (via the cluster means from 6.3.5) to one of the two main clusters and subsequently one of the six sub-clusters. The measure of utility in table 6.10 examines the number of EDs in the swapped populations that change cluster.

Table 6.10: Fitting Clusters after LDS and RRS, based on Cluster Centres from the Unperturbed Data

	LDS 10%	LDS 25%	LDS 50%	RRS 10%	RRS 25%	RRS 50%
Number of EDs changed main cluster	2% 59	3% 80	3% 111	2% 65	3% 96	3% 117
Number of EDs changed sub-cluster	28% 884	30% 937	51% 1594	28% 869	35% 1093	37% 1148

The changes relative to the cluster size are very small. It is hard to detect any difference between RRS and LDS. The amount of changes at sub-cluster level is much greater than for the main clusters.

6.3.7 Conclusion: Impact on Geodemographic Classifications

In conclusion, both the ED and ward classifications appear to be resistant to the effects of swapping in terms of the fit of the swapped populations to the unperturbed clusters. LDS fits the unperturbed clusters very well with the vast majority of wards falling in the same cluster but with RRS, the fit is not so good with approximately half of the wards falling in a different cluster. Despite this, the overall broad pattern for RRS is still much the same as before swapping. Creating the clusters from scratch on the other hand, is very unreliable and swapping has a much greater impact in this context. This indicates some information on the original (pre-swapping) pattern helps to retain utility. Any swapping rate over 10% severely distorts the data in both cases.

6.4 Impact on Estimates for Multilevel Models

An important statistical procedure performed using census data is to fit models for inference and prediction. These inferences may be used, for example, to formulate important policy decisions and thus it is essential that the model is accurate. Estimates for models which are spatial in nature, such as multilevel models, are likely to be distorted by geographical perturbation. Multilevel models recognize the existence of spatial clusters in the dataset, and may be used to separate unit and area level effects by treating the groups (areas) as a random sample from a population of groups. There are many examples of multilevel models that make use of census data; Moon et al. (2005) study the impact of area on health, Johnston et al. (2005) analyse voting behaviour and neighbourhood effects and Goldstein and Noden (2003) look at social segregation in schooling. Most of these studies use census variables to explain a non-census response. However Heady et al. (2003) have looked at multilevel models relating to small area data where the response has been a census variable including;

- The proportion of households with dependent children which contain one parent families
- The proportion of households without a car
- The proportion of households with more than one resident per room
- Proportion of economically active households containing a resident of working age who is unable to work due to sickness

Moreover Brunsdon et al. (1998) and Barnett et al. (2001) analyse the census variable limiting long term illness in terms of other census variables including density, unemployment and socio-economic grouping. For the remainder of this chapter, we refer to limiting long term illness as 'LLTI'.

With regard to geographical perturbation, the relationships between the geography and attribute variables are distorted, so the multilevel model that fits the original data may not be applicable after perturbation. More specifically, in the case of swapping, the data are thought to become more homogeneous as illustrated by the LISA maps in section 6.2.2. This means the differences between areas may be reduced and the area level variation may no longer be significant (thus a single level model should be used). LDS might be expected to show smaller differences to the unperturbed model when compared with RRS because households are moved shorter distances. At the individual level, no change in the parameter estimates would be expected because there is no interaction with geography; for example the effect of age on LLTI at the individual level should be unchanged after swapping. Consequently the lowest level of the model will be EDs. This section examines how the two methods (RRS and LDS) impact on the outcome of a typical multilevel model for census data. The analysis will involve fitting a multilevel model in MLWIN to the unperturbed census data in a

stepwise procedure. A Poisson multilevel model is proposed based on the LLTI variable. The results will then be compared with the same model specifications applied to the RRS and LDS populations. Different sampling fractions will be studied.

6.4.1 Model Specification

The response variable will be LLTI. This variable is a recent addition in the 1991 census. It is a popular variable used in census analysis since it covers a broad range of illnesses which may have resulted from exposure to certain work-related risks, poor housing conditions or stress (Brunsdon et al., 1998). It is well known that older age groups and males that are more likely to suffer from LLTI than females and the young. As in Barnett et al. (2001), premature LLTI will be analysed where the population at risk is all individuals in households under the age of 65. This is because for these younger age groups, the variable is likely to vary geographically and show strong associations in relation to socio-economic phenomena. This allows us to study in more depth changes between the unperturbed and swapped populations.

LLTI can be measured as a binary response; 'yes' or 'no' at the individual level or alternatively as counts within an area; i.e. percentage of individuals in the area who have a LLTI. Since census users only have access to aggregate outputs rather than microdata, the lowest level in the model will be small areas / neighbourhoods or more specifically Enumeration Districts. Many studies of LLTI using multilevel models take into account area context such as whether area level deprivation (Congdon, 1995) or rurality (Barnett et al., 2001) has an effect over and above the individual or neighbourhood level socio-economic factors. The population for analysis will represent one LAD in Hampshire. A very large dataset would cause problems in MLWIN. Southampton LAD is a suitable candidate as it consists of 29,009 households and is spatially heterogeneous being a major urban area in England.

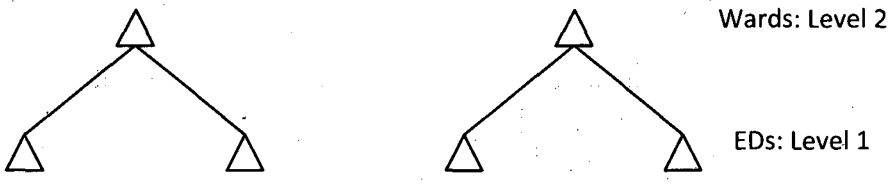
Poisson Multilevel Modelling

The **response will be the count of 40-49 year old males with LLTI in each ED**. Since the population at risk is large and the number of cases are relatively small, the Poisson distribution can be used to model the distribution across EDs. Using only a single level model would ignore the geographical structure of the data, and would ignore the fact that:

(a) smaller areas have a tendency to have similar rates of LLTI to their neighbours, i.e. EDs in the same ward having similar rates of illness

(b) counts of LLTI tend to differ according to the 'rurality' of the wards. Higher rates of illness are generally associated with more rural areas.

These hypotheses form the rationale for developing a multilevel model whose structure is as follows:



Let $LLTI_{ij}$ be the Poisson distributed response for ED *i* in ward *j*. Assume that the $LLTI_{ij}$ are independent Poisson random variables with expected rate π_{ij} such that $LLTI_{ij} \sim \text{Poisson}(\pi_{ij})$. $LLTI_{ij}$ might typically be standardised by age and sex as in Barnett et. al. (2001), but for simplicity only one age and sex group will be considered. A logarithmic transformation will be used to prevent the model from predicting negative numbers of LLTI and also means that the explanatory variables have a multiplicative effect. A single level Poisson model can be written following the notation of Rasbash et al. (2004) and Goldstein (2003):

$$\left.\begin{aligned} & LLTI_{ij} \sim Poisson(\pi_{ij}) \\ & \log(\pi_{ij}) = offs_{ij} + \beta_0.1 + \beta_1 predictor_{ij} \\ & \text{var}(LLTI \mid \pi_{ij}) = \pi_{ij} \end{aligned}\right\} \qquad (6.4)$$

β_0 is an intercept parameter and β_1 the slope parameter associated with the predictor$_{ij}$. $offs_{ij}$ is the logged area population which allows comparison of rates rather than number of cases which is generally sensible as otherwise the model will simply predict that more cases are seen with more

people at risk. The model assumes no under- or over-dispersion (i.e. $E(LLTI_{ij}) = Var(LLTI_{ij}) = \pi_{ij}$) and as in Leyland and Goldstein (2001), only Poisson variation in the response:

$$\left.\begin{aligned} & LLTI_{ij} = \pi_{ij} + e_{0ij} x^{*}_{0ij} \qquad \text{where } e_{0ij} \text{ are the residuals} \\ & x^{*}_{0ij} = \pi_{ij}^{0.5} \\ & \text{So } E(LLTI_{ij}) = E(\pi_{ij} + e_{0ij}\pi_{ij}^{0.5}) = \pi_{ij} + E(e_{0ij})\pi_{ij}^{0.5} \\ & \text{And } Var(LLTI_{ij}) = Var(\pi_{ij} + e_{0ij}\pi_{ij}^{0.5}) = 0 + (\pi_{ij}^{0.5})^2 Var(e_{0ij}) = \pi_{ij}.Var(e_{0ij}) \end{aligned}\right\} \quad (6.5)$$

Therefore it follows that the residuals must have $E(e_{0ij}) = 0$ and $Var(e_{0ij}) = 1$ if only Poisson variation is present. This model can be extended to the two-level case by assuming that the EDs are nested in wards.

$$\left.\begin{aligned} & LLTI_{ij} \sim Poisson(\pi_{ij}) \\ & \log(\pi_{ij}) = offs_{ij} + \beta_{0j}.1 + \beta_{1j} predictor_{ij} \\ & \beta_{0j} = \beta_0 + u_{0j} \\ & \beta_{1j} = \beta_1 + u_{1j} \\ & \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) \quad \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \\ & \text{var}(LLTI \mid \pi_{ij}) = \pi_{ij} \end{aligned}\right\} \quad (6.6)$$

The multilevel will study whether LLTI at ED level varies between wards and within different classes of ward. The classification divides wards into urban, rural or suburban depending on the outcome of the area classification (section 6.3; table 6.5). The model 6.6 will also be extended further to include predictors at ward level.

Maximum likelihood estimation is computationally intensive for Poisson models and so quasi-likelihood methods are used. In MLWIN, marginal quasi-likelihood (MQL) or predictive quasi-likelihood (PQL) can be implemented. Following the advice of Rasbash et al. (2004) MQL will be used first and followed by PQL using the MQL estimates. MQL is a crude approximation and may bias estimates downwards if the sample size within level 2 units are small or the response proportion is extreme. Therefore PQL offers improved estimates but these are less stable and can have convergence problems.

Six explanatory variables will be considered (all continuous);

- UNEMP: Proportion of unemployed

(based on the 1991 census variable *econprim* – '6' is unemployed)

- SINGLE: Proportion of single parents

(based on the 1991 census variable *famtype* – '7' lone parent with dependent children)

- CROWD: Proportion of households where there is overcrowding

(based on number of persons being greater than the number of rooms)

- RENT: Proportion of rented households (privately or otherwise)

(based on the 1991 census variable *tenure* – anything but '1' or '2')

- NOAMEN: Proportion of households without central heating

(based on the 1991 census variable *cenheat* – '3' no central heating)

- UNSKILL: Proportion of unskilled workers

(based on the 1991 census variable *soclass* – '6' is unskilled)

UNEMP is an indicator of economic well-being in the EDs whereas *NOAMEN* is an indicator of affluence/poverty. *SINGLE* takes into account household composition in the EDs. *CROWD* takes into account cramped housing conditions which may cause illness. *UNSKILL* attempts to take into account occupations which may lead to higher rates of LLTI. Finally *RENT* is another indicator of lifestyle which may affect LLTI.

The procedure for fitting the model for 40-49 males with LLTI in each ED will be as follows:

MODEL 1: Single Level Model (to identify significant predictors)
MODEL 2: Ward Level Variation (is there any significant variation in illness between wards?)
MODEL 3: Ward Level Predictors (attempting to explain the ward level variation, is there still any remaining variation?)
MODEL 4: Interactions Model (do the ED predictors vary in different types of ward?)

6.4.2 Results: Impact of LDS and RRS on Multilevel Model Estimates

Preliminary analysis showed some degree of correlation between the variables (at most 0.4) so the most significant will be included first and then the remaining predictors added back in. The multilevel models were then fitted extending from the single level model. Results are shown in tables 6.11 to 6.14.

Model 1: Single Level Model (three predictors were significant from the original six)

$$LLTI_{ij} \sim Poisson(\pi_{ij})$$

$$\log(\pi_{ij}) = offs_{ij} + \beta_0.1 + \beta_1 rent_{ij} + \beta_2 noamen_{ij} + \beta_3 unskill_{ij}$$

$$\text{var}(LLTI_{ij} \mid \pi_{ij}) = \pi_{ij}$$

Model 2: Ward Level Variation

$$LLTI_{ij} \sim Poisson(\pi_{ij})$$

$$\log(\pi_{ij}) = offs_{ij} + \beta_{0j}.1 + \beta_1 rent_{ij} + \beta_2 noamen_{ij} + \beta_3 unskill_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [\sigma^2_{0u}]$$

$$\text{var}(LLTI_{ij} \mid \pi_{ij}) = \pi_{ij}$$

Model 3: Ward Level Predictors

$$LLTI_{ij} \sim Poisson(\pi_{ij})$$

$$\log(\pi_{ij}) = offs_{ij} + \beta_{0j}.1 + \beta_1 rent_{ij} + \beta_2 noamen_{ij} + \beta_3 unskill_{ij} + \beta_4 urban_j + \beta_5 suburban_j$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [\sigma^2_{0u}]$$

$$\text{var}(LLTI_{ij} \mid \pi_{ij}) = \pi_{ij}$$

Model 4: Interactions Model

$$LLTI_{ij} \sim Poisson(\pi_{ij})$$

$$\log(\pi_{ij}) = offs_{ij} + \beta_{0j} urban_j + \beta_{1j} rural_j + \beta_2 rent.urban_{ij} + \beta_3 rent.rural_{ij} + \beta_4 noamen.urban_{ij} + \beta_5 noamen.rural_{ij} + \beta_6 unskill.urban_{ij} + \beta_7 unskill.rural_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$\text{var}(LLTI_{ij} \mid \pi_{ij}) = \pi_{ij}$$

Table 6.11: (Impact on a Multilevel Model) Single Level Model

MODEL 1: Parameter Estimates with Standard Errors in brackets	constant	rent	noamen	unskill
Unperturbed	-2.977 (0.032)	1.536 (0.106)	0.643 (0.153)	0.839 (0.281)
LDS 10%	-2.962 (0.033)	1.556 (0.108)	0.559 (0.154)	0.681 (0.278)
RRS 10%	-2.972 (0.032)	1.601 (0.102)	0.508 (0.161)	0.893 (0.283)
LDS 25%	-2.948 (0.031)	1.567 (0.106)	0.590 (0.154)	0.793 (0.269)
RRS 25%	-2.974 (0.033)	1.531 (0.115)	0.602 (0.160)	0.996 (0.294)
LDS 50%	-2.969 (0.033)	1.619 (0.109)	0.507 (0.163)	0.710 (0.292)
RRS 50%	-2.957 (0.033)	1.472 (0.108)	0.591 (0.159)	1.074 (0.293)

RENT: Proportion of rented households, NOAMEN: Proportion of households without central heating, UNSKILL: Proportion of unskilled workers

Table 6.12: (Impact on a Multilevel Model) Ward Level Variation

MODEL 2: Parameter Estimates with Standard Errors in brackets	constant	rent	noamen	unskill	Level 2 Variation
Unperturbed	-2.968 (0.039)	1.506 (0.127)	0.484 (0.179)	0.609 (0.292)	0.033 (0.012)
LDS 10%	-2.959 (0.034)	1.559 (0.111)	0.378 (0.173)	0.467 (0.290)	0.032 (0.010)
RRS 10%	-2.974 (0.035)	1.576 (0.122)	0.408 (0.183)	0.739 (0.304)	0.026 (0.011)
LDS 25%	-2.971 (0.036)	1.520 (0.125)	0.386 (0.169)	0.501 (0.288)	0.042 (0.010)
RRS 25%	-2.981 (0.036)	1.550 (0.126)	0.516 (0.174)	0.900 (0.292)	0.015 (0.008)
LDS 50%	-2.951 (0.037)	1.602 (0.125)	0.329 (0.175)	0.529 (0.305)	0.022 (0.008)
RRS 50%	-2.957 (0.033)	1.488 (0.124)	0.452 (0.164)	0.963 (0.294)	0.024 (0.009)

RENT: Proportion of rented households, NOAMEN: Proportion of households without central heating, UNSKILL: Proportion of unskilled workers

Table 6.13: (Impact on a Multilevel Model) Ward Level Predictors

MODEL 3: Parameter Estimates with Standard Errors in brackets	constant	rent	noamen	unskill	Level 2 Variation	urban	suburban
Unperturbed	-3.343 (0.070)	1.486 (0.122)	0.400 (0.161)	0.460 (0.286)	0.024 (0.009)	0.462 (0.072)	0.425 (0.086)
LDS 10%	-3.380 (0.077)	1.512 (0.122)	0.300 (0.168)	0.296 (0.311)	0.023 (0.008)	0.514 (0.081)	0.506 (0.092)
RRS 10%	-3.357 (0.068)	1.551 (0.117)	0.290 (0.169)	0.561 (0.293)	0.019 (0.008)	0.480 (0.070)	0.413 (0.081)
LDS 25%	-3.324 (0.070)	1.495 (0.122)	0.297 (0.167)	0.356 (0.285)	0.028 (0.009)	0.302 (0.082)	0.109 (0.070)
RRS 25%	-3.054 (0.055)	1.503 (0.125)	0.413 (0.179)	0.783 (0.309)	0.010 (0.005)	0.163 (0.057)	-0.150 (0.083)
LDS 50%	-3.352 (0.073)	1.577 (0.122)	0.241 (0.174)	0.367 (0.304)	0.019 (0.008)	0.192 (0.062)	0.103 (0.071)
RRS 50%	-2.978 (0.057)	1.466 (0.126)	0.392 (0.170)	0.830 (0.274)	0.017 (0.008)	0.091 (0.058)	-0.301 (0.086)

RENT: Proportion of rented households, NOAMEN: Proportion of households without central heating, UNSKILL: Proportion of unskilled workers

Table 6.14: (Impact on a Multilevel Model) Interactions Model

MODEL 4 Parameter Estimates with Standard Errors in brackets	urban	rural	urban variation	rural variation	urban. rent	rural. rent	urban. noamen	rural. noamen	urban. unskill	rural. unskill
Unperturbed	-2.889 (0.039)	-3.361 (0.132)	0.024 (0.008)	0.081 (0.042)	1.511 (0.132)	1.012 (0.574)	0.404 (0.176)	0.610 (0.774)	0.417 (0.290)	1.289 (2.019)
LDS 10%	-2.878 (0.040)	-3.339 (0.122)	0.021 (0.009)	0.075 (0.034)	1.577 (0.128)	0.991 (0.544)	0.278 (0.178)	0.602 (0.803)	0.283 (0.306)	-0.014 (1.820)
RRS 10%	-2.901 (0.040)	-3.302 (0.125)	0.015 (0.007)	0.088 (0.049)	1.605 (0.124)	1.078 (0.540)	0.313 (0.183)	0.343 (0.827)	0.573 (0.290)	-0.450 (2.122)
LDS 25%	-2.839 (0.037)	-3.300 (0.119)	0.027 (0.009)	0.069 (0.051)	1.519 (0.125)	1.305 (0.519)	0.320 (0.170)	0.236 (0.776)	0.338 (0.287)	0.847 (2.049)
RRS 25%	-2.947 (0.035)	-2.996 (0.118)	0.013 (0.006)	0.060 (0.032)	1.620 (0.129)	0.472 (0.544)	0.425 (0.180)	0.207 (0.761)	0.800 (0.308)	-1.811 (1.947)
LDS 50%	-2.873 (0.038)	-3.391 (0.123)	0.019 (0.006)	0.062 (0.055)	1.585 (0.125)	1.627 (0.515)	0.236 (0.177)	0.704 (0.800)	0.407 (0.307)	-1.055 (2.010)
RRS 50%	-2.923 (0.036)	-3.177 (0.135)	0.019 (0.008)	0.064 (0.040)	1.541 (0.120)	0.839 (0.497)	0.438 (0.174)	0.111 (0.819)	0.753 (0.313)	1.507 (1.727)

RENT: Proportion of rented households, NOAMEN: Proportion of households without central heating, UNSKILL: Proportion of unskilled workers

Before interpreting the models, we note that the average count of LLTI by ED is 2.22 with a variance of approximately 1.73 (so no strong evidence of under- or over-dispersion). The average count of males aged 40-49 in EDs (the population at risk) is 35 with a large standard deviation, therefore it is sensible to adjust the base count per ED using an offset[29].

Table 6.15(a): Population at Risk (LLTI as response in multilevel model)

	Mean	Min	Max	Std
Population at Risk (40-49 year old males in EDs)	35	1	91	13

Table 6.15(b): Counts of LLTI by ED

	Mean	Min	Max	Std
Counts of LLTI by ED	2.22	0	10	1.73

- Model 1

The parameter effects are multiplicative in a Poisson model, and appear reasonable although their effects are very small because of the small average count of LLTI; there would have to be a substantial increase in the proportion of the predictors to produce a change in the cases of LLTI. However such a substantial increase in proportions is feasible for all three predictors:

	Proportion renting by ED	Proportion with no amenities by ED	Proportion of unskilled workers by ED
Minimum value	0%	0%	0%
Maximum value	50%	58%	92%

For example:

- For a 10% proportion for all three predictors, the number of cases of LLTI would be 0.3 for the unperturbed model and also for LDS50 and RRS50, with a population at risk set to the average of 35.

[29] The offset adjusts the model to allow for the population at risk; otherwise the model would simply predict more cases of long term illness for larger populations. An offset is used here because of the large variation in population at risk across EDs.

- For a 50% proportion for all three predictors, the number of cases of LLTI would be 1.1 for the unperturbed model, 1.0 for LDS50, and 1.2 for RRS50, for the same population at risk set to the average of 35.

This is a positive result. Even after swapping 50% of the records, in both cases, the changes in the model are very small compared to the unperturbed data and the general interpretation would be the same.

- Model 2

Model 2 shows significant variation in LLTI between wards but this variation is very small at 0.033. RRS consistently reduces the size of this variation whereas for LDS, the variation actually increases on one occasion for the 25% sample. There is evidence in this model to suggest that sampling variation may be more important in this context. For example, RR25 decreases the ward level variation to 0.015 but then it increases again to 0.024 for RRS50.

- Model 3

Model 3 shows that there is a significant effect of the type of ward on LLTI but again this effect is very small. Table 6.16 illustrates just how small the variation between urban and rural wards is and shows there is slightly more change to the RRS50 model compared to the LDS50 model but in terms of general interpretation of the model estimates, the impact of swapping is tiny. We might be concerned if the increase for RRS50 was 10 more cases for rural for example, rather than the small value of 0.15. Since the difference in variation, although significant, is very small, it is perhaps not the best example to pick up any differences between LDS and RRS.

Table 6.16: Interpretation of Multilevel Model 3 for LDS50 and RRS50

	Unperturbed	LDS50	RRS50
Increase in predictors from 10% to 40% (for count of 35 at risk)	0.33 more cases for urban 0.33 more cases for rural	0.23 more cases for urban 0.21 more cases for rural	0.42 more cases for urban 0.29 more cases for rural
LLTI risk in urban and suburban wards compared to rural wards (averaged over models with 10% proportions and 40% proportions)	0.02 more cases for rural	0.03 more cases for rural	0.15 more cases for rural

- Model 4

Model 4 also shows significant variation within urban and rural wards. Their covariance is zero as would be expected. The general trend is for the urban and rural variation to decrease for both RRS and LDS, as the sampling fraction increases. An example interpretation of the effect of urban compared to rural wards, for LDS50 and RRS50 compared to the unperturbed model, is shown in table 6.17.

Table 6.17: Interpretation of Multilevel Model 4 for LDS50 and RRS50

	Unperturbed	LDS50	RRS50
LLTI risk compared to urban wards (all predictors 10% proportion)	0.11 more cases than rural wards	0.15 more cases than rural wards	0.08 more cases than rural wards
(all predictors 60% proportion)	0.12 more cases than rural	0.67 more cases than rural	0.44 more cases than rural

6.4.3 Conclusions from Multilevel Modelling of Census Data

The multilevel model used was at a small scale, studying differences between LLTI in EDs. The response looked at a specific target group; 40-49 year old males only. This kind of analysis might be used in practice to target certain types of EDs to reduce illness rates. We saw, for example, how EDs with a large rental population, who are unskilled with low access to amenities (50% proportions for all predictors) have greater rates of LLTI than those EDs which have a low or non-existent rental population, who are mainly skilled and have access to amenities (10% proportions for all predictors). The outcome of these models was unchanged after both RRS and LDS, even when swapping 50% of households. Moreover the variation between wards and the effect of urban and rural wards, although small but significant, also showed little change after applying RRS and LDS and was still significant. In terms of comparing LDS and RRS, LDS does seem to retain the level 2 variation to a greater degree than RRS which in some cases halves it. In conclusion, the impact of both swapping methods was negligible in terms of general interpretation of these particular models, with LDS generally preserving the model slightly better than RRS. In terms of taking this work forward, a multilevel model which shows much larger differences in rural/urban variation (or whatever the level one structure is) would be helpful to draw more conclusive results on the impact of RRS and LDS.

6.5 Discussion

In this chapter an attempt has been made to study a variety of analyses that might be affected by geographical perturbation with specific reference to the RRS and LDS methods. In addition, different properties of the data were studied; in section 6.2 the underlying trends in the data were examined, in section 6.3 summary measures of the data were studied, and in section 6.4 inferences from a multilevel model were considered. First some general remarks are summarised regarding swapping. Then some general themes that have emerged from the analysis are discussed in regard to the differences between RRS and LDS.

Results showed that sampling fraction has a large impact on the quality of the data after swapping. At the 80% level for example, there was significant distortion to the spatial rankings and to the LISA maps. The geodemographic classifications were also very distorted for sampling fractions larger than 10%. However providing some information about the unperturbed data that is non-disclosive can potentially help the user to obtain very accurate results despite a large amount of swapping, e.g. the cluster centres of an area classification could be published (if non-disclosive), or the cluster which an ED or ward falls in (at the higher aggregate level). This would be an alternative to the user creating a classification from scratch for example, which proved to give very unreliable findings. A general result also found and that makes sense, is that models that were very significant or data that demonstrated strong patterns/trends, were more resistant against the effects of swapping.

Returning to the main objective of the chapter, which was to compare the impact on utility between LDS and RRS, we now make some general remarks that can be summarised from the results as a whole. Area classifications were better preserved by LDS than RRS; this is likely to be because density is a big indicator of household type and LDS involves swapping with households in nearby areas and of similar household density. Spatial autocorrelations were better preserved with LDS than RRS for probably the same reasons. In terms of spatial rankings, LDS consistently performs better than RRS with fewer LSOAs or wards moving rank group. The multilevel model was disappointing in that the scale of the parameter estimates meant that no major differences could be seen between LDS and RRS with both methods generally producing very accurate results, close to the unperturbed models. There is the possibility that RRS had a tendency to reduce ward level variation more than LDS but no firm conclusions could be made from the multilevel models as there were occasions where the variation actually increased. This would be a possible area for further work.

Chapter 7 Conclusions and Further Research

7.1 Summary

The aim for NSIs is to publish as much accurate data as is possible to meet user demand without compromising disclosure risk. The specific objective of this thesis has been to develop a new methodology to address the spatial nature of disclosure risk in census data arising from the problem of geographical differencing and small area outputs. This is an important problem amongst the SDC literature as demonstrated in chapter 2 and many NSIs are moving towards implementation of a flexible output system exemplified by the recent ABS (2006 Census) table builder.

In chapter 3, a new framework of geographical perturbation methods was identified for census data. This research focuses specifically on the spatial nature of disclosure risks in census data. Geographical perturbation has a number of advantages not least that these methods are pre-tabular in nature. This results in both consistent and additive outputs from the protected microdata, benefits which are amongst the most important priorities for users. The new framework encompasses fully-tested methods such as RRS, geomasking techniques for application to census data as well as improved variants of RRS such as LDS. The methods all have in common the modification of spatial point locations in order to create uncertainty into the attribute-geography relationships, thus addressing disclosure risk from a spatial perspective. Which method is used would depend on the requirements of the statistical agency in respect to the advantages and disadvantages of each method in terms of ease of implementation, speed and flexibility. Zone-independent methods in particular have been developed in this thesis to resolve the geographical differencing problem which may occur in a flexible tabulation scenario and to

move away from SDC methods which rely on pre-determined output zones. Zone-independent methods are much more resilient to future reaggregation challenges.

Displacement methods offer an alternative geographical perturbation approach whereby households have unrestricted movement in the census region. Risk-utility outcomes were similar to swapping for the same sampling fraction and for the same mean perturbation distance. Only a preliminary investigation of displacement for census data was carried out in this thesis, but it offers an advantage for those particular types of records which might be difficult to pair for swapping, e.g. large households and communal establishments. If a combination of SDC methods were to be applied to census data, then these hard-to-match records could be displaced with the remainder of the sampled households swapped.

LDS was the most promising new method and is an improvement over RRS being zone-independent. Therefore it provides stronger protection against differencing. Swapping records at the 'local' level and in proportion to population density has the important benefit that it permits small area tables to be produced with less disclosure risk. Although more noise is added at the local level, similar levels of protection and damage to those seen for more conventional methods are observed for larger zones. In chapter 6, utility was explored by studying the impact of swapping on ESDA techniques, area classifications and changes to a multilevel model. Particularly for the smaller swap rates, the most significant patterns in the data were still remained with LDS (unlike RRS) as indicated by the LISA maps and fewer changes were noticed in spatial rank. The area classifications showed a similar picture with the 10% swap retaining most of the pattern in the case of LDS with much more change shown after RRS. However assigning the EDs and wards to predetermined clusters (from the unperturbed classification) showed very little deviation in terms of homogeneity measures for both LDS and RRS. In terms of the impact on a multilevel model, little change was seen in the parameter estimates of the predictors and the general interpretation was unchanged even for as high as a 50% swap (for RRS and LDS). It was the variation within urban and rural wards and their interacting predictors which were affected by swapping but mostly by the level of swapping rather than whether LDS or RRS was used.

LDS has in addition, other benefits over RRS. By swapping shorter distances, the likelihood of edit failures occurring is diminished (for example swapping a farming household into an urban area). Moreover this means that the complicated effect on Origin-Destination tables is also reduced since the relationship on distance travelled for residence and workplace is less likely to be distorted (if swapping were applied to all types of data).

In terms of a disclosure control strategy for a census, geographical perturbation is primarily a spatial SDC method and so does not change the disclosure risk arising from special uniques (uniques independent of geography). However these special uniques are usually rare and most risk can generally be attributed to geography. A statistical organization would normally desire that disclosure risk should be less than 50% so that the odds are against an intruder finding a true unique. Ideally the disclosure risk would be even smaller than this, at around 10% or less. Swapping a 10% sample under either method was very ineffective and the percentage of true uniques remained above 80%. This is straightforward to deduce and thus not releasing the swapping rate to the public seems logical. Even swapping 25% of the data resulted in a disclosure risk still above 50% at all levels of geography and this helps to explain why utility remained so high in the LISA maps at this level. To reduce disclosure risk sufficiently would probably mean swapping a very high proportion of records under either method. Organizations may consider it unacceptable to implement such high swapping levels, which means it is unlikely that swapping would ever be used as a sole protection method. However, for many NSIs, a combination of SDC methods often provides the solution. This allows different methods to target the different types of disclosure risk and, in addition, the combination of methods may ameliorate the biases originating from different types of SDC methods. For example, LDS (and displacement for unusual records) may be combined with (unbiased) small cell rounding. Small cell rounding is applied independently to tables so on the whole, outputs are additive and consistent which pleases users and protects confidentiality.

7.2 Further Research

If LDS were to be considered for application to a Census, the first most important area for further research would be to conduct tests on real census data. Whilst the synthetic data (chapter 4) was realistic by nature of the microsimulation process, drawing households from the 1991 SAR, it omitted 12-person households and communal establishments. Moreover only 30% of the microdata records were unique (although 92% of records were unique within an LAD). Although these limitations may be minor, it is important to get accurate results to compare different SDC methods fairly. In addition, in this thesis, little attention has been paid to the variety of census outputs since the focus has been on the scenario of providing flexible tabulation outputs from the underlying census microdata. A full evaluation of the method for application to real census data would have to include suitability for protecting other outputs such as the SARs and Origin-destination tables. These data are based on the same base population and are linked. Thus if the small area tables (based on the fully enumerated population) were compared to the SAR (based

on a sample) and protected by different methods, the protection could be undone by differencing if consideration is not given to the outputs as a whole.

In this thesis, disclosure risk was measured in terms of cell uniques. Risk assessment may, for example, be extended by looking at disclosure from zeros and other small cells. Some NSIs consider cell uniques in the margins to be most disclosive, representing attribute disclosure since the respondent must take that particular variable characteristic. Measurement of utility on the other hand is also far from straightforward. Since LDS is dependent on population density, it would be a logical extension to study the effects in regions of high and low density more closely. In addition, the impact of sampling variation could be determined more precisely as well as the effect on utility of targeting specific records.

In terms of specific extensions to the methodology; a promising approach not experimented with is *targeted local density swapping*. This is likely to further reduce risk subject to what the NSI deems 'risky records'. Displacement methods were only briefly studied and there is potential for further work here, in particular, to look at moving records to 'matched' areas which attempt to retain utility. Additionally we could consider in more detail the rearrangements approach.

Finally we may also consider application of this work to other high sample fraction data sources; a topic that has been outside the scope of this thesis. Gutmann and Stern (2007) discuss the challenges of confidentiality relating to mapped data (geoprivacy). There are other fields where confidentiality is important in a spatial context such as analytical outputs from health or education data (non-census). A typical example is geodemographic classifications for ethnicity or education profiling. Mateos et al. (2007) investigate name-based ethnicity classifications for Britain where individuals are classified into categories of Cultural, Ethnic and Linguistic groups based on the probable origins of names. Data are collected at very fine spatial scales. This has many disclosure implications when there are very few people of a certain type. Moreover geodemographic tools can be applied in education profiling to examine and profile students who apply to higher education institutions to ensure and action widening participation strategies. However geodemographic indicators may threaten privacy when appended to other data such as university application successes. In these contexts, geographical perturbation could be implemented (using units other than households) to introduce a level of uncertainty into the geographical-attribute relationships to protect against disclosure.

7.3 Conclusion

With the next round of Censuses coming up in 2011 in the UK, statistical agencies will be searching for appropriate methodologies for disclosure control of their data. Given the demand for small area data and flexible outputs, methods that address the spatial nature of risk will be particularly advantageous and haven't been previously explored in the SDC literature. The zone-independent methods developed provide greater protection against differencing and arbitrary boundary changes. Geographical perturbation methods such as displacement and swapping may be considered as variants of the well-tested RRS approach which has been applied to both UK and US Censuses. Displacement offers a solution for hard to match households. LDS, which has been discussed in-depth in this thesis, reduces small area risk when applied at the local level, and in implementation in proportion to density improves the overall R-U outcome. Pre-tabular methods such as these are very flexible and the sampling fraction can be adjusted to give the protection required. Essentially, LDS employs greater spatial intelligence and whether or not record swapping alone is judged to provide sufficient protection, it is a powerful disclosure control method which may provide a more efficient balance of risk and utility.

Appendices

A2.1 Linked Tables and Differencing

The following shows how linked tables can be compared to obtain the full three-dimensional table.

Figure A2.1: Example of Differencing via Linked Tables

(a) Location of business and sex of self-employed shopkeepers

	Centre	Outskirts	Total
Male	19	6	25
Female	3	6	9
Total	22	12	34

(b) Financial position and sex of self-employed shopkeepers

	Weak	Strong	Total
Male	13	12	25
Female	3	6	9
Total	16	18	34

(c) Location of business and financial position of self-employed shopkeepers

	Weak	Strong	Total
Centre	7	15	22
Outskirts	9	3	12
Total	16	18	34

Solution

This problem can be solved in a number of ways but here we take an approach using bounds for each cell and narrowing down the intervals using the table constraints.

Figure A2.2: Solution to Differencing from Linked Tables

Cell Bounds	Weak centre	Strong centre	Weak outskirts	Strong outskirts
Male	0 – 7	0 – 15	0 - 9	0 – 3
Female	0 - 7	0 - 15	0 - 9	0 - 3

From table (c)

Cell Bounds	Weak centre	Strong centre	Weak outskirts	Strong outskirts
Male	0 – 13	0 – 12	0 – 13	0 – 12
Female	0 - 3	0 - 6	0 – 3	0 – 6

From table (b)

Cell Bounds	Weak centre	Strong centre	Weak outskirts	Strong outskirts
Male	0 – 7	0 – 12	0 - 9	0 – 3
Female	0 - 3	0 - 6	0 - 3	0 - 3

Keep the smallest intervals

Cell Bounds	Weak centre	Strong centre	Weak outskirts	Strong outskirts
Male	4 – 7	9 – 12	0 - 9	0 – 3
Female	0 - 3	3 - 6	0 - 3	0 - 3
Totals (table c)	7	15	9	3

Narrow down the intervals further using margins from table (c)

Solution	Weak centre	Strong centre	Totals (table a)	Weak outskirts	Strong outskirts	Totals (table a)
Male	7	12	19	6	0	6
Female	0	3	3	3	3	6

Narrow down the intervals again using margins from table (a) - enough to work out the answer

A2.2 Semi-linked Tables and Differencing

Here is an example based around the data given in Algranati and Kadane (2004), to show how semi-linked tables based on different sources can be pieced together to reveal a higher dimensional dataset. The aim of the report was to show whether race, either of the defendant or of the victim, was an important determinant of the Department of Justice's decision on whether to seek the death penalty – a controversial issue in the US. The higher dimensional dataset would have been confidential.

Figure A2.3: Semi-linked Tables and Example of Differencing

Table (1) – Counts of crimes by Federal District by Ethnicity of Defendant (Rhode Island shown only)

Federal District	Number of crimes	Crime	Ethnicity
Rhode Island	4	18 USC 1959 (a)	Black
Rhode Island	1	18 USC 1959 (a)	Hispanic

Table (2) – Recommendations by Federal District by Ethnicity (Rhode Island shown only)

Federal District	Ethnicity	Cases	Recommendation
Rhode Island	Black	4	No recommendation
Rhode Island	Hispanic	1	No recommendation

Table (3) – Ethnicity of defendant among cases of exclusively Hispanic victims (Rhode Island only)

Federal District	Number of Cases	Ethnicity
Rhode Island	4	Black
Rhode Island	1	Hispanic

Table (4) – Ethnicity among cases with a single victim (Rhode Island shown only)

Federal District	Number of cases	Ethnicity
Rhode Island	5	Hispanic

Higher-dimensional table

Federal District	Ethnicity	Cases	Recommendation	Crime	Ethnicity of victim	Number of victims
Rhode Island	Black	4	No recommendation	18 USC 1959 (a)	Hispanic	1
Rhode Island	Black	4	No recommendation	18 USC 1959 (a)	Hispanic	1
Rhode Island	Black	4	No recommendation	18 USC 1959 (a)	Hispanic	1
Rhode Island	Black	4	No recommendation	18 USC 1959 (a)	Hispanic	1
Rhode Island	Hispanic	1	No recommendation	18 USC 1959 (a)	Hispanic	1

A5.1 Statistics about the Swaps

Table 1 shows where the household records have moved to after swapping using the LDS and RRS methods. (3) was calculated independently of (2).

Total households = 595,174	LDS10	RRS10	LDS25	RRS25	LDS50	RRS50
(1) Percentage of households with a different **geographical reference** after swapping	55,531 (9.3%)	55,520 (9.3%)	116,829 (20%)	135,718 (23%)	297,417 (50%)	298,595 (50%)
(2) Percentage of households in a different **ED** after swapping	42,387 (7%)	55,520 (9%)	88,570 (15%)	103,400 (17%)	136,067 (23%)	246,499 (41%)
(3) Percentage of households in a different **ward** after swapping	21,746 (4%)	51,786 (9%)	45,644 (8%)	16,668 (3%)	64,942 (11%)	44,604 (7%)

Part of the reason for the shortfall in numbers of swapped records for example for (1), LDS10 only resulted in 9.3% swaps rather than 10% because either suitable matches could not be found or households with the same reference were swapped (note that 15% of households in the synthetic population were repeated three times and approx. 1% four or more times.) These results are consistent with Boyd and Vickers (1999), where 9.7% of records were actually swapped for a 10% swap rate.

Although LDS is a local swap, similar amount of households are swapped between EDs, probably because of the population density effect. LDS is generally more predictable than RRS. With LDS, there is a pattern of roughly half the swaps between EDs also moving between wards. With RRS, this is not so predictable. In addition, a different initial sample was selected for each of the swaps so much of the results depend on where the initial households are located (i.e. if they are near a ward boundary for example).

A5.2 Sampling Variation

Sampling variation is potentially an important concern with geographical perturbation as there are many possible ways in which the sample can be selected from the population. Boyd and Stokes (1999) studied the effect of sampling variation on RRS examining how ED proportions of an attribute vary for different swap rates. Swap rates of 1%, 3%, 5%. 10% and 20% were examined. As an example, table (1) shows results for the employment variable using a 20% swap and five different runs:

Table (1): Effect of Different Samples for Swapping on the Percentages of Employment

	Employed	Unemployed	Inactive	NCR
Original	38.4	13.2	29.5	18.9
Run1	38.6	10.3	32.2	18.9
Run2	38.6	11.6	30.8	18.9
Run3	39.8	12.5	28.7	19.0
Run4	38.9	13.0	29.2	18.9

In general, the larger swap rates had the most variation as might be expected. The variation was also dependent on the variable as illustrated in table (2):

Table (2): Effect of Different Samples for Swapping on the Percentages of Certain Attributes

	LAD average	Largest increase over all EDs in LAD	Largest decrease over all EDs in LAD
Proportion Single	43.1	1.0	2.0
Unemployment	5.1	1.1	3.0
Ethnic minorities	5.0	1.3	15.7
Born outside the UK	5.7	8.0	<1
Public housing	27	5.7	6.1

The authors also mention that cases of increase and decrease as large as 1% were rare with most EDs showing no change.

A5.3 Full Results for the Differenced OAs and EDs

Note that many of the OAs contain very few numbers of households whereas the EDs contain more sensible numbers. This is due to the synthetic population created by a microsimulation technique (see chapter 3) which was based around EDs to sample appropriate numbers of households. Within the EDs the locations of households were artificially generated, thus some OAs contain very small numbers. This is not always reflective of the true census population since the England & Wales OAs were designed specifically to contain a minimum number of households.

Table: Differenced Areas (OAs from EDs) in terms of number of households

OA	ED
2	129
30	191
50	201
117	254
2	100
53	247
69	199
20	241
90	197
41	188
9	256
5	275
1	143
64	248
40	216
2	217
1	94
5	172
6	136
82	225
113	195
1	274
1	261
16	110
64	237
88	227
1	199
1	247
83	215
3	176
1	223
111	208
42	234

18	186
59	208
80	265
43	182
77	200
18	99
11	123
30	183
35	273

A6.1 The List of 41 variables used in the Area Classifications

DEMOGRAPHIC	
Babies	Age 0-4: Percentage of resident population aged 0-4
Children	Age 5-14: Percentage of resident population aged 5-14
Midage	Age 25-44: Percentage of resident population aged 25-44
Oldage	Age 45-64: Percentage of resident population aged 45-64
Elderly	Age 65+: Percentage of resident population aged 65+
Indpak	Indian, Pakistani or Bangladeshi: Percentage of people identifying as Indian, Pakistani or Bangladeshi
Black	Black African, Black Caribbean or Other Black: Percentage of people identifying as Black African, Black Caribbean or Other Black
BornoutUK	Born Outside UK: Percentage of people not born in the UK
Hden	Population Density: Population Density (number of people per hectare)
HOUSEHOLD COMPOSITION	
Divorced	Separated/Divorced: Percentage of residents 16+ who are not living in a couple and are separated/divorced
Singpens	Single Person Household: Percentage of households which are single pensioner households
Dallpens	*All pensioner households: Percentage of households with only pensioners*
Singpar	Lone Parent Household: Percentage of households which are lone parent households with dependent children
Couplenochild	Two Adults No Children: Percentage of households whichare cohabiting or married couple households with no children
familnondepchild	Households with Non-dependent Children: Percentage of households comprising one family and no others with non-dependent children living with
HOUSING	
Not used	*Rent (Public): Percentage of households that are resident in public sector rented accommodation*
Rentprivate	Rent (Private): Percentage of households that are resident in private/other rented accommodation
Terraced	Terraced Housing: Percentage of all household spaces which are terraced
Detached	Detached Housing: Percentage of all household spaces which are detached

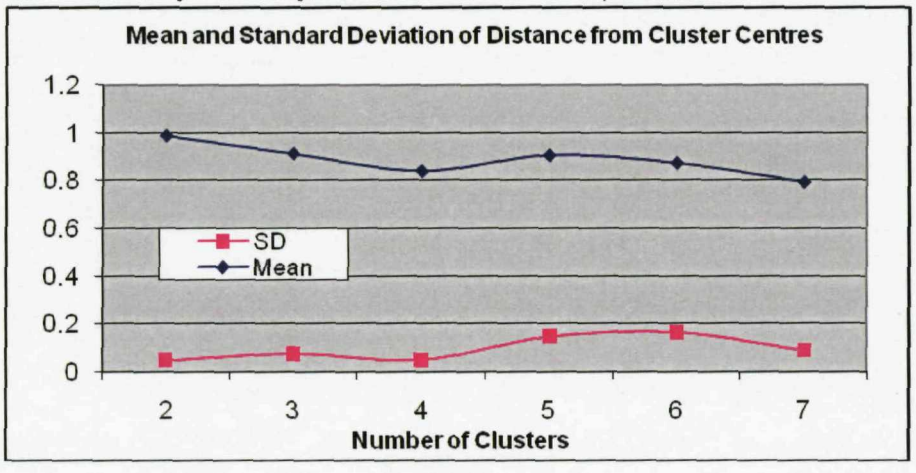
Flat	All Flats: Percentage of household spaces which are flats
Nocenheat	No Central Heating: Percentage of occupied household spaces without central heating
Roomsnum	Average House Size: Average house size (rooms per household)
Persinhh	People per Room: The average number of people per room
SOCIO-ECONOMIC	
Higherqual	HE Qualification: Percentage of people aged between 16-74 with a higher education qualification
Routineocc	Routine/Semi-Routine Occupation: Percentage of people aged 16-74 in employment working in routine or semi-routine occupations
Twocars	2+ Car Household: Percentage of households with 2 or more cars
Publictrans	Public Transport to Work: Percentage of people aged 16-74 in employment
Workshome	Work from Home: Percentage of people aged 16-74 in emplooyment who work mainly from home
SIR	LLTI (SIR): percentage of people who reported suffering from a Limiting Long Term Illness (Standardised Illness Ratio, standardised by age).
(doesn't exist)	*Provide Unpaid Care: Percentage of people who provide unpaid care*
EMPLOYMENT	
Student	Students (full-time): Percentage of people aged 16-74 who are students
Unemployed	Unemployed: Percentage of economically active people aged 16-74 who are unemployed
Parttime	Working Part-time: Percentage of economically active people aged 16-74 who work part time
Lookhome	Economically Inactive Looking after Family: Percentage of economically inactive people aged 16-74 who are looking after the home
(the following have different definitions for the 1991 Census....)	
Agricfish	*Agriculture/Fishing Employment: Percentage of all people aged 16-74 in employment working in agriculture and fishing*
Construc	*Construction Employment: Percentage of all people aged 16-74 in employment working in construction*
Manuf	*Manufacturing Employment: Percentage of all people aged 16-74 in employment working in manufacturing*
Hotelcat	*Hotel and Catering Employment: Percentage of all people aged 16-74 in employment working in hotel and catering*
Services	*Services Employment: Percentgae of all people aged 16-74 in employment working in services*
Financial	*Financial Intermediation Employment: Percentage of all people aged 16-74 in employment working in financial intermediation*
Transport	*Transport Employment: Percentage of all people aged 16-74 in employment working in transport*

A6.2 Statistics for Grouping Wards in the Unperturbed Data

Statistics for Number of Clusters (Grouping Wards in the Unperturbed Data)

Number of clusters	Mean distance to cluster	Standard Deviation
2	0.9884	0.0485
3	0.9136	0.0763
4	0.8417	0.0509
5	0.9075	0.1456
6	0.8748	0.163
7	0.7952	0.0882

Mean and SD of Distance from Cluster Centres in Unperturbed Data



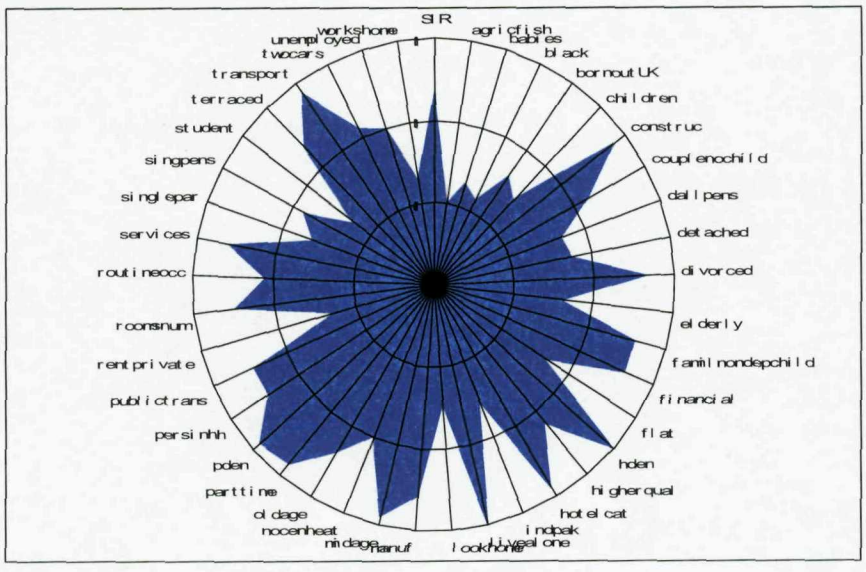
Size of each cluster (Grouping Wards in the Unperturbed Data)

Number of Clusters	Size of each Cluster
2	96 171
3	109 93 65
4	36 75 95 61
5	99 34 10 69 55
6	30 86 53 45 13 40
7	8 106 49 6 24 25 49

The *mean distance to cluster centre* decreases until five clusters are created and moreover at this point, the standard deviation starts increasing. Also taking into consideration the size of each cluster; clusters containing very small numbers of wards may be inaccurate or unreliable. The next stage involved splitting the clusters further to create a sub-classification. The divisions were decided based on the same statistics as above. Clusters with fewer than ten wards were not considered.

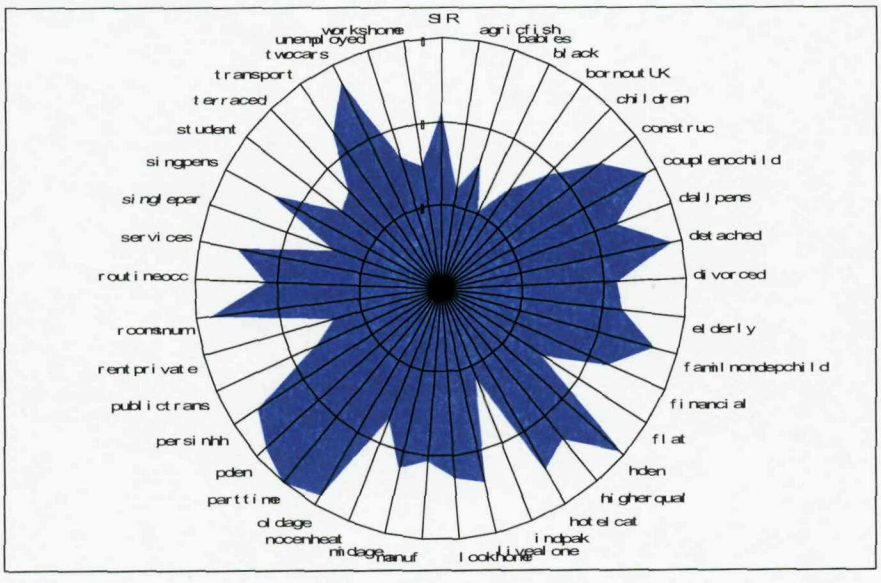
A6.3 Summarising the Clusters (grouping wards in the unperturbed data) – Radial Plots

Cluster 1a: (36 wards)



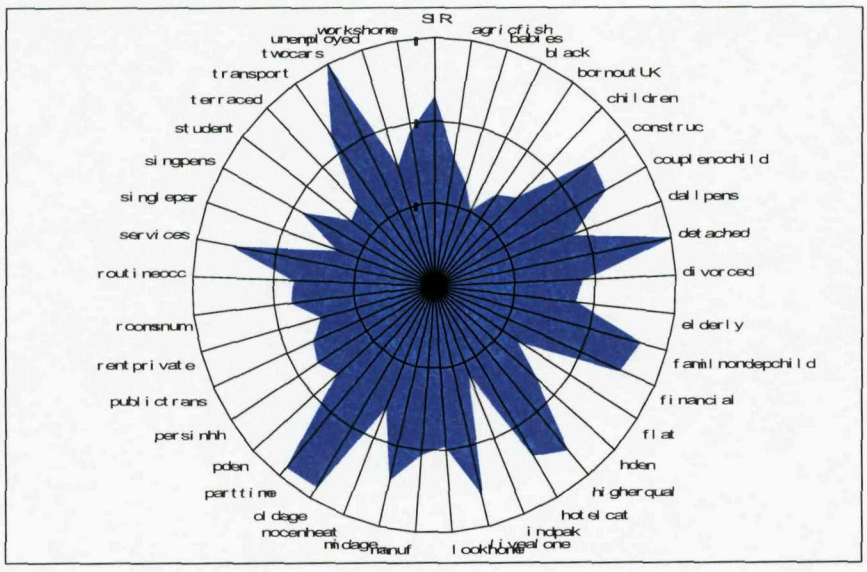
Distinctive Variables
- Urban / High density - People living alone - Many aged 25-44 - Working in service industries - Higher than average ethnic minority groups

Cluster 2a: (31 wards)



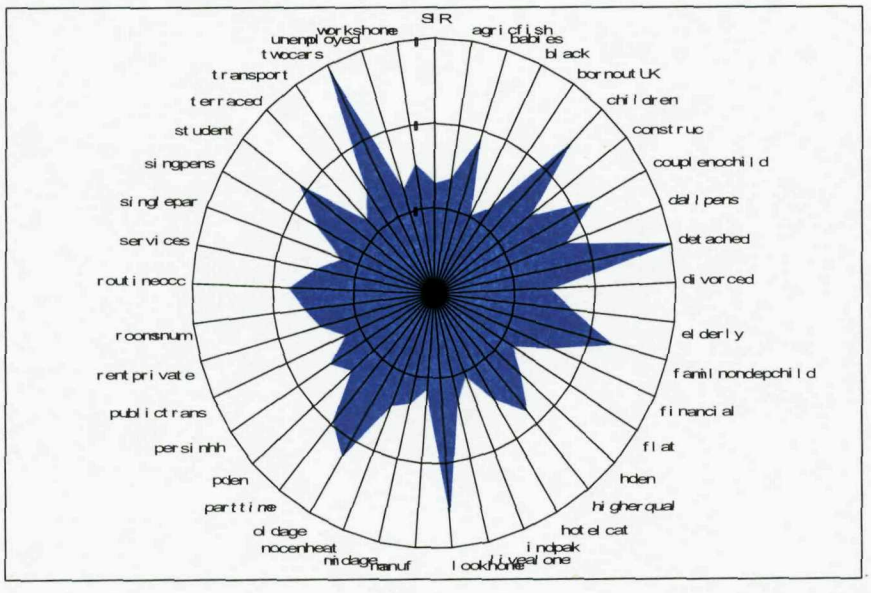
Distinctive Variables
- Two + car households - Detached housing - Urban / high density -Part-time workers - Many aged 45 to 64 -Few flats, single parents, ethnic minority groups, children

Cluster 2b: (44 wards)



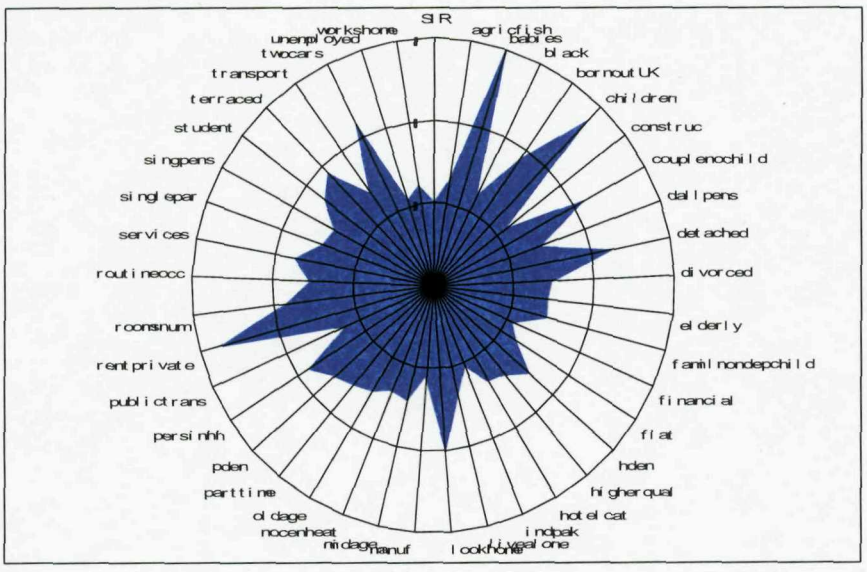
Distinctive Variables
- Two + car households - Detached housing - Rural / Low density - Many aged 45 to 64 - Few flats, single parents, ethnic minority groups, children

Cluster 3a (35 wards)



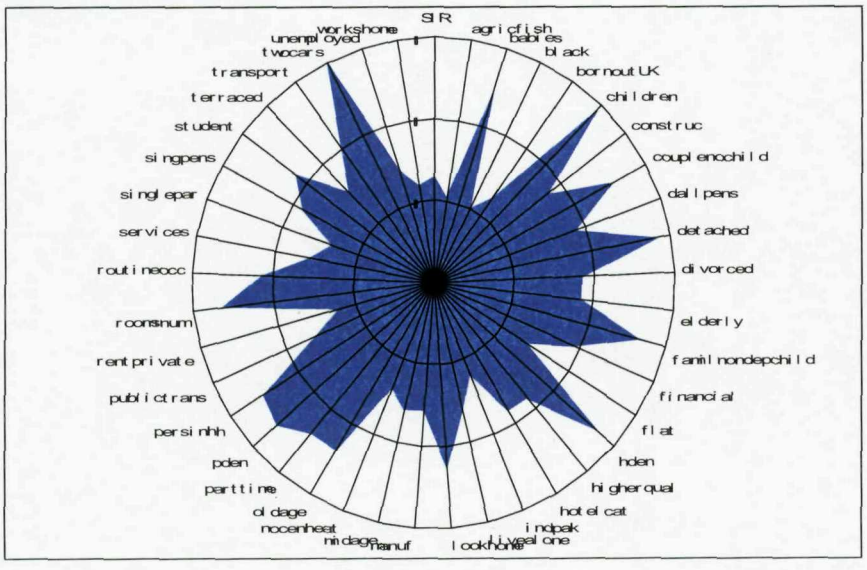
Distinctive Variables
- Two + car households - Detached housing - Looking after home - Many children (5 - 14) - Rural / low density -Low on renting privately, flats, ethnic minority groups, SIR (indicating low LLTI)

Cluster 3b: (7 wards)



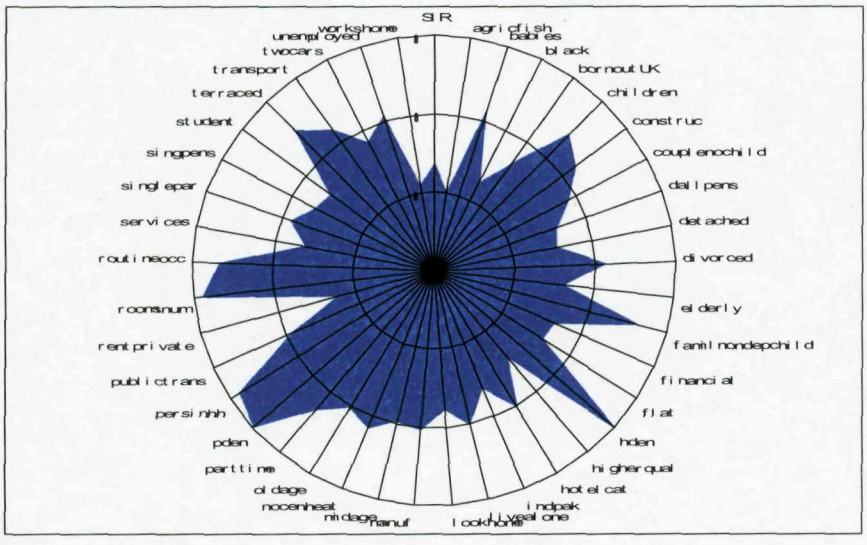
Distinctive Variables
-High proportions of babies and children - High renting privately - Low density / rural -Few flats, two car households, low LLTI (SIR)

Cluster 3c: (53 wards)



Distinctive Variables
-High proportions of babies and children - Detached housing - Suburban - Families with non-dependent children -Few flats, single parents, low LLTI

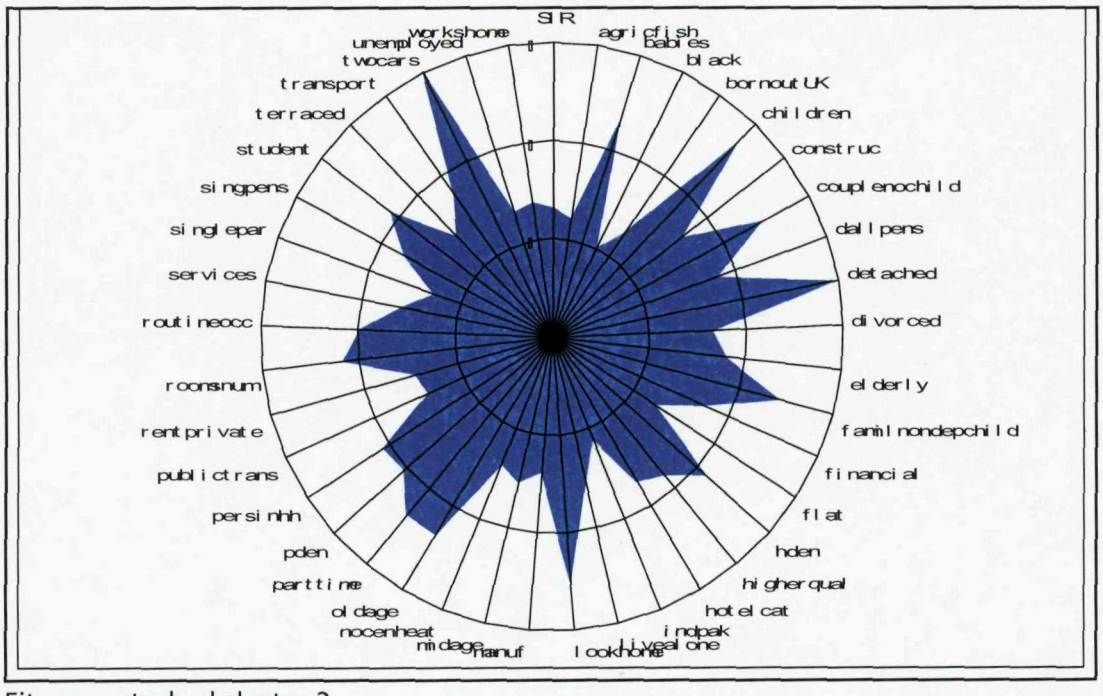
Cluster 4a: (61 wards)



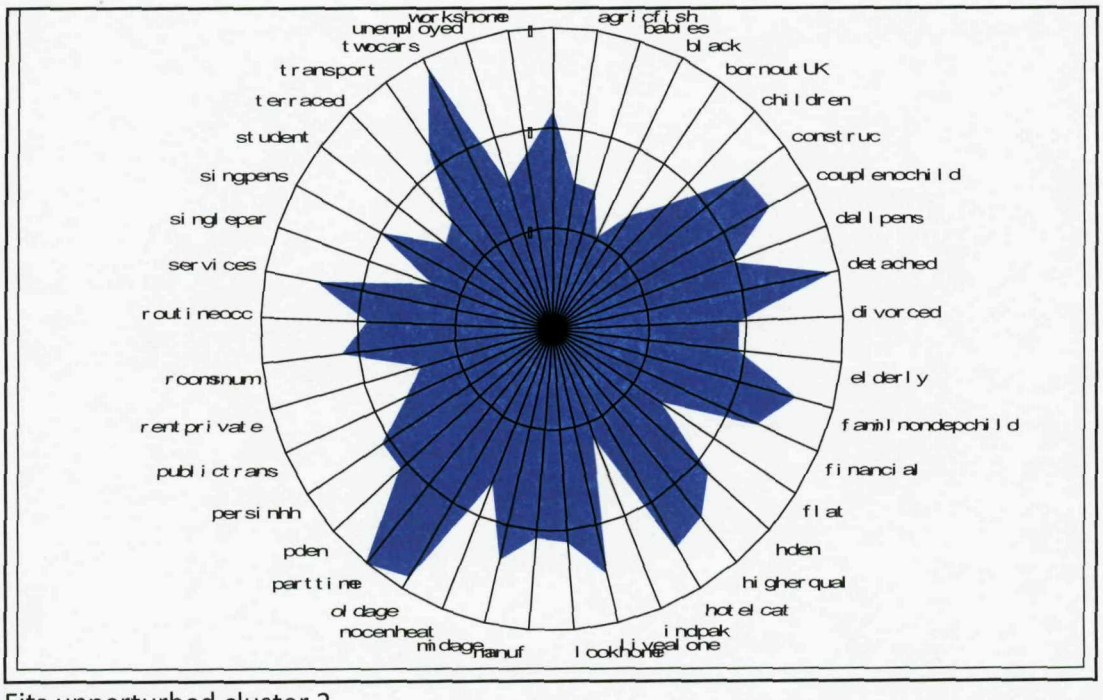
Distinctive Variables
-Urban / high density - Terraced housing - Number of people in household high -All age groups and many children -Low on detached housing, pensioner households, working at home

A6.4 Fit of Swapped Data to Ward Classification

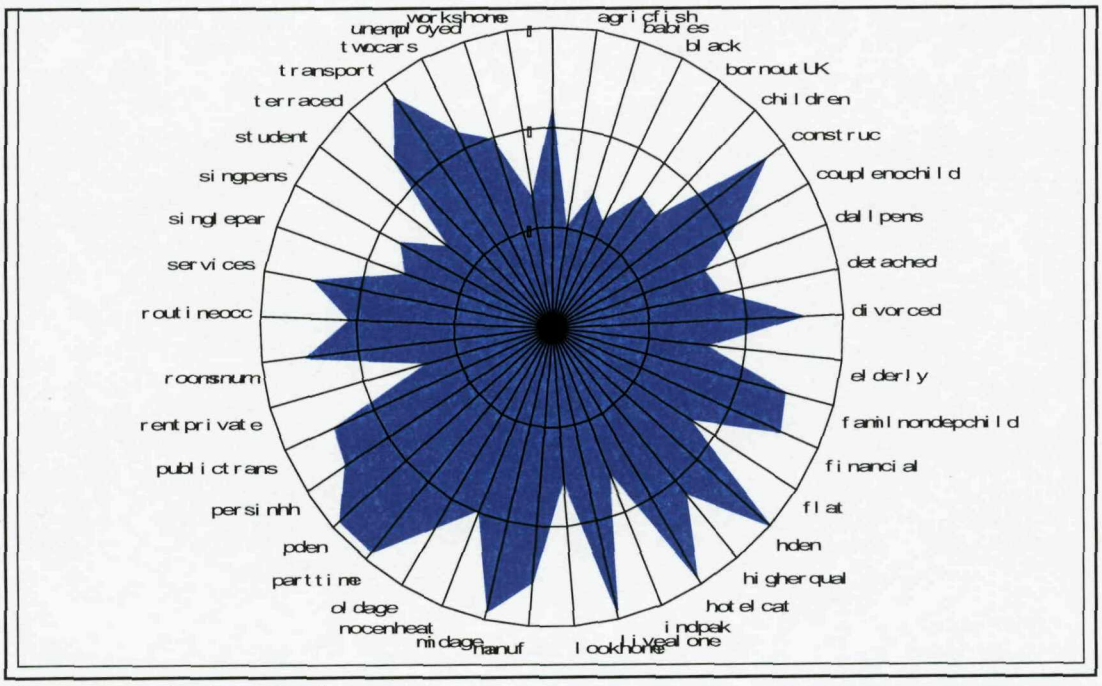
After LDS 10%



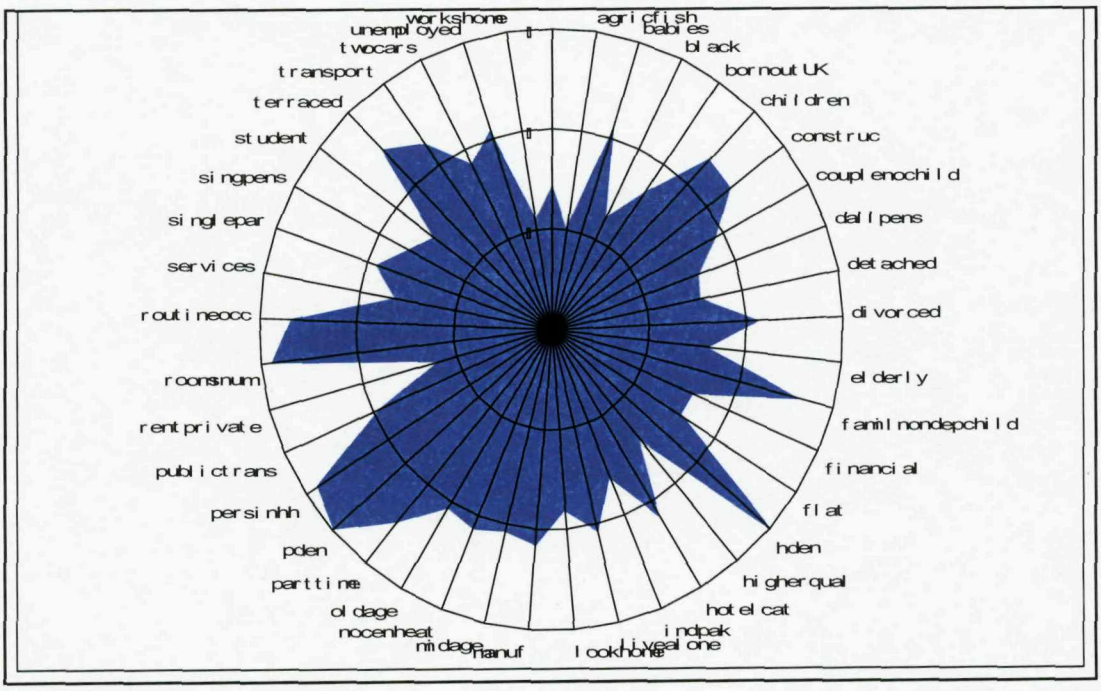
Fits unperturbed cluster 3



Fits unperturbed cluster 2

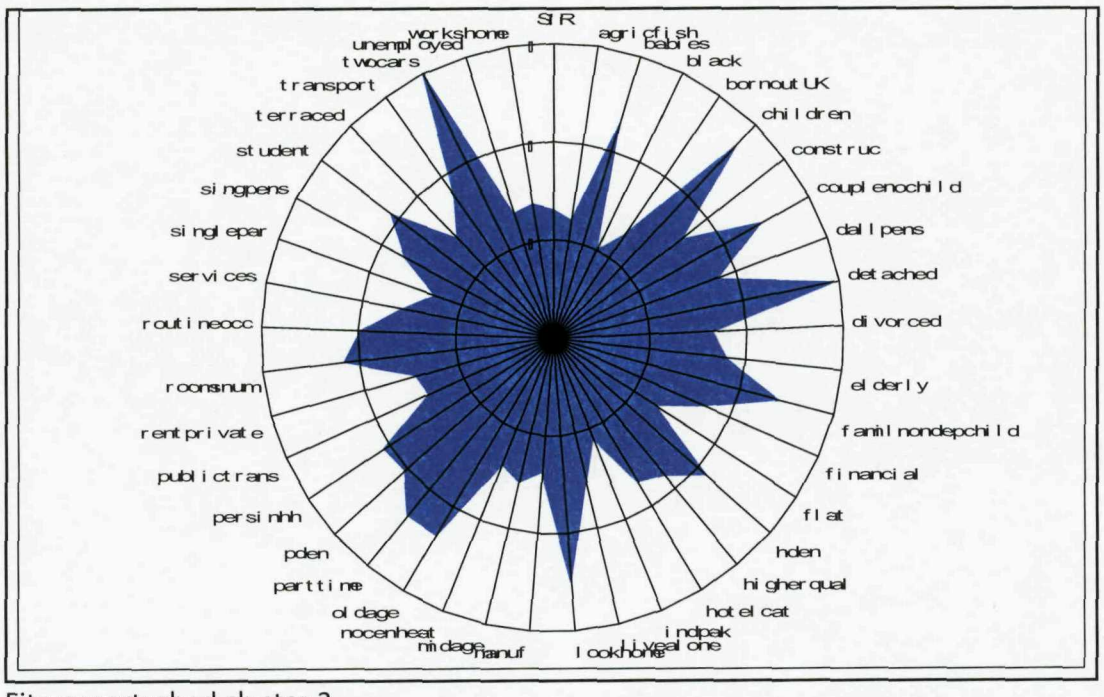


Fits unperturbed cluster 1

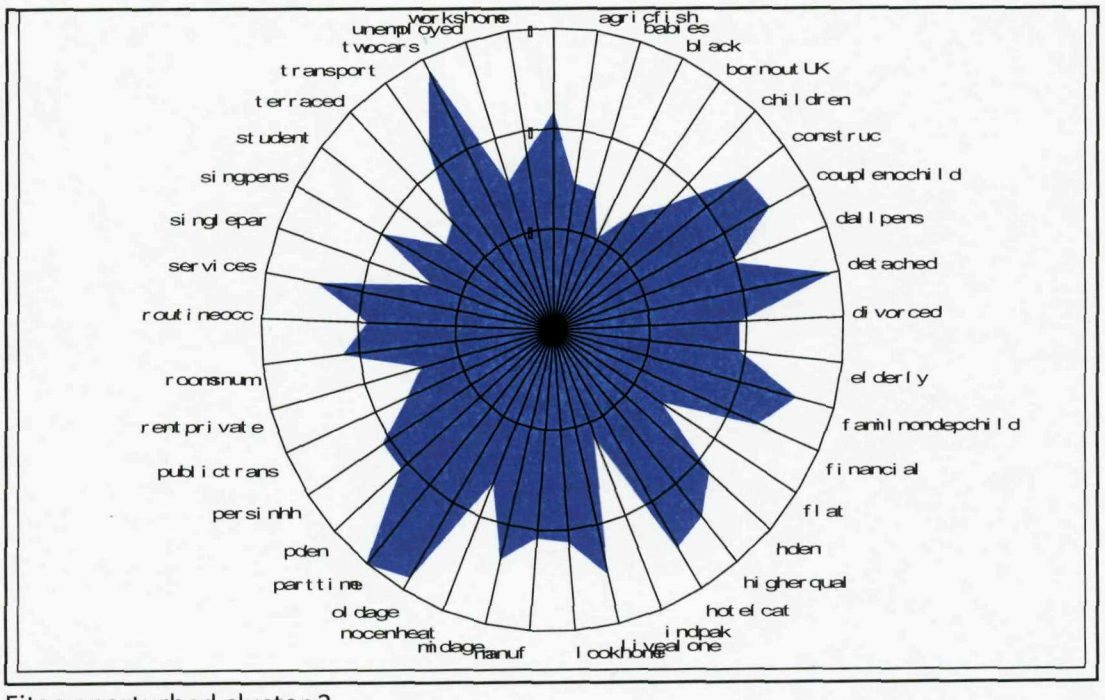


Fits unperturbed cluster 4

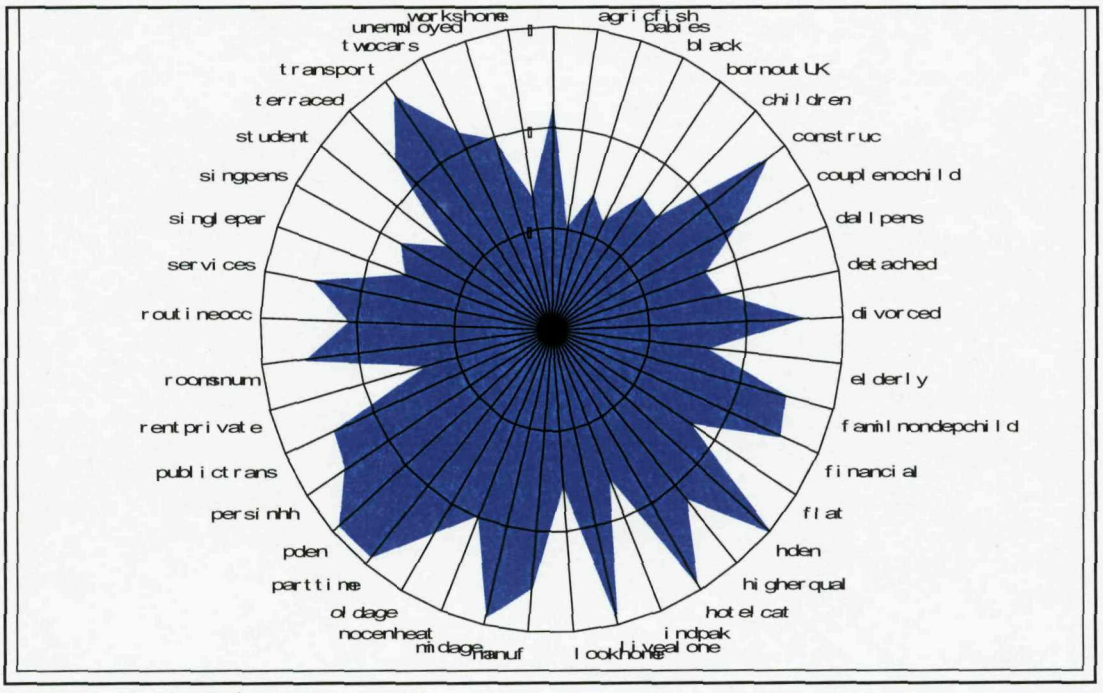
After RRS 10%



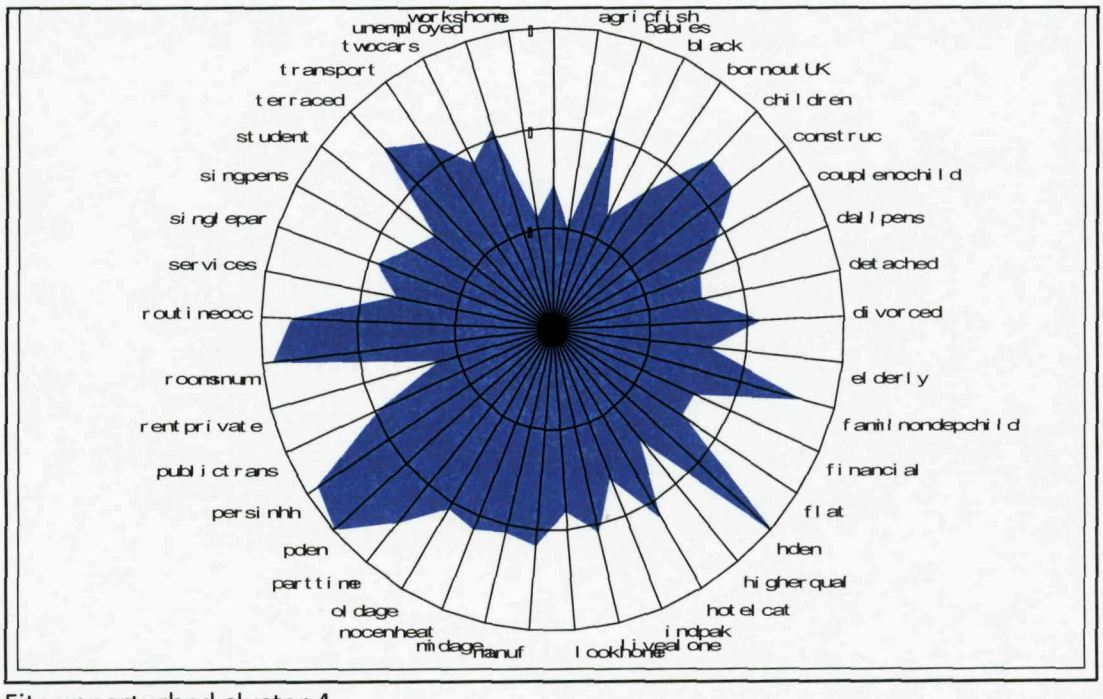
Fits unperturbed cluster 3



Fits unperturbed cluster 2



Fits unperturbed cluster 1



Fits unperturbed cluster 4

Glossary

Cell uniques – cell counts of one in tables

Control/match variables – set of variables on which swapping pairs must take the same value (usually key variables)

Differencing – the comparison of two tabular outputs to produce a previously unpublished table

Donor household – a sampled household for swapping, to be paired with a 'recipient' household

Geographical perturbation – modifying the spatial location of a household unit

Noise – creating uncertainty around household location via geographical perturbation

Original data – data without SDC applied

(Output) zone – generic term to describe a non-specific geography; e.g. wards, output areas, enumeration districts

Perturbation rate – total percentage of records swapped in the population

Protected data – data that has had SDC applied to it

Slivers – small output zones which have been geographically differencing

References

Abowd, J.M. and S. Woodcock (2004) 'Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data' in *Privacy in Statistical Databases,* J.Domingo-Ferrer and V.Torra (eds.) Berlin: Springer-Verlag, pp 290-297.

Algranati, D.J. and J.B. Kadane, (2004) 'Extracting Confidential Information from Public Documents: The 2000 Department of Justice Report on the Federal Use of the Death Penalty in the United States', *Journal of Official Statistics*, 20, pp 97-113.

Anselin, L. (1995) 'Local Indicators of Spatial Association', *Geographical Analysis*, 27 (2), pp 93-115.

Armitage, P. and D. Brown (2003) 'Neighbourhood Statistics in England and Wales: Disclosure Control Problems and Solutions', Proceedings Joint UNECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April).

Armstrong, M.P., G. Rushton, and D.L. Zimmerman (1999) 'Geographically Masking Health Data to Preserve Confidentiality', *Statistics in Medicine,* 18, pp 497-525.

Ballas, D. and G.P. Clarke (2001) 'Modelling the Local Impacts of National Social Policies: A Spatial Microsimulation Approach', *Environment and Planning C: Government and Policy,* 19, pp 587-606.

Ballas, D., G.P. Clarke, and I. Turton (1999) 'Exploring Microsimulation Methodologies for the Estimation of Household Attributes', 4th International Conference on GeoComputation, (Virginia, USA, 25-28 July).

Barnett, S., P. Roderick, D. Martin, I. Diamond, and H. Wrigley (2002) 'The Inter-Relationships between Three Proxies of Health Care Need at the Small Area Level: An Urban-Rural Comparison', *Journal of Epidemiology and Community Health*, 56, pp 754– 761.

Bethlehem, J.G., W.J. Keller, and J. Pannekoek (1990) 'Disclosure Control of Microdata', *Journal of the American Statistical Association,* 85, pp 38-45.

Birkin, M., C. Brunsdon, S. Carver, T. Champion, M. Charlton, G. Clarke, and P. Rees (1995) *Census Users' Handbook,* Cambridge: Pearson Professional Ltd.

Birkin, M. and G. Clarke (1995) 'Using Microsimulation Methods to Synthesize Census Data', in *Census Users' Handbook,* S.Openshaw (ed.) London: GeoInformation International pp 363-387.

Boudreau, J. (2005) Pers.comm. Census Confidentiality Unit, Statistics Canada.

Boyd, M. and P. Stokes (1999) 'Report on the Impact of Sampling Variation on Swapping', Internal Report, Office for National Statistics.

Boyd, M. and P. Vickers (1999) 'Record Swapping - A Possible Disclosure Control Approach for the 2001 UK Census', in Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, working paper 26, (Luxembourg).

Bracken, I. and D. Martin (1989) 'The Generation of Spatial Population Distributions from Census Centroid Data', *Environment and Planning A,* 21,(4) pp 537-543.

Brown, D. (2003) 'Different Approaches to Disclosure Control Problems Associated with Geography', Proceedings of Joint UNECE/Eurostat work session on Statistical Data Confidentiality, working paper 14, (Luxembourg).

Brunsdon, C., A.S. Fotheringham and M.E. Charlton, (1998) 'Geographically Weighted Regression - Modelling Spatial Non-Stationarity', *Journal of the Royal Statistical Society, Series D - The Statistician*, 47 (3), pp 431-443.

Bycroft, C., A. Staggemeier and J.J. Salazar (2005), 'Controlled Rounding Implementation', in Joint UNECE/Eurostat Conference on Statistical Data Confidentiality, (Geneva, November).

Champion, A. (1994) 'Population Change and Migration in Britain since 1981: Evidence for Continuing Deconcentration', *Environmental Planning A,* 26 (10), pp 1501-1520.

Chen, G. and S. Keller-McNulty (1998) 'Estimation of Identification Disclosure Risk in Microdata', *Journal of Official Statistics,* 14 (1), pp 79-95.

Chowdhury, S., G. Duncan, R. Krishnan, S. Roehrig, and S. Mukherjee (1996) 'Disclosure Detection in Multivariate Categorical Databases: Auditing Confidentiality Protection Through Two New Matrix Operators', *Management Science,* 45 (12), pp 1710–1723.

Christie, S. and D. Fone (2003) 'Does Car Ownership Reflect Socio-Economic Disadvantage in Rural Areas? A Cross-Sectional Geographical Study in Wales', *UK Public Health*, 117 (2), pp 112-116.

Clarke, G.P. (1996) *Microsimulation for Urban and Regional Policy Analysis*, London: Pion.

Congdon, P. (1995) 'The Impact of Area Context on Long Term Illness and Premature Mortality: An Illustration of Multi-Level Analysis', *Regional Studies*, 29, pp 327-344.

Corscadden, L. (2005) Pers.comm. Statistics New Zealand.

Cox, L. and G. Sande (1979) 'Techniques for Preserving Statistical Confidentiality', in Proceedings of the 42nd work session of the International Statistical Institute, (Manilla).

Cressie, N. (1991) *Statistics for Spatial Data,* New York: Wiley.

Dalenius, T. (1977) 'Towards a Methodology for Statistical Disclosure Control', *Statistisk Tidskrift* 15, pp 429-444.

Debenham, J., G. Clarke, and J. Stillwell (2003) 'Extending Geodemographic Classification: A New Regional Prototype', *Environment and Planning A,* 35 (6), pp 1025-1050.

DeBerg, M., M. Kreveld, M. Overmars, and O. Schwarzkopf (2000) *Computational Geometry: Algorithms and Applications*, (2nd ed.), New York: Springer-Verlag,

Denning, D. (1980) 'Secure Statistical Databases with Random Sample Queries', *ACM Transactions on Database Systems,* 5 (3), pp 291-315.

DeWolf, P. (2005) Pers.Comm. Statistics Netherlands.

Diggle, P. (2003) *Statistical Analysis of Spatial Point Patterns,* (2nd ed.) New York: Academic Press.

Duke-Williams, O. and P. Rees (1998a) '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, pp 579-605.

Duke-Williams, O. and P.H. Rees, (1998b) 'Factors Affecting Confidentiality Risks Involved in Releasing Census Data for Small Areas', in Domingo-Ferrer, J. (ed.), Statistical Data Protection: Proceedings of the Conference, (Lisbon, 25 to 27 March 1998). Office for Official Publications of the European Communities, Luxembourg. pp 369-379.

Duncan, G., S. Keller-McNulty., and S. Stokes, (2001) 'Disclosure Risk vs. Data Utility: the R-U Confidentiality Map', Technical Report LA-UR-01-6428, Statistical Sciences Group, Los Alamos National Laboratory.

Duncan, G. and D. Lambert (1989) 'The Risk of Disclosure for Microdata', *Journal of Business and Economic Statistics,* 7, pp 207-217.

Elliot, M. (2000) 'DIS: A New Approach to the Measurement of Statistical Disclosure Risk.' *International Journal of Risk Management,* 2 (4), pp 39-48.

Elliot, M. and A. Dale (1999) 'Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk', Netherlands Official Statistics, Special Issue SDC, Vol 14 Spring 1999 *http://www.cbs.nl/NR/rdonlyres/C3E1B07E-1893-4809-9955-50DEA2B9ADA6/0/nos991.pdf* (accessed August 2007).

Everitt, B.S., S. Landau, and M. Leese, (2001) *Cluster Analysis,* (4th ed.) London: Edward Arnold.

FCSM (1994) 'Report on Statistical Disclosure and Disclosure-Avoidance Techniques', Federal Committee on Statistical Methodology. Technical Report Statistical Policy Working Paper 2. *htp://www.fcsm.gov/working-papers/sw2.html.*

Fellegi, I. (1972) 'On the Question of Statistical Confidentiality', *Journal of the American Statistical Association,* 67 (337), pp 7-18.

Fellegi, I. and J. Phillips (1974) 'Statistical Confidentiality: Some Theory and Applications to Data Dissemmination', *Annals of Economic and Social Measurement,* 3 (2), pp 399-409

Fienberg, S. and J. McIntyre, (2004) 'Data Swapping: Variations on a Theme' in *Privacy in Statistical Databases,* Lecture Notes in Computer Science, Dalenius and Reiss (eds.), Berlin: Springer,

Fischetti, M. and J. Salazar (1998) 'Computational Experience with the Controlled Rounding Problem in Statistical Disclosure Control', *Journal of Official Statistics,* 14 (4), pp 553-565.

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

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton, (1998) 'Geographically Weighted Regression: a Natural Evolution of the Expansion Method for Spatial Data Analysis', *Environment and Planning A,* 30 (11), pp 1905-1927.

Fotheringham, A.S., M. Charlton, and C. Brunsdon, (1996) 'The Geography of Parameter Space - An Investigation of Spatial Non-Stationarity', *International Journal of Geographical Information Systems*, 10 (5), pp 605-627.

Fraser, B. and J. Wooton, (2005) 'A Proposed Method for Confidentialising Tabular Output to Protect against Differencing', Joint UNECE/Eurostat conference on Statistical Data Confidentiality, (Geneva).

Gatrell, A., T. Bailey, P. Diggle, and B. Rowlingson (1996) 'Spatial Point Pattern Analysis and its Application in Geographical Epidemiology', *Transactions of the Institute of British Geographers*, 21, pp 256-274.

Giessing, S. (2005) Pers.comm. Statistical Federal Office Germany.

Goldstein, H. (2003), *Multilevel Statistical Models*, London: Arnold, (3rd ed.).

Goldstein, H. and P. Noden, (2003) 'Modelling Social Segregation', *Oxford Review of Education,* 29 (2), pp 225-237.

Gutmann, M.P., and P.C., Stern (2007) *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*, Washington: National Academy Press, National Research Council.

Heady, J. and P. Clarke (2003) 'Model-based Small Area Estimation Series', No.2 Small Area Estimation Report, ONS report, January.

Heldal, J. (2003) 'Rounding as a Confidentiality Measure for Frequency Tables in Stat-bank Norway', in UNECE/ Eurostat work session on Statistical Data Confidentiality, Luxembourg.

Hilbert, D. and S. Cohn-Vossen, (1999) *Geometry and the Imagination*, New York: Chelsea, pp. 33-35.

Hirschfield, A. and K. Bowers (1997) 'The Effect of Social Cohesion on Levels of Recorded Crime in Disadvantaged Areas', *Urban Studies,* 34 (8), pp 1275-1295.

Johnston, RJ, C. Propper, S.M. Burgess, R. Sarker, A.A. Bolster, and K. Jones, (2005) 'Spatial Scale and the Neighbourhood Effect: Multinomial Models of Voting at Two Recent British General Elections', *British Journal of Political Science,* 35 (3), pp 487-514.

Karr, A. (2006) 'Combining Statistical Disclosure Methods for Microdata', Technical report, National Institute of Statistical Sciences. *www.niss.org/downloadabletechreports.html.*

Keller-McNulty, S. and E. Unger (1993) 'Database Systems: Inferential Security', *Journal of Official Statistics,* 9, pp 475-499.

Kwan, M., I. Casas, and B. Schmitz (2004) 'Protection of GeoPrivacy and Accuracy of Spatial Information: How Effective are Geographical Masks?', *Cartographica*, 39 (2), pp 15-28.

Lambert, D. (1993) 'Measures of Disclosure Risk and Harm', *Journal of Official Statistics,* 9, pp 313-331.

Leitner, M. and A. Curtis, (2006) 'A First Step Towards a Framework for Presenting the Location of Confidential Point Data on Maps – Results of an Empirical Perceptual Study', *International Journal of Geographical Information Science*, 20 (7), pp 813 - 822 .

Leyland, A. H. and Goldstein, H. (2001) *Multilevel Modelling of Health Statistics, Leyland and Goldstein (eds.)* ' (eds.) London: Wiley,.

March, M. and D. Norris (1987) 'Disclosure Avoidance Techniques in the Canadian Censuses of Population and Agriculture', in Proceedings of the Survey Research Methods Section, ASA.

Markkula, J. (2003) 'Geographic Personal Data, its Privacy Protection and Prospects in a Location-Based Service Environment', *Jyvaskyla Studies in Computing,* 30, pp 109

Martin, D. (2000) 'Towards the Geographies of the 2001 UK Census of Population', *Transactions of the Institute of British Geographers,* 25, pp 321-332.

Martin, D. (1998) 'Optimizing Census Geography: The Separation of Collection and Output Geographies', *International Journal of Geographical Information Science,* 12, pp 673-685.

Martin, D. (1989) 'Mapping Population Data from Zone Centroid Locations', *Transactions of the Institute of British Geographers,* 14(1), pp 90-97.

Mateos, P. (2007) 'A Review of Name-Based Ethnicity Classification Methods and their Potential in Population Studies', *Population Space and Place,* 13 (4), pp 243-263.

McElroy, L. (2003) 'Privacy Impact Assessments and Confidentiality Protection Techniques for Statistical Data Dissemination', in 17th International Roundtable on Business Survey Frames, (Italy).

Mertz, J. (1991) 'Microsimulation - A Survey of Principles, Developments and Applications', *International Journal of Forecasting,* 7, pp 77-104.

Mesev, V. (1998) 'The Use of Census Data in Urban Image Classification', *Photogrammetric Engineering and Remote Sensing,* 64, pp 431-438.

Moon, G., S.V. Subramanian, K. Jones, C. Duncan, and L. Twigg, (2005) 'Area-Based Studies and the Evaluation of Multilevel Influences on Health Outcomes', in Ann Bowling and Shah Ebrahim (Eds.), *Handbook of Health Research Methods*, Open University Press, pp. 266-292,

Moran, P.A.P. (1950) 'Notes on Continuous Stochastic Phenomena', *Biometrika,* 37, pp17-23.

Morphet, C. (1993) 'The Mapping of Small Area Census Data - A Consideration of the Role of Enumeration District Boundaries', *Environment and Planning A,* 25 (9), pp 1267-1277.

Munnich, R. and S. Josef (2003) 'On the Simulation of Complex Universes in the Case of Applying the German Microcensus', DACSEIS research paper series 4.

ONS (2006, Accessed August 2007). Review of the Dissemination of Health Statistics: Confidentiality Guidance. Office for National Statistics. *http://www.statistics.gov.uk/about/Consultations/downloads/Health_Stats/Health_Stats_Report.pdf.*

ONS (2004a, Accessed January 2005). 1991 Census Datasets. Office for National Statistics *http://census.ac.uk/cdu/Datasets/1991CEnsusdatasets/.*

ONS (2004b, Accessed January 2005). Disclosure Protection Census 2001. Office for National Statistics *http://www.statistics.gov.uk/census2001/discloseprotect.asp.*

ONS (2002, Accessed February 2005). Standard Area Statistics. Office for National Statistics *www.statistics.gov.uk/census2001/op11.asp.*

ONS (2001, Accessed April 2005). Census 2001 Review and Evaluation: Edit and Imputation. Office for National Statistics *http://www.statistics.gov.uk/census2001-/pdfs/editandimpexecsumm.pdf.*

Openshaw, S. and C. Wymer (1995) 'Classifying and Regionalizing Census Data' in *Census Users' Handbook*.

Orcutt, G. (1957) 'A New Type of Socio-Economic System', *Review of Economics and Statistics*, 58, pp 773-797.

Paass, G. (1988) 'Disclosure Risk and Disclosure Avoidance for Microdata', *Journal of Business and Economic Statistics,* 6 (4), pp 487-500.

Rasbash, J., F. Steele, W. Browne and B. Prosser, (2004) 'A User's Guide to MLWIN', version 2.0, Centre for Multilevel Modelling, Institute of Education. (Downloaded with MLWIN).

Rao, J. N. K. (2003) *Small Area Estimation*, New York: Wiley-Interscience.

Reijneveld, S. (1998) 'The Impact of Individual and Area Characteristics on Urban Socio-Economic Differences in Health and Smoking', *International Journal of Epidemiology,* 27 (1), pp 33-40.

Rhind, D., K. Cole, M. Armstrong, L. Chow, and S. Openshaw (1991) 'An On-line, Secure and Infinitely Flexible Database System for the National Census of Population' South East Regional Research Laboratory, Birkbeck College, University of London (working paper 14).

Robertson, D. (1993) 'Cell Suppression at Statistics Canada', in Proceedings of the 1993 Annual Research Conference, Bureau of the Census, pp. 107-131.

Schulte Nordholt, E. (2001) 'Statistical Disclosure Control (SDC) in Practice: Some Examples of Official Statistics of Statistics Netherlands', *Statistical Journal of the UNECE*, 18 (4), pp. 321-328.

Shlomo, N., (2006) 'Statistical Disclosure Control Methods for Census Frequency Tables', Southampton Statistical Sciences Research Institute, Methodology working paper M06/03.

Shlomo, N. (2005a) 'Assessment of Statistical Disclosure Control Methods for the 2001 UK Census', Joint ECE/Eurostat work session on statistical data confidentiality, working paper 19, Geneva.

Shlomo, N. (2005b), 'Statistical Disclosure Control for Census Outputs' in BSPS Day Meeting on Disclosure Control, LSE, September.

Shlomo, N. and C. Young (2006a) 'Quality Measures for Disclosure Controlled Statistical Data', in Q2006 European Conference on Quality in Survey Statistics (Cardiff).

Shlomo. N, and C. Young, (2006b) 'Statistical Disclosure Control Methods Through a Risk-Utility Framework' in *Privacy in Statistical Databases,* pp. 68-81.

Simonoff, J. (1996) *Smoothing Methods in Statistics*, Springer-Verlag.

Skinner, C., C. Marsh, S. Openshaw, and C. Wymer (1994) 'Disclosure Control for Census Microdata', *Journal of Official Statistics,* 10, pp 31-51.

Steel, P. and L. Zayatz (2003) 'The Effects of the Disclosure Limitation Procedure on Census 2000 Tabular Data Products (abridged)', Technical Report, US Census Bureau.

Steele, F., J. Brown, and R. Chambers, (2002) 'A Controlled Donor Imputation System for a One-Number Census', *Journal of the Royal Statistical Society*, Series A, 165, part 3, pp 495-522.

Stinchcomb, D. (2004) 'Comprehensive GIS Application for West Nile Virus Surveillance', 2004 Health GIS Conference Proceedings, Available online at *http://gis.esri.com/library/userconf/health04/index.html.*

Tammilehto-Luode, M. (2001) 'Disclosure Control for Demographic Statistics', Redefined Guidelines and Development of Methods at Statistics Finland. In ECE/Eurostat work session on Statistical Data Confidentiality (Skopje), working paper 14.

Tickle, M., E. Kay, H. Worthington, and A. Blinkhorn (2000) 'Predicting Population Dental Disease Experience at a Small Area Level using Census and Health Service Data', *Journal of Public Health Medicine,* 3, pp 368-374.

Tobler, W. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography,* 46 (3) pp234-240.

Tranmer, M. and D. Steel (1998) 'Using Census Data to Investigate the Causes of The Ecological Fallacy', *Environment and Planning A,* 30 (5), pp 817-831.

VanWey, L.K., R.R. Rindfuss, M.P. Gutmann, B. Entwisle and D.L. Balk (2005) 'Confidentiality and Spatially Explicit Data: Concerns and *Challenges', in* Proceedings of the National Academy of Sciences of the United States of America: Spatial Demography Special Feature, *102 (43*) Available online at: *http://www.pnas.org/cgi/reprint/102/43/15337.*

Vickers, D. and P.Rees, (2007) 'Creating the National Statistics 2001 Output Area Classification', *Journal of the Royal Statistical Society Series A*, 170 (2), pp 379 – 403.

Voas, D. and P. Williamson (2001) 'The Diversity of Diversity: A Critique of Geodemographic Classification', *AREA,* 33 (1), pp 63-76.

Voas, D. and P. Williamson (2000). 'An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata', *International Journal of Population Geography,* 6 pp 349-366.

Wallace M, J. Charlton and C. Denham (1995) 'The New OPCS Area Classification' *Population Trends,* 79, pp 15-30.

Wallace, M. and C. Denham (1996) 'The ONS Classification of Local and Health Authorities of Great Britain', in Studies on Medical and Population Subjects, ONS. Number 59.

Walters, K., E. Breeze, P. Wilkinson, G. Price, C. Bulpitt, and A. Fletcher (2004) 'Local Area Deprivation and Urban-Rural Differences in Anxiety and Depression among People Older than 75 years in Britain', *American Journal of Public Health,* 94 (10), pp 1768-1774.

Wand, M. (1994) 'Fast Computation of Multivariate Kernel Estimators', *Journal of Computational and Graphical Statistics,* (3), pp 433-445.

Willenborg, L. and T. De Waal (2001) *Elements of Statistical Disclosure Control* Lecture Notes in Statistics, Volume 155. New York: Springer-Verlag.

Willenborg, L. and T. De Waal (1996) *Statistical Disclosure Control in Practice,* Number 111 in Lecture Notes in Statistics. New York: Springer-Verlag.

Williamson, P., M. Birkin, and P. Rees (1998) 'The Estimation of Population Microdata by using Data from Small Area Statistics and Samples of Anoymised Records,' *Environment & Planning A,* 30 (5), pp785-816.

Wolf, D. (2001) 'The Role of Microsimulation in Longitudinal Data Analysis', Papers in the Microsimulation Series, Centre for Policy Research, New York.

Wu, F. and D. Martin (2002) 'Urban Expansion and Simulation of Southeast England using Population Surface Modelling and Cellular Automata', *Environment and Planning A,* 34, (10), pp 1855-1876.

Zayatz, L. (2006) 'Disclosure Avoidance Practices and Research at the US Census Bureau: An Update', US Census Bureau. Downloadable from *http://www.census.gov/srd/www/byyear.html.*

Zayatz, L. (2003) 'Disclosure Limitation for Census 2000 Tabular Data' in Joint UNECE/Eurostat work session on statistical data confidentiality working paper 15, (Luxembourg 7-9 April 2003).

Zimmerman, D.L. and C. Pavlik, (2006), 'Quantifying the Effects of Mask Metadata Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data', Technical Report 369, University of Iowa, Department of Statistics and Actuarial Science; Available online at: *http://www.stat.uiowa.edu/techrep/tr369.pdf.*

***General Web References**

ACORN (*http://www.caci.co.uk/acorn/)*

NeSS (http://neighbourhood.statistics.gov.uk)

Ordnance Survey (http://www.ordnancesurvey.co.uk/oswebsite/products/addresspoint/)

Census Geography Look-up Directories
http://census.ac.uk/cdu/datasets/lookup_tables/postal/postcode_enumeration_district_directory.htm

One Number Census (*http://www.statistics.gov.uk/census2001/editimputevrep.asp)*

Commercial websites using census data:
www.upmystreet.com

www.checkmyfile.com/Guest/NeighbourhoodSearch.asp
www.neighbourhoodstatistics.gov.uk