



UNIVERSITY OF LEEDS

This is a repository copy of *The Development of Web-Based Interface to Census Interaction Data*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/5024/>

Monograph:

Stillwell, J. and Duke-Williams, O. (2000) *The Development of Web-Based Interface to Census Interaction Data*. Working Paper. School of Geography , University of Leeds.

School of Geography Working Paper 00/04

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

WORKING PAPER 00/04

**The Development of a Web-based Interface to
Census Interaction Data**

John Stillwell and Oliver Duke-Williams
Centre for Computational Geography,
School of Geography,
University of Leeds,
Leeds LS2 9JT
j.stillwell@geog.leeds.ac.uk
o.duke-williams@geog.leeds.ac.uk

Paper presented at the ESRC 2001 Census Development Programme Workshop on
'Project Deliverables and Results', Royal Statistical Society, 1 November 2000

TABLE OF CONTENTS

| | Page |
|---|------|
| CONTENTS | iii |
| LIST OF TABLES | iv |
| LIST OF FIGURES | iv |
| ABSTRACT | v |
| 1. INTRODUCTION | 1 |
| 2. OBJECTIVES AND REQUIREMENTS | 2 |
| 3. INTERACTION DATA SETS | 6 |
| 3.1 1991 Special Migration Statistics | 6 |
| 3.2 1991 Special Workplace Statistics | 10 |
| 3.3 1981 Special Migration Statistics | 11 |
| 3.4 1981 Special Workplace Statistics | 14 |
| 3.5 Studies using the SMS and SWS | 15 |
| 4. SYSTEM ARCHITECTURE | 16 |
| 4.1 Overcoming the Limitations of the Web | 16 |
| 4.2 Dynamic Web Pages | 17 |
| 4.3 Maintaining State | 19 |
| 4.4 Database Uses | 20 |
| 4.5 Overview of system architecture | 21 |
| 4.6 Provision of Advanced Features | 22 |
| 4.7 Use of metadata | 25 |
| 5. USING THE INTERFACE | 44 |
| 5.1 Login and the General User Interface | 44 |
| 5.2 Selection of geographical areas | 45 |
| 5.3 Selection of Variables | 50 |
| 5.4 Extraction and Output | 51 |
| 6. 2001 CENSUS INTERACTION DATA | 59 |
| 7. CONCLUSIONS | 63 |
| REFERENCES | 66 |

LIST OF TABLES

| | Page |
|--|------|
| 1. Fields in 'meta_datasets' and their uses | 27 |
| 2. Sample entries for fields in 'meta_datasets' for 1991 SMS Set 1 | 29 |
| 3. Fields in 'meta_tables' and their uses | 31 |
| 4. Sample entries for fields in 'meta_tables' for Tables 1 and 2 of the 1991 SMS Set 1 | 31 |
| 5. Basic layout of the table defined in Table 4, sample record 1 | 32 |
| 6. Basic layout of the table defined in Table 4, sample record 2 | 33 |
| 7. Fields in 'meta_vars' and their uses | 34 |
| 8. Fields in 'meta_geog' and their uses | 36 |
| 9. Sample entry in 'meta_geog' | 37 |
| 10. Fields in 'meta_geoglut' and their uses | 39 |
| 11. Part of the ddatabase table lut_gbward 1991 | 39 |
| 12. Fields in 'meta_users' and their uses | 41 |
| 13. Fields in 'meta_dataperms' and their uses | 43 |
| 14. Fields in 'meta_usage' and their uses | 43 |

LIST OF FIGURES

| | |
|---|----|
| 1. Examples of using PHP and HTML to write messages | 19 |
| 2. WICID system architecture | 24 |
| 3. Metadata structure | 28 |
| 4. WICID Login Screen | 47 |
| 5. WICID Welcome Screen | 47 |
| 6. WICID General Query Interface | 48 |
| 7. WICID Geographical Area Status | 48 |
| 8. WICID Area Selection Tools | 49 |
| 9. WICID List Selection of Countries | 49 |
| 10. WICID Flow Type Selection | 52 |
| 11. WICID Data Selection Tools | 52 |
| 12. WICID Data Set Selection | 53 |
| 13. WICID Table Selection | 53 |
| 14. WICID Variable Selection | 54 |
| 15. WICID Query Completed | 54 |
| 16. WICID Simple Query | 55 |
| 17. WICID More Complex Query | 55 |
| 18. WICID Output Selection | 57 |
| 19. WICID Labels Selection | 57 |
| 20. WICID Output Preview | 58 |
| 21. WICID Basic Statistics | 58 |

ABSTRACT

This project involves the development of a Web interface to origin-destination statistics from the 1991 Census (in a form that will be compatible with planned 2001 outputs). It provides the user with a set of screen-based tools for setting the parameters governing each data extraction (data set, areas, variables) in the form of a query. Traffic light icons are used to signal what the user has set so far and what remains to be done. There are options to extract different types of flow data and to generate output in different formats. The system can now be used to access the interaction flow data contained in the 1991 Special Migration Statistics Sets 1 and 2 and Special Workplace Statistics Set C. WICID has been demonstrated at the Origin-Destination Statistics Roadshows organised by GRO Scotland and held during May/June 2000 and the Census Offices have expressed interest in using the software in the Census Access Project.

1 Introduction

Successive censuses since 1961 in Great Britain have asked respondents questions about their change of usual residence and their daily commuting behaviour. Much of the data collected on interaction flows in recent censuses have been incorporated into various published census products such as the National and Regional Migration Reports or the Workplace and Transport to Work Reports. It was not until the 1981 Census that the Special Migration Statistics (SMS) and the Special Workplace Statistics (SWS) produced by OPCS were purchased by the ESRC on behalf of the academic community. Equivalent data sets were obtained in 1991 and it is proposed that sets of origin-destination statistics will be produced from the 2001 Census.

The SMS are invaluable for understanding migration patterns and characteristics of migrants between particular places and for defining the characteristics of origin and destination areas. They provide researchers with detailed spatial information about the movement of key population sub-groups. They provide evidence of the volume and composition of migration flows occurring between and within urban and rural areas as well as those streams of migrants entering GB from abroad. They support the construction of sub-national population projections. Likewise, the SWS have immense value to transport planners and modellers as well as those concerned with understanding commuting patterns. The concept of 'self-containment' requires the use of workplace (and sometimes migration) statistics to determine travel to work areas, housing market areas, functional urban areas. Both the SMS and the SWS are required to understand the relationship between migration and commuting.

Despite their perceived importance and potential, the SMS and SWS are census products that have been underused by researchers and planners hitherto. Both data sets are large and complex and their use has been constrained for two reasons in particular: limitations associated with the flows available in the data sets and difficulties associated with data extraction. The overall aim of this Census Development Programme project is associated with the latter problem and thus to increase the use of the special migration and commuting data that will be produced from the 2001 Census.

This paper describes the progress that has been made on the development of an Internet-based software system to access data from previous censuses. The paper contains a summary of the data sets involved and a short review of their use in academic studies (Section 3). An outline of the system architecture and of the structure of the metadata required to develop the information system is presented (Section 4) and the process of building a query to extract data is exemplified using a sequence of Web pages (Section 5). A short summary of the proposed tabulations for 2001 is also included (Section 6). We begin, however, by presenting the specific objectives and the requirements of the system in the next section.

2 Objectives and Requirements

The project has the following objectives:

- to provide users with access to SMS and SWS data via the Internet;
- to develop an easy to use interface for data extraction;
- to generate a library of popular sub-sets of interaction data;
- to provide optional facilities to display, analyse and output the data; and
- to provide access to other census-based interaction data.

The size and complexity of the SMS and SWS origin-destination data sets present a considerable challenge to users wishing to undertake computer-based analyses of migration and commuting patterns. Various software systems have been developed in the past. *MATPAC* was developed to handle the 1981 Census data whereas a choice of software systems has been available for users to extract 1991 SMS and SWS interaction data via MIDAS (subsequently MIMAS). *SMSTAB* is the software developed initially at the University of Leeds by Oliver Duke-Williams and made available via MIDAS for use by academics to access the SMS data. *SMSTAB* was converted into *SWSTAB* to provide access to the SWS Set C. *QUANVERT* is the alternative software platform for providing flexible and rapid access to the SMS and SWS Set C. *SMSTAB* and *SWSTAB* both operate through preparing files containing sets of line commands. *QUANVERT* is a commercial product that has been adapted to do much the same task.

There is no doubt that the availability of information has been transformed by the development of the Internet, connecting 'servers' with 'clients' and thus allowing clients immediate access to the resources held on servers. The World Wide Web has become one of the most efficient channels for transferring information and our project seeks to take advantage of this powerful technology. A 'client-server' application has been developed that allows registered users to extract interaction data online but also permits potential users to browse the system, discover exactly what data sets are available and how data can be extracted. The essence of the project is to provide an interface that is both secure and very user-friendly. This implies that it should be 'interactive' and 'menu-driven' rather than command-driven, and that screens should be easy to understand. Our intention is to provide a range of pathways to the data but

not in such a way as to over-complicate the procedures involved. In fact, a key principle is to minimise the information and parameters that are required in building an extraction query which might then be run either online or in batch mode.

Interaction flow data is unlike stock data in that it is not possible to simply add the total outflows from one area to the total outflows from a neighbouring area to give an aggregate total. Flow data is non-additive because of the need to remove the flows taking place between the two areas that are included in the respective totals. Given this problem, one of the most challenging requirements is to develop a system that allows for the flexible aggregation of origin and destination areas. Such a facility is very useful because it enables users to extract detailed spatial information for small areas in which they may have a particular interest whilst simultaneously generating interaction flows for larger areas that are of less direct importance but may provide valuable contextual or comparative information. For example, a user may be particularly interested in the flows between wards of a major city, but may also wish to generate aggregate flows between each ward and the surrounding districts as well as flows from each ward to other regions throughout GB. Allowing flexibility in the specification of both the origin and destination geographies means that it becomes possible to close the system or to account for the remaining flows. However, it is necessary to incorporate some limitations on the number of geographical scales at which area aggregation is feasible.

We envisage the likelihood that there will be frequent demands for specific national sub-sets of the interaction data, particularly for more aggregate geographies. The inter-regional flows of males, females or persons by age group in 1990-91 are

examples of flow matrices likely to be popular amongst users. The intention is to use the WICID system to create and validate a limited number of popular data sub-sets that would then be available directly from a library. Downloading files would take a minimum of mouse clicks. Access to the queries used to extract the data would also be available to users to check the precise specifications of the data. The library facility would be expanded when appropriate to allow users to deposit new sub-sets of data and the associated queries.

Whilst accurate data extraction is the predominant function of the WICID system, users of migration and commuting data often require raw census data to be manipulated in some way before being interpreted, analysed or visualised. Consequently, one of the objectives of the project is to provide, as a set of options, some data manipulation or analytical functionality. On the one hand, this will involve the provision of a set of tools for ranking and sorting the data and the computation of some simple descriptive statistics (e.g. mean, standard deviation, maximum, minimum, range) for the sub-set of flows extracted. On the other hand, we envisage the need to provide tools to derive new variables based on the data extracted (e.g. net migration or migration efficiency) or requiring other information such as populations (e.g. migration intensities, commuting velocities). The second requirement implies the storage in the database of appropriate populations at risk for use as denominators as well as careful explanation of the ways in which new variables are derived.

Portability and adaptability are other important dimensions. WICID is being designed in such a way that the software will be portable to another server (e.g. irwell at MIMAS) and adaptable to provide access to the origin-destination statistics from the

2001 Census when they become available. Work up to the present time has focused on building a system that provides access to the 1991 Census SMS and SWS. However, a further objective of the project is to provide access to other interaction data sets relating to both the 1991 and 1981 Censuses, details of which are outlined in the next section.

3 Interaction Data Sets

3.1 1991 Special Migration Statistics

The 1991 SMS are a set of statistics relating to the characteristics of individual migrants or migrant households in the 12 month period preceding the 1991 Census. A migrant is defined as a resident who has a different usual address one year before the Census to that at the time of the Census. A wholly moving household, as the name implies, is a household where all the usually resident members aged one and over are migrants. Full details of the organisation and content of the SMS are given in OPCS and GRO(S) (1993a, 1993c) and have been summarised by Flowerdew and Green (1993). Two sets of 1991 SMS were created:

- *SMS Set 1* provides migration flow statistics for *wards* (England and Wales) and postcode sectors (Scotland); and
- *SMS Set 2* provides migration flow statistics for *districts* in Great Britain.

The SMS 1 data set is extremely large, consisting of counts for the 10,933 units involved (8,895 wards in England, 945 wards in Wales and 1,003 postcode sectors in Scotland). In addition, flows from 97 foreign countries or country groups to Great Britain are included, giving a matrix of 11,030 origins by 10,933 destinations - over

120 million potential flows. In fact, the SMS 1 matrices are quite sparsely populated as only a small percentage of the potential number of origin-destination pairs have flows associated with them. For example, only about 1% of the potential number of origin-destination pairs in Great Britain in SMS 1 were non-zero (Cole, 1995). The SMS 1 involve two tables containing 12 counts or variables:

- Table 1: All migrants: age by sex (five broad age groups) (10 counts); and
- Table 2: Wholly moving households and residents in them (2 counts).

The SMS 2 data set consists of counts associated with the migration flows between the districts of Great Britain. There are 459 districts involved (366 in England, 37 in Wales and 56 in Scotland). In addition, flows from foreign countries to districts in Great Britain are included. Associated with each flow is a variable number of counts: the data given in SMS 1, together with more detailed socio-economic details about migrants when a suppression threshold permits. The SMS 2 comprises 12 tables containing 93 counts:

- Table 1: All migrants: age (5 broad age groups) by sex (10 counts);
- Table 2: Wholly moving households and residents in wholly moving households: counts (2 counts);
- Table 3: All migrants: age (5 year groups) by sex (38 counts);
- Table 4: All migrants: marital status by sex (6 counts);
- Table 5: All migrants: ethnic group (4 counts);
- Table 6: All migrants: whether resident in households by whether suffering from limiting long term illness (4 counts);

- Table 7: All migrants aged 16+: economic position (7 counts);
- Table 8: Wholly moving households: tenure (3 counts);
- Table 8S: Wholly moving households: tenure (4 counts);
- Table 9: Wholly moving households: sex and economic position of head (7 counts);
- Table 10: Residents in wholly moving households: sex and economic position of head (7 counts);
- Table 11S: All migrants: Gaelic speakers (1 count); and
- Table 11W: All migrants: Welsh speakers (1 count).

Details of the SMS are outlined in Rees and Duke-Williams (1994, 1995). The SMS data are derived from the 100% sample and include imputed households. Unlike the published Migration Topic Reports (OPCS 1994), no 10% variables (e.g., social class/SEG, industry or occupation) are available. For reasons of privacy and confidentiality, migrant data in SMS 2 are suppressed when there is a flow of less than 10 people between two districts, and details of wholly moving households are suppressed when there are less than 10 households in a given flow. In addition, details are suppressed when there are 10 or more migrants, but all of them members of the same (wholly moving) household. The suppression thresholds used mean that a large amount of the potential data in SMS 2 is not released. Rees and Duke-Williams (1995) have investigated the problem of suppression and have identified the total extent of the problem as involving:

- flows between 135,916 origin-destination pairs of districts suppressed because there are less than 10 migrants;

- flows between 110,268 origin-destination pairs of districts suppressed because there are less than 10 wholly moving households; and
- the flow from Havering in Greater London to Carmarthen in Dyfed which involved more than 10 migrants but all were part of the same (wholly moving) household.

Rees and Duke-Williams (1997) have estimated the data for suppressed flows in Tables 3-10.

Simpson and Middleton (1999) subsequently looked at the impact of the 2% underenumeration in the 1991 Census and identified three further elements 'missing' from the SMS data in addition to the suppressed (unpublished) flows. These are due to:

- unit-response - migrants among 1.2 million unrecorded residents - estimated to be between 218,000 and 376,000 individuals;
- item non-response - residents who were recorded as migrants in the Census but with origin unknown - 326,000 migrants had origins unknown; and
- reporting error - those residents who were recorded in the Census but whose migrant status was wrongly recorded - approximately 500,000 people.

Simpson and Middleton estimate plausible correction for systematic bias introduced by each aspect and apply their adjustments to the flows corrected for suppression by Rees and Duke Williams (1997). This provides a further set of adjusted data relating to Table 3 with 15-19 as a single group. The final adjustment made by Simpson and

Middleton was to add 10% to all flows as an allowance for the net impact of mis-reporting of migrants as measured in the Census quality check. This adjustment is not included in the tables reported in the published paper since the authors recognised that the equal enlargement of every flow might be misleading.

WICID will provide users with access to the adjusted data sets from the respective tables in SMS 2 as well as the raw data in each of the tables in the SMS 1 and 2. An option will be available of adding 10% to the SMS 2 Table 3 counts adjusted by Simpson and Middleton. Permission will be sought from ONS/GRO(S) to provide a sample set of aggregate data, blurred to preserve confidentiality, that will be available unregistered users who want to explore how the information system operates.

3.2 1991 Special Workplace Statistics

The SWS derive from two questions in the 1991 Census about place of usual residence and place of work. There are three sets of tables that comprise the 1991 SWS but only one that involves interaction flow data. SWS Set A provide a set of statistics relating to the employed/self *employed population resident* in each electoral ward of England and Wales and postcode sector in Scotland. SWS Set B provide a set of statistics relating to the employed/self *employed population by workplace* in each electoral ward of England and Wales and postcode sector in Scotland (including people who live in each ward/postcode sector and work at home/do not have a fixed workplace/did not specify a workplace). It is the third set of tables, SWS Set C that provide the data relating to *journey-to-work flows* within and between wards in England and Wales and postcode sectors in Scotland. For each origin-destination pair, there are 9 tables in SWS C containing 274 counts as follows:

- Table 1: Economic position and age: employees and self employed (54 counts);
- Table 2(1): Hours worked (10 counts);
- Table 2(2): Family position (12 counts);
- Table 4: Distance to work: employees and self employed with workplace coded (16 counts);
- Table 5: Transport to work: employees and self employed (22 counts);
- Table 6: Cars available in households: employees and self employed in households (10 counts);
- Table 7: Occupation (sub-major groups): employees and self employed (48 counts);
- Table 8: Social class and socio-economic group: employees and self employed (54 counts); and
- Table 9: Industry divisions: employees and self employed (48 counts).

Since the data were obtained from a 10% sample, problems of adjustment to preserve confidentiality are not encountered in the SWS and users will have access to the raw counts in each of the tables through WICID.

3.3 1981 Special Migration Statistics

As in 1991, those filling in the Census forms in 1981 were asked to give the address one year before the Census, with the postcode if known, for each person who had changed his or her usual address within the previous 12 months. Using the Central Postcode Directory (CPD), all the postcodes of previous residence were linked to

wards and to higher level geographies for publication in printed reports. Ballard and Norris (1983) indicate that between 1 and 2% of addresses were referenced to an incorrect ward.

The 1981 SMS are also divided into two sets. Set 1 is comprised of 148 cell counts for both individual and household-based units for one-year migrants in six tables. The four tables based on 100% counts of individual migrants are as follows:

- Table 1: Migrants usually resident/formerly resident by ward aged 16 and over by economic position and sex (16 counts);
- Table 2: Migrants in private households usually resident/formerly resident by ward aged 16 and over by economic position by household tenure and sex (44 counts);
- Table 3: Migrants usually resident/formerly resident by ward by marital status and sex (8 counts); and
- Table 4: Migrants usually resident/formerly resident by ward by age and sex (36 counts).

The two tables based on wholly moving households or persons in them are as follows:

- Table 5: Wholly moving households within/into/from each ward by economic position by age of head of household (32 counts); and
- Table 6: Persons in wholly moving usually resident/formerly resident by ward by economic position of head of household by age of head of household (32 counts).

Five types of migration data were available from SMS Set 1 in principle: flows within each ward; flows from each ward to each district; flows from each district to each ward; flows from outside Great Britain to each ward; and flows from origin 'not stated' to each ward. This basic structure appeared straightforward but, as Cole and Squires (1987) have indicated, confidentiality constraints imposed a more complex structure on the data set. The critical constraint was that if the number of migrants in any one of these categories fails to reach 25, the flow may be 'thresholded' up into the next highest geographical level. Thus, a flow of 20 migrants within a ward, for example, may become part of a flow from the ward to a district or county remainder. This 'cascading' of flow data presents significant complications when attempting to extract meaningful information from the data set and to examine the pattern and structure of migration flows into and out of districts. For these reasons, Cole and Squires suggested that the SMS Set 1 was a very limited data set for conducting detailed migration research and that the county is the most refined spatial scale at which flows can be considered as reliable. The thresholding problem is the major reason why this data set has received such limited attention.

The 1981 SMS Set 2 were slightly less geographically complex than SMS Set 1 with separate counts for males and females migrating within each ward or between each pair of wards. These flows represent a count of the population and were not subject to data adjustment or thresholding. However, the data set also included flows from origin districts to destination wards where the origin could not be determined to ward level. In a companion CDP project, a method is being developed to retrieve the 1981 SMS 2 data and to re-estimate 1980-81 flows between wards as defined in 1991. The

task is facilitated by the re-tabulation of 1981 migration data for small areas in Scotland in 1991. When the complete national ward-to-ward matrices for males and females have been estimated, the data will be made available through the WICID interface, allowing users to conduct their own analyses of change between 1981 and 1991. It will be possible to re-use the techniques developed for this project when the 2001 origin-destination data sets become available to generate interaction flows for the 1980-81, 1990-91 and 2000-01 for a consistent set of small areas.

3.4 1981 Special Workplace Statistics

Those completing the 1981 Census forms were asked to record the address, with the postcode if known, of the place of work of each person aged 16 or over in a job in the week prior to the Census. A 10% sample of the census forms were taken and workplace postcodes coded and referenced to wards. As in 1991, local workplace statistics for 1981 are in three sets: Set A provides counts of persons aged 16 and over in a job 'last week' according to workplace; Set B provides counts of persons aged 16 and over in a job 'last week' according to usual residence; and Set C provides counts of trips between usual residence and workplace. Since the data was taken from a 10% sample, there was no need to make further adjustments for confidentiality and the data are not subject to thresholding. Details of the SWS are found in OPCS/GRO(S) (1993c, 1993d).

SWS Set C consists of 172 counts for each origin-destination, providing information on the mode of transport to work, social class, socio-economic group, occupation, industrial division and age structure of males and females aged 16 and over in employment. These data were supplied in five 1981 SWS Set C tables and it is the

data in each of these tables that will be re-estimated for 1991 wards in the companion CDP project. Our intention is that, as with the re-estimated SMS Set 1 data, the re-estimated SWS Set C data will be available to users through WICID.

3.5 Studies using the SMS and SWS

Whilst the SMS and SWS have been used by ONS and in local government for various functions, both sets of data remain under-used in both empirical and modelling migration research. One recent example of empirical work based on 1991 SMS is the updating of The Scottish Office's 1992 study of Scottish Rural Life (Williams *et al.*, 1999). The availability of the SMS allowed a much fuller analysis of migration than in the previous report showing that all of the districts in rural Scotland had net in-migration in the year preceding the 1991 Census. The 1991 SMS have been used to analyse migration patterns in Scotland more generally by Forster (1998), highlighting the fact that different migrant sub-groups, especially different life-cycle groups, have different movement patterns. Earlier, Champion (1994) and Champion and Atkins (1996) used the SMS to analyse migration change in Britain as a whole whilst Flowerdew and Boyle (1992) used the 1981 SMS and SWS to compare inter-ward migration and commuting in Hereford and Worcester.

These latter data from the 1981 Census were also used by Flowerdew and Boyle (1995) to calibrate a Poisson regression model for short-distance migration within the same county. Subsequently, these authors used the 1991 SMS Set 1 to fit Poisson models to inter-ward flows, attempting to overcome the under-dispersion problem resulting from modelling zero and very small flows (Boyle *et al.*, 1998). Analysis at this scale provides the potential for more detailed generalization about migration

processes such as suburbanization, counterurbanization, intra-urban mobility, rural depopulation and the relationship between housing and demographic change at the local level. Other examples of the use of the 1991 SMS in modelling work include the application of the Poisson approach to migration flows into the Scottish highlands and islands, from the remainder of Britain, between 1990 and 1991 (Boyle, 1995) and the calibration of spatial interaction models on the entire 1991 SMS Set 1 inter-ward matrix using parallel programming techniques on a high performance computer (Turton and Openshaw, 1998).

The use of SWS for academic research appears to be even less than the SMS. Frost *et al.* (1996, 1997, 1998) have used the SWS to look at the energy consumption implications of changing journeys to work in London, Birmingham and Manchester, whilst Spence and Frost (1995) found that the propensity of residents to work locally has marginally fallen over time, but certainly not at a rate that might be expected given the statistics on average work journey lengths.

4 System Architecture

4.1 Overcoming the limitations of the web

All Web-based applications have to overcome the limitations of the traditional Web model. In order to describe these limitations, and the way in which the WICID system deals with them, it is necessary to understand the fundamental model of data delivery via the Web. A Web server, in the simplest form, waits for a request to arrive from a client (i.e. a user's browser). The request will generally be for a particular file, such as an HTML file (a page of marked-up text) or an image. The server will look for the file in its local filestore, and when found, will send a copy back to the client.

This works well for collections of information which will not change, but the model has significant drawbacks: the system is *simple*, *static* and *stateless*.

The system is *simple* in the sense that the basic interface toolkit provided by HTML – the mark-up language used to write Web pages – is limited. It can be used effectively to display and provide multimedia material, but the elements provided to allow interaction with a user are poor. These elements include buttons and tick boxes, but they are rudimentary when compared to the elements that are typically used when developing a stand-alone program with a graphical interface. The term *static* means that the files are not modified in any way before being sent back to the client. Any serious application will require output to be dynamic, so that it can be modified to suit the needs of each user. For example, a national Web site giving details of films being shown at local cinemas needs to be able to send details about only those cinemas which are local to each user, rather than simply sending a huge list of all films being shown at all cinemas in the country. Finally, *stateless* means that all requests from clients are treated independently by the server, regardless of previous pages viewed. In the case of the cinema example, it is necessary to pass data from one Web page (e.g. a search form) about the user's whereabouts to another Web page (the results page) and thus to *maintain state* within the server.

4.2 Dynamic Web pages

The problem of providing local information about cinemas could be solved simply by providing a set of links for each town or city to a pre-prepared static page listing the current films being shown. However, as applications grow in complexity, it becomes increasingly necessary to overcome the limitations of stasis and statelessness.

Dynamism is often provided by adding a scripting facility to the Web server. This extends the traditional Web model. When the Web server receives a request for a dynamic page, it can determine that in order to satisfy the request it must run a script (or program). This script will have various inputs (data supplied by the user, such as his or her location) and will generate output that can be understood by the client browser (i.e. in the form of an HTML file). Although the Web server handles requests for dynamic content in a different way, the browser receives the results and displays them in exactly the same fashion.

WICID uses the language PHP (<http://www.php.net>) to provide server-side scripting facilities. This is a language which has a C-like syntax and supports object-oriented programming. It contains a wide variety of features designed to ease the implementation of Web-based applications, in particular allowing the programmer to send queries to a database management system (DBMS) and then retrieve the results and act on them. The language can either be embedded within a normal HTML file, or can encapsulate an entire page. Figure 1 shows a simple example of using PHP to write the message 'Hello world', both with an embedded statement (Figure 1a) and with PHP producing the whole page (Figure 1b). In both cases, the file containing the script would need to be given an extension other than '.html', in order to indicate to the Web server that it must be treated in a special manner. Typically, the extension '.php' is used.

| | |
|---|---|
| <pre><HTML> <HEAD><TITLE>Hello</TITLE></ HEAD> <BODY> <P><?php echo "Hello world"?></P> </BODY> </HTML></pre> | <pre><?php echo "<HTML>" ; echo "<HEAD><TITLE>Hello</TITLE>< /HEAD>" ; echo "<BODY>" ; echo "<P>Hello world</P>" ; echo "</BODY>" ; echo "</HTML>" ; ?></pre> |
| <p>(a) A 'Hello world' script using a PHP statement embedded in an HTML file</p> | <p>(b) A 'Hello world' script written entirely in PHP</p> |

Figure 1: Examples of using PHP and HTML to write messages

Both scripts in Figure 1 will produce the same output whenever called, and thus do not exhibit any dynamism. However, complex scripts can contain as much program branching, conditional logic and calls to databases as are desired.

4.3 Maintaining state

A server-side language such as PHP can thus produce dynamic Web pages with relative ease. However, the Web model remains stateless. In order to develop a complex application, it is necessary to pass information between pages giving the current state of whatever variables are required to produce the ultimate required output. In an application such as WICID, large amounts of data must be held. With small applications, it is possible to pass information explicitly between different Web pages (for example, coded within a suffix to the web URL); however, as applications become larger, this becomes impractical.

The solution to this problem is to use a system in which data can be held on the server detailing the state of each session. A session in this case is each individual instance of a user trying to build and execute queries with WICID. In order to do this, it is

necessary to identify each user (determine who they are), authenticate them (make sure that they are who they say they are) and then keep track of the current state of the application from that user's perspective. In WICID, this is done with a code library called PHPLIB (<http://phplib.netuse.de>). This library is written in the PHP language, and contains a series of functions that implement the required session management features. The use of PHP, including features provided by the PHPLIB library, therefore permits the development of a dynamic Web application, in which state can be maintained for the duration of a session. A session begins when a user logs in to the WICID system, and ends either when he/she logs out, or after a certain inactivity time limit has expired.

4.4 Database uses

The general purpose of WICID is to provide access to a large body of data. These data are stored in a database, and the Web interface is intended to help users build a query that will return a subset of data. When the query is ready, it is passed to a DBMS, which processes the query, and sends the results back to the WICID system. The DBMS chosen for WICID is PostgreSQL (<http://www.postgresql.org>). As the name suggests, the database uses SQL as its query language. The choice of an SQL database is important since this means that it is relatively simple to substitute PostgreSQL with any other SQL-compliant database system.

The DBMS is used not only for storing the primary data to which WICID provides access (i.e. the census interaction data), but also for storing and updating two other important bodies of data. Firstly, the DBMS also stores *metadata* describing the interaction data. These metadata include information such as the geographic level at

which the interaction data are coded (e.g. 1991 districts in the case of the 1991 SMS Set 2), and the permissions which are required for a user to be able to access a given data set. The metadata are fully described in Section 4.7. The second additional data set contains those data required to maintain the state of each WICID session. Thus, there is a database table which contains a record for each current session, holding the current values of all session variables. Although the current development version of WICID uses the same DBMS to store a wide variety of data (primary interaction data, supporting metadata and application session data), it is not necessary to use the same DBMS for all these purposes. In fact, there are often advantages to using a different DBMS – perhaps running on different physical machines, in different locations – for different purposes, and thus spreading the computational load more efficiently.

4.5 Overview of system architecture

WICID uses the server-side language PHP, together with the library PHPLIB, to provide dynamic content, in which individual users can generate and execute complex queries. The task of data management and retrieval is carried out by an SQL database management system, PostgreSQL. Figure 2 illustrates this model.

The model is divided into four sections, which from top to bottom show the user's view of the system, parts controlled by the Web server, parts controlled by PHP, and parts controlled by the DBMS. Thus, the user sees only a normal Web server, which returns requested documents, just like any other Web server. At the bottom end of Figure 2 there is an illustration of PHP passing SQL statements to a DBMS, and receiving results in return. The actual way in which the DBMS processes the SQL commands, and the way in which the data are actually stored, is not relevant to PHP.

Of course, the PHP scripts which form the application do far more than simply form SQL statements and send them to a DBMS, and may also interact with other programs – the interaction with a DBMS is shown as one example of the function and scope of PHP within the WICID system.

The Web server portion of Figure 2 shows the server retrieving files from a filestore, and either sending them directly back to the user, or onto PHP for processing. This part of the Figure is a simplified view of the system, in order to illustrate this process. In fact, PHP is effectively a part of the Web server (it is one of many modules compiled into the server) rather than an independent entity.

4.6 Provision of advanced features

The system described above is suitable for generating dynamic content based on interactions with a user. However, it is worth noting that the arrow linking PHP with the Web server in Figure 2 is labelled ‘Output in the form of a ‘plain’ file’. Thus far, we have described a system in which all output which is sent back to the user must be in a form which is compatible that user’s Web browser – in other words, in standard HTML. Such a system is still constrained by the first limitation outlined in Section 4.1, namely simplicity. In order to provide advanced features in WICID at a later stage, such as mapping facilities or other interactive elements, it is necessary to use additional tools such as Java or Javascript. Javascript can be used directly, by simply including it in the output produced by PHP, modifying some template code as appropriate. Javascript is a language which can be embedded within a Web page, and run on the client browser. The use of Javascript may allow the inclusion of

sophisticated interactive selection tools in the interface, or to incorporate some level of validation of submitted data in order to check for errors.

The use of Java would require slightly more sophisticated handling. Java is a powerful programming language and may be used to provide features such as mapping tools. Whilst it is easy to include a reference to a Java applet within a page produced by PHP, and also to provide query-specific parameters to that applet, it will generally be the case that it will be necessary for the applet to send data back to the PHP application. It is this aspect which is likely to present difficulties. It is likely that Java applets will set values in a database table using their own database connectivity functions, and that the PHP application will read value from this database and update its own state whenever necessary.

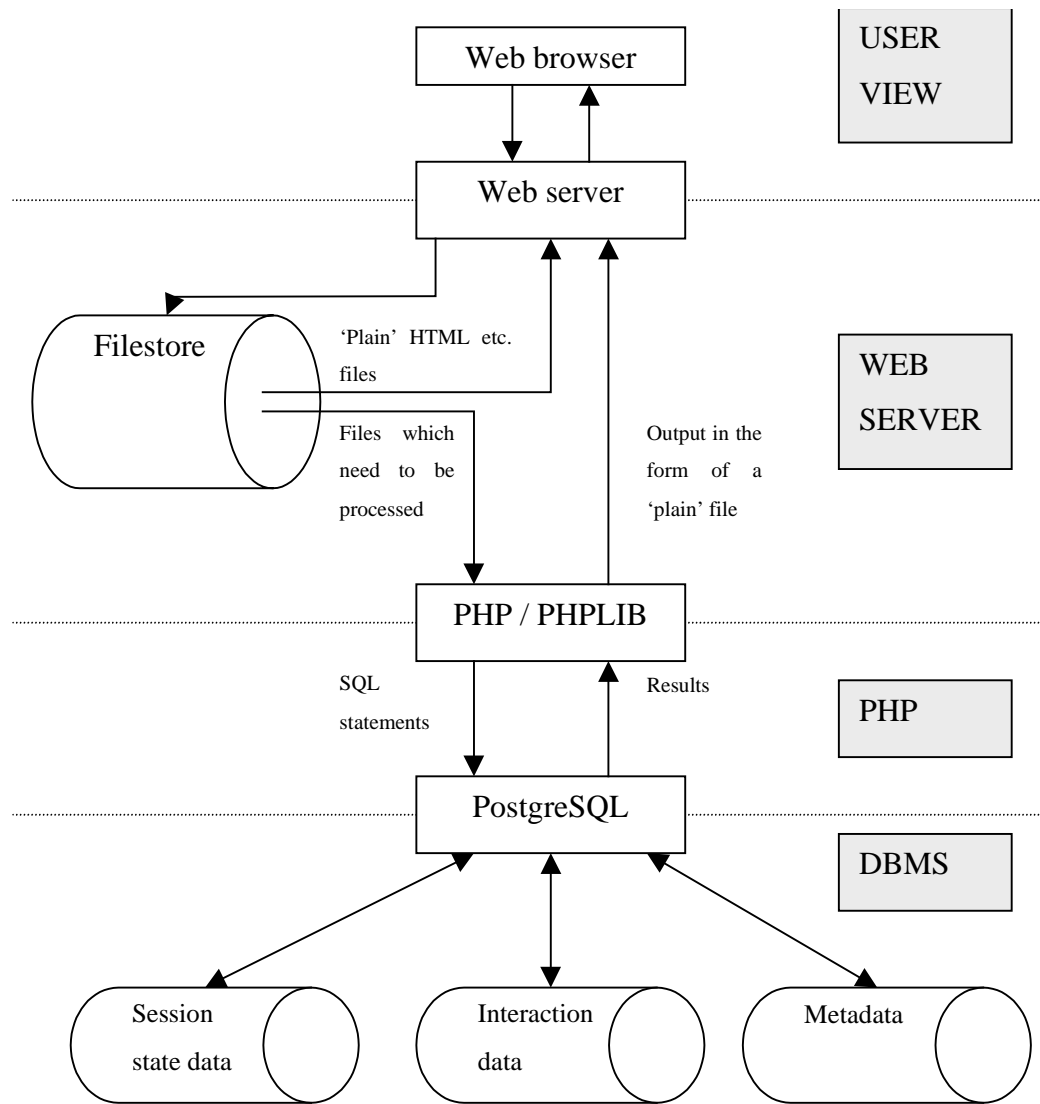


Figure 2: WICID system architecture

4.7 Use of metadata

As well as the main migration and commuting data sets, there is also a body of metadata used by WICID. As indicated in Section 4.4, these metadata include information such as the details of the geographies for which the main data sets are coded. This section of the paper describes the general structure of the metadata system in WICID.

WICID has been designed to have few (or no) hard-coded assumptions about the data that it handles. Instead of pre-determined assumptions, all aspects of the primary data sets are described in the metadata, and are looked-up whenever a dynamic page is produced. The main exceptions to this data-independence come in the form of the Help system, background information, and narrative material included in the various pages, all of which are written for the census interaction data specifically. However, this data-independence should allow WICID to work with any interaction data set that is defined in a manner similar to the Census interaction data sets, including those Census data sets that do not yet exist, such as the 2001 origin-destination statistics. By the phrase ‘defined in a manner similar to the Census interaction data sets’, we mean that the data can be represented as a series of vectors, in which each vector describes a flow between a single origin and a single destination. The vector should have a basic structure that consists of an origin identifier, a destination identifier, and a set of fields giving information about the flow between the origin and the destination. This information will generally be a set of counts disaggregating the flow in some manner, and it is assumed that the data fields can be divided into one or more portions, described as ‘tables’, that disaggregate the flow (or subsets of it) in different ways. At present, WICID assumes a rule that the vectors must be aggregate

observations and must be coded with a unique pairing of origin and destination. However, this rule does not lead to many assumptions within the application code, and could be over-ridden if necessary.

Given the general data structure outlined above, WICID enforces a set of requirements concerning the types of fields used to hold the data, and the names of all fields used in the database. This is done so that the name of any field can be derived internally wherever needed using a fixed set of rules. Data to be loaded into the system must therefore be coded appropriately and all fields given their 'correct' names.

The WICID metadata system (Figure 3) is built around two main tables, 'meta_datasets' and 'meta_tables' and a number of other tables: 'meta_vars', 'meta_geog', 'meta_geogluts', 'meta_users', 'meta_usage' and 'meta_dataperms'. The table 'meta_datasets' includes an entry for each interaction data set held, whilst 'meta_tables' contains a list of all tables (that is, sets of flow dis-aggregations) contained within all data sets. It may seem counter-intuitive that the details of the flow disaggregations contained within each data set are not held in the table 'meta_datasets', which purports to describe each dataset. However, as there are different numbers of tables in each data set, it is not possible to construct a (sensible) database table that is suitable for describing each data set *and* the data contained within it. Table 1 lists the fields in the metadata table 'meta_datasets', whilst Table 2 gives an example of the entries for one dataset – Set 1 of the 1991 SMS. It is worth noting that the metadata in this table do not mention the actual flow disaggregation tables contained within each data set.

Table 1: Fields in ‘meta_datasets’ and their uses

| Field | Use |
|--------------|---|
| datatype | Unique data set identifier – (the primary key) |
| permclass | A permission class for this data set – makes it feasible for different users to be given differential access to data |
| orig_geog | A number identifying the geographic system used to code origins in this data set |
| dest_geog | A number identifying the geographic system used to code destinations in this data set – this can be different to the origin coding if required. |
| rename | The name of the relation (database table) actually holding the data |
| label | A short label to use for this data |
| desc | A longer description of the data |
| notice | A specific copyright notice to use on any output from this data set |

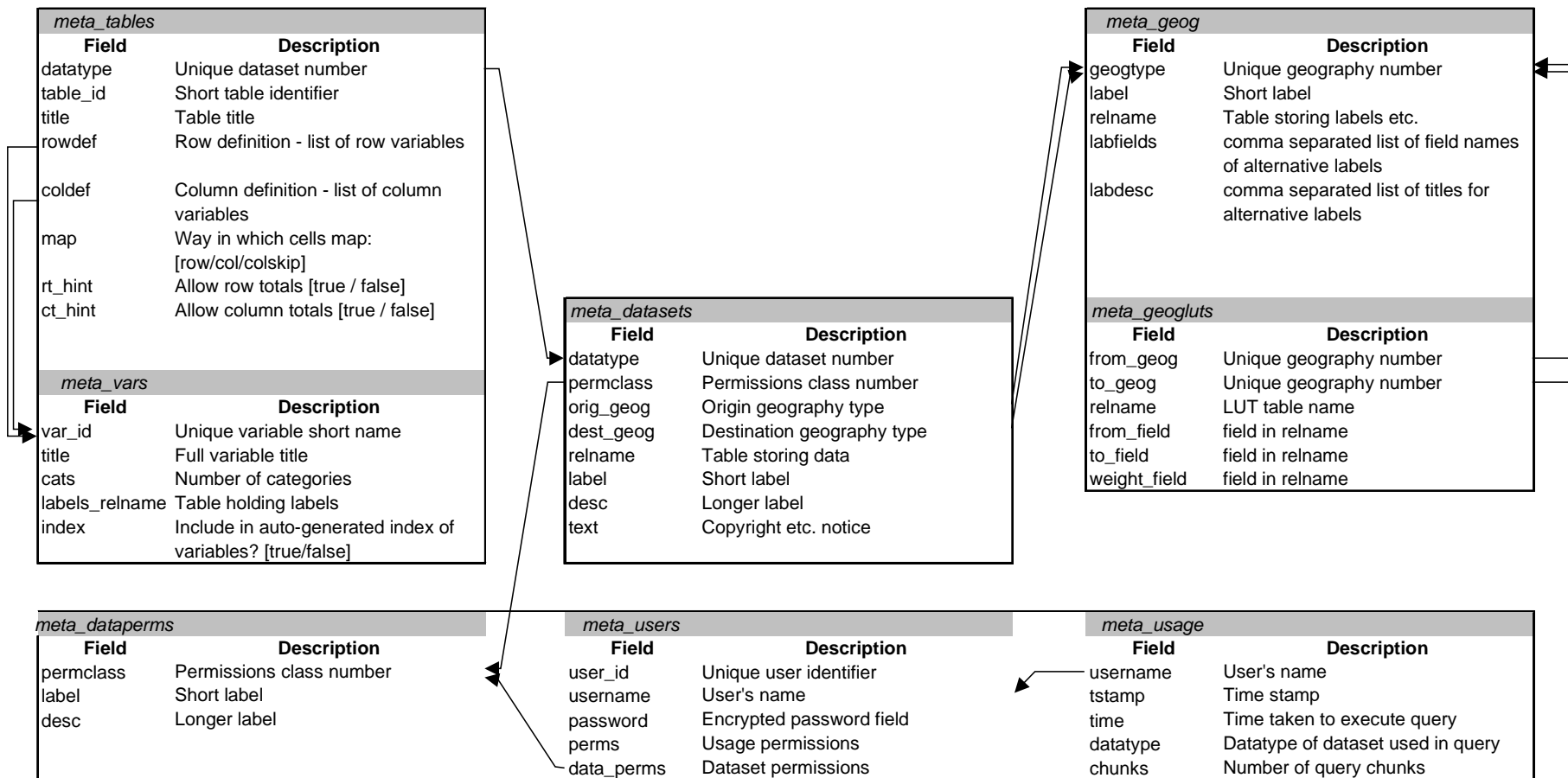


Figure 3: Metadata structure

It can be seen in Table 2 that the first four values are numeric. It should be noted that the first value (datatype) forms the primary key (i.e. a unique identifier) for this table, whilst the next three are all – in SQL terms – foreign keys. This means that they can be used to look-up further information in other tables. The interaction data are being held in a table called ‘data_sms1’, and the ‘label’ and ‘desc’ fields give some basic description of what is in the data set. Finally, the ‘notice’ field is used to hold a copyright notice. It should be noted that this last field is not obeying the strict database rules of ‘normalisation’. It is likely that this exact notice will be used for more than one data set, and thus if it is held in full for each case, we will have data redundancy. The table should really store a foreign key, which is used to look up an entry in a separate table listing the known copyright notice types.

Table 2: Sample entries for fields in ‘meta_datasets’ for 1991 SMS Set 1

| Field | Sample record |
|--------------|--|
| datatype | 2 |
| permclass | 2 |
| orig_geog | 2 |
| dest_geog | 2 |
| relname | data_sms1 |
| label | 1991 SMS Set 1 |
| desc | Ward level migration data from the 1991 Census |
| notice | Source: The 1991 Census, Crown Copyright, ESRC/JISC Purchase |

The structure of the second main table, ‘meta_tables’ is shown in Table 3, with two sample entries – the two tables present in the 1991 SMS, Set 1 – being shown in Table 4. There is one record in ‘meta_tables’ for each flow disaggregation table present in the data sets held in the system. The link between ‘meta_tables’ and ‘meta_datasets’ can be seen in the first line. This field is used as a foreign key, and it must take a value that exists for

the field 'datatype' in the table 'meta_datasets'. In Table 4 both records have a datatype of 2, the same value defined for 1991 SMS Set 1 in 'meta_datasets' in Table 2. The remaining fields describe the flow disaggregation tables. The 'table_id' field must be unique given the value of datatype, and is used to derive field names in the database table that holds the data vectors. This field is generally based on the table numbering used in published documentation, but it must be a valid string to use as part of an SQL field name. The examples shown in Table 4 are simple numeric cases, but it is also legitimate to use letters as well – for example tables 11S and 11W in 1991 SMS Set 2. It is not possible, however, to use punctuation characters such as parentheses.

Table 3: Fields in ‘meta_tables’ and their uses

| Field | Use |
|--------------|---|
| datatype | The datatype of the dataset in which this table can be found |
| table_id | A unique identifier for this table |
| title | The title of this table |
| rowdef | A text string listing the variables which define the rows of this table |
| coldef | A text string listing the variables which define the columns of this table |
| map | A keyword describing the mapping of data for this table in the data vectors of the data set in which this table can be found (see text for description) |
| rt_hint | A ‘true-false’ toggle hinting whether row totals would be sensible for this table |
| ct_hint | A ‘true-false’ toggle hinting whether column totals would be sensible for this table |

Table 4: Sample entries in ‘meta_tables’ for Tables 1 and 2 of the 1991 SMS Set 1

| Field | Sample record 1 | Sample record 2 |
|--------------|---|--|
| datatype | 2 | 2 |
| table_id | 1 | 2 |
| title | All migrants: age (5 broad age groups) by sex | Wholly Moving Households and residents in Wholly Moving Households: counts |
| rowdef | age1 | count1 |
| coldef | sex1 | wmhh1+res_in_wmhh1 |
| map | row | row |
| rt_hint | t | f |
| ct_hint | t | f |

The next three lines define the flow disaggregation table. These lines include a row definition, a column definition and a definition of the order in which data is held in the data vectors. The row and column definitions together specify one or more sub-tables, thus allowing fairly complex concatenated tables, as found in the 1991 SAS and LBS, and to a lesser extent, in the 1991 interaction data sets. Each sub-table is defined as an n -way cross-tabulation, with a minimum value for n of 2. Thus, each sub-table must have a row variable and a column variable, although it is possible to define a dummy variable as one of the dimensions.

The first record has a simple table definition, consisting of a single sub-table component. This sub-table is 2-way crosstabulation using the variable ‘age1’ for rows, and the variable ‘sex1’ for columns. These (and all) variable labels are referenced to an additional metadata table, ‘meta_vars’. This metadata table is described fully below, but it is useful to understand at this point that it stores information about each variable including the number of categories for each variable. In the case of this example, the variable ‘sex1’ has (unsurprisingly) 2 categories, whilst ‘age1’ is a variable giving a broad disaggregation of age with 5 categories. Thus, this record defines a table with five row categories and two column categories, as illustrated in Table 5.

Table 5: Basic layout of the table defined in Table 4, sample record 1.

| | Sex1[1] | Sex1[2] |
|---------|---------|---------|
| Age1[1] | | |
| Age1[2] | | |
| Age1[3] | | |
| Age1[4] | | |
| Age1[5] | | |

This definition is almost complete, but one area for ambiguity remains – the WICID system does not ‘know’ at this stage how the data to fill this table are stored. There are 10 fields which might reasonably be numbered from 1 to 10, with the assumption that the first field corresponds to the combination age1[1] and sex1[1]. However, the ordering of subsequent fields is not obvious. Is the second field the combination ‘age1[1] and sex1[2]’, or the combination ‘age1[2] and sex1[1]’?. The simplest solution would be to impose an ordering pattern which is used consistently. However, a variety of systems of cell ordering already exist – namely those used in published documentation for the 1991 (and other) interaction data sets. It is necessary to preserve these existing field ordering

patterns in order to maintain consistency with the published documentation. For this reason, the field ‘map’ is used in the metadata table ‘meta_tables’. This field uses a keyword to describe the field ordering pattern. There are three alternative keywords (‘row’, ‘col’ and ‘colskip’) covering all patterns in present use. If additional data sets introduce further patterns, a new keyword would need to be introduced, and suitable processing code added to WICID’s table layout and interpretation routines.

The second record in Table 4 defines a more complicated table. It can be seen that the ‘coldef’ field for this record refers to two variables – ‘wmhh1’ and ‘res_in_wmhh1’ – and that these are joined with a ‘+’ character. Again, details about the actual variables are held in the table ‘meta_vars’. This format indicates an overall table which is composed of two component sub-tables. The first sub-table is specified by the row variable ‘count1’ and the first column variable ‘wmhh1’, whilst the second is specified (again) by the row variable ‘count1’, and by the second column variable, ‘res_in_wmhh1’. The row variable is also an example of the use of a dummy variable – ‘count1’, which is simply a frequency count – to allow a univariate table to be described as a 2-way crosstabulation. This is because both column variables have a single category: they are univariate observations. Table 6 shows the basic layout of the table defined in the second record.

Table 6: Basic layout of the table defined in Table 4, sample record 2

| | | |
|-----------|----------|-----------------|
| | wmhh1[1] | res_in_wmhh1[1] |
| count1[1] | | |

Variable names may also be joined using commas in the row and column definition fields, in order to define a nested set of variables. For example, were the ‘rowdef’ field in

a table to be defined as ‘age1,sex1’, then WICID would understand this to mean that rows should use the categories of sex1 iterated within each category of age1.

The final two fields in the ‘meta_tables’ metadata table are ‘rt_hint’ and ‘ct_hint’. These are Boolean fields which can optionally be used to draw an enhanced table. They hint to the system whether or not row and col (respectively) totals would be valid. In the examples in Table 4, these are both set to ‘t’ (i.e. true) for the age by sex table in record 1. This is because the totals of either of these variables (for a given case of the other variable) would be meaningful. On the other hand, both variables are set to ‘f’ in the second record, because neither row nor column totals would be meaningful – column totals would be a total of a single value, and therefore unnecessary, whilst a row total would be a total of two fields measuring different things.

As explained above, flow disaggregation tables are defined in terms of sub-table blocks, which are cross-tabulations of two or more variables. These variables are defined in the metadata table ‘meta_vars’. Table 7 describes the fields in this metadata table.

Table 7: Fields in ‘meta_vars’ and their uses

| Field | Use |
|----------------|--|
| var_id | A unique identifier for this variable |
| title | The title of this variable |
| cats | A count of the number of categories in this variable |
| labels_relname | The name of a database table which holds the labels for each category |
| index | A Boolean toggle indicating whether or not this variable should be included in an index of known variables |

The use of most of these fields should be obvious. The identifiers defined in the ‘var_id’ field are the ones used to identify variables in the ‘rowdef’ and ‘coldef’ fields of the

metadata table 'meta_tables'. There are a number of cases in which some fundamental variable in the data is represented in alternative ways, for example the use of different groupings of age. Alternative classifications are all represented as separate variables in WICID, and thus there are several age variables defined in 'meta_vars' with the identifiers 'age1', 'age2' and so on. It should be noted that the numeric suffix in these names refers to alternative classifications rather than to categories within a generic 'age' variable.

The 'title' field in 'meta_vars' gives a full title for each variable, whilst the 'cats' field shows the number of categories within the variable. The next field listed 'labels_relname', holds the name of a database table which holds all the labels for this variable. It is currently assumed that the database table has a suitable structure – namely that the labels are held in a field called 'label', and that there are a sufficient number of labels. A polished version of the WICID system should provide administrative functions for adding new variables, which only permit a new variable to be registered in 'meta_vars' if the labels database table has been checked and found to be suitable. The final field of 'meta_vars' is called 'index'. This is a Boolean field (true or false) which indicates whether or not a variable should be included in an index. It is intended to provide a list of all variables present in the WICID system, to allow users to search for flow disaggregation tables on the basis of the presence of some preferred variable. However, some variables are used as dummy variables to facilitate table layout. It is probable that some dummy variables will not be meaningful in any general context, and

that it would be confusing to include them in an index of all variables. The ‘index’ field allows such variables to be identified, by setting the value to ‘false’.

The metadata table ‘meta_vars’ holds data about all the variables by which data are classified with the important exception of geographical area. The geography data is held in two metadata tables, ‘meta_geog’ and a supporting table, ‘meta_geoglut’. Table 8 shows the structure of ‘meta_geog’. A record is held in ‘meta_geog’ for each geography that is used in the WICID system. A *geography* is any particular set of areas (e.g. wards, counties, parliamentary constituencies) that are used to subdivide¹ the country.

Table 8: Fields in ‘meta_geog’ and their uses

| Field | Use |
|--------------|---|
| geogtype | Unique numerical identifier for a geography |
| label | The name of that geography |
| rename | The name of a database table holding labels information for that geography |
| labfields | A list of alternative labels to use |
| labdesc | A list of descriptions of the labels listed in the field ‘labfields’ |
| input_only | A Boolean toggle indicating whether or not the geography should only be used to select areas and not for output |

The field ‘geogtype’ is used to uniquely identify each geography. Table 9 lists the entries in ‘meta_geog’ for one particular geography: geogtype 2. This is the geography used to define both origins and destinations in the sample entry from ‘meta_datasets’ shown in Table 2. Each geography is given a full title in the field ‘label’. It can be seen from the example in Table 9 that a date has been included in the label. Administrative geographies are subject to regular change in the UK, and thus it is important to identify the time period (of data collection) to which a particular geography is attached. However, the

relationships between geographies and dates are complex, and they are not currently well modelled in the WICID system. The reference to ‘1991’ in the label field in the example in Table 9 is – as far as the database is concerned – an arbitrary set of characters with no special meaning. It is currently up to the database administrator to make sure that data sets are defined in the metadata table ‘meta_datasets’ as having appropriate base geographies.

Table 9: Sample entry in ‘meta_geog’

| Field | Sample entry |
|--------------|---------------------------|
| geogtype | 2 |
| label | GB wards 1991 |
| relname | geog_gbward1991 |
| labfields | opcs_code, label |
| labdesc | OPCS/ONS code, Place name |
| input_only | f |

The field ‘relname’ in ‘meta_geog’ refers to a database table holding a set of place identifiers for the geography. The present structure assumes that all geographies use a numeric sequence number as their primary identifier – all interaction data sets in WICID have origin and destination coded numerically. Attached to these sequence numbers can be any number of alternative identifiers, such as the OPCS/ONS ids (e.g. 08DAFA) familiar to most existing users of Census data, or place names (e.g. Aireborough). Different geographies will have different valid sets of identifiers, and therefore the fields ‘labfields’ and ‘labdesc’ are used to hold a list of fields within the database table defined by ‘relname’ and a list of matching text descriptions respectively.

¹ In fact, a ‘geography’ does not have to imply a subdivision: (where appropriate) the single categories ‘UK’ and ‘Great Britain’ could equally be used as geographies.

The final field in 'meta_geog' is called 'input_only'. This is not currently used by WICID, but intended to facilitate future use. Some geographies are meaningful to users, but cannot easily be used to output data. The main example of this is postal geography. By setting 'input_only' to 'true', it would be possible to identify geographies which could be used by the user to select a set of areas in another geography. Using the example of postal geography, a user may be able to type in a postcode, and generate output for the set of wards which best approximate that postcode area.

The example of the use of postal geography introduces the topic of translation from one geography to another. The ability to aggregate data to different geographies is one of the key facilities provided by WICID. In order to perform such aggregations, it is necessary to store metadata listing the relationships between different geographies. These metadata are held in the metadata table 'meta_geoglut', the fields of which are described in Table 10.

The fields 'from_geogtype' and 'to_geogtype' hold geogtype identifiers, as listed in the metadata table 'meta_geog'. For any aggregation between two geographies to occur, a suitable record must exist with the two relevant geogtypes. The aggregation is performed using a look-up table (LUT). The LUT to be used is defined in the field 'lut_relname'. Table 11 shows part of the LUT 'lut_gbward1991', which is the database table specified in the field 'lut_relname' in Table 10. This table is used as a LUT linking wards with other geographies, in this case districts and counties. The first five entries are shown, and

it can be seen that the first five wards all form part of both the first district and the first county.

Table 10: Fields in ‘meta_geoglut’ and their uses

| Field | Use |
|---------------|---|
| from_geogtype | The geogtype (as listed in meta_geog) of a geography which can be aggregated to some other geography |
| to_geogtype | The geogtype (as listed in meta_geog) of the geography to which the from_geogtype will be aggregated |
| lut_relname | The name of a database table holding the look-up table which details the mapping needed for the aggregation |
| from_field | The name of a field in the table mentioned in ‘lut_relname’ which lists identifiers of the geography specified by the ‘from_geogtype’ |
| to_field | The name of a field in the table mentioned in ‘lut_relname’ which lists identifiers of the geography specified by the ‘to_geogtype’ |
| weight_field | The name of a field in the table mentioned in ‘lut_relname’ gives a weight for each individual <i>from-to-to</i> mapping |

Table 11: Part of the database table lut_gbward1991

| ward | district | district_weight | county | county_weight |
|------|----------|-----------------|--------|---------------|
| 1 | 1 | 100 | 1 | 100 |
| 2 | 1 | 100 | 1 | 100 |
| 3 | 1 | 100 | 1 | 100 |
| 4 | 1 | 100 | 1 | 100 |
| 5 | 1 | 100 | 1 | 100 |
| ... | ... | ... | ... | ... |

Whereas in other cases WICID makes some assumptions about the presence of various fields with particular names, in the case of the LUTs, the field names to be used are defined using the entries in ‘from_field’ and ‘to_field’. This is because a single database table may hold the LUTs for several aggregations, and thus fewer assumptions can be made about field names. In practice, it is likely that all aggregations from a particular geography will be held in the same database table. The final field used in ‘meta_geoglut’ is ‘weight_field’. This is used to define a third column in the LUT which holds a variable

giving a weight for each entry in the LUT. In Table 11, all weights are shown as being 100. Thus, for example, the first line in the table indicates that 100% of the count for ward 1 should be added to the total for the first district (along with 100% of the count for wards 2, 3, 4, 5 and so on). At present all aggregations use a consistent hierarchy of areas, such that all weights in all LUTs are set to 100. As all the weights are equal they can effectively be ignored, and these may be termed ‘unweighted LUTs’. The WICID system has been written with the expectation that at some stage it will be necessary to use mappings which cannot be achieved with unweighted LUTs. Instead, some fraction of the count for a ‘from’ area will be added to one ‘to’ area, with the remainder being added to a second ‘to’ area. An example of this would be the use of a post-1998 local authority geography, in which some wards have been split up and the components added to different new areas. If we wish to tabulate data that has 1991 base geographies at 1998 LA level, then we would need to use a weighted look-up table. This process is not ideal, and it would be preferable to use an alternative data set which has been modelled and re-estimated for the 1998 geography, but weighted LUTs allow a tabulation to be performed where no such data set exists.

There are three remaining tables in the metadata system, ‘meta_dataperms’, ‘meta_users’ and ‘meta_usage’. These are used to hold information about users, and to link users to specific data sets. Table 12 shows the fields used in the table ‘meta_users’. This table has an entry for all users of the WICID system. There are two fields used to identify the user – ‘user_id’ and ‘username’. The second of these is the userid which is used in combination with a password when the user logs in to the system. The former identifier

is used internally, and is not revealed to the user. This internal id has certain properties: it is of a fixed length, and it is generated when a new user is created in the system using a seeded cryptographic function (md5) which should make the the value for any user difficult to predict. This ‘secret’ identifier is used as part of the WICID security system.

Table 12: Fields in ‘meta_users’ and their uses

| Field | Use |
|--------------|-------------------------------|
| user_id | Unique user id |
| username | User’s name |
| password | Password field |
| perms | Interface permissions of user |
| data_perms | Data permissions of user |

The next field shown in Table 12 is the ‘password’ field. This holds an encrypted form of the user’s password. The final two fields are ‘perms’ and ‘data_perms’. These list the permissions that each user has within the WICID system. The ‘perms’ field allows administrative users to be identified. These users can then be permitted to carry out functions such as creating new general users. The ‘data_perms’ field lists the classes of data to which each user has access. Whilst it is hoped that in future access can be granted as a whole to all Census data sets, the present situation is not ideal. Different Census data sets has separate registration processes, and the WICID system therefore recognises the fact that different users may have different levels of permission to use datasets. The metadata table ‘meta_dataperms’ lists all known permission classes. This table is illustrated in Table 13. The table structure is fairly simple, in that it consists only of a unique identifying field, and two descriptive fields. The data permissions are referred to as ‘classes’, because it is often the case that more than one data set is covered by a particular registration procedure. Thus there may be one permission class referring to all

1991 interaction data sets, and a second one referring to all 1981 data sets. It can be seen in Table 2 that all data sets are identified in the metadata table 'meta_datasets' as belonging to a particular permission class.

Table 13: Fields in ‘meta_dataperms’ and their uses

| Field | Use |
|--------------|--|
| permclass | Unique identifier for permission class |
| label | Short label of permission type |
| desc | Longer description of permission type |

The final metadata table is ‘meta_usage’. This table is used to log all data extractions performed by WICID. The table fields are shown in Table 14. Each WICID query is split up internally into requests for data from different data sets, and then a series of 1 or more individual SQL queries are generated that extract the required data from that dataset. These individual queries are referred to as ‘chunks’. Whenever a data extraction occurs, the details of it are recorded in ‘meta_usage’. The field ‘username’ records the name of the user making the extraction, whilst the field ‘tstamp’ logs the precise date and time that the extraction was made. Within WICID, the time taken to carry out the extraction for a given data set (in one or more SQL chunk) is recorded. This value is logged in the field ‘time’. The final two fields record the dataset from which the data was extracted, and the number of chunks into which the query was split. The latter information is kept to assist the process of monitoring performance of the query building process.

Table 14: Fields in ‘meta_usage’ and their uses

| Field | Sample entry |
|--------------|---|
| username | User’s name |
| tstamp | Time and date that this entry was logged |
| time | Time taken in seconds to execute query |
| datatype | Datatype of dataset used in query |
| chunks | Number of chunks into which the query was split |

5 Using the Interface

In this section of the paper, the current version of the WICID interface is illustrated using a selection of screens or Webpages. Some of the features that appear on particular pages have not yet been implemented.

5.1 Login and the general user interface

The initial page to which potential users of WICID will be directed is the Login page (Figure 4) where users registered to extract data from the SMS and SWS will input their userids and passwords and where unregistered users will be able to type in a 'guest' userid and password to browse the system and access the sample of 'blurred' data. Once logged into WICID, the first page to appear will be a welcome page (Figure 5), with links to the general query interface, the library of popular data sets (and their associated queries) and pages of information about the interface and the data.

There are three lines of information at the top of this page, as with all the pages that follow. The first line provides a monitor of the current page being used together with a summary of the path taken to arrive at the page. There is also a Help link available in the top right hand corner of the page. The second line performs three functions: there is a reminder of the userid being used; a customise; and a Logout now link that allows the user to quit the system at any time. The customise link will allow a user to set or modify settings such as a preferred email address to use when emailing results files or colour schemes to use in generated graphs. The third line provides five links, three of which take the user to other pages within WICID, such as the WICID project page, the general query

interface ([WICID query](#)) and the [Current query](#). It is envisaged that this line will also contain a link to a page containing other census Web sites and a link that allows either a current query to be saved or an existing query to be restored.

The general query interface (Figure 6) is where the user can identify whether the query has been completed or not and whether data extraction can go ahead. Each data extraction query requires the selection of geographical areas ([Geography](#)) and tables or flows ([Data](#)). Both parts of the query must be prepared before any data can be extracted. The two parts may be completed in any order, can be revisited and adapted if required. Traffic light icons are used to indicate whether each section 'Needs attention' or is 'Ready'. When each part has been prepared satisfactorily, the traffic light icon will be green and when both traffic lights are green, it becomes possible to proceed to the extraction phase.

5.2 Selection of geographical areas

Origin and destination areas must be selected in order to submit a query. The current selection status of the geographical areas is shown on the Geography page (Figure 7) together with the number of types of flow selected. These will all be zero in the first instance. Links are available to [Select origins](#), [Select destinations](#) or [Select flow types](#). In many cases, users will require data for the same sets of origin and destination areas; consequently, a facility is available to set the destinations equal to the origins selected. However, origins and destinations sets may also be different and WICID allows the user to select combinations of origin or destination areas from four different spatial levels: wards, districts, counties and regions. Figure 8 illustrates the area selection page that

contains a number of optional area selection methods. Quick selection allows all origins at one of the four spatial scales to be selected. List selection allows the user to select origins from a list at a specified level (Figure 9) and is likely to be useful when the user does not know the sequence numbers, OPCS codes or names of areas for which data is required. Finally, there is the type in box selection option which requires either the area sequence number, its code or name to be input and is seen as being a useful option when users require data for only a very limited set of origins or destinations or for experienced users who are familiar with area codes. In addition to the links to the different types of area selection pages, the area selection page itself also contains facilities to unselect all the areas previously selected, and to show or edit the existing area lists.



Figure 4: WICID login screen

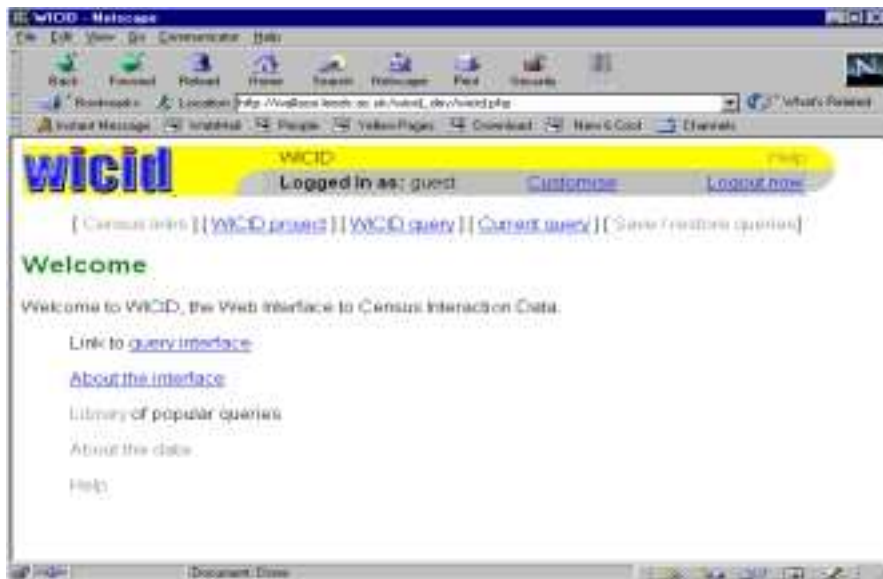
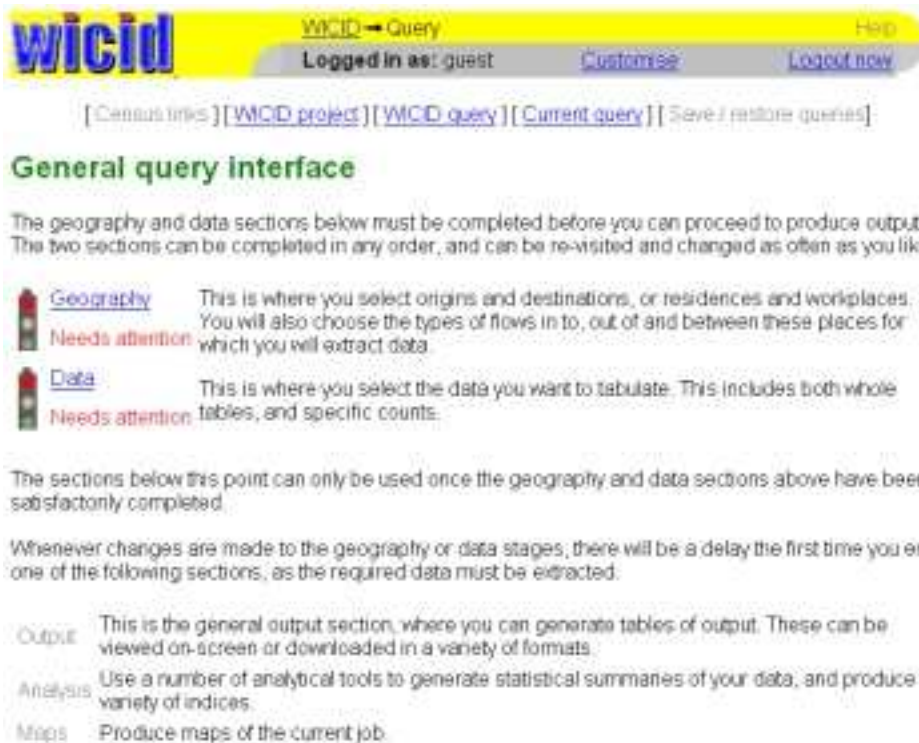


Figure 5: WICID welcome screen



wicid WICID - Query Help
 Logged in as: guest Customise Logout now

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

General query interface

The geography and data sections below must be completed before you can proceed to produce output. The two sections can be completed in any order, and can be re-visited and changed as often as you like.

Geography *Needs attention* This is where you select origins and destinations, or residences and workplaces. You will also choose the types of flows in to, out of and between these places for which you will extract data.

Data *Needs attention* This is where you select the data you want to tabulate. This includes both whole tables, and specific counts.

The sections below this point can only be used once the geography and data sections above have been satisfactorily completed.

Whenever changes are made to the geography or data stages, there will be a delay the first time you enter one of the following sections, as the required data must be extracted:

- Output** This is the general output section, where you can generate tables of output. These can be viewed on-screen or downloaded in a variety of formats.
- Analysis** Use a number of analytical tools to generate statistical summaries of your data, and produce a variety of indices.
- Maps** Produce maps of the current job.

Figure 6: WICID general query interface



wicid WICID - Query - Geography Help
 Logged in as: guest Customise Logout now

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

Geography

Mirrored origins and destinations

Origins and Destinations

- 0 origins currently selected. [Select origins](#) [Edit list](#)
- 0 destinations currently selected. [Select destinations](#) [Edit list](#)

Flow types

- 0 types currently selected. [Select flow types](#)

What now?

You need to select some origins and destinations in order to submit a query, you also need to identify the flow types in which you are interested.

If the option *Mirrored origins and destinations* is selected, then any selections made for origins will automatically be replicated as selected destinations, and vice versa. Note that this will overwrite any existing selections.

Go back to general [query interface](#) section.

Figure 7: WICID: geographical area status



Figure 8: WICID area selection tools



Figure 9: WICID: list selection of counties

Once the origins and destination areas or area lists have been selected, the remaining piece of information required to complete the Geography section is the choice of the type of flow required. Figure 10 illustrates the flow type selection page. Here, the user is confronted with a schematic table of possible flow types, and tick boxes with which to select flows between origins and destinations and within common areas, total outflows from all origins, total inflows to all destinations and/or total migration flows taking place in the system. WICID computes the number of counts associated with each of these options and indicates these values in the right hand column on the page. Once options have been selected, the values are highlighted. There is also a filter available to allow the option of selecting all flows, both between and within areas to fine tune the previous selections. An update button beneath the table will allow the flow types identified on the page to be selected.

5.3 Selection of variables

The Data section of WICID enables the user to select data counts in one of three different ways (Figure 11). Quick selection provides a shortcut to allow the user to select some aggregate data (all migrants or commuters in 1990-91). The table selection option allows the user to select counts from a table as defined in the published documentation whereas the variable selection option allows selection from a list of relevant variables from all tables. The table selection option is currently operational and Figure 12 indicates the page where the required data set is chosen. Clicking on one of the data set options takes you to page containing the appropriate list of tables (Figure 13) and clicking on the required table presents a layout of the table with tick boxes for each of the counts. Figure 14

illustrates the table layout for SMS Set 2 Table 1: All migrants: age (5 broad age groups and sex).

5.4 Extraction and output

When the Geography and Data sections of the query have both been completed, the user is guided back to the general query interface to check that both traffic lights are on green (Figure 15). At this point, it would be wise for users to check the full specification of their current query using the link on the third line from the top of the page. Figure 16 illustrates a relatively simple query in which the user has specified the same 8 metropolitan counties as origins and destinations and has requested migration flows between these areas from the sample data set. Figure 17 is a more complex query in which 13 origins (9 districts of South and West Yorkshire and 4 counties of Humberside, North Yorkshire, South Yorkshire and West Yorkshire) and 35 destinations (2 counties and 33 boroughs of Greater London) are selected. The sample migration data is selected but the flow types requested in this case include total outflows, total inflows and total migrants overall, as well as inter-area flows.

There are a number of ways in which the flows between origins and destinations can be tabulated. This page allows you to select which flowtypes you want WICID to produce.

The following table shows some of the possible flow relationships, and allows you to select the ones that you wish to include in your output. Note that the figure is for illustration only - it does not necessarily reflect the numbers of origins and destinations that you have selected, or the layout of your output.

You have selected 8 origins and 6 destinations. Of these, there are 0 areas which are in both the origin and destination groups.

| | Destinations | | | | Click button to include | Flow description | Potential number of flows to be produced (including any zero flows) |
|---------|----------------|--------------------------|--------------------------|--------------------------|--------------------------|--|---|
| | D ₁ | D ₂ | D ₃ | Totals | | | |
| Origins | O ₁ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | Flows between origins and destinations and within common areas | 48 |
| | O ₂ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | Total outflows from origin areas | 3 |
| | O ₃ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | Total inflows to destination areas | 6 |
| | Totals | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | Total of all flows in system | 1 |

Additional filters
 Mij flows:

Figure 10: WICID flow type selection

The screenshot shows the WICID web interface. At the top, there is a navigation bar with 'wicid' logo, 'WICID → Query → Data', and 'Help'. Below this, it says 'Logged in as: guest' with links for 'Customise' and 'Logout now'. A breadcrumb trail shows: '(Census trics) | (WICID project) | (WICID query) | (Current query) | (Save / restore queries)'. A status bar indicates '0 data items currently selected' with links for 'Unselect all' and 'Edit list'. The main content area is titled 'Quick selections' and lists two options: 'All migrants (1991 SMS)' and 'All employees and self-employed (1991 SMS)'. Below this is the 'Selection tools' section, which includes 'Select by table' and 'Select by variable' with brief descriptions. At the bottom, there is an 'Additional variables' section with a link to 'Select other data' and a description of how to use additional variables for places.

Figure 11: WICID data selection tools

wicid WICID → Query → Data → Select by table [Select dataset] [Help](#)
 Logged in as: guest [Customise](#) [Logout now](#)

[[Census links](#)] [[WICID project](#)] [[WICID query](#)] [[Current query](#)] [[Save / restore queries](#)]

0 data items currently selected [[Unselect all](#)] [[Edit list](#)]

Select a dataset from the list below. You will be shown a list of tables from that dataset

The datasets highlighted with the symbol: ➔ are the ones to which you currently have access. You can examine the table structures in other datasets, but will not be able to pick data from them.

- ➔ [Sample data](#) Sample dataset - blurred count of inter-district migrants from 1991 Census
- [1991 SMS Set 1](#) Ward level migration data from the 1991 Census
- [1991 SMS Set 2](#) District level migration data from the 1991 Census
- [1991 SWS Set C](#) Ward level journey-to-work data from the 1991 Census

What now?

You must select some data before you can produce any output!

- Select a data set using the above menu
- OR go back to the general [data selection](#) page
- OR go back to the general [query](#) interface

Figure 12: WICID data set selection

wicid WICID → Query → Data → Select by table [Select table] [Help](#)
 Logged in as: guest [Customise](#) [Logout now](#)

[[Census links](#)] [[WICID project](#)] [[WICID query](#)] [[Current query](#)] [[Save / restore queries](#)]

0 data items currently selected [[Unselect all](#)] [[Edit list](#)]

- Table 1 [All migrants: age \(5 broad age groups\) by sex](#)
- Table 2 [Wholly Moving Households and residents in Wholly Moving Households: counts](#)
- Table 3 [All migrants: age \(5 year groups\) by sex](#)
- Table 4 [All migrants: marital status by sex](#)
- Table 5 [All migrants: ethnic group](#)
- Table 6 [All migrants: whether resident in households by whether suffering from limiting long term illness](#)
- Table 7 [All migrants aged 16+: economic position](#)
- Table 8 [Wholly Moving Households: tenure](#)
- Table 8S [Wholly Moving Households: tenure](#)
- Table 9 [Wholly Moving Households: sex and economic position of head](#)
- Table 10 [Residents in Wholly Moving Households: sex and economic position of head](#)
- Table 11S [All migrants: Gaelic speakers](#)
- Table 11W [All migrants: Welsh speakers](#)

What now?

You must select some data before you can produce any output!

- Select one of the above tables
- OR go back to the [dataset selection](#) list to choose a different data set
- OR go back to the general [data selection](#) page
- OR go back to the general [query](#) interface

Figure 13: WICID table selection

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

0 data items currently selected [Unselect all] [Edit list]

Table 1 All migrants: age (5 broad age groups) by sex

| | Male | Female |
|--------------|------|--------|
| 1-15 | 1 | 2 |
| 16-29 | 3 | 4 |
| 30-44 | 5 | 6 |
| 44-Pers. age | 7 | 8 |
| Pers. age+ | 9 | 10 |

What now?

You must select some data before you can produce any output!



- Select some values from the table structure
- OR go back to the [table selection](#) list to choose a different table
- OR go back to the [dataset selection](#) list to choose a different data set
- OR go back to the [general data selection](#) page
- OR go back to the general [main](#) interface

Figure 14: WICID variable selection

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

General query interface

The geography and data sections below must be completed before you can proceed to produce output. The two sections can be completed in any order, and can be re-visited and changed as often as you like.

-  [Geography](#) This is where you select origins and destinations, or residences and workplaces. You will also choose the types of flows in to, out of and between these places for which you will extract data.
-  [Data](#) This is where you select the data you want to tabulate. This includes both whole tables, and specific counts.

The sections below this point can only be used once the geography and data sections above have been satisfactorily completed.

Whenever changes are made to the geography or data stages, there will be a delay the first time you enter one of the following sections, as the required data must be extracted.

- [Output](#) This is the general output section, where you can generate tables of output. These can be viewed on-screen or downloaded in a variety of formats.
- [Analysis](#) Use a number of analytical tools to generate statistical summaries of your data, and produce a variety of indices.

Figure 15: WICID query completed

wicid WICID → Query → Status Help
 Logged in as: guest Customise Logout now

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

Summary of current query

| | |
|---------------------|--|
| Origins | 8 GB counties 1991: (Sequence number,Place name) 1 Inner London -8,West Yorkshire |
| Destinations | 8 GB counties 1991: (Sequence number,Place name) 1 Inner London -8,West Yorkshire |
| Flow types | 1 M ₁ |
| Variables | 1 Sample data tff |

Go back to [previous page](#)

Figure 16: WICID simple query

wicid WICID → Query → Status Help
 Logged in as: guest Customise Logout now

[Census links] [WICID project] [WICID query] [Current query] [Save / restore queries]

Summary of current query

| | |
|---------------------|---|
| Origins | 13 GB districts 1991: (Sequence number,OPCS/CNS code,Place name) 49,05CC,Barnsley -52,05CG,Sheffield -66,06CX,Bradford -69,06CB,Wakefield GB counties 1991: (Sequence number,Place name) 6,6outh Yorkshire -8,West Yorkshire -20,Humber-side -37,North Yorkshire |
| Destinations | 35 GB counties 1991: (Sequence number,Place name) 1 Inner London -3,Outer London GB districts 1991: (Sequence number,OPCS/CNS code,Place name) 1,01AA,City Of London -33,02BK,Waltham Forest |
| Flow types | 4 M ₁ M ₂ M ₁₀ M ₁₂ |
| Variables | 1 Sample data tff |

Go back to [previous page](#)

Figure 17: WICID more complex query

Clicking on the [Output](#) link on the query interface page takes the user to the output page which allows a variety of alternative layout and labelling options to be selected (Figure 18). Output can be generated in matrix form as an origin-destination pair list. Published table frameworks and bespoke layout options allowing users to design their own output will be implemented in due course. Format options allow output to be produced as an HTML table or in a comma separated (.csv) file. Delivery may be to the screen, to a file for downloading or as a file (generated from a batch run) for sending via email. A separate label selection page (Figure 19) allows the user to specify using tick boxes what form of origin and destination labels are required on the output. Once these output selections have been made, output may be previewed (Figure 20) before being finally produced.

One other option that follows data extraction and has been implemented only in rudimentary form at this stage is the facility to undertake some analysis on the data. Figure 21 indicates that the current interface allows a set of basic statistics to be produced for the flows selected in the query. These include the minimum and maximum values, the mean, the median and the standard deviation of the data selected. Further work is required to develop this section of WICID and to offer users a range of options for computing derived variables, for generating new area-specific or system-wide indicators and for fitting spatial interaction models.

Query output - plan

This page allows you to alter the characteristics of the output.

| Layout options | General options |
|---|---|
| <p>Output layout</p> <p><input checked="" type="radio"/> Origin - destination matrix [Advanced settings...]</p> <p><input type="radio"/> Origin - destination pair list [Advanced settings...]</p> <p><input type="radio"/> Published table frameworks [Advanced settings...]</p> <p><input type="radio"/> Bespoke layout [Design...]</p> <p>Output format</p> <p><input checked="" type="radio"/> HTML table [Advanced settings...]</p> <p><input type="radio"/> Comma separated values [Advanced settings...]</p> <p>Delivery method</p> <p><input checked="" type="radio"/> On screen</p> <p><input type="radio"/> File download</p> <p><input type="radio"/> Via email</p> | <p>Origin and destination labels</p> <p>Current labels to use in output [Change]</p> <ul style="list-style-type: none"> • GB districts 1991 Sequence number • GB counties 1991 Sequence number |

Figure 18: WICID output selection

wicid [WICID](#) → [Query](#) → [Output](#) [Help](#)
 Logged in as: guest [Customise](#) [Logout now](#)

[[Census links](#)] [[WICID project](#)] [[WICID query](#)] [[Current query](#)] [[Save / restore queries](#)]

Labels selection

This page allows you to choose the labels that you want to use for origins and destinations, and the order in which they will appear.

GB districts 1991

| Label | Label position | | | Not used |
|-----------------|----------------------------------|-----------------------|-----------------------|----------------------------------|
| | 1 | 2 | 3 | |
| Sequence number | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> |
| OPCS/ONS code | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Place name | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

GB counties 1991

| Label | Label position | | Not used |
|-----------------|----------------------------------|-----------------------|----------------------------------|
| | 1 | 2 | |
| Sequence number | <input type="radio"/> | <input type="radio"/> | <input checked="" type="radio"/> |
| Place name | <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figure 19: WICID labels selection

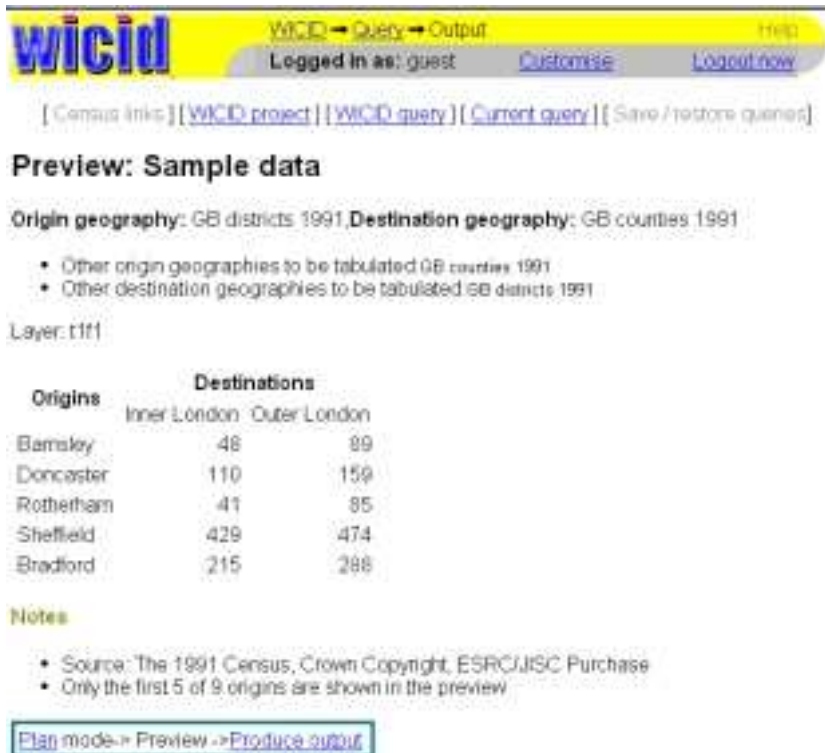


Figure 20: WICID output preview

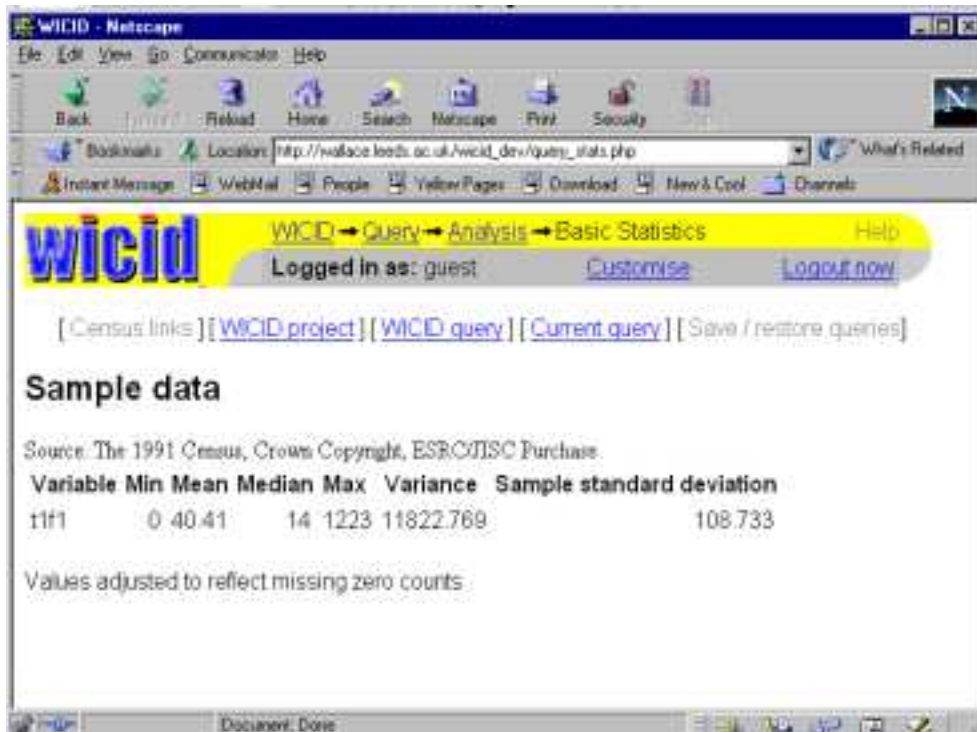


Figure 21: WICID basic statistics

6 2001 Census Interaction Data

WICID is being developed with a view to providing an interface to the origin-destination statistics produced from the 2001 Census. The Census Offices have proposed that the standard products will be SMS and the SWS for England and Wales plus Special Commuting Statistics (SCS) for Scotland - based on journeys to places of work and study. These data sets would contain the 'flow' elements equivalent to the 1991 SMS and SWS. It is proposed that some or all of the output for migrants by zone of origin and workers/commuters by zone of destination would be moved into the standard tables. This means that counts of out-migrants as well as in-migrants will be available for different zones. In preparation for the 2001 Census, the Census Offices have consulted on a wide range of issues relating to the origin-destination statistics and their proposed tabulations (ONS/GROS/NISRA, 1999; 2000). It is proposed that origin-destination products for 2001 will be based on data processed for 100% of census records and it is also assumed that a set of postcode-based Output Areas will be created for the whole of the UK for *Area Statistics*.

In order to ensure consistency with the rest of the range of census output, the SMS and SWS/SCS will be produced from the same output database. The initial data capture process involves completed census forms being scanned and the images converted to 'data' by a combination of recognition software and manual intervention. Valid postcodes will be captured on scanning; invalid or partial postcodes will be imputed using a (ONS/GROS/NISRA, 1999). Other adjustments will be made that result from the *One*

Number Census (ONC) process which aims to add records to the database for missed households and for missed individuals in otherwise enumerated households. As in previous censuses, the Census Offices intend to make a number of amendments to the data so as to prevent users of the data from making any deduction about a household or individual record. The method of statistical disclosure control has not yet been decided.

SMS Set 1 will include the flows between local authority areas, into local authority areas from outside the UK, into local authority areas when no usual residence one year ago was given or where origin is unknown. Throughout the tables, the term ‘sex’ used in 1991 is replaced by the term ‘gender’. The set of proposed tables for SMS 1 are likely to be as follows:

- Table 1: Migrants by age and gender (40 variables);
- Table 2: Migrants in households by family status and gender (18 variables);
- Table 3: Migrants by ethnic group and gender (14 variables);
- Table 4: Migrants with or without limiting long term illness by gender (8 variables);
- Table 5: Migrants aged 16 to 75 by economic position and gender (20 variables);
- Table 6: Moving groups of migrants in households (8 variables);
- Table 7: Moving groups of migrants in households by tenure (16 variables);
- Table 8: Moving groups of migrants in households by gender and economic position of head of group (80 variables);
- Table 9: Moving groups of migrants in households by NS-SEC of head of group (96 variables) ; and

- Table 10: Migrants resident in Scotland/Wales/Ni who know Gaelic/Welsh/Irish family status and gender (18 variables).

SMS Set 2 in 2001 are due to contain the flows between wards or postcode sectors in Scotland. Four tables comprising Set 2 are proposed as follows:

- Table 1: Migrants by age group and gender (22 variables);
- Table 2: Moving groups of migrants in households (4 variables);
- Table 3: Migrants by ethnic group and gender (4 variables); and
- Table 4: Moving groups of migrants in households by NS-SEC of head of group (24 variables).

As in 1991, it is proposed that there will be three sets of journey to work/study tables but unlike 1991, the three sets of tables will include interaction flows. SWS/SCS 1 will include flows between local authorities; SWS/SCS 2 will include flows between wards or postal sectors and SWS/SCS 3 will include flows between output areas. The proposed tables for SWS 1 are as follows:

- Table 1: Workers by broad age group and gender (10 variables);
- Table 2: Workers by family status and gender (18 variables);
- Table 3: Workers by family status and hours worked (18 variables);
- Table 4: Workers by method of travel to work (11 variables);
- Table 5: Workers by NS-SEC and gender (24 variables);

- Table 6: Workers by NS-SEC of head of group and gender (32 variables); and
- Table 7: Workers by ethnic group (7 variables).

Proposed tables for SWS 2 are as follows:

- Table 1: Workers by broad age group (5 variables);
- Table 2: Workers by gender (2 variables);
- Table 3: Workers by family status (7 variables);
- Table 4: Workers by hours worked and gender (6 variables);
- Table 5: Workers by method of travel to work/study (11 variables); and
- Table 6: Workers by NS-SEC (12 variables)

There is only one table proposed for SWS 3:

- Table 1: Workers by method of travel to work/study (11 variables).

The tables proposed for the SCS 1 are as follows:

- Table 1: Students and schoolchildren by age and gender (10 variables);
- Table 2: Students and schoolchildren by family status and gender (18 variables);
and
- Table 3: Students and schoolchildren by ethnic group (7 variables).

For SCS 2, there are three tables proposed as follows:

- Table 1: Students and schoolchildren by age group (5 variables);
- Table 2: Students and schoolchildren by gender (2 variables); and
- Table 3: Students and schoolchildren by family status (7 variables).

The proposed table for SCS 3 is as follows:

- Table 1: Students and schoolchildren by method of travel to study.

It is envisaged that WICID will be able to provide the interface to these data in 2003 although some additional work will be necessary to construct the relevant pages and modify the software.

7 Conclusions

This paper has reviewed the data sets that will be accessed by WICID, has explained in detail the relationships between the software components of the information system and has illustrated the framework of metadata that underpins the system. The paper has also illustrated how the user will interact with the interface. Some progress has been made on the development of the WICID interface since the previous workshop. The project has now reached a point where access to the following data sets is a reality:

- 1991 SMS 1 Tables 1 and 2 (12 variables);

- 1991 SMS 2 Tables 1-12 (93 variables); and
- 1991 SWS C Tables 1-9 (274 variables).

By the end of the project, we intend to provide access to the following data sets:

- 1991 SMS 2 Tables 3-10 adjusted for suppression (89 variables);
- 1991 SMS 2 Table 3 adjusted for suppression, underenumeration and mis-reporting (38 variables);
- 1981 SMS 2 Table 1 re-estimated for 1991 ward boundaries (2 variables); and
- 1981 SWS C Tables 1-5 re-estimated for 1991 ward boundaries (variables to be determined).

Some data verification has been undertaken using the raw interaction flows extracted from the 1991 SMS. This has involved running equivalent queries with SMSTAB and with WICID and checking for consistency. In future, we intend to check unsuppressed flows for aggregate geographical areas (e.g regions) extracted from WICID with those available in the published 1991 Census reports.

One of the major developments of the interface since the last workshop has been the programming of a mechanism to allow the flexible selection of origin and destination areas and the associated organisation of metadata. Now that this has been solved, future efforts can be focused on implementing some of the following:

- ❑ the library of popular data sets and their associated queries;
- ❑ the save and restore queries facility;
- ❑ the select by variable facility
- ❑ the analysis section;
- ❑ the mapping section;
- ❑ the Help system; and
- ❑ the links to other census web sites.

References

- Ballard, B. and Norris, P. (1983) User needs – an overview, Chapter 3 in Rhind, D. (ed) *A Census User's Handbook*, Methuen London: 89-113.
- Boyle, P. (1995) Modelling population movement into the Scottish highlands and islands from the remainder of Britain, 1990-1991, *Scottish Geographical Magazine*, 111(1).
- Boyle, P.J., Flowerdew, R. and Shen, J. (1998) Modelling inter-ward migration in Hereford and Worcester: the importance of housing growth and tenure, *Regional Studies*, 32(2): 113-32.
- Champion, A.G. (1994) Population change and migration in Briatin since 1981: evidence for continuing deconcentration, *Environment and Planning A*, 10: 1501-20.
- Champion, A.G. and Atkins, D.J. (1996) The counterurbanisation cascade: an analysis of the 1991 Census Special Migration Statistics for Great Britain, Seminar Paper 66, Department of Geography, University of Newcastle upon Tyne.
- Cole, K. (1995) Why are the Special Workplace and Migration Statistics Special? *MIDAS Newsletter 3*, MIDAS, Manchester.
- Cole, K. and Squires, S. (1987) The Special Migration Statistics from the 1981 Census: the data and their implications, Paper presented at the LAMSAC workshop, County Hall, London.
- Duke-Williams, O. (1997) A Guide to Using the 1991 Special Migration and Workplace Statistics on MIDAS, *Working Paper*, School of Geography, University of Leeds, Leeds.
- Flowerdew, R. and Boyle, P. (1992) Migration trends for the West Midlands: suburbanisation, counterurbanisation or rural depopulation?, Chapter 9 in Stillwell, J., Rees, P. and Boden, P. (eds.) *Migration Processes and Patterns Volume 2: Population Redistribution in the United Kingdom*, Belhaven Press, London: 144-161.
- Flowerdew, R.; Boyle, P. J. (1995) Migration models incorporating interdependence of movers, *Environment and Planning A*, 27(9): 1,493-502.
- Flowerdew, R. and Green, A. (1993) Migration, transport and workplace statistics from the 1991 Census, in Dale A. and Marsh, C. (eds) *The 1991 Census User's Guide*, HMSO, London: 269-94.

- Frost, M., Linneker, B. and Spence, N. (1996) The spatial externalities of car-based worktravel emissions in Greater London, 1981 and 1991, *Transport Policy*, 3, 187-200.
- Frost, M., Linneker, B. J. and Spence, N. (1997) The energy consumption implications of changing worktravel in London, Birmingham and Manchester: 1981 and 1991, *Transportation Research A*, 31(1): 1-19.
- Frost, M., Linneker, B. and Spence, N. (1998) Excess or wasteful commuting in a selection of British cities, *Transportation Research A*, 32, 529-538.
- Forster, E. (1998) Exploring internal migration in Scotland: getting underneath the patterns and unpacking the processes, Poster Paper, Popfest Online 1(1).
- OPCS/GRO(S) (1993a) *1991 Census User Guide 35 Special Migration Statistics: Prospectus*, OPCS and GRO(S).
- OPCS/GRO(S) (1993b) *1991 Census User Guide 36 Special Workplace Statistics: Prospectus*, OPCS and GRO(S).
- OPCS/GRO(S) (1993c) *1991 Census User Guide 51 Special Migration Statistics: Cell Numbering Layouts*, OPCS and GRO(S).
- OPCS/GRO(S) (1993d) *1991 Census User Guide 52 Special Workplace Statistics: Cell Numbering Layouts*, OPCS and GRO(S).
- ONS/GROS/NISRA (1999) *2001 Census: Output: A Discussion Paper. Origin-Destination Output (Workplace/commuting and migration)*, Unpublished Paper.
- ONS/GROS/NISRA (2000) *Origin-Destination Statistics A Discussion Paper*, Consultations, June 2000.
- Rees, P.H. and Duke-Williams, O. (1994) The Special Migration Statistics: a vital resource for research into British migration, Working Paper 94/20, School of Geography, University of Leeds, Leeds.
- Rees, P.H. and Duke-Williams, O. (1995) The story of the British special migration statistics, *Scottish Geographical Magazine*, 11: 13-26.
- Rees, P.H. and Duke-Williams, O. (1997) Methods for estimating missing data on migrants in the 1991 British Census, *International Journal of Population Geography*, 3: 323-368.
- Simpson, S. and Middleton, E. (1999) Undercount of migration in the UK 1991 Census and its impact on counterurbanisation and population projections, *International Journal of Population Geography*, 5: 387-405.

- Spence, N. A. and Frost, M. (1995) Worktravel responses to changing workplaces and changing residences, in Brochie, J. *et al* (eds) *Cities in Competition: Productive and Sustainable Cities for the 21st Century*, Longman: Melbourne, 359-381.
- Turton, I. and Openshaw, S. (1998) *Human Systems Modelling: Results of the HPC Initiative at Leeds*, Centre for Computational Geography, School of Geography, University of Leeds.
- Williams, N., Shucksmith, M., Edmond, H. and Gemmell, A. (1999) *Scottish Rural Life Update: A Revised Socio-Economic Profile of Rural Scotland*, Rural Affairs and Natural Heritage Research Findings No.4, Scottish Office Central Research Unit, The Scottish Office.