WORKING PAPER 98/7

THE 2001 CENSUS: What do we really, really want?

Edited by Philip Rees.

PUBLISHED SEPTEMBER 1998

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

WORKING PAPER 98/7

THE 2001 CENSUS: WHAT DO WE REALLY, REALLY WANT?

Edited by Philip Rees

Contributions by
Seraphim Alvanides (University of Leeds)
Bob Barr (University of Manchester)
Paul Boyle (University of Leeds)
Angela Dale (University of Manchester)
Brian Dodgeon (University of London)
Oliver Duke-Williams (University of Leeds)
Martin Frost (Kings College London)
Heather Joshi (University of London)
David Martin (University of Southampton)
Donald Morse (University of Edinburgh)
Stan Openshaw (University of Leeds)
Phil Rees (University of Leeds)
Ian Turton (University of Leeds)
Simon Whalley (University of Leeds)
Paul Williamson (University of Liverpool)

A Report to
The Economic and Social Research Council (ESRC) and
The Joint Information Systems Committee (JISC)
by the participants
in the Fourth Workshop Planning for the 2001 Census
13-14 May 1998
held at the School of Geography, University of Leeds, Leeds, UK

This report summarises the presentations made and discussions held at the Fourth of the ESRC/JISC *Workshops Planning for the 2001 Census*. The report presents views of expert census users and summarises the recommendations to ESRC and JISC about what kinds of data from the 2001 Census should be requested from the UK Census Offices. The Workshops are supported by ESRC Award H507265031.

PUBLISHED SEPTEMBER 1998

For further copies contact the Working Paper Secretary,
School of Geography, University of Leeds, Leeds LS2 9JT
Telephone +44 (0)113 233 3300

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

ACKNOWLEDGEMENTS

ESRC and JISC very generously supported this series of *Workshops Planning for the 2001 Census - Determining Academic Community Needs and Strategy*, under ESRC Award H507265031. The UK Census Offices have been able to accept invitations to attend the Workshops and receive the inevitable mix of brickbats and bouquets. Chris Denham (ONS), Frank Thomas (GROS) and Robert Beatty (NISRA) represented their respective UK Census Offices in a good spirit of constructive consultation. This reports on the second part of the Joint Workshop on Census Issues held at the University of Leeds, 12-14 May 1998. The first part of the Workshop focused on *The One Number Census: a Research Workshop* and was organised by Ludi Simpson on behalf of the Office for National Statistics, Advisory Groups and 2001 Census Working Groups and the ESRC funded research project *Quantifying the impact of census non-response on social policy and social research* under award R000236963. Sincere thanks are due Ludi Simpson for his efficient organisation and liaison.

Several colleagues gave their time to chair the different sessions of the Workshops and thanks are due to Paul Williamson (University of Liverpool), David Martin (University of Southampton), Paul Boyle (University of Leeds). The administrative arrangements for the Workshop were very efficiently organised by Christine Macdonald (University of Leeds) and spacious rooms and quality refreshments and meals were provided by Weetwood Hall, the Village Hotel and the Travel Inn.

ABSTRACT

The Census of Population is a very large exercise in data collection and processing. In 2001 some 25 million households in the United Kingdom will be contacted and asked to provide answers to a simple questionnaire of 25 to 30 questions. Such a task is likely to cost £125-150 millions to the Census Offices. Purchase of the data for academic research purposes is likely to cost ESRC and JISC some £1.5 to £2 millions directly and an equivalent amount indirectly on support over the following decade. It is therefore essential that the Population Census is very carefully planned beforehand and that the greatest possible value is extracted from the data collected.

This edited collection of papers reports on presentations and discussions in the Fourth and Final Workshop in the series *Workshops Planning for the 2001 Census - Determining Academic Community Needs and Strategy.* The Fourth Workshop was entitled *The 2001 Census: What do we really really want?*. The aim was to gather together and summarise the principal recommendations of the First (Geography), Second (Interfaces) and Third (Special Data Sets) Workshops. The Workshop was twinned with another on *The One Number Census: A Research Workshop*, organised by Ludi Simpson. The One Number Census project is a major undertaking by the Census Offices to deal with anticipated underenumeration by estimating how many households and people are missed by the standard enumeration.

Part 1 of the report on *Look Up Tables and Area Statistics* contains chapters by David Martin on the output geography proposals for 2001, by Seraphim Alvanides and Stan Openshaw on further developments to the methods being used to define output areas and by Bob Barr on what the Look Up Tables associated with the 2001 Census should be like. These chapters contain key recommendations on census output geography. Part 2 of the report puts forward recommendations for the preparation of *Microdata - Samples of Anonymised Records and Longitudinal Data* from the 2001 Census. Angela Dale summarises the conclusions of the SARs Sub-Group of the Census Offices' Output Working Group. Brian Dodgeon and Heather Joshi document the essential features of the 2001 Census Link to the Longitudinal Study and make a final plea for some new questions. Part 3 of the report reviews proposals for the improvement of *Interaction Statistics from the 2001 UK Census*. Paul Boyle and Phil Rees make radical proposals for revamping the provision of *Migration Statistics*. Martin Frost concentrates on ways of improving the accuracy of the *Workplace Statistics*. The fourth part of the report gathers together recommendations about information technology interfaces to census data, arguing that the tools and infrastructure are now in place to make networked and standalone access to the different types of data so much easier for the new user. Donald Morse and Alistair Towers review what interfaces to boundary data should look like. James Harris argues for the development of interfaces based on general data standards and the Web to access census statistics. Oliver Duke-Williams outlines how complex migration statistics can be presented for access in a simpler and easier to use interface. Ian Turton identifies how current software developments in Java progamming will make possible delivering easy to use interfaces to microdata very simple. Finally, Paul Williamson describes a design of a data dictionary for all census data sets. In Part 5, recommendations are summarised. Phil Rees reports on the views of 140 respondents drawn from the different corners of the academic community. The final pages try to draw out some general points from the very large number of recommendations made in the Fourth Workshop.

LIST OF SPEAKERS

Seraphim Alvanides
Geography, University of Leeds, Leeds LS2 9JT
Tel: 0113 233 3348, Fax 0113 233 3308, Email: s.alvanides@geog.leeds.ac.uk

Bob Barr
Geography, University of Manchester, Manchester M13 9PL
Tel: 0161 275 3648, Fax: 01925 750911, Email: r.barr@man.ac.uk

Paul Boyle
Geography, University of Leeds, Leeds LS2 9JT
Tel: 0113 233 3325, Fax 0113 233 3308, Email: p.boyle@geog.leeds.ac.uk

Angela Dale
CCSR, Faculty of Economic and Social Studies, University of Manchester M13 9Pl
Geography, University of Leeds, Leeds LS2 9JT
Tel: 0161 275 4876, Fax: 0161 275 4722, Email: Angela.Dale@mailhost.mcc.ac.uk

Brian Dodgeon
Centre for Longitudinal Studies, Institute of Education, 20 Bedford Way, London WC1 0AL
Tel: 0171 580 1122, Fax: 0171 612 6089, Email: ls@ioe.ul.ac.uk

Oliver Duke-Williams
Geography, University of Leeds, Leeds LS2 9JT
Tel: 0113 233 3309, Fax 0113 233 3308, Email: o.duke-williams@geog.leeds.ac.uk

Martin Frost
Kings College London, The Strand, London WC2R 2LS
Tel: 0171 873 2622, Fax: 0171 873 2287, Email: martin.frost@kcl.ac.uk

James Harris
Manchester Computing, University of Manchester, Manchester M13 9PL
Tel: 0161 275 6060, Fax: 0171 275 6040, Email: james.harris@man.ac.uk

David Martin
Geography, University of Southampton, Highfield, Southampton SO17 1BJ
Tel: 01703 593 808, Fax: 01703 593 729, Email: d.j.martin@soton.ac.uk

Donald Morse
Data Library, Computing Service, University of Edinburgh, George Square, Edinburgh EH8 9LJ
Tel: 0131 651 1241, Fax: 0131 650 6547, Email: donald.morse@ed.ac.uk

Philip Rees
Geography, University of Leeds, Leeds LS2 9JT
Tel: 0113 233 3341, Fax 0113 233 3308, Email: p.rees@geog.leeds.ac.uk

Ian Turton
Geography, University of Leeds, Leeds LS2 9JT
Tel: 0113 233 3309, Fax 0113 233 3308, Email: i.turton@geog.leeds.ac.uk

Paul Williamson
Geography, University of Liverpool, Roxby Building, Liverpool L69 3BX
Tel: 0151 794 2854, Fax: 0151 794 2866, Email: william@liv.ac.uk

LIST OF PARTICIPANTS

Seraphim Alvanides
School of Geography, University of Leeds, Leeds, LS2 9JT

Felicity Andrew
PPRU, Oxfordshire County Council, County Hall, OXFORD, OX1 1ND

Richard Arnold
Dept Epidemiology & Public Health, Imperial College, Norfolk Place, London, W2 1PG

Colin Bainbridge
Policy Development, North Yorkshire County Council, County Hall, Northallerton, DL7 8AH

Greg Ball
The Environment Department, Shropshire County Council, The Shirehall, Abbey Foregate, Shrewsbury, SY2 6ND

Robert Barr
School of Geography, University of Manchester, Oxford Road, Manchester, M13 9PL

Robert Beatty
Northern Ireland Statistics and Research, Arches Centre, 11-13 Bloomfield Avenue Belfast, BT5 5HD

Amanda Bellringer
General Register Office for Scotland, Ladywell House, Ladywell Road, Edinburgh, EH12 7TF

Jennifer Boag
Falkirk Council, Chief Executives Office, Municipal Building, Falkirk, FK1 5RS

Tim Bolderson
Global Business Network, 209 Business Design Centre, 52 Upper Street, London, N1 0QH

John Boyle
Economic Affairs Division, Illiam Dhone House, 2 Circular Road, Douglas, Isle of Man, IM1 1PQ

Paul Boyle
Department of Geography, University of Leeds, Leeds, LS2 9JT

James Brown
Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ

Chris Brunsdon
Dept. of Town & Country Planning, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU

Lisa Buckner
Room 4200W, Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 5RR

Julian Calder
D2/09 Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ

Jackie Carter
Manchester Computing, University of Manchester, Oxford Road, Manchester, M13 9PL

Ray Chambers
Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ

Tony Champion
Department of Geography, The University, Daysh Building, Newcastle Upon Tyne, NE1 7RU

Roma Chappell
Population Estimates Unit, Room 2300, ONS, Segensworth Road, Titchfield, Fareham, PO15 5RR

John Charlton
Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ

Martin Charlton
Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU

Jim Clark
General Register Office for Scotland, Room 1/1/1, Ladywell House, Ladywell Road, Edinburgh, EH12 7TF

Samantha Cockings
Dept of Epidemiology & Public Health, Imperial College at St Mary's, Norfolk Place, London, W2 1PG

Garnet Compton
General Register Office for Scotland, Ladywell House, Ladywell Road, Edinburgh, EH12 7TF

Mike Coombes
NE.RRL., CURDS, University of Newcastle, Claremont Bridge, Newcastle upon Tyne, NE1 7RU

Marie Cruddas
Office for National Statistics, Segensworth Rd, Titchfield, Fareham, PO15 5RR

Angela Dale
CCSR, Faculty of Economic and Social Studies, University of Manchester, Dover Street, Manchester, M13 9PL

Hywel Davies
Dept of Demographic & Social Studies, London Research Centre, 81 Black Prince Road, London, SE1 7SZ

Chris Denham
Census Services Branch, Office for National Statistics, Segensworth Road, Titchfield, Fareham, PO15 5RR

Roger Denton
Norwich City Council, Chief Execs. Department, City Hall, Norwich, NR2 3BE

Roger Dewhurst
49 Brownsvill Road, Heaton Moor, Stockport, Greater Manchester, SK4 4PF

Ian Diamond
Department of Social Statistics, University of Southampton, Southampton, S09 5NA

John Dixie
Office for National Statistics, Segensworth Road, Titchfield, Fareham, PO15 5RR

Brian Dodgeon
Centre for Longtitudinal Studies, Institute of Education, 20 Bedford Way, London, WC1 OAL

Daniel Dorling
Department of Geography, University of Bristol, University Road, Bristol, BS8 1SS

Keith Dugmore
8 Hugh Street, London, SW1V 1RP

Oliver Duke-Williams
School of Geography, Universtiy of Leeds, Leeds, LS2 9JT

Jason Dykes, Department of Geography, University of Leicester, Bennet Building, Leicester, LE1 7RU

Aodan Edmonds
Department of Geography, Trinity College, Dublin, Ireland

Piers Elias
Tees Valley Joint Strategy Unit, PO Box 199, Melrose House, Melrose Street, Middlesborough, TS1 2XF

Heather Eyre
School of Geography, Universtiy of Leeds, Leeds, LS2 9JT

Jan Finnigan
Tyne & Wear Research and Information, Room 146, Civic Centre, Newcastle upon Tyne & Wear, NE1 8QN

Graham Foster
Directorate of Planning Services, The Town Hall, Hornton Street, London, W8 7HX

A Stewart Fotheringham
North East Regional Research Lab, CURDS, University of Newcastle, Newcastle upon Tyne, NE1 7RU

John Fox
ONS, 1 Drummond Gate, London, SW1V 2QQ

Forest Frankovitch
NHS Executive North West, 930-932 Birchwood Bowland, Birchwood, Warrington,WA3 7QW

Jan Freeke
Planning Department, Glasgow City Council, 231 George Street, Glasgow, G1 1RX

Martin Frost
Department of Geography, Kings College London, The Strand, London, WC2R 2LS

Linda Frost
Planning Studies, Envi & Development, Manchester City Council, PO Box 463, Manchester, M60 3NY

Chris Gardiner
School of Urban and Regional Studies, Sheffield Hallam University, Sheffield, S1 1WB

Isobel Gibson
41 Jubilee Close, Ledbury, Herefordshire, HR8 2XA

Gillian Goddard
Room 443 B, Department of Health, Skipton House, 80 London Road, London, SE1 6LH

Robin Hall
Planning and Estate Department, Chesterfield Borough Council, Town Hall, Chesterfield, S40 1LP

James Harris
Manchester Computing, University of Manchester, Oxford Road, Manchester, M13 9PL

Tom Hennell
NHS Executive North West, 930 - 932 Birchwood Bowland, Birchwood, Warrington, WA3 7QW

John Hollis
London Research Centre, 81 Black Prince Road, London, SE1 7SZ

Stuart Holroyd
The Environment Department, Shropshire County Council, The Shirehall, Abbey Foregate, Shrewsbury, SY2 6ND

Jan Howard
Land Use Strategy, Milton Keynes Council, Civic Offices, PO Box 114, 1 Saxon Gate East, Milton Keynes, MK9 3HQ

Eileen Howes
London Research Centre, 81 Black Prince Road, London, SE1 7SZ

Norman Jamieson
City of Edinburgh Council, City Development Dept., 1 Cockburn Street, Edinburgh, EH1 1ZH

Stephen Jones
Kirklees Metropolitan Council, Devt. Unit, 2nd Floor, Civic Centre 111, Huddersfield, HD1 2EY

Graham C Jones
Room 4300W, Office for National Statistics, Segensworth Road, Titchfield, Fareham, PO15 5RR

Tim, Jones
Office for National Statistics, Zone D2/10,/1, Drummond Gate, London, SW1V 2QQ

Heather Joshi
Centre for Longitudinal Studies, Institute of Education, 20 Bedford Way, London, WC1 OAL

Mandy Kingston
The NENE Centre for Research, NENE University College, Northampton Park Campus, Boughton Green Road, Northampton, NN2 7AH

Catherine Knight
Research and Intelligence Unit, Cheshire County Council, County Hall, Chester, CH1 1SF

Neil Lander-Brinkey
Office for National Statistics, Census, Population & Health Group, Segensworth Road, Titchfield, Fareham, P016 7UU

Rachel Leeser
London Research Centre, 81 Black Prince Road, London, SE1 7SZ

Rob Lewis
London Research Centre, Parliament House, 81 Black Prince Road, London, SE1 75Z

David Lloyd
Prescribing Support Unit, Brunswich Court, Bridge Street, Leeds, LS2 7RJ

Kevin Lynch
Social statistics Research Unit, The City University, Northampton Square, London, EC1V OHB

Kenneth MacKinnon
Ivy Cottage, Ferintosh, The Black Isle, Ross-shire, IV7 8HX

Nick MacReady
Ordnance Survey,Romsey Road, Southampton, SO16 4GU

Pat Mann
ONS, Segensworth Road, Titchfield, Fareham, PO15 5RR

David Martin
Department of Geography, University of Southampton, Highfield, Southampton, S017 1BJ

Peter Maxson
Dept of Planning, Trans & Econ. Strategy, Warwick County Council, PO Box 43, Shire Hall, Warwick, CV34 4SX

Mark McConaghy
Room B4/11, Office for National Statistics, Drummond Gate, Pimlico, London, SW1V 2QQ

Elizabeth Middleton
CCSR, Faculty of Economic and Social Studies, University of Manchester, Dover Street, Manchester, M13 9PL

Robert Moore
Census Advisory Committee, Bennachie, Carmel Road, Holywell, Flintshire, CH8 7DD

Roger V Morgan
Room 328, The Town Hall, Hornton Street, London, WB 7NX

Donald Morse
UKBORDERS Service, Data Library, Computing Service, University of Edinburgh, George Square, Edinburgh, EH8 9LJ

Joanna Southworth
School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS

Malcolm Spittle
Policy Development Unit, Environmental. Services, North Yorks C. C, County Hall, Northallerton, DL7 8AH

Andy Teague
Office for National Statistics, Census, Titchfield, Fareham, PO15 5RR

Frank Thomas
General Register Office Scotland Census, Ladywell House, Ladywell Road, Edinburgh, EH12 7TF

Alistair Towers
Data Library, Main Library, George Square, Edinburgh, EH8 9LJ

Eddie Turnbull
General Register for Scotland, Room 1/1/2, Ladywell House, Ladywell Road, Edinburgh, EH12 7TF

Steve Turner
Tees Valley Joint Strategy Unit, PO Box 199, Melrose House, Melrose Street, Middlesborough, TS1 2XF

Ian Turton
The School of Geography, University of Leeds, Leeds, LS2 9JT

Nigel Walford
School of Geography, Kingston Polytechnic, Penrhyn Road, Kingston Upon Thames, Surrey, KT1 2EE

Judith Walton
Office for National Statistics, Room 2300, Segensworth Road, Titchfield, Fareham, PO15 5RR

Margaret Ward
The Data Archive, University of Essex, Wivenhoe Park, Colchester, C04 3SQ

Simon Whalley
Flat 4, 199 Kirkstall Lane, Headingley, Leeds, LS6 3EJ

Dick Wiggins
Department of Sociology, City University, Northampton Square, London, EC1V 0HB

Paul Williamson
Department of Geography, University of Liverpool, Roxby Building, Liverpool, L69 3BX
Tom Wilson
School of Geography, University of Leeds, Leeds, LS2 9JT

Bill Wood
Chief Executive's Office, Essex County Council, County Hall, Chelmsford, Essex, CM1 1LX

INTRODUCTION

Phil Rees

This document reports on the presentations and discussions of the Fourth ESRC/JISC Workshop Planning for the 2001 Census. The Fourth and Final Workshop was held on 13 and 14th May 1998 at the University of Leeds. At the Final Workshop the recommendations from the First, Second and Third Workshops and the Census User Survey of Views on the 2001 Census were debated and revised for presentation to ESRC/JISC and, for information, to the UK Census Offices.

The views expressed in the papers and discussion are those of the individuals concerned and do not represent the official positions of the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC), the Office for National Statistics (ONS), the General Register Office Scotland (GROS) or the Northern Ireland Statistics and Research Agency (NISRA).

However, the purpose of the Workshop was to present ideas, proposals and suggestions for the production of datasets from the 2001 Census, to be given serious consideration by both ESRC/JISC and the UK Census Offices for adoption in the lead up to the 2001 Census.

The report is divided into four main parts, preceded by this Introduction and succeeded by a Summary Part Five. Each part contains reports on the presentations made at the Workshop, on the discussions of each presentation and on the recommendations put forward in small group discussion. Part One of the report makes recommendations about the geography which will underpin the production of area statistics and the look up tables which will link households and output areas to other geographies. Part Two of the Report contains the presentations and discussions relating to the 2001 Census Samples of Anonymised Records and to the addition of a 2001 Census link to the Longitudinal Study. Part Three concerns Interaction Statistics from the 2001 UK Census. These include Workplace Statistics and the Migration Statistics. The lessons of experience with the 1991 Special Interaction Statistics are addressed and the authors put forward proposals for improvement. Part Four of the report makes recommendations about interfaces to and metadata about census statistics. The Final Part summarises the recommendations put forward by speakers and discussants.

At the conclusion the report, the points made in the Workshop discussion are gathered together. This discussion took place after the main presentations. Recommendations throughout the report are highlighted for easier reference in italics and given a number and a letter representing the data set to which they refer: G for geography, S for the SARs, L for the LS, M for the Migration Statistics, W for the Workplace Statistics, B for boundary data, I for Interfaces and Metadata and Q for recommendations arising from the user questionnaire survey.

PART I: LOOK UP TABLES AND AREA STATISTICS

CHAPTER 1
OUTPUT GEOGRAPHY FROM THE 2001 UK CENSUS: RECOMMENDATIONS

David Martin

1.1 Introduction

Various commentators have noted widespread dissatisfaction with aspects of previous census output geographies. Particularly, Coombes (1995) proposes seven 'tests' of census geography, which illustrate the multiple and conflicting criteria which might be used, while others such as Openshaw (1995) are critical of the 'neglect of GIS' in 1991 census outputs. Many alternative geographies have been proposed, and aspects of the debate concerning 2001 geography are collected in Rees (1997). The basic issues concern the separation of the output geography, based on output areas (OAs) from the collection geography, based on enumeration districts (EDs), and the rules for the creation of OAs. In particular, it is noted that EDs are created to maximise enumeration efficiency, which rarely provides an optimal geography for data output. This brief paper discusses matters arising from experiments conducted by the author in association with the Office for National Statistics (ONS) on 2001 output geography design procedures, and makes a series of recommendations for 2001 geography.

1.2 Current developments

An Arc/Info application was developed at ONS for the design of EDs for the 1997 census test, known as the geography area planning system (GAPS), which is described more fully in Martin (1998). GAPS permitted interactive placement of ED boundaries using Ordnance Survey ADDRESS-POINT, Boundary-Line and raster 1:10000 mapping and the 1991 census geography as represented by ED-line with real-time computation of expected population sizes and enumeration difficulty gradings. For the first time, GAPS permitted the production of individual maps and enumeration record books containing address lists for each ED.

On the basis of the GAPS GIS, a series of experiments were conducted using external programs to construct a separate output geography for the areas included in the 1997 test, which are described in Martin (1997). Essentially, the automated zoning procedure (AZP) described by Openshaw and Rao (1995) was adapted to the design of OAs by the aggregation of unit postcodes within wards. This entailed the creation of unit postcode boundaries, by Thiessen polygon creation around individual address locations, constrained within ED boundaries and subsequently merged to form unit postcode polygons. The contiguity matrix from these postcode polygons was then used in combination with specially tabulated unit postcode level output from the 1991 census database as input to the zone design program. Target population sizes and a threshold size, a tenure-based homogeneity measure, and crude shape constraint were used. The program easily constructs many more, more uniformly sized, above-threshold OAs than 1991 EDs for the same area, even when higher population thresholds are used. This work demonstrated the practicality of directly computing an output geography in the ONS environment, drawing on the data sources already held within GAPS. A new phase of work is now under way in which a working prototype output area production system (OAPS) is currently under development. The new system will be fully integrated with ONS' office systems and offer the potential for testing alternative boundary constraints, target population sizes and other design criteria. Development of this system has involved experimentation with the creation of more sophisticated unit postcode polygons by the combination of Ordnance Survey Land-Line data with the existing data series noted above.

It should be noted that no formal decision has yet been taken as to the nature of 2001 output geography, but that the general approach described here is currently being actively pursued through consultation with users during summer 1998. The survey of census users reported by Rees (1998) showed a majority of (academic) respondents broadly favouring this type of output geography over the use of 1991 or 2001 EDs as 2001 OAs.

1.3 Discussion

Rees' (1998) recommendation 9 is that 'the UK Census Offices should design output areas for small area statistics based on unit postcodes that fit into statutory areas on a best fit basis.' The second most favoured option in Rees' survey was to use 1991 EDs, but this would actually be extremely difficult to put into practice. In the 1991 geography, 1981 EDs were to be retained as much as possible, but statutory boundary changes and substantive population changes necessitated modification of the boundaries of no less than 68% of 1981 EDs. The very nature of the census requires that data be reported for statutory areas. In Scotland in 1991 unit postcode approximations to statutory areas were considered acceptable, but there is currently little support for such approximations from local authorities in England and Wales.

The approach described above uses unit postcodes (or part postcodes where they are split by statutory boundaries) as the basic building block of the output geography. There has been longstanding demand for closer integration between postcode and census geographies, a clear recommendation of the Chorley Report (DoE, 1987), but widely articulated by contemporary census users (eg. Dugmore, 1996). There are still no definitive boundaries for unit postcode polygons in England and Wales, although these have existed for a long time in Scotland, and the creation of these boundaries is a significant stage in the zone design process. It should be noted in passing, that convergence of the approaches used throughout the UK would be enormously helpful to census users and might encourage more truly national analyses of the data.

During the course of discussions and workshops concerning 2001 output geography commentators such as Barr and Openshaw have argued that 2001 geography should be derived from individual addresses and large scale digital mapping without recourse to postcodes as intermediate building blocks. The current work on the creation of OAPS suggests that such an approach would also be extremely difficult to implement, not due to any conceptual or technical barriers, but simply due to the complexity of identifying and extracting suitable digital features from (eg.) Land-Line to form meaningful polygon boundaries, which directly affect the contiguity matrix used in output area design. Road boundaries and major physical features simply do not provide sufficient density of topologically structured information to adquately subdivide residential geography. Meaningful but much less significant features such as garden fences and field boundaries are very difficult to extract from Land-Line in a structured way. It seems likely that achieving agreement on the construction rules which would effectively create a national dataset of computable address polygons could prove even more elusive than agreement on a single census output geography! Despite their weaknesses and tendency to change over time, postcodes provide a grouping of residential addresses which is strongly related to built form, and which remain critical for many organisations' use of census outputs. A range of options are available for the specific rules to be used in zone design, with some parameters such as threshold population sizes yet to be set by political decisions external to the geography design process. Full integration of OAPS with the ONS office system will permit the use of more sophisticated boundary shape controls by giving access to the boundary information held in the GAPS GIS. The use of social homogeneity measures remains a matter for careful consideration, both in principle and practice. Whatever design constraints and building blocks are used, the provision of a full set of digital boundary and centroid products should be an essential component of 2001 output, and it is strongly suggested that a

generalised boundary set, suitable for medium-scale thematic mapping, be made available at very low cost as part of the basic census outputs.

Another option raised during academic discussions of 2001 geography has been that of user-defined output areas. The approach being developed here is in no way incompatible with the creation of such output areas, and it potentially provides the basis for an ongoing national population GIS to be maintained at ONS beyond the census. The inclusion of an address-level base to such a system would effectively provide a baseline for management of all future boundary change and redistricting, and would permit change over time to be traced. Although such a system would by necessity reside within ONS, it is to be hoped that a framework might be developed for collaborative work between ONS and the academic community in which these 'data in a safe setting' (Marsh et al., 1994) might form the basis for innovative new research.

1.4 Recommendations

The following recommendations take account of development work at ONS and academic debate concerning census output geography held over the last two years, but remain the author's own. In this brief paper it has not been possible to address every aspect of the debate, but these recommendations draw on the wider discussion, and not just the issues selected for comment above.

Recommendation G1. *There should be separate collection and output geographies. The requirements of each boundary system are very different, and there are no longer any technical reasons why they should be kept the same.*

Recommendation G2. *There should be a standard output geography, which will form the basis for the major area-based statistical outputs. It will be necessary for this standard geography to match statutory boundaries in place at the time of the census. This should not rule out the subsequent availability of larger custom output areas which may prove invaluable for certain kinds of study.*

Recommendation G3. *The standard output geography should match unit postcode geography as far as possible. As discussed above, the postcode geography is still heavily used by many organisations and, equally important in this context, provides a series of building blocks from which an output geography can be constructed.*

Recommendation G4. *Output geography design should be conducted post-enumeration, in order to ensure maximum possible agreement between the different databases involved in its creation. This should make possible recommendations 5 and 6:*

Recommendation G5. *There should be no below-threshold OAs, although the level of threshholding is not yet known, and there may be a requirement to keep parishes in some outputs, even though there are likely to be sub-threshold parishes.*

Recommendation G6. *There should be a series of wholly digital output geography and ancillary products, at more than one level of generalisation.. Automated creation of output geography makes possible the creation of a single national geography dataset, which is fully internally consistent and integrated with the small area statistics, or equivalent. It should contain directly computed centroids for postcodes and OAs, and should be available at more than one level of generalisation, with a low cost dataset directly available for thematic mapping applications.*

Recommendation G7. *Directly comparable geography systems should be used throughout the UK, particularly concerning the rules for OA construction including target population size and relationships with statutory and postcode geographies.*

Recommendation G8. *The 2001 census geography systems should form the basis for an integrated national population GIS, to be maintained by the census offices after 2001. Such a system would permit the direct recomputation of population information for all subsequent boundary reorganisations, and provide the definitive geographic base for all ONS data linkage from the address level upwards.*

Acknowledgement

Much of the work reported here has been undertaken in association with the Census Division at ONS. Grateful thanks are offered for the guidance and assistance of ONS staff, but responsibility for all views and opinions expressed in this paper remains that of the author.

References

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

Department of the Environment (1987) *Handling geographical information: the report of the Committee of Enquiry chaired by Lord Chorley*. London: HMSO

Dugmore, K. (1996) What do users want from the 2001 Census? In *Looking towards the 2001 Census*. OPCS Occasional Paper 46. London: OPCS 21-23

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

Martin, D. (1997) From enumeration districts to output areas: experiments in the automated creation of a census output geography. *Population Trends* **87**, 36-42

Martin, D. (1998) Census geography 2001: designed by and for GIS? In Carver, S. (ed.) *Innovations in GIS 5* London: Taylor and Francis

Openshaw, S. (1995) The future of the census. In OPENSHAW, S., ed, *Census Users' Handbook*. Cambridge: GeoInformation International 389-411

Openshaw, S. and Rao, L. (1995) Algorithms for reengineering 1991 Census geography. *Environment and Planning A* **27**, 425-46

Rees, P. (ed.) (1997) *The debate about the geography of the 2001 census: collected papers from 1995-6* Working paper 97/1, School of Geography, University of Leeds

Rees, P. (1998) What do you want from the 2001 Census? Results of an ESRC/JISC survey of user views Paper presented at the Annual Conference of the RGS/IBG, University of Surrey, 6-8 January

CHAPTER 2
SOME FURTHER EXPERIMENTS WITH DESIGNING OUTPUT AREAS FOR THE 2001 UK CENSUS

Stan Openshaw, Seraphim Alvanides and Simon Whalley

2.1 Introduction

Martin (1997, 1998) describes a method for automating the design of census output geography and presents some results he obtained for a test area. This paper seeks to confirm these results and build upon this pioneering approach via the use of finer resolution spatial data and a more sophisticated version of zone design algorithms than was used by Martin. The topic is very important because of proposals by the Office of National Statistics to consider using output areas in England and Wales based on a new and uniform geography that nest within wards. This new geography would have consistently defined properties, resulting in as small as possible zones of similar size, compact shapes and reasonably homogenous in terms of a few key variables. Additionally, the digital representations of these new output areas will be accurately known (without any need for digitisation) and their composition in terms of postal addresses precisely defined. If these plans materialise, it will constitute the most significant advance in census geography for 30 years.

The 2001 census will be the first UK census for which the definition of the areas used to report the small area census results (the output areas) could be designed by a computer procedure to have a consistent set of common properties. The release of small area census data for census enumeration districts was a major innovation in census outputs that dates back to the 1961 census. Census enumeration districts (EDs) are a Victorian concept and reflected the desire to divide up the country so as to equalise the work load of the census enumerators. As such it was traditionally the smallest geographical entity for which census data could be fairly readily provided. The 1961 small area statistics for EDs experiment was a consequence of the computerisation of the census and has ever since been repeated for each subsequent census. The task of designing the EDs (traditionally a paper exercise based on large scale paper maps) was also broadened to allow local authorities to make an input if they so wished. Increasingly census EDs became a very useful census data reporting framework that has been used in a wide range of research, planning and commercial geodemographic applications. It can be classed as geographical information since each census ED has been given an approximate 100m grid-reference.

Unfortunately census EDs are a non-ideal basic spatial unit reporting small area census data for a number of reasons:

1. they are not homogenous,
2. they vary greatly in population size,
3. they are not stable over time,
4. they vary in shape and physical size,
5. they are not designed in any formal or consistent way to conform to an explicit set of fixed rules,
6. they only poorly relate to unit postcode geography in England and Wales and the situation in Scotland is fast becoming chaotic due to changes occurring in postal geography,
7. the digital boundaries and the maps belong to agencies other than the census or as Openshaw (1996) put it the "arcs" are owned by the OS and the "info" by ONS as the census has not been traditionally viewed as integrated geographical data source,
8. end-users have not previously been permitted to design their own EDs but have to make do with the official ones and

9. the approximate x, y coordinate point references for Eds are an inadequate form of spatial representation for an increasing number of purposes which require more accurate levels of spatial resolution.

Until fairly recently nothing much could be done about it because of a mix of technical, methodological, and data reasons. However, the situation is now quite different. There are a number of major changes which are very relevant, in particular:

1. GIS is now a universally available and mature technology,
2. the OS Address Point product is complete and this claims to provide an accurate 1m grid-reference for each postal address in the UK.
3. the users of census data are far more experienced and aware than previously and have increasingly demanding and more diverse expectations,
4. much more is now known about the degrees of geographical freedom involved in zone design as a result of research by Openshaw and Rao (1995), Openshaw and Alvanides (1998a,b), and Martin (1997, 1998), and others;
5. computer speeds have increased dramatically since the 1991 census was taken and are continuing to do so, whilst hardware prices have fallen equally dramatically making what would have been prohibitively expensive in 1991 extremely very affordable by 1998, and
6. the census faces an increasing number of alternative "census like" data products that offer far high levels of geographical precision and flexible outputs.

In an age when users are more aware of modifiable areal unit problems and zone design tools (such as ZDES) are becoming widely available over the internet, it is no longer acceptable to merely ignore the zone design aspects involved in the creation of census output areas on the grounds that it is not possible or feasible or affordable to do anything different or that most users do not care and are happy with the traditional products. If users claim so now, are they likely to be so unconcerned when the 2001 census results are eventually produced? Once ONS had no choice about how to perform this task, now they do and it would be quite remarkable if they were to continue to use the same old legacy style of output areas in 2001 as they did in 1991. Whether or not they are prepared to allow census users complete freedom to design their own output geographies subject to various confidentiality restrictions is still unknown and may be regarded as too ambitious for 2001. However, if not, then the onus is placed firmly on ONS to perform the output area design process in a far more rigorous and consistent manner than has historically been practical and seek to identify an output geography that is at a sufficiently fine level of resolution that it allows flexible re-aggregation sufficient to meet existing and still emerging new user needs.

This paper presents a methodology and a case study of how this task can be performed. Section 2 describes the zone design method being proposed here and how it builds on previous work by Martin (1997) and Openshaw and Rao (1995). Section 3 presents an empirical case study based on simulated Address Point census data. Finally Section 4 makes a number of suggestions about how to design better census output areas relevant to the 2001 census.

2.2 Zone DEsign System (ZDES)

The original automated zone design program (AZP) algorithm of Openshaw (1977) was originally developed to explore modifiable areal unit effects. Subsequently it was used to experiment with zone design. However, AZP was far ahead of its time and it was not until the mid 1990's that the increased availability of digital map data and fast workstation hardware resulted in its revival). The original AZP technology has been extended to form the ZDES system based on the work of Openshaw and Rao

(1995), but subsequently dramatically extended and improved; see Openshaw and Alvanides (1998a, b). ZDES in common with AZP views the zone design task as a special kind of combinatorial optimisation problem. The aim is to optimise a function of the data generated by a zoning system defining an aggregation of N original zones into M regions or output zones (M<N). Expressed in mathematical notation it looks like any other mathematical optimisation problem, viz

optimise F(Z)

where F(Z) is a general function of Z except Z is not a simple set of linear or non linear parameters but defines an aggregation of N initial zones into M output zones. Additionally, there are implicit constraints on Z such that each of the original N zones have to be assigned to exactly one output zone and all the members of the same output zone have to be connected so that when the internal boundaries are dissolved they form a single polygon. This optimisation task might be categorised as a constrained non-linear integer optimisation problem. It can only be solved via heuristic methods that may not find the global optimum result; indeed, there is no way of knowing whether there is a single global optimum result to find! The view here is that finding a global optimum may be less relevant than finding extremely 'good' results, however 'goodness' is to be measured. In practice the principle of caveat emptor needs to be applied.

The general function F(Z) need not be continuous nor convex for all feasible values of Z. Despite this complexity, the fairly simple Monte Carlo optimisation method, formerly called AZP (Openshaw, 1976), seems capable of providing what appear to be good solutions to many zone design functions. The other innovation with AZP was the benefit of viewing F(Z) as being any relevant function, thus suggesting that a single algorithm could in principle solve any zone design problem, ranging from electoral redistricting to location-allocation modelling simply by plugging in a different objective function. Additionally, the original ZDES code has also been extended so that ZDES v5 will now handle a wide range of unconstrained and constrained zone design problems. Note that the term constrained relates here to applications in which there are extra user defined constraints imposed on the zone design process which are additional to those contiguity and coverage restrictions already implicit in ZDES. The ZDES software can be downloaded from a WWW site together with some recent papers (ZDES, 1998). At present, we are working on developing UNIX portable and NT versions which will work on any UNIX or NT system supporting Arc/Info release 7.0 or above.

2.3 Designing census output zones

2.3.1 Overview of the process

Martin (1997) performed his research within the ONS data confidentiality barrier. This greatly complicated the research task and it is a tribute to his skill and enthusiasm that he managed to achieve so much. His building blocks were unit postcodes and he tried to retain ED boundaries wherever it was sensible to do so because of the belief that at least some of the boundaries had a more general utility. Each ward was treated separately. Wards are currently the smallest "official" geography in Britain and this restriction breaks the census output area design task into about 10,420 separate pieces as each ward has to be processed separately. Seemingly this is a very daunting task until it is realised that the same procedure is applied to each ward and the entire process can be automated. Additionally, this type of problem is very well suited for a parallel supercomputer as a separate processor can be assigned to each ward if necessary. Martin's use of postcode building blocks and his attempt to preserve ED boundaries may be questioned as unit postcodes are inherently unstable building blocks, they do not nest into wards in England and Wales, and the old ED 1991 boundaries are of immensely variable meaningfulness. Moreover, the use of linear features from OS digital databases also carries with it additional, perhaps large, data cost and royalty penalties which may well be passed on to the end users.

The current belief is that the use of land line data may well cost far more than any value it adds to the output zones. This is because the most important linear barrier features (e.g. rivers, railway lines, major roads, open spaces) are already implicit in the distribution of address points and these aspects may be more readily (and far more cheaply) handled by the choice of an appropriate objective function for the zone design process.

Another aspect is that unit postcode geography is probably reaching the end of its life. Postcodes are not viewed as a zoning system, they are optimised for postal delivery, they are unstable, and there is no standard specified for them. Nevertheless, postcodes are an extremely convenient geography. Postcodes trivialised the historic address matching task and provide a very easy way of attaching approximate X,Y coordinates to data relating to postal addresses and of matching address data to census EDs. However, now that a British Standard for addresses exists and there is a national address level gazetteer for the UK (Address Point), it becomes clear that postcodes are no longer so effective and may have already reached the end of their really useful life. As address matching methods improve over the next few years, postcode geography will become far less relevant and will increasingly come to be viewed as what they always were, namely a quick and dirty fix to a problem that once could not be solved in any other way. Accordingly, it is argued that the obvious building blocks for the 2001 and subsequent census output areas are the individual addresses in the OS Address Point file. These are the smallest, most accurate, and most stable basic spatial units that can exist. The 2001 census should, ideally, be based on an updated and corrected set of address points either based on the OS Address Point product or on an ONS created version using hand held GPS or some equivalent technology capable of yielding sub 1 metre locational references for all the addresses in the UK.

Address points are, by far, the most obvious building blocks for the 2001 census and for designing small area census and output area geography. They have postcodes attached so they can also be linked to postcode data and they should increasingly be machine address matchable. Unfortunately, the Address Point data are not yet available for academic research, so one of the authors (Whalley) spent many hours digitising all the address points within a ward in order to create a simulated address point coverage for a part of Sheffield.

2.3.2 Choice of constraints and zone design function

The question now is what properties should an ideal census output geography possess?
Cliff et al (1975) suggest that in general a good zoning system should be

1. as simple as possible,
2. homogenous and
3. compact.

Wise et al (1997) use similar criteria to (2) and (3) and suggest that the zones should also be of equal size. Martin (1997) uses population size, shape, and homogeneity. He writes "..the population objective is based on the minimisation of the total squared difference between output area population sizes and target population size ... the boundary constraint seeks to minimise the total squared boundary length of the output area ... the homogeneity constraint ... seeks to minimise the sum of the squared differences ..." (p.14)

In effect in all these zone design applications the objective function is the weighted sum of up to three different design functions. This is, unfortunately, not the best or most appropriate way of creating zoning systems that meet a set of design constraints. There is also another problem. In a census context obtaining identically sized zones is far less important than ensuring the smallest one exceeds a ONS specified minimum size. Similarly, there is no need to optimise homogeneity or squared boundary lengths per se because the desired optimal target values are not unknown. There is merely a need to

ensure that all the output zones exceed a minimum comparable degree of compactness and homogeneity. It is far more important to avoid the occasional straggling or spindly output zone that it is to insist that all zones are approximately circular in shape which is a highly unrealistic goal.

In essence the design functions used by Cliff et al., Wise et al., and Martin should be re-cast as inequality constraints. The homogeneity constraint can be set at some reasonably high arbitrary value, say 75%. The shape constraint is far more problematical because it is difficult to know what limits to use and it may not matter much, or at all, if occasionally a strangely shaped zone is produced. One solution would be to convert it into an objective function and then seek to minimise it; however, this may still produce highly eccentric shapes. Instead it is suggested that a better alternative is to use a population weighted accessibility function instead. A compact zoning system will tend to have a minimum sum of within zone travel distances around the point of minimum aggregate travel. This can be simply expressed as a set of M separate P-median problems, one for the members of each of the M output zones. Optimising the sum of these M P-median problems will tend to produce naturally compact zones that automatically adapts to the local distribution of address points.

2.3.3 Handling constraints in zone design

The problem that Martin, Wise et al and Cliff et al were previously unable to solve is how to optimise one design function (within zone population weighted distances) subject to inequality constraints on the values of the other zone design functions (size and homogeneity). It is this problem that ZDES can now solve. In ZDES some constraints are implicit in the algorithm and can never be violated (e.g. contiguity). Other design constraints that the user may wish to impose on the zoning systems are of the more usual type characteristic of mathematical programming, but even here the nature of the zone design task creates additional complications. These user-defined constraints can be applied to each of the individual M output zones, and, or, else relate to some characteristics of the data generated by all the output zones. For example, a minimum zone size inequality constraint can be applied to all M zones and an additional inequality restriction placed on the nature of the data generated by the complete set of zones; viz. that skewness is less than some threshold value. Examples of the latter are provided in Openshaw (1978a) when he generates zoning systems that maximise a correlation coefficient (as the objective function), subject to the data being approximately normally distributed and with zero spatial autocorrelation; both the latter are specified as inequality constraints specifying ranges of acceptable values. In a census output area design context the constraints can be restricted to size and homogeneity.

In general, there are two ways of handling these user defined constraints:

1. convert them into a single weighted function or
2. handle them via a far more sophisticated penalty function method borrowed from non-linear optimisation literature.

Most geographers have adopted the first approach, perhaps, without realising its key deficiencies.

Thus Cliff et al (1975), Wise et al (1997), and Martin (1997) all use a sum of weighted constraint violations. For example, the SAGE system of Wise et al (1997) allows the user to provide differential weights to three zone design functions (homogeneity, equality, and compactness). The user supplies weights to them (0 to 100%) as a means of trying to balance the different objectives. The problem with all these methods is that each function is measured in different units and they need to be standardised in an appropriate way. Martin (1997) writes "Setting each weight to 1.0 means that each is given an equal weighting, such that a 1% overall improvement in one measure is of equal attractiveness to a 1% overall improvement in another" (p14). The problem is that this scaling is not straightforward. The principal criticisms can be summarised as follows.

1. The zone design functions of the Cliff-Wise-Martin type are not constraints in a conventional mathematical optimisation sense; instead they are really multiple separate objective functions, the weighted sum of which are to be minimised in the hope that this delivers a good result. Unfortunately, there is no assurance that either any or all of these design functions will meet whatever minimum or maximum limits are placed upon them. The overall result is uncontrolled inconsistency from one ward to the next in relation to the design criteria.
2. The quality of the solutions depend entirely on the respective weighting given to the competing design functions and especially how each of them are scaled or standardised so that the previous quote from Martin (1997) is satisfied.
3. Function standardisation is not easy and it is almost impossible to determine how best to scale the competing functions. For example, how do you relate within zone sum of squares to squared boundary length of the output areas to squared deviations from a target population size? Each function has a different mean and variance that constantly change as the zoning system is modified.
4. In essence the design functions are being used as equality constraints but with extreme and quite possibly unreasonable target values (e.g. minimising the within zone sum of squares is equivalent to an equality constraint of zero and a goal of complete homogeneity) but sometimes with far more attainable, but impossible to exactly meet, targets (e.g. minimise deviations from mean zone size). Is it necessary or sensible to have equality constraints with right hand values (i.e. targets) of zero?
5. If non-zero targets are used then the problem reduces to finding a feasible zoning system that satisfied a set of inequality constraints. However, there are potentially multiple different solutions that may achieve this goal and for it to work there has to be an overall objective function to define which of the feasible solutions is best.
6. However, if the intention is to use multiple objective functions in zone design then consideration should be given to various multiple objective function handling methods, such as goal programming, in which there is an explicit trade-off between alternative functions. The problem is that this process is usually very complex and non-automatic involving interactive trade-offs.

A far better and simpler strategy is to select one of the design functions as the objective function and treat the others as equality or inequality constraints, setting realistic, explicit, and attainable target values for them.

2.3.4 *Penalty function methods used in non-linear optimisation*

Once the zone design problem is thought of as a mathematical optimisation problem (i.e. an objective function with constraints) then there are various methods for handling constraints on non-linear optimisation problems that may be used in a zone design context. The simplest is to add a penalty function to the objective function that reflects these constraint violations. The weighting given to these constraint violations is then gradually increased so that, hopefully, a sequence of unconstrained optimisations will gradually move towards a solution of the original constrained problem.

One way of operationalising this approach is as follows.
Let C_j be the j^{th} constraint violation; for example, if a minimum size constraint is applied so that for any zone j

$$C_j = \max (\text{target} - S_j, 0)$$

where
target is the minimum household size so that the constraint being imposed is in fact $S_j \geq$ target
and S_j is the household size of zone j

So, C_j either has a value of zero if S_j equals or exceeds the size threshold or is set equal to the degree by which the constraint is violated (i.e. the number of households that needs to be added to zone j to reach the size threshold).

A penalty function approach would now seek to re-arrange the zones in order to optimise a new objective function that has a penalty term which represents the constraint violations: For example, minimise a function such as

$$F(Z) + r_k^{-1}(\sum_j C_j(Z)^2$$

for successive values of the positive parameter r_k chosen so that it slowly moves to zero. The hope is that this sequence of unconstrained problems will converge with a solution to the constrained problem. The problem with these and related methods is that they tend to involve terms that can easily create very large numbers which causes major difficulties to the optimiser with results that are highly sensitive to small changes making them hard to handle. Also, the differential scaling issue involved in trading off one set of constraints against another is still not being properly addressed. Finally, the optimisation method used to solve a particular unconstrained problem may easily become stuck and experiments indicate that this is particularly likely in a zone design context.

2.3.5 Powell-Fletcher method

Powell (1969) describes a far superior alternative to handling constraints. He developed a modified penalty function approach that has two sets of controlling parameters. This was later generalised to handle inequality constraints; see Fletcher (1987). It involves optimising a new function

$$\Phi(Z,\sigma,\theta) = F(Z) + 0.5\sum_i^n \sigma_i(C_i(Z) - \theta_i)^2$$

where

$\Phi(Z,\sigma,\theta)$ is a penalty function that is dependent on the zoning system X, σ_i and θ_i;

σ_i and θ_i are a series of parameters that are estimated to ensure gradual satisfaction of the constraints;

$C_i(Z)$ is a constraint violation which is some function of the zoning system and, or the zones in the zoning system;

F(Z) is the objective function and Z is the set of unknowns in the optimisation which is in fact the zoning system.

The required solution can usually be obtained for moderate values of the parameters, if one exists. The attraction is that it is easy to change these parameters to generate a suitable sequence of unconstrained problems. The key point here is that the σ_i and θ_i control parameters are changed slowly to ensure that, if at all possible, a constrained solution will be obtained. Fletcher and Powell provide convergence proofs and it only requires that the magnitude of the objective function and the various constraint functions are scaled so as to have similar magnitudes. As Openshaw (1978b) demonstrated it works extremely well on zone design problems and so far no better methods have been developed. It has the added benefit that this penalty function approach avoids many of the local minima that the original AZP algorithm is prone to discover because it can go and up down function gradients depending on the respective values of the control parameters. A version of ZDES has now been developed which implements this approach in a form suitable for designing census output areas. The question now is how well does it work.

2.4 Census output area experiments

2.4.1 Data creation

The study region used here is a ward in Sheffield that was selected because of the diversity of the characteristics it contained. The labour involved in creating simulated Address Point data was sufficiently great as to preclude examination of more than one ward. The address points were digitised from a Sheffield City large scale map (1:2000) representing the centroids of properties identified on the map (See Figures 1 and 5). The number of households at each address was inserted as an attribute; this assumes that each house number on the map represents one household. The resulting dataset replicates the Address Point data to sub-meter accuracy and consists of 5406 points representing 7700 households.

As no real 1991 individual census data were available for this area it was necessary to create other variables used to represent a measure of zonal homogeneity. The homogeneity criteria probably only really need to be at a fairly gross level of detail because it is unreasonable to expect high levels of social homogeneity even at this scale. Additionally, there may be some concern that designing zones to be "too homogenous" may cause subsequent statistical problems in analysing the data (affecting variances). On the other hand complete lack of any control for homogeneity is also undesirable, because systematic geographical differences in the levels of homogeneity do need to be controlled to some extent. Probably the best and most relevant variables to use are, therefore, either tenure type or building type. Here the former is used as it could be captured directly from the map that was used for the digitisation. There are two house tenures; council or privately owned and these are applied via the homogeneity criteria as used by Martin (1997).

The last stage of data creation involved the generation of two sets of 5406 thiessen polygons each, that would be used as building blocks. The first set was thiessen polygons restricted to linear features in order to experiment with Martin's (1997) approach; due to lack of OS data, ED boundaries were used as linear features (Figure 2). The second set was based on free thiessen polygons without any restrictions to their shape, clipped only to the ward boundary (Figure 6). The thiessen polygons were examined for inconsistencies and entered into ZDES for the creation of the contiguity matrix and the necessary attribute tables.

2.4.2 Objective function and constraints

The objective function most suitable here is the sum of population weighted within zone distances around the point of minimum aggregate travel for each output area, the so called P- median point. These accessibilities are summed for all M output regions to provide a useful indicator of output area compaction. Thus, the objective function is:

$$\text{minimise} \sum_{j}^{M} \sum_{i \in Mj}^{N} P_i D_{ij}$$

where
Pi is the population value for small area *Ni* (in this case the number of households);
Dij is the Euclidean distance between the address point and the population centroid of the output area *Mj* of which i is a member, note that this P-median centroid depends on the current membership of region j.

In addition the following constraints are used. The percentage homogeneity of any output zone is calculated as

$$H_i = 100 \, (\max (X_1, X_2) / (X_1 + X_2))$$

where X_1 is council tenure and X_2 is non-council tenure in zone i.
For current purposes each output zone is constrained so that: $H_i > 75$
where H_i is the homogeneity measure previously defined . A second set of size constraints are applied so that every zone exceeds an arbitrary size of 25, more formally: $S_j \geq 25$
where S_j is the number of households in output zone j. Note that with M output zones there are 2M inequality constraints that have to be satisfied before an acceptable solution is found.

2.4.3 An automated census output areas procedure

A brief outline of the principal stages in an automated census output area design process is shown in Figure 11. This is a modification of Martin (1997) but with a few key changes in the mechanisms for handling the constraints and it has a different objective function. It may be of passing interest to note that most of this processing is performed automatically inside the ZDES system. The choice of constraint values is fixed by the global design process, but the number of output areas is variable and is specific to each ward. In addition, the ward specific nesting could be replaced by a higher level of geography. ZDES compute times are linear in N and almost indifferent to the magnitude of M. So a Local Authority nesting would be quite feasible and the results would be better because of the reduced effects of ward boundaries.

The best way of dealing with the problem of how many output zones can be engineered is to guess the maximum number (given a minimum size level) and then slowly reduce it until a feasible and visually pleasing solution is obtained. For the 1991 census an output area had a minimum household count of 16. The objective here is to generate as many output areas as possible, given the design constraints. It is useful to note that the principal constraint on the nature of the output areas used to report census data is the need to preserve the confidentiality of the data. Currently this is defined only indirectly via minimum household and people size counts. This being the case, the obvious strategy from a user's point of view is to seek the maximum number of output zones that meet these restrictions. This is a far more useful design objective than merely insisting that most zones are of broadly the same size. In the limit, both two approaches are identical; for instance when all zones exactly match the ONS minimum (population and household) size. This objective is probably unrealistic but there is no reason to assume that zoning systems cannot be engineered that come close to meeting this goal.

The reason for wishing to maximise the number of output areas is, therefore, simply because this will maximise the subsequent utility of the small area data. Having more rather than fewer output areas will give those census users denied an opportunity to design their own geographies, maximum flexibility in the re-aggregation of the published small area census data. Indeed, if the areas are sufficiently small then there may well be no need for any other form of output geography as users could re-engineer these small zones to match their needs, also using ZDES but with the OAs as the building blocks.

2.4.4 Discussion of the results

Martin (1998) notes that in the three experiments he performed it was sometimes possible to more than double the numbers of output areas. In the Sandwell ward of Birmingham the 50 1991 EDs could be increased to 111, in Petersfield ward from 25 to 48 but in the Craven ward from 19 to 18. However, these increases are almost certainly smaller than a more sophisticated constrained handling zone design method would manage, as is demonstrated here.

Table 2.1 shows the statistical properties of the results of our experiments with the restricted thiessen polygons for the Sheffield ward, together with the ED properties for comparisons. The maximum number of households in output areas was arbitrarily set to 25, well above the official census figure of 16; the same figure being the minimum zone size constraint. As can be seen there comes a point where

no feasible result can be obtained. This can be fine tuned to whatever arbitrary degree of precision is desired. Selected outputs from these runs were mapped in figures 3 and 4 for 50 and 200 output zones respectively. The algorithm performed well, but the belief was that it might perform better with the non-restricted thiessens.

Table 2.2 shows the results for the unrestricted thiessens. Again the algorithm performed extremely well suggesting that about 290 output areas could be identified that met the design constraints (cf. 36 1991 census eds). However, as Figures 7 to 10 show, it is clear that some of the maps with the larger numbers of output areas have started to disintegrate and no longer show compact polygon shapes. The explanation is probably due to the homogeneity constraint. It would be sensible to select less than the maximum number of output areas. In this case somewhere between 200 (figure 8) and 250 (figure 9) output areas. Some indication of the shape disintegration can be seen in table 2.3 via a crude shape index that measures the internal arcs length divided by the number of polygons.

Table 2.1: Output area statistics with ED boundary barriers for Sharrow ward, Sheffield

Unit type	Number of Units	Status	Minimum Households	Maximum Households	Mean Households	*in_shape index*
EDs	36	n/a	73	408	214	*600*
ZDES	50	OK	40	269	154	*545*
ZDES	100	OK	26	206	77	*368*
ZDES	150	OK	25	141	51	*297*
ZDES	200	OK	25	109	39	*263*
ZDES	250	*no f.s.*	25	80	31	*280*

Notes: *in_shape:* (internal arcs length) / (number of polygons); *no f.s.:* no feasible solution

Table 2.2: Output area statistics without boundary barriers for Sharrow ward, Sheffield

Number of Units	Status	Minimum Households	Maximum Households	Mean Households	*In_shape index*
50	OK	65	295	154	*530*
100	OK	27	171	77	*362*
150	OK	25	129	51	*295*
200	OK	25	83	38	*264*
240	OK	25	125	32	*284*
250	OK	25	170	31	*273*
275	OK	25	62	28	*267*
285	OK	25	72	27	*291*
290	OK	25	60	27	*304*
295	*no f.s.*	25	63	26	*302*

Notes: *no f.s.:* no feasible solution

Table 2.3 Output Area Shapes

Number of Units	Shape Measure
50	530
100	362
150	295
200	264
240	284
250	273
275	267
285	291
290	304

2.5 Conclusions

The results confirm and add further support for the creation of a new style of output area geography consisting of zones that have been consistently defined, satisfy confidentiality restrictions, and yet are small enough to be used as building blocks in other user specific geographies and re-aggregations. The use of a more sophisticated zone design algorithm that permits a re-specification of the problem in an optimisation framework improves the quality of the results and puts the entire process on a stronger methodological basis. The same technology could also be used to create a second tier of geography by re-assembling the small areas into a second tier of new geography that was 1991 ED like or ward-like in terms of numbers but consistently defined to have common properties. This is an important point. If output areas 4 to 8 times smaller than 1991 EDs can be created then it would be fairly easy to re-zone these output areas to provide uniform geographies at other scales of resolution without risking differencing problems. It may also permit best fitting to historic census geographies that are now defunct (i.e. 1981 wards, frozen postcode sectors for 1991, parishes for 1971, etc.). Finally, these methods could also be used as a basis for an entirely user controlled flexible output geographic design process merely by adding an explicit confidentiality risk constraint to the zones that are produced. This is possible now or soon could be but may be this will be more appropriate for the 2011 rather than the 2001 census or for user testing in 2005.

If the new output areas are to be successful then there are five conditions that may need to be met: (1) a global agreement as to the design criteria and the values of the constraints, (2) the availability of a digital representation of the resulting areas (created from the zone outputs), (3) the creation of a fully connected topology without edges to permit subsequent re-aggregation to be easily performed, (4) the diffusion of the necessary zone design software to allow census end users to re-engineer their census geographies and (5) an end-user community sufficiently aware, now, well in advance of the census, to see the benefits of such a revolution in four or five years time. It is this latter condition that is probably the hardest of all to meet and it requires a proactive rather than a reactive response by ONS.

Recommendation G9. *The Census Offices should take the findings of the research carried out by Openshaw, Alvanides and Whalley into account when refining and finalising the procedures for defining output areas for the 2001 Census.*

Acknowledgements

The authors would like to acknowledge the use of 1991 Census data purchased by ESRC/JISC, for most of the analysis and mapping needs in this paper. We would also like to thank the GIS Unit, Planning Department of the Sheffield City Council that kindly provided the 1:2000 map for the modelling of the Sharrow ward address data. All data and maps hold Crown Copyright.

References

Cliff, A.D., Haggett, P., Ord, K, Bassett, K., Davies, R. (1975) *Elements of Spatial Structure* CUP, Cambridge

Fletcher R (1987) *Practical Methods of Optimisation*. Chichester, Wiley.

Martin D. (1997) Implementing an automated census output geography design procedure. Department of Geography, University of Southampton, Southampton. Draft 20/01/97 (Copies obtained by the author)

Martin D. (1998) Optimising census geography: the separation of collection and output geographies, *International Journal of Geographical Information Science* 12 (in press)

Openshaw, S. (1977) A geographical solution to scale and aggregation problems in region-building, partitioning, and spatial modelling. *Transactions of the Institute of British Geographers* 2, 459-72

Openshaw S. (1978a) An empirical study of some zone design criteria, *Environment and Planning A* 10, 781-794.

Openshaw, S. (1978b) An optimal zoning approach to the study of spatially aggregated data. In Masser I., Brown P. (eds.) *Spatial representation and spatial interaction*. Boston MA, Martinus Nijhoff, 95-113

Openshaw, S. (1984) The modifiable areal unit problem. *Concepts and Techniques in Modern Geography 38*. Norwich, UK, GeoBooks

Openshaw, S. (1996) *Census Users Handbook.* Cambridge, GeoInformation International.

Openshaw, S. (1996) Developing GIS relevant zone based spatial analysis methods. In P.Longley, M.Batty (eds.) *Spatial analysis: modelling in a GIS environment*. Cambridge, GeoInformation International: p55-73

Openshaw, S., Alvanides, S. (1998) Applying GeoComputation to the analysis of spatial distributions. In P.Longley, M.F.Goodchild, D.J.Maguire, D.W.Rhind (eds.) *GIS: Principles, Techniques, Management and Applications* GeoInformation International (forthcoming)

Openshaw, S., Rao, L. (1995) Algorithms for re-engineering 1991 Census geography. *Environment and Planning A* **27**, 425-46

Openshaw, S., Schmidt, J. (1996) Parallel simulated annealing and genetic algorithms for re-engineering zoning systems. *Geographical Systems* **3**, 201-20

Powell, M. (1969) 'A method for nonlinear constraints in minimisation problems', in R Fletcher (ed.) *Optimisation* Academic Press, London 283-298

Wise, S., Haining, R., Ma, J. (1997) 'Regionalisation tools for the exploratory spatial analysis of health data', in A Getis and M.M. Fischer (eds.) *Recent Developments in Spatial Analysis*, Springer, Berlin 83-100

ZDES (1998) htttp://www.geog.leeds.ac.uk/research/ccg/zdes3.html.

Figure 2.1: Tenure types, ED boundaries and distribution of Synthetic address points for the Sharrow ward, Sheffield (urban)

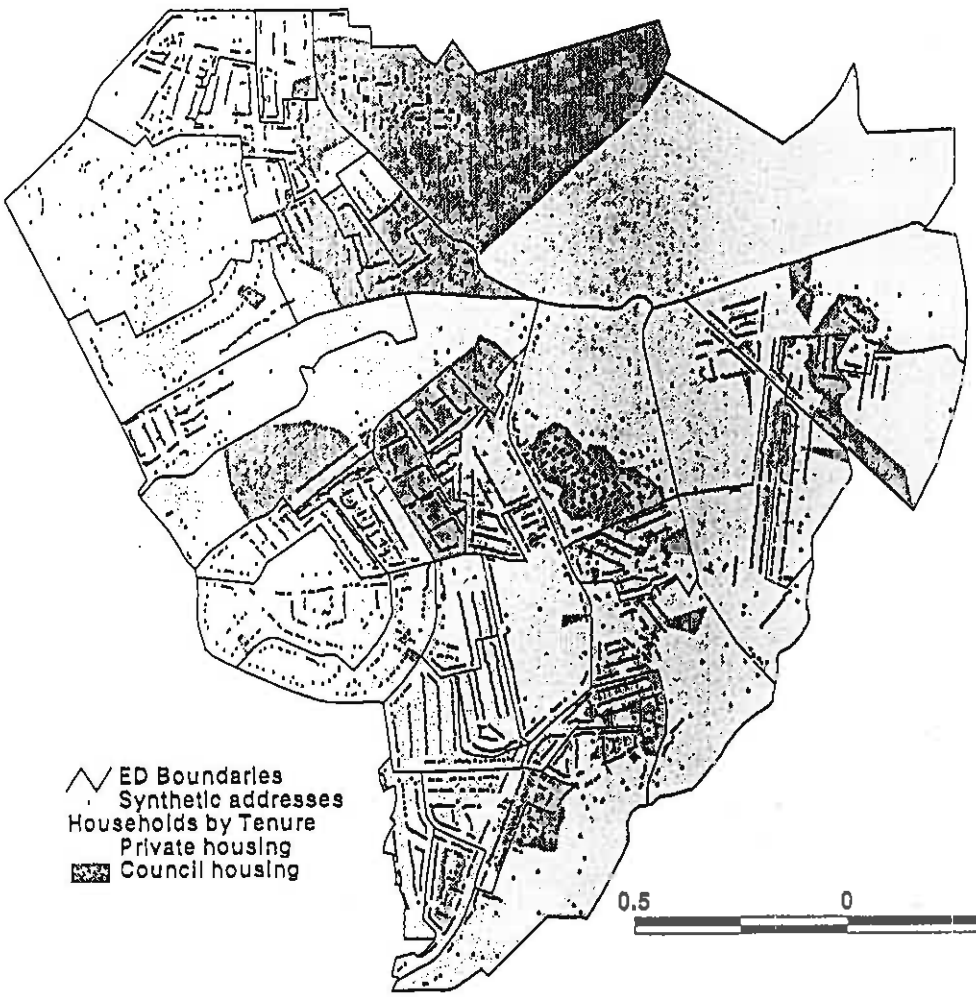
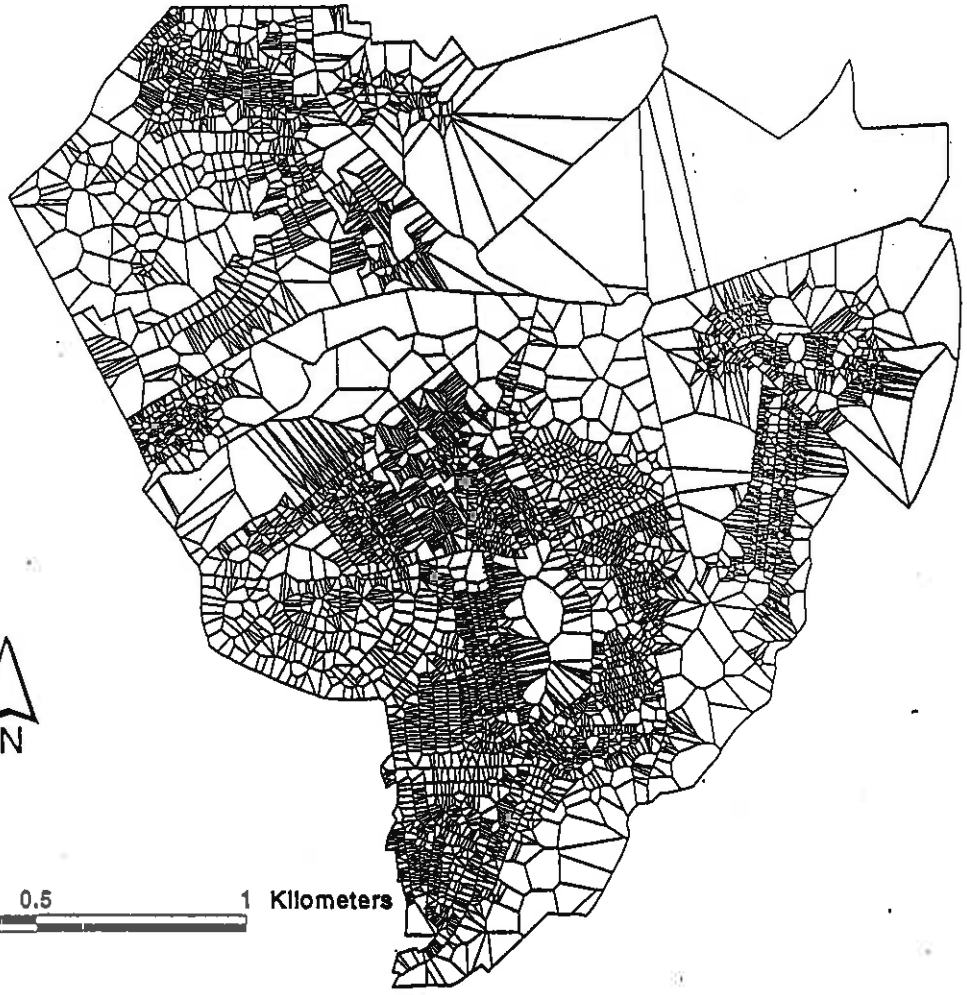


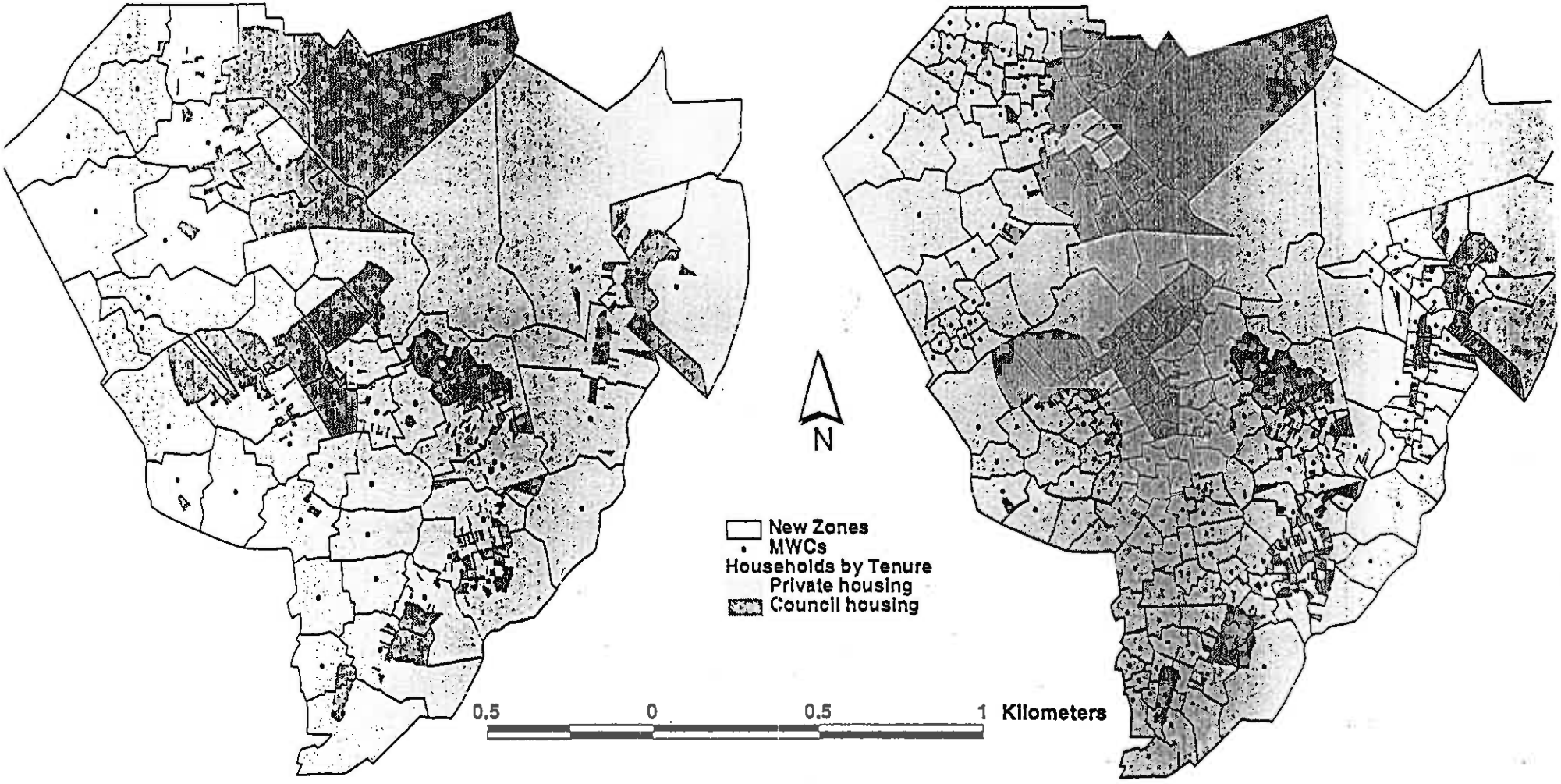
Figure 2.2: The 5406 thiessen polygons resulting from the Synthetic address points, using ED boundaries as barriers



Source: 1991 Census, Crown Copyright ESRC/JISC purchase
ED-line boundary data ESRC/JISC purchase

'igure 2.3: Zone Design result: 50 new zones (with barriers and associated MWCs

Figure 2.4 **Zone Design result: 200 new zones (with barriers) and Associated MWCs**



Source: 1991 Census, Crown Copyright ESRC/JISC purchase
ED-line boundary data ESRC/JISC purchase

Figure 2.5: Tenure types, ED boundaries and distribution of Synthetic address points for the Sharrow ward, Sheffield (urban)

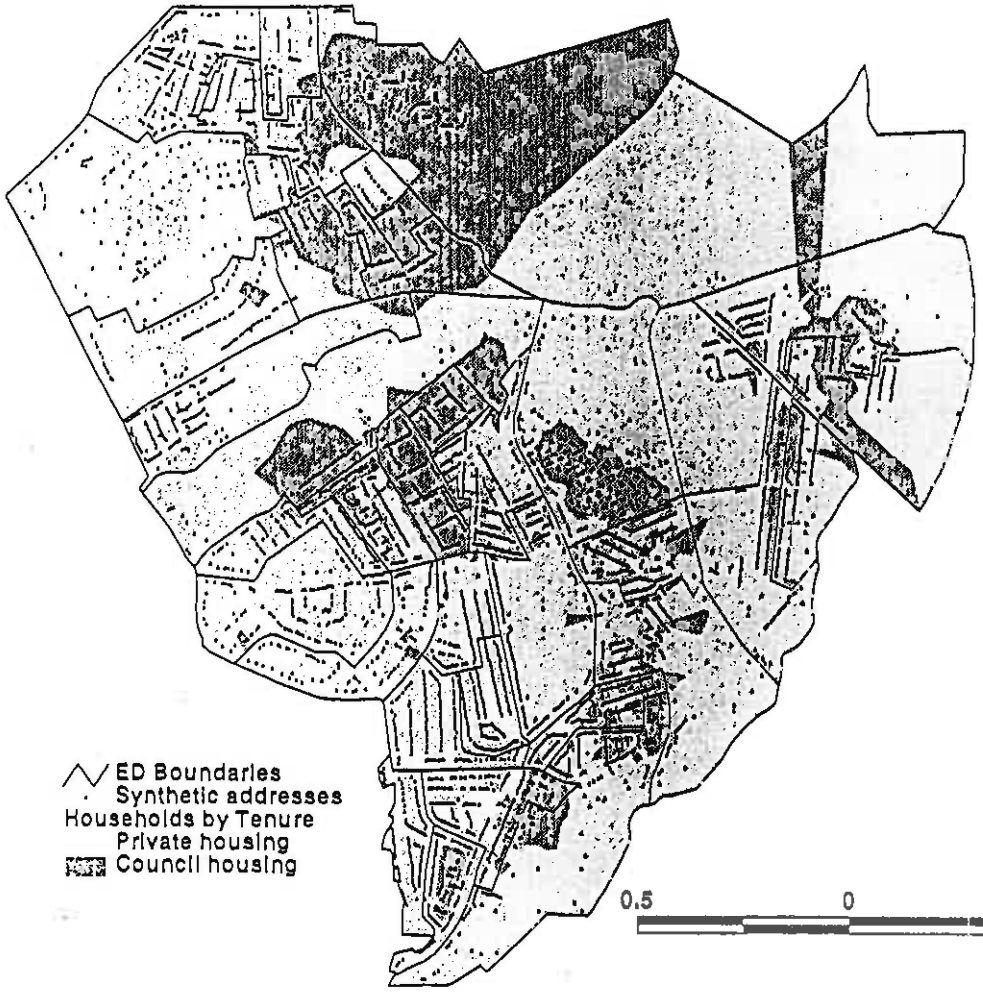
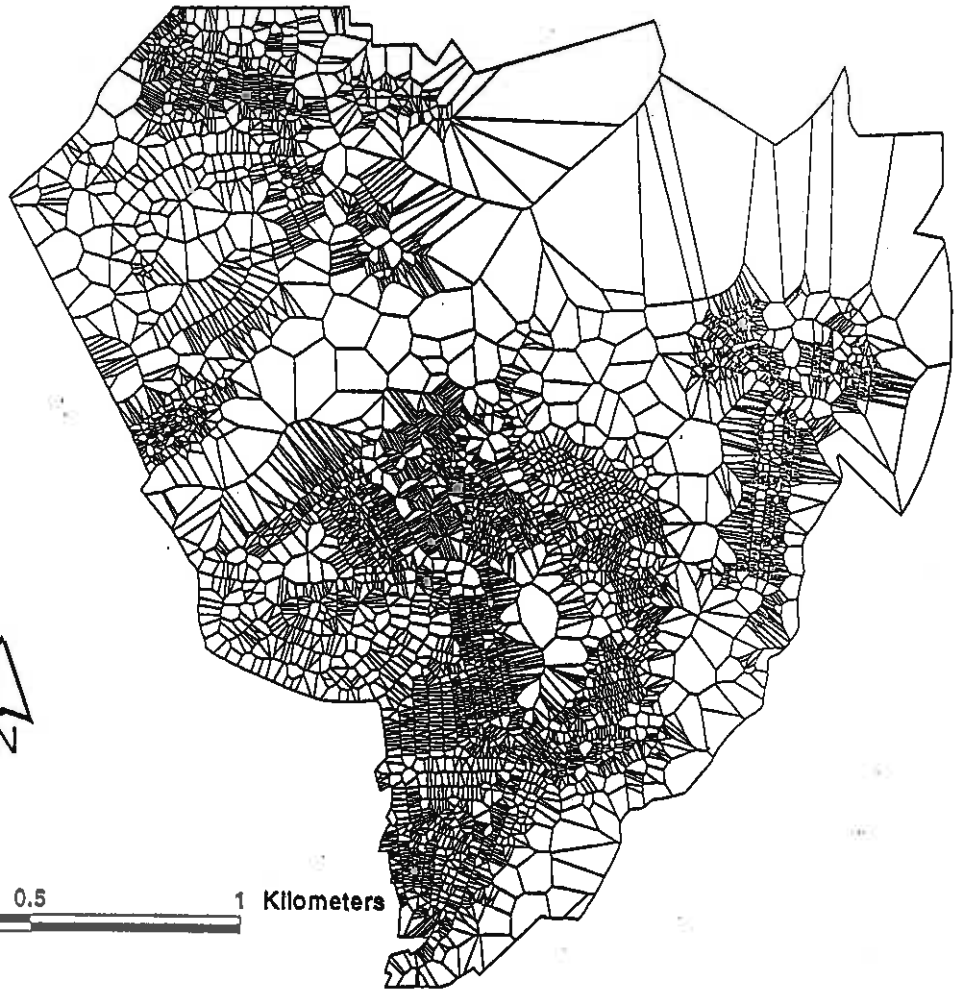


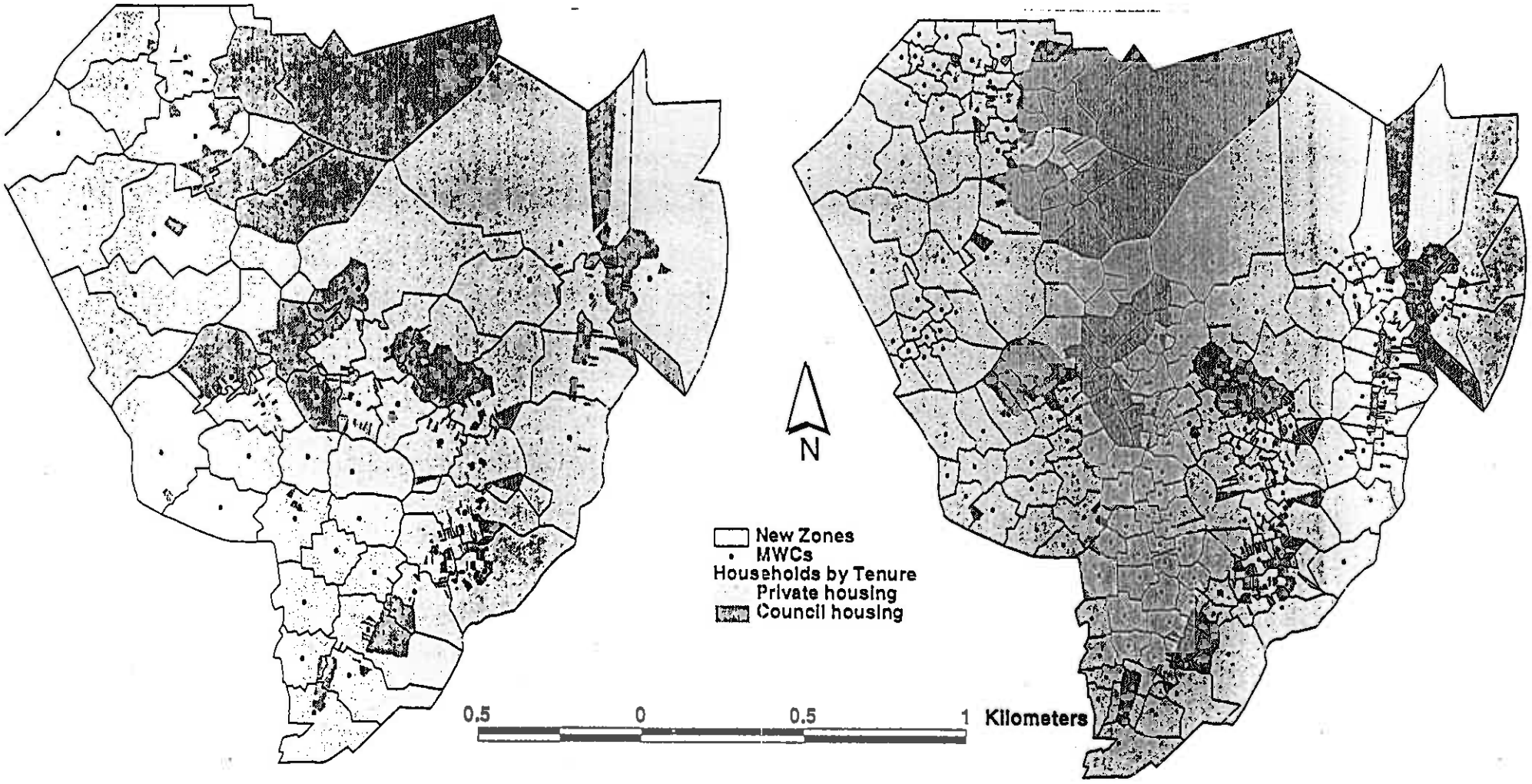
Figure 2.6: The 5406 thiessen polygons resulting from the Synthetic Address points without barriers



Source: 1991 Census, Crown Copyright ESRC/JISC purchase
ED-line boundary data ESRC/JISC purchase

Figure 2.7: Zone Design result: 50 new zones (no barriers) and associated MWCs

Figure 2.8: Zone Design result: 200 new zones (no barriers) and associated MWCs



Source 1991 Census, Crown Copyright ESRC/JISC purchase
ED-line boundary data ESRC/JISC purchase

Figure 2.9: Zone Dessign result: 250 new zones (no barriers)

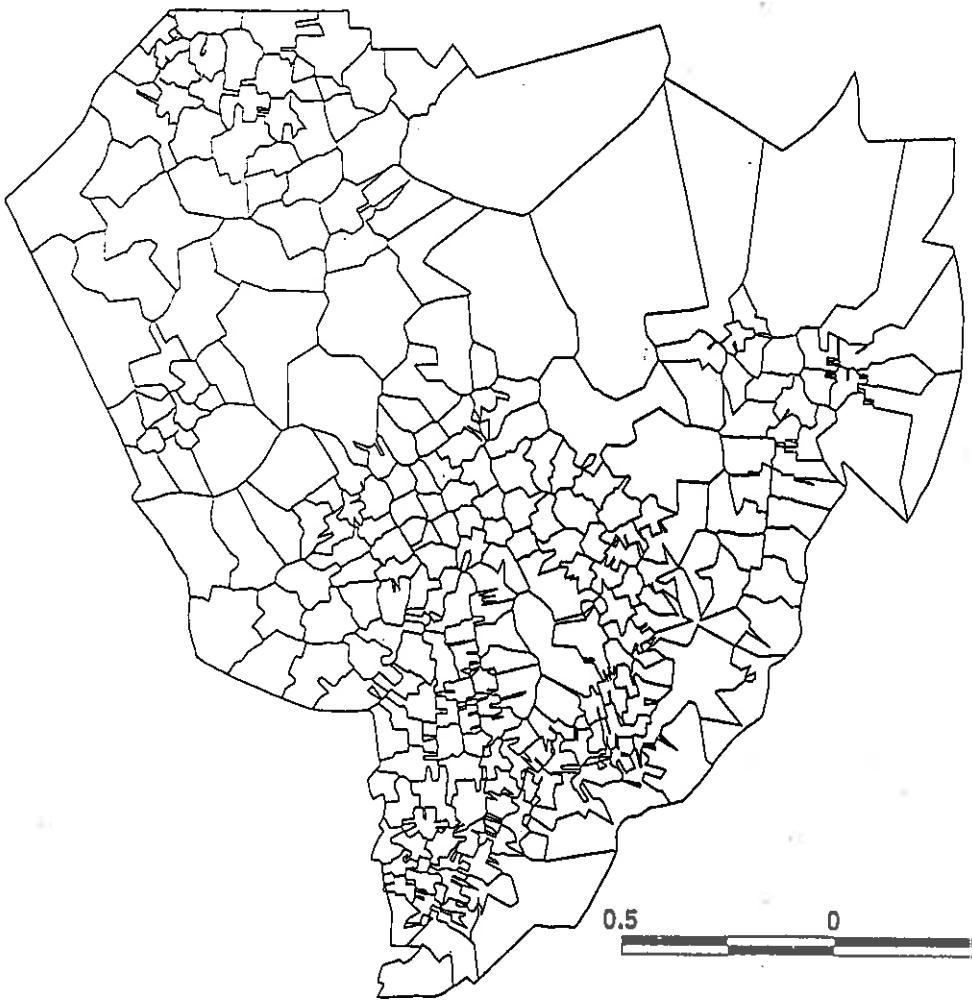
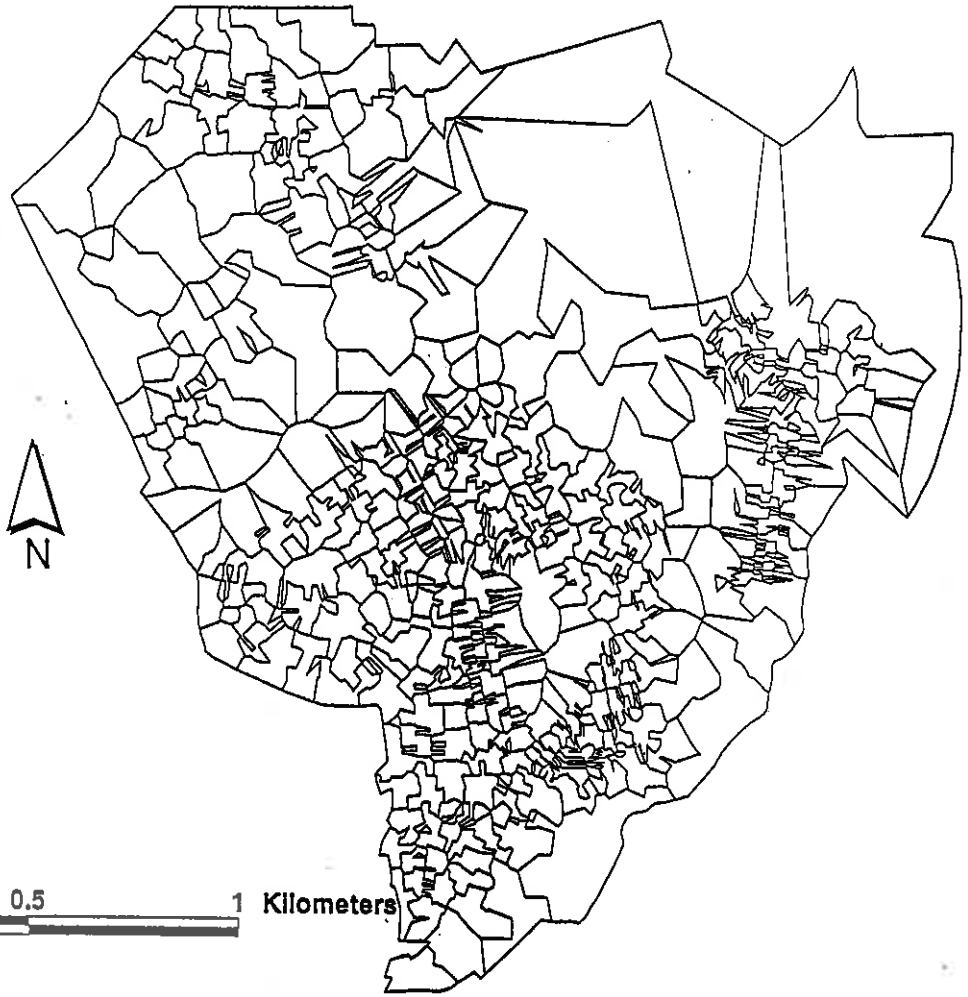


Figure 2.10: Zone Design result: 290 new zones (no barriers)



Source 1991 Census, Crown Copyright ESRC/JISC purchase
ED-line boundary data ESRC/JISC purchase

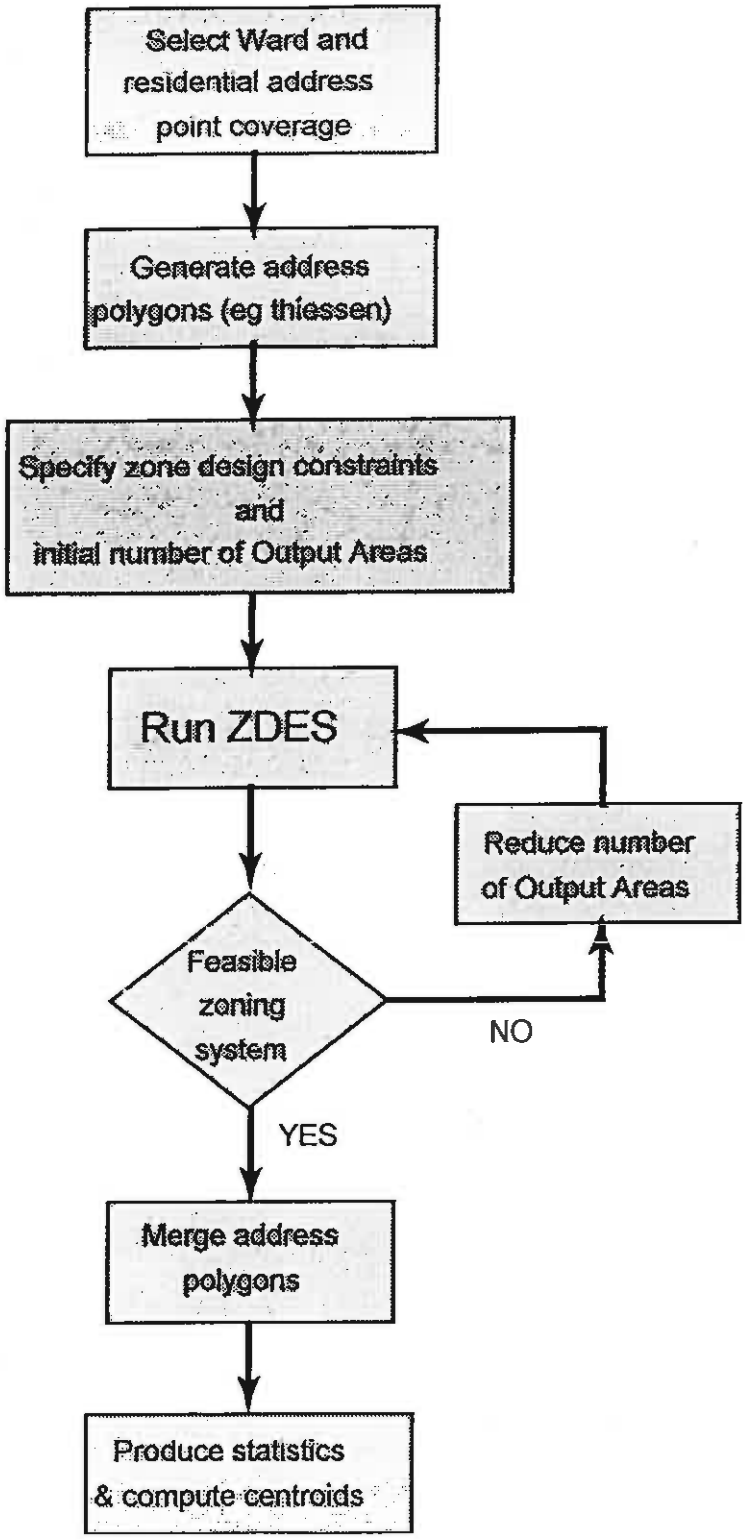


Figure 2.11 Overview of an automated census geography design process

CHAPTER 3
LOOK UP TABLES FOR THE 2001 CENSUS: RECOMMENDATIONS

Bob Barr

In this Chapter Bob Barr puts forward recommendations about geographic look up tables that should be acquired from the Census Offices and Mapping Agencies by ESRC/JISC. Look up tables link one set of entities in a system with another. So for example, there is a look up table that links addresses to unit postcodes or to output areas as defined in Chapter 1. Look up tables always have their converse in membership lists, lists of one type of entities belonging to another. So for each unit postcode there is a list of addresses that constitute its membership and for each output area there is a list of unit postcodes that make it up. The recommendations address some general organisational issues, the issue of standards and quality control, the metadata issue, the need for a master database containing all census look up tables and spin-off tables needed.

3.1 General considerations

Recommendation G1. *The role of the academic sector and the access to look up tables for the academic community needs to be clarified.*

Recommendation G2. *The academic sector should licence all major look up tables widely used outside. In addition to the CPD, PAF and Address Point are key data sets.*

Recommendation G3. *Look up tables should have a higher priority than boundary data both for initial acquisition and maintenance and support.*

3.2 Standards and quality

Recommendation G4. *The consistency of look up tables with boundary data should always be checked.*

Recommendation G5. *The academic sector should onsider providing look up table support outside the sector, possibly in partnership, as part of the National Geographic Data Forum (NGDF).*

Recommendation G6. *A standard for look up tables and associated data should be established as part of the NGDF.*

Recommendation G7. *A recommendation for the appropriate authority to maintain the Standard Geographic Base (SGB) look up tables is needed.*

3.3 The importance of metadata

Recommendation G8. *Metadata, lineage information and temporal change data for look up tables shoulld be published.*

Recommendation G9. *Transparent fully documented and independently conducted quality assurance procedures are required.*

Recommendation G10. *An error/anomaly reporting service should exist ensuring that errors for all census related tables are centrally recorded and disseminated from a single point of contact.*

3.4 Integrated master look up database

Recommendation G11. *An integrated geographical identifier database should be establisjed and placed in the public domain following the Australian, Canadian and US models.*

Recommendation G12. *The look up table database and master name table should be integrated into CASWEB (Census Area Statistics Web interface - see Chapter 12).*

Recommendation G13. *Where ESRC or JISC money is used to develop new look up tables, in particular historic links, funding should include the cost of work required to integrate the new tables into the system used for current tables.*

3.5 Associated tables

Recommendation G14. *Consideration should be given to the production of analytically defined tables such as contiguity tables. Such tables can be maintained in "virtual form" and produced on the fly if necessary.*

PART 2: MICRODATA - SAMPLES OF ANONYMISED RECORDS AND LONGITUDINAL STUDY

CHAPTER 5
SARS FROM THE 2001 CENSUS: RECOMMENDATIONS

Angela Dale

As a result of an ESRC Development Grant two papers have been produced which set out in detail:

(1) the results of a detailed survey of user and non-user requirements and
(2) initial work to establish the risk associated with implementing these requirements.

Based on this work and following discussions with ONS and the SARs 2001 Output Working Group the following recommendations are being made, subject to further disclosure work and discussions with users.

Recommendation S1. *The sample size of the Individual SAR should be increased and the population threshold for geography should be decreased.*

Users would like to be able to maximise the number of 2001 local authority areas that remain separately identifiable. This would significantly increase the policy value of the Individual SAR by avoiding the current difficulty where two or more local districts combined into a single SAR area. In 1991 there was a threshold of 120K population. Of particular importance, the move to Unitary Authorities will require a change in the geographical base of the 2001 SARs by comparison with 1991. Particular attention will be paid to the difficulties incurred in Scotland.

It is proposed that the threshold of 120K population can be safely reduced. If it was lowered to 90K then we estimate that 67% of LAs using 2001 boundaries would meet the criterion; at 60K this would go up to 90%. At the moment further work is being done to establish where a safe threshold can be drawn.

It is also proposed that there should be an increase in sample size from 2% to 3%. This would increase the precision of estimates at smaller areas.

Recommendation S3. *Additional information about head of household should be given in the Individual SAR and summary household measures should be provided in Household and Individual SARs.*

The addition of extra variables relating to the head of household has been suggested by users. However they greatly increase the risk of identification if they are assumed to be available for matching to an outside file. However, it is proposed that the most important variable relating to head of household is ethnic group. If this were to be added to the SARs then, because of homogeneity within the household, the additional risk would be minimised. Additional work is currently being undertaken to assess this possibility.

Summary household measures, especially descriptions of household composition, are particularly valuable and pose little threat to confidentiality. We therefore propose that a household classifications is added to the 2001 SARs. This should be:

i. the MRS lifestage variable; or
ii. the DETR (DOE) household classification; or
iii. an academic classification derived to reflect broad categories of household by stage of lifecycle and composition.

Recommendation S3. *A case for a Third SAR at a much smaller level of geography should be investigated.*

There have been a number of requests for a third SAR at a much smaller level of geography. There are arguments for basing it at 5K population size - equivalent to about 42% of all wards and a much larger percentage of the population . The benefits of a SAR at this low level of geography are:

i. academic analysis allowing effect of area to be included (eg multilevel modelling)
ii. providing building blocks to aggregate to different boundaries – ie health authorities, old LAs and new LAs
iii. analysis within LA or HAs.

With a substantial reduction in individual level detail, a case can be made for the third SAR with geography down to large ward size.

It will, however, be important to clarify the analytic value of this SAR in the light of an on-line tabulation service with 100% data.

Recommendation S4. *A Household SAR should be requested from the 2001 Census of the same detail as in 1991, except that the regional definition to be used should be that of the new Government Office Regions.*

There is little scope for increasing detail on the Household SAR An increase in sample size could only be achieved with loss of the existing regional geography. This is not being recommended.

However, it is proposed that the 2001 SARs should reflect the Government Office Regions in recognition of the policy requirement for identifying these areas. Although this would mean some loss of continuity from 1991 when standard regions were used, it is more important to get meaningful geography for 2001.

Recommendation S5. *Attention should be given to respecifying some individual variables (such as migration origin) and to the specification of any new variables introduced in 2001.*

Once the overall parameters of the 2001 SARs have been agreed it will be necessary to consider changes to individual variables. These include more detail on migration variables. In addition it is assumed that any new questions asked in the 2001 Census would be available on the SARs.

CHAPTER 5
THE ONS LONGITUDINAL STUDY LINK TO THE ENGLAND AND WALES CENSUS: RECOMMENDATIONS

Brian Dodgeon and Heather Joshi

The authors summarised the recommendations put forward at the Third Workshop for the construction of the Link of the Longitudinal Study to the 2001 Census. At the time of the Workshop the conclusion of the ONS Review of the LS was not known. However, all researchers consulted had stressed that the LS was the jewel in the Crown of census data sets, the envy of researchers across the world. It is therefore with some confidence that the authors put forward their recommendations.

5.1 Longitudinal Study (LS) Link to the 2001 Census

Recommendation L1. *The LS Link to the 2001 Census should be funded - each succeeding Census adds exponentially to the value of the LS.*

Recommendation L2. *To this effect, efforts should be made to maintain awareness in Whitehall and Parliament of the importance of the LS link to the 2001 Census.*

5.2 Implications of the one number census

Recommendation L3. *As the imputation of missing Census data will occur earlier than hitherto, the LS link should be to raw data, before imputation and editing are done, or else all imputed data should be clearly flagged.*

Recommendation L4. *A census volume on imputation should be produced to help users understand the methodology by which the process of imputation was carried out.*

5.3 New 2001 Census variables

Recommendation L5. *A question should be asked abut all intra-household relationships. Currently we cannot be sure of the relationship of another household member to the LS member if neither is the head of household.*

Recommendation L6. *An income question should be asked. LS Users favour having enough detail in the banding to be able to approximate both those in higher tax brackets and those eligible for means tested benefits. Net income may be more useful than Gross, for comparability with other EU countries.*

Recommendation L7. *The educational attainment question should capture the complete educational history, which may be achieved through multi-ticking of the proposed question.*

Recommendation L8. *Better information on handicap and disability is desirable, as the Limiting Long Term Illness question lacks explanatory detail. Codifying different medical conditions could be problematic, but an alternative might be to codify "quality of life" indicators such as the ability to walk, cook, do shopping, etc.*

Recommendation L9. *On ethnicity, more information would be helpful on those with mixed parentage, and the separation of those of Irish descent would also be desirable.*

Recommendation L10. *A smoking question would be enormously useful for LS research into health outcomes.*

Recommendation L11. *The employment status question should be expanded to clarify the duration of unemployment for those seeking work.*

Recommendation L12. *There should be a question to ascertain if English is the respondent's first language.*

5.4 Issues of continuity and consistency

Recommendation L13. *It is absolutely essential that the Occupation and Industry questions should enable coding in full detail, and not in the proposed reduced form of 13 and 29 tick-box categories respectively. The implications of this proposal would be disastrous, as the ability to do meaningful research into occupational mortality would be lost: this was one of the primary reasons for the creation of the LS. Recognising that coding of open questions is expensive even with advances in technology, we recommend that the utility of census data for all users would be better served by continuing the practice of 10% processing of the "hard-to-code" between questions, additional coding of all LS members, rather than opting for 100% coding of much less detailed information.*

Recommendation L14. *Occupation should be double-coded with the 1991 classification as well as the 2001 coding. This should also apply to Industry. It should also be possible to compare RG Social Class and the new ONS Social Classification user two occupational classifications.*

Recommendation L15. *To the LS should be added derived variables summarising sequences of employment status, housing tenure, region of residence, household/family position, occupation and marital status.*

Recommendation L16. *A bank of complex derived variables from previous projects should be added to the LS.*

Recommendation L17. *Look up tables should be prepared that link the smallest output areas in the 1971, 1981, 1991 and 2001 Censuses.*

5.5 Documentation

Recommendation L18. *The LS Technical Volume should be updated.*

Recommendation L19. *The LS Data Dictionary should be updated.*

Recommendation L20. *Guides should be produced on changes in classifications between Censuses (particularly in occupations and social class), on the comparability of derived variables across Censuses, and on the complex derived variables added to the LS.*

5.6 User interfaces

Recommendation L21. *Interactive tutorial material introducing the LS to potential users should be placed on the World Wide Web.*

Recommendation L22. *Possibilities for micro-analysis in a safe environment to be kept under review, for exampl to allow a much greater range of statistical packages.*

Recommendation L23. *To avoid the introduction of any more user charges to individual academic projects.*

PART 3: INTERACTION STATISTICS FROM THE 2001 UK CENSUS

CHAPTER 6
MIGRATION STATISTICS FROM THE 2001 UK CENSUS: WHAT DO WE REALLY, REALLY WANT

Paul Boyle and Phil Rees

This paper updates a similar paper presented in a previous meeting (Rees 1997) and we focus primarily on the additions to that paper. We emphasise that: careful attention needs to be given to the accuracy of these data; not enough use was made of the migration data from the 1991 census; some simple changes to the migration data handling and output will improve these data; user-friendly software must be provided for accessing these data; separate support is required to encourage and facilitate academic research using interaction data.

6.1 Introduction

The importance of studying intra-national migration has long been recognised in a variety of academic disciplines, including geography, economics and sociology, especially since migration is a more important element of population redistribution than natural increase in many developed countries including the UK. There is a growing need for accurate migration estimates for various reasons including: population forecasting (the potential distribution of the 4.4 million new households is an important topic currently); the role of migration in the changing geography of production and the new spatial division of labour; the role of migration as the balancing mechanism in regional labour markets; and the link between migration and changing household circumstances. Consequently, there is a strong demand for detailed and accurate migration estimates which can only be provided from the census.

Of course, there are problems with the migration information provided in censuses which are difficult, if not impossible to solve. Migration is only measured in one year out of every ten and emigrants are ignored. Multiple migrations within the year prior to the census are also missed as are migrants who die in the year. We also know nothing about the characteristics of the migrants at the start of the year, which is crucial if we wish to examine the way in which people's circumstances change with migration.

Even so, the migration information from the census provide an excellent opportunity to study some of the most important migration questions as it is comprehensive and allows migrant characteristics to be identified. In fact, much of the migration data available to academics from the 1991 census was ignored and there are a number of reasons for this. The interaction data are probably the most complicated census data output and in 1991, and especially 1981, the extraction of these flow data was extremely complicated. No group was funded to provide specialist academic advice and support.

6.2 Measuring Migration

Some improvements to the way migration is measured in a census could be made. Children under one are not assigned a migrant status but they have a starting address in the year (place of usual residence at time of birth) and a usual address at the time of the census. Counts of such infant migrants are an essential input to subnational population estimation and projection.

Recommendation M1. *Consideration should be given again to extending the migration question to capture infant migrants, as recommended by the Migration Question Sub-Group. If a question*

revision is ruled out, migration tables should report the migration status of the mother (or other parent/guardian if no mother is resident in the household) for children under one.

There is also little information captured about the status of the migrant at the start of the year at the origin location, apart from those characteristics which can be easily inferred (e.g. age one year ago) or which do not change (e.g. country of birth). For example, the important relationship between employment status and migration cannot properly be measured because only status after the migration is known. Other important characteristics one year ago might include marital status and tenure.

Recommendation M2. *Consideration should be given again to expanding the relevant questions to capture economic position one year ago, as recommended by the Migration Question Sub-Group.*

Increasingly studies are being attempted that integrate census data from different nations. However, these kinds of comparison are difficult because censuses are carried out in different years and the migration questions vary. Migration is usually measured by comparing the current address with the address one year ago. However, in the US a five year interval has been used. And in some countries the duration of residence is also included.

Recommendation 3. *Consideration should be given to the definition and implementation of an internationally agreed definition of migration.*

It has been estimated that in the 1991 Census of Great Britain, 0.9% more persons migrated in the year before the Census than were recorded therein. An additional problem occurs when tables of migration flows are created. Many residents report that they had a different address one year ago but fail to provide details. The origin not stated category comprised 6% of migrants in the 1991 Great Britain Census. In most migration tables the number of migrants with an origin not stated are reported so that users can make their own adjustments. However, under the plans for a one number census (Jones 1997; ONS, GROS and NISRA 1997, pp.1-12) consideration should be given to imputing the missing origin information.

Recommendation M4. *Consideration should be given to the problem of undercounting specific to the migration statistics, because respondents systematically failed to report a different address one year before the census and because some respondents who did report a different address failed to report origin details.*

Recommendation M5. *Imputed migrants should be flagged.*

Recommendation M6. *Any one number census methodology must account for migration (approximately 2% of the population?) during the six weeks between the census date and the post census survey.*

Recommendation M7. *Any one number census methodology must give serious attention to the identification of missed migrant origins.*

The analysis of migration by household structure is relatively undeveloped. The UK Censuses report the migration of wholly moving households and of households where the head has migrated, together with the residents of such households. However, there are other types of migration: individual migrants moving into households with non-migrant members and migrants moving between, into or out of communal establishments. There is also the important migration flow of students from parental homes to university/college residences and from university/college residences to first career residences on graduation. This issue has been exposed by the Migration Question Sub-Group and the decision to

record students as resident at their term time address will make possible the recording of these important, hitherto undercounted flows.

Recommendation M8. *A careful study should be undertaken of ways it will be possible to classify migrants by household status from the 2001 Census, with a view to recommending the kinds of new tables that might be produced. This study should involve collaboration between the Census Offices and customer sectors.*

Recommendation M9. *Serious attention should be given to the potential use of household units in the census (rather than wholly moving households).*

Recommendation M10. *A careful study should be undertaken of ways it will be possible to classify migrants by student status from the 2001 Census, with a view to recommending the kinds of new tables that might be produced. This study should involve collaboration between the Census Offices and customer sectors.*

6.3 National Migration Statistics

The 1991 Census National Migration Statistics (OPCS and GROS 1994a, 1994b and 1994c) are organised in two parts and the tables cover Great Britain only. A separate volume on Migration was published on Northern Ireland (CONI 1994). Harmonisation and integration of these into a volume covering the United Kingdom is needed.

Recommendation M11. *The National Migration Statistics should be consolidated into a United Kingdom product by harmonising and integrating the Great Britain and Northern Ireland tables.*

A large number of immigrants arrive in Great Britain from the Irish Republic, making this an important migration flow.

Recommendation M12. *Careful attention should be given to the flows to and from the Irish Republic.*

From the 32 by 32 matrices in National Migration Statistics 1 and 2 researchers can extract interregional matrices and inter-country matrices, but not matrices for inter-county flows or flows between any other comparable classification at the same scale (e.g. NUTS 2 region). The Special Migration Statistics had to be used to generate those flow tables. There is a need to produce a wider set of migration flow tables as part of the National Migration Statistics.

Recommendation M13. *Plan to produce a much wider set of computer readable flow tables in the National Migration Statistics using a suite of geographies, ranging from standard regions, through counties (or equivalent) to local government units (unitary authorities, districts).*

The nature of the National Migration Statistics needs therefore to be re-thought. The published volume in 2001 should provide definitions, explanations and summary tables but the detailed tables can be published as a library of computer file form on CD-ROM. The number, format and contents of the files need thorough discussion between the census offices and census users.

Recommendation M14. *Plan the National Migration Statistics publication as a single volume containing selected summaries, definitions, explanations and analysis and a catalogue of the suite of more detailed tables placed on an accompanying CD ROM.*

6.4 Regional Migration Statistics

The Regional Migration Statistics from the 1991 Census were originally to be published as a set of volumes, one per region (as from the 1971 and 1981 Censuses) but the decision was made to scrap paper publication because of cost and they were published as computer readable tables. The academic community copies are stored on the MIDAS system of Manchester Computing as a set of files.

Recommendation M15. *Plan the Regional Migration Statistics publication as a single volume containing selected summaries, definitions, explanations and analysis and a catalogue of the suite of more detailed tables placed on an accompanying CD ROM. Alternatively, the National and Regional Migration Statistics could be planned as an integrated volume.*

6.5 Local Base Statistics and Small Area Statistics

There is relatively little migration information incorporated in the Local Base Statistics (LBS) or Small Area Statistics (SAS). There is considerable scope for extending the crossclassifications of migrants to cover all the dimensions tackled in the National and Regional Migration Statistics, perhaps with broader coding of the variables. The type of move (TYMO) classification is developed more elaborately at smaller scales. In the Local Base Statistics (Table L15) the following TYMO classification is used

a. Moved within wards
b. Between wards but within district
c. Between districts but within county
d. Between counties but within region
e. Between regions or from Scotland
f. From outside GB
g. Between neighbouring districts
h. Between neighbouring counties/Scottish regions.

The classification is perhaps over elaborate: flows g and h are not needed to complete the picture of in-migration. More importantly "Migrants from the area of residence to the rest of GB" are omitted. There are also no statistics on migrants with origin not stated. These have been merged into the "Moved within wards" category. It is therefore not possible to measure the balance of inflowing and outflowing internal migrants properly. The tables cannot easily be used in population change analysis or population estimation. Out-migrant flows are missing because of the way the LBS and SAS were processed area by area. It is impossible to count out-migrants from an area until the whole national census has been processed. If the "one number census" approach is adopted for the 2001 Census, out-migrants can be counted for all areas, including those of the smallest scale.

Recommendation M16. *Improve the provision of Migration Area Statistics by tabulating out-migrant flows, by adopting a simplified type of move classification and by expanding considerably the crossclassifications of migrants. These could be provided as additional tables in the general area statistics produced from the census or as part of the Special Migration Statistics.*

6.6 Special Migration Statistics

The *Set 1* array has a very large origin-destination face or matrix. The origins are composed of four types: (1) 9930 wards in England and Wales, (2) 1003 pseudo postcode sectors in Scotland, (3) 96

areas outside Great Britain (individual countries or groupings of countries or other areas such as Northern Ireland), and (4) an origin-not-stated category. The destinations are made up of the first two types of area. The third dimension is made of broad age-sex groups (Table SMS M01) and the categories of wholly moving households and residents in such households (Table SMS M02).

The *Set 2* array has a much smaller origin-destination face or matrix. The origins consist of three types of area: (1) 459 local government districts in Great Britain, (2) 96 areas outside Great Britain (as in Set 1) and (3) an origin-not-stated category. The attribute dimension can be divided into two parts. The first part consists of the broad age-sex groups and wholly moving household categories of Set 1 plus a more detailed age-sex classification (19 age groups by sex). These data are published without modification and are extremely valuable for demographic analysis, population change analysis, population estimation and forecasting.

The second set of attributes by which the Set 2 district migration flows are classified consist of a further 38 to 40 counts organised in 8 or 9 tables (the exact statistics available differ by country). These data were regarded by the Census Offices, despite the very broad coding used, to be a threat to the confidentiality of individual census microdata. As a result, these data were suppressed when the total number of migrants between an origin and destination was fewer than ten. This meant that, although a majority of migrants were reported in these tables, data for only a minority of flows were available. No sensible analysis or aggregation of these flow data could be carried out. Any analysis had to be confined to examining the total flows into and out of districts by these socio-economic characteristics (Champion 1996). The difficulties posed by this wholesale suppression in the second part of the SMS Set 2 challenged Rees and Duke-Williams (1995b, 1997) to "reverse engineer" the suppression and to reconstruct the flow array virtually in its entirety.

So what should be done about the confidentiality problem if equivalent data sets are produced from the 2001 Census. The alternative protection devices which have been suggested include (1) a small amount of record swapping in master database, as part of the one number census approach, (2) sampling, as has been used with other census interaction data set, the Special Workplace Statistics, (3) rounding to a small number base and (4) random perturbation. The first two devices are applied before table production while the third and fourth devices are applied after table production. Because origin-destination matrices for wards/sectors and districts contain a majority of very small flows, methods (2), (3) and (4) will all introduce error. Record swapping is being canvassed widely because although it introduces a small amount of error, the errors should be self-cancelling when aggregation is carried out. However, the research to establish the viability of this method has still to be done. Sampling has the advantage of being used successfully in the past and of making possible tables with more detailed socio-economic coding if the precedent of SWS coding is followed (see Flowerdew and Green 1993 for details).

Recommendation M17. *If it is still considered essential to apply additional protection measures to some migration tables at the smaller geographic scales, sampling should be used in preference to suppression, rounding or random perturbation.*

Consideration should be given to the provision of a library of migration flow matrices for widely used sets of origins/destinations, the structure of which users can immediately grasp and use. A suite of flow matrices at each scale should be available. These could be linked to the flow matrices produced in the National, Regional and Local Migration Statistics.

Recommendation M18. *Once the structure of the National, Regional, Local and Special Migration Statistics is agreed, a user friendly interface to all data sets should be developed building on experience with the 1991 Census migration data.*

Recommendation M19. ***The interface should allow user-defined geographies to be easily specified.***

Recommendation M20. *The interface should also provide some simple interaction measures.*

In previous discussion between the Census Offices and census users, there has been a strong demand for a flexible tabulation system which users can employ to produce their own designed tables. The ingredients needed for a flexible tabulations system are:

(1) a method for assessing the confidentiality risk of a particular table request;
(2) a system which users can use to design their requests before submission (e.g. the software linked to a dummy database resembling the census database) and
(3) administrative arrangements between ESRC/JISC and the Census Offices for funding table requests in a cost-effective way.

Recommendation M21. *Consideration should be given to the development a flexible tabulation service provided by the Census Offices or a designated agency to provide any additional migration tables needed by users.*

Recommendation M22. *To reduce the cost and speed the delivery of a flexible tabulation service, both the tabulation/analysis software and a dummy census dataset should be released to the academic community.*

Recommendation M23. *The SMS and SWS should be produced in a compatible format as possible improving the user-friendliness of the data sets.*

6.7 Sample Of Anonymised Records

Much of the recent work on migration has taken advantage of the data provided in the Sample of Anonymised Records (SAR). These individual level data were especially useful for allowing migrant characteristics to be determined. The data were frustrating in some ways though.

Recommendation M24. *More geography is needed for the migration origin (currently region). The ability to construct inter-SAR-area matrices would have been useful.*

Recommendation M25. *An area-type measure would be also be useful at both the origin and the destination (calculated from data at the ward-level).*

6.8 Centre For Interaction Data (CID)

The complexity of the 1991 census interaction data was one reason why relatively little use was made of these statistics. This could be improved considerably by providing specialist support.

Recommendation M26. *An ESRC-funded specialist support centre dealing with interaction data from the census (migration and commuting).*

It is also important, if the decision to alter the provision of the SMS is made, for potential users to be able to influence the output that is available as standard. Requirements for the data extraction software also need to be given attention.

Recommendation M27. *A survey of user-needs for interaction data would be carried out prior to the census, to aid the production of a user-friendly interface for users.*

There is no reason why such software could not be made available to Local Authorities and discussions with ONS are required to consider using such a centre in a data support role for the Local Authorities.

Recommendation M28. The possibility of providing some support to local authorities for the provision of census interaction data would be investigated.

To improve the use of these data (perhaps even in teaching environments, rather than research alone), careful attention must be given to the software used to extract the data.

Recommendation M29. CID would produce software to provide flexible access and analysis of the standard interaction data supplied for wards and districts.

Recommendation M30. The possibility of providing this software as a web-based service would be investigated.

The centre would also provide continuous support for users. This would encompass specific problem-solving, but also the general provision of training courses.

Recommendation M31. CID would offer on-line academic support for interaction data extraction and analysis; short courses for introducing the software; an occasional newsletter to explain the tables that have been produced.

If the provision of interaction data is altered from the 1991 situation, with users having much more flexibility in being able to determine their own tables, help will be required to deal with confidentiality problems. This may involve some complex calculations and the centre would offer advice on this problem.

Recommendation M32. CID would also act in a similar way to the Longitudinal Study group, providing help with confidentiality in table definitions. CID would co-ordinate user discussions about the most helpful tables that could be produced.

A critical component of interaction data modelling is the calculation of inter- and intra-area distances. These can be calculated between population-weighted centroids for areas, but these are inaccurate if the straight-line distance crosses estuaries or large rivers. Methods for dealing with these types of problem are already being considered at Leeds and a set of alternative distance measures will be provided at a range of spatial scales. Most obviously, different distances may be chosen for the SMS and SWS data.

Recommendation M33. CID would also provide support calculating alternative distances between pairs of units (wards, districts etc.) that are vital in interaction modelling exercises.

The longer term goal of the centre would be to collect migration information from a range of sources, providing advice on use and analysis.

Recommendation M34. In the longer term, CID would become a specialist centre providing support for other forms of interaction data from national and international sources.

6.9 Conclusion

Compared to other forms of census data the interaction matrices are difficult to use and analyse. The recommendations here are designed firstly to provide more accurate and useful information from the 2001 data. Secondly, the output of flow statistics is difficult because of the sheer size of the matrices

involved. Here we suggest that computer-readable output is far more useful than very large volumes. Finally, a key goal is to encourage more academic use of these data and this may only be achieved by the provision of expert support. Overall, we would hope to see a more flexible interaction data service for 2001.

Reference

Rees P. (1997) Migration statistics from the 2001 Census: what do we want? In P. Rees (ed.) Third Workshop on the 2001 Census - Special Datasets: What Do We Want? Working Paper 97/9, School of Geography, University of Leeds, Leeds LS2 9JT, UK.

CHAPTER 7
WORKPLACE STATISTICS FROM THE 2001 UK CENSUS: RECOMMENDATIONS

Martin Frost

Much of what follows is based on the discussion which took place as part of the Third Census Workshop and was reported in the subsequent Working Paper (Rees, 1997). At that meeting a range of Census Users and representatives from ONS had the opportunity to propose a set of recommendations about the form and content of the Workplace Statistics for 2001.

Recommendation W1. The SWS Sets A and B should be included in the Area Statistics.

In essence the reasoning here was that following the spirit of a 'one number' approach to universal adjustments of the Census database and the proposal that all tables should be processed at 100% coverage, it made little sense to separate off what are essentially static data relating to the characteristics of residents and workplaces within zones from the other information on residents conventionally contained in the SAS or LBS tables. Integration of these data would ensure better standards of comparability and would simplify access. This recommendation was made subject to the reservation that there should be comparability between the counts of residents and workplaces contained in these static tables with zones totals derived from the flow data contained in Set C.

This implies that decisions about the use of wards, 'frozen' wards or postcode areas would be driven primarily by the general Area Statistics rather than by special considerations related to the Workplace Data

Recommendation W2. *The area specification chosen for the SWS Set C academic purchase should be the same as that used in the Area Statistics*

In 1991 a range of possible combinations of ward, ED and postcode areas were potentially available to users who ordered the Workplace Statistics. In the event the ESRC purchase was specified as a 'symmetric' ward to ward set in England and Wales, symmetric postcode areas in Scotland. This recommendation follows the logic of maximising comparability between the Workplace and Area data in any purchase of 2001 tables, with an assumption that a similar 'symmetric' flow table will be ordered.

Recommendation W3. *Because of the trade-off between table details and modification/suppression there should be a range of output specifications adjusted to spatial scale*

This raises the awkward unknown of what kind of suppression might be applied to an extremely sparse 100% flow table of work journeys. Given that flows of employees between pairs of wards (or similar areas) will contain large quantities of small numbers (probably below 5 movers in the majority of cases) suppression will probably become a major issue in the use of these data. This was largely avoided in previous Censuses on the grounds that the data were only a 10% sample. In previous Censuses any flow data purchased by ESRC was for small areas (wards) only, allowing the user to aggregate to zones of their own choice. This recommendation aims to provide users with a range of larger 'standard' spatial units (e.g. Districts or Counties) so that suppression rules can be applied to these zones as a whole.

Recommendation W4. Users should be involved closely in discussions on modification/suppression, and on imputation of workplaces of missing persons

This recommendation clearly develops the discussion of the previous paragraph but also adds one further important issue within the Workplace Statistics. At the time of Workshop 3 the principal imputation problem specific to the Workplace Statistics was seen as the difficulty of coping with Census forms which have no specified workplace location or an unusable workplace postcode. In the past various 'rule-based' approaches have been used to try to construct usable postcodes from those that are incomplete or inaccurate. To date it is not clear what procedures are to be followed in 2001 to cope with these problems.

It is becoming increasingly clear, however, that this is not the only imputation issue that will affect the Workplace Statistics in 2001. At the time of writing (more may be revealed on May 12/13th!) it is not clear how the suggested adjustments to the Census database to allow for under-enumeration will cope with allocating imputed workplaces. This is not dissimilar to the problem of allocating imputed residents to households but has the particular problem that workplace choice is highly sensitive to spatial location as well as the attributes of the resident in question. This is clearly a problem that requires urgent discussion.

Recommendation W5. *Detailed metadata on coding practice, modification/suppression and imputation should be provided*

This is part of an attempt to encourage wider use of these data. In the past they have been used by a relatively small group of individual researchers who have built up specialist knowledge often over a considerable number of Censuses. Part of this knowledge has been about the procedures used for coding, imputation etc. This proposal is geared towards using the techniques generally proposed to open up availability of all forms of Census information to ensure that more people can find out more easily about the characteristics of the Workplace data in the hope that this will stimulate greater use.

This leaves open the question of software which might be used to ease people's access to this potentially large data set. Although ONS have indicated a willingness to consider a variety of ways in which users might interact with the data to specify table constructions of their choice, it is not clear how access to the data will be improved in 2001. This is another topic in need of continuing discussion.

Reference

Rees, P (1997) (ed.) *The 2001 Census - Special Datasets: What do we want?* School of Geography, University of Leeds Working Paper 97/9

PART 4: INTERFACES TO CENSUS DATA

CHAPTER 8
INTERFACE(S) TO DIGITISED BOUNDARY DATA FROM THE 2001 UK CENSUS: RECOMMENDATIONS

Donald Morse and Alistair Towers

8.1 Introduction

The paper is in two parts. Part one discusses what interface(s) are required to give the most straightforward but comprehensive access to all the digitised boundary data (DBD) held in the UKBORDERS database. Currently under development are a 'generic' interface – which provides access to all the data held in the system, and the first of a set of 'express' options which will provide direct access to key subsets, such as the 2001 boundary data. Part two puts forward some recommendations on what 2001 Census DBD – content, quality, format and delivery - would best meet the needs of the academic community.

8.2 Interface Recommendations

The census is a key data set for a growing number of academics from an increasingly wide range of disciplines. The use of GIS and desktop mapping in the analysis and presentation of census and related statistics has also grown apace, particularly since the digitisation of the 1991 Census boundaries coupled with the maturation of desktop mapping software. UKBORDERS, one of the suite of services provided by EDINA[1], is a Census Unit funded under the 1991 ESRC Census Initiative to provide password-controlled access to digitised boundary (DBD) data for census and other geographies which contribute to research and teaching needs within the UK Higher Education community. UKBORDERS holds a wide variety of DBD - from postcode unit boundaries to national outlines; there is also a considerable spread over time, from the mid-nineteenth century to the present. Most, but by no means all of the coverages held relate to areas for which census data are available, if only, in the case of earlier censuses, in hard-copy published form.

Arranging straightforward information on, and access to, such a large and complex database presents a host of challenges. The most difficult is the need to accommodate continually growing content covering a range of geographies and time periods. At the same time, because UKBORDERS has a remit to serve the whole of the UK higher education community, it seeks to provide equally for the needs and abilities of a very wide ranges of users - from those with a high degree of technical sophistication and experience of working in a high-level GIS environment, to those with little (sometimes no) knowledge of what is required to make rudimentary use of DBD.

A basic description of the functions within the UKBORDERS system is as follows. A request (step 1) is processed by the 'Perl Gateway' which then (step 2) sets the parameters required by querying the Ingres MetaDatabase where information describing the DBD held in ArcInfo is stored. The 'Perl Gateway' then uses the information acquired (step 3) from the MetaDatabase to extract the required DBD from ArcInfo and return the output to the user.

As is the case with most interfaces to online services provided for the academic community, the UKBORDERS user interface currently under development is graphics-based and is designed for use with a 'web browser' such as Netscape or Microsoft Explorer. It has to be recognised, however, that not all users make use of such interfaces: some because they do not have a suitable computer, or

[1] EDINA is one of three JISC-funded National Data Centres which provide online access to digital data

software (although this will presumably be substantially less so over the next few years). For others who have a clear idea of the coverage they require, 'mining' down through several layers of interface can be unnecessarily time-consuming and often frustrating. Indeed, it is arguable that even when it might be considered that a graphical interface is most useful, over-dependency on objects rather than text can cause confusion and sometimes-unnecessary stress.[2]

Most of the functionality of the UKBORDERS interface is compatible with and can be used in a standard telnet emulator. At the same time, while the new browser version continues to make extensive use of text-based menus, 'step-down' maps are being added as alternative aids where appropriate. It remains, however, that the first priority of a service such as UKBORDERS is to provide access to all the data currently held. New data, including boundary sets constructed from census building blocks, are added to the UKBORDERS stock on a regular basis. Adding graphical steps to an interface providing access to these can be considerably more resource-expensive and cause more delay in provision than slotting in text references which appear in scrolling menus. An administrative, form-based interface to the metadatabase has recently been constructed which allows the straightforward incorporation of new boundary sets.

8.3 Use of digitised boundary data

A vital consideration for many academics is the measurement of change. The continual fall in price, coupled with the increasing power and 'user-friendliness' of both desktop-computer software and hardware over the last decade or so, is set to continue into the forseeable future. As a result, more use is being made of technology in research and teaching by academics in disciplines which hitherto have not been synonymous with new technology. In the humanities, for example, historians and others interested in population dynamics over the longer term have become interested in the mapping of digitised census and related data in their analysis of change.

As might be expected, geographers are by far the largest user group by discipline of the DBD supplied through UKBORDERS – 900 out of more than 1200 users, representing about seventy-five percent of the total user community. Of the remainder in terms of absolute numbers and percentages of the total the largest number come from medical departments, with social sciences in general, computing, environmental studies and planning (and architecture) also contributing relatively high numbers in this context.

So far as can be ascertained, the predominant use of boundary data supplied through UKBORDERS is in thematic mapping, and this apparently applies to geographers as well as 'the rest'. Table 1 shows the results of a survey on data use undertaken by UKBORDERS. Of three hundred users polled, 35 responded.

---

[2] *Wood, M. 1993 p.111*

Table 1. Use of DBD

Use	No. of Users
Thematic Mapping	19
Context Maps	14
Statistical Graphs	2
Total	35

8.4 The Generic Interface

It is the purpose of the UKBORDERS Generic Interface to provide straightforward access to all of the data held in the system. As well as providing the ability to map 2001 data, the supply of DBD for 2001 will add to the utility of those for 1991 and previous censuses. Users who access the data via the UKBORDERS Generic Interface will be able to extract 2001 DBD along with DBD from earlier censuses – 1991 and 1981, for example – and will be informed over what period of time the areas they select are valid. This will help them determine whether the areas with which they wish to work are compatible over time; if not, then the metadata supplied by UKBORDERS will help them decide which boundary sets are most appropriate for the statistics they wish to map.

The 'Generic' interface to the UKBORDERS database, currently under development can be described in schematic form as follows. In order to extract one or more files (which may contain one or more coverages) users must make decisions in four basic stages:

1. **Set Constraints**
   By default, all of the data held in UKBORDERS is selected. Users may choose to modify this blanket selection by constraining one or more of:
   - A *time period* of interest – any data which is time-stamped for any part of the period set is selected
   - One or more of the five *geography* 'types' held in UKBORDERS – Census, Administrative, Electoral, Postal or 'Other'
   - One or more of the three *countries* currently represented in UKBORDERS – England, Wales and Scotland[3]. (This constraint may also be set by clicking on a map consisting of the boundary outlines of the three countries).

1. **Select File(s)**
   - Depending on the constraints set in Stage 1, Stage 2 immediately provides a file list *or* a list of the level(s) of aggregation available (for example, ED, Ward, District) selecting from which then produces a list of available files

1. **Select Data Transformation(s)**
   - *Generalisation* (default 'off')
   - *Precision* - where appropriate in relation to …
   - *Format* - Arc/Info ungenerate, polys; Arc/Info ungenerate, lines; Arc/Info ungenerate, points; Arc/Info export; Autocad DXF; MapInfo; and MapInfo Native Table (for use in Microsoft Excel)

[3] Supply of DBD for Northern Ireland has not yet been agreed with OSNI

Preview and / or Download output.

- Provision of a 'preview' option to allow the user to confirm that they have specified, and the system has delivered, the required coverage, is regarded by the UKBORDERS team as a prerequisite

5. The 'Express' Interface

The 'Express' interface contains the same set of procedures as the 'Generic' interface except for the omission in Stage 1 of the first two steps – time period and geography type. In addition, because most aggregations for any particular census can be hierarchically nested, a set of 'step down' clickable maps is provided in Stage 2 as an optional selection tool

Confirmation that the DBD supplied to users is 'fit for purpose' is provided in the creation of a metadata file detailing source(s), originator(s), ownership, requisite copyright statement, secondary processing / derivation, etc., along with a résumé of the choices made by the user in selecting her / his output.

8.5 Boundary Data Recommendations

8.5.1 The 1991 DBD

The 'raw' data for 1991 as received by UKBORDERS were digitised by EDLINE (England & Wales) and GRO(S) (Scotland) from a variety of source scale mapping. These cannot now be attributed because detailed records of which maps (and, thus, scales) were used for particular areas have apparently not been retained. Digitisation of Enumeration Districts was completed for England & Wales and Postcode Units for Scotland. Higher aggregates were then derived at UKBORDERS from these basic 'building blocks' using indices[4] supplied by EDLINE and GRO(S). The process of deriving higher aggregates involves the dissolution of unwanted internal boundaries. For example, the creation of the boundary for the Local Government District of Sheffield involves combining 1,057 EDs and then dissolving all of the vertices, which are not required for the District boundary. Similarly, Sheffield contains 29 Wards, the production of the boundaries for which involves the combining of the appropriate EDs for each ward, dissolving those vertices which do not delineate the ward boundaries and outputting the result into a file whose extent is that of the District of Sheffield.

Given the appropriate look-up set, because higher aggregate areas are built using EDs / OAs as 'building blocks', a variety of new intercensal census and non-census areas can also be constructed. This is particularly true for Scotland where postcode-based Output Areas are used as the base unit. For example, the introduction of Unitary Authorities in Scotland in 1996 meant that the local government regional and district areas previously used as higher aggregate census reporting areas became administratively 'redundant', while the census statistics also lost their usefulness for most new authorities. An index supplied by GRO(S) allowed the new areas to be built with a high degree of accuracy, providing a very useful new administrative boundary set for academics as well as local and national government. In this instance, Local Base Statistics were also produced for the new areas, which extended the 'shelf-life' of the census for those interested in measuring population for resource allocation, electoral and related purposes.

Constructing boundary sets is one thing, making them useable by a diverse user community is another. Users of DBD in Higher Education require coverages for use in a variety of hardware environments and with a number (and different versions) of software packages. In order to meet user demand and increase the utility of the DBD on offer, further derivation occurs in the output of requested coverages in a choice of formats. The boundary data are held in UKBORDERS in Arc/Info but are offered to users in

[4] The 'indices' referred to are the *Area Master File* (England and Wales) and *Output Area to Higher Aggregate Index* (Scotland). Both are in effect look-up tables.

a variety of other formats: Arc/Info ungenerate – polys; Arc/Info ungenerate – lines; Arc/Info ungenerate – points; Arc/Info export; Autocad DXF; and MapInfo. It is intended to offer other formats for which there is demand, such as MapInfo Native Table (for use in Microsoft Excel), MapViewer and Atlas in the near future. It is worth noting that because some software cannot handle the number of nodes present in the polygons constructed for some higher aggregates from the DBD as supplied; generalisation is necessary to allow any use of the data by particular users. DBD are held in UKBORDERS in double precision. Some of the (versions of the) software packages used in the community can only handle single precision, so UKBORDERS offers DBD in both single and double precision to allow users the choice which best meets their requirements. At the same time, the system will not allow the extraction of inappropriate precision and format combinations – MapInfo 3.0, for example, does not use double precision, so only single precision can be extracted when the DBD are requested in that format.

8.5.2 DBD Requirements for 2001 – Content, Quality, Format, Delivery

Because of the wide variety of uses which the higher education community makes of DBD, a basic requirement is that they should offer as high a degree of flexibility as possible. It is a working assumption of the census offices that there will be an increased demand for customised geographies for 2001. The model used in 1991 of supplying DBD for the lowest aggregate reporting units along with indexes which allow these units to be used as basic building blocks in the construction of higher aggregates is a highly successful example of the way in which a wide range of customised geographies can be constructed efficiently and at low cost. The use of indexes caters for the building of libraries of 'standard' census higher aggregate areas, but also greatly increases the utility of Census DBD by making possible the construction of 'ad hoc' and new statutory areas for the mapping of statistics for 'non-standard' census, census area-related and non-census aggregations. This is particularly true where the basic reporting units, as was the case for Scotland, are based on postcodes.

An important consideration for many users in higher education is the measurement of change. While it is recognised that intercensal change is not necessarily a basic requirement of users in other sectors, the utility of the 2001 DBD in academic teaching and research (which often leads directly to the creation of value-added census products) would be greatly enhanced if they were constructed so as to make them as 'time-flexible' as possible. On the assumption that the basic DBD building blocks for 2001 will be polygon- (as opposed to address-) based flexibility could best be achieved by supplying the data with time-stamp attribute information attached to arcs rather than polygons.[5] This would act as a comprehensive record of boundary change and would allow the efficient 're-use' of arcs in the construction of new polygons when boundaries were redefined from time to time. After all, census boundaries are, effectively, only 'valid' for the day that the census is taken. Thereafter, they are essentially a collection of administrative, electoral, postal, etc., boundaries for areas whose validity over time is determined by other agencies and factors. Time-stamped arcs would allow greater precision in the building of new higher aggregates for other purposes, regardless of whether full sets of census statistics were produced to match them, thus increasing the utility and value to both user community and vendor.

As stated above, many (if not most) users of boundary data in the higher education community require them for the production of thematic maps. Because of this, the recent introduction of a sophisticated generalisation option in the UKBORDERS system has proved popular with users - 25% of all extractions were generalised in each of the first two months of its availability.[6] The generalisation option is an important development in our continuing strategy to encourage use of the DBD by as wide a range of higher education community users as possible. This should in no way deflect attention from that fact that highly detailed DBD are still a requirement for many, particularly 'high-end' GIS users.

---

[5] Gregory, I. and Southall, H. 1998

[6] Mackaness, W., Edwardes, A. and Urwin, 1998

Given the demonstrable integrity of the DBD produced by the UKBORDERS generalisation algorithm, the needs of the higher education community would best be served by following the precedent set in 1991 of delivering highly detailed DBD to directly meet 'high end' needs, which would at the same time provide the raw material for the provision of quality-assured generalised data through UKBORDERS. Indeed, a somewhat unexpected bonus of the UKBORDERS generalisation module is that it is a precise quality control tool for ungeneralised boundaries. The process of generalisation picks up and flags errors in the original data because it is designed not to proceed when it comes across any defect – slivers, unclosed polygons, self-intersection, etc.

As regards the format in which the 2001 boundary should be delivered to the higher education community, UKBORDERS has no over-riding preference. We will undertake to convert the data from any of the established proprietary formats to those which our user community requires. Currently, the most requested formats are Arc export, ArcView Shape files and MapInfo. Whether this will still be the case in 2001 remains to be seen. But it must be recognised that in whatever format the data are delivered, delivering them on to users *must* involve some transformation. The transformations that we currently offer users, such as generalisation and format conversion, maintain the data in a 'fit for purpose' state.[7] This is an essential requirement that we are committed to continue.

We look forward to the early delivery of the 2001 digitised boundary data to the academic community, so that our users may gear themselves up to the mapping of the 2001 statistics as soon as they are made available - particularly if, in the event of a One Number Census, the statistics take a little longer to produce than in the past!

References

Burnhill, P. and Morse, D. J. *Census Geography* Paper given at 'ESRC Workshop on the 1991 Census and Large Government Datasets', 6-10 September, 1993, University of Manchester

Denham C (1993) "Census Geography - an overview", in Dale, A & Marsh, C (Eds.) *The 1991 Census User's Guide*, London, HMSO

Gregory, I. And Southall, H. 1998 'Putting the Past in its Place' in *Progress in GIS: Proceedings of the GISRUK Conference, Edinburgh*

Mackaness, W. Edwardes, A. and Urwin, 1998 T. 'Self-Evaluating Generalisation Algorithms to Automatically Derive Multi-scale Boundary Sets' in Proceedings of the ESRC Mini-Development Projects Workshop, held at the Office for National Statistics, Titchfield.

Medyckyj-Scott, D. and Morris, B. 1998. 'The virtual map library: providing access to Ordnance Survey digital map data via the WWW for the UK Higher Education Community'. *Computers, Environment and Urban Systems*. (in press)

Medyckyj-Scott, D. and H. M. Hearnshaw (Eds). 1993. *Human Factors in Geographic Information Systems*. London: Belhaven Press.

Openshaw, S. 1995 'The future of the census' in Openshaw, S. (Ed). *Census Users' Handbook* Cambridge: GeoInformation International

Wood, M. 'Interacting with Maps' in Medyckyj-Scott, D. and H. M. Hearnshaw (eds). 1993. *Human Factors in Geographic Information Systems*. London: Belhaven Press

[7] However, while we have demonstrated this so far as generalisation is concerned, we do not know of any systematic study into the effects of format conversion; we suspect there are no effects in converting data from one proprietary format to another which renders the 'unfit for purpose'.

CHAPTER 9
WEB BASED INTERFACES TO AREA STATISTICS FROM THE 2001 UK CENSUS: RECOMMENDATIONS

James Harris

This paper will address the digital dissemination of census data in the UK. A number of important weaknesses in the 1991 model of data access are identified and solutions explored in the context of the Casweb experimental Web-based interface to 1991 Census area statistics.

It is clear from the volume and nature of user support queries handled by the Census Dissemination Unit that present facilities for accessing census data fail to meet the requirements of a significant proportion of users. While there are a variety of packages that can be used to access and manipulate Census data, the most common means of access for academia is by using SASPAC over the network on a remote UNIX server, such as the MIDAS machine. Not only does such a system require the user to have some familiarity with the UNIX environment, but the inherent structure of the software also makes significant demands in terms of learning overhead and prior knowledge about the structure of the data. While SASPAC can also be used in the more familiar environment of the desktop PC (and, indeed, a more user-friendly interface is now available) this does not address the problem for academic users who will continue to access the Census from a central data resource over the Internet.

Another problem with the current system - also to some extent a legacy of decisions taken back in the late 1980s - is that the data is bound up in an arcane proprietary format. While this makes for rapid data retrieval, the technological constraints are very different on today's hardware, and it is a pity that the format of the data precludes the use of powerful and functionality-rich industry standard statistics and GIS packages to subset and manipulate the raw data. Any new system must address this issue, and while compatibility with prior releases of census data will remain a priority, supporting open data formats will prove to be the best defence against future dependence on legacy systems.

The priorities for any new system can be summarised as follows:

- Flexible, secure, platform independent on-line access
- Intuitive, easy-to-use software
- Industry standard (and non-proprietary) data formats
- Integrated interface and analytical toolkit

The Casweb system consists of a large, relationally structured database of Census counts and metadata which is held centrally on an NT server at Manchester Computing. A Web interface allows users to formulate data extraction queries through a series of menus and maps accessed via the Web. The system is therefore built around a client-server architecture with the user logging in from their desktop computer and interacting with the Census database using a Web browser. Casweb is a secure system, with users having to authenticate themselves by user name and password to gain access to the interface. The map-based front-end places spatially-referenced census data within its geographical context and incorporates a number of spatial data resources including the census boundaries and digital map data.

The system has been implemented across a range of development environments and various platform/software combinations have been evaluated during the course of the project. The Web interface uses a combination of HTML forms, server-side scripting and proprietary server software to pass SQL queries to the database via CGI and the Web server API (see Figure 9.1). The implications

and advantages of employing advanced methods for user interaction and data retrieval were discussed in the Workshop.

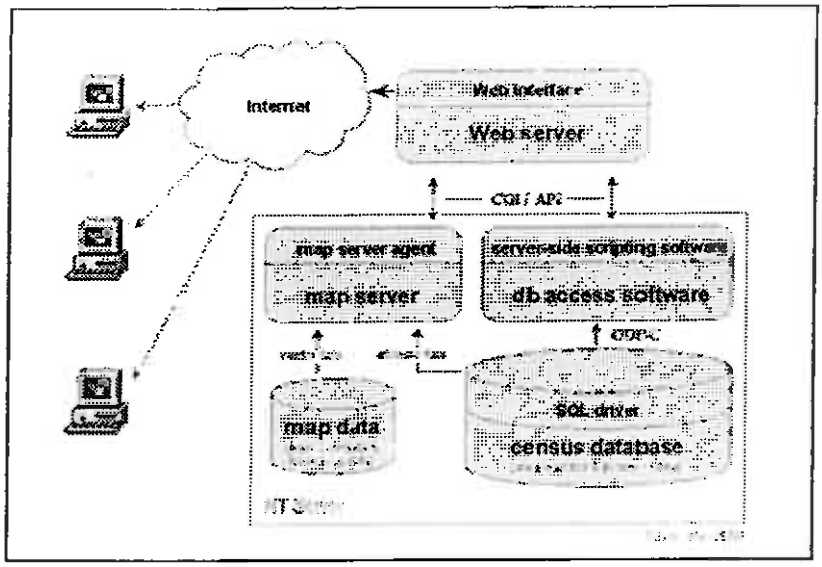


Figure 9.1 The Casweb system for accessing census area statistics

CHAPTER 10
INTERFACES TO INTERACTION DATA

Oliver Duke-Williams

10.1 Background - the Census and interaction data

Most questions asked in the 1991 Census were about individuals or thehouseholds they live in; however there were two questions which dealtspecifically with locations other than the household. Employees and theself-employed were asked about their place of work, and people who hadmigrated over the previous year were asked about their location of formerresidence. The responses to these two questions form the basis of the interaction data in the Census. These data were included in a number of datasets in different forms - the Small Area Statistics (SAS) and Local Base Statistics (LBS) contained some tables giving information about residents in particular areas who had been migrants, and the Sample Of Anonymised Records (SAR) included variables for both workplace and location of former residence. The main datasets used for analysis of interaction data, however, were the Special Migration Statistics (SMS) and the Special Workplace Statistics (SWS), and it is these two sets of data that are considered in this paper. Strictly speaking, the terms SMS and SWS refer to an extraction system used to produce sets of data for customers, although the terms have been generally used in the academic community to refer to the data themselves, specifically the ESRC/JISC purchased data, which included the following data:

- The SMS Set 1 - ward level origin-destination interaction data
- The SMS Set 2 - district level origin-destination interaction data
- The SWS Set C - ward level residence-workplace interaction data

Sets A and B of the workplace statistics are origin and destination specific, and thus are not within the scope of an assessment of interfaces to interaction data. The nature of the ESRC/JISC purchase of the SMS data is described in fuller detail by Rees and Duke-Williams (1995).

This paper briefly describes the interfaces used in 1991 to study the above datasets and then considers the aspects of interaction data which differentiate them from other types of data. Given this background, the paper then makes proposals about necessary features of future interfaces to interaction datasets.

10.2 Interfaces to the 1991 interaction data

Two programs were used as interfaces to the 1991 datasets, *Quanvert* produced by Quantime Ltd, and *smstab/swstab*(hereafter referred to simply as *smstab)* produced by the Centre for Computational Geography at the University of Leeds. The latter was initially produced solely to perform quality assurance checks on the SMS data, but was later adapted to act as a user-interface to both datasets. Table 1 shows the size of the datasets as originally supplied and as used by the interface programs. The file sizes are large and therefore impractical to handle directly with programs such as SPSS. Subsets of the data for local areas might effectively be used with such packages, although the process of loading data from the original files into SPSS may require substantial effort due to the hierarchical data structure.

Table 10.1 File sizes and characteristics

Dataset	File sizes			Cases[a]	Variables[b]
	Original[c]	smstab[d]	Quanvert[e]		
SMS 1	171.5Mb	56.5Mb	94.7Mb	2,336,318	12
SMS 2	281.8Mb	44.7Mb	116.9Mb	362,896	94[f]
SWS C	2.5Gb[g]	163.2Mb	481.4Mb	1,156,670[g]	274

Notes:
a. Cases in original data. Some cases are redundant.
b. Variables in original data. Both smstab and Quanvert added some derived variables.
c. Files having been stripped of trailing blanks.
d. Combined size of data file and index file.
e. Total size of directory used for each dataset.
f. Some variables are only included for destinations in Wales or Scotland.
g. Estimated: original file no longer available.

Quanvert and smstab provided very different interfaces to the datasets, with Quanvert providing an interactive front end to allow users to generate and run queries and smstab requiring users to supply the query directly in the form of either command line arguments or a command file. (Quanvert permitted the keystrokes from an interactive session to be recorded and then used directly as input as a command file, however such files were difficult to modify, especially for novice users). Quanvert provided a low level conceptual model of the data, with basic queries returning the number of records which matched a particular query; in order to tabulate the actual numbers of migrants with particular characteristics, a 'special table' has to be defined. In contrast, smstab generated origin-destination observations and flow matrices directly, but provided no mechanism for querying the number of records which met given criteria.

10.3 What makes the SMS and SWS different from other Census datasets?

The assessment of existing interfaces and proposals for future interfaces can be broadly divided into two categories: interface elements common to all Census datasets, and elements which address problems which are specific to the interaction data. In order to assess the latter group, we need to consider what aspects of the interaction data are different from other datasets. The SMS and SWS both feature aggregated tables similar to the SAS and LBS, but these tables are far fewer in number, and generally simpler and less detailed, being cross-tabulations of only 2 or 3 variables. Both datasets have a level of geographical detail that is comparable to the LBS, and thus much more detailed than the SAR. The implication of this is that identification of the data that the user requires is not a significant problem (although obviously there is no task so simple that a badly designed user interface can't render it near-impossible), whereas selection of the geographic regions of interest is a significant problem.

There are two issues of area selection in interaction data that are unique; firstly the need to select two sets of areas rather than one, and secondly the fact that special attention has to be paid to the way that areas are aggregated to form new zones. In the first case, the user has to select both origins and destinations, although the importance of this may not be immediately apparent if either the origins and destinations are identical, or if one of the sets is simply 'the rest of the country'. However, where the origins and destinations are different, the importance can simply be stated as: if it's difficult or annoying to select one set of areas, then it will be doubly so to select two sets of areas.

A fundamental difference between interaction data and other types of data lies in the way in which they can be aggregated. If a user wishes to create a new zone X from areas A and B, then the total number of widget-makers in the new zone will simply be equal to the sum of the values of the two input areas. However, if we want to generate the total of a variable telling us how many people work in widget

factories outside their area of residence, then we can not simply add up the "works outside area" values for areas *A* and *B*. This is because some of the people who work outside *A* will work in *B*, and *vice versa* and we do not want to count these people when we calculate the number of widget-makers who work outside the area *X*. As long as an interaction matrix is available (as, by definition, it is with the interaction datasets) rather than simply a set of totals for origins and destinations then the new values can be calculated correctly. The task is a simple one in theory, but one which can be difficult in practice, especially if the data user is not familiar with any programming languages. For this reason, it is highly desirable for the user interface to provide aggregation facilities.

10.4 Interface issues

The interface that the user sees is only part of the overall interface to a dataset, and the user interface can only be as powerful as the underlying structure permits it to be. This section of the paper therefore not only makes recommendations about necessary features from a users point of view, but also makes proposals about efficient methods of implementing such an interface. One of the oft-stated aims of both the academic community and the Census Offices has been that the level of access to Census data should be improved, and that we should strive to have more people using Census data. In this spirit, the proposals made here aim to make few assumptions about technical capabilities of users, in order to permit a high degree of flexibility.

User requirements in an interface vary widely. There will be novice users who are unfamiliar with the data and need significant prompting, people who want only a few variables to be used as a minor element of an otherwise unrelated project and who do not want to spend a long time learning how to use a new package, advanced users who know what they are doing and want a "minimum fuss" method of accessing the data and so on. It is unlikely that any one interface can meet all of the user requirements adequately, and thus if access to these datasets is to be widened it is likely that more than one interface will be necessary. Having said that, it is to be hoped that interfaces provide a reasonable degree of ability to be configured to suit the user -- there is no need to provide two interfaces where one configurable one will do.

Recommendation I1. *The design of a system to use the datasets should expect to allow alternative interfaces to be used by different users, and for interfaces to be capable of being modified to suit individual users' preferences.*

The most significant feature of interaction datasets from the user's point of view is the selection of the areas of interest. The SMS Set 1 and SWS Set C are both ward level datasets, and scrolling through 10,000 ward codes to find the required code in an interactive interface is not an enviable task. Just as different users have different requirements for features in an interface, so different queries will want to handle geography in different ways. It is unlikely that a single mechanism for selecting areas of interest will be the most efficient method for all tasks. However, users will not want to switch to an alternate interface simply to handle geography in different ways.

Recommendation I2. *The interface(s) should have multiple complementary ways of selecting geographic areas of interest.*

By this, it is suggested that within an interface there should be more than one way of selecting the areas of interest. This may range from simply typing a code or set of codes into a particular field (which is probably the most efficient approach for those who know what they are doing), to using an interactive map to select areas. In many cases the user will want to exploit the hierarchical nature of the Census geography -- they will want to select all wards in a given district, or all districts in a given county etc., and thus it would be advantageous to allow for this possibility. This might be implemented either using a series of pull down menus, or via a map interface where the user can browse at a high level and then,

for example, click on a district polygon and carry out some 'select all wards' procedure, without having to see / download a ward level map.

Recommendation I3. *The interface(s) should exploit the hierarchy of British geography so that areas in a high level of the geography can give direct access to all their component low level areas.*

This proposal makes implicit reference to one of the key issues in Census geography -- the fact that there are not one but many geographies of interest to users. As well as the hierarchical groupings of wards, districts, counties and so on, there are alternative groupings of areas using characteristics such as postal geography. In addition, there are non-contiguous geographies based on characteristics of areas rather than there locations: a user may, for example, want to study migration between types of places – major cities, other cities, towns, rural areas etc. As discussed in the section considering the differences between interaction data and other types of data, there are specific problems in the aggregation of interaction data. For this reason, it is vital that the interfaces to interaction data incorporate an ability to re-aggregate data according to the user's requirements.

Recommendation I4. *The interface(s) should allow the user to re-aggregate data. There should be a library of commonly used aggregations of areas, and also a mechanism for the user to generate their own aggregation of "building block" areas.*

Barr *et al.* (1998) propose a central database of look-up tables allowing users to convert between various geographies, and it would therefore be sensible to allow interfaces to exploit such a database. The principle of connectivity between different Census outputs and projects is one which offers much scope for improving the power and flexibility of interfaces. For example, a user may wish to design an aggregation of areas based on some demographic characteristics of input areas. It would be much simpler to do this within the context of an aggregation design tool that could load some data from the 2001 equivalent of the SAS and perform some simple operations on the data than it would be to have to do it externally. However, it is unwise to try and aim for monolithic interfaces which can do everything, at the expense of being difficult to use, high in CPU memory demand and intimidating to users by dint of having too many features.

Recommendation I5. *Interfaces to all datasets should be designed with the knowledge that they are not the sole window on Census data. Ideally they should be network-aware and capable of exchanging data, but at a more mundane level they should at least be able to save and load some common data formats.*

Some of the above suggestions have made reference to map based interfaces. These are desirable because it make the selection of areas more intuitive, although it is perhaps less useful that a simple list mechanism at local levels. For example, a user may wish to select Loughborough as an area of interest, and have a vague idea that it is in Leicestershire. They will, hopefully, be able to locate Leicestershire from a national map, but at a more focussed level, may have to query all units to find out which one is Loughborough. his is, of course, a trick question, as the name "Loughborough" is not used as either a ward or district name. Therefore there is the additional proposal, applying to all Census outputs that:

Recommendation I6. *A gazetteer is available linking Census geography to other commonly used place names.*

However, there are some other problems which need to be considered. Firstly, as described above, it is in the nature of interaction datasets that two sets of areas have to be selected. With a map interface this may involve zooming in to one area to select some origins, and then zooming out and zooming in on another area to select some destinations. This may be time consuming, and highlights the problems with map interfaces of selecting spatially separated areas. Furthermore, map interfaces also have

higher demands on the capabilities of the machines on which they run, and also pose difficulties for users with visual problems, as they are incompatible with screen-readers, and also for users who have difficulties in using a mouse to drive software, as menu based interfaces are generally more amenable to keyboard-only input if required.

Recommendation I7. A map interface is desirable, but should not be the sole interface method for selecting areas of interest. This proposal follows logically from Recommendation 2.

One of the inevitable consequences of having to select lots of areas, is that it may well take some time to set up a query. It is therefore vital that there be some mechanism for the user to save a query in some sort of scripting language. Such a file must be able to be saved in a plaintext form by the user, so that they can modify it without access to the various interface clients, and can submit it to some batch processor if required. The latter may be necessary if a user wants to carry out a large number of similar queries. The most effective way of doing this would be if all interface clients produced as their output a query framed in some standard language.

Recommendation I8. All interface clients should produce output in the form of a query written in a common language. This language should be understood be all interfaces, and should be a high level language specified for the purpose of querying the Census interaction datasets.

The purpose of having a common language that is understood by all clients is that a user should be able to save a query that has been produced by one client, and load that query into an alternative client, that might be more suitable for the job in hand. Alternatively this may be seen in terms of two users sharing queries that they have set up. If the first user has used one client to set up the query, then the second user should not be obliged to use the same client to modify it. The point of stating that a high level language should be used is to suggest that the user should not be faced with a query written in SQL or a similar database querying language. Rather, they should see a command file which indicates what data is to be retrieved for which origins and destinations, and what output formats should be used, with terms which make sense within the context of the SMS or SWS or future equivalents.

Recommendation I8 introduced the term "client" into this paper. This suggests a client-server structure. This is one method of implementing a system with multiple interfaces, and would also facilitate the use of a common query language: all clients would produce a query which would be passed on to a server which would interpret the query and extract data from a database. However it would also be possible to have a system in which all interfaces incorporate a module which understands the query language and can extract data directly. Ultimately, it may be the case that one or more interfaces incorporate the ability to extract data themselves from some datafiles, but others pass on requests (maybe via a module which understands the Census interaction data query language) to an industry standard database engine, which would extract the data, presumably with greater speed.

One of the key splits in the model of data provision is between the use of a central data server, and the distribution of data directly to the user, whether via the Internet, on CD-ROM, by carrier pigeon, or any other method. However, there is no reason why both markets can not be served by the same product. A client-server structure could be run with the server either on the same machine as the client, or an a remote machine. Given the above discussion of a client-server model versus a complete interface model, it may be the case that a CD-ROM contains datafiles which a local interface can query directly, together with tools to set up the data for a more efficient DBMS approach. At present the amounts of data involved still make it impractical for most users to have a copy of a dataset of the size of the ESRC/JISC purchase, but for smaller subsets it should be possible to distribute data on CD-ROM or a similar medium.

An advantage of central provision is that both version control of the data is possible (*i.e.* if problems are discovered with the data, then a revised dataset can be installed) as well as central management of derived variables that researchers wish to share with the rest of the academic community. These advantages can also be seen as explicit disadvantages of a dataset distributed to users -- it is not possible to update the dataset without having to print and distribute a new batch of CDs. Furthermore, end users of distributed media may not be aware that the dataset they are using is out of date. However, it is increasingly likely that end users will have some level of connection to the Internet, and thus the local client may be able to query a remote server from time to time in order to check the version number of the most up to date data file.

Recommendation I9. *There should be a central facility providing access to the data as well as distributed "personal" versions. The central facility should manage version control of the data for revisions and derived data, and respond to meta-queries as well as data-queries from remote clients.*

The emergence of World Wide Web in recent years provides solutions to many of the problems posed by the suggested set up. Web browsers can connect to either remote or local servers, and thus there is no reason why a version of the data distributed on CD-ROM could not come complete with both a web server, a web browser and support scripts and documentation. In practice, web applets written in Java or similar languages can optionally read files on a local machine (subject to normal file permissions) and thus it would be feasible to distribute a Java interface to the data which would directly query some local datafiles. A significant advantage of this is that (subject to Java "settling") the interface would then be platform independent.

With a Web based structure the user would be able to use either the supplied tools, or incorporate the database with an existing server and use his/her own preferred browser. This suggestion has the significant advantage that the key elements of the system - the server, the browser and the script interpreter (usually perl) have already been implemented in many forms for all popular hardware and software platforms. The remaining element of the proposed structure is the database, if a client-server set up is used. It is likely that different databases will be installed on different servers, especially where different operating systems are involved (most probably a Unix system in the case of national level servers, and NT systems for smaller tasks). This problem is not insurmountable if there is a common mechanism for interacting with the database, and thus it is is suggested that industry standard databases are used, which can "talk" one of the familiar database protocols. This may still prove problematic in the case of distributing a basic querying package along with the data on a CD-ROM, as there will be a need to distribute a database system as well as the web server and other tools.

Recommendation I10. *The most logical structure for the database system is for it to be implemented using an industry standard database connected to a Web server, with Web browsers acting as interface clients. At least one interface should be provided which can query some local datafiles without having to interact with an external database system.*

The above proposals have dealt with the way in which users can extract data that they require, but have not dealt with form of that data. The experience of the use and provision of interfaces to the 1991 data is that however many forms of output are available, a user will always be able to suggest an alternative one that they desire. It is essential that all interfaces are as flexible as possible in the way that they present data.

Recommendation I11. *Interfaces should allow multiple views of the data, including origin-destination flow pairs, flow matrices, origin or destination constrained views etc. They should also allow, as far as possible, cross-tabulation of the data in a flexible manner.*

The multiplicity of ways in which users want to conceptually view the data is reflected in the wide variety of formats in which they will wish to save the extracted data.

Recommendation 112. The interfaces should allow users to save data in a wide variety of forms. This should include common exchangeable forms such as comma separated values as well as forms compatible with popular applications which may be used to report or post-process the data, such as SPSS, SAS or the Microsoft Office range of products. Output should not be limited to proprietary data formats, and no assumptions should be made that the output form that a user might want is related to the type of computer on which they are performing the extraction.

10.5 Conclusion

This paper has made a number of proposals about interfaces to handle interaction data from the Census. If these proposals are accepted and interaction datasets are produced from the 2001 Census, then further work will be required in order to sensibly specifiy the mechanisms used to implement these proposals. The form of interaction data from the 2001 Census is currently under debate, and many of the issues are discussed by Rees (1997}.

The proposals about the features of interfaces from the users point of view centre around the geography of the datasets, due to the assertion that this is the most complicated element of the datasets. The goal of the interface(s) should be to provide both an efficient way of selecting areas of interest, and an elegant way of re-aggregating data correctly. Additional proposals refer to methods of implementing a system to query the database, and suggest that a common approach can satisfy both users who prefer local access to the data and a system of central provision of data. Furthermore, the use of platform independent technologies can serve to produce common solutions which therefore reduce the development costs of the system.

References

Barr, R. and Harris, J. and Cole, K. and Dawes, I. (1998) Integrating look-up tables for census access: a database approach, *ESRC Census Projects Workshop, ONS Titchfield, 16-17 April 1998*, ESRC.

Rees P. (1997) Migration statistics from the 2001 Census: What do we want?, Working Paper 97/06, School Of Geography, University of Leeds, UK.

Rees, P. and Duke-Williams, O. (1995) The story of the British special migration statistics, *Scottish Geographical Magazine*, 111, 1, 13-26.

CHAPTER 11
SARS INTERFACES 2001

Ian Turton

11.1 Background

A major UK census milestone was reached in 1993 with the release of Britain's first ever, official, sample of anonymised census microdata. In fact this amounts to a 3\% sample of the 1991 census being released in anonymous form, without any names and addresses and only coded to large geographical areas. This followed almost a decade of discussion and debate about the content, need and confidentiality aspects of possible microdata from the census.

Turton and Openshaw (1995) agree with Marsh (1993) when she summarised the advantages of the SAR as follows "In the final analysis, however, the value of the SAR lies in the fact that, in contrast to tabulations produced by the Census Offices, we do not have to plan in advance all the useful ways in which the data can be presented. ... Almost every census user will have experienced the frustration of finding that the table they required was slightly different from the one they in the end had to use." (p297). This is a slight exaggeration. Nevertheless it is the historically relatively small number of essentially non-spatial census users who will probably benefit most from the SARs, indeed this group may well be instrumental in opening up the census data resource to a broader social science constituency (Openshaw 1995).

11.2 Lessons from 1991

The SAR is a big dataset, though not massive by modern standards. The original SAR data files amount to 126 Megabytes (79Mb for the 2% sample of individuals and 47Mb for the 1% sample of households). If the data are loaded for SPSS then the system files amount to 93Mb and for SAS then the system files go up to 176Mb.

The original intent in 1991 was to load the SAR data files on to a large mainframe at Manchester Computing Centre (MCC) to provide a table output service for users using the same relational database package model 204) as used by the UK census agencies in processing the 1991 census. Indeed, the Census Microdata Unit at the University of Manchester was established in 1992 to provide three sorts of dissemination service relating to the SARs:

1. an online service, available over a network using database software such as SIR and Quanvert and statistical packages such as SPSS and SAS;
2. distribution of the raw SAR data; and
3. a customised tabulation service.

Turton and Openshaw (1995) describe a fourth diffusion and dissemination path that was developed by an ESRC research project (1992-94). The idea was to develop portable data access software designed for use with the SAR and which can be transferred using ftp over JANET to any unix workstation with sufficient disc space. The aim was to provide an alternative standard means of accessing and using the SAR that was likely to become increasingly relevant during the 1990s. Indeed the continued fall in cost of workstation hardware emphasised the importance of storing the SAR data in a form suitable for a unix workstation and a PC environment. This provided academics (and others who buy the data) with a very different access path, one based on the distribution of the SAR data with associated specialist

software designed to allow both expert and non–expert users to easily and quickly obtain maximum benefit from the SAR datasets. Indeed, it was thought likely that both experienced and enthusiastic SAR users would increasingly want the SAR data available on their local unix and PC systems and that they would find specially developed SAR relevant software of considerable practical assistance, to complement the widely available general purpose statistical packages.

In the dissemination of the 1991 census data it was seen as important that "the data are: easily accessed with a minimum of alien computer control language, analysis is fast, the data is largely self - documenting, and a wide range of different social scientist skill levels can be catered for. Access has to be easy, straight forward and intuitively obvious, but also table creation is not sufficient by itself, so new tools are needed to help users cope with the SARs." Turton and Openshaw (1995).

In 1991 the SARs were accessed by a variety of software packages, though the majority were text based, reflecting the slower network speeds common at the time. Commercial packages such as SPSS, SIR and SAS that were already familiar to academic users were provided at MIDAS. MIDAS also provided a commercial package, Quanvert which proved unpopular with users as it was less than obvious how to use it and it did not treat all variables equally, for example geographic areas and occupations had to be row not column variables. USAR was provided as free software to anyone who had legal access to the SARs. As a new package this caused some problems in take up from established users of survey data but proved very popular with novice users as it was menu driven and would run on a variety of machines from PCs to Unix mainframes.

11.3 Background to the 2001 SARs

11.3.1 Potential SARs in 2001

Dale and Elliot (1998) discuss a series of possible SAR formats for the 2001 Census. Their key proposal is for the level of geographic detail to be increased by a lowering of the threshold size in the individual sample to 90 thousand people (from 120,000 in 1991). They also suggest that increasing the sample size from 2% to 3% would have little effect on confidentiality. They also report experiments on the potential for a third SAR with a more detailed geography, of the order of wards, but with many variables recoded.

The effect of these proposals on the software required to access the SARs would be to increase the size and complexity of the datasets to be handled. But this is not of any real consequence as since 1992 machine speeds have increased by at least an order of magnitude and a further doubling or tripling can be expected by the time the 2001 SARs are available.

11.3.2 User requirements

Users of the 2001 SARs are likely to have many of the same requirements as users of the 1991 SARs. Brown and Dale (1998) report the results of a survey of SAR users and non-SAR users. The majority of users were making use of the SARs as a research tool, with only one third using them for teaching. When questioned on ease of use by Brown and Dale (1998) only 21% of respondents said that existing systems were easy to use, though in some cases this related to the documentation or the Unix system used at Manchester. There were also complaints as to the time taken to produce multiple tables. Users showed a large spread of systems that were used to access the SARs, ranging from a large central service (MIDAS) to stand alone PCs. This shows the rapid advances in computing power that have occurred since the 1991 SARs were first thought of. Many users expressed a preference for the data to be delivered either on CD-ROM or via the world wide web.

11.4 Proposals

11.4.1 One package or many?

While it would be possible to produce the SARs in a format that could be accessed by only one package, this would be very unwise. A single package will never satisfy the needs of all users, it would be too complex for novices and too limiting for the experienced user. There is also the question as to how do we check the accuracy of the package, if there is only one answer then how do we know if it is the correct one?

Recommendation I13. *More than one package is essential to the viability of the SARs in order to provide a robust and flexible solution.*

11.4.2 Specialist or general software?

There are a variety of general statistical packages available that are capable of handling the whole 1991 SAR or some subset of it. However many of these packages have problems associated with them, some can only use a subset of the data, fine if that is all you need, but a problem when suddenly you need to compare your results to the national figures. Other packages have restrictions on how some of the variables can be used, for instance the district variable in the 1991 individual SAR has too many categories for some packages to handle.

By contrast USAR which was developed specially for the SARs has none of these problems associated with it. Users can use as much or as little of the dataset as they wish. No variable is restricted and all variables can be arbitrarily grouped . USAR also adds extra features that allow users to explore the data more fully (Turton and Openshaw 1994, Openshaw and Turton 1996). So while many packages allow users access to the SARs only specialised software allows the full richness of the data to be fully utilised.

Recommendation I14. *Specialist software allows more users more access to more of the SARs.*

Following on from this proposal the remainder of this paper will consider only what features a specially design package must provide to be of use to the user community.

11.4.3 How much should this cost?

At present the 1991 SARs can be accessed by a variety of commercial and free academically created packages. While cost is seldom an issue for a centrally provided academic service, it is to be expected that in 2001 more academic sites will wish to make use of the data locally, and the concerns of the public sector as to the costs of the data as well as the cost of software must be considered (Brown and Dale 1998). It would seem that the cheapest method of providing access to the SARs is via academically produced software, provided that any concerns over quality, reliability, performance and usefulness can be met.

Recommendation I15. *At least one method of accessing the SARs must be free or very cheap, so as not to deter users.*

11.4.4 Ease of use and flexibility

It is of no use to anyone if the software package produced is so flexible and feature loaded that it is impossible for a novice user to produce a table in less than 5 minutes from first meeting the package.

However a package must be capable of being extended so that an expert user can carry out the more complex tasks that they commonly turn to SPSS or SIR for at present. This must include the ability to produce new variables and if necessary the ability to add new functions easily. By using the remote methods of Java adding links to other existing statistical packages is relatively easy. The industry standard CORBA can also be used to combine different packages. The standard system should provide everything that 90% of users require.

Recommendation I16. *An interface must be provided that allows an average undergraduate to produce a moderately complex table with less than five minutes instruction.*

Recommendation I17. *The systems functionality should be easily extensible to meet new and unforeseen needs, as well as minority users.*

11.4.5 Local or central provision?

As the power of desktop PCs increases more and more users will wish to store and access the SARs from their desktop. However there will always be users that either prefer to allow a central site to take care of their computing, or who have very compute intensive tasks that will need a more powerful machine. These users will want access to a package that can run on as many machines as possible, which will also be a requirement in many departments. For example the CCG already has 3 types of Unix, and 3 versions of MS windows to deal with in a single building. Users however will not want to learn a different package for each machine, their support teams will also not want to support 6 different machines. It will probably make sense therefore to develop a specialist program in Java for portability. This also allows a single copy of the program to be developed that can be placed on a CD-ROM to be distributed without regard for the target machine. The package could be produced as a stand alone application or as an applet that a user could load from a local disk (or a remote site, if a suitable security model can be devised) with their existing web browser. Java also has the advantage that experienced programmers can customise the program as they require, and users who require extra features but lack programming knowledge will have no difficulty finding a Java programmer at a reasonable cost within their own institution.

Recommendation I18. *Any package developed must be extremely portable between machines. There is no longer any need to tie users to a central system or an approved operating system.*

11.4.6 Outputs

There are now a bewildering variety of outputs that users require from simple tables, to graphs and maps. The package developed should not attempt to provide all or even many of these possible outputs. Instead the package should provide a simple variety of common interchange formats, e.g. rich text format for tables, comma separated values for graphs, arc/info ungenerate format for maps, etc. Again the modular object-orientated format of Java would make it very simple to provide the hooks that users could use to add their own output formats or even links to packages that were Java compliant.

Recommendation I19. *Users must be able to use the output of the SARs in a form readable by any program of their choice.*

11.4.7 Documentation

In the 1991 SARs documentation was provided as hard copy and dealt only with a few specific packages. In 2001 it makes more sense for the documents relating to the SAR variables and coding to be provided in HTML both from a central site and in a form that can be bundled in with other

programs. This will allow a user to check a definition while the program is running where ever they are, as the documents could either be locally available or the central site could be reached over the internet.

Recommendation I20. *Documentation should be made available in machine readable form at no cost and with no restrictions on copying to developers and users to make access to the SARs as easy as possible.*

14.4.8 Beyond the SARs

There are more survey datasets available to the academic, public and commercial sectors than the SARs. It is therefore important that suitable software is provided either in the package or as an easy to use add on to import a variety of datasets. In the academic sector this includes the General Household Survey, Labour Force Survey and New Earnings Survey. Local government departments may be interested in using it to access council tax registers or other internal datasets. Whereas commercial sector companies and academic departments may be interested in carrying out similar research on the SARs and lifestyle databases.

Recommendation I21. The package developed should be able to read other survey datasets easily.

11.5 Conclusions

In conclusion it is likely that in 2001 the SARs will be larger and more complex than in 1991. However many of the packages used to analyse the 1991 SARs lacked features that researchers needed or were working at the extreme limits of the package or were close to too slow to be useful. In 1991 an academically produced package provided the cheapest, easiest and most flexible means of analysing the SARs. There is no reason why a similar package building on the developments in computing power and language can not be produced for the 2001 SARs. If possible work should start on this development before the 2001 census, rather than waiting for the SAR to be released as happened with the 1991 census. The package developed must be easy to use, cheap, powerful, portable and flexible in both its inputs and outputs. If all these aims can be achieved then the 2001 SARs will reach a much larger audience than the 1991 SARs managed.

References

Brown M. and A. Dale (1998) A survey of SAR users, their requirements for the 2001 SARs and their views on dissemination and support. *Working Paper No. 6*,The Cathie Marsh Centre for Census and Survey Research, Faculty of Economics, University of Manchester, Manchester, M13 9PL

Dale A. and M. Elliot (1998) A report on the disclosure risk of proposals for SARs from the 2001 Census. *Working Paper No. 5*,The Cathie Marsh Centre for Census and Survey Research, Faculty of Economics, University of Manchester, Manchester, M13 9PL

Marsh, C. (1993) The sample of anonymised records. In A. Dale and C. Marsh (Eds.) *The 1991 Census User's Handbook*. HMSO, London. Pp. 295--311.

Openshaw, S. (ed.) (1995) *Census Users' Manual*. GeoInformation International, Cambridge.

Openshaw, S. and I. Turton (1996) New opportunities for geographical census analysis using individual level data. *Area* **28**, 8, 167-176.

Turton, I. and S. Openshaw (1994) A Step-by-Step Guide to Accessing the 1991 SAR via USAR. *Working Paper 94/6*, School of Geography, University of Leeds.

Turton, I. and S. Openshaw (1995) Putting the 1991 Census Sample of Anonymised Records on your Unix workstation. *Environment and Planning A* **27**, 391-411.

CHAPTER 12
METADATA FROM THE 2001 UK CENSUS: RECOMMENDATIONS

Paul Williamson

Metadata from the 2001 UK Census: Recommendations

12.1 1991 Census outputs

To the end user one of the most important improvements in the 1991 Census compared to previous Censuses was the greatly increased range, quantity and detail of Census output. Two Samples of Anonymised Records were released for the first time, the various special topic volumes all improved in coverage upon their 1981 counterparts, and additional special topic volumes, not previously compiled, were also released. However, perhaps the single most important change was the increase in the number of statistics available for small geographic areas.

a) increase in cell counts

In 1981 the fifty three published tables of Census Small Area Statistics contained 5,500 counts. For the 1991 Census, the number of SAS tables was increased by just over half to eighty seven, whilst the number of counts trebled to over 15,500. However, whereas for the 1981 Census the SAS were the only source of information at ward level, for 1991 both the SAS and LBS provide ward level data. Taking the LBS into account, there was a greater than ninefold increase in the number of table counts available between Censuses for small areas. As a result, the task of remembering in detail the full wealth of information available from the Census passed beyond the average researcher's reach.

b) increase in table complexity

When designing table layout for the SAS and the LBS, the OPCS had to strike a compromise between table 'legibility' and efficiency of page usage. Ideally, each released cross-tabulation of Census data should be counted as a separate table. However, even for the 1981 SAS this was not possible. Instead, cross-tabulations sharing common elements have had to be concatenated (jointed) into larger, single, tables.

Although the need for table concatenation is understandable, it does make for increased table complexity. This complexity is naturally greater for the 1991 Census SAS and LBS output, given that a greater than ninefold increase in counts was matched by a less than fourfold increase in the number of tables. The layout of SAS Table 46 from the 1991 Census clearly illustrates the increased complexity of table (See Table 1). There are numerous ways in which this table may be split into individual cross-tabulations. Table 1 presents a solution compatible with the table classification approach adopted in the construction of MetaC91 (described below).

12.2 OPCS indexing of 1991 Census outputs

a) published indexes

In order to extract the maximum benefit from the large increase in available small area Census data, a cross-referencing tool is needed, which enables the user to quickly and accurately identify the presence or absence of a given piece of information in the SAS/LBS Census output. A number of possible candidates for this role exist for the 1991 Census. OPCS User Guides 24 and 25 offer, respectively, indexes of table contents in the LBS and SAS. However, in these indexes a maximum of one sub-entry is given for each entry hindering searches for a three or four way cross-tabulation. Additionally, the indexes are not comprehensive. Hence not only is the tracking down of the detailed cross-tabulation of long-term illness by age by sex (SAS Table 12) not directly possible, but even the existence of a cross-tabulation of long-term illness by either of age or sex cannot be verified (see Table 2a). OPCS User Guide 38 offers a more comprehensive index of table contents (an entry for age is present, but not for sex), but again contains only one level of sub-entry and omits any reference to tables containing 10% processed Census counts (see Table 2b). A final index published by the OPCS/GROS is contained in the County/Region Reports (see Table 2c). However, not only is the one-level sub-entry approach retained, but there is still a lack of comprehensiveness. A researcher examining the index in a county/region report could be forgiven for thinking that a cross-tabulation of limiting long-term illness by sex does not exist, when in fact seven tables contain such cross-tabulations (LBS Tables 6, 12, 13, 18, 44, 47 and 75).

The remaining potential published 'index' of Census LBS and SAS table contents is a full set of LBS/SAS table layouts, whether in the form of a Census County/Region Report (OPCS 1992; GROS 1993) or a SASPAC91 User Manual Part 2 (SASPAC 1992). However, thumbing through the 186 LBS/SAS tables and the greater than fifty thousand cross-tabulated cell counts is both a laborious and an error prone method of checking to see if a desired piece of information (cross-tabulation) is available.

What is needed, then, by the general Census user is a database about a database, a meta-database of Census variables and tables, containing information on *what* data are reported in *which* tables from the 1991 Census Local Base and Small Area Statistics and elsewhere.

b) in-house indexes

The OPCS were able to provide information on Census table creation in the form of a glossary of all Census variables, as a series of TAU (programming) statements, or via their Table and Information Monitoring System (TIMS).

The *glossary* contains a record for each unique Census variable, detailing Census variable name, definition and method of derivation direct from Census data or via combination of other derived Census variables. However, although the ideal data source in almost every respect, the Glossary falls down through the lack of any link between variable records and the Census table(s) with which they are associated.

TAU (programming) statements were used by OPCS to generate the derived variables required for Census output. TAU statements contain the same information as the Glossary in a less immediately accessible form, but with the addition in some cases of references to associated Census tables.

The OPCS's *Table Information and Monitoring System* is actually an on-line database which defines tables in terms of their component variables. The combination of information contained in the Glossary

and in TIMS is that which, ideally, Census users would like access to. However, TIMS did not, at the time when MetaC91 was being developed, hold details for all of the LBS and SAS tables.

To summarise, OPCS's in-house machine readable Census data indexes were:

- Piecemeal
- Not up-to-date!
- Not user friendly
- Not readily available to the public

12.3 Features of MetaC91

In response to the need for a more comprehensive index of the 1991 LBS/SAS Census output, the meta-database MetaC91 was created by an academic working at the University of Leeds. MetaC91 boasts a number of features which include:

- Windows (or DOS) based
- Small in size (1.4Mbytes for Windows version)
- User friendly - learn in minutes!
- Fast
- Multiple search types supported
- Cheap (£5 - reproduction costs only)
- FREEly disseminable (i.e. no site licence fees, no forms to fill in...)

12.4 Problems of MetaC91

However, addressing a number of the problems with pre-existing Census indexes outlined above, MetaC91 is not without its own problems:

- Took time to compile (2-3 person months)
- Released at least one year after release of first LBS/SAS data
- Non-standard definition of keyterms/variables
- Only 99.9% accurate
- No graphic representation of Table layouts (neither SASPAC cell numbers nor 'real' tables containing Census counts)
- 1991 Census only (not 1971/81 etc)
- Coverage of LBS/SAS only (no coverage of special topic reports, which came out after MetaC91 produced)
- No coverage of LBS Table 100 (Students) (released after MetaC91 produced)
- Non-updatable

12.5 Recommendations for a 'meta-index' of 2001 Census outputs

Drawing upon the experiences of developing MetaC91, as outlined above, the following conclusions are drawn.

- In order to allow users to extract the maximum benefit from available Census output, a cross-referencing tool is needed, which enables the user to quickly and accurately identify the availability of a required piece of information in SAS, LBS and other Census output.

- Given the inherent limitations of any paper based indexing system, such a cross-referencing tool should be based upon a machine-readable index of 2001 Census output.

- Such a cross-referencing tool could be used as a highly valuable marketing tool for ONS Census products

12.6 Recommendations for a Census output meta-index:

Recommendation I22. *A machine-readable 'meta-index' of <u>all</u> 2001 Census outputs should be regarded as an essential and necessary part of any ONS Census output dissemination strategy.*

Recommendation I23. *ONS should give serious consideration to the creation and maintenance of <u>one</u> up-to-date machine readable Census data index, which incorporates all intended published output, from SAS/LBS through to special topic tables and the SARs, and which is used as the basis of both in-house production of output and public marketing of Census outputs*

Recommendation I24. *The meta-index should be as comprehensive as possible.*

Ideally, any such 'meta-index' should be a straight copy of Table Information Management System (or equivalent) used by ONS in the development of the 2001 Census, thus saving in software development costs. The version of the meta-index released to users (or placed on a Web site) could simply be 'crippled' to ensure that information from certain 'sensitive' fields (e.g. who is in charge of programming the tabulation, completion date due etc.) was made inaccessible/removed.

In addition, the meta-index should include the type of information likely to be included in a 2001 Census 'definitions' volume.

Recommendation I25. *The meta-index should be as freely (cheaply) available as possible.*

Free dissemination of the 2001 Census 'meta-index' is encouraged for two reasons:
a) the meta-index could be viewed as a machine readable marketing catalogue of ONS Census products.
b) free dissemination allows ONS to continually up-date the meta-index as additional Census outputs are produced in response to consumer demand

Recommendation I26. *Recommended 2001 Census 'meta-index' functionality. It should be readily disseminated, be user friendly, support multiple search criteria, contain additional 'meta-information' and be timely.*

These features are now explained in more detail.

a) Readily disseminated

- Platform independent, ensuring that the software will run without adaptation across a wide range of hardware platforms including PCs, Macs and unix systems

The simplest route to platform independence is the creation of a web-based meta-indexing tool which will run on any host machine capable or running a Web-browser.

- Available on a wide range of media (floppy disk; CD-ROM; via the ONS web-site)
- Minimal hardware requirements (to permit dissemination to widest range of potential users)

b) User friendly

- User friendly - learn in minutes! (i.e. point and click based)
- Fast

c) Multiple search criteria supported

Search based upon a combination of one or more of:

- Variable (e.g. age by sex)
- Specific variable aggregations (e.g. single year of age by sex)

The results of searches for specific tabulations should identify only Census output tables containing the specific tabulation required, not Census output tables in which the required variables occur independently of each other, due to concatenation of a number of different tabulations

- spatial scale
- Availability by country in the UK (e.g. Gaelic only in Scotland...)
- Type of Census output (SAS; LBS; special topic volume; SARs...)
- 100 or 10% coded
- Population base of tabulation (i.e. counts for all residents/private households/communal establishments)

d) additional 'meta-information'

- Use of 'official' variable names (e.g. *MarStatt* for marital status, as in 1991 SAR), standard to all Census products
- Range of aggregations of a variable used in Census output (i.e. a listing of all of the possible aggregations of marital status categories used in official Census output -see appended Table 2 for example)
- Hyper-text links to definitions of terms - what, exactly, is a 'dependant' (i.e. machine readable version of currently published *Definitions* volume)
- Captured graphic images of 'real' tables containing actual Census data [e.g. Isle of Wight for SAS/LBS, although obvious problem for Special Topic tables in that need for publication would then be obviated].

Often users have found that apparent ambiguities in table layout are clarified when a table containing 'real' data is consulted.

- Captured graphic images of numbered table cell layout (especially useful for users of 2001 equivalent of SASPAC)

e) Timeliness

- First version of meta-index should be released with or, preferably, before issue of Census data
- Meta-index should be continually updated as special topic volumes and additional LBS/SAS tables are issued, in response to customer demand.

The planned release of a number of versions of the meta-index is to be encouraged, as such a production schedule ensures the maximum flexibility in Census office response to changing customer demand for Census outputs. In this context it should be noted that a web-based meta-index, freely

downloadable from the ONS web-site, has the advantage of being readily updatable, whilst obviating the need for ordering/distribution and minimising costs to users.

OPTIONAL EXTRAS

- Coverage of 1971/81/91 Census data
- Coverage of other government datasets/surveys (e.g. age by tenure also available in the quarterly LFS) - again, a good 'marketing' tool
- Graphical (or other) representation of range of aggregations available for a given variable (e.g. tenure - see appended Figure 1)
- Interface with SASPAC (or equivalent) to obtain specific counts for specific areas

(In which case include an indication of confidence intervals associated with counts at various spatial levels)

- Cost of data order 'calculator' (i.e. highlight desired cells and fill in blanks in pro forma order form 'wizard', including guide through name and number of areas required)
- Front end to Census data tabulation package (i.e. define and extract your own tables - confidentiality constraints permitting)

References

Williamson P., Rees P. and Birkin M. (1994) 'A meta-database of census variables and tables', *ESRC Data Archive bulletin*, 55, S1-S3

Williamson P, Rees P and Birkin M (1995) 'Indexing the census: a by-product of the simulation of whole populations by means of SAS and SAR data', *Environment and Planning A*, 27, 413-424

Table 12.1 The layout of 1991 Census SAS Table 46

Table S46 Households with dependent children; **housing:** Households with dependent children; residents in such households												
		Persons per room		Amenities			Tenure					
	Total House-Holds	Over 1 and up to 1.5 persons per room	Over 1.5. Persons per room	Lack or Share use of bath/ shower and/or inside WC	No central heating	not self-contained accommodation	Owner occupied	Rented privately	Rented from a housing association	Rented from a local authority or new town	No car	Total Persons in Households
Household composition **All households with dependent children** **Households of 1 adult with 1 or more Dependent children** Dependent child(ren) aged 0-4 only Dependent child(ren) aged 5 and over only Dependent child(ren) aged 0-4 and 5 and Over	1	2		3		4	5				6	7
Dependent children in households of 1 adult with 1 or more dependent children All dependent children Dependent children aged 0 – 4 Dependent children aged 5 – 15 Dependent children aged 0 – 17	8	9		10		11	12				13	14
Other households with dependent Children Dependent child(ren) aged 0-4 only Dependent child(ren) aged 5 and over only Dependent child(ren) aged 0-4 and 5 and Over	15	16		17		18	19				20	21
Persons in other households with dependent Children All adults All dependent children Dependent children aged 0 – 4 Dependent children aged 5 – 15 Dependent children aged 0 – 17	22	23		24		25	26				27	28
Households with 3 or more dependent Children Households with 3 or more persons aged 0 – 15	29	30		31		32	33				34	35
Households with 4 or more dependent Children Households with 4 or more persons aged 0 – 15	36	37		38		39	40				41	42

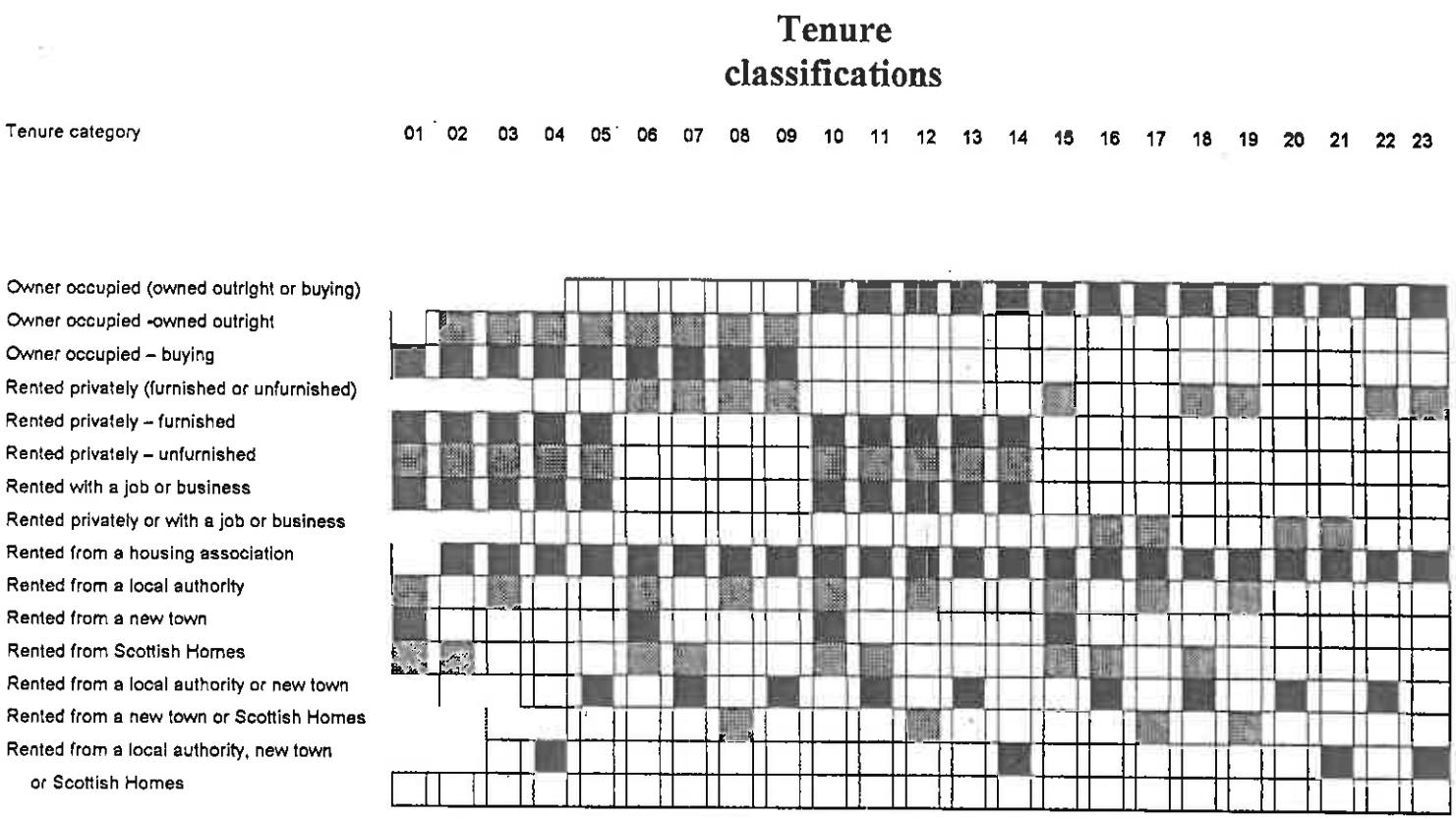
Table 12.2 The indexing of limiting long-term illness in OPCS/GROS publications

Table 3a Index extract from OPCS/GROS User Guide 24		Table 3b Index extract from OPCS/GROS User Guide 38		Table 3c Index extract from OPCS/GROS County/Region Report	
100 %					
(Limiting) long-term illness		(Limiting) long-term illness		(Limiting) long-term illness	
		age	12-14	age	12-14
Dependants (sic)	29	Dependants	29	Dependants	29
by economic position	14	Economic position	14	Economic position	14
by ethnic group	49	ethnic group of head of household	49	ethnic group of head of household	49
by floor level accommodation	68	Floor level of accommodation (lowest) (Scotland only)	68	floor level of accommodation (lowest) (Scotland only)	68
		household composition	44	Household composition	44
		housing characteristics	47	housing characteristics	47
imputed residents	18	Imputed residents	18	imputed residents	18
in commercial establishments (sic)	4.13	In communal establishments	4.13	in communal establishments	4.13
in households	12.44	In households	12	in households	12
		occupancy norm (Scotland only)	70	Occupancy norm (Scotland only)	70
Pensioners	47	Pensioners	47	Pensioners	47
		percentage of households	(I)	Percentage of Households	(I)
		percentage of residents	(E)	Percentage of residents	(E)
10 % Sample					
by hours worked	75			hours worked weekly	75
by SEG of head of household	86			socio-economic group of head of household	86
Source: OPCS/GROS (1992a) 1991 Census User Guide 24: Local Base Statistics Cell Numbering Layouts, p.vii.		Source: OPCS/GROS(1992c) 1991 Census User Guide 38: Local Base Statistics, Small Area Statistics – explanatory notes, p. iii		Source: GROS (1993) 1991 Census: Report for Shetland Islands. Part 1, p. 31 and Part 2, p. 16	

Table 12.3 Aggregations of Marital Status categories used in 1991 LBS/SAS

MarStatt	Marital status			
MarStatt01	*MarStatt02*	*MarStatt03*	*MarStatt04*	*MarStatt05*
Single	Single	Single, widowed or divorced		
Married	Married	Married	Married Cohabiting	Married
Widowed	Widowed or divorced			
Divorced				

Figure 12.1 Tenure Classification



Note: Some tenure classifications exclude the category of rented with a job or business from published counts (e.g. Tenure07)

PART 5: SUMMARY

CHAPTER 13
QUESTIONS AND THE CONTENT OF AREA STATISTICS FOR THE 2001 UK CENSUS: RECOMMENDATIONS

Phil Rees

13.1 BACKGROUND

The Economic and Social Research Council together with the Joint Information Systems Committee (of the Higher Education Funding Councils) have funded the purchase of computer readable data from the 1991 Census for use in academic research and teaching. The data have been supported by two Census Programmes (1991-96 and 1996-2001), which ensure smooth delivery of the data to the individual user, free at the point of use. These Census Programmes have funded four data units which maintain and develop the different data sets and provide help, documentation and training.

If a similar programme of activity is to be funded based on the 2001 Census, then careful justification of census data purchases and support must be assembled. In order to put together the case, ESRC/JISC have funded a set of consultation activities within the academic community, including a series of Workshops in 1997 at which participants have discussed proposals for the 2001 Census (Rees 1997a, 1997b). They have also commissioned a survey of census user opinion which was carried out during August and September 1997. The survey involved mailing sets of questionnaires to all departments in UK Universities carrying out social science research, to all other institutions recognised for receipt of grants by ESRC and to all members of the various census user lists held by the Census Programme Units. The questionnaire was advertised via the *census-uk@mailbase* list and made available on the *http://census.ac.uk/* Web site maintained by the Census Dissemination Unit at Manchester Computing. The questionnaire was distributed in the week beginning the 11th August 1997. Return of completed questionnaires was requested by the 30th September, with the deadline extended by request to the 30th November to cater for census users who had been missed by the first mailouts. In total some 140 census researchers completed a return.

The target population for the survey was users of computer readable census data. Since completion was voluntary, the returns are not a statistically representative sample of the target population. Virtually all respondents (95%) were based in universities. Respondents were drawn from the full range of grades in academic research, with representation from senior academic, junior academic and research track staff. The distribution of respondents by discipline was wide: geography and related disciplines provide 40% of respondents, 22% were from the social disciplines with 13% involved in a health related field. The survey respondents represent active and knowledgeable census users, whose views need to be taken seriously by both ESRC/JISC and the UK Census Offices (the Office for National Statistics, the General Register Office Scotland and the Northern Ireland Statistics and Research Agency).

The main aim of the survey was to inform ESRC/JISC of the views of researchers in supported institutions about census data needs in 2001. The results are being used as input to ESRC/JISC's strategy for census data procurement. The paper replicates, for the 2001 Census, the work for the 1991 Census by the ESRC Working Group on the 1991 Census (Marsh *et al.* 1988). The recommendations also draw on the discussions in a series of consultative workshops that have been held in 1997-8 (Rees 1997a, 1997b). The questionnaire also yielded information of use to the UK Census Offices[1] in their proposals for the 2001 Census, which will be incorporated in a White Paper in the autumn of 1998. The first part of the questionnaire was therefore designed explicitly to gauge the need for particular questions and bundles of questions.

The questionnaire covered four major topics: (1) the topics and questions proposed, (2) proposed changes in concepts for the 2001 Census, (3) proposed changes in outputs for the 2001 Census, (4) the shape of the ESRC/JISC Census Programme for the 2001 Census. This paper focuses on the views expressed about which questions should be asked and about how area statsitics should be produced. Fuller accounts are given in Rees (1998a, 1998b). These two aspects are now considered in turn. Recommendations are made on the basis of survey findings and other consultations. These recommendations are addressed directly to ESRC/JISC and thence to the UK Census Offices.

13.2 TOPICS AND QUESTIONS PROPOSED

Table 13.1 reports user views on household topics proposed for the 2001 Census, while Table 13.2 lists the views on the individual topics. The topics and associated questions were taken from the June 1997 Test carried out by the Census Offices. The Census Test used a questionnaire with 11 household questions and 32 individual questions. The questions had been thoroughly debated over the previous 18 months (1995-97) in the Census Offices' Content Working Group, which included representatives of all census customer sectors. Some questions from the June 1997 Test which were certain to be asked in the 2001 Census such as names of household residents, names of visitors and their usual address, sex, date of birth, and marital status, were not included in the questionnaire on user views.

The entries in Tables 13.1 and 13.2 have been organised by their "approval" rating by respondents. The approval rating is the percentage viewing the question as either "essential" to their research or "highly desirable". Those entries in italics in the table indicate proposals for new questions which were not asked in the 1991 Census or for which significant extensions are being proposed (e.g. to the education questions). If the budget for the 2001 Census is under severe pressure, then the rankings provided in Tables 13.1 and 13.2 show which questions can be omitted by the UK Census Offices with least damage to academic research.

13.2.1 Support for household topics

The most strongly supported topics are tenure, accommodation type and number of cars available, which achieved approval ratings of 80% or more. The other household topics included in the 1997 Test and the 1991 Census scored between 50 and 70%. There was least support for the question on *exclusive use of bath/shower/toilet*, probably because very few dwelling units now lack this feature. Two topics fail to achieve 50% approval, though a majority of respondents still saw these topics as "of interest", "highly desirable" or "essential". There is not great support for the new question on *garden or yard* nor for the extension of the question about *lowest floor level of accommodation*, previously used in Scotland. The rating of this last topic reflects the contextual specificity of some census questions. In Scotland such a question has strong relevance because of the strong presence of tenement and other multi-storey structures in the housing market. Most respondents were from Higher Education Institutions outside Scotland and probably saw the question as irrelevant to the analysis of housing markets with much lower shares of multi-storey housing. If questions are to be dropped from the Census then the garden and floor level are the best candidates, as there is much stronger support for the other household questions and all individual questions.

Respondents made a number of useful comments about the wording and coding of the households questions. There were requests for distinguishing the energy source for central heating, on the size of gardens and for recognition of informal sub-tenancy as a tenure (important for marginalised individuals). Most concern was expressed about the way "rooms" are defined. This issue is very familiar to the Census Offices who have carried out extensive research into the accuracy of census responses. The Census Validation Survey from the 1991 Census (Heady *et al.* 1996) concluded that

the rooms question was the least accurate in the 1991 Census. Experiments with different question forms have been carried out by the Social Survey Division of the Office for National Statistics. The conclusion of this work was that it is extremely difficult to achieve completely standard answers, that the current question form was probably satisfactory and users should be aware of the likely error levels.

Table 13.1 Rating of household topics proposed for the 2001 census

Q.	Topic	N	% of respondents answering question				
			Essential	Highly desirable	Of interest	Very low priority	Not of interest
	70%+ "approval"						
H9	Tenure	121	72	14	9	0	5
H8	Cars or vans available	119	57	26	11	2	4
H1	Type of accommodation	118	51	32	10	3	3
	50-<70% "approval"						
H2	Sharing of accommodation	118	35	31	24	4	6
H4	Rooms	117	40	23	21	8	9
H10	Landlord	119	35	24	28	6	8
H6	Central heating	120	25	26	31	9	9
H11	Furnished/unfurnished accommodation	118	22	29	29	9	11
H3	Exclusive use of bath/shower/toilet	118	24	26	23	13	14
	<50% "approval"						
H7	*Garden or yard (new)*	*119*	*14*	*29*	*31*	*13*	*13*
H5	*Lowest floor level of accommodation (extended)*	*119*	*11*	*22*	*27*	*19*	*22*

Notes:

1. The rows are arranged in descending order of % essential and % highly desirable taken together.
2. N = number of responses = base for percentages (i.e. the row % are computed using this base).
3. New or extended topics are indicated in italics

13.2.2 Individual topics

Table 13.2 provides census user views on the individual topics to be covered in the 2001 Census. Virtually all questions (23 out of 26) win 50% "approval" and 13 questions were seen as "essential" or "highly desirable" by 70% of respondents. When Table 13.2 is examined in detail, the new questions proposed receive very different levels of support. There is overwhelming support for the *income* question: only two other questions were regarded as more essential/highly desirable for respondents' research (*employment status*, *main job)*. There is also enthusiasm for a *general health* question and good support for the *number of paid jobs* and *years since paid job* questions and the two questions on *unpaid help*. There is, however, less support for the question on *number of people employed* which some respondents saw as demanding knowledge that many people would not have. Finally, among the new questions least enthusiasm was expressed for the question on religion. Only 41% saw this question as "essential" or "highly desirable", compared with the equivalent 92% who supported the income question. In addition some respondents expressed the worry that a question on religious group would prove contentious and therefore a threat to census coverage.

Recommendation Q1. *A question on broad banded income should be included in the 2001 Census.*

13.2.3 Comments on question content and wording

Respondents were invited to comment on question wording and coding. Most frequently mentioned were the wording or coding on the ethnic, religion, language, carers and mode of travel questions. First, the question about relationship is discussed.

13.2.3.1 The relationship question

In designing the census user questionnaire, users were not asked to comment on the relationship question on the assumption that the full matrix form of the relationship question would be asked. This question asks each household member to record their relationship to every other member and is straightforward for household members to answer. However, the evaluation of the 1997 Census Test revealed significant errors in completion of the matrix, which need expensive correction in the editing process (Dixie 1998). As a result abbreviated versions are being been tested which are easier to complete and provide most of the relationships between household members. In the Third ESRC/JISC Workshop (Rees 1997b) participants stressed the value of the full matrix question for any research that needed to identify sub-units within the household (families, couples, step relations). A dozen survey respondents directly referred to the need for full relationship information, giving strong support to the full matrix version.

Recommendation Q2. *The UK Census Offices should include the full relationship question in the 2001 Census (using the matrix format).*

Table 13.2 Rating of individual topics proposed for the 2001 census

Q.	Topic	N	% of respondents answering question				
			Essential	Highly desirable	Of interest	Very low priority	Not of interest
	70%+ approval						
I17	Employment status	126	82	14	5	0	0
I21	Main job	127	78	15	5	1	2
I31	*Total gross income (new)*	*127*	*67*	*25*	*4*	*2*	*2*
I8	Ethnic group	126	75	9	8	4	5
I14	*Educational qualifications (extended)*	*125*	*60*	*22*	*14*	*1*	*3*
I19	*Years since paid job (new)*	*126*	*41*	*39*	*15*	*2*	*4*
I24	Hours worked in main job	125	50	24	18	4	5
I7	Country of birth	125	57	17	17	4	6
I15	Usual address one year ago	122	48	25	16	7	5
I12	Long-term limiting illness	126	55	18	18	6	4
I18	*Number of paid jobs (new)*	*124*	*33*	*39*	*21*	*2*	*5*
I5	Student/schoolchild status	124	43	29	19	7	2
I11	*General health (new)*	*126*	*39*	*32*	*18*	*5*	*7*
	50-<70% "approval"						
I29	Mode of travel to work	127	46	24	15	10	6
I22	Main things done in job	125	40	26	26	2	6
I23	Supervision/management responsibilities	124	36	30	24	5	6
I10	*Provision of unpaid help (new)*	*125*	*25*	*34*	*22*	*10*	*9*
I26	What organisation or company makes/does	123	22	36	24	11	8
I25	Organisation or company of main job	124	23	34	27	10	7
I13	*Receipt of unpaid personal help (new)*	123	22	35	24	8	11
I30	Other activities in last week	125	27	30	23	11	9
I28	Address of workplace	125	37	20	18	13	13
I27	*Number of people employed (new)*	122	22	29	29	12	8
	<50% "approval"						
I6	Term-time address	121	28	20	26	13	13
I9	*Religious group (new)*	*122*	*23*	*18*	*34*	*12*	*12*
L1	Language (Irish, Gaelic, Welsh)	125	22	14	26	17	22

Notes:

1. The rows are in descending order of % essential + % highly desirable.
2. N = number of responses = base for percentages (i.e. the row % are computed using this base).
3. New or extended topics are indicated in italics

13.2.3.2 The ethnic question

There was considerable support for the recognition of the Irish as a separate ethnic group. Nine respondents asked that the Irish be recognised as a separate ethnic group. The need to identify the Irish was acknowledged in the 1991 Census but only at the output stage.

Recommendation Q3. *The UK Census Offices should add "Irish" as a category to the ethnic question.*

Note that if Irish were to be added to the list of ethnic categories, then it might be necessary to break down the "White" category into "White British", "White Irish", "White European" and "White Other"

A second area of concern was that in 2001 individuals be given more explicit opportunities to say what their choice of the "Mixed ethnic group" category meant. In the 1991 Census, the write-in responses indicated that people of mixed ethnicity were an important group and an attempt was made to classify the responses. However, few tables were produced on the basis of the more detailed answers because of the uncertainty about the exact meaning of many write-in answers.

Figure 13.1: The ethnic question recommended to ONS Advisory Groups, April 1998

1 **What is your ethnic group?**

◆ First choose one section from a to e.

a **White**

If White, please indicate your cultural background

- ☐ English
- ☐ Irish
- ☐ Scottish
- ☐ Welsh
- ☐ Other European
- ☐ Any other cultural background

b **Mixed**

If Mixed, please indicate your cultural background

- ☐ White British and Black Caribbean
- ☐ White British and Black African
- ☐ White British and Asian
- ☐ Any other cultural background

c **Black or Black British**

If Black or Black British, please indicate your cultural background

- ☐ Caribbean
- ☐ African
- ☐ Any other cultural background

d **Asian or Asian British**

If Asian or Asian British, please indicate your cultural background

- ☐ African-Indian
- ☐ Indian
- ☐ Pakistani
- ☐ Bangladeshi
- ☐ Any other cultural background

e **Any other group?**

Please indicate your cultural background

- ☐ Chinese
- ☐ Japanese
- ☐ Philippino (sic) (Filipino is the correct form)
- ☐ Vietnamese
- ☐ Any other cultural background

Source: Dixie 1998.

Recommendation Q4. *The UK Census Offices modify the ethnic question in the 2001 Census so that respondents identifying themselves as belonging to the "Mixed ethnic group" can more clearly record the nature of their parentage.*

Further tests of alternative forms of the ethnic question have been carried out by the Social Survey Division of the Office for National Statistics and the Questions and Contents Working Group sub-group on the Ethnic Question have commented on the alternatives (Dixie 1998). The recommended form of the question is now in the form shown in Figure 13.1, which incorporates both of the recommendations put forward by academic community respondents.

13.2.3.4 The question on religion

Although this question failed to receive great support from academic census users, its proponents were enthusiastic in putting the case for such a question. One respondent wrote as follows:

> "I think a question about religion should be included throughout the UK for the following reasons: a) There is currently no way of differentiating migrants from Northern Ireland or their descendants in terms of coming from a Protestant or Catholic background. This is regrettable as there is some evidence that they have different experiences settling in Britain. b) The CRE report suggests that there continues to be differentiated socio-economic experiences for Catholics compared with Protestants in Scotland. c) I support the case of some Muslim groups for the inclusion of a question on religion. d) For all these reasons 'Christian' as a category should be sub-divided as it is in the NI Census."

The observation about the vagueness of the term "Christian" in the question form used in England, Scotland and Wales was echoed by several other respondents, and the question used in Northern Ireland which recognises five Christian categories was commended. This point is being considered carefully by the UK Census Offices (Dixie 1998, p.3). The most recently tested question has the following categories: None, Christian, Buddhist, Hindu, Sikh, Islam, Jain, Jewish, Zoroastrian, Any other religious faith. As a result of testing the categories "Jain" and "Zoroastrian" may be dropped and Christian denominations added.

13.2.3.5 Language question

The language question was not widely supported because the 1997 Census Test form was used only outside England. Several respondents commented on the desirability of a mother tongue question, the answers to which would help Local Education Authorities in providing help in learning English and in gauging how much documentation should be produced in other languages. A question on *language spoken most often at home* has is to be tested by ONS Social Survey Division.

13.2.3.6 Mode of travel question

Several respondents were concerned about incorrect answers to the question on mode of travel to work. One respondent identified two problems: that the terminology for the various forms of urban rail transport varies from one part of the country to another, and that journeys to work may get confused with journeys between two residences.

13.2.3.7 Carers' questions

These questions contain the phrase "substantial unpaid personal help". Several respondents were concerned that the word "substantial" was undefined and suggested that some indication of hours of help per week be added to the question or accompanying instructions.

13.2.3.8 The income question

The question that drew most concern about coding or wording was the income question. Practically every respondent who provided comments (57 in total) called for the top category used in the June 1997 Census Test to be sub-divided to capture more detail at the upper end of the income spectrum. The banding used in the Census Test was as follows: (per year) nil, less than £3,000, £3,000 to £5,999, £6,000 to £9,999, £10,000 to £14,999, £15,000 to £24,999 and £25000 or more.

There were lots of different suggestions about the higher income bands that should be used in the 2001 Census. These will clearly need to be adjusted nearer the date to likely income levels in 2001 and might need to be expressed in euros rather than pounds sterling! One respondent made a sensible suggestion for determining income bands in 2001 that would overcome the problem of changing monetary values, any currency change, the bias of other suggestions from what must be a very high income group of respondents and maintain comparability over time.

Recommendation Q5. *The UK Census Offices revise the coding of the proposed question on income to differentiate those with higher incomes more finely.*

13.3 Outputs from the 2001 census

This section of the paper reports on academic census users' reactions to the proposals on census output geography and the form of the area statistics to be released. Views on Special datasets and and the new proposals for more flexible forms of output are described in Rees (1998a, 1998b).

13.3.1 Area statistics from the 2001 Census

13.3.1.1 Use of area statistics from the 1991 Census

The survey questionnaire asked respondents about their use of census data sets. Unsurprisingly, the most heavily used data were the Area Statistics with 69% reporting use of the 1991 SAS, compared with 31% for the Individual SAR, only 12% for the Special Migration Statistics Set 1, 16% for the Special Workplace Statistics Set C, 20% for the Longitudinal Study 1981-91 link, 66% the Digital Boundary Data for GB and 34% the ED/PC Directory. Use of the latter two datasets waswlinked to use of the Small Area Statistics.

Table 13.3 reports in more detail the datasets used by survey respondents. The numbers reported in the table cannot be summed directly to give the percentage of respondents using area statistics as very many respondents used more than one data set. The sum of datasets by scale gives some idea of multi-use: 140 respondents use 261 datasets between them or an average of 1.9 datasets each.

Table 13.3: Numbers and percentages of respondents using 1991 Census area statistics

Area Statistics	N (out of 140)	%
Dataset		
1991 Census Small Area Statistics (SAS)	97	69
1991 Census Local Base Statistics (LBS)	80	57
1991 Census Statistics Derived from SAS or LBS	51	36
Other Area Statistics (please specify)	15	11
Scale		
1991 Census SAS for EDs, OAs	83	59
1991 Census LBS for Wards/Postal Sectors	85	61
1991 Census SAS or LBS, for districts, counties	83	59
1991 Census SAS for Constituencies	10	7
Country		
1991 Census SAS or LBS, England and Wales	83	59
1991 Census SAS or LBS, Scotland	53	38
1991 Census SAS, Northern Ireland	15	11

Notes: EDs = enumeration districts; OAs = Output Areas; N = number of responses

13.3.1.2 Proposals for output areas from the 2001 Census

There has been considerable debate about proposals for *output areas* from the 2001 Census (discussed in the First ESRC/JISC Workshop in January 1997 - see Rees 1997a). This debate arose from the diversity of geography for which needed census data at a fine spatial scale (Rees 1997c). During 1996 the case was debated in the Output Working Group of the UK Census Offices for and against the provision of more than one small area statistics, using different geographical bases and investigated by different researchers (Martin 1997; Rees and Duke-Williams 1997). Eventually, a consensus was reached between census user sectors and Census Office statisticians. The following points have gained considerable support and are put forward as Recommendations.

Recommendation Q6: *There should be a uniform geography across whole UK.*

Recommendation Q7: *Only one standard SAS set should be produced at the smallest scale.*

Recommendation Q8: *Output areas should, in principle, be separated from collection areas.*

Recommendation Q9: *Output areas should be based on aggregations of unit postcodes.*

Recommendation Q10: *Output areas should fit wards current in 2001 and so aggregate to administrative areas.*

Recommendation Q11: *The demand for small area statistics for other geographies should be achieved through look up tables linking output areas to other areas after 2001.*

Recommendation Q12: *Look up tables linking output areas to 1991 Census areas should be provided.*

Recommendation Q13: *Digital boundaries should be output concurrently with the census statistics.*

Recommendation Q14: *The detail of output statistics should be adjusted to geographical scale.*

These recommendations reflect needs for both administrative and postal geographies. In the 1987 survey (Marsh *et al.* 1988) there was overwhelming support for postcode based small area geography but only some of this need was met in the outputs from the 1991 Census. Postal sector SAS and the ED/PC directory were produced in England and Wales, while in Scotland Output Areas were collections of unit postcodes. Output areas in 2001 could be designed to be of uniform size making comparison of statistical parameters more valid. They can be defined from the same GIS base being used for ED planning. They can be designed so that confidentiality population thresholds are all met. Their boundaries and postcode constitutions can be published prior to the production of census statistics. Statistics for larger areas can be produced by aggregating Output Area statistics, on either an exact fit or a best fit basis.

The consultations have ruled out the possibility that more than one set of small area statistics would be produced, because of the perceived risk of disclosure of information about identifiable individuals. Whatever decisions are taken about the form of output areas from the 2001 Census, there are likely to be directories or look up tables linking to various current and historic geographies.

The ESRC/JISC Questionnaire attempted to gauge the degree to which this consensus matched census user needs in one crucial respect: the *nature of the output areas* to be used with the 2001 Census. Respondents were asked to say how important the alternative geographies were for their research and to rank them. The options for output areas are as follows: (1) use enumeration districts defined for collection purposes in 2001, (2) use output areas built up from unit postcodes, and (3) use enumeration districts or output areas from the 1991 Census again. The first alternative would repeat what has

usually occurred at previous censuses. The second alternative emerged from the discussions of the Output Working Group and associated research by Martin (1997), as discussed above. The third alternative preserves temporal comparability and was suggested by Dorling in the First ESRC/JISC Workshop.

Table 13.4 shows that there was little support for using *collection areas from the 2001 Census* as the output geography with 36% seeing this option as essential or highly desirable for their research and only 15% ranking this as the best option (second panel). Opinion was split between the *postcode option* and *1991 Census reporting areas* with 77% regarding postcode based output areas as essential or highly desirable for their research and 69% seeing 1991 Census reporting areas similarly. However, when asked to rank the three output area options (Table 13.5), respondents voted with an absolute majority (58%) for the postcode based option.

Table 13.4: The rating of output area options by respondents

Option	No. of responses =100%	Rating category (%)				
		Essential	Highly desirable	Of interest	Very low priority	Not of interest
2001 Enumeration Districts	88	19	17	33	26	5
Output Areas built from postcodes	99	38	39	15	2	5
1991 EDs/Oas	94	33	36	22	7	1

Table 13.5: The ranking of output area options by respondents

Option	N=100 %	Rank (%)		
		1	2	3
2001 EDs	69	15	30	55
OAs built from postcodes	69	58	25	17
1991 EDs/OAs	69	31	46	24

One of the most important functions of a census is to measure population change. This has always been quite difficult using British censuses because the definitions of administrative, electoral, postal and census collection areas change for very good reasons. Many solutions have been proposed: grid cells (Bracken and Martin 1995, aggregating areas from two censuses to zones with common boundaries (produced from the 1981 Census), freezing an output geography for one census and using it in subsequent censuses (Dorling 1995), use of look up tables that link the geography at one point in time to another (as currently implemented via twice yearly updates of the ED/PC directory by ONS) or designing current output areas to nest into previous units, as 1991 OAs in Scotland nested within the 1981 ED.

Respondents were asked next how useful would the particular solutions to the *time comparability* problem be for their research? The solutions are listed in Table 13.6. The first two are look up tables which link 2001 Output areas to two geographies used in 1991, either grid cells (Bracken and Martin 1995) or enumeration districts (EDs)/output areas (OAs). The third alternative involves the definition of zones which can be formed from both 1991 and 2001 building blocks. This alternative is usually satisfactory in areas of little change but leads to large zones in areas where much change has occurred.

The fourth alternative was to design 2001 OAs to fit into 1991 EDs/OAs. Finally, there was the solution of using the same ED/OA boundaries as in 1991.

Table 13.6: Rating of solutions to the time comparability problem (%)

Proposal	N=100 %	Essential	Highly desirable	Of interest	Very low priority	Not of interest
Option 1: Look up tables linking 2001 Oas and grid cells	81	14	27	41	11	7
Option 2: Look up tables linking 2001 Oas and 1991 EDs/Oas	80	38	41	15	4	3
Option 3: Common zones from 2001 Oas and 1991 EDs/Oas	78	14	36	40	6	4
Option 4: 2001 OAs that aggregate to 1991 EDs/Oas	83	18	42	28	11	1
Option 5: 1991 EDs/OAs as 2001 output areas	78	13	40	22	23	3

Notes:
1. N = number of responses
2. ED = enumeration district (census collection area)
3. OA = output area (census publication area)

There was greatest support for the creation of look up tables that linked 2001 output areas to 1991 enumeration districts and output areas (option 2). Some 79% (Table 13.6) saw this option as essential or highly desirable. Option 4, involving design of output areas that aggregated to 1991 EDs/OAs, had an approval rating of 60%. Option 5, use of 1991 enumeration districts and output areas as the 2001 small area geography, was given an approval rating of 53%. The least preferred option was to provide a link to previous geographies via grid cells (option 1). There were many varied comments on the way in which temporal comparability of the geographic units used to report the results of successive censuses could be achieved. However, all were agreed that achievement of a sound means of comparison before publication of the 2001 Census was vital.

The contents of the 2001 SAS have not yet been discussed in detail, and their content will be determined through consultations in 1998 and 1999. They would take into account new topics, new codings and the 100% processing of census returns in 2001. More detailed tables resembling the 1991 Census Local Base Statistics could be published for larger areas (wards, local government areas). The statistics will be subject to anti-disclosure measures: these would include minimum thresholds for the number of persons and households (as in 1991), but the Census Offices are investigating secondary protection measures (in their "belt and braces" approach), and are interested in user preferences, given the difficulties experienced by users in handling randomly perturbed cell counts in previous censuses.

Table 13.7 lists the options currently being considered and the acceptability ratings assigned by respondents. The only protection device that received much support was the *random adjustment of cell counts* in tables, used with the small area statistics in the past three censuses. Better the devil you know than the one you don't. Some 63% of respondents found this measure acceptable or highly acceptable, whereas the equivalent percentages for rounding of cell numbers, suppression of low cell counts and record swapping were 40%, 41% and 18% respectively (Table 13.7). The latter figure is particularly important as this protection measure has been often suggested by Census Office statisticians as worth experimenting with. However, currently users express hostility to the device, with 57% regarding this measure as unacceptable.

Table 13.7: The acceptability of disclosure protection measures (%)

Proposal	N=100 %	Highly acceptable	Acceptable	Indifferent	Un acceptable	Not of interest
Small random changes to cell counts (e.g. -1, 0, +1)	88	16	47	15	19	3
Rounding of cell numbers (e.g. to base 3 or base 5)	86	7	33	26	28	7
Suppression of cells with low counts	88	8	33	17	39	3
Swapping of a small number of records between areas	88	2	16	21	57	5

Notes: N = number of responses

Recommendation Q15. *The UK Census Offices are seriously urged to accept that use of population size thresholds, imputation in connection with the census and "natural" error are sufficient protection for the confidentiality of census data and therefore to abandon as unnecessary any further protection measures. If such a view is not accepted, then the Census Offices should continue to use small random changes to cell counts, as in previous censuses or record swapping over short distances only.*

13.3.6 Boundary data from the 2001 Census

Boundary data in digital form are provided by the UKBORDERS service at the University of Edinburgh and by the MIDAS service at the University of Manchester. The services collaborate in the dissemination of digital boundary data, in its quality assurance and in the development of user interfaces and the provision of mapping software. The ED boundaries for England and Wales and the OA boundaries for Scotland are held in a large ARC/INFO and INGRES database at Edinburgh from which user extractions via the UKBORDERS interface can be generated at a variety of scales and formats, while at Manchester an extraction system called DBD91 can be used to select boundary sets for England and Wales. The UKBORDERS service has added libraries of pre-selected boundaries commonly. It is hoped that in 2001 that arrangements can be made to add Northern Ireland boundaries.

Some 44% of respondents report use of digital boundary data for England and Wales and 22% use boundaries for Scotland (Table 13.8). These usage levels approach those of the area statistics and are clearly linked with the mapping of counts and derived statistics from the SAS and LBS. The most popular scales (Table 13.9) were wards/postal sectors (41% reporting use), followed by enumeration districts/output areas (36%), districts (32%) and counties/Scottish regions (21%). Out of the 74 respondents who answered the question, a remarkable 87% saw boundary data as essential or highly desirable for their research (Table 13.10).

Table 13.8: Number and percentage of respondents using 1991 Census Boundary data sets

Boundary data	N	%
1991 Census Digital Boundary Data, England and Wales	62	44
1991 Census Digital Boundary Data, Scotland	31	22
1981 Census Digital Boundary Data, Great Britain	25	18
Other Digital boundary data	5	4
OS Digital Maps as background	16	11
OPCS/GROS ED planning maps (microfilm/paper copies)	17	12

Notes: N = number of responses (out of 140)

Table 13.9: The scales at which boundary data were used

Scales used	N	%
Enumeration districts/Output Areas	51	36
Wards/Postal Sectors	58	41
Districts	45	32
Counties/Scottish Regions	30	21
Regions (Standard or Government Office)	19	14
Other zones	8	6

Notes: N = number of responses (out of 140)

Table 13.10: Importance of Boundary Data for research (%)

N=100%	Essential	Highly desirable	Of interest	Very low priority	Not of interest
74	73	14	11	1	1

Notes: N = number of responses (out of 140)

Provision of boundary data in digital form is therefore a vital requirement for the 2001 Census. The boundary data associated with the 1991 Census came to the academic community from a variety of suppliers, both Census Offices and commercial agencies. In the run up to the 2001 Census the situation is likely to be very different with ONS, GROS and NISRA all having their own Census GIS systems (or access to them) for input planning and output area design. It is very likely that the boundaries of output areas will be generated prior to the census and should be released to users well before the area statistics for testing with 1991 Census converted to the new geographies. One uniform agreement between the Census Offices and ESRC/JISC should be possible, provided the Census Offices themselves agree suitable licensing arrangements with the mapping and address agencies (Ordnance Survey Great Britain, Ordnance Survey Northern Ireland and the Royal Mail).

Most of the comments made by respondents referred to the existing databases and support arrangements rather than those for 2001, but virtually all of the points are relevant in preparing for the 2001 Census. There was general appreciation for the boundary data and services delivered with the 1991 Census. However, respondents stressed the case for better display interfaces to the boundary data, and the need for quality assurance of the data in 2001. However, by far the commonest request that library of common coverages in generalised formats (fewer digits) should be provided from the outset. One respondent went further and provided a framework for the systematic production of common coverages.

In fact, the UKBORDERS service is currently busy preparing such standard coverages and has collaborated with the Department of Geography at the University of Edinburgh in developing a method of generalising digital boundary coverages that avoid the topological problems of current algorithms (Mackaness, Edwardes and Urwin 1998).

A number of other issues were raised by respondents, which need to be taken on board in preparing for the 2001 Census. The first concerned provision of paper and microfilm versions of census boundaries on topographic map backgrounds. The 1991-96 ESRC/JISC Census Programme purchased ten copies of the microfilms of the original ED planning maps for England and Wales from ONS and distributed them to ten University libraries for consultation by census users. Paper maps were purchased by

individual user or user departments or university libraries. The quality of reproduction was poor as several respondents recognised.

However, the solution to this problem is not to request better paper or microfilm versions in 2001 because they will probably not exist. Instead the academic community should seek access to the digital versions of background topographic maps which can be merged electronically with the census boundaries to provide the detailed information needed by researchers. Negotiations have been proceeding between Ordnance Survey Great Britain and the Joint Information Systems Committee over the past 18 months and 1998 may bring an agreement to fruition.

One problem with such a solution using digital cartography was flagged up by a vigilant user:

> "The licensing arrangements in 1991 were restrictive over reproduction of the boundaries. It is very likely that the World Wide Web will be the medium of choice for delivery of 'papers' containing maps, especially if the maps need to be in colour. The licensing of the 2001 boundaries should consider the role of the web carefully. The availability of an 'official' generalised set of boundaries might assist in this." [Respondent 51]

Finally, a respondent reminded us that boundaries never stay constant and the ESRC/JISC Programme should have a strategy for keeping boundary data up-to-date.

These responses can be distilled into three main recommendations.

Recommendation Q16. *ESRC/JISC should negotiate the licensing of digital boundary data associated with the 2001 Census with the UK Census Offices and digital topographic data as background to census boundaries with the UK Mapping Agencies.*

Recommendation Q17. *ESRC/JISC should negotiate the licensing of generalised common coverages of digital boundary data as well as the detailed boundary data for the smallest units.*

Recommendation Q18. *ESRC/JISC should negotiate the licensing of updated common coverages of digital boundary data.*

13.3.7 Look up tables for the 2001 Census

Since the 1991 Census was taken, the importance of having good Look Up Tables, which relate one small area geography to another, has become very clear. One of the most used products from the 1991 Census has been the Enumeration District/Postcode Directory which enables users to make estimates of SAS data for postal zones or to aggregate data georeferenced by unit postcode to the census geography. The importance of updating such Look Up Tables has also been realised and the ESRC/JISC Census Programme now purchases an annual update of the ED/PC Directory. Users have also worked hard to develop their own Look Up Tables and these have been deposited with the Census Dissemination Unit at Manchester Computing for general use. Research is currently under way on placing Look Up Tables in a more general relational database framework (Barr *et al.* 1998).

Look up tables relating census small areas to other geographies were widely used by researchers with 47 reporting use of the 1991 Census ED/PC Directory for England and Wales, for example (Table 13.11). Of the 70 respondents replying to the question on importance of look up tables to their own research, 84% saw them as essential or highly desirable (Table 13.12).

Table 13.11: Number and percentage of respondents using the 1991 Census Look Up Tables

Look Up Tables	N	%
1991 Census Area Master File	27	19
1991 Census ED/1991 PC Directory, England and Wales	47	34
1991 Census ED/post-1991 PC Directory, England and Wales	30	21
1991 Census Indexes, Scotland	9	6
1991 Census Ward/Functional Region, Great Britain	18	13
1991 Census Ward/Localities, Great Britain	15	11
1991 Census ED/1981 Census Ward, Great Britain	22	16
Other Look Up Tables	4	3

Notes: N = number of responses (out of 140)

Table 13.12: Importance of Look Up Tables for research (%)

N=100%	Essential	Highly desirable	Of interest	Very low priority	Not of interest
70	46	39	9	7	0

Notes: N = number of responses (out of 140)

The role of look up tables was identified as vital in much research by one respondent. Several respondents were concerned about the need for quality control of look up tables and urged better endeavours in connection with the 2001 Census. Two respondents outlined general schemes of look up tables that would be needed in connection with the 2001 Census:

> "A well organised family of look up tables (or general relational database) will be essential in 2001. It should include: (1) the OA/PC directory updated as PCs change; (2) OA/wards, LGDs updated as wards/LGDs change; (3) 1991 ED/OA to 2001 OA look up table so that change over the 1991-2001 can be studied." [Respondent 27]

> "A connected 'best fitting' effort to create Tables from 2001 OAs to 1991 EDs/Wards is essential - and a vital follow through of this will be to then generate Tables of 2001 OAs to all the other pre 2001 geographies." [Respondent 47]

Cole (1997) has recently reviewed the need for a gold standard for centroids and lookup tables which puts detail on the comments of survey respondents. He argued that there had been no consistent and systematic plan for producing and verifying the look up tables associated with the 1991 Census. Tremendous progress had been made but a great deal of improvement was still needed. For 2001 he suggests a programme of centroid production should be followed (see Cole 1997 for further details) and the following programme of look up table production:

- Look up tables should be updated to reflect changes in administrative and postal geography from 2001 onwards.
- Look up tables should be generated to link 2001 output areas to previous census geographies.
- Look up tables should be generated to link previous census geographies to 2001 output areas.
- Where possible, development of a set of look up tables should be integrated across all countries to facilitate generation of UK files (e.g. UK postcode to higher area index file).
- Integration of look up tables with digitised boundary data products should be planned.
- Boundary changes should be time stamped.
- Head counts for individual postcodes should be supplied.
- Address based referencing products should be developed and maintained (i.e. link individual

addresses to 2001 output areas).
- Immediate access to all the georeferencing tools should be provided under a single licensing agreement.

Recommendation Q19. *ESRC/JISC should negotiate the licensing of a family of look up tables and associated products with the UK Census Offices and Mapping Agencies.*

13.5 Conclusions

The shape of academic opinion about what is wanted from the 2001 Census is clear. They want data that are comprehensive across topics, comparable over time, easy to access, and accurate! They generally appreciate the value of the census data from the 1991 Census produced by the UK Census Offices and delivered to them by the ESRC/JISC Census Programme. They are acutely aware of deficiencies in data sets and have made their views known through the medium of the user survey. However, they also able to put forward solutions to those deficiencies and to contribute to the development of the necessary methodology.

In general, responses to the proposals under development by the Census Offices and the ESRC/JISC Census Programme are very positive. However, census users want as much continuity and comparability with earlier censuses as possible. This raises important issues for the planning of outputs from the 2001 Census. The Census is often caricatured as only being for "geographers", with the observation that other social scientists could make do with a large household survey. However; this stereotype needs extending to include "historians" as well. Census users need comparable social information on people over space and time.

References

Barr R, Harris J, Cole K and Dawes I (1998) Integrating look-up tables for census access: a database approach. Paper presented at the ESRC Census Projects Workshop, 16/17 April, Titchfield.

Bracken I. and Martin D. (1995) Linkage of the 1981 and 1991 UK Censuses using surface modelling concepts. *Environment and Planning A 27, 349-364.*

Cole K. (1997) The need for a gold standard for centroids and lookup tables. Chapter 9 in Rees P. (ed.) (1997a) *The 2001 Census: what geography do we want?* A Report to the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC) by the participants in the First ESRC/JISC Workshop Planning for the 2001 Census, 22-23 January 1997, held at the University of Leeds

Dixie J (1998) Progress on small-scale testing of possible census questions. *Advisory Group paper 98-02*, Office for National Statistics, Titchfield.

Dorling D. (1995) *A new social atlas of Britain.* John Wiley, Chichester, Sussex.

Heady P. *et al.* (1996) *Census validation survey: quality checks*. London: HMSO.

Mackaness W, Edwardes A, Urwin T (1998) Self evaluating generalization algorithms to automatically derive multi scale boundary sets. Paper presented at the ESRC Census Projects Workshop, 16/17 April, Titchfield.

Marsh C., S. Arber, N. Wrigley, D. Rhind and M. Bulmer (1988) Research Policy and Review 23. 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.

Martin D. (1997) From enumeration districts to output areas: experiments in the automated creation of a census output geography. *Population Trends*, 88, 36-42.

Rees P. (ed.) (1997a) *The 2001 Census: what geography do we want?* A Report to the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC) by the participants in the First ESRC/JISC Workshop Planning for the 2001 Census, 22-23 January 1997, held at the University of Leeds

Rees P (ed.) (1997b) Third Workshop: the 2001 Census - Special datasets: what do we want. A report to the ESRC and JISC based on contributions to the Third ESRC/JISC Workshop Planning for the 2001 Census, September 1997, Royal Geographical Society, London. *Working Paper 97/9*, School of Geography, University of Leeds, Leeds, UK.

Rees P. (ed.) (1997c) The debate about the geography of the 2001 Census: collected papers from 1995-6. Report prepared for the First ESRC/JISC Workshop Planning for the 2001 Census, January 1997, *Working Paper 97/1*, School of Geography, University of Leeds.

Rees P (1998a) What do you want from the 2001 census? Results of an ESRC/JISC survey of user viewsPaper presented at the Annual Conference of the Royal Geographical Society (with the Institute of British Geographers), organised by Kingston University and held at Guildford, 5th-8th January 1998. Session (QMRG/PGRG) on Population and Migration Information.

Rees P (1998b) Views of the academic community on the 2001 Census. *Environment and Planning A*, forthcoming.

Rees P. and O. Duke-Williams (1997) Experiments with and recommendations for the creation and release of Small Area Statistics from national censuses. Report prepared for the Statistical Disclosure Control Project, ESPRIT Programme of the European Union. Available from the School of Geography, University of Leeds.

CHAPTER 14
SUMMARY OF DISCUSSION, RECOMMENDATIONS AND USER VIEWS

14.1 Discussion on look up tables and area statistics

The following questions and points were raised in the discussion.

14.1.1 Output areas

Dave Rossiter (Oxford University) asked when output areas might become available, expressing concern that design plans might mean delay.

Bob Barr (University of Manchester) stressed the importance of tying output area boundaries to significant street geography.

Phil Rees (University of Leeds) asked for reassurance that the Census Offices (ONS, GROS and NISRA) and Mapping Agencies (OS, OSNI) were seeing the design of census output areas as a joint endeavour that would be harmonised under the principles agreed in the discussions of the Output Working Group in 1995-97.

Chris Denham (ONS) responded that there was an ongoing review of geographic systems at ONS, which would produce definitive recommendations soon, based on the principles and the work of David Martin. The proposed output areas would be smaller than 1991 EDs on average, and would satisfy the majority of users.

14.1.2 Look up tables

James Brown (University of Southampton) stressed the need for look up tables built from the unit postcode as well as the address.

Danny Dorling (University of Bristol) asked that output areas be constrained to fit inside 1991 wards so that change be measured easily. Failing that he emphasized that there should be an accurate look up table that linked 2001 output areas and 1991 wards.

14.2 Discussion and recommendations on microdata

The following points relevant to the SARs 2001 Working Group were identified in the discussion by Angela Dale (Manchester).

1. It is important to leave enough time for consultation with Scottish users about Scottish geography *(Ken Mackinnon, Black Isle)*.

2. There was much support for adding the ethnic group of the Head of Household to the Individual SAR *(Ceri Peach, Oxford)*. Arguments advanced include the fact that this gives comparability with the SAS and that it provides a proxy for language needs.

3. The geographical detail for the migration and workplace variables needs consideration *(Paul Boyle, Leeds; Tony Champion, Newcastle)*. This should form a second step in deliberations and get detailed

consideration after the size and structure of the files have been agreed. It is important to bear in mind a fall-back position of the addition of an area classification to reflect the type of area from which someone moved.

4. Strong arguments were advanced for a third SAR. For example, *Danny Dorling (Bristol)* raised a basic question: "is it area or class which affects mortality most?" To answer this question 5-year age groups would be adequate but the data is not in the LBS.

There was a clear message from ONS that the third SAR will need strong justification in the following areas: importance of analytic work which goes beyond tabulations and risk of disclosure perceived risk of disclosure.

14.3 Discussion and recommendations on the interaction statistics

There was substantial support in discussion for the recommendations relating to interaction statistics and for the need for a centre of expertise in these special data. One of the key roles of experts in migration and commuting analysis will be to ensure that the right data are requested from the 2001 Census and to guide the development of systems that will enable new users to access these valuable data.

14.4 Discussion on interfaces to census data

There was debate about the merits of a networked solution to census data access, favoured by academic participants and a desktop solution, favoured by local government users, represented by Hywel Davies (London Research Centre). The interactive, intuitive nature of desk top packages running under Windows was very attractive to users. The academic proposals for Web based interfaces were aiming for that capability without losing the advantages of central provision and data management. The Census Offices could see a role for both provision of the data on suitable packet media (CD ROM, DVD) with a suitable, user friendly interface and for the dissemination of data via the Web. If a dataset is fixed and immutable, then packet delivery to the client desk top is probably a sensible strategy. However, Census data sets do evolve overe time for several reasons: (1) many data errors are only found when the data have been fully distributed, (2) new geographical statistics are developed as geographies change. Even in 1998, 7 years after the 1991 Census, the Census Offices and the academic community are producing new SAS datasets for the new local government areas which came into force on 1st April 1998.