WORKING PAPER 97/1

THE DEBATE ABOUT THE GEOGRAPHY OF THE 2001 CENSUS: Collected Papers from 1995-6

Edited by Philip Rees

Background papers for the
First ESRC/JISC Workshop Planning for the 2001 Census on
"The 2001 Census – What Geography Do We Want?"
22/23 January 1997
School of Geography, University of Leeds, Leeds, UK

PUBLISHED JANUARY 1997

For further copies contact the Working Paper Secretary,
School of Geography, University of Leeds, Leeds, LS2 9JT
Telephone 0113 233 3300

Views expressed in Working Papers are
those of the author(s) and not necessarily
those of the School of Geography.

CONTENTS

ABSTRACT

This collection of papers reports on the discussions that have been held during 1996 between the Census Offices and Census Users about the geographical frameworks to be adopted for the 2001 Census. The papers were originally presented at meetings during the year between Census Office staff and the staff of user organisations, including universities. The papers have been collected together as background briefing for participants in the First in the series of Workshops Planning for the 2001 Census - Determining Academic Community Needs and Strategy, funded by the Economic and Social Research Council and the Joint Information System Committee (Award H507265031).

CONTRIBUTORS

Oliver Duke-Williams
School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom
Tel +44 (0)113 233 3309, Fax +44 (0)113 233 3309, E-mail oliver@geog.leeds.ac.uk

David Martin
Department of Geography, University of Southampton, Highfield, Southampton SO17 1BJ
Tel +44 (0)1703 592215, Fax +44(0)113 593295, E-mail david.martin@soton.ac.uk

Phil Rees
School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom
Tel +44 (0)113 233 3341, Fax +44 (0)113 233 6757, E-mail phil@geog.leeds.ac.uk

Frank Thomas
General Register Office (Scotland), Ladywell House, Edinburgh EH12 7TF
Tel +44 (0)131 314 4217, GTN 7166 217, Fax +44 (0)131 314 4344, GTN 7166 344

ACKNOWLEDGEMENTS

Preparation of the items included in this collection was supported by awards from the European Union under the Statistical Disclosure Control programme and from ESRC for Co-ordination of the ESRC/JISC 1991 Census of Population (Award H507345001). The Census Offices (ONS, GROS and NISRA) have given permission to reproduce various Output Working Group papers.

1. THE DEBATE ABOUT CENSUS GEOGRAPHY: OVERVIEW AND ISSUES

Phil Rees, January 1997

This collection of papers reports on the discussions that have been held during 1996 between the Census Offices and Census Users about the geographical frameworks to be adopted for the 2001 Census. The papers were originally presented at meetings during the year between Census Office staff and the staff of user organisations, including universities. The papers have been collected together as background briefing for participants in the first in the series of *Workshops Planning for the 2001 Census - Determining Academic Community Needs and Strategy*, funded by the Economic and Social Research Council and the Joint Information System Committee (Award H507265031). Some editing and pruning of the papers has been carried out where material repeats earlier arguments or is peripheral, and to achieve a uniform style. The references have been consolidated into one list. Some editorial comments are inserted at the start of papers to indicate the context in which each was produced.

The second paper, *The Geography of Outputs from the 2001 Census*, was designed to kick start the process of deciding and designing the geography to be used for publishing area statistics from the 2001 Census. It tries to summarise the opinions of census users expressed in a series of debriefing meetings which the Census Offices had conducted in 1994-95, picking out the improvements which were felt to be necessary. It also sets out some of the problems that would need to be solved if users' full wish list were to be realised. Paper 2 also introduces a set of definitions of terms used in the debate about census geography, Three of the terms - multiple geographies, flexible geographies and building blocks -have been interpreted in different ways during the debate. Looking back over the debate, my understanding of the terms is as follows. *Building blocks* are geographical units used to build areas for output of statistics, for which no statistics are published. *Multiple geographies* are a set of standard geographical areas for which statistics aree published and released for general use. *Flexible geographies* are a set of additional geographical areas, defined by user or user communities, for which statistics can be requested but the statistics will be subject to additional protection or vetting before release.

The third paper, *Flexible geographies and area aggregation: designing small areas for outputs from the 2001 Census*, refines the arguments of the previous paper and adds a careful consideration of the basic spatial units for aggregation. It suggests that, if flexible aggregation from a residential address base incorporating exact grid references (geocodes) is not possible because of cost or contractual problems, the unit postcode has very many advantages as the basic spatial unit for aggregation to census output areas.

The fourth paper, *The Geography of Outputs from the 2001 Census*, is the Census Office responds to the arguments of census-users presented in papers 2 and 3. Most of the principles proposed in Paper 2 are accepted but some are shown to be in conflict. In particular, the *confidentiality principle* conflicts with the *principle of matching geographies to user needs*. The paper demonstrates how differencing of two sets of small area statistics can lead, in theory, to the extraction by users of SAS for areas with populations (of individuals or households or both) below thresholds set to protect confidentiality. The proposed solution is to stick with one set of nested areas, as in past censuses, but to use the unit postcode as the building block, providing varying levels of statistical detail dependency on spatial scale.

The fifth paper, *The Geography of Outputs from the 1991 Census: Output Units and Look Up Tables*, moves the debate on in two respects. It considers the task of designing output areas for the 2001 Census, accepting the principle of separation of input and output areas (principle 4 in Paper 2) and the principle of common geographies (principle 6 in Paper 2) and using the unit postcode building block and hierarchical system proposed in Paper 4. Martin has worked in detail on this design task and this is reported in his paper "Implementing an automated census output geography", presented to the First ESRC/JISC Workshop Planning for the 2001 Census. The paper also considers in more detail the alternatives for matching census geographies over time and argues for a systematic programme of look up table creation based on the (changing) unit postcode.

The sixth paper, *The Flexible Geography Task of the Statistical Disclosure Project: Progress, Thinking and Issues*, is a report on work being carried out at the University of Leeds as part of an EU programme on Statistical Disclosure Control. The Leeds work attempts to quantify the extent of the differencing problem faced when more than one set of area statistics is published. The approach is to construct a synthetic household and person database for the Yorkshire and Humberside in which the households are precisely located by 1991 Census enumeration district and an OS grid reference. It is then possible to overlay any output geography on this database and construct alternative sets of area statistics. The system is then used to search for differenced areas and report the person and household numbers in such areas. The paper also develops ideas about the output area design process and about look up tables further.

The seventh paper, *Issues Concerning the Flexible Geography Task of the Statistical Disclosure Control Project* reports on discussions between the Leeds and ONS SDC project teams, measuring the disclosure risk for tables.

The eighth paper, *A Note on Look Up Tables*, proposes a general definition for look up tables and a first list that might be produced in connection with the 2001 Census.

Of course, the debate about census geography had been going on for many years before 1996, in fact ever since machine readable statistics for small areas had been made available from the 1971 Census by the Office of Population Censuses and Surveys. Prior to the 1991 Census, a multidisciplinary group organised by Neil Wrigley compiled the views of the academic community on the nature of the forthcoming census and many of their recommendations concerned geography (Marsh *et al.* 1988). After the 1991 Census a number of reviews of the geography have appeared. Chris Denham, the main architect of the geographical schemes for collection and output in England and Wales, has provided a comprehensive account of the development of census geography from 1961 and a guide to the organisation of geographical outputs from the 1991 Census (Denham 1993). Barr (1993) provides a review of the schemes used and a comparison with United States practice. Both these pieces appear in the *Census User's Guide* of Dale and Marsh (1993), which is essential reading for any serious census user. In the later *Census Users' Handbook* (Openshaw 1995), Mike Coombes provides a systematic account of what the purposes and uses of census geography are (Coombes 1995). He also proposes a set of seven tests as a user's benchmark against which to evaluate any approach to the geography of census output. It is useful to reproduce these here and for readers to keep them in mind as they digest the proposals and counter-proposals of the debate (Coombes 1995, p121).

1. *Are the building blocks the smallest that the confidentiality restrictions deem to be possible (to allow maximum flexibility of aggregation)?*
2. *Is each of the areas in a set of building blocks, or other set of areas, defined on a consistent basis across the country?*
3. *Does this set of areas represent (parts of) 'real world' entities, such as settlements, which can be recognised using these boundaries?*
4. *Does the set of areas allow comparison with previous census(es) at this level, or for some minimal grouping of areas to create consistent boundaries?*
5. *Does the set of areas cover the whole of the country without leaving any locations whose data are too sparse to allow them to be published?*
6. *Are the boundaries of these areas available in digital form?*
7. *Can these areas be readily and accurately linked by their location coding to the areas used in many non-census datasets?*

The aim of the Workshop held on the 22/23 January is to develop recommendations for developing the small area geographies to be used in the 2001 Census. There are good prospects that most of Coombes' questions will be answered in the affirmative.

2. THE GEOGRAPHY OF OUTPUTS FROM THE 2001 CENSUS: A FRAMEWORK

Phil Rees, November 1995

This paper was prepared for the meeting of the Working Group on Output, Geography and Confidentiality (OWG) held on Friday 24th November 1995, Room 136B, Department of Health, Skipton House, 80 London Road, Elephant & Castle, London. It sets out a framework for planning the outputs of counts and tables for geographical areas from the 2001 Census. A set of principles is proposed which have been implicitly accepted by both suppliers and users in discussion during 1994 and 1995. These principles then lead to a set of alternative proposals which need conceptual development, retrospective empirical evaluation using 1991 Census data and future testing in preparation for the 2001 Census.

2.1 Aims and definitions

This paper aims to set out a framework for the debate about the small area geography to be adopted for the next Census of Population in the United Kingdom in April/May 2001. It does not propose any particular solution but tries to clarify what the debate should be about, what alternatives need to be explored and the sets of benefits, costs and risks associated with each of the alternatives. The paper seeks to contribute to a co-operative dialogue between suppliers (Census Offices) and users (academics, local government, central government, business, libraries, the public). However, it represents the official views of neither the Census Offices nor the Economic and Social Research Council. It is envisaged that the framework should be co-operatively revised by the Census Offices and Customer Sectors over the next year (1995-96) to become the framework for detailed output planning.

Some definitions are needed at the outset in order to clarify the debate.

By **geographies** we mean alternative systems of aggregation of individual and household data from the census to areas whose spatial extent is known, either as collections of point references or as the territory within known boundaries.

By **small area** is meant areas that smaller than the units used for local government (though they need not necessarily add exactly to them).

By **multiple geographies** is meant the existence of more than one set of small area boundaries for which small area statistics are published.

By **flexible geographies** is meant the ability of users to define their own geographical systems for which they can produce their own sets of statistical tables.

By **small area statistics** is meant a suite of different sets of fixed tables of census counts. These SAS may be very simple for small areas of minimum size but quite detailed for the largest areas. This extends the distinction between small area statistics and local base statistics developed very usefully in the outputs from the 1991 Census of Population for Great Britain.

By **addresses** is meant the description of the location of residential accommodation (whether for private households or in communal establishments) used by the postal service (Royal Mail).

By **co-ordinates** is meant the system of rectangular co-ordinates forming the National Grids of Great Britain and of Ireland used by the Ordnance Surveys of Great Britain and Northern Ireland respectively.

The agencies which have responsibility for the census in the United Kingdom are as follows.

(1) The new **Office for National Statistics**, takes responsibility for the Census in England and Wales and co-ordinates the Census in the UK., which was formed by a merger in 1996 between the Central Statistical Office and the Office of Population Censuses and Surveys (OPCS).

(2) The **General Register Office Scotland** (GROS) takes responsibility for the Census in Scotland.

(3) The **Northern Ireland Statistics and Research Agency** takes responsibility for the Census in Northern Ireland.

The customers who use the outputs of the Census are currently divided into the following sectors.

(i) **Central Government departments**
(ii) **Local Government department**
(iii) **Businesses/marketing agencies** and
(iv) **Academic community/Higher Education**

In future, some formal recognition might be given to
(v) **Libraries**
(vi) **Schools/Secondary Education** and
(vii) **The public**

2.2 Principles

The following **principles** have come to be accepted in the Census supplier and user communities as a result of experiences and innovations introduced in the 1971, 1981 and 1991 Censuses, and following on from the critical evaluation of the strengths and weaknesses of the approaches adopted. Each principle is spelt out and then comments are made on the precedents for each principle in past UK census practice.

(1) The confidentiality principle
The outputs associated with each geography should not compromise the confidentiality of the underlying individual and household records.

(2) The data protection principle
There will need to be protection measures applied to counts for small **areas** that will ensure that confidentiality is protected.

(3) The trade-off principle
There will need to be a trade-off between the size of area and detail of output provided: for the smallest areas fewer counts can be provided; for larger areas more counts can be output.

(4) The principle of separation of input and output areas
The geography of input areas (for collection of the census) and the geography of output areas (for publication of the census) should be regarded as separate.

(5) The principle of matching geographies to user needs
There can be no single geography which satisfies all users. The corollary is that there several different geographies will need to be produced.

(6) The principle of common geographies
The geographies produced should be common across all parts of the United Kingdom.

(7) The principle of user-defined geographies
There is a need to create systems for generating small area geographies that can be tailored by users to their specific purposes.

(8) The principle of delivering the census as a GIS
The Small Area Statistics should be delivered simultaneously with the appropriate digital boundary data either as an integrated geographical information system or as suitable input to such a system. Note that most of the work of preparing the boundary systems can be done between now and 2001.

There are many precedents in past census practice in the United Kingdom for each of these principles.

Principle (1), preserving confidentiality, is a duty of the Census Offices with respect to all statistical abstracts. In the past a great deal of caution in producing outputs was exercised because the risks of disclosure were not well understood. One of the goals of the Census Development Project and associated academic projects should be to understand what the risks of disclosure are when employing various protection methods.

Principle (2), that small area data need protection, follows from the first principle. A variety of methods have been used to protect data but these have not been systematically evaluated and compared in terms of their costs and benefits.

Principle (3), that there has to be a trade-off between size of geographical unit and detail released, has been applied in the production of small area statistics in the Censuses of 1971, 1981 and 1991. Fewer counts are produced the smaller the scale of area. For the part postcode unit only counts of total persons and households are produced. For enumeration districts circa 9,000 counts were produced in 1991, while for wards circa 20,000 counts were produced in the Local Base Statistics.

Principle (4), the separation of input and output areas, was adopted in Scotland for all SAS outputs in 1991, where output areas were designed as aggregations of unit postcodes that could match 1981 enumeration districts. In both England and Wales and Northern Ireland SAS outputs were produced that were not based on the collection area: postcode sector tables in both countries and also grid square data in Northern Ireland.

Principle (5), of matching geographies with user community needs, has emerged in the user debriefing (1991 Census) and user consultation (2001 Census) meetings held by the census offices over the 1994-5 period. Each user community has its preferred small area geography. Local and central government insist on exact fits with statutory areas: local government districts, electoral wards and Parliamentary constituencies. Business and marketing agencies insist on areas built from postal units. Academic researchers use both these types of area but are interested in measuring change over time, which has proved hard to achieve because of both changes in collection areas or the postal coding system.

Principle (6) calls for common, agreed approaches by each of the UK's Census Offices. Because the Census has been carried out by three different agencies, there have been considerable difficulties in comparing topics across England and Wales, Scotland and Northern Ireland. The 1991 Census saw considerable progress in Northern Ireland where the joint Census Office, ESRC and Queen's University project has produced Small Area Statistics for enumeration districts in the province, defined on the same basis as in England and Wales.

Principle (7), that there should be a system for supplying user-defined geographies, has a precedent in the provision of tables to researchers from the Longitudinal Study of England and Wales. The data are kept in a safe setting, the OPCS computer system, and user-requested tables are only released after vetting for confidentiality risks. The challenge in 2001 is to extend this system to the production of area based tables and to make the checking procedures much less labour intensive.

2.3 Multiple standard geographies

There are at least four types of small area geography for which some census users have been lobbying:

(1) small areas that fit exactly into administrative units;
(2) small areas that are exact aggregations of unit postcodes;
(3) small areas that are regular spaces defined by the National Grid system of co-ordinates; and

(4) small areas that match those used in previous censuses.

2.3.1 Administrative small areas

Units used for local government and for local and national elections are established by statute. These include local government districts and local electoral wards. Other areas such as counties or regions can be built up from them. SAS data are needed by local governments for all sorts of administrative and planning purposes.

In the 1991 Census for England and Wales and Northern Ireland small area data were produced exactly for electoral wards. In Scotland small area data for wards were produced on a best fit basis by aggregating unit postcode data.

Within wards the areas used were the census collection areas called enumeration districts in England and Wales. These have no function beyond that being used in the census to provide equal workloads and a convenient, easily recognisable patch within which enumerators distribute and collect census forms.

2.3.2 Postcode based areas

A great many databases of non-Census data hold postcoded information, which needs to be linked and related to census variables. This requires use of small areas based on postcodes. The need is paramount in business and medical research applications.

The smallest areas for which SAS tables were produced in Scotland were Output Areas (OAs) based on aggregations of unit postcodes. These OAs sum to postal sectors, postal districts and postal areas. SAS data for postal sectors were produced in England and Wales in addition to that made available for wards. The production of both postal sector and collection area geographies by OPCS for the 1991 Census implies an acceptance that the risk posed by differencing these data sets is low. Differencing is discussed in more detail in section 5.2.

2.3.3 Grid based areas

Grid based SAS data were produced in Great Britain from the 1971 Census and a limited number of small area counts can, in principle, be produced from the 1971, 1981 and 1991 Censuses of Northern Ireland. Grid areas lend themselves to mapping, to linkage to environmental variables and to comparisons over time because the Grid framework is time-independent. No grid square data were produced from the 1991 Census of Great Britain, but Bracken and Martin (1995) created a system for estimating grid data from enumeration district SAS from the 1981 and 1991 Censuses, which enables researchers to make comparisons over time.

2.3.4 Historically consistent small areas

One of the major questions that a census is designed to answer is "how much change has occurred since the last census?"

The Output Areas in Scotland were designed to be aggregated easily to form equivalents of the enumeration districts in the 1981 Census. In England and Wales and in Northern Ireland comparable areas can only be formed by aggregating EDs (the 1971-81 Census Tracts) or by carrying out estimation to common units (Bracken and Martin 1995 or Dorling 1995). A large percentage of both EDs and wards changed their boundaries between censuses.

The case for each of these geographies is unassailable. In planning for 2001 it is imperative that either the production of many standard geographies is proved to be a safe thing to do with existing protection procedures or that the production is made safe through additional protection procedures.

Will the production of many standard SAS data sets pose problems at time of production? The most difficult work involved is to build the correct lookup tables from the basic building blocks to be used in the Census (either residential addresses or part-postcode units). All this work can be done between now and 2001. In 2001 the Census Offices should have fast tabulation packages available which can produce the SAS very speedily from the Master files. Processing time should not be a constraint. The great advantage to the Census Offices would be that they would have several tailored products to sell rather than just one.

2.4 Flexible geographies

Standard geographies for reporting census results may not satisfy all needs. There will inevitably be revisions of the standard geographies in the intercensal period, for which small area statistics will need to be produced. For example, new electoral wards will be defined by the Boundary Commissions. New policy areas may be developed. Research may demand the use of areas defined as homogeneous with respect to particular variables.

There should therefore be a method for producing tables for user-defined geography. Clearly, because of the dangers to confidentiality that such a facility poses it should be kept behind a security fence and outputs assessed for confidentiality risk before release.

2.5 Safe data versus data in a safe setting

Marsh, Dale and Skinner (1994) have made a very useful distinction, in the context of census microdata, between **safe data and data in a safe setting**. Safe data are individual records sufficiently protected by anonymity, sampling, broad coding, and legal conditions that then can be released to users. The Samples of Anonymised Records from the 1991 Census constitute such a safe data set. Data in a safe setting are individual records in their original fully coded form from which statistical tables can be produced, the production of which can only occur in a safe setting (in a secure computer designated by the Census Offices). Security procedures include the vetting of researchers with access to the data, the logging and monitoring of table requests and the checking of table outputs. The Longitudinal Study of OPCS (England and Wales) is data in a safe setting. Another example of data produced in a safe setting is the production of special tabulations to order by the Census Offices.

Such a distinction can be applied to counts and tables for geographical areas. Small area statistics for multiple standard geographies should be released as safe data. The flexible production of counts or tables for user-defined geographies should be data produced in a safe setting. Production of data in a safe setting is currently a very expensive operation: Census Office statisticians must check outputs and specialist staff have to run the database jobs to produce tables. There would have to be considerable automation of the whole process through the development of security software in order to reduce costs. The tabulation production should be a task for users to do, with the software making decisions about its release from the safe setting and making reports to Census Office personnel in the case of potential problems.

2.6. Building blocks

To produce multiple small area statistics, two approaches have been suggested.

(1) Construct tables for each geography using the household and communal establishment returns themselves. The residential addresses of the country would be precoded using all likely standard systems: postcode, census collection area, OS grid co-ordinate, previous census code. Fast database packages would then be used to produce the necessary small area statistics (sets of preset tables).

(2) Produce small area statistics for a very small area geography and then construct small area statistics for any other areas by aggregating these SAS for basic building blocks. The basic Building Blocks (BBs) would not be unit postcodes but the part postcode units formed by overlaying postcodes on administrative geography. These part-postcode units are used in the ED to Postcode lookup tables constructed for the 1991 Census. The aggregation of SAS from these BBs could be accomplished using the aggregation functions built into software such as SASPAC (Small Area Statistics Package).

2.7. The confidentiality problem

The heart of the confidentiality problem is the risk that a crosstabulation within the Small Area Statistics contains single counts which enable an observer to recognise an individual living in the area and to infer further information not used in that recognition.

2.7.1 Thresholding and table specification

The primary protection of the individual or household data is the requirement that areas contain more than a minimum number of people or households. In the 1991 Census, the threshold size was set at 50 persons and 16 households. If enumeration districts (England and Wales, Northern Ireland) did not meet these thresholds, they were amalgamated with neighbouring areas. In Scotland, output areas could be designed to meet the thresholds so this slight difficulty did not occur.

A second protection was to vary the degree of detail available for areas depending on the population they contained. The smaller the area, the fewer counts that could be produced, the larger the area the more counts that could be produced. Practice in the 1991 Census is instructive in this respect.

(1) A count of number of resident households was released for part-postcode units (in the ED/PC directories).

(2) In Northern Ireland circa 800 counts are available for grid based areas defined by users. The areas can be as small as 100 metre grid squares in urban areas.

(3) The Small Area Statistics make available circa 9,000 counts organised in 86 tables for the smallest areas upwards (enumeration districts in England and Wales and Northern Ireland, output areas in Scotland). The areas contain about 370 people on average.

(4) Local Base Statistics make available circa 20,000 counts organised in 99 tables for the second smallest areas upwards (wards in England and Wales, postal sectors in Scotland). The areas contain about 5,000 people on average.

The principle of adjusting the quantity of information released to users to match the size of area has been built into the output policy of the UK Census Offices.

2.7.2 The differencing problem

The problem with thresholding as a protection principle is that the release of several small area data sets may make it possible to derive other data sets which violate the thresholding rules. This can occur if the counts for one set of areas are subtracted from another set of areas using a different geography. This has been termed the **differencing problem.**

There are two different situations at issue. The first is that the two geographies are roughly the same scale and generally overlap. The second is that one geography has much smaller areas than a second geography so that subtraction of wholly nesting smaller areas from larger areas might yield tables for area differences which violate the thresholds. For example, postal sectors and wards are two boundary networks of the same size that overlap and could be differenced. Enumeration district and postal sector data could also be differenced though no one has, to my knowledge attempted such an exercise.

General overlapping does not pose a problem because estimation techniques must be used to construct tables for the intersecting areas. These estimation techniques might use area shares or population shares (e.g. from the part postcode unit data). When the two geographies are similar in scale, the estimates can be very inaccurate.

Nesting poses more dangers and this led the Census Offices to adopt additional protection devices for the statistics published for the smallest areas. These protection devices are discussed below.

2.8 Protection alternatives

2.8.1 Safe setting

The only feasible protection for a system that allows the production of tables for any user defined geography is to place the system in a safe setting and to have means to vet output before release (cf. The Longitudinal Study). However, extension of current LS or special tabulations practice will not meet user demands for fast turnaround of such requests. Three features need to be built into the system.

(1) A very fast database/tabulation package is needed. Open tender to software firms, academic centres and other national statistical agencies should yield a suitable product.

(2) A safe sample dataset (anonymised and heavily perturbed in lots of different ways) should be released to users so that they develop their tabulation jobs using the fast database/tabulation package. The safe sample dataset will differ from the SAR in containing the full classifications of each variable that will be used in the safe setting.

(3) A package that designs the zones that meet user needs and links to the fast tabulation package, reporting the risks involved in release of the results for vetting by the Census Offices.

2.8.2 Safe data

A careful assessment, through rigorous experimentation with the relevant 1991 Census data, needs to be made of the efficacy and costs of the following measures that have been used or which are proposed to protect tables of data release for small geographical areas.

(1) Minimum person and household **thresholds.**
(2) **Blurring** of counts through random perturbation.
(3) **Suppression** of parts of tables.
(4) **Broad coding** of variables.
(5) **Controlled rounding** of tables.
(6) Aggregation of tables from very small to small areas by **point in polygon matching**.
(7) Reporting ratio variables or derived indicators rather than counts.

The evaluation of each of these measures must take into account "natural" sources of error:

(a) Reporting errors (investigated via a Census Validation Survey)
(b) Question ambiguity
(c) Non-response and the need for imputation
(d) Missing households which are imputed
(e) Missing persons and households (under-enumeration)
(f) Decay of relevance of data over time as a result of population change.

The aim in holding a Census and disseminating the results is to reduce "natural" errors to a minimum, but a recognition that such errors contribute to data protection would be useful.

2.9 Final Remarks

This paper has attempted to set out a framework for thinking about the geography of output from the next census. It offers no set of prescriptions for what should happen but tries to put forward in a systematic way the options that should be considered. A great deal of research and development will be needed to bring many of these ideas to fruition. The academic community should be able and be prepared to contribute to that research and development if it wishes to see some of the proposals put forward here become reality.

3. FLEXIBLE GEOGRAPHIES AND AREA AGGREGATION: DESIGNING SMALL AREAS FOR OUTPUTS FROM THE 2001 CENSUS

Philip Rees and David Martin
April 1996

This paper was presented at a meeting organised by the Royal Statistical Society and British Society for Population Studies on The 2001 Census: Outputs and Geography, held on 18 April 1996 at the Royal Statistical Society, Errol Street, London. It revises the arguments presented in the preceding paper and adds to them by considering the building blocks that could be used in constructing the output geography of the 2001 Census. The maps and diagrams in the original paper, not central to the points made, have been omitted in this version.

The first section of the paper defines the issues involved in delivering statistics for small areas. The second section outlines a framework for decision making, large parts of which have now achieved consensus agreement (Rees 1995, Thomas 1996a). The third section describes the strategy being adopted in a research project investigating the feasibility of protecting small area data produced for several small area geographies. Section four of the paper discusses the basic spatial units or building blocks that might used to construct output areas, while the fifth section reviews the alternative candidates. The sixth section connects the problem of output area definition to work on a British standard for addresses. The key message of the paper that ways must be found to produce the variety of geographic data sets which users are asking for.

3.1 Background and issues

A variety of outputs can be expected from the 2001 census. These include printed volumes of statistics such as (1) key statistics for areas, (2) published national topic reports and (3) published area reports (for unitary areas and regions), and machine readable data such as (4) flexible tables derived from the master database (data in a safe setting), (5) pre-specified tables for geographic areas (safe data), (6) sample microdata (safe data), (7) interaction or flow data (safe data) and (8) boundary/georeference data (safe data). Users are already requesting easier and cheaper access to special tables, small area statistics for more than one geography, better migration and journey to work flow data, more microdata sets and simultaneous delivery of the georeference data with census statistics. In this paper discussion focuses on pre-specified tables for geographic areas, the product called Small Area Statistics (SAS) in the 1966, 1971 and 1981 Censuses and extended in 1991 to include Local Base Statistics (LBS).

In providing new outputs from the 2001 Census, the Census Offices must balance freedom of information against the right to privacy. They have a statutory duty to place output statistics before Parliament and to preserve confidentiality of the original enumeration records for 100 years. On the other hand there are intense pressures from users to obtain more detailed and specific statistics and from the Treasury to increase the commercial return from Census outputs.

The following questions arise concerning the production of pre-specified tables (SAS/LBS). (1) For what kind of areas are SAS/LBS needed? (2) How will the SAS/LBS be protected? (3) What will be the contents of the SAS/LBS? (4) How will the SAS/LBS be delivered? In the present paper we concentrate on the first two questions and consider a framework for answering them in the next section.

3.2 A framework for choosing and designing small area geographies

The framework described here was first outlined by Rees (1995) and then taken up and then taken up and modified by Thomas (1996a). The accounts of the frameworks are very similar except they are expressed from the point of view of the provider and user. Here we set out definitions of key terms and briefly present the principles, so that attention can be concentrated on the three principles on which agreement has not yet been reached.

3.2.1 Definitions

Some definitions of key terms are needed in order to clarify the debate.

Geographies are systems of aggregation of individual and household data to areas whose spatial extent is known, either as sets of points or as the territory within known boundaries.

Small areas are territories smaller than the units used for local government (though they need not necessarily add exactly to them).

Standard geographies are a set of small area boundaries for which small area statistics are published.

Flexible geographies are areal aggregations defined by users for which they can produce their own sets of statistical tables.

Small Area Statistics are a suite of different sets of fixed tables of census counts. These SAS may be very simple for small areas of minimum size but quite detailed for the largest areas. This extends the distinction between small area statistics and local base statistics developed very usefully in the outputs from the 1991 Census of Population for Great Britain.

Addresses are descriptions of the location of residential accommodation (whether for private households or in communal establishments) used by the postal service (Royal Mail).

Co-ordinates are pairs of numbers that indicate geographical location in the National Grids of Great Britain and of Ireland used by the Ordnance Surveys of Great Britain and Northern Ireland respectively.

A *threshold* is the minimum number of households or residents or other units or combinations of households and residents for which a given output will be produced.

Differencing is the derivation of a third set of statistics by comparing a first and second set, which may produce statistics for populations which fall below pre-defined thresholds.

3.2.2 Principles for designing SAS output geography

The following principles have come to be accepted by Census takers and Census users as a result of experience and innovations introduced in the past three censuses and as a result of discussions leading up to the 2001 Census. There is agreement between census providers and users on the following *six* principles.

(1) The *confidentiality* principle
The outputs of each geography should not compromise the confidentiality of the individual and household records.

(2) The *non-disclosure* principle
Protection measures will need to be applied to counts for small areas to ensure confidentiality.

(3) The *threshold* principle
There will need to be a trade-off between the size of area and detail of output provided.

(4) The principle of *separation* of input and output areas
The geographies of input areas (collection areas) and of output areas (publication areas) should be separate.

(5) The principle of *consistent* geographies
The geographies produced should be common across all parts of the United Kingdom.

(6) The principle of *delivering the Census as a GIS*

The SAS should be delivered simultaneously with digital boundary data.

The next *three* principles are still the subject of debate and of research.

(7) The principle of *matching standard geographies to user need*
Several different geographies will be needed to match general user needs.

(8) The principle of *user-defined flexible* geographies
A system is needed for users to generate geographies tailored to their specific purposes.

(9) The *approximation* principle
Output areas must fit areas of interest within an agreed tolerance.

The arguments supporting principles (1) to (6) are rehearsed in Thomas (1996a). Here we concentrate on principles (7) to (9) which are still the subject of debate.

3.2.3 Standard geographies

Principle (7) proposes that a small set of standard geographies be used for the production of small area statistics from the 2001 Census. User consultation in debriefing sessions after the 1991 Census revealed that different customer sectors had different preferences for small areas. We can identify four different types of area for which substantial groups have expressed a preference and which can be defined *before* the 2001 Census.

3.2.3.1 Small areas fitting into *administrative* units

Central government departments carry out statutory duties which involve the allocation of resources to subnational areas (e.g. the allocation of local authority block grants by the Department of the Environment). Exact census statistics are needed for local government areas as input to the resource allocation formulae. Even slight approximations can result in the misallocation of millions of pounds of grant. Within local government areas similar resource allocation takes place to electoral ward areas, although the processes used are not standard across all local governments. Local councillors represent electoral ward populations, which they require to be allocated fair shares of district resources using the best population statistics. In addition, the National Health Service has similar needs for census statistics for allocating funds to health districts.

Figure 3.1 provides an illustration of small area geography from the 1991 Census for Leeds and Wakefield in West Yorkshire. The small areas were census collections areas called enumeration districts (EDs), which nest exactly into the electoral wards current for the local elections of May 1991 and hence into the local government district boundaries. We note that EDs are not administrative units and will not necessarily be needed as output areas in 2001 (because of the acceptance of principle 4 for the conceptual separation of input and output areas). If collection areas are not used in 2001, then a set of sub-ward areas that nest exactly into wards will need to be designed.

3.2.3.2 Small areas that are aggregations of unit *postcodes*

Marketing analysis companies and consultancies do extensive work for commercial clients. A lot of this work involves the bringing together of client databases of address information and census information. The requirement is for small area statistics for areas built exactly from unit postcode. A similar need is expressed by medical researchers who wish to measure health risks: the case records of mortality and morbidity are address based and are matched most easily with small area populations at risk for aggregations of unit postcodes.

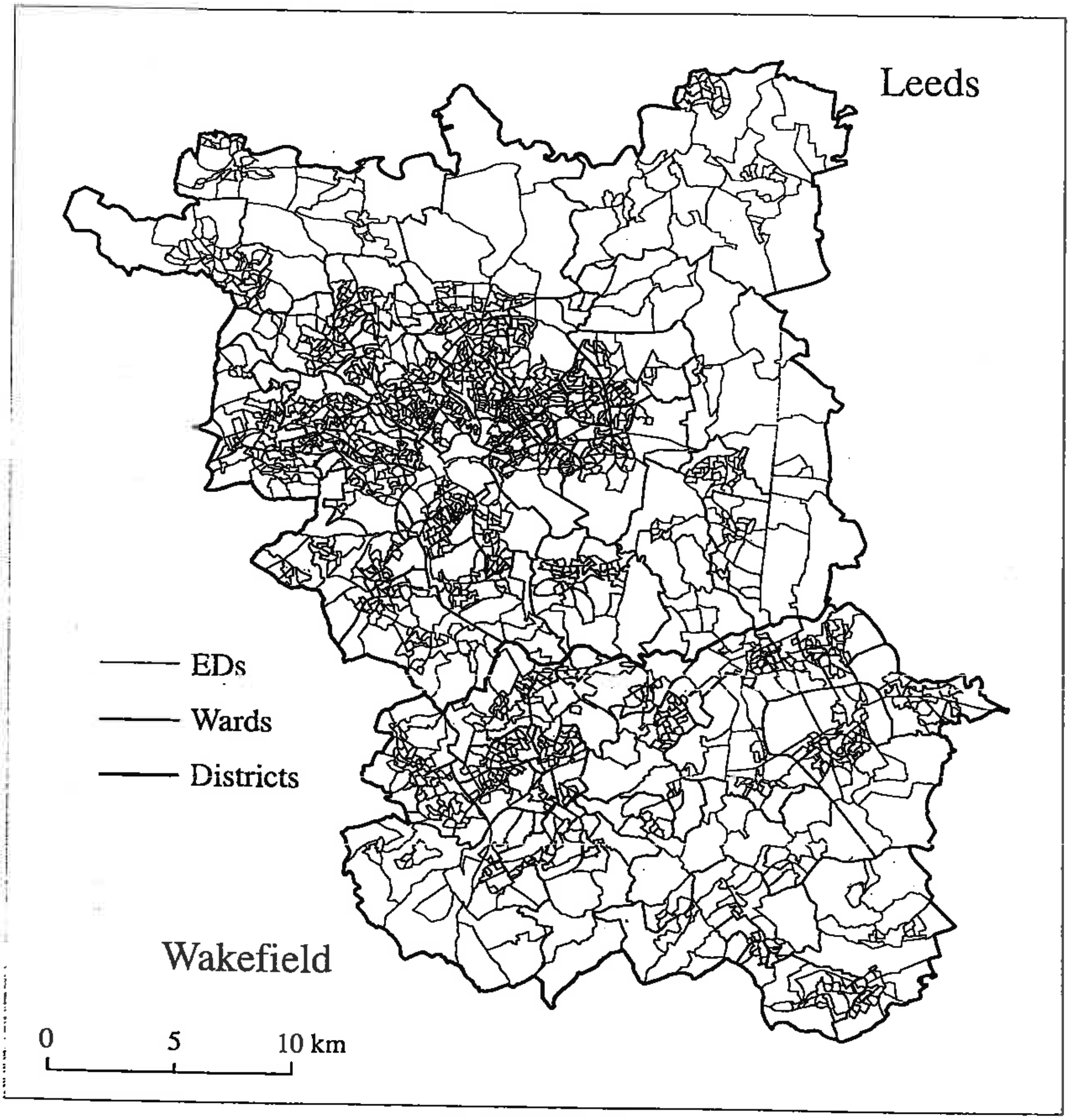


Figure 3.1: Example of administrative small areas: Leeds and Wakefield - 1991 Census wards and EDs

Figure 3.2 depicts the postal geography of the Leeds and Wakefield areas. Comparison with Figure 3.1 shows that there is virtually no relationships between postal and administrative geographies. We note that there exists no standard set of sub-postal sector small areas for use with 1991 Census data. A set of sub-sector areas will therefore need to be designed for use with the 2001 Census. Note that these need not be similar to the pseudo-EDs proposed for the 1991 Census, which were aggregations of unit postcodes that closely fit the EDs used in the Small Area Statistics.

3.2.3.3 Small areas that are regular spaces defined by National *Grid* co-ordinates

Environmental scientists gather data on the properties of air, water or land on a grid basis and wish to link their data to census populations reported on the same regular basis. In 1971 Small Area Statistics were produced for kilometre squares for Great Britain and in Northern Ireland a software system was developed for producing a small number of frequency counts for grid based areas (1 kilometre grid squares in rural areas and 100 metre grid squares in urban areas). For 2001 the design issue would be to decide what size grid squares should be used: 1 km for all areas or with smaller grid squares for denser areas.

Figure 3.3 provides an example of grid area map. The data derive from the population surface counts developed by Bracken and Martin (1995). The grid areas are 200 metre squares which provide the fine areal mesh needed in urban areas.

3.2.3.4 Small areas *comparable over time*

The first use of census data by virtually all users is a comparison of population levels and compositions since the last census. The absence of comparable set of small areas between 1981 and 1991 in England and Wales has caused users much difficulty although imaginative solutions have been developed. Dorling (1995) developed a look up table between 1991 EDs and 1981 ward units; Bracken and Martin (1995) produced an algorithm for linking both 1981 EDs and 1991 EDs to 200 metre square grid areas. In Scotland the use of postcode aggregations as Output Areas meant that areas comparable with 1981 EDs could be constructed directly.

Figure 3.4 shows the geography of the 1981 Census for Leeds and Wakefield. Enumeration district boundaries are not available in digital form on a national basis and so are not depicted. A close comparison with Figure 1 will reveal that the ward boundaries of Leeds Metropolitan districts have not changed between 1981 and 1991 but both the number and boundaries of Wakefield wards changed. One key advantage of producing Small Area Statistics for 1991 small areas is that all the digital boundaries for these areas already exist so that principle (6) is readily achievable.

Assuming that the case for producing small area data for these four standard geographies is accepted and implemented, will this satisfy all possible user needs? The answer is no: there will always be users who have good reasons for seeking statistics for special purpose areas. There is also another need: to generate small area statistics for administrative and postal areas, the boundaries of which change. Major administrative changes of district level boundaries occurred in the 1960s, 1970s and 1990s. It would be wise to anticipate another reorganisation in the 2000s. In fact, electoral divisions are always systematically revised in intercensal periods, using the evidence of the latest census in order to preserve the equality of electoral populations for wards and constituencies (Parliamentary and European).

There are two ways of producing statistics for these new areas. The first method is to develop very accurate look up tables using the correct population base counts which would enable users to aggregate one of the standard geographies into a new version of the standard geography or a user-required geography. The second method is to generate exact counts in a safe setting (on a Census Office computer) and check that confidentiality is not breached. Section 2.4 describes the structure of such lookup tables, while section 2.5 describes the features required of a system for generating flexible geographies.

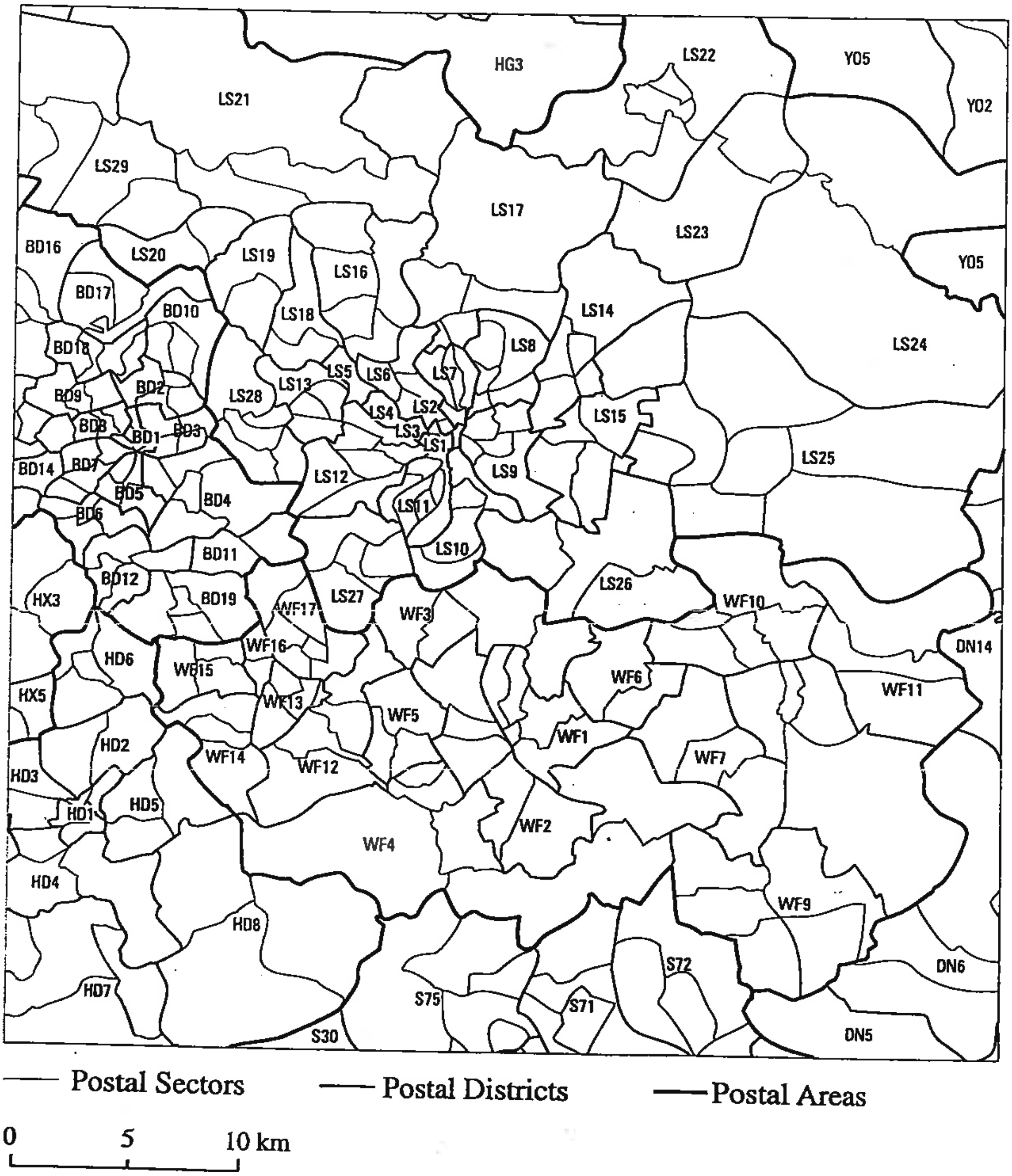


Figure 3.2: Example of postal small areas:
Leeds and Wakefield area - 1991 postal districts and sectors

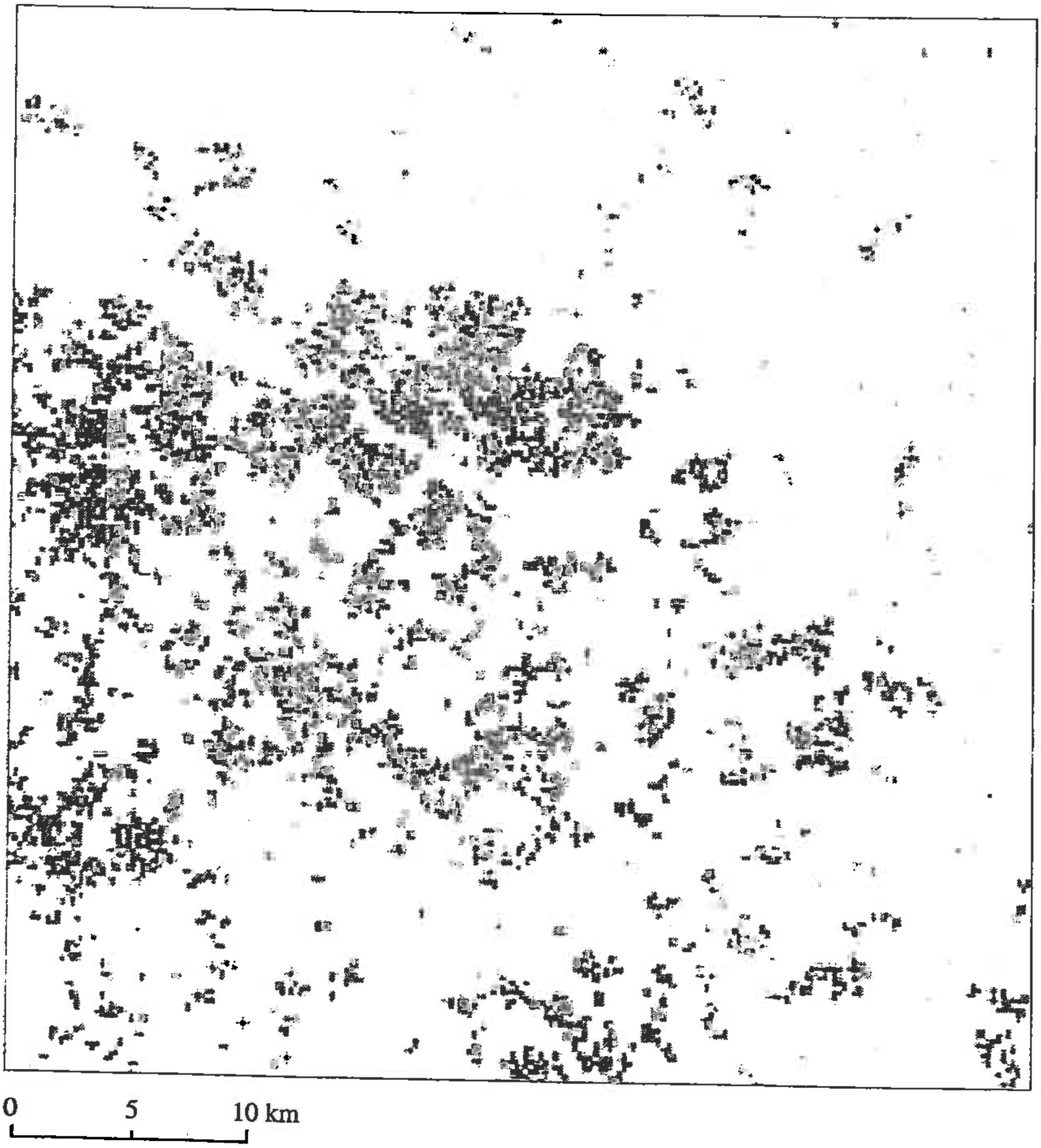


Figure 3.3: Example of grid square small areas: 1991 Census raster population surface (200m^2)

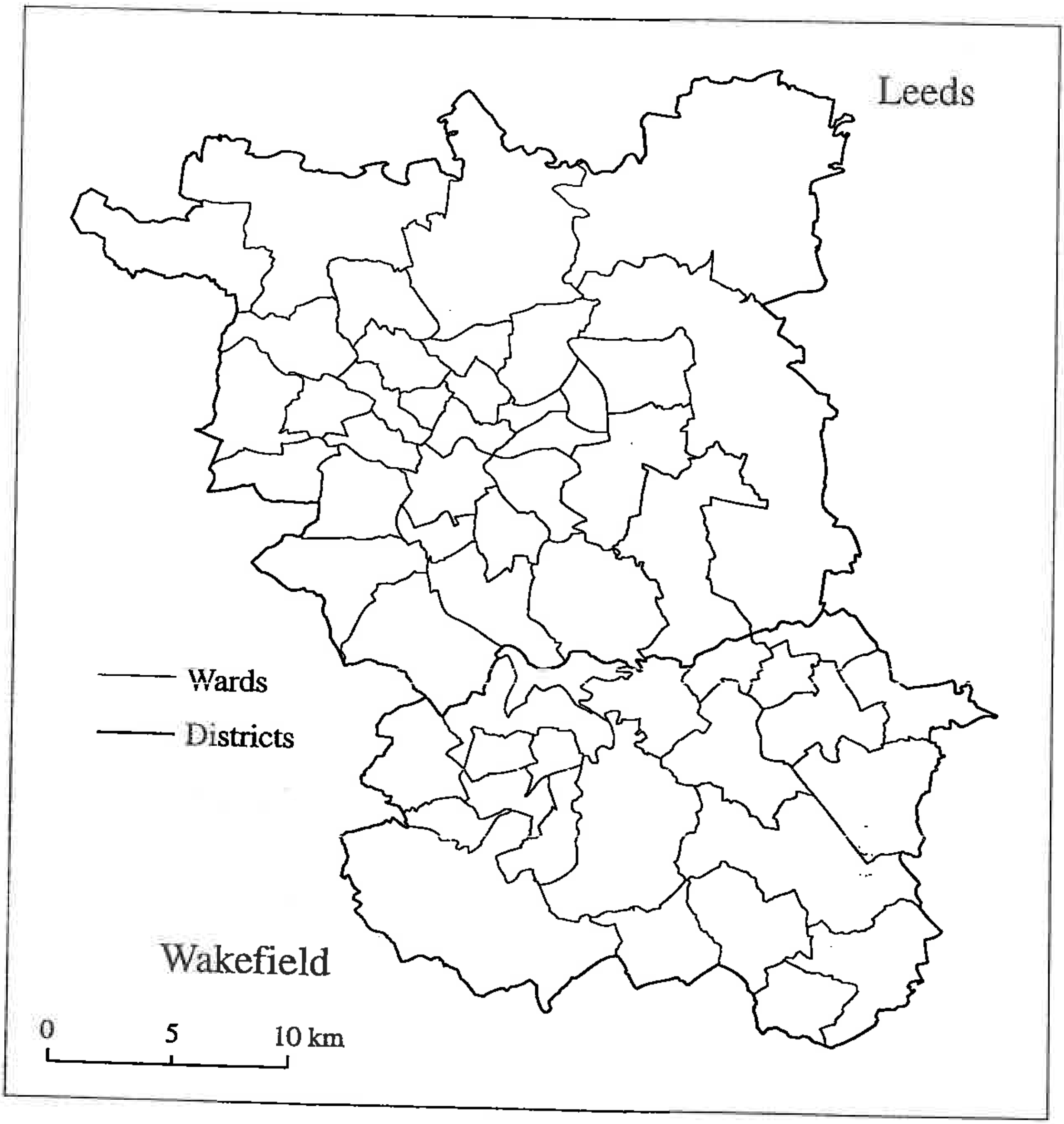


Figure 3.4: Example of the geography of a previous census: Leeds and Wakefield area - 1981 wards

3.2.4 The need for look up tables

Look up tables are widely used information handling. They consist of lists of areas organised and labelled using one classification which provide for each area details of its membership of another classification. There are several kinds of look up tables:

(1) tables implementing perfect aggregation which does exist
(2) tables assuming perfect aggregation which may not exist
(3) tables involving imperfect aggregation using population counts
(4) tables involving imperfect aggregation using surrogate measures.

It is useful to provide examples of each type to fix ideas.

The list of 1991 Census Enumeration District codes for England and Wales is an example of a look up table of the first type. These contain within them the codes for the ward, district and county that the ED fits into.

A list of 1991 Census Scottish Output Areas with their electoral ward memberships attached (used to develop ward SAS on a best-fit basis) is an example of a look up table of the second type. Because the smaller areas do not fit into the larger areas exactly, perfect aggregation is not achieved.

The 1991 Census Enumeration District/Postcode Directory in England and Wales is an example of a look table of the third type. It relates two geographies: the collection areas for the census and unit postcodes used by the Royal Mail at the time of the census. The look table lists areas which are the intersections of EDs and postcodes and includes information on the number of households in the areas. Using this information fractional weights can be computed for converting ED SAS into unit postcode SAS. The unit postcode SAS can then be aggregated to larger postal units defined by the user and relevant ratio statistics computed.

An example of a look up table of the fourth kind would be a list of ED and policy area intersections which includes a surrogate variable for allocating the ED population across the intersections. Land area derived from GIS overlap analysis is one such surrogate variable used in the West Yorkshire Population Model and Information System (Rees 1995). Test have shown that using land area does not result in very accurate aggregations for area sets of similar size.

If it were decided not to produce SAS for the four standard geographies, then we would strongly recommend production of a series of look up tables of the third type. One very useful extension of the ED/PC directory information would be to provide more population base counts on which to base the aggregation of SAS tables. At a minimum these should include besides the number of resident households, the number of residents in households and the total number of residents. At maximum the look up tables could include all of the separate resident or household bases for SAS tables.

3.2.5 Flexible geographies

Why might a system for producing tables for geographies other than the standard be needed? There are two reasons: firstly, standard geographies may not satisfy all needs; and secondly there will be revisions of the standard geographies in the intercensal period. For example, new electoral wards will be defined by the various Boundary Commissions. New policy areas may be developed. There should therefore be a method for producing tables for user-defined geographies.

This should be part of a larger flexible tabulation system in which users can request the tables they need.

There should therefore be a method for producing tables for user-defined geography. Clearly, because of the dangers to confidentiality that such a facility poses it should be kept behind a security fence and outputs assessed for confidentiality risk before release.

Marsh, Dale and Skinner (1994) have made a very useful distinction, in the context of census microdata, between *safe data* and *data in a safe setting*. Safe data are individual records sufficiently protected by anonymity, sampling, broad coding, and legal conditions that then can be released to users. The Samples of Anonymised Records from the 1991 Census constitute such a safe data set. Data in a safe setting are individual records in their original fully coded form from which statistical tables can be produced, the production of which can only occur in a safe setting (in a secure computer designated by the Census Offices). Security procedures include the vetting of researchers with access to the data, the logging and monitoring of table requests and the checking of table outputs. The Longitudinal Study of OPCS (England and Wales) is data in a safe setting. Another example of data produced in a safe setting is the production of special tabulations to order by the Census Offices.

Such a distinction can be applied to counts and tables for geographical areas. Small area statistics for standard geographies should be released as safe data. The flexible production of counts or tables for user-defined geographies should be data produced in a safe setting. Production of data in a safe setting is currently a very expensive operation: Census Office statisticians must check outputs and specialist staff only have to run the database jobs to produce tables. There would have to be considerable automation of the whole process through the development of security software in order to reduce costs. The tabulation production should be a task for users to do, with the software making decisions about its release from the safe setting and making reports to Census Office personnel in the case of potential problems.

The only feasible protection for a system that allows the production of tables for any user defined geography is to place the system in a safe setting and to have means to vet output before release (cf. The Longitudinal Study). However, extension of current LS or special tabulations practice will not meet user demands for fast turnaround of such requests. Three features need to be built into the system.

(1) A very fast database/tabulation package is needed. Open tender to software firms, academic centres and other national statistical agencies should yield a suitable product.

(2) A safe sample dataset (anonymised and heavily perturbed in lots of different ways) should be released to users so that they develop their tabulation jobs using the fast database/tabulation package. The safe sample dataset will differ from the SAR in containing the full classifications of each variable that will be used in the safe setting.

(3) A package that designs the zones that meet user needs and links to the fast tabulation package, reporting the risks involved in release of the results for vetting by the Census Offices.

3.3 Problems and solutions

We have made three proposals about the geography of the SAS to be produced from the 2001 Census. Firstly, we have proposed that four sets of SAS be produced and disseminated corresponding to four standard geographies for which there is proven demand by different user communities. Secondly, we have proposed that a system be designed for flexible tabulation (including SAS tables) of the 2001 Census for other geographies that users might need to use or to cater for inevitable changes in two of the standard geographies. For these two proposals to be successful means of making the data safe from disclosure have to be developed. In the event that such methods cannot be successfully developed, a third proposal is that accurate lookup tables be developed that would enable users to create SAS for their geographies from one standard geography (which would probably involve the intersection of administrative and postal geographies).

In the next sub-section, we outline the main threat of disclosure, and then analyse the protection devices that have been used in different contexts.

3.3.1 The differencing problem

The main protection device used by Census Offices is the population threshold for geographic areas. A minimum size is set for the count of residents or of households, or both. If the counts for an area fall below the threshold, then no tables can be published for that area. In the 1991 Census an area had to have at least 16 households *and* 50 residents. However, if tables have been produced for more than one geography, it is possible to produce tables for areas with populations below the thresholds by subtracting tables for one geography from another.

There are two different situations in which differencing can occur: (1) where two geographies of roughly the same spatial *overlap* (e.g. postal sectors and wards) and (2) where one geography has much smaller areas which *nest* into a second geography (e.g. enumeration districts partially nest into postal sectors). In general overlapping does not pose a problem, though there are situations where deductions can be made. Nesting poses more dangers and additional protection devices for SAS are needed.

3.3.2 Devices to create safe data

How can census tables for geographical areas be protected so that they are safe to release despite the differencing threat? To help discuss the alternative ways this might be accomplished we first spell out the components of a protection system. This is laid out in Figure 3.5.

Stage (0) involves the *collection of the census records*. The aim will be to collect these as accurately as possible. Inevitably, there will be "natural" errors due to incomplete coverage, inaccurate responses or decay in relevance over time through population change.

Stage (1) provides the option for deliberately *modifying records* by swapping whole or part records. This is a drastic device to protect confidentiality and very infrequently applied. Users consulted about this technique have not favoured it.

Stage (2) offers the choice of *correcting the coverage errors* through imputation of the characteristics of households for which there is firm evidence of existence and/or through estimation of missing whole households or of missing individuals within households. The current intention for the 2001 Census is to implement imputation for both absent and missing households.

Stage (3) involves a decision about the *Basic Spatial Unit* (BSU) for which counts would be constructed. The second half of the paper discusses the options.

Stage (4) is concerned with the *aggregation of counts or tables* for BSUs to the various output areas. This aggregation can be "perfect" if BSUs fit exactly into output areas but may be "imperfect" if they do not. One method of "imperfect" aggregation that has been suggested involves assigning whole BSUs to output areas if the BSU centroids fall within the output area boundaries. Such point-in-polygon assignment has the advantage of not requiring digital boundaries for BSUs but is not very accurate if output areas are close in size to BSUs. Imperfect aggregation can be regarded as a means of protecting small area statistics.

Stage (5) defines a minimum *threshold* size for output areas in terms of the resident population (individuals and households). No small area statistics are published for areas that fall below the threshold. This is a widely used protection device, which has the merit of being easy for all to understand.

Stage (6) introduces the widely used protection device of *broad coding*: aggregating the categories used for variables employed in tables. Simple broad classifications are used in tables for small geographical areas; detailed narrow classifications are used in tables for large geographical areas.

Substantial experience has been built up by the Census Offices in applying stage (5), *Thresholding* and stage (6) *Table design*. For example, in England and Wales, if collection areas (enumeration districts) failed to reach thresholds, they were joined with neighbouring areas. In general a strategy was adopted that produced fewer counts the smaller the area, the more counts the larger the area. In the 1991 Census at the very smallest scale a count of resident households was released for part-postcode units (in the ED/PC directories). At a slightly larger scale in Northern Ireland about 800 counts are available for grid based areas defined by users. The areas can be as small as 100 metre

Stage	Activity and options
(0)	Collect the census records (0.1) Assess coverage (0.2) Assess accuracy (0.3) Assess timeliness
(1)	Adjust the records: (1.1) Either leave records alone (1.2) Or perturb the records through record swapping
(2)	Add missing records: (2.1) through imputation (2.2) through estimation
(3)	Form Basic Spatial Units (BSUs): (3.1) Either use addresses (households and communal establishments) (3.2) Or construct counts/tables for very small areas
(4)	Aggregate counts/tables to the target geography: (4.1) Aggregate BSUs using perfect hierarchy (4.2) Aggregate through point-in-polygon technique
(5)	Set minimum person/household thresholds.
(6)	Modify the tables by broad coding variables.
(7)	Modify the cell counts in tables: (7.1) Either perturb cell counts randomly. (7.2) Or suppress cell counts. (7.3) Or round cell counts with control to totals.
(8)	Carry out risk assessment on tables.
(9)	Recycle steps (1) to (7) taking new decisions.

Figure 3.5: A protection system for small area statistics

grid squares in urban areas. In the 1991 Census Small Area Statistics some 9,000 counts organised in 86 tables are made available for areas containing an average of about 370 people. In the 1991 Census, the Local Base Statistics make available circa 20,000 counts organised in 99 tables for areas containing an average of about 5,000 people.

Stage (7) makes *adjustments to the cell counts* in tables produced at the previous stage to protect confidentiality. There are three alternatives available: random perturbation of cell counts, rounding of cell counts to larger multiples and suppression of cells which contain uniques. The first alternative has been employed in the censuses of 1971, 1981 and 1991. It has the disadvantage of creating inconsistency between tables (which should have the same totals) and between nested areas (tables for smaller areas should add up to the larger). These features have disturbed non-expert users but expert users have found ways to adjust and constrain derived statistics so as to restore consistency. Rounding of cell counts has not been systematically implemented in UK censuses, but is implicit when tables are based on a sample of the census records. The final device of suppression has been extensively applied to the Special Migration Statistics (Set 2 for inter-district flows) and has rendered much of that dataset unusable. Users are strongly resisting any application of this protection device in future.

Stage (8) involves *assessing the disclosure risk* associated with the set of small area statistics produced. This disclosure risk is in part a function of the number of low counts (ones or zeroes particularly) in the table, and in part a function of the prior knowledge of person wishing to breach confidentiality.

The process of design of small area statistics then recycles and different options are tested out.

3.3.3 Protection for data in a safe setting

There are likely to demands for many more output geographies than proposed above to the standard set and users will also wish to create different tables. A system that allows the production of tables for any user defined geography must be placed in a safe setting which allows the vetting output before release. The precedents for such a system are the very sensitive *Longitudinal Study* datasets, linking the 1971, 1981 and 1991 Censuses, and the Special Customised Tabulations that can be ordered by users from the Census Offices. The problem with both these systems has been their high cost to users. Academic access to the Longitudinal Study requires an extensive support programme funded by ESRC and JISC, but which has the capacity to service about 40 projects simultaneously. The Special Tabulations service from the 1991 Census, although extremely valuable to researchers, has involved costs in the £1000s per table and long queues. Users would prefer costs in the £100s and short queues, of course.

Three features need to be built into an improved system: *(1) a very fast database/tabulation package; (2) a safe sample dataset;* and *(3) a zone design and risk assessment package.* Open tender should yield a suitable database/tabulation product, though careful evaluation is needed of the extent to which standard packages in the marketplace meet Census tabulation needs (e.g. do they allow variables to have a very large number of categories so that geography can be treated like any other variable in table design). The reason for suggesting that a safe sample dataset be released to users would be to enable them to develop their own tabulation jobs using the fast tabulation package for submission to the Census Offices' system. The Samples of Anonymised Records can fulfil this function, although not perfectly because the SARs will embody a substantial degree of broad coding. The final element in this flexible tabulation system is a set of routines that design the zones that meet user needs, that assesses and reports the risks involved in release of the results for vetting by the Census Offices.

3.3.4 Statistical Disclosure Control (SDC) project

Research into the issues discussed in section 3 of the Chapter is currently under way at the University of Leeds. The research is funded under European Union's ESPRIT (Information Technology) programme. A network of institutions is carrying out research into Statistical Disclosure Control. The leading partner is a group at Statistics Netherlands led by Leon Willenborg. Other partners include the University of Manchester (led by Angela Dale), the University of Southampton (Chris Skinner), Office of Population Censuses and Surveys (Andy Teague, David Thorogood and Jan

Thomas), the University of Eindhoven, the University of Padua, the Italian Statistical Office (ISTAT) and the University of Leeds (principal investigators Phil Rees and Stan Openshaw)

3.3.5 Flexible geography task of the SDC project

The SDC project at the Centre for Computational Geography, University of Leeds is called the *Flexible Geography (FG) Workpackage*. This project has two tasks for 1996 and 1997 respectively.

The first task, Task FG-1 concerns the safe release of SAS for standard geographies. The aim of this task is to carry out *experiments* and produce reports and recommendations for the production of statistical tables for geographical areas from the next Census round. To this end we are building a software system (a set of connected programs) that implements the protection system for small area statistics set out in Figure 5 above.

The second, Task FG-2, concerns the development of a system for assessing statistical disclosure risks in spatially aggregated data for a suite of geographies. The aim is to design a *red-orange-green* light system for automating the decisions about release of geographical tables from the Master database of a census.

3.3.6 The first experiment in Task FG-1

Once the protection system for small area statistics is working, we will carry out experiments using a simulated database of households and individual records. The experiment assumes that the 1991 Census Sample of Anonymised Records constitutes a complete Master file for a census which could be assigned a real geographic location on a probabilistic basis. A Master file for the population of Yorkshire and Humberside, one of the UK standard regions, has been constructed by sampling households in the 1 % SAR for each census collection area (enumeration district) in the region and by sampling communal establishment individuals from the 2% SAR. The households and individuals are then assigned a random but exact 1 metre location within the enumeration district using a random vector and distance from the ED centroid and a point-in-polygon technique.

We will then construct a representative sample of SAS tables from this database. These constitute a "true" set of tables for this synthetic population. Software will be developed for each protection measure proposed. There will a volume control for each measure that can be turned up or down in the experiments.

The sample of tables from Small area statistics for each of the four multiple standard geographies will be generated for the study area (administrative, postal, grid and historic). The table sets will then be systematically differenced and the degree of disclosure in practice determined. The experiment will be repeated using a variety of options at each stage in the protection system and a variety of settings of the protection device.

3.4. Basic spatial units for areal aggregation

In the late 1980s, there was considerable discussion of 'basic spatial units' (or BSUs) (e.g. Openshaw, 1990), which were described as a single set of small geographical areas from which most other areal geographies could be defined. It should by now be apparent that no such set of areas exists for the purposes of the 2001 Census, as different users require aggregation geographies which are based on completely different criteria. In the past, the geography of census EDs became a de facto standard which was widely used for the organization of other geographical datasets. It is suggested here that potential census users are now quite strongly committed to other georeferencing systems and that integration with these other developments is at least as important as the establishment of another new (incompatible) geography for 2001 data output.

3.4.1 New georeferencing systems

Growth in the use of postcode- and address-based information has played a significant role in the development of the present situation (Raper *et al.*, 1992; Martin and Higgs, 1996). Despite the increasing computing power available to census customers, and the range of user-defined geographical areas which may be required, there will still be sizeable demand for areally aggregated data for one or more standard output geographies. These geographies might be constructed directly from the collection geography, as in the past, or might be entirely new aggregations derived from the individual-level census data held by the Census offices.

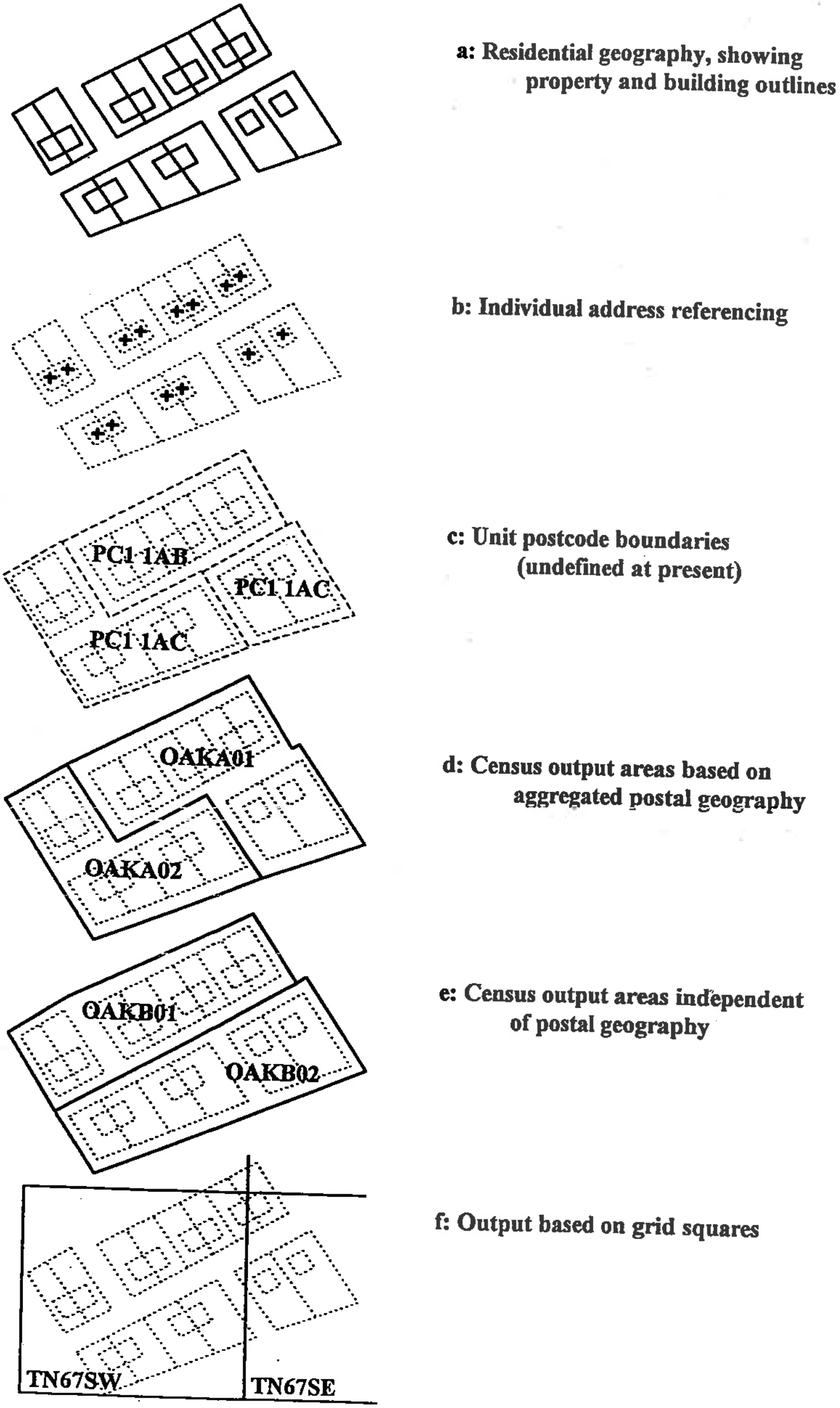
For many purposes the most attractive option will be to build areally aggregated datasets which 'fit' into widely used standard geographies, and it is therefore necessary to consider the most appropriate geographical objects from which such output geographies can be defined. Candidates clearly include the individual address, the unit postcode, the part postcode unit, the street, 1991 or 2001 enumeration district, ward, grid square, locality, and others, including hybrid objects constructed from the intersection of these basic building blocks. In the past, the manual boundary definition process used in census geography design has been a major obstacle to the preparation of multiple output geographies, but the existence of address-level referencing in 2001 offers the potential to create multiple aggregated geographies automatically. The constraints are therefore no longer primarily technical ones, but are concerned with confidentiality, production costs and user demand. For the first time in 2001 complete sub-metre geographical referencing will be available at the time of census design, using Ordnance Survey's ADDRESS-POINT product (Elliott *et al.*, 1993). ADDRESS-POINT provides a full postal address with postcode, a 0.1m resolution grid reference, and a unique reference number (OSAPR) for each property. Puckey (1995) outlines a number of output options which become possible if census data for individual households are linked with ADDRESS-POINT references, but suggests that such linkage would probably need to be conducted during the processing cycle, by inputting all or part of each address and matching with the geography base. This approach is strongly recommended here as the basis for the Census offices' internal database from which other output geographies will be derived. Some consideration will now be given to some of the major strategies for standard aggregated geographies.

3.4.2 New data handling techniques

The massive expansion of geographical information systems applications in the late 1980s and early 1990s has meant that many more users of census data have powerful tools for the manipulation of different geographies (Martin, 1996). These users are often able to explore a variety of methods for the allocation (for example) of postcodes to enumeration districts, but each of these can only provide an estimate for situations in which the definitive relationships are not published as part of the standard census outputs. The growth in the widespread use of digital geographical datasets during the same period has meant that there are more potential georeferencing systems than ever before for data derived from the census. In understanding likely user requirements for aggregate data in 2001, it is necessary to be aware of the ways in which potential census users are already working with extensive geographically referenced datasets. For many organizations, the ability to associate census data with their existing operational databases will be a major consideration. This has become particularly relevant as more organizations (both public and commercial) have sought to integrate census data with internal information in order to profile market or service catchment areas, construct geodemographic indicators, or characterize clients (see, for example, Birkin, 1995). In this context, there is also likely to be a considerable market for new ancillary products related to the 2001 census outputs, which provide clear definitions of the relationships between different sets of areal aggregations. While preserving census confidentiality, it should be possible to utilize the full address-level census database within the Census offices in order to provide information to users on the scale of mismatches and errors which may occur when working with a particular geography.

3.5 Candidates for BSUs

Diagrammatic examples of each of the options discussed here are given in Figure 3.6, which illustrates how the different types of geographical referencing might relate to the small area of residential geography, for which property and building outlines are shown in Figure 3.6a.



a: Residential geography, showing property and building outlines

b: Individual address referencing

c: Unit postcode boundaries (undefined at present)

d: Census output areas based on aggregated postal geography

e: Census output areas independent of postal geography

f: Output based on grid squares

Figure 3.6 Illustration of alternative output geographies at a microscale

3.5.1 Individual addresses

Individual address references such as those provided by ADDRESS-POINT are illustrated in Figure 3.6b. Although the individual address provides an important basic component from which areal aggregations may be constructed, it seems unlikely that any set of zones much smaller than 1991 EDs would provide an acceptable confidentiality threshold for output without extensive suppression. The relationship between postal addresses and household counts from the census will vary considerably between different types of neighbourhood, with a general one-to-one correspondence likely in outer suburban areas, an excess of households in inner urban areas where many properties are subdivided, and possibly an excess of properties over households in some areas where second or vacation homes comprise a significant proportion of the dwelling stock. Address referencing forms the basis for the functioning of many organizations who are current or potential census users, and a difficulty with the 1991 output has been the lack of any readily available means of placing individual addresses unambiguously within EDs for the majority of users without access to ADDRESS-POINT and digital ED boundaries. Not only has the use of address-level applications increased during the 1990s, but may be expected to grow still further beyond 2001. There would therefore be considerable demand for any product which allowed the automatic assignment of addresses to one or more of the standard output geographies. This might take the form of a detailed look up table or address list, as discussed earlier. For any geography based on newly designed EDs, this would be complex, but for aggregations based around address components such as postcodes (as performed in Scotland in 1991), assignment could be a relatively straightforward hierarchical process. The need for such assignment is of sufficient importance to be a relevant consideration in the design of output geographies.

3.5.2 Unit postcodes

Unit postcodes are the smallest division of the postal geography, and typically contain around 14 postal addresses. There are around 1.7 million unit postcodes, and these are the most widely used and understood small area references. The postcode forms the basis for the operations of many census-using organizations, and it is therefore important that there should be some clear relationship between census output and the postcode system. In 1991 this has been provided by the ED/postcode directory, which gives the count of households falling within each unique ED/postcode intersection, and allocates each postcode into a 'pseudo-ED', representing the ED in which the majority of its population live. There is still no definitive set of areal boundaries (such as those illustrated by Figure 3.6c) for unit postcodes in England and Wales, although these are available in Scotland. The creation of such boundaries by automated means is now entirely feasible, although some important design considerations remain unresolved, such as: 'should boundaries incorporate higher level administrative boundaries or major physical features?' and 'should they cover the entire land area?'. The determination of an agreed definition in conjunction with Ordnance Survey and Royal Mail would appear to be a most urgent priority in order to ensure that decisions taken concerning census outputs are not incompatible with major developments in population georeferencing which will inevitably take place during the lifecycle of the 2001 census data. It is not possible to provide a 'perfect' direct match between postcodes and larger areal units such as parliamentary constituencies or local government areas, but the likely mismatches at the edges of such zones will be very small relative to total zone sizes and populations, and it should be carefully considered whether acceptance of such marginal errors (especially when they can be quantified at the time of zone design) is an acceptable price to pay for a more widely accessible system of basic building blocks.

The unit postcode therefore has some considerable attractions as a basic building block for other output geographies, such as that illustrated in Figure 3.6d. Postcodes provide a convenient intermediate zone for which household data may be aggregated, and then further aggregation rules may be applied in order to form groups of postcodes which correspond to higher level geographies. At the lowest level, it may be feasible to provide a small number of the most basic counts (populations, households, age structure), with further details being added at the higher levels, trading off geographical for statistical detail.

3.5.3 Census-specific output areas

For previous censuses, an ED geography has been devised primarily for the management of the enumeration process, and EDs have then been used for output. The separation of input and output geographies now has widespread acceptance, but should a new standard set of census output areas such as those shown in Figure 6.6e be defined? It may be more appropriate not to attempt to devise a completely new ED-level output geography which is unique to the census, but to focus instead on outputs for existing small areas at the lowest level and statutory units such as wards as the next 'standard' geography. At the smallest area level, there would be considerable utility in using the 1991 ED boundaries as far as possible to permit analysis of change and to avoid unnecessary effort for users in moving between one essentially arbitrary output geography and another. Where extensive new development has taken place, 1991 EDs may be subdivided to provide smaller geographical units which meet 2001 confidentiality constraints. This 'new' geography would therefore be known in advance to users, and a small 'update' of modified boundaries would allow existing users to convert to the new geography immediately. A similar range of statistical tabulations might be provided as for the 1991 SAS, again promoting comparability across the censuses. Implicitly, such a route would set up the 1991 boundaries as a 'frozen' census geography, and there might be many good arguments why it is not ideal for this purpose. Nevertheless, true comparability across time can only be established if one boundary set is taken as the starting point, and using a previous census geography does at least have the advantage of providing comparability at once, rather than waiting for the same set of issues to be discussed in ten years' time!

3.5.4 Grid squares

An alternative frozen geography is aggregation to regular grid squares such as those used in 1971, and illustrated by Figure 3.6f. These have the advantages of permanence and the absence of digital boundaries, but suffer from high rates of suppression due to the wide range in population densities, and the inability to recombine them into any other standard zonal geography. Such an aggregate geography would be very easy to produce from an ADDRESS-POINT database, and should perhaps be reinstated as a geography for intercensal comparison, avoiding many of the interpretation problems inherent in the irregular zonal systems.

3.6 British standard 7666

The new British Standard 7666 (Cushnie, 1994) defines a national standard for address referencing in three parts: part 1 deals with the specification of a national street gazetteer, part 2 deals with a land and property gazetteer and part 3 concerns property addressing. Each of these aspects has relevance to the 2001 census. A street gazetteer may offer one of the simplest methods for the allocation of address-based information into census geographies, which is not dependent on postal geography (which tends to change more often). The use of the standard for any address-based products from the census is to be commended in so far as it will help to cement a common standard for address referencing and avoid many of the existing frustrations when attempting to integrate address-level datasets from different sources.

A full BS7666 address includes a 'locality' identifier, but there is as yet no clear definition of what exactly comprises a locality, and initial trial implementations of the standard such as that in the City of Bristol described by Yeoman (1995) have tended to use either ad hoc neighbourhoods or existing (1991) census wards. It would seem particularly important that the likely implications of BS7666 are therefore taken into account in the design of 2001 output geography: perhaps the census provides the ideal basis from which to derive a national set of localities which are aggregations of postcodes or other output areas.

3.7 Conclusions

In this chapter we have identified a number of principles for the design of census output geography - some of which are widely accepted, and some of which are still the subject of debate. The latter are the matching of standard geographies to user needs, the availability of user-defined flexible geographies, and the approximation principle. We have argued in favour of multiple standard geographies, but emphasis should be placed on integration with existing georeferencing systems rather than on the derivation of an entirely new scheme for 2001. The postcode system has major attractions at the small area level, in so far as it is readily integrated with address- and postcode-based systems already in use, and provides a national system of small, widely understood basic spatial units. It is assumed that there will be a requirement to produce aggregate data for larger statutory areas, and this may be performed either by approximate aggregations of the postcode geography with known error characteristics, or as separate aggregations from the household data. Comparability with past geographies is a very important aspect, and it is suggested that the a modified 1991 geography be used as far as possible, rather than the creation of a new 2001 small area geography which is unrelated to any of the existing systems. In addition, the grid-based aggregations offer a more abstract stable geography for long-term intercensal comparison, which would be relatively simple to produce from the detailed 2001 database.

Whichever standard geographies are adopted, a most important new product will be a set of comprehensive look up tables which show the interrelationship between the areal units used (degree of overlaps, populations involved). This should certainly incorporate the type of information currently provided in the ED/postcode directory, but with the valuable addition of a new higher resolution centroid for each postcode (or other basic building block), which could be derived from ADDRESS-POINT, and would become an important dataset in its own right. Such a directory might be linked to an address register or street gazetteer providing allocations of recognizable spatial features into the standard aggregate geographies. Whatever action is taken in this direction, it is imperative that current developments in georeferencing standards are taken into account in order to ensure that the census output geographies provide the lead, rather than adding to the confusion concerning spatial referencing.

Standard geographies will not satisfy all user needs, and no geography is likely to remain an unaltered 'standard' for an entire intercensal period. We therefore see a need for a mechanism for the creation of user-defined geographies, and draw the distinction between data published for standard geographies which may be considered *safe data*, and an intelligent system for tabulation of *data in a safe setting* using user-defined geographies. A prototype protection system has been presented which provides a framework for overcoming the potential differencing problem created by the existence of multiple output geographies. The implementation of such a system is the subject of current research.

4. THE GEOGRAPHY OF OUTPUTS FROM THE 2001 CENSUS

Frank Thomas, June 1996

This paper was issued as Paper 16 of the Output Working Group for consideration by the Planning Board of the Census Offices. An initial version had been presented at the second meeting of the Output Working Group on the 22 April 1996 at the Department of Health, Skipton House, Elephant and Castle, London. The paper responds to the proposals made in the preceding two papers.

4.1 Introduction

A number of key users of Census output have proposed that the output geography for 2001 be more 'flexible' than for 1991. A flexible output geography would for any given output entail

- identifying areas smaller than those for which the output would normally be released; these areas could in principle, be single households or communal establishments,
- aggregating these areas to various sets of 'output areas' to match each user's geography as closely as possible.

The notion of flexible geography, when examined, is, in effect, a double geography with the given output available for two area type but with that for the smaller area type heavily modified. The alternative is to produce a simpler output for the lower area, thus creating the beginnings of a hierarchy of combinations of outputs and area types. The output geography for 1991 (both in England and Wales and in Scotland) can largely be classified as belonging to the second (hierarchical) category.

Neither option achieves the ultimate goal of giving the user 'any output for any areas'. What flexible geography offers that hierarchical geography does not is a version of the output in question for building blocks that are, strictly speaking, below the threshold set for the output. A second threshold must be set for these building blocks. This increase in output available has a price

- the output for building blocks has to be heavily modified, and flexibility at the level of the single address is unachievable,
- devising a modification scheme to protect a single output will be difficult, to produce one that protects several will be highly complex,
- the modification needed has to be done after tabulation, so it will be impossible to remove inconsistencies between tables and areas,
- critics of the Census could seize on the second (building block) threshold as crucial rather than the original threshold.

While the Census Offices should do all they can to maximise the use of the Census (and increase revenue) the benefits of flexible geography are illusory. The Policy board is recommended to reject the proposal for flexible geography, as it stands, and agree instead that the Census Offices provide output for an extended hierarchy of combinations of area type and output.

4.2 Summary

The rest of the paper gives an account of the principles that govern the output geography for the 2001 Census. these conflict with each other but have, on the whole, been accepted by both suppliers and users in discussion since the last Census in 1991. The principles cover issues of

- Confidentiality and non-disclosure
- Thresholds, statistical detail and disclosure control: the need for a threshold for a given specification of a table, set of tables or other output; how thresholds may be breached by differencing; the various ways in which modification and other methods of preventing disclosure may be done
- Separation from input geography
- Consistency of output geography throughout the UK

- Precision: what do users want, can approximations be acceptable, should there be consistency in the outputs produced
- The Census as a GIS

Apart from the above principles, the choice of output geography for 2001 will depend on cost and practicalities. Practicalities are dealt with so far as possible. Possible technical developments for 2001 are briefly described with more detail in Annex C. The two main options - flexible geography and hierarchical geography - are described and an assessment made of each against the principles given earlier. The paper concludes by stating what may be done next to refine the needs of users for output geography following whatever choice of option is made.

4.3 Origin of paper and acknowledgements

The paper contains much of the material in the one (OWG 9) prepared by Phil Rees *[paper 2 in this collection]*. A first version of the paper was presented by Frank Thomas as OWG 12 at the meeting of OWG on 22 April. Additional material has its origins in papers written within the Census Offices. This paper is solely the responsibility of its author.

4.4 Principles

The principles in Table 1 have in the main come to be accepted in the Census supplier and user communities from critical evaluation of the strengths and weaknesses of the approaches adopted in the 1971, 1981 and 1991 Censuses. In this section, each principle is spelt out and comments are made on the precedents for each principle in past UK Census practice. Comments are also made, where appropriate, on how a principle may be developed for 2001. It is a matter of judgement how much weight is given to any of these principles relative to the others.

4.4.1 Confidentiality and non-disclosure - principles (1) and (2)

Principle (1) preserving confidentiality, is a duty of the Census Offices with respect to all statistical abstracts. In the past a great deal of caution in producing outputs was exercised because the risks of disclosure were not well understood. One of the goals of the Census Development Programme and associated academic projects should be to understand what the risks of disclosure are when using methods to prevent them.

Principle (2), that small area data need protection, follows from the first principle. A variety of methods have been used to protect data but these have not, until recently, been systematically evaluated and compared in terms of their costs and benefits.

4.4.2 Threshold, statistical detail and disclosure control - principle (3)

If there were no need to adhere to the *confidentiality* and *non-disclosure* principles then output geography would reduce the technicalities of assigning Census records to the users' areas of interest. However, output geography as determined by thresholds is one of the means used to prevent disclosure. It is suitable choice of threshold, statistical detail and degree of modification that achieves the required balance between utility of the output and prevention of disclosure.

4.4.3 Thresholding

The primary protection of the individual or household data is that areas for which a given output is to be produced should contain more than a minimum number of people or households. Thresholding is an easily explained technique and the need for it is generally accepted by users of census data. In the 1991 Census, the threshold size for Small Area Statistics (SAS) was set at 50 persons and 16 households. If enumeration districts (England and Wales, Northern Ireland) did not meet these thresholds they were amalgamated with neighbouring areas. In Scotland, output areas were designed to meet the thresholds so this slight difficulty did not occur. A similar approach was used for Local Base Statistics (LBS) for wards in England and Wales and postcode sectors in Scotland. The

Table 4.1: The principles governing output geography

	Principle	Explanation
(1)	Confidentiality	The outputs associated with each geography should not compromise the confidentiality of the underlying individual and household records
(2)	Non-disclosure	There will need to be protection measures applied to counts for small areas that will ensure that information about individuals and households is not systematically disclosed. Examples of such measures are thresholds and modification of statistics.
(3)	Thresholds, statistical detail and modification	A given output can only be provided for areas with at least a given population; for the smallest areas fewer counts can be provided; for larger areas more counts can be output. Statistics for small areas also need to be modified.
(4)	Separation of input and output areas	The geography of input areas (for collection of the Census) and the geography of outputs areas (for publication of the Census) should be regarded as separate.
(5)	Consistent geographies	At least one consistently defined geography should be available for all parts of the United Kingdom.
(6)	Matching geographies to user needs	There can be no single geography which fully satisfies all users. The corollary is that several different geographies may need to be produced.
(7)	User-defined geographies	There should be systems for generating small area geographies that can be tailored to specific purposes (including any unknown at the time of the Census).
(8)	Approximation	Output must fit users' areas to within an agreed tolerance.
(9)	Consistent tables	Tables that have been modified must, as far as possible, present statistics that are consistent for different tables for an area and for an area and its constituent smaller areas.
(10)	Delivering the Census as a GIS	The Small Area Statistics should be delivered simultaneously with the appropriate digital boundary data either as an integrated geographical information system or as suitable input to such a system.

thresholds for LBS were set at 1000 persons and 320 households and wards and sectors were merged if necessary to meet thresholds. An alternative used in 1981 to merging below-threshold areas was to suppress their output (together with any equivalent cells for higher areas which would otherwise allow cells for the blanked area to be found by differencing). Both merging and suppression have drawbacks. Suppression leaves gaps in the output geography while merging creates output areas that may straddle boundaries that the unmerged areas may have been designed to respect.

4.4.4 Table specification

A measure complementary to thresholding is to vary the degree of detail available for areas depending on the population they contained. The smaller the area the fewer counts that could be produced the larger the area the more counts that could be produced. The choice of threshold is a function of the detail of the variables within a table.

4.4.5 The differencing problem

Thresholding alone is insufficient as a protection principle if the statistics in question are being produced for more that one set of geographies and the thresholds applied independently to each. It may be possible to derive the same statistics for areas that are generated from the original geographies and that are below threshold. This has been termed the **differencing** problem.

The simplest case of differencing arises when a pair of areas from different geographies differs by a population below the threshold for the output in question. More complex cases arise when comparisons are between a group of areas from one geography and a group from the other. There are even more subtle risks. For examples of differencing with two overlapping geographies see Annex B. If an area for which statistics can be deduced is below threshold then we violate Principle (3). The more zero cells there are in the output, the greater this risk becomes.

For the 1991 Census in England and Wales, Enumeration Districts (EDs) used for output were not exact aggregations of postcodes and there was a proposal to produce SAS for aggregations of postcodes that best approximated to each ED. The proposal was rejected because the risk was too great that SAS could be derived for below threshold areas formed by differencing EDs and their postcode approximations (see Census Newsletter No 15, 18 February 1991).

4.4.6 Modification of cells in output

Another method used in 1991 to prevent disclosure was the modification of counts after tabulation in the LBS and SAS for areas below local authority district levels. Cells in tables were perturbed by the quasi-random addition of -1, 0 or 1. In 1991, the large number of cells in the LBS was recognised as leading to an increased risk of disclosure and, therefore, the modification process was applied twice. Modification was only used where counts were based on 100% processed data. Tables produced on the 10% data presented less of a disclosure risk because the identity of the sampled population was not disclosed.

For 2001, if the output for smaller areas is aggregated into larger areas (as described later), it may be necessary to divide methods of post-tabulation modification into those done before aggregation and those done after.

4.4.7 Alternative methods of modification

Post-tabulation measures can include rounding to the nearest 5 or 10, and the suppression of low counts. Measures undertaken *prior* to tabulation include the swapping of records between similar cases in different areas, undertaking additional imputation, and switching data for individual records. *Pre-tabulation modification* offers the benefit that outputs will be consistent.

4.4.8 Other methods of preventing disclosure

In addition to thresholding and modification, other methods of protecting confidentiality have been suggested and are used elsewhere. These include reporting ratio variables or derived indicators rather than counts. This should offer protection for larger areas, but in smaller areas e.g. with fewer than 100 households, ratios expressed as percentages are just as disclosing as absolute values.

There is also a measure of protection given by the errors created in any large scale data collection exercise such as the Census. The types of error are listed below. There is also the decay of relevance of data over time as a result of population change. However it is likely that interest in confidentiality and non-disclosure will increase at the time of the 2001 Census. The more static members of the population may feel their 10 year old data is still very relevant should it be disclosed.

Other methods are given in the report referred to in Annex C.

4.4.9 Separation from input geography - principle (4)

Principle (4) was adopted in Scotland for all SAS outputs in 1991, where output areas (OAs) were designed as aggregations of unit postcodes that could match 1981 enumeration districts - the areas that were used for that Census. Also, in both England and Wales and Northern Ireland, SAS outputs were produced that were not based on the collection area: these areas were postcode sectors in both countries and also grid squares in Northern Ireland.

To increase the range from which output geographies can be chosen, the Census Offices accept Principle (4) i.e. the separation of collection and output areas.

If any Census output is based on a sample, selecting the sample should take account of the output geography rather than the input geography.

4.4.10 Consistency of output geography throughout the UK - principle (5)

Principle (5) calls for a common, agreed approach by each of the UK's Census Offices. Because the Census has been carried out by 3 different agencies, there have been difficulties in comparing topics across England and Wales, Scotland and Northern Ireland. The 1991 Census saw considerable progress in Northern Ireland where the joint Census Office, ESRC and Queen's University project has produced SAS for EDs in the province, with geography defined on the same basis as in England and Wales. SAS for Scotland was produced for the postcode-based OA but could be aggregated to 1981 EDs (also postcode-based).

The Census Offices have agreed that there should be at least one geography common throughout the UK. This in turn calls for agreement among users on the basis of the common geography.

4.4.11 Precision - principles (6), (7), (8) and (9): user needs

Principle (6), of matching geographies to user needs, has long been enshrined in Census Office practice and has been confirmed in the user debriefing (1991 Census) and user consultation (2001 Census) meetings held by the Census offices over the 1994-95. Each user community has its preferred small area geography.

- Output has been produced for *statutory* areas. In particular, in the 1991 Census for England and Wales and Northern Ireland small area data were produced exactly for electoral wards. In Scotland SAS for wards were produced on a best fit basis by aggregating data for Output Areas. SAS and LBS were produced exactly for Scottish local authorities (LAs) by means of splitting any postcodes that straddled LA boundaries. The postcode-based OAs could then be contrived to nest within Las. For 2001, local and central government state a requirement for outputs for *statutory* areas: local government districts, electoral wards and Parliamentary constituencies.
- Output has also been produced for *postal geography*. The smallest areas for which SAS tables were produced in Scotland were Output Areas (OAs) based on aggregations of unit postcodes. Except where mergers to achieve thresholds produce OAs that straddle boundaries, OAs sum to postal sectors, postal districts and postal areas. SAS data for postal sectors were produced in

England and Wales in addition to that made available for wards. For 2001, business and marketing agencies will continue to need output for areas built from *postcodes*. A great many databases of non-Census data hold postcoded information which needs to be linked and related to Census variables. This requires use of small areas based on postcodes. The need is also paramount in applications in some of the public sector (e.g. the Health Service), and in medical research.

- SAS for *grid squares* was produced in Great Britain from the 1971 Census and a limited number of small area counts can, in principle, be produced from the 1971, 1981 and 1991 Censuses of Northern Ireland. Grid squares lend themselves to mapping, to linkage to environmental variables and to comparisons over time because the grid framework is time-independent. No grid square output was produced from the 1991 Census of Great Britain but Bracken and Martin (1995) created a system for estimating grid data from enumeration district SAS from the 1981 and 1991 Censuses which enables researchers to make comparisons over time.
- Output has also been geared to the need to *compare one Census with its predecessor*. In previous censuses, EDs that were unchanged between Censuses were identified as such in the SAS. However, a large percentage of both EDs and wards changed their boundaries between Censuses. In England and Wales and in Northern Ireland comparable areas can only be formed by aggregating EDs (the 1971-81 Census Tracts) or by carrying out estimation to common units (Bracken and Martin 1995 or Dorling 1995). As mentioned earlier, the Output Areas in Scotland in 1991 were designed to be aggregated easily to form equivalents of the enumeration districts in the 1981 Census.

Areas that don't match administrative, postal or grid geographies or those of previous Censuses will no doubt be specified and output requested for them. *Principle (7)* implies that there should be a system for supplying geographies which cannot be known at the time the output system is set up. In the past, this need has generally been met by providing SAS for areas small enough to be aggregated into the areas needed by the user and partly by the Census Offices re-aggregating the output required.

- One example of an unknown geography is that used for a *future* Census (or equivalent). Boundaries of the areas used in one Census are liable to have been changed considerably by the time of the next one. The aim here might be either to define the output geography for 2001 in terms of units that change little or are small enough to aggregated not too imprecisely to the areas used for the output of statistics from future Censuses.
- Another example of a geography unknown at the time of the census may arise when administrative areas are later revised and statistics required for the new areas. Examples where such needs were met are production of 1981 SAS for wards created in 1983 in Scotland and of 22 of the 1991 LBS tables for new local authorities in Wales and Scotland.

4.4.12 Approximation

Principle (8) embodies the notion that there are various sources of error in the Census but nevertheless for many purposes the output from it is good enough for the purposes to which it is put. Although these sources of error may help in providing protection against disclosure, the aim must be to reduce them. Paper OWG 4, *Confidentiality and Geography*, Andy Teague, OPCS, gave some estimates for sources of error in the 1991 Census:

Persons missed entirely	2 per cent
People imputed	1.5 per cent
Wrong answers to questions	1 to 2 per cent
Missing answers	0.5 to 1 per cent
Validation errors	0.5 to 1 per cent

Against this must be set a figure of 2.4 per cent for the error in using in Scotland aggregates of OAs to wards instead of exact statistics. (Actually the comparison is between OA aggregations and *postcode* aggregations; it is impossible precisely to estimate the difference between OA aggregations and exact fit. The percentage quoted is the average difference between the 2 aggregations for each of the 1158 wards in Scotland at the time of the Census.) It should be possible to reduce this error even with OAs of the 1991 variety with more preparation of the geographic database to reduce the number of OAs straddling ward boundaries.

A further source of noise is modification. The modification applied to 1991 SAS for each constituent area accumulates in the resulting totals. Unlike the 1981 practice, modification was not 'self-cancelling' as areas are aggregated. Taking 1991 SAS Table 2 for an aggregation of 40 OAs (equivalent to a ward roughly), the total for the table will be about 4,300 persons with a 95% confidence limit of ±60 (1.4%). For an individual cell the average value is 60 with a 95% confidence interval of ±6 (10%). These errors can be reduced if independently modified SAS for areas larger than OAs are used in the aggregation. Combining SAS cells to produce a proportion is also problematic. Estimates of male unemployment for an aggregation of 6 EDs can vary with a 95% confidence interval of ±8% on an estimate of 26% (calculations supplied by a user in Kingston upon Hull City Council).

4.4.13 Consistency

Counts occurring more than once within a table or set of tables for an area can be inconsistent. In 1991, principle (9) was partly followed in that the LBS and SAS underwent consistent modification for any area for which both outputs were produced. Also any areas that featured in more than one geography would be identically modified. For example, a one-ED civil parish would be an area in both the ED geography and the civil parish geography. The SAS produced for the area in each geography was contrived to be the same. The main reason for providing consistency was to prevent the risk of disclosure by comparing two sets of comparable output that had been differently modified.

Other inconsistencies between areas (e.g. between a ward and the aggregation of constituent EDs) or between tables for the same area were not removed.

4.4.14 The Census as a GIS - principle (9)

The SAS for 1991 and earlier years included a grid reference for every ED or OA. In 1991, GRO(S) produced the boundaries of OAs (and constituent postcodes) as well as Census output. Adopting principle (10) means that more should be done to produce statistical and geographic information as a single output. Most of the work of preparing the geographic output can be done between now and the release of 2001 statistical output.

The Census Offices will also examine ways of meeting the requirements of Principle (10) i.e. supplying the output of the Census as suitable for input into a GIS with both statistical and geographic output.

4.5 Possible technical developments for 2001

There are several possible technical developments relevant to output geography for the 2001 Census. A summary is given below and details in Annex C.

General developments in computing especially in Geographic Information Systems may increase the ability to 'crack' the methods adopted to prevent disclosure. There are also opportunities to develop new Census products.

Postcode boundaries and Address-Point could provide more flexible geographic output.

There is a distinction between **safe data and data in a safe setting**. The former is a product released by the Census Offices as being *safe* having had various measures taken to prevent disclosure (e.g. Samples of Anonymised Records, SARs). The latter is a dataset not released but from which output may be produced if disclosure rules are not breached (e.g. the Longitudinal Study of ONS for England and Wales).

The European Statistical Disclosure Control (SDC) programme, partly based in the UK, investigations.

The UK Standard Geographic Base is a project that, if implemented, will adopt a variety of area types as 'core spatial units'.

4.6 Options

The options for 2001 can be divided into 2 classes: flexible geography and hierarchical geography. Output geography for 1991 both for England and Wales and for Scotland fell into the latter class.

4.6.1 Flexible geography

Flexible geography is the construction of output units (OUs) to match the various needs of users by assembling smaller building blocks (BBs). Output for BBs is held in a *safe setting* from which *safe data* i.e. the output for OUs is produced. For the outputs in question (to be denoted SAS in more generalised manner of Annex A) each OU must contain at least the number of households or residents set as *thresholds* to minimise the risk of disclosure of information about individual households or residents. A BB will generally contain fewer than the threshold number of households and residents. A particular case of the BB is the household or communal establishment.

Differencing (see section 4.4..5) will reveal SAS for BBs, which must therefore have the additional protection of modification (see section 4.4.6-4.4.8). There are 3 options: modification can be carried out before or after aggregation of SAS to OUs, or both before and after.

If modification is done before aggregation only, then BBs are, in effect, OUs because SAS is revealed for them exactly as it is held in the safe setting.

If modification is done after aggregation only, the modification given to BBs is at least twice that givens to OUs. The most straightforward way in which a BB may be revealed is by a differencing of 2 OUs identical except for one BB. Assuming modification is applied by perturbing the cells in SAS, then the cells in the revealed SAS for the BB will have been perturbed twice. However, the same BB may be revealed in many different ways (by comparing many different pairs of OUs). Differences in the resulting modification patterns will reveal the true values of cells and perhaps the modification method itself.

So, modification must be carried out both before and after aggregation of BBs to OUs. The OUs will get two doses, the first to SAS for each of its constituent BBs and the second to the aggregated SAS. The revealed BB will be protected by the first dose and (at least - or as little as) twice the second applied at the OU level.

Flexible geography for SAS is therefore, in effect, a double geography - at OU and BB level. In presenting our proposals to the public, we may now have to state that the SAS was available for two types of area and give the second threshold for the BBs, but add that SAS for this smaller area is modified more.

There are implications, or requirements, for the modification scheme adopted. The second dose for an OU

- should be a function of the combination of BBs being aggregated; thereby the resulting modification will be invariant for the OU even when it appears in a second or subsequent geography;
- may also be contrived to lessen as the number of BBs in an OU increases (but this will lessen the protection given to BBs revealed by differing large OUs).

A more important requirement of the modification scheme is that, if the BBs were a lot smaller than OUs, the first dose of modification will have to be fairly 'heavy' compared with the second dose. Ideally, this modification should 'self-cancel' as BBs are aggregated to OUs. This, however, presupposes a particular aggregation i.e. a particular geography of OUs and the self-cancelling may become self-exacerbation for other geographies. Self-exacerbation will cut across one of the reasons for having a flexible geography in that one geography is 'preferred' to others. More 'noise' at the BB level also means that the real difference in SAS between two OUs that differ by one BB will be largely obscured by 'noise' - another lessening of the potential of flexible geography.

In the double geography of OU and BB, therefore, the size of the BB must for presentational and practical reasons be related to that of the OU. If the SAS were the 1991 SAS and the OUs were the size of 1991 EDs , then there would be little point in using individual addresses as BBs. (Note that the SAS for single addresses would have to be modified before aggregation.) The modification scheme required would remove any advantages of the flexibility gained. The reduction in the percentage quoted as 2.4% in section 4.4.12 would only be at the expense of increasing the 'noise' of ±8% in section 4.4.12. Also there would be no solution to the problem of inconsistencies between tables caused by post-tabulation modification.

A final point about any SAS-OU-BB combination is that if other SAS-OU-BB combinations are produced, care will have to be taken to ensure confidentiality is not breached because of near comparability in any two SASs and the corresponding sets of areas. This probably implies that the OUs in one combination should be the BBs of another combination in a series or hierarchy.

4.6.2 Hierarchical geography

An alternative approach is to make all output safe by making BBs OUs and reducing SAS for the smaller OU i.e. we have combinations SAS_1 - OU_1 and SAS_2 - OU_2. We may also stipulate that the OU_2s nest into the OU_1s. This builds on the scheme for output geography in 1991.

Modification is still required for smaller areas. First, the modification of SAS_2 for OU_2 (equivalent to BB in the 'flexible' double geography) need not be as 'heavy' as for the BB because it is not SAS_1 that is revealed but the smaller SAS_2.

Second, there is no need for post-aggregation modification because there is no aggregation. All output produced is safe so pre-tabulation modification can be considered, possibly by swapping records that are common in a set of core variables. The pairs of records for swapping could be selected using a similar methodology to that used for imputing missing data. (In fact, such swapping may not always be necessary in areas already 'modified' by 'excessive' editing and imputation.) If the swaps are kept local they will be self-cancelling in that larger OUs are more likely to contain pairs. This modification will automatically achieve consistency between SAS_1 and SAS_2 for and OU_1. (Consistency between LBS and SASS for, say, a ward was achieved with post-tabulation modification but not easily.)

If required the user would have to apportion the fuller SAS_1 to the smaller OU_2 using SAS_2.

The concept of the SAS_1-OU_1 and SAS_2-OU_2 combinations can be extended to a hierarchy of outputs and areas. In 1991, the hierarchy looked like

Threshold number of households/residents	Geography identifier (England and Wales)	Geography identifier (Scotland)	Data released
1/1	Part-postcode unit (in ED/PC directory)	Postcode unit (on Census 91 Postcode Index)	Number of households and residents
16/50	ED; and SAS Ward	OA; and SAS Sector	SAS
320/1000	LBS Ward	LBS Sector	LBS

The easiest means of extending the 1991 hierarchy would be to incorporate the postcode into the list of areas as it is the only readily available way of pushing the hierarchy into its smaller end. There are a variety of area types already used in the Census for the larger end of the hierarchy.

Table 4.2 below is given as an illustrative example only (it is not intended as a detailed proposal). The thresholds are expressed as numbers of households but could instead be numbers of residents or combinations of households and residents. Areas in the second column falling below threshold would have to be merged (rather than suppressed).

Table 4.2: A possible hierarchical output geography for 2001

Threshold number of households	Geographical identifier	Data released (cumulative)
1	Address	No. of individuals
10	Postcode	No. of households and individuals, house type, floor level of households' accommodation
20	2-3 Postcodes	Central heating; 0,1,2 or 3 or more cars
50	5 Postcodes	No. of rooms (top sliced), % tenure
100	'Output Area'	Age, marital status, long-term illness
500	'Neighbourhood'	Activity last week, education level
2,000	Ward	'Detailed' ethnic group, income
20,000	District	All variables in detail

At particular thresholds, new variables and more detail on already present variables would become available. Each variable would need to be considered for both sensitivity and its use for identifying individuals (e.g. age may not be sensitive but it may help identify individuals).

Table 4.3 below summarises how a flexible (double) geography and the two corresponding levels of a hierarchical geography measure up to the principles in section 4.4.

Table 4.3: Assessment of options against principles

	Principle	Flexible (double) geography	Two levels of a hierarchical geography
(1) (2) (3)	Confidentiality Non-disclosure Thresholds, statistical detail and modification	Gain in statistical detail at BB level	Less modification needed; simpler scheme to present with only a single threshold for SAS
(4)	Separation of input and output areas	No difference between options	No difference between options
(5)	Consistent geographies	No difference between options	No difference between options
(6) (7)	Matching geographies to user needs User-defined geographies	The same area types can feature in both options	The same area types can feature in both options
(8)	Approximation	For a given output, greater accuracy is achieved but at the expense of statistical 'noise' already (i.e. in 1991) high	Good at the expense of statistical detail; possible need to apportion to lower levels
(9)	Consistent tables	Impossible	Possible
(10)	Delivering the Census as a GIS	No difference between options	No difference between options

4.7 What next?

Both schemes of output geography described in this paper are, in effect, hierarchical. Whichever is chosen research may be needed to establish exactly what users (and prospective users) actually want in the way of precision of geography for given output relative to sources of noise including coverage. Users should be invited to prioritise their needs to allow the Census Offices to meet these where possible. To this end a sub-group of the OWG will be set up. It may be that discussions will be useful even though it is some time before topics to go on the Census form are decided. In particular, the sub-group could deliberate on the combinations of SAS-OU(?-BB) and modification may be practicable. Assumptions will have to be made about the content of the Census.

Users may also be prepared to support any measures to improve the geographic base in the UK for example any proposals to align postal and statutory geographies.

Annex 4A: Definitions

Geographies Alternative systems of aggregation of individual and household data from the Census to areas whose spatial extent is known, either as collections of point references or as the territory within known boundaries.

Small areas Areas smaller than the units used for local government (though they need not necessarily add exactly to them).

Multiple geographies The existence of more than one set of small area boundaries covering the same territory for which small area statistics are published.

Flexible geographies The ability of users to define their own geographical systems for which they can produce their own sets of statistical tables.

Small area statistics Any set of fixed tables of counts or other Census statistics. SAS may be very simple for small areas of minimum size but quite detailed for the largest areas. This extends the distinction between small area statistics and local base statistics developed very usefully in the outputs from the 1991 Census of Population for Great Britain. (The term 'SAS' is used in its more specific 1991 sense up to section 4.5.)

Address The description of the location of residential accommodation (whether for private households or in communal establishments) used by the postal service (Royal Mail).

Threshold The minimum number of households or residents or other units or combinations of households, residents, etc. for which a given output will be produced.

Differencing Deriving from statistics produced for multiple geographies at least part of the same statistics for a below-threshold small area.

Building block A small area for which statistics are produced for aggregation into larger areas.

Below-threshold building block A building block which contains fewer households or residents or both for release of the statistics produced; the area would be combined with others to form above-threshold output areas.

Pre-tabulation modification The perturbation of data in Census records before tabulation.

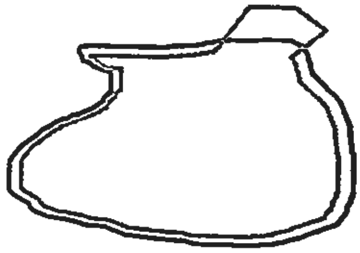
Post-tabulation modification The perturbation of output after tabulation or aggregation to areas for output; with below-threshold building blocks this concept can be further divided into pre-aggregation and post-aggregation modification - the aggregation in question being that of building blocks into output areas.

Annex 4B: The differencing problem

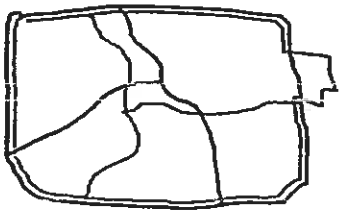
The differencing problem

In all these examples disclosure of below-threshold information is certainly or almost certainly revealed.

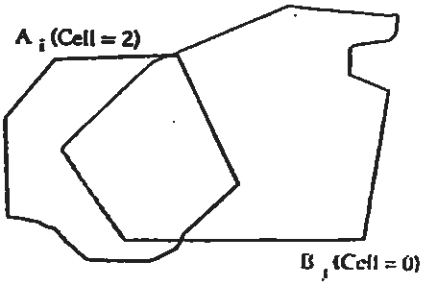
Example 1 Area in Geography A and area in Geography B differ by less than threshold



Example 2 Group of areas in Geography A and group of areas in Geography B differ by less than threshold

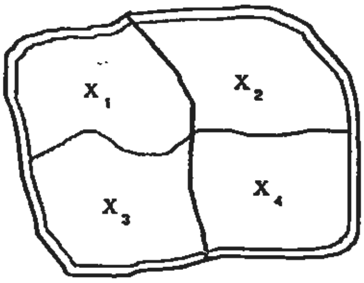


Example 3 Area A_i in Geography A overlaps area B_j in Geography B and given cell in output for A_i is non-zero but corresponding cell for B_j is zero



A second area in Geography A with the cell = 0 could slice off a bit more of the area in B

Example 4 Group of areas in Geography A and group of areas in Geography B cover same territory and each area is just above threshold



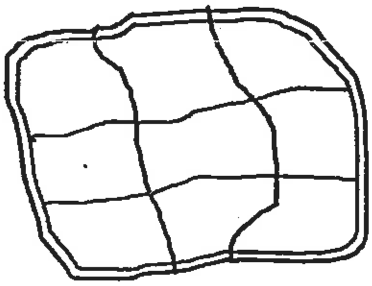
There are 4 equations in 4 unknowns

$$x_1 + x_2 = c_1$$
$$x_3 + x_4 = c_2$$
$$x_1 + x_3 = c_3$$
$$x_2 + x_4 = c_4$$

but the 4th can be derived from the other three so the set cannot be solved uniquely, unless a third geography were to give, say

$$x_2 + x_3 = c_5$$

Example 5 If there were 3 areas from each geography covering the same territory then there would 6 equations (less one as above) for possibly 9 unknowns.



Number of equations increases when a c_i is zero: equation with c_i on right side of equation is replaced by $x_j = 0$ for each x_j on left side.

A third geography would also give more equations.

Annex 4C: Technical developments

4C.1 General developments in computing

The demand for multiple small geographies, combined with the increasing availability of powerful computers, leads to an increased risk of indirect disclosure through differencing tables for slightly different areas.

4C.2 Geographical Information Systems

Unforeseen changes in GIS and IT by 2001 (remember in 1985 Desk-top PCs were rare) means many of the issues in this paper may have to be revisited. Developments in GIS have many implications for the presentation of Census data and confidentiality.

If GIS graphics became universally available on PCs and such output was standard, then presentations may involve defining a land area as a pixel and colour coding across the spectrum of the percentage of the population having a particular characteristic in that pixel. Confidentiality may be protected by statistical smoothing of pixel data and presentation colours, with the geographical detail restricted by the area level represented by each pixel.

Although developments in GIS technology are likely to offer alternative ways of presenting data which avoid certain disclosure problems, there will almost certainly still be a demand for data to be available in tables produced by the Census Offices.

4C.3 Using postcode boundaries and Address-Point

The view of some users, that Census Offices should provide more flexible geographic output, stems in part from the perceived advantages of newly available geographic products such as Ordnance Survey's Address-Point, and advances in geographical information systems.

Address-Point's primary purpose within the Census operation in England and Wales is not directly to enhance output, but to help provide a more cost-effective and efficient method of planning collection areas and providing enumerators with address lists. However, to improve output, it may be possible to derive Thiessen polygons around postcodes to reflect underlying geography (using one of the OS digital mapping products) and use these to derive polygons that are aggregations of neighbouring postcodes.

GRO(S) already has digitised Scottish postcode boundaries, and so has not bought Address-Point. However GRO(S) is interested in the automatic production of enumeration address lists, and may use a product based on the Royal Mail's Address Manager, although the possibility of using Local Authority Council Tax valuation lists is being investigated.

Census products that use Address-Point may cost more than any similar products that don't.

4C.4 Safe data versus data in a safe setting

Dale and Marsh in *The 1991 Census User's Guide* (HMSO) have made a very useful distinction in the context of Census microdata, between safe data and data in a safe setting. *Safe data* are individual records sufficiently protected by anonymity, sampling, broad coding and legal conditions that they can be released to users. The Samples of Anonymised Records from the 1991 Census constitute such a safe dataset. *Data in a safe setting* are individual records in their original fully coded form from which statistical tables can be produced, the production of which can only occur in a safe setting (in a secure computer designated by the Census Offices). Security procedures include the vetting of researchers with access to the data, the logging and monitoring of table requests and the checking of table outputs. The Longitudinal Study of OPCS (England and Wales) is data in a safe setting. Another example of data produced in a safe setting is the production of special tabulations to order by the Census Offices.

Such a distinction can be applied to counts and tables for geographical areas. Small area statistics for a standard geography should be released as safe data. Counts or tables for user-defined geographies could be data produced in a safe setting. Production of data in a safe setting is currently a very expensive operation. Census Office statisticians must check outputs and only specialist staff can run the programs to produce tables. There would have to be considerable automation of the whole process through the development of security software in order to reduce costs. The tabulation production should be a task for users to do. The security software would make decisions about its release or non-release from the safe setting and making reports to Census Office personnel in borderline cases.

The development of an automatic safe-setting system from which to produce outputs would be difficult, but would have to be on the following lines:

- A very fast database/tabulation package is needed. Open tender to software firms, academic centres and other national statistical agencies should yield a suitable product.
- A safe sample dataset (anonymised and heavily perturbed in lots of different ways) should be released to users so that they develop their tabulation jobs using the fast database/tabulation package. The safe sample dataset will differ from the SAR in containing the full classifications of each variable that will be used in the safe setting.
- A package that designs the zones that meets user needs and links to the fast tabulation package, reporting the risks involved in release of the results for vetting by the Census Offices.

4C.5 ARGUS

The ARGUS-based projects which are currently beginning may offer an improved means of highlighting cells which pose a disclosure risk in tables providing an automatic or semi-automatic process for altering coding or modifying cells. Other systems developed as part of this may help to identify records which are rare or unique, allowing these to be 'modified', recoded, or suppressed.

4C.6 GSS Task Force on Disclosure

The Government Statistical Service Methods Committee has set up a Task Force on Disclosure. The Task Force produced a first report in December 1995. Further work is recommended and at least one further report may be produced.

4C.7 UK Standard Geographic Base

Output geography for the 2001 Census will need to take account of a report that will appear shortly containing recommendations for a set of 'core spatial units'. This set includes both wards and postcodes.

5. THE GEOGRAPHY OF OUTPUTS FROM THE 2001 CENSUS: OUTPUT UNITS AND LOOK UP TABLES

Phil Rees, David Martin and Oliver Duke-Williams, September 1996

This paper was prepared in response to the Thomas proposals in the previous paper. It has had limited circulation within the Census Offices but has not yet been presented at an Output Working Group meeting. The paper takes the debate about the nature of output units to be used in the 2001 Census following on from Output Working Group paper OWG9 (Rees 1995), Output Working Group OWG12 (Thomas 1996a), Rees and Martin 1996 and Output Working Group paper OWG16 (Thomas 1996b). It accepts the principles as agreed in paper OWG12. The paper comments on the discussion in OWG 16 on the assessment of the "Flexible Geography" and "Hierarchical Geography" options and puts forward some practical proposals for the production of Small Area Statistics.

5.1 The debate to date

5.1.1 The request for four output geographies

It was suggested in Rees (1995) and in Rees and Martin (1996) that users wanted 2001 SAS output for four types of output units (OUs):

(i) OUs that sum exactly to administrative geographies (2001 wards and districts)
(ii) OUs that sum exactly to postal geographies (postal sectors in 2001)
(iii) OUs that are part of a regular raster geography (grid squares defined by the OS National Grid)
(iv) OUs that match areas used in previous censuses (1981, 1991).

5.1.2 Four geographies OK for larger small areas?

Precedent suggests that SAS can be provided for "large" small areas for each of these geographies. SAS for wards and postal sectors were both provided in 1991. Kilometre square SAS were provided with ED SAS in Great Britain in 1971 and in Northern Ireland in 1991. In 1974-75 Census tables were republished for new districts. SAS were provided for Census Tracts that connected 1971 and 1981 EDs in England and Wales. In 1991 SAS were provided for Output Areas in Scotland that can be summed to 1981 EDs. Confidentiality of these statistics is protected for all areas smaller than districts by random blurring of cell values in SAS Tables.

5.1.3 The Thomas arguments

The issue that therefore divides users from the Census Offices is whether SAS can be provided for "small" small areas for these four geographies.

In OWG 16 Thomas (1996b) argues this would compromise confidentiality. Users could difference two or more SAS geographies and therefore there would have to be heavy modification of outputs (double blurring) and heavy restrictions on the degree of detail supplied in such multiple SAS. Thomas argues that a hierarchical system similar to that used in 1991 but common across the whole UK would make possible the production of several versions of the SAS with less detailed tables for smaller areas and more detailed tables for larger areas.

Thomas also argues that pre-modification of OU records (by swapping records or part records or by modifying records or part records) would be preferable to modifying output tables. The main gain would be that all OU tables in the hierarchy would be consistent, at least up to district scale. This argument for pre-tabulation modification has been put both by the Census Office statisticians - Teague, Dixie - and by academic researchers - Cole, Martin.

We comment on these arguments and make some further points.

5.1.4 The need for SAS for four geographies

Although we feel sure such a need for the four kinds of OUs is strong, there is variation and disagreement between user groups as to whether the SAS needs always to be exact (exact fit of building blocks into the areas).

The principal group demanding "exact" statistics are local government representatives in England and Wales, who are very conscious of the role of Census statistics in resource allocation formulae used by the Department of the Environment in deciding grants. Experiments by Martin have indicated that summation of census records by postcode unit to electoral wards on a best fit basis give totals very close to those derived by adding up enumeration district numbers. By best fit is meant the assignment of each postcode unit to the ward it falls inside either wholly or partially, choosing the ward the largest part falls in.

Academic researchers would certainly settle for good approximations to three of the four geographies. A systematic agreed programme for producing postal, grid and historical geographies, even if approximate, would be a considerable improvement on past situations.

5.1.5 New areas, new dangers

The dangers of differencing are set out in OWG 16 as theoretical possibilities. The jury, however, we would argue is still out in terms of deciding whether the empirical occurrence of differenced areas constitute a real threat to the confidentiality of Census data.

Rees and Duke-Williams (1996) are currently carrying out experiments using 1991 Census data to discover whether the fears are real or not. The experiments involve

(i) creation of a synthetic census population of individuals and households for Yorkshire and Humberside by factoring up the private household population of the 1% SAR and the communal establishment population in the 2% SAR;

(ii) random assignment of these private households and communal establishment rolls to enumeration districts in Yorkshire and Humberside using the relevant SAS counts as totals; and

(iii) random assignment of exact locations within EDs using a routine incorporating an efficient point-in-polygon algorithm distributing households around ED centroids within ED boundaries.

This synthetic population can be counted for any non-ED geography, the boundaries of which are known. In the experiment, the postal sector geography and the ED geography are overlain, and sets of differenced areas generated. The population and household counts for differenced areas have the distribution shown in Table 5.1.

Table 5.1: The distribution of population and household counts for differenced areas formed by overlapping EDs and postal sectors

Population in differenced areas	Number of areas	% of areas (without error)	Number of households in differenced areas	Number of areas	% of areas (without error)
200+	1015	100	48+	1015	100
100-199	0	0	32-47	0	0
50-99	0	0	16-31	0	0
<50	0	0	<16	0	0
Error	1	-	Error	1	-
Total	1016	100	Total	1016	100

This table shows that differencing *small* areas (EDs) and *medium* sized areas (postal sectors) does not produce areas for which SAS tables can be derived which are sub-threshold. The Census Offices can be sure that confidentiality will not be violated if they supply SAS from the 2001 Census for both ED-sized OUs and postal sectors again, as in 1991. Rees and Duke-Williams intend to extend these experiments by overlapping EDs and grid squares of small size to investigate what happens when two sets of *small* areas are overlapped.

5.1.6 SAS for different levels in standard hierarchy

Thomas (1996b, OWG15) makes a number of suggestions about how SAS output would vary with different thresholds. This is a useful extension of 1991 practice but probably needs to be focused now on practical proposals.

5.1.7 Modification of records before tabulation

This is an attractive idea but considerable experimentation is needed to make sure that any scheme does not distort the SAS tables derived as a consequence to a significant degree. This can be done with the Yorkshire and Humberside experimental system of Rees and Duke-Williams (1996).

5.2 Building blocks and output units

There is still a great deal of confusion about the role of building blocks and their relation to output units which must be removed.

To make progress, let us accept Thomas's proposal that just one set of output units (OUs) be created. What should be their characteristics and how should they be designed? Here output units are defined as areas for which SAS will be published. They nest into larger areas such as wards or postal sectors. Building blocks (BBs) are fundamental geographic units from OUs are built. The suggestion in Thomas (1996b) is that the BB should be the postcode unit (PC) rather than the part-postcode unit (PPU) created in the 1991 Census in England and Wales in the ED/PC directory. The proposal is that the errors in summing unit postcodes to wards and districts are very small indeed and can be ignored in practice. Accepting this argument, Figure 5.1 shows the advantage of such BBs: they sum to areas in the administrative hierarchy (wards, unitary authorities/local government districts, statistical regions) and in the postal hierarchy (postal sectors, postal districts and postal areas).

Note that the PCs are collections of addresses and codes rather than territorial areas. The territory around PCs must defined in a separate operation through the recognition of property parcels, through a digitising operation or through a GIS operation such Thiessen polygon creation constrained to include all household locations assigned to a PC. It is not necessary to build SAS tables for these BBs: all that it is necessary to do is to make sure that all enumeration entities in the census (private households, communal establishments) can be labelled by the postcode unit of their residential address and the administrative area that the address falls in (ward, unitary authority/local government district). It is necessary to assign an Ordnance Survey National Grid reference to the address so that aggregations to grid squares can be made. It is also necessary to add to the enumeration record the geocodes of any area that cannot be deduced from/linked to the PC.

A key issue is the accuracy and cost of this census address base. The address base is traditionally built up by using previous census lists, the Royal Mail Central Postcode Directory, enumerators returns (which record any missing addresses within assigned territories) and assignment of centroids at the time of the census. An alternative base that is currently under development is Ordnance Survey's Address Point product which supplies the address centroid (as opposed to the PC centroid). The Address Point product is being used in the design of collection areas for the 2001 Census, but its use in census output is currently not planned. Users have argued strongly for using this new product. The Local Government community has negotiated a Service Level Agreement for access to OS digital products including Address Point. The academic community (led by the Joint Information Systems Committee, supported by ESRC and other research councils) are currently conducting similar negotiations. Could not these licences somehow allow the Census Offices to supply users with the more accurate information they may well have already paid for? However, decisions about the address base of the census do not affect the subsequent arguments in this paper.

Figure 5.1: The standard hierarchies of census areas

Administrative Hierarchy	Postal Hierarchy	(Small) Area Statistics Sets
postcode units		(S)AS0
output units		(S)AS1
wards	postal sectors	(S)AS2
unitary authorities/LGDs		(S)AS3
	postal districts	(S)AS3
	postal areas	(S)AS3
regions		(S)AS3
countries		(S)AS4
United Kingdom		(S)AS4

In Figure 5.1 the administrative and postal hierarchies are set out. Associated with the various spatial levels are sets of Small Area Statistics (SAS), the detail of which increases from small areas (Level 0) to the largest (Level 4). Note that the term *Small* is rather confusing (particularly to students) when the statistics apply to large areas. Why don't we just drop the adjective and call them just *Area Statistics*?

Note that the Figure 5.1 proposal accepts the danger posed by differencing of OUs and postal sectors is negligible. To our knowledge no user has attempted to compute SAS for the differenced areas.

All of the units listed in Figure 5.1 are defined outside the census operation itself, with one exception, the Output Unit, a sub-ward unit. The OU is the only one that needs to be defined in association with the census process.

Note that the external units (both administrative and postal) are all subject to systematic and substantial post-census change, which needs to be planned for. We take up this issue later.

5.3 How should OUs be defined?

Several questions have to be answered here.

(i) What are the thresholds for population and households?
(ii) Who should do the design of OUs?
(iii) What data should be used in the design?
(iv) What criteria should be used?

This presents us with a classic regionalisation problem. Some tentative thoughts are as follows.

(i) *Thresholds.* In the 1991 Census a joint thresholds of 50 persons and 16 households was used. Areas with fewer people or households or both were merged with neighbouring areas. However, if

OUs are to be designed prior to the 2001 Census which seems sensible, then perhaps thresholds should be set higher than these quite low limits so that post-Census merging is unnecessary. However, if Address Point is used as the basis for the OU design, then counts of residential addresses can provide some indication of household numbers in areas of considerable change.

(ii) *Designers.* The choices for designers of OUs are either the Census Offices centrally who could apply uniform methodology or Local Government Authorities, who have local knowledge and responsibilities. Past experience suggests that central design will ensure uniformity and consistency. It is clear, however, that the design process must be automated so that a series of computer operations will yield proposals, given a set of design goals. The computer algorithm which generates the OUs can be run and tested in advance of knowledge of exact constraints (e.g. ward or district boundaries) and then re-run when the position for 2001 is fixed.

(iii) *Data.* Any regionalisation will require definition of the operational taxonomic units (OTUs) to be classified into regions and the attributes to be used in that classification. The logic of current thinking would seem to indicate that PCs be used as OTUs and that 1991 Census attributes be used to characterise these. This implies quite a lot of reprocessing of the 1991 Census master file to generate these indicators so that practical considerations may dictate use of only simple counts of people and households. If 2001 Census data were to be used to design OUs, this would imply unacceptable delay in outputs.

(iv) *Criteria.* Some of the criteria which should govern the design of OUs are well established. Other criteria are subject to debate.

- OUs should have above threshold populations and households.
- OUs should recognise ward and local government and other statutory boundaries.
- OUs should be spatially compact and single polygons.
- OUs should recognise important physical and social barriers.
- OUs should be socioeconomically homogeneous.

The regionalisation problem is to design 100,000-300,000 Output Units within 10,000 wards from about 1,600,000 postcode units. All numbers here are approximate but give the right order of magnitude. Each ward would contain about 10-30 Output Units and each OU would be made up of about 5-16 PCs. The OUs would range in size from circa 100 to 400 people. The algorithm for regionalisation will require a contiguity matrix between PCs within wards, to ensure that only contiguous PCs are grouped together. There are several different methods for generating the contiguity matrix which Martin is currently investigating. The design of Output Units is a formidable task but would be constrained within a ward, merely being repeated across the 10,000 wards of the country. Use of software based on products such as ZDES developed by Openshaw and Rao (1995) would be advantageous.

5.4 Coping with changing geographies: existing solutions

One of the four geographies which Rees (1995) identified as important to academics and others was a geography that matched geographies used in past censuses. In fact, such a requirement turns out to be much more general. Both the geographic hierarchies set out in Figure 5.1 are subject to change after the census. In the case of wards. These changes are driven by the first-past-the -post nature of the British electoral system which is based on constituencies (wards for local elections, Parliamentary constituencies for national elections, European constituencies for the European Parliament). Electoral areas are adjusted to be of more equal size, to correct for the population shifts that have taken place: only in this way can the rough equality of electors' votes be preserved. Continuous changes occur in residential addresses through new dwelling construction and old dwelling demolition precipitate changes in unit postcodes which are lists of addresses.

A variety of solutions to the problem of matching geographies between censuses have been designed with respect to the 1971, 1981 and 1991 Censuses.

5.4.1 Use of unvarying geography

Grid square SAS was produced in 1971 for Great Britain but not in 1981 or 1991. The gap was filled by Bracken and Martin (1995) who devised procedures for assigning EDs in 1981 and 1991 to common 200 metre square units, and provided a set of converted counts. The 1971 grid square SAS is available to the UK academic community via the ESRC Data Archive at the University of Essex, while the Bracken/Martin 1981/91 counts are available via the MIDAS service of Manchester Computing (University of Manchester).

In Northern Ireland a limited number of circa 800 counts have been provided for user defined grid geographies for 1 kilometre grid square areas in rural areas and 100 metre grid squares in urban areas. These have not been purchased for academic community use because of the high price/outputs ratio.

5.4.2 Creating lowest common denominator matching areas

GRO(S) built their collection units in 1981 using unit postcodes. In 1991 output areas were also based on unit postcodes, and were matched as far as possible with 1981 EDs. Where this was not feasible, aggregations of one or both geographies could be made to form matching areas. This approach works well for perhaps 80% of the territory but leads to a lot of work where there has been considerable change in either population or postal geography.

A similar approach was adopted by OPCS to match 1971 and 1981 EDs: census tracts were defined to be the lowest common denominator areas between censuses. However, this had the disadvantage that only big common areas could be used in regions of population change.

5.4.3 Fixing on one geography and converting to it

This approach has been adopted by Census Researchers at the University of Newcastle (Coombes, Champion, Dorling - see Dorling 1995). The fixed geography adopted was the 1981 Census ward, and 1991 Census SAS for EDs in England and Wales and for Output Areas in Scotland related to the 1981 Census ward. A look up table that gives the 1981 Census ward membership of each ED and OA in the 1991 GB Census is available on the MIDAS service of Manchester Computing. The look up was defined on a best fit basis using point-in-polygon routines in ARC: the points were the 1991 ED or OA centroid (population weighted centres of gravity) and the polygons were the 1981 Census ward boundaries.

The disadvantage of this approach is that it uses a geography which becomes increasingly out of date as time passes, and works well only when the units to be aggregated from the current census are small in relation to the target geographies of the previous census.

EUROSTAT's GISCO (Geographical Information System for the Community) also adopts this frozen geography approach for its NUTS 5 database. Change from an early 1980s geography is monitored by defining the additions and subtractions of subareas over time. Such a system works well when the geography adopted is relatively stable but very difficult to support when the chosen geography is always radically revised as direct result of the census itself. This is the case for electoral divisions such as wards under the first-past-the-post system of elections.

5.4.4 Look up tables that link census and subsequent geographies

At each census the relationship between census geography and postal geography is known. SAS are published for postal areas, for output areas, which are aggregations of unit postcodes, in Scotland, for postal sectors (e.g. LS6 9, SO17 1) in England and Wales, Scotland and Northern Ireland. In England and Wales and in Northern Ireland, a look up table was created that linked the 1991 Census Enumeration District (ED) and unit postcodes (PC): the ED/PC directory.

However, the postal geography changes as the Royal Mail assigns new postcodes to new addresses or reorganises existing postcodes (as happened in Manchester in 1993 for example). The Royal Mail updates its database on addresses and postcodes frequently. ONS and GRO(S) also use Royal Mail updates to revise their Central Postcode Database and thus their ED/PC Directories (ONS) or Output Area to Postcode Index (GROS).

The ED/PC directory is a list of ED and PC intersections or overlaps for which a count of households falling in the intersection (at the 1991 Census) is provided along with the grid reference for the postcode centroid. This list can be organised in either ED or PC order. The list (which is a look up table) can be used to generate 1991 Census counts (or whole SAS) for current units based on postal geography, or can be used to sum contemporary postcoded information to match 1991 Census Eds. The number of households in each ED/PC intersection can be used as a weight in the conversion process. The ED/PC directory can be converted fairly easily into a gazetteer file for input to the aggregation routine in SASPAC, for example. Such a process makes the assumption that the distribution of households across ED/PC intersections can be used to distribute SAS counts in general.

5.5 Coping with changing geographies: a plan

It seems to us that the fourth solution to the changing geography problem is the most general - that is, we should develop a series of look up tables that link the 2001 Census Output Units to other geographies required and to geographies that will change over the decade. This assumes that the proposal for multiple geographies at the smallest scale will not be approved, and that the scheme of Figure 5.1 will be the one implemented. Confidentiality would be preserved because any additional SAS would be estimates only and any differencing operations would not yield useful statistics.

So, the sequence of events in the production of SAS for OUs and associated look up tables would be the following.

(i) Agree that OUs be aggregations of the postcode units that fit exactly as possible within wards.

(ii) Carry out a design study to work out methods for defining OUs according to agreed criteria - minimum population and household size, contiguity, compactness, socio-economic homogeneity and recognition of interaction communities.

(iii) When OUs for 2001 are known (2000?), publish their composition in terms of PCs. The ED/PC directory for 2000 could then be used to generate a look up table for 1991 Census EDs and OAs to 2001 Census OUs. This table could be employed by users to generate 1991 Census SAS for all 2001 areas (Figure 1) prior to the 2001 Census.

(iv) Produce key look up files on an annual basis that link OUs to changing geographies. The OU/PC directory would one such file. Another would be an OU to ward directory to reflect ward changes (enabling links to the vital electoral and administrative statistics to be made). Alternatively, a ward/PC directory could be generated, leaving users the jobs of making the OU/PC/ward link.

(v) Produce for the 2001 Census the boundaries of PCs, which can be dissolved to provide OU boundaries and those of higher order geographies.

(vi) Update the PC boundaries each year as the OU/PC directory is updated so that users can match their new estimates with the boundary set.

In this plan very careful attention needs to be given to the exact specification and content accuracy of the look up tables, and the base from which they were to be produced. Aspects include accuracy of the OU/PC link, accuracy of the PC centroid, the need to have counts of the main SAS tables totals not just the number of private households. At a minimum these table totals should include (a) total households, (b) total residents, (c) residents in households and (d) residents aged 16 or more.

5.6 Work for user communities

User communities would purchase the Area Statistics (AS) and digital boundary data (DBD) for OUs in the UK in say 2002 and look up tables and updated DBD in each year from 2001 to 2010. User communities might wish to then organise once only conversion of the OU AS to the new geographies.

In the academic community the ESRC/JISC Census Programme supported two units - the Census Dissemination Unit (CDU), Manchester Computing, University of Manchester and the UKBORDERS project, Edinburgh University - who have fulfilled this role. The CDU have created new higher order SAS datasets from the Census purchase (e.g. postal district SAS from the postal sector SAS, a GB ward SAS from the county and Scottish region ward SAS and many others). This programme should be extended to cover changing geographies (e.g. a 2002 postal sector AS, a 2006 ward AS). The UKBORDERS project provides users with a tool for creating their own higher order geographies, as does the KINDS project (Manchester Metropolitan/Salford/Manchester Universities). However, experience has shown that users also value systematic libraries of boundary files, made available for FTP in different formats.

6. THE FLEXIBLE GEOGRAPHY TASK OF THE STATISTICAL DISCLOSURE CONTROL PROJECT: PROGRESS AND THINKING

Oliver Duke-Williams and Phil Rees, October 1996

This paper was prepared as a briefing document for a meeting of participants in the Flexible Geography Task of the Statistical Disclosure Control (SDC) Project. The SDC project is funded under the ESPRIT (Information Technology) programme of the European Union and is a network of National Statistical Offices and Universities led by the Central Bureau of Statistics (CBS or Statistics Netherlands). The School of Geography at the University of Leeds is funded to investigate methods for providing disclosure control for Geographical Statistics in collaboration with the Office for National Statistics, Census Population and Health Group. The first sub-task, FG-1, tackled by the project has been to assess the risks posed by differencing by carrying out experiments with a synthetic household and individual database constructed from 1991 Census SAS data for Yorkshire and Humberside and the 1% SAR for Great Britain.

6.1 Background

The Statistical Disclosure Control (SDC) programme is a network of projects headed by Statistics Netherlands, funded by the European Commission's Fourth Framework ESPRIT initiative and overseen by the Statistical Office of the European Communities. The aim of the programme is to improve the methods and practice of SDC and to deliver usable software for use by National Statistical Offices.

The Leeds project within this programme is concerned with *Flexible Geography*. The overall aim of the project is to develop means of delivering statistics for geographic areas, in a safe form, that meet user needs. The first Flexible Geography task, FG-1, seeks to determine whether sets of tables for geographical areas can be safely released and in what form. In the UK context we are specifically investigating whether it is safe to release Small Area Statistics for administrative, postal, grid and historic geographies. The second Flexible Geography task FG-2, seeks to provide National Statistical Offices with methods for testing whether additional user demands for geographies beyond this basic set can be safely supplied. Originally, the ambition was to provide methods that applied to any kind of tables, but this is now seen as too ambitious. Alternative ways forward for FG-2 are discussed later.

The content of this paper is informed by the extensive discussion of SDC issues at the recent 3rd International seminar on Statistical Confidentiality held on October 2-4 in Bled, Slovenia (organized by the Statistical Office of the Republic of Slovenia). At the Bled Seminar papers were presented by Rees and Duke-Williams (1996) and by Rees, Martin & Duke-Williams (1996).

6.2 Flexible Geography Task FG-1

6.2.1 Method

A system of databases and software has been set up to test the safety of producing more than one set of Small Area Statistics. The elements in the system are shown in Figure 6.1 and are as follows:

(1) *A synthetic household and persons database* is created by SYNTHPOP which is a random replication of the Household (1%) SARs records to form the circa 2 million households and 5 million people living in Yorkshire and Humberside in 1991. These household and individual records are assigned to enumeration districts according to SAS counts, and within EDs are assigned exact grid locations randomly. This is a highly artificial population but one that can be counted and processed like the real census population.

(2) *Boundary data sets held in ARC format* for alternative geographies. Any geography must be described as a set of boundaries in a commom co-ordinate system and is held as a set of ARC files. The population can be counted within any geographic unit in the boundary set. Boundary sets can be overlapped and the properties of the intersection areas examined.

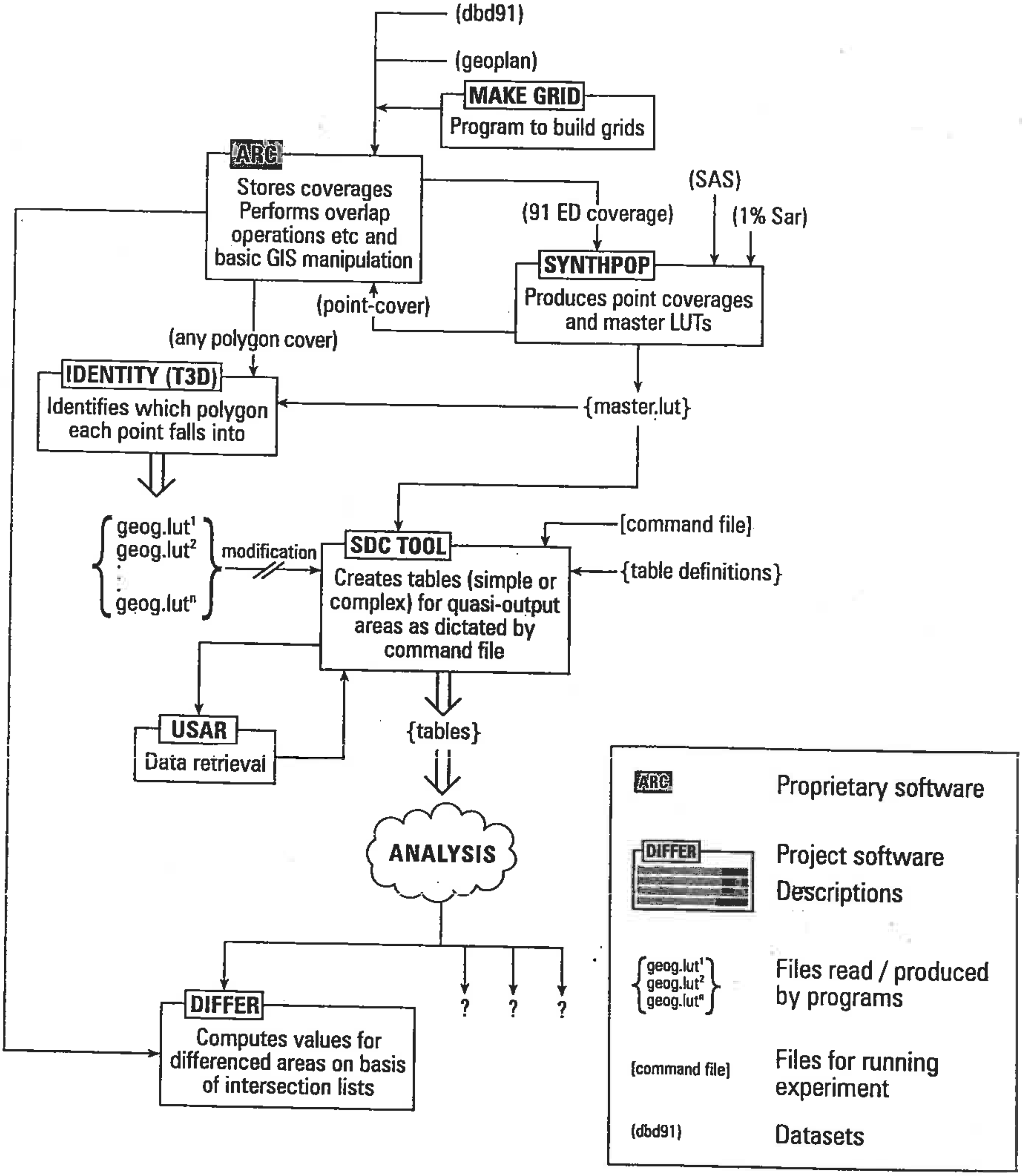

(dbd91)
(geoplan)
MAKE GRID
Program to build grids
ARC
Stores coverages Performs overlap operations etc and basic GIS manipulation
(SAS)
(91 ED coverage)
(1% Sar)
SYNTHPOP
Produces point coverages and master LUTs
(point-cover)
(any polygon cover)
IDENTITY (T3D)
Identifies which polygon each point falls into
{master.lut}
geog.lut¹
geog.lut²
geog.lutⁿ
modification
SDC TOOL
Creates tables (simple or complex) for quasi-output areas as dictated by command file
[command file]
{table definitions}
USAR
Data retrieval
{tables}
ANALYSIS
DIFFER
Computes values for differenced areas on basis of intersection lists
?
?
?
ARC
Proprietary software
DIFFER
Project software Descriptions
geog.lut¹ geog.lut² geog.lutⁿ
Files read / produced by programs
[command file]
Files for running experiment
(dbd91)
Datasets

Figure 6.1: The flexible geography experimental system

(3) *The GIS package ARC* is used to overlap boundary sets and identify the overlap/links between one set of areas and another in order. To count how many households and individuals are contained in the intersections is a function of ARC but to speed the processing a separate routine called IDENTITY has been written to do this point-in-polygon matching on the Edinburgh T3D parallel computer.

(4) *Small Area Statistics are generated in two steps* for (a sample of eight tables from the full set of 86). In the first step, the fast tabulation package USAR (Turton and Openshaw 1995) is used to construct the component simple crosstabs. In the second step, the component crosstabs are assembled together as SAS tables and sub totals and grand totals added by a FORTRAN routine that emulates operations normally carried out in spreadsheets. SDCTOOL controls this operation and will also contain the following routines.

(5) *Routines for modifying the all counts in the table to prevent disclosure* will then be applied. These have yet to be written and discussion is needed as to what techniques should be tested. In the broader context of the SDC programme, other groups are working on re-coding (Technical University Eindhoven), cell suppression (University of Padova) and controlled rounding (University of Padova). Their work is to incorporated in the μ-Argus (microdata SDC) and τ-Argus (table SDC) packages being produced by Statistics Netherlands. So we propose to implement only the current random perturbation of cell counts method, for the time being. Note that current practice uses minimum population and households counts as a joint protection device with random perturbation of cells.

(6) *Routines for reporting the results of experiments* in which the counts in two geographies are differenced. Currently, a FORTRAN routine, DIFFER, produces simple statistics for differenced areas. We have also developed a method for assembling cross tabulations into the tables of the standard SAS used in the 1991 Census as mentioned earlier.

6.2.2 Completed Experiments

We have begun running experiments using this system and reported them in the Bled Seminar (Rees and Duke-Williams 1996). Tables 6.1 to 6.3 show the distribution of population and household counts for areas formed by differencing two SAS geographies. The results are preliminary but we do not expect them to change substantially as the "errors" are reduced.

Table 6.1 reports on analysis of the differenced areas formed by subtracting sets of wholly contained EDs from postal sectors or postal sectors from larger sets of EDs containing the postal sector. Figure 6.2 shows how differences are computed. Of the 773 postal sectors, no differenced areas are generated, mainly because they are on the Yorkshire and Humberside boundary which defines which EDs were included in the study. Differencing is possible for 509 sectors, each of which has an inner and outer set of EDs. The inner set consists of EDs that fall wholly within the sector while the outer set of EDs wholly contains the sector. No differenced areas in this analysis contain sub-threshold numbers of persons or households. We conclude that release of SAS at small and medium scale is safe on this criterion. Both sets of SAS are protected by random perturbation of cells.

In Table 6.2 the comparison is different in that we use the ED/Postcode Directory for 1991. The census and postcode geographies match more closely because the two data sets refer to the same dates and because the directory reports the major intersections only. When the differencing is done using two boundary sets as in Table 6.1 lots of sliver intersections are produced as a result of differences in accuracy of digitising. These slivers, ironically, protect the data by making it more difficult to identify differenced areas. A much larger number of differenced areas (1396 compared with 1016) can be recognised as a result. In this comparison 6 differenced areas fall below the persons' threshold and 5 below the households' threshold. The percentage of differenced areas below threshold, however, is very small - only 0.4%. This represents only 1 area in 250, and the associated SAS tables are still protected by random perturbation of interior cells.

Table 6.1: The properties of areas formed by differencing 1991 Census Enumeration Districts and 1995 Postal Sectors in Yorkshire and Humberside

Postal sector category	No. of postal sectors	% of total
Differencing not possible	262	33.9
No wholly contained EDs	50	
On edge of region	212	
Differencing possible	511	66.1
No data (new PSs)	2	
With data	509	
Total	773	100.0

Population size	No. of differenced areas	% of total
Above Threshold		
200+	1018	100.0
100-199	0	0.0
50-99	0	0.0
Below Threshold		
<50	0	0.0
Total	1018	100.0

Household size	No. of differenced areas	% of total
Above Threshold		
48+	1018	100.0
32-47	0	0.0
16-31	0	0.0
Below Threshold		
<16	0	0.0
Total	1018	100.0

Table 6.2: The properties of areas formed by differencing 1991 Census Enumeration Districts and 1991 Census Postal Sectors in Yorkshire and Humberside

Population count	No. of differenced areas	% of areas
Above Threshold		
200+	1366	97.9
100-199	9	0.6
50-99	3	0.2
Below Threshold		
<50	6	0.4
Errors	12	0.8
Total	1396	100.0

Household	No. of differenced areas	% of areas
Above Threshold		
48+	1372	98.4
32-47	3	0.2
16-31	1	0.1
Below Threshold		
<16	5	0.4
Errors	13	0.9
Total	1394	100.0

Table 6.3: The properties of areas formed by differencing 1991 Census Enumeration Districts and 1 kilometre grid squares in Yorkshire and Humberside

Population count	No. of differenced areas	% of areas	Household count	No. of differenced areas	% of areas
A. Differenced areas formed by subtracting EDs from grid squares					
Above Threshold					
200+	1716	99.4	48+	1720	99.7
100-199	5	0.3	32-47	2	0.1
50-99	3	0.2	16-31	2	0.1
Below Threshold					
<50	1	0.1	<16	2	0.1
Error	1	0.1	Error	0	0.0
Total	1726	100.0	Total	1726	100.0
B. Differenced areas formed by subtracting grid squares from EDs					
Above Threshold					
200+	584	40.1	48+	694	58.2
1 0-199	162	13.6	32.47	64	5.4
50-99	80	6.7	16.31	69	5.8
Below Threshold					
<50	263	22.1	<16	250	21.0
Error	101	8.5	Error	106	8.9
Total	1190	100.0		1193	100.0

Notes:

1. Errors are negative numbers sometimes produced when Bracken/Martin Census population counts by 200 metre grid square are used. These grid counts are estimates produced by the allocation of whole ED or partial ED counts to grid squares.

Table 6.4: Comparisons planned between different geographies

Different geography	Standard SAS geography: 1991 EDs
Postal sectors 95	✓
Postal sectors 91	✓
1 km grid sq.	✓
0.2 km grid sq.	*
5 km grid sq.	*
10 km grid sq.	*
Mixed small and large grid sq.	*

Notes: ✓ = experiment completed * = experiment planned

Table 6.3 reports the results of comparing ED counts and counts for grid squares of 1 kilometre in size. If we subtract EDs from grid squares, only 1 differenced area falls below the persons' threshold and 2 below the threshold for households. However, when we reverse the differencing this benign situation changes and we find that just over 20% of differenced areas fall below the thresholds. The reason for this is that many grid squares fit inside low density rural Eds. Note that in practice SAS for large numbers of rural grid squares would be suppressed because their populations fall below the confidentiality thresholds.

6.2.3 Planned experiments

The simulated database can be used to carry out many more experiments using grid square networks, whose boundaries can be easily generated. Table 6.4 lays out a suggested set of further experiments.

The grid cells of different size will provide an idea of the effect of spatial scale on the extent to which differenced areas fall below threshold. The final experiment seeks to adjust grid cell size to the density of settlement (as does the grid cell output from the Northern Ireland Census, for example).

6.2.4 Protection measures to be applied to the SAS tables

Practice at the 1971, 1981 and 1991 Censuses was to carry out "blurring" or random perturbation of cell counts by adding +1, 0, or -1 to the count in a quasi-random pattern. Counts that start as zero are left unmodified. The distribution of adjustments is symmetric - i.e. an equal number (overall) of cells are raised by +1 as are lowered by -1. However, the frequency of modification compared with no modification (adding zero to the cell count) is kept confidential.

Similar perturbation has been applied to the "big brother" of the SAS - the Local Base Statistics at ward (England and Wales, Northern Ireland) or sector (Scotland) scale. All published accounts of the protection measures suggests that "double blurring" has been applied. However, Rees and Duke-Williams (1994, p.15) have cast doubt on this statement. They compared a true of count of migrants resident in wards (England and Wales) and sectors (Scotland) derived from the Special Migration Statistics which was not perturbed with the equivalent count from the LBS which was perturbed. This comparison showed that only single blurring had been applied. Clear and unambiguous statements about protection measures should be made in connection with the 2001 Census.

Cole (1993) has outlined the difficulties produced as a result of modification of all counts through random perturbation. He recommends that ED SAS counts be used only when SAS at a higher spatial scale is not available, and that statistics derived from ED SAS be treated with considerable caution. The main problems occur when the EDs are small or suppressed (being sub-threshold). The Output Area SAS in Scotland suffered particularly from small populations. Small population sizes in addition make use of the 10% sample statistics at ED scale very doubtful because of the very wide confidence bands. How might these difficulties be reduced in the 2001 Census SAS, while retaining proper protection?

A number of proposals for improving SAS outputs in 2001 have been put forward.

(i) The topics (occupation, employment) previously sampled should be published at a 100% level.

(ii) Output units should be defined independently of collection areas so that they can be designed, pre-census, to be above threshold (avoiding suppression).

(iii) The output units should be defined to be well above the thresholds and reasonably equal in size.

(iv) Linkage to previous geographies should be achieved through look-up tables, not by producing very small lowest common denominator areas (as was the case in Scotland in 1991).

(v) The protection device should be changed from random perturbation of cell counts to another protection device such as record swapping.

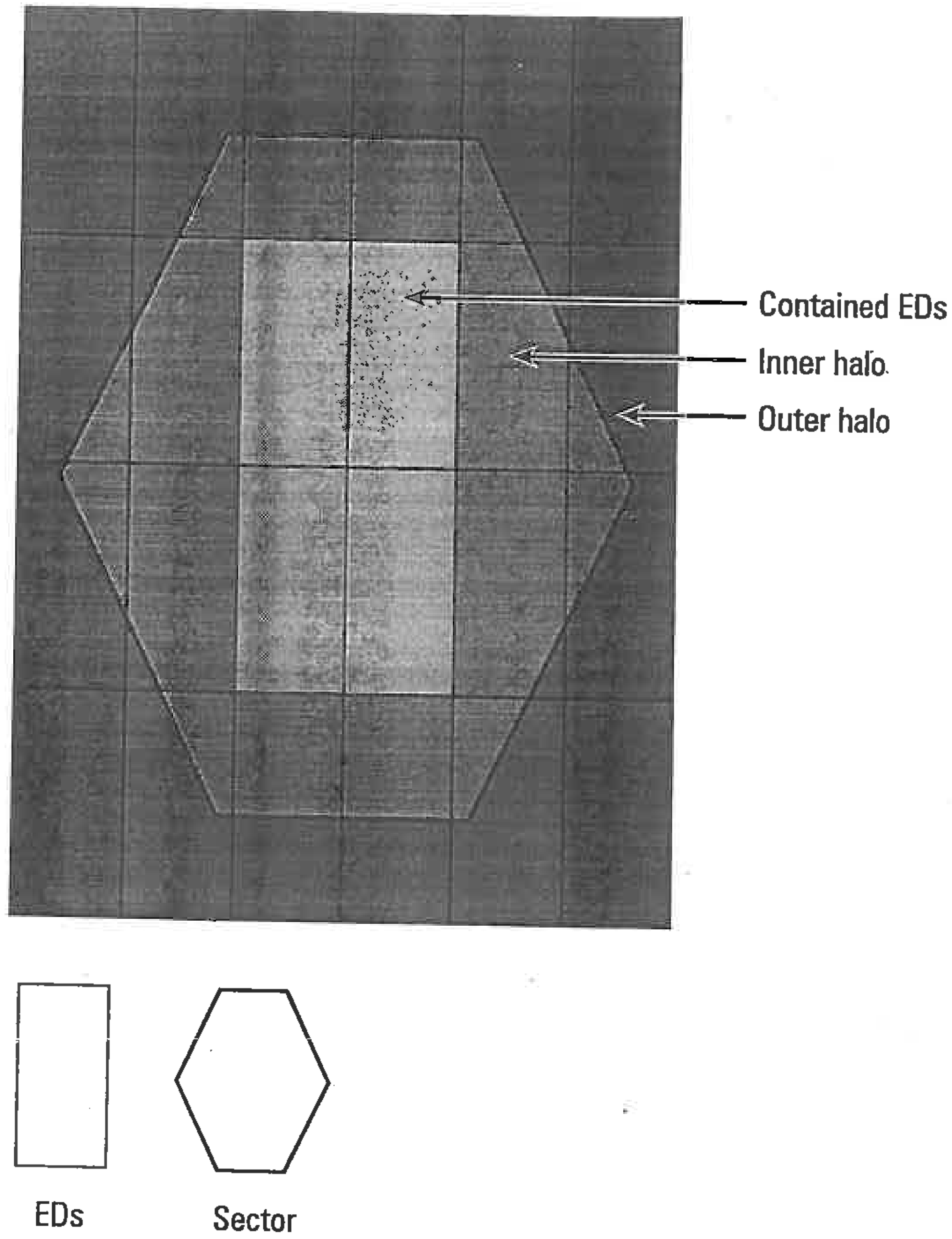


Figure 6.2: How to compute differences

Proposal (i) is being considered by the Content Working Group in connection with design of the occupation question in particular. Proposal (ii) has been accepted by the Outputs Working Group. Proposals (iii) and (iv) are made in Rees, Martin and Duke-Williams (1996). We consider proposal (v) here in more detail.

6.2.5 Record swapping as a protection measure

Proposal (v) has been used by other National Statistical Offices (e.g. US Bureau of the Census). What happens is that records are swapped between geographical units or in other words the geographical identifiers of two records are swapped. Swaps are confined to records with a set of key variables in common. If all variables had to be common (extremely unlikely over 24 variables), then nothing would change.

The advantage of record swapping is that all subsequent statistics could be generated on the basis of a fixed, perturbed master database. The tables in Area Statistics would be fully consistent - table totals that should agree would be the same. Tables would sum consistently from small areas to give large area tables.

The disadvantage of record swapping across geographical areas is that it would alter micro-geographical patterns. The distortion could be limited by confining swapping within geographical limits - no swaps across wards or perhaps across output units (only between building blocks, the unit postcode).

Record swapping could be seen as a supplement to two record addition operations proposed for the 2001 census:

(i) record imputation, where a record is duplicated from another nearby for housing units certified by as occupied by enumerators.

(ii) record estimation, where an estimate of the undercount of households is made using information from the Census Validation Survey and from the age-sex structure of the population and records are duplicated from others in the same area or pool of areas.

The Yorkshire and Humberside experimental system can be used to design record swapping methods which distort the true geography as little as possible.

6.2.6 Summary of FG-1 sub-tasks

(i) Complete experiments to overlap alternative geographies and evaluate differenced areas.

(ii) Implement the random perturbation measure and show the effects at different aggregations on SAS tables.

(iii) Implement a record swapping method and show the effects at different aggregations on the SAS tables.

(iv) Write a report making recommendations about how Small Area Statistics should be developed from the 2001 Census.

6.3 Flexible Geography Task FG-2

The original intention was to develop a system for producing tables for alternative geographies, providing assessment of the safety of table requests. The system would be located in a safe environment (e.g. ONS computer system). It was envisaged that users would submit table requests and that these could be swiftly assessed as safe (green light) or unsafe (red light) automatically, or as ambiguous (orange light) and needing expert assessment.

It is likely from the results of our differencing experiments to date that most user generated geographies at small area scale will prove to be unsafe. In addition other groups in the SDC project have reported that the problem of linked tables, of which alternative sets of geographical tables are a special case, is proving very difficult to solve.

The solution proposed in Rees, Martin and Duke-Williams (1996) is that instead a family of look up tables be developed to make possible the conversion of standard OU SAS to new geographies.

Flexible Geography task FG-2 should therefore be re-oriented to developing the methods for (i) designing output units and (ii) building look up tables. We explain what is involved.

6.3.1 Design of output units

Rees, Martin and Duke-Williams (1996) define the task of designing output units as being one of combining building blocks (postcode units) within a ward into a set of equal sized and homogeneous units. Algorithms for doing this have been developed by Openshaw (see Openshaw and Rao 1995). These algorithms search for the optimal grouping into M regions of N zones to maximise or minimize an objective function subject to a set of constraints. The search methods include a heuristic, the original Automatic Zoning Procedure (AZP) of Openshaw (1977), simulated annealing and tabu search.

The typical OU design problem would be to combine 200 unit postcode areas into 20 output units so that the heterogeneity of OU population and household sizes and intra-zone distances were minimised (or the OU populations sizes were as similar as possible and their shapes as compact as possible). Minimum size thresholds should probably be set some way above confidentiality thresholds so as to reduce the degree of revision needed post-Census. The equal size goal is designed to provide roughly equivalent "confidence bands" around derived statistics. The compactness criterion is designed to maximise the "geographic representativeness" of OUs and to avoid the appearance of gerrymandering.

It would be possible to test out the Openshaw algorithms on the Yorkshire and Humberside experimental system using unit postcode aggregations of the underlying populations as building blocks.

The design process involves several steps.

(i) Designing methods of constructing boundaries for unit postcodes (building blocks) and for generating contiguity information for all building blocks.

(ii) Agreeing the objective functions to be used in the design process. The software to be used, ZDES, is very flexible so the results of using different objective functions can be explored.

(iii) Testing the design using the Yorkshire and Humberside experimental system.

(iv) Testing the design using the 1997 Census Test areas.

(v) Presentation review and revision of the Output Units proposals in consultation with user groups.

(vi) Rolling out the design process to the whole United Kingdom. This last step could be carried out with the 1996 postcodes but would need revision later in the decade closer to the 2001 Census.

6.3.2 Look up tables

We first discuss what these should look like, and then outline the family of look up tables (LUTS) that will be needed.

There are three types of LUTS: (i) exact LUTs, (ii) best fit LUTs and (iii) weighted LUTs. The first type simply links the small area to the larger area it fits exactly. For example, wards fit exactly inside one and one only district (unitary authority). The census code provides such a look up table e.g. area 08DAFJ01, an enumeration district code for county 08 (West Yorkshire), district DA (Leeds), ward FJ (Cookridge), ED 01 (the first). The second type links the small area to another set of areas with which it has overlaps and assigns a weight to each overlap. The weight may be based on the respective areas, on a count of residential addresses or a count of resident households or residents. An example of such a LUT is the enumeration district/postcode directory which has entries for every ED/PC intersection and a count of associated households, which can be used to weight the conversion of ED SAS counts to PC SAS counts.

Three kinds of look up tables will be needed in connection with Output Unit SAS from the 2001 Census. The first are LUTs that convert SAS in previous censuses to the 2001 Census OU geography. The second are LUTs that convert OU SAS to other geographies at the time of the 2001 Census. The third are LUTS that update, during the 2001-2011 period, the OU SAS to both the standard geographies and the alternate geographies.

Table 6.5 sets out a list of the LUTs which will be needed. The first set provides for backwards compatibility over time and would be used to generate new versions of the 1971, 1981 and 1991 SAS for 2001 Output Units and Postal Sectors. These LUTs could be constructed wither by the Census Offices or by the academic community, building on previous exercises carried out by the University of Newcastle, for example.

The second set of LUTs provide the links between 2001 Output Units and other geographies for which SAS are not to be produced and which are not simple aggregations of OUs. This list is illustrative and there could be many more such LUTs.

The third set of LUTs maintain the link, year by year, between the OU and the two main geographies, administrative/electoral and postal for which SAS are published.

The best methods for constructing LUTs - such as from ED to PC directories, or from unit postcode best fits - need to be investigated. Some final points need to be made. The LUTs for small area geography are large files with 1000s of entries. Means of quality checking of the information need to be developed. Also the degree of approximation involved in constructing estimated SAS needs to be assessed.

6.4 Conclusions

This paper has reviewed progress made on the Flexible Geography tasks contained within the Statistical Disclosure Control programme funded by the European Union's ESPRIT project. The experimental system constructed for a UK region is already yielding valuable results on the "new areas, new dangers" theme. Proposals for carrying the work forward over the remainder of the first year of the project (FG-1 task) have been made. Revisions have been made to the direction of work for the second year (FG-2 task) which are the subject of discussion with the Office for National Statistics. These revisions will produce, we feel, deliverables that can be used by National Statistical Offices.

Table 6.5: A family of look up tables

Census	Geography	Year(s)	Geography	Coverage
LUTs to provide backwards compatibility in time				
1971	Wards	2001	OUs	Great Britain
1981	Wards	2001	OUs	Great Britain
1991	EDs	2001	OUs	England Wales
1991	OAs	2001	OUs	Scotland
1991	EDs	2001	OUs	Northern Ireland
1991	Postal Sectors	2001	Postal Sectors	United Kingdom
LUTs to provide alternative geographies in 2001				
2001	OUs	-	km grid squares	United Kingdom
2001	OUs	2001	Civil Parishes	England
2001	OUs	2001	Communities	Wales
2001	OUs	2001	Civil parishes	Scotland
2001	OUs	2001	Postcodes	United Kingdom
LUTs to provide forwards compatibility in time				
2001	OUs	2002-10	Wards	United Kingdom
2001	OUs	2002-10	Postcodes	United Kingdom

7. ISSUES CONCERNING THE FLEXIBLE GEOGRAPHY TASK OF THE STATISTICAL DISCLOSURE CONTROL PROJECT

Oliver Duke-Williams and Phil Rees, October 1996

This note spells out issues raised in discussions about flexible geography at the School of Geography held on 22 October and at the Office for National Statistics (ONS) held on 23 October 1996. Involved in the discussions were the SDC project team at the School of Geography at Leeds (Phil Rees, Oliver Duke-Williams and Stan Openshaw), members of the 2001 Census Development team (Chris Denham, David Thorogood, John Puckey, Andy Teague) and David Martin (on research leave from the department of Geography at the University of Southampton).

7.1 Measuring the risk of disclosure

In order for us to perform any critical comparison of different protection devices it is necessary for us to have a quantitative measure of risk of disclosure. Here, 'disclosure' is said to have taken place if any published Census data allow someone to discover information about an individual over and above that which the 'discloser' already knows about the individual.

The model of risk is based on considering the number of uniques in the data. A unique is an individual who has unique values for a number of key variables. As the number of key variables used to make this assessment increases, so the likelihood of uniques occurring increases. The choice of these variables and the justification of the choice (or perhaps more importantly, the justification of exclusion of particular variables from the set of key variables) is important.

It should be noted that uniqueness should be considered in terms of the output tables rather than the input data. For example, suppose the most detailed grouping of age was five-year age groups. If two individuals were unique when assessed on all other key variables, and were aged 22 and 23 respectively, then they would not have to be considered as unique for the purpose of risk assessment, because they would be reported in the same age category.

This concept of using variable groupings as detailed in an output table rather than the input data becomes more complicated when more than one table is considered (as must be the case). The 'mask' used to filter data into different groups must be made up of the intersection of all uses of the variable under question. The most complicated variable (in that in has the most different groupings) is age, which has 57 different representations in the 1991 SAS and LBS tables (Williamson 1993). However, only one grouping initially needs to be considered, as it is more detailed than any other grouping (the one used in table L38). This grouping has single years of age 0 to 90, and a final 90+ category. Nevertheless, the question of intersections of groupings may remain. There are 23 age groupings which are only subsets of the possible age range, typically starting at age 16. The interaction of age with other variables may be important. Does the assessment of uniqueness have to take into account the relationships which exist between variables? Suppose some individual is found who has unique characteristics based on some set of variables, but no table exists which cross-tabulates the variables in a way that this uniqueness can be discovered. Does this still represent a dangerous case?

Table 7.1: An example data set

Case	a	b	c	d	e
1	1	2	3	1	2
2	1	2	3	1	1
3	1	3	2	1	2
4	1	3	2	1	1

Table 7.1 illustrates this. This shows a dataset with four cases, each of which have values for five variables (a to e). If all five variables are considered together, then all four records are unique. However, suppose only two output tables are generated from this data; table 'a' using variables a, b and c, and table 'b' using variables d and e. Neither table would have any cells containing a value of

1. Is any data still at risk? However, a more complicated case can be made that there is a third table, table 'c', which uses variables a, d and e. Again, no cells in this table would have a value of 1, but does this create a potentially dangerous linkage between all five variables?

A formal model of assessing uniqueness of records must therefore be established.

7.2 Risk in tables

This sections assumes that some model of uniqueness has been set up, and that records in the original data can be flagged as either unique (and therefore at risk) or not unique.

7.2.1 Measuring risk in tables

For any table, the 'raw risk' can be assessed as n/m, where n is the number of cells with a count of 1, and m is the total number of persons (or households) tabulated. Any process designed to find some 'ideal' set of tables and output areas could seek to minimise the total raw risk for all tables in all areas. It is perhaps a matter of debate whether this measure of risk should be reduced to zero, or minimised to some 'acceptable risk level', possibly a similar level to that which was present and therefore deemed to be acceptable in the 1991 Census.

However, this model of risk over-estimates the likelihood of an individual being disclosed because of various forms of error and uncertainty which are either introduced to the data deliberately or believed to be unavoidable. In each case, the raw risk value can be multiplied by some factor between 0 and 1 in order to reflect the role of error in protecting the data. A list of such factors might include:

whole record imputation
If k% of records are imputed, then the risk should be multiplied by (100-k) / 100

estimation
If l% of records are estimated, then the risk should be multiplied by (100-l) / 100

record swapping
If m% of records are swapped, then the risk should be multiplied by (100-m) / 100

random perturbation
If x% of cells are shifted by + or - 1, then the risk should be multiplied by (100-x) / 100

geographic blurring
If y% of records are sampled from outside the fixed boundaries of the zone in question, then the risk should be multiplied by (100-y) / 100

part record imputation
Any variable used in a (non wholly-imputed) record may have imputed (due to errors or omissions on the original form). This is likely to happen to different degrees with each variable, and so there should be some factoring down based on the variables used. For all variables in the cross-tabulation, a proportion z might be imputed, and so the raw risk can be modified by:

$\{(100-z_{x1})/100\}) \{(100-z_{x2})/100\} \ldots \{(100-z_{xn})\}/100\}$

Some thought needs to be given to the level of independence of these factors. Clearly, records can not be imputed and estimated, and record swapping would presumably be done on 'real' records only. Thus simple multiplication by all factors might not be the best approach. The effects of record swapping will also vary depending on the swapping algorithm used. For example, if records are swapped between output units within a ward, then it this part of the risk calculation should be dropped as output geographies become ward-sized or greater.

When more than one table is used, the above assessment of risk must be repeated for all tables, and then the results summed.

The effect of cell blurring on the raw risk can be seen in two ways. Firstly, it could be argued that it effectively reduces the risk to 0, because no-one can confidently say that any particular unique in a table (i.e. a count of 1 in a table cell) is correct. On the other hand, a more acceptable approach might be to consider the probability of any particular cell being a genuine count of 1. For the given blurring algorithm implemented, cell counts of 0 to x would all be considered as being at risk, if x is the maximum possible value to which the algorithm might convert an initial count of 1. The total risk calculated for cells of each at risk value should then be summed.

7.2.2 Measuring the damage caused by protection devices

A second model must be established to determine the amount of damage that each prospective SDC scenario with a particular combination of protection devices and chosen parameters would do to the data. This can be fairly easily implemented by producing two versions of all tables - a true table compiled form the original data, and an output table which is the one that might be expected to be published. A measure of divergence for the `published' table can be generated (probably using a chi-squared measure - the sum of differences between observed and expected values squared divided by the expected).

7.3 Issues discussed with the Office of National Statistics at Titchfield

7.3.1 Flexible Geography Tasks of SDC project

The results produced thus far at Leeds were presented and discussed. It was suggested that we carry out more experiments for the same geographies but with alternative thresholds. Thresholds set at 1/2 and twice the 1991 ones were suggested as suitable test values.

Some concern was voiced about our intentions to include random perturbation of interior cell counts, or cell blurring, amongst the set of SDC techniques to be tested. Chris Denham felt that it would make it difficult for ONS to comment on our results, as they would be unwilling to say whether or not our blurring algorithm produced similar results to that used for the 1991. However, we still believed that it was vital to our experiments to be able to compare various alternative approaches to the 'traditional' blurring approach.

7.3.2 General Proposals in the draft OWG paper

Discussion of the OWG paper led on to a rather more wide ranging discussion about the possible output geography of the 2001 Census. The idea that the problem of defining (sub-ward) output units could be deferred until after the generation of ward level statistics was suggested and met with mixed views.

It was generally felt that output units built up from unit post codes (UPCs) would have great utility:

1. They can be linked to other datasets based on UPCs.
2. Their commercial value would therefore be considerably enhanced.
3. They can be designed to meet external criteria relevant for users.

However there are a number of problems which are important and have to be overcome.

Establishing polygons around UPCs is something which is difficult to do. David Martin is carrying out various experiments at ONS into methods of doing this. These include defining Thiessen polygons around the grid centroids assigned to UPCs. The point was raised: what happens when a new address is added to the dataset which falls outside the polygon? The counter-point is that this is simply part of a general revision process that goes on to subsequent to the census. Unit postcode definitions and boundaries would need to be redefined.

A more significant problem are problems with the Address-point data; it is difficult to predict the populations of UPC areas prior to the Census, because it is difficult to differentiate between commercial and residential properties in Address-point. Furthermore, Address-point obviously has limited coverage: `dead' codes will not be removed immediately and likewise there will be a time lag in adding new codes. Therefore, any creation of output areas based on UPCs will require significant post-Census work in accounting for anomalies in Address-point, as well as performing required aggregation of UPCs into areas above whatever threshold is used.

Another consideration is that if output units were to be based on UPCs, then the process would involve the creation of a new address--postcode database, together with a set of digital boundaries. Whilst this would be a viable commercial opportunity for ONS, it might present difficulties in that it would be inconsistent with the Central Postcode Directory (CPD); ONS would be in the undesirable position of having two similar products which gave conflicting results.

8. A NOTE ON LOOK UP TABLES

Phil Rees, November, 1996

This note spells out in more detail the nature of the family of look up tables proposed in the Rees, Martin and Duke-Williams paper (paper 5 in this collection). The note is extracted from correspondence between Phil Rees and Frank Thomas in November 1996.

Look up tables can also be called "geographical pointers", "geographical indexes" or "geographical gazetteers". At some stage we need to adopt an agreed names - maybe geographical lookup indexes and tables (GLINT). They have the general form shown in Table 1 when applied to Output Units.

The LUT consists of a list of source geography Output Units, identified by suitable alphanumeric codes, labels or indexes ($U(i)$). Each OU intersects with a number ($N(i)$) of target areas (TA), which may be 0 or 1 or 2 or more. If N(i) is 0, this indicates that the target geography does not completely cover the OU geography. If all N(i)'s are one, then the LUT is a simple index that gives the OU membership of a TA which is a hierarchical aggregation of OUs. If $N(i) > 1$ then this indicates the two geographies overlap. There will be $N(i)$ entries of target area codes, $T_j(i)$, for each OU.

Associated with each intersection (entry in the table) are a set of weights that can be applied when converting OU counts to TA counts. In principle, there are K such weights - examples are the household count in the ED to PC directory in England and Wales, and the recently published population count. It would be very useful to extend the list of published weights to some other table population bases (e.g. economically active residents).

The LUT can be organised in different ways from that set out in Table 8.1. For example, the ED to PC directory has records that look in principle like this

$$\begin{array}{lll}
\vdots & \vdots & \vdots \\
U(i) & T_1(i) & W_1^1(i) \\
\vdots & \vdots & \vdots \\
U(i) & T_j(i) & W_2^1(i) \\
\vdots & \vdots & \vdots \\
U(i) & T_{N(i)}(i) & W_{N(i)}^1(i) \\
\vdots & \vdots & \vdots
\end{array}$$

Each entry is a source geography/target geography pair or intersections. This has the advantage that the file can be re-sorted to provide the LUT to convert from target geography to source geography. Gazetteer files used with the SASPAC software have this type of format.

Given agreement that OU SAS for wards and sectors can all be produced, we need a family of lookup tables to be used in converting OU SAS to other output geographies. Table 8.2 is a tentative first list. The lists do not prescribe any particular method for the construction of the LUT contents. This will depend on the address data base available, the postcode data base available, their costs, their accuracies, licence conditions and so on. However, I do note that all Census Offices either now have or will have soon good GIS systems to handle the process of LUT construction.

Table 8.1: **Formal description of a geographical lookup and index table (GLINT)**

Source Geography		Target Geography					
Output Unit Code	No. of Target Areas Intersected	Target Area Codes	weights for source and target area intersections				
			1	...	k	...	K
$U(1)$							
	$N(1)$						
		$T_1(1)$	$W^1_1(1)$	...	$W^k_1(1)$	...	$W^K_1(1)$
		$\vdots$	$\vdots$				$\vdots$
		$T_j(1)$	$W^1_j(1)$	...	$W^k_j(1)$	...	$W^K_j(1)$
		$\vdots$	$\vdots$				$\vdots$
		$T_{N(1)}(1)$	$W^1_{N(1)}(1)$	...	$W^k_{N(1)}(1)$	...	$W^K_{N(1)}(1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$U(i)$							
	$N(i)$						
		$T_1(i)$	$W^1_1(i)$	...	$W^k_1(i)$	...	$W^K_1(i)$
		$\vdots$	$\vdots$				$\vdots$
		$T_j(i)$	$W^1_j(i)$	...	$W^k_j(i)$	...	$W^K_j(i)$
		$\vdots$	$\vdots$				$\vdots$
		$T_{N(i)}(i)$	$W^1_{N(i)}(i)$	...	$W^k_{N(i)}(i)$	...	$W^K_{N(i)}(i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$				$\vdots$
$U(n)$							
	$N(n)$						
		$T_1(n)$	$W^1_1(n)$	...	$W^k_1(n)$	...	$W^K_1(n)$
		$\vdots$	$\vdots$				$\vdots$
		$T_j(n)$	$W^1_j(n)$	...	$W^k_j(n)$	...	$W^K_j(n)$
		$\vdots$	$\vdots$				$\vdots$
		$T_{N(n)}(n)$	$W^1_{N(n)}(n)$	...	$W^k_{N(n)}(n)$	...	$W^K_{N(n)}(n)$

Definitions:-

$U(i)$ = Code for output unit i
$N(i)$ = Number of target areas intersected by unit i
$Tj(i)$ = the jth target area intersected by output unit i
$W^k_j(i)$ = value of weight k for output unit i and target area j

Table 8.2: Suggested list of look up Tables

LUTs to link to the 1991 Census areas

1. 2001 version of ED to PC Directory (England & Wales)
2. 2001 version of ED to PC Directory (Northern Ireland)
3. 2001 version of OA to PC Directory (Scotland)

The following will exist as a result of OU design:

4. 2001 PC to OU Index (England & Wales)
5. 2001 PC to OU Index (Northern Ireland)
6. 2001 PC to OU Index (Scotland)

It should therefore be possible to write software to link LUTs 1 and 4, 2 and 5, 3 and 6, and produce the following LUTs

7. 1991 ED to 2001 OU (England & Wales)
8. 1991 ED to 2001 OU (Northern Ireland)
9. 1991 ED to 2001 OU (Scotland)

All other LUTs for larger scale geographies can be built from LUTs 7,8,9
e.g. 1991 ward to 2001 ward (UK)
which might be used to convert data only available at ward scale (e.g. vital statistics).

For earlier Censuses, the academic community should be able to build on the Newcastle work (Dorling et al.) linking 1971 EDs and 1991 EDs and OAs to 1981 Census Wards in Great Britain to produce a LUT for 1971 EDs and 2001 OUs together with a LUT for 1981 EDs and 2001 OUs.

LUTs to link to other 2001 Census areas: fixed

There will be just one set of fixed LUTs - those that connect the OU to grid areas.

10. 2001 OU to grid area at various scales (UK)

LUTs to link to hierarchy of 2001 Census administrative and postal areas: fixed

11. 2001 OU to wards, LGD/UAs, regions, countries, health areas etc. = indexes/area master files (UK)
12. 2001 OU to sectors, districts, areas (post-towns) = indexes/area master files (UK)

LUTs to link to administrative and postal geographies over time

13. 2001 OU to wards, LGD/UAs, regions, countries, health areas etc. - update for year 2001 +*n* (annual)
14. 2001 OU to PC directory, - update for year 2001+*n* (twice yearly).

REFERENCES

Barr R (1993) Census geography: a review. Chapter 3, Part II in Dale A and Marsh C (eds.) *The 1991 Census User's Guide*. HMSO, London. Pp.70-83.

Birkin M (1995), Customer targeting, geodemographics and lifestyle approaches, In: Longley, P. and Clarke, G (eds) *GIS for business and service planning*, GeoInformation International, Cambridge, 104-149.

Bracken I and Martin D (1995) Linkage of the 1981 and 1991 UK Censuses using surface modelling concepts. *Environment and Planning A, 27, 379-390.*

Coombes M (1995) Dealing with census geography: principles, practices and possibilities. Chapter 4 in Openshaw S (ed.) *Census Users' Handbook*. GeoInformation International, Cambridge. Pp.111-132.

Cole K (1993) The 1991 Local Base and Small Area Statistics. Chapter 8 in A Dale and C Marsh (eds.) *The 1991 Census User's Guide*. HMSO, London. Pp. 201-247.

Cushnie J (1994), A British Standard is published, *Mapping Awareness* 8 (5), 40-43.

Dale A and Marsh C (eds.) (1993) *The 1991 Census User's Guide*. HMSO, London.

Denham C (1993) Census geography: an overview. Chapter 3, Part I in Dale A and Marsh C (eds.) *The 1991 Census User's Guide*. HMSO, London. Pp.52-69.

Dorling D (1995) Visualizing chnaging social structure from a census. *Environment and Planning A, 27, 353-378*

Elliott L, McCallum D and Pretty S. (1993) ADDRESS-POINT: fusion of OS mapping precision and Royal Mail postal addresses, *Proceedings of AGI 93 conference* AGI: London 1.9.1-1.9.13.

Lander Brinkley N (1996), 2001 Census: a geography for outputs, an operational overview. *Output Working Paper OWG 13*, 2001 Census programme, Office for National Statistics, Census Division.

Marsh C, Arber S, Wrigley N, Rhind D and Bulmer M (1988) The view of academic social scientists on the 1991 UK Census of Population: a report of the Economic and Social Research Council working group. *Environment and Planning A, 20, 851-889.*

Marsh C, Dale A and Skinner C (1994) Safe data versus safe settings: access to microdata from the British Census. *International Statistical Review*, 62, 35-53.

Martin D (1996) *Geographical information systems: socioeconomic applications.* Second Edition Routledge: London.

Martin D and Higgs G (1996), Population georeferencing in England and Wales *Environment and Planning A* (in press).

Openshaw S (1990) Spatial referencing for the user in the 1990s, *Mapping Awareness* 4 (2), 24-29.

Openshaw S (1997) Algorithm 3: a procedure to generate pseudo random aggregations of *N* zones into *M* zones where *M* is less than *N*. *Environment and Planning A*, 9, 1423-1428.

Openshaw S and Rao L (1995) Algorithms for re-engineering 1991 census geography. *Environment and Planning A*, 27, March, 425-446.

Puckey J (1995), Census geography in the year 2000, *Proceedings of AGI 95 conference* AGI: London 8.1.1-8.1.5.

Raper J, Rhind D and Shepherd J (1992), *Postcodes: the new geography,* Longman: London.

Rees P (1995) The geography of outputs from the 2001 Census: a framework. *Output Working Group Paper OWG 9*, 2001 Census programme, Office for National Statistics, Census Division.

Rees P and Duke-Williams O (1994) The special migration statistics: an vital resource for research into British migration. *Working Paper 94/20* School of Geography, University of Leeds, Leeds, UK.

Rees P and Duke-Williams O (1996) A protection system for small area statistics. In I Trsinar and S Dujic (eds.) *Statistical Confidentiality.* Proceedings of the Third International Seminar on Statistical Confidentiality. Statistical Office of the Republic of Slovenia, Ljubljana. Pp.80-92.

Rees P and Martin D (1996) Flexible geographies and area aggregation: designing small areas for outputs from the 2001 census. Paper presented at the Royal Statistical Society/British Society of Population Studies meeting on *The 2001 Census,* held at the Royal Statistical Society, Errol St, London, April 1996.

Rees P, Martin D and Duke-Williams O (1996) The geography of outputs from the 2001 Census: output units and look up tables. Paper prepared for the Output Working Group, Census Development Programme, UK Census Offices.

Thomas F (1996a) The geography of outputs from the 2001 Census. *Output Working Group Paper 12,* Census Development Programme, Office for National Statistics, Titchfield, Hampshire.

Thomas F (1996b) The geography of outputs from the 2001 Census. *Output Working Group Paper 15*, Census Development Programme, Office for National Statistics, Titchfield, Hampshire.

Turton I and Openshaw S (1995) Putting the 1991 sample of anonymised records on your UNIX workstation. *Environment and Planning A,* 27, March, 391-411.

Yeoman B (1995) Addressing your information - The Bristol trials of BS7666. *Proceedings of AGI 95 conference* AGI: London 10.1.1-10.1.4.