Chapter 7
Census Interaction Data and the Means of Access

Oliver Duke-Williams, Vassilis Routsis and John Stillwell

Abstract

The 2011 Census Origin-Destination Statistics collated and released by the Office for National Statistics are much more extensive than those generated from previous censuses, covering student mobility and second home ownership as well as migration and journey to work flows. This chapter explains some of the basic concepts underpinning these data, outlines the questions on the census questionnaire from which the counts of different flows are derived and indicates what data are available under different access conditions at different levels of geography. The chapter also provides some guidance to users who want to use the online interface, *WICID*, to build queries to extract and download the flow data they require for subsequent analysis.

7.1 Introduction

Flow data – often referred to as 'origin-destination data' or ‘interaction data’ – are a specialised form of aggregate data that involve movements (e.g. of goods or people) from one geographical location to another. In a population census context, as we shall see in due course, the data sets are about flows of individuals or households but, more generally, flow data might also include the movement of physical goods such as manufactured commodities or of intangible assets such as flows of capital between banks, or of telecommunications between countries.

Recent population censuses in the United Kingdom (UK) have captured the movement of people or groups of people in households between places of usual residence in the 12 months before each census, together with the movement of people between where they live and where they work. These flow or interaction data sets, referred to collectively as the Origin-Destination Statistics (ODS) by the Office for National Statistics (ONS) in the context of the 2011 Census, also contain counts of students and those with second residences as well as migrants and commuters.

This chapter outlines the different types of flow data that are available from the 2011 Census and explains how their characteristics have changed since previous censuses. Changes in the process of statistical disclosure control mean that whilst there is no small cell adjustment of the 2011 flow data, a three-tier hierarchy of access conditions – open, safeguarded and secure – has been introduced. The chapter will also exemplify how users can access flow data at different levels using the *WICID* software interface that is part of the UK Data Service. Chapters 27 and 28 of this book provide examples of how flow data are being used in research to understand the changing patterns of internal migration and commuting to work between 2001 and 2011.

7.2 Census flow data: concepts and definitions

Census flow data are generated from the responses to certain questions on a census form. In recent decades, two sets of questions have been used to generate flow data relating to migration and to commuting, but new questions in the 2011 Census have permitted the expansion of this somewhat. Results from the 'flow' questions on the census form can be seen in many census data outputs: there are 'origin' (outflow) and 'destination' (inflow) fields to varying levels of geography in cross-sectional and longitudinal microdata sets, and there are a number of aggregate observations in the small area data. This chapter, however, focuses only on those outputs published specifically as part of flow data products that are large and complex (relative to other census data sets) and less widely used.

Flow data sets from the 1966 (sample), 1971 and 1981 Censuses were produced on demand at cost to the customer (Denham and Rhind, 1983, p. 67); more latterly they have been produced as part of planned outputs, and distributed by ONS. Rees *et al.* (2002), writing before the 2001 flow data were released, described the migration outputs produced as part of the 1991 Census and compared them to expected outputs from the 2001 Census; a similar comparison of 1991 and expected 2001 journey to work outputs is reported in Cole *et al.* (2002).

In order to fully understand census migration data, it is useful to set out some concepts and definitions. The first of these is that of 'usual residence', the address at which a person lives all or most of the time. For most people, this idea is wholly

unambiguous, but there are many individuals for whom the concept is more blurred – for example students who may be considered to have both a parental domicile or 'home' address as well as a ‘term-time’ address, seasonal workers who may spend large parts of the year at an address other than their 'home' and persons with second residences who may regularly spend a short time at different 'homes'. The children of divorced parents are another specific population subgroup whose place of usual residence may be divided between two or more locations.

Migration data – as with any other data relating to some change of status – can be viewed at varying levels of granularity. We may view a migration, a permanent change of usual address, to be a discrete event, and try to capture and record all such events over a time period; these are known as 'moves or event data'. Alternatively, we may look at the net effect of an individual’s change in location over a time period in order to ask whether a change has occurred at some stage during that period; these counts are referred to as 'migrant or transition data'. Census migration data in the UK adopt the latter approach, and look at the net effect of migration over a one year period. The two approaches give different counts of migration (Rees, 1977). Census migration data compare a person's address at census date with that 12 months previously, and where there is a difference, a single act of migration is recorded. In practice, a person may have changed address more than once during the transition period; each of these intermediate moves would be recorded by event data, but would not be recognised in transition data.

Further differences arise due to the demographic accounting framework imposed. For a person to be identified as a migrant in census data, it is necessary for them to be present and recorded in the data both at the beginning and end of the transition period. Persons who have left the country in the transition period will not be recorded (and will thus not be identified as migrants), nor will persons who have changed residence during the transition period but who have died before census day. Likewise, infants born during the period will not be included as migrants, an issue that has caused ONS to generate estimates of migrants aged under one year of age in the last two censuses from data on births and migration rates of women of child-bearing age. Assuming that each migration event can be captured, events data have a number of advantages over transition data, not least in identifying the true propensity to migrate, and can be used to estimate transition counts. Events data are

typically captured through population registers or administrative sources (such as the National Health Service Central Register) and cannot be recorded by an instrument such as a census as easily as transition data. The incidence of these two types of migration data, together with lifetime migration, in countries around the world are documented in Bell *et al.* (2014).

Journey to work data are in many ways simpler than migration data as they do not measure change over a transition period. However, they also have their own ambiguities, relating to both place of work, and mode of travel to work. For many people, their place of work is fixed, and can thus easily be described. However, for many other people this is not so easy, as their workplace may vary from day to day, on either a predictable or an unpredictable fashion. The pattern of journey to work may also vary from day to day, both for people with and without fixed workplaces: a journey pattern may include intermediate locations such as a school or caring commitment, or may involve retail destinations as part of the trip; separate journeys to work and returning home may take different routes and may involve different types of transport. Again, a census form necessarily has to try and frame these ideas within a limited amount of space, typically with only a single question possible, and cannot collect the same level of trip event level data as would be possible in a specialised travel survey. Problems arise for the census agencies over how to deal with homeworkers *vis à vis* commuters in the broadest sense and, more specifically, the distinction between those who work at home and those who work from home when reporting mode of transport.

7.3 The 2011 Census flow data questions

Migration questions

A number of questions relating to migration were included on the 2011 Census form. The same main question wording and the same response categories were used in England and Wales (Q21, Form H1), in Scotland (Q10, Form H0), and in Northern Ireland (Q13, Form H4). The response categories were: ‘The address on the front of the questionnaire’; ‘Student term time/boarding school address in the UK’; ‘Another address in the UK’; and ‘Outside the UK’. For the second and third options, space was given to write in a UK address; for the last option space was given to write in a

country name. Those persons who indicated via this question that 12 months prior to the census they were not living at the address on the front of the census form (that is, the person's usual residence at the time of the census) were identified as migrants. This definition thus includes all persons who have changed usual address, both within the UK and those who had entered the UK.

Very similar questions about migration (with minor variation in the question wording, and in the tick box options) have been used in all recent UK Censuses. In the 1971 Census, two questions were included, one asking about usual address one year previously, and the second asking about usual address five years previously. A number of other questions on the 2011 Census form related specifically to international immigrants, including questions identifying how long someone had been in the UK, and the length of intended stay.

A sequence of questions were asked that are particularly relevant to the generation of census flow data. Question 9 on the form used in England and Wales asked for each individual's country of birth, with tick-box options for the most common responses and an additional write-in option. Persons who were not born in the UK were then asked in the next question to give the month and year of most recent arrival in the UK, with a note indicating that short visits outside the UK should be ignored. Finally, those persons indicating that they had been born outside the UK *and* who had most recently arrived within the last year were asked to state the length of intended stay in the UK, with three options: less than six months, six to twelve months, and twelve months or more. Equivalent questions were asked in Northern Ireland, although in Scotland only the month and year of arrival was asked. Where data were collected about both time of arrival and length of intended stay, it is possible to use the definition adopted by the United Nations (1998) that a long-term migrant is one who has moved to another country for at least twelve months, and to restrict tables of results to long-term migrants only. This definition relates to movement across international boundaries; it does not require twelve month residency at any particular residential address within the country.

The question about a respondent's usual address one year prior to the census is often referred to as ‘the migration question’ as it is the key identifier of a 'migrant'. Similar questions are used in many censuses around the world, with some variation in the length of time over which migration is observed (Bell *et al.,* 2014). The

responses to the question allow the generation of a flow matrix from origin (location of usual address one year prior to the census) to destination (location of usual address at the time of the census). Given that locations are fundamentally captured at the address level (assuming that addresses given are correct and complete), they can be aggregated to any level of geography. In the case of migration data, the most fine-grained level of aggregation is the Output Area (OA) level. This is often a far more detailed level than required for analysis, but gives the option of arbitrary re-aggregation into any other geography that can be represented as an amalgamation of these units. Spatial aggregation is frequently required for data to meet confidentiality constraints and is the reason why ONS release only aggregate flows at this geographical scale.

Student questions

As described above, one of the response options on the migration question was to identify a named UK address as being a student term-time or boarding school address. This tick box option had not been included in prior UK censuses (someone with this status would have simply recorded their address as being 'another address in the UK') but is of a significance belied by its small footprint on the physical page. By filtering on this option, it has been possible for the national statistical agencies to generate a subset of migration statistics relating to students, specifically to those persons who had a student address (or boarding school address) one year prior to the census. Of these, some will still have been students at the time of the census, whereas others will have completed their studies. The response to this question thus gives us some idea of how students and recent former students migrate. This is of interest as students are both highly mobile and often poorly captured in data The 1991 Census recorded students at their parental domicile, for example, and a special table (Table 100) was released for student flows, whereas term-time address was used in 2001 (Duke-Williams, 2009)..

In an ideal world, the question about address one year ago might be supplemented with others asking about other statuses one year ago, e.g. employment status, occupation and housing tenure, all of which would help analysts to understand the context of the migration transition as well as the spatial aspects. However, such questions would of course burden the respondent and significantly increase the length of the form, and this is deemed impractical by the census

agencies. However, the 'student' tick box does allow analysts to gain a richer idea of the context of migration, although may of course prompt further questions when interpreting the results. Data disaggregated by student address status one year ago have only been available from the 2011 Census.

Journey to work and place of study questions

Another question that captures address data is that relating to the respondent's place of work. Unlike the migration question, however, the wording of the commuting question in 2011 varied in different parts of the UK. In England and Wales, Q40 asked *"In your main job, what is the address of your workplace?",* with space to write in an address. Tick box categories allowed people to indicate that they worked mainly at or from home, on an offshore installation, or that they have no fixed place of work. The responses to this question allow an origin-destination matrix to be constructed of 'home' to 'workplace' locations.

In Scotland (Q11) and in Northern Ireland (Q43), the census form asked a different question: *"What address do you travel to for your main job or course of study (including school)?".* This captures data on the journey to work in the same way that the question used in England and Wales does, but additionally captures data on the journey to a place of study for students and schoolchildren. The tick box options had slightly different wording as required to refer to both groups, but there was also an additional tick box option: 'Not currently working or studying'. This fundamentally changes the audience for the question. Whereas the question in England and Wales only addresses those self-employed or in employment[1], the question used elsewhere in the UK can be answered by all respondents. A result of this is that care must be taken if comparing results in England and Wales with those in the rest of the UK to ensure that a like-with-like comparison is being made.

For many people in Scotland and Northern Ireland, the distinction between place of work and place of study is not problematic in that people have one of these but not both, and the form design is unambiguous. However, for anyone who both works and studies, a problem arises in that only one address can be given. The form wording directs people to use the place at which they spend most time, but this is

[1] For England and Wales, the question was preceded by a routing question that asked *"If you had a job last week"*, with those persons who did have a job then directed to the 'place of work' question.

subjective, and the balance may well vary over the course of the year. In flow data sets produced from the 2001 Census, the additional availability of data about journey to a place of study led to the creation of different data products for Scotland and for the rest of the UK. For residences in Scotland, journeys to work and place of study were reported in the 2001 Special Travel Statistics (STS), whereas for residences in England, Wales and Northern Ireland, results were reported in the 2001 Special Workplace Statistics (SWS). The 2001 STS could be broadly thought of as a superset – in terms of table structure and content – of the 2001 SWS, but direct comparison between the two sources was awkward for users.

As with the migration data, since these locations are gathered at the address level, they can be freely aggregated at the data processing stage and the finest level at which the 'home' location is observed is the OA level. At the workplace (or destination) end, however, OAs are not the ideal spatial units to use for these data. Some workplaces are physically large (for example, an airport) but do not have permanent residents, and thus cannot easily be accommodated in a residential focussed geography. Some parts of the country – for example, the City of London – have large numbers of workers, but very few residents. In England and Wales, an alternative geography was developed (Martin *et al.,* 2013) of Workplace Zones (WPZs or WZs also often used in documentation) (ONS, 2014). WPZs are designed such that they can be smaller than OAs in areas to which a large number of workers travel (thus allowing the City of London to be split up into small areas), but larger than OAs in residential areas where there are fewer employers (reducing problems of flows of very few people). Consequently, there are 53,578 WPZs in England and Wales, as compared to 181,408 OAs, with WPZs nesting hierarchically into middle super output areas (MSOAs). The journey to work data are thus somewhat more spatially complex than the migration data, in that their base origin and destination geographies are not the same and the matrices of flows from OAs to WPZs are asymmetric.

Whilst various census questions relate to employment, e.g. occupation undertaken, responsibilities and hours worked, one particular question refers directly to the interaction between home and work, namely a question on the method of transport used to travel to work. This question has been included (with evolving response categories) on all censuses from 1966 onwards and is of major importance

for use in transport planning. Chapter 28 in this book examines how modal split has changed between 20001 and 2011 at local authority district scale in England and Wales.

Questions on the workplace have been asked in all recent UK censuses; the overloading of the question to allow place of study to be recorded was used in Scotland for the first time in 2001, and in both Scotland and Northern Ireland in 2011. Like any question that has a written answer (as opposed to a tick box selection), detail on occupation and addresses are difficult and expensive to code. Workplace data were coded for a 10% sample of census forms in 1966, 1971 and 1981. The 1961 Census used a short form/long form approach, with 10% of households receiving a long form (with workplace and occupation questions only appearing on the long form), and all other households receiving a short form; this is the only time that the UK has used this two form approach (ONS, Undated a).

Second residence questions

The 2011 Census in England and Wales included two related 'flow data' questions that have not previously been asked in UK censuses. The first of these (Q5) asked *"Do you stay at another address for more than 30 days a year?"*, with space to write in either a UK address or the name of a country outside the UK. A follow-up question (Q6) then asked about the nature of that address, with options of: 'Armed forces base address'; 'Another address when working away from home'; 'Student's home address'; 'Student's term time address'; 'Another parent or guardian's address'; 'Holiday home'; and 'Other'. Similar questions were trialled in tests in Scotland prior to the 2011 Census (NRS, 2007), but were not used in the full census. The test census form used in Scotland included similar identification of reason for having a second residence, but added two further questions regarding the number of days per week and weeks per year that the second residence was used.

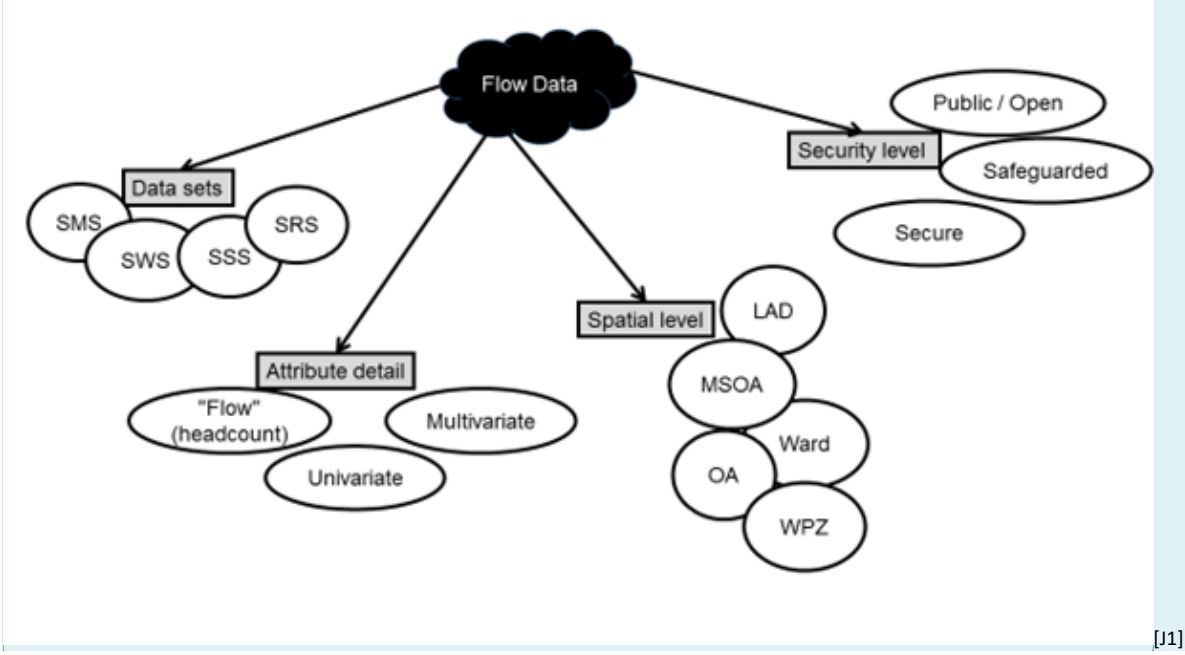
As the response categories suggest, the nature of a 'second residence' can be diverse and covers a broad range of the mobility spectrum (Bell and Ward, 2000), from annual/seasonal movements in the case of a holiday home to weekly or weekly (or more frequent) moves between two locations in the case of commuting/work based residences, and of children alternating time between separated parents. The processing of the data gave rise to a number of different matrices which comprise

the Second Residence Statistics (SRS): from location of usual residence to location of second residence; from location of second residence to location of workplace (for persons who had a second address for work purposes); and a combined category of either primary or secondary residence to workplace location. There is little precedence for a wide ranging question on second residences in the census, although some aspects are picked up in censuses elsewhere. For example, the 2010 Census of Switzerland[2] included a question for employed persons *"From which address do you normally leave for work?"*, with provision for the respondent to indicate whether it was the same address as used on the front of the census form, or a different address, to be written in the space provided, with a parallel question for school children and students. No previous UK census has included a question of this kind.

7.4 Origin-Destination Statistics in 2011

A large number of tables of flow data have been produced as part of the outputs of the 2011 Census. They can be primarily grouped into four families of tables: the Special Migration Statistics (2011 SMS); the Special Workplace Statistics (2011 SWS); the Special Residence Statistics (2011 SRS); and the Special Student Statistics (2011 SSS). Within these families, groupings can also be considered in terms of the structure of the data tables, the spatial level and the security or access level. This set of groupings are summarised in **Error! Reference source not found.**.

[2] http://unstats.un.org/unsd/demographic/sources/census/quest/CHE2010en.pdf



[J1]

Figure 0.1 2011 Census flow data outputs

Table structure

The structure of the data tables describes the amount of attribute detail used to tabulate results. There are three levels: at the most aggregate, flows are reported in terms of total numbers of persons only and are referred to as 'flow' or 'headcount' tables. There is scope for ambiguity with this terminology: in this chapter the term ‘flow data tables’ refers generically to all data tables from flow data outputs, regardless of structure, whilst the term ‘flow tables’ refers to a specific subset of flow data tables, namely those that report total counts only. ‘Univariate’ tables report a given flow disaggregated by categories of a single variable (such as age), whereas ‘multivariate’ tables report a given flow by cross-tabulated categories of two or more variables (such as age and sex). The terms 'univariate' and 'multivariate' are used slightly loosely in this respect, in that the data are already disaggregated by two further characteristics, an origin location and a destination location.

Spatial structure

The spatial level or geography used to report flow data also varies between tables of output. This is represented in part in Figure 7.1, although the actual set of geographies used varies within and between countries of the UK. At the finest scale, OAs are used across the whole of the UK to tabulate flow data. WPZs are the finest geography used to tabulate workplace destinations in England and Wales but were

not constructed in Scotland or Northern Ireland at the same time as those for England and Wales.

Both these base polygons can be aggregated into larger units, although they have separate pathways. OAs can be aggregated into Lower Layer Super Output Areas (LSOAs), which in turn can be aggregated in to Middle Layer Super Output Areas (MSOAs). WPZs can be aggregated into MSOAs but they do not nest within LSOAs. MSOAs are thus the lowest level geography at which migration and commuting data can be easily compared with an entirely consistent geography, for the majority of journey to work observations.. Many tables are made available at the local authority district (LAD) level, a composite of several different administrative units across parts of the UK: London Boroughs (plus the City of London), Metropolitan Districts, Non-metropolitan Districts, Unitary Authorities (plus the Isles of Scilly), Council Areas and District Council Areas, as explained in Chapter 6. For the 2011 flow data, it is common for the City of London to be reported as an aggregate unit together with City of Westminster, and for the Isles of Scilly to be similarly aggregated with the Cornwall UA, due to small residential population sizes in both cases.

The 2011 flow data offer an improvement over the 2001 outputs in the handling of overseas origins. Whereas migration data in 1991 included country of origin (major origin countries, and regionally grouped smaller countries), in 2001 this level of reporting was removed, with only 'total from overseas' counts being included. The 2011 migration data are generally divided into 'UK' and 'non-UK' variants, with the non-UK versions tabulating international flows. For the non-UK tables, some have a broad geography, in which 59 countries or groups of countries are recognised, whilst some are 'detailed' and use a standard coding in which up to 184 countries or groups of countries can be identified. As with other levels of spatial detail, there is a trade-off between detail and ease of access to the data.

Geographic scope

As well as the reporting geographies used to disaggregate output tables, it is useful to consider the geographic scope of the different outputs. This is more complex with

the 2011 outputs than with outputs form earlier censuses. Some outputs can only be generated for some countries, due to differences in questions asked. Thus, flow data tables relating to second residences are only available for people with (primary) residences in England and Wales, as the relevant questions were not asked of respondents living elsewhere in the UK. A question about Welsh language skills was asked in Wales, and a flow data set disaggregated by Welsh language ability is available for residences in Wales only. Similar flow tables about language ability in Scotland were published as part of the outputs from earlier censuses, but have not yet been published as part of the 2011 outputs.

Regardless of the main scope of any table, the outputs usually contain cross-border flows. Thus, tables published for residences in Scotland might potentially include flows to destinations in England (or Wales or Northern Ireland), and so on. Cross-border flows are more problematic in the 2011 outputs than was the case with outputs from earlier censuses as they sometimes feature asymmetric geographies: thus flows within a country may be tabulated at ward to ward level, but cross-border flows will be tabulated at a ward to district level. Some tables have been specified at a UK level, whereas others are only published for England and Wales, Scotland or Northern Ireland.

Summary of tables

Error! Reference source not found. provides a summary of the numbers of flow data tables produced at different levels of attribute detail, by country and by 'family' of outputs. The total of 223 tables represents a considerable expansion on the number of tables published as part of the 2001 Census (across three spatial levels) which comprised a total of 16 migration tables, 14 journey to work tables and 14 'travel' tables (Stillwell and Duke-Williams, 2007).

Table 7.1 Summary of origin-destination tables from 2011 Census

	SMS			SWS		SRS	SSS		Total
Attribute detail	UK	EW	W	UK	EW	EW	UK	SC	
Flow count	5	-	-	4	6	24	3	-	42
Univariate	22	13	2	43	38	34	9	3	164
Multivariate	12	2	-	-	2	-	1	-	17

Total	39	15	2	47	46	58	13	3	223

Whereas the geographies used in the 2001 and earlier outputs were relatively straightforward, those used in the 2011 outputs are more diverse. Not only are a wider range of geographies used, but the names and scope of these geographies varies across component parts of the UK, resulting in a set of output tables which in some cases only apply to part of the UK.

7.5 **Security and access control** summarises the range of geographies used and the total number of zones in each system.

Table 7.2 Summary of the most common 2011 Census geographies used in flow data tables

Name	**Breakdown**	**Remarks**
United Kingdom Output Areas	Total: 232,296 England: 171,372 Wales: 10,036 Scotland: 46,351 Northern Ireland: 4,537*	* ONS never used OAs for Northern Ireland in the official 2011 Census data dissemination. In most of the 2011 Census flow data sets containing lower level geographies, Super Output Areas (SOAs) were used instead.
United Kingdom Workplace Zones	Total: 100,819 England: 50,868 Wales: 2,710 Scotland: 46,351* Northern Ireland: 890*	* As of January 2016 no WPZs had been released for Scotland and Northern Ireland. As part of this geography data set, Scotland OAs and Northern Ireland SOAs were added.
United Kingdom Lower Layer Super Output Areas (LSOA)	Total: 42,143 England: 32,844 Wales: 1,909 Scotland: 6,500* Northern Ireland: 890*	* Northern Ireland SOAs and Scotland Data Zones (DZs) have been used in this geography data set.
United Kingdom Middle Super Output Areas (MSOA)	Total: 9,326 England: 6,791 Wales: 410 Scotland: 1,235* Northern Ireland: 890*	*Only England and Wales MSOAs have been released. As part of this geography data set, Scotland Intermediate Zones (IZs) and Northern Ireland SOAs were added.
United Kingdom Wards	Total: 9,505 England: 7,689 Wales: 881 Scotland: 353 Northern Ireland: 582	
United Kingdom Merged Local Authorities	Total: 404 England: 324 Wales: 22 Scotland: 32 Northern Ireland: 26	
United Kingdom	Total: 12	* Only England regions have been released. In

Regions	England: 9 Wales: 1* Scotland: 1* Northern Ireland: 1*	WICID, Wales, Scotland and Northern Ireland are represented by their country geographical code.

7.5 Security and access control

Perhaps the most significant development in the 2011 origin-destination data has been the introduction of a revised approach to data security. Various methods of preserving confidentiality have been used in past censuses; whilst some similar approaches were retained in the 2011 outputs, much of the emphasis has been placed on controlling availability of the data. In previous censuses, a range of statistical disclosure control approaches were used to 'protect' the contents of tables of results (Duke-Williams and Stillwell, 2007). Outputs from the 2001 Census were (for residences in England and Wales, and in Northern Ireland) protected using an approach called ‘small cell adjustment method’ (SCAM) that probabilistically altered values of 1 and 2 to 0 or 3.

SCAM proved problematic for flow data in particular, as these data are characterised by having a very large proportion of the non-zero flows being small values. This method was dropped for 2011 Census data; instead, the liability was moved down from ONS to census users who are now responsible for the protection of low flow counts in their research outputs.

ONS has used a method called record swapping to help protect the data. Record swapping is a pre-tabulation method applied to all the 2011 Census data (including the microdata) and involves swapping information between a pair of households within a proximate area (ONS, 2012). This method is not limited to flows below three so all counts have the same potential to contain swapped records. Record swapping ensures that any 'attacker' making a claim to have identified an individual within the census outputs could be met with an argument that it is possible that the 'person' in question may in fact be contained within a swapped record: thus, there is always uncertainty that any given data point is accurate. An advantage of record swapping over post-tabulation forms of disclosure control is that there is consistency (in terms of values aggregating to the same total) within and between

published data tables, a feature that is not possible with approaches such as SCAM which undermine the concept of a ‘one-number’ census.

The record swapping approach to disclosure control is not sufficient to fully protect all of the data; there remain a large number of small values within the outputs, some of which are 'genuine' non-swapped values, and thus may provide a risk of disclosure. The proportion of records that were swapped is unknown, and thus the level of uncertainty in the data cannot be quantified. Further measures have been put into place to protect the small values that frequently occur in flow data, namely controlling who can access different sets of outputs.

The 2011 Census origin-destination data outputs have been divided into three groups (ONS, Undated b) that fit with an existing ONS taxonomy of access control:

- *Public* data, also referred to by users as 'open data' and as 'Open Government License (OGL) data', are openly available to all users, from two main sources as outlined below;
- *Safeguarded* data are not considered to be personal (in the context of a legal definition of 'personal data'), but for which there may be a risk of disclosure if they are linked to other data (including *a priori* knowledge); and
- *Secure* (or 'controlled') data may be identifiable, and thus are potentially disclosive. In the case of origin-destination data, secure data have a high level of attribute and/or spatial detail, giving the risk that someone may claim to be able to identify individuals (subject to record swapping).

The number of tables in each access category are indicated in Table 7.3. Certain access conditions have been put in place for safeguarded and secure data. Safeguarded data are available to users who are identified as being from the public sector, subject to publication restrictions. Users are advised by ONS – via an agreed text displayed on the website providing access to the data – that small values (that is, values below three) in outputs must be protected. ONS suggest that data *could* be rounded, but note that this is only one possible method. In practice, responsible users will need to take care that any measures applied cannot be reversed (for example, by reference to unmodified sub-totals). Alternative approaches are up to the user, but might include cell suppression or aggregation. Controls on publication of small numbers also apply to non-tabular outputs, such as graphs or maps. A

consequence of the restriction on publication of small numbers is that the data cannot be re-distributed in their original form. Researchers outside the public sector who wish to use the safeguarded data are obliged to visit the ONS Virtual Microdata Lab (VML), a safe-setting controlled access facility where they can also access the secure tables.

The 'secure' data sets are available for use under controlled conditions by researchers who have Approved Researcher accreditation, and who have had a specific project approved. Data are currently available in comma separated value and (in some cases) SPSS formats. As with other controlled-access data sets, outputs produced by researchers must satisfy disclosure risk assessments before they can be published or otherwise disseminated by researchers. Researchers are required to book a terminal in advance, and then travel to one of the ONS offices at which the VML can be accessed.

Table 7.3 2011 Census flow tables in each access category

	Open	Safeguarded	Secure	Total
SMS	5	51	176	232
SWS	9	84	193	286
SSS	1	15	33	49
SRS	6	52	117	175
Total	21	202	519	742

7.3 Obtaining flow data

As outlined in the previous section, there are different routes to using the census origin destination data that are dependent on the security or access rating of each table. The public flow data are available through NOMIS[3] (Townsend *et al.,* 1987) and through *WICID*[4] (Stillwell and Duke-Williams, 2003). *WICID* (Web-based Interface to Census Interaction Data) has been developed and extended through a

[3] https://www.nomisweb.co.uk/census/2011

[4] https://wicid.ukdataservice.ac.uk/

series of ESRC funded projects, and is part of the UK Data Service Census Support. The open nature of these data mean that it is likely that redistribution will also occur through other routes as well over time.

WICID also provides access to the safeguarded flow data and is the primary access route for these data for users in the public sector. As well as providing an interface to plan and download data extracts, the system also hosts download files in both CSV format and in a format suitable for use in conjunction with *SASPAC* (Rhind, 1984), a system which is widely adopted in local authorities. The secure flow data can only be accessed via the ONS VML and a user must have Approved Researcher status. The VML is also the access route for non-public sector users of safeguarded flow data.

This section focuses on the Web-based Interface to Census Interaction data (*WICID*) that is available via the UK Data Service website and describes how it can be used to create a data extract. Users are stepped through the process of constructing a query in two main phases, after which the design and structure of the download can be selected. The two main phases are data selection and geography selection. We explain each of these in turn.

In the first step, the user is required to select some data of interest. A number of approaches (selection tools) are available (Figure 7.2). *Quick Selection* allows simple access to a number of pre-defined totals (typically, total persons in a flow) for a variety of tables. The data are structured as consisting of tables (grouped into sets), each made up of a number of variables. The related metadata permits two selection methods: *Select by variable* and *Select by dataset and table*. The *Select by variable* tool lists all identified thematic variables. Thus, the term ‘age’ is listed as a single variable, rather than the large number of variants of age reflecting different age groups used in different tables. Having select a variable, the user is iteratively shown a list of matching tables, together with a modified list of variables, limited to those that are used in conjunction with the currently selected variable(s). The *Select by dataset and table* tool allows browsing through all tables. In both tools, having selected a table, the user is shown a representation of the table, from which all or some of the table cells can be selected and added to the query.

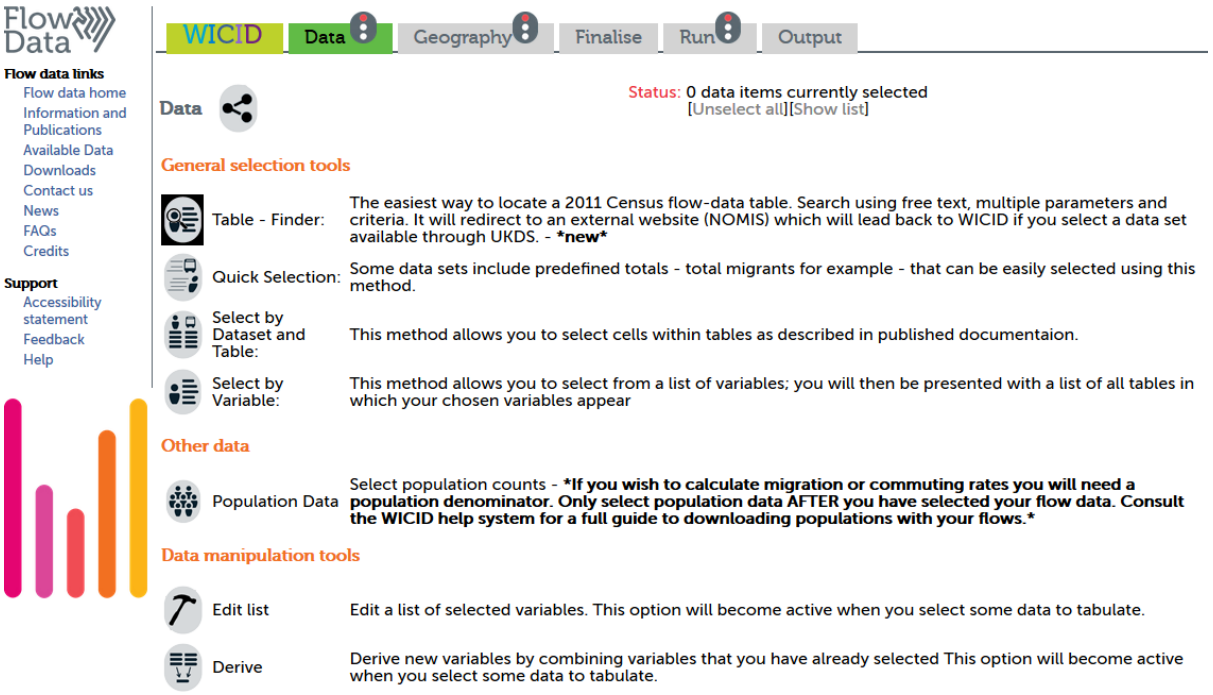


Figure 7.2 The data selection tools interface in *WICID*

In contrast with the outputs from 1991 and 2001, data in a much larger number of tables were disseminated as part of the 2011 Census Flow Data release, often with somewhat generic titles. This has made browsing by dataset and table more impractical. Working together with ONS and NOMIS, a new 2011 Census 'table-finder' tool was developed for this purpose. *WICID* provides easy access to this tool by integrating control from and to it. As part of this implementation, a prominent button that redirects to the *Table-Finder* has been placed in the main *WICID* data selection page (Figure 7.2).

The *Table-Finder* tool provides interactive selection of 2011 Census tables by allowing users to search using free text and multiple parameters and criteria. All 2011 Census open and safeguarded flow-data tables contain links back to *WICID* when selected. Users can also choose to visit *NOMIS* webpages instead (open data only) or contact ONS to request permission to access the data set in their VML. It may be worth noting that the *Table-Finder* only lists tables with their natively supported geographies, therefore tables of flows aggregated to higher level geographies which are supported by *WICID* will not be displayed.

The *WICID* link redirects back to the *WICID* system at the table information webpage; this page provides detailed information on the census table that has been selected such as totals, supported geographies (both native and via aggregation), variables and citation information. It also contains a button to select the table via the *WICID* queryable tool and, wherever available, buttons to download the table in CSV or *SASPAC* format via the downloads page. Access control remains the same; users will need to login via Federated Access in order to be able to query safeguarded data, while a simple anonymous guest-login is enough to acquire access to the open data sets.

Having selected data to tabulate, the user is then directed to select geographical areas of interest as part of the second phase of query building. As with the data phase, there are a number of tools in the geography phase to assist the user. The *Quick selection* tool allows the user to add all areas in a chosen geography (e.g. all wards, all districts, *et cetera*) to their query. The *List selection* tool allows the user to select a geography, and then a list is shown of all components of that geography, from which the user can select one or more areas. This is practical for geographies with a relatively small number of areas, but looking through a list is impractical for geographies with thousands or tens of thousands of areas. In such cases, the *Type-in* box selection method is more helpful – users can enter an area name (or wildcarded partial name) or, if known, alphanumeric code, and will be shown a list of all matches, from which they can select one or more areas.

A related tool is *Postcode selection*, which allows the user to type in a full or partial postcode, and generate a list of all areas that intersect with that postcode. For all tools, the user is initially shown a list of geographies in order to choose one from which to select areas. This list is modified based on the currently selected data; thus, the user can only select areas from geographies that are compatible with their previous data selection. As well as familiar geographies, there are also aspatial ‘geographies’ which group together various totals and special categories for some data sets and represent these as ‘areas’. Thus, for example, a user might select all flows from a set of regular areas to the aspatial destination ‘workplace unknown’.

For origin-destination data, it is necessary to select two sets of areas. In many cases, the user will want to carry out a symmetric selection – that is, the set of origin and destination areas will be identical. A short cut tool, ‘Copy selection’, enables this

to done easily, setting the destinations to be the same as the currently selected origins, or *vice versa*. Once the Geography part of the query has been completed, the query can be run to generate the flows. Figure 7.3 shows a *WICID* query that has been constructed in order to extract and download total migration flows within and between 12 regions of the UK from SMS Table MM01CUK_all which contains data within and between LADs.

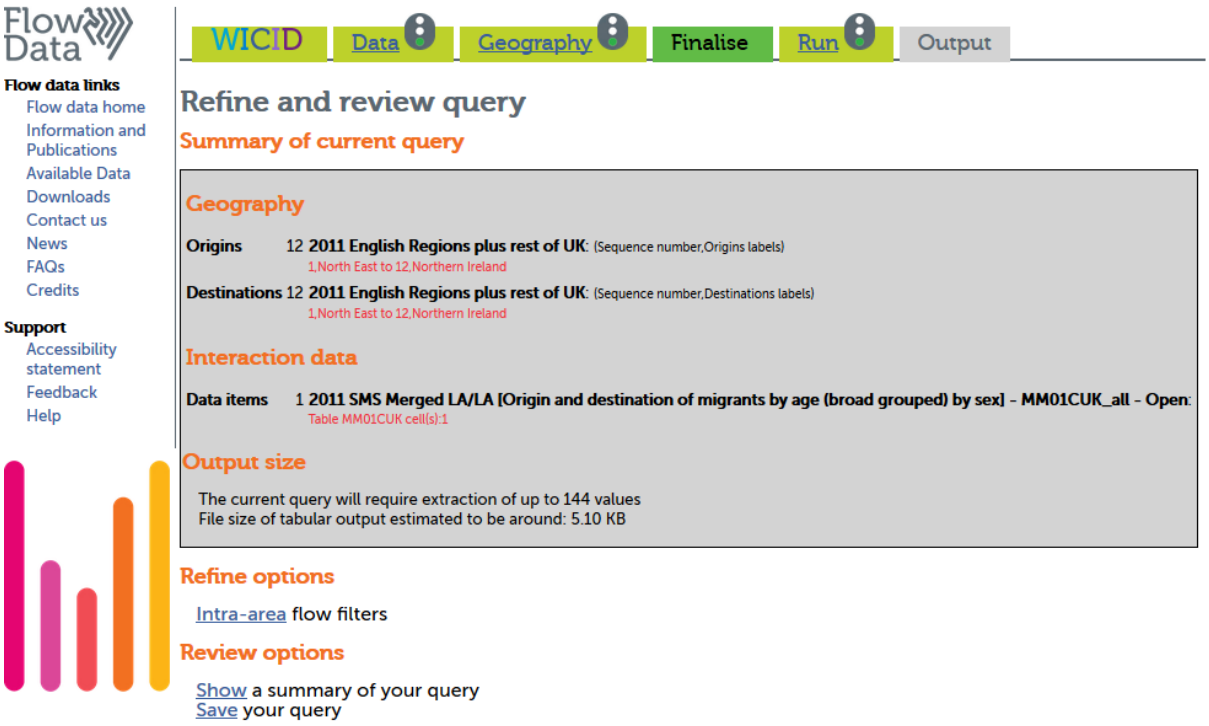


Figure 7.3 *WICID* query for extraction of regional migration flows from 2011 SWS

WICID includes a large number of look-up tables converting from one geography to another, which permit on-the-fly aggregation of base areas. Thus, whilst any particular dataset has been formally published at one spatial scale, *WICID* can seamlessly present it as also being available for aggregates of that base geography, as indicated in the example shown in Figure 7.3. Within the *List selection* mode it is also possible to exploit these look-up tables in another way. Having initially selected a geography for which results are to be tabulated, the user has the option of selecting an additional aggregate geography to use to help select areas. For example, a user might select wards as the initial geography, but then rather than

typing in ward names or selecting from a list of wards, would select districts as the selection geography. In this scenario, the user would be shown a list of districts, and on choosing one, all wards within that district would be added to the query.

7.4 Conclusions

The 2011 origin-destination outputs offer a number of advantages over the outputs of previous censuses. In particular, they include entirely new data sets (SSS and SRS) which will allow researchers to explore mobility patterns in new ways; there are a larger number of output tables than have previously been the case, especially at a spatially detailed level; and the SCAM disclosure control approach used with the 2001 outputs, which was particularly problematic in the case of flow data, has not been repeated.

These advantages have come with certain costs. Most significantly, the trade off for more detailed data and less heavy-handed disclosure control has been the introduction of access controls. For public sector users, access is not especially onerous: the data are available for download and for on-line use. The restrictions on publication may prove somewhat more tricky, given that spatially detailed flow data are characterised by a large number of cells with small values, including large numbers of cells with the value 1 and 2. Dealing with this will require a two-pronged attack for educators and trainers: first, it will be important to make sure that users of the data are aware of the requirements; and second, advice will need to be given over how users might modify the data appropriately.

A secondary problem is posed by the large number of tables: these are considerably harder to navigate than has been the case in the past. The situation is complicated by the fact that some of these tables are overlapping in the way that they provide data for different parts of the UK, with variant scales for cross-border flows. It will be helpful to derive portmanteau data sets that group together data from different source tables to a UK level to aid the user experience. A part of the problem with table navigation seems to arise from the simple state of being flow data tables: the table titles used by ONS attempt to incorporate labelling of the spatial structure and base population of the table as well as the actual content (that is, the cross-

classifying variables in the table), and become long, unwieldy and hard to differentiate.

The origin-destination tables are invariably amongst the last of the outputs from any census since they draw on data collected separately by each of the national statistical agencies and require conflation (by ONS). Given their characteristic of possessing a dual geography, they are also some of the most difficult data sets to process, disseminate and analyse. The *WICID* interface attempts to facilitate access to these data sets and we hope that taking the time to understand them will be rewarding for researchers.

Acknowledgements[JS2]

References

Bell, M., Charles-Edwards, E., Kupiszewska, D., Kupiszewski, M., Stillwell, J. and Zhu, Y. (2014) Internal migration data around the world: Assessing contemporary practice. *Population, Space and Place,* 21(1): 1-17*.*

Bell, M. and Ward, G. (2000) Comparing temporary mobility with permanent migration. *Tourism Geographies,* 2(1): 87-107.

Cole, K., Frost, M. and Thomas, F. (2002) Workplace data from the census. In Rees, P., Martin, D. and Williamson, P. (eds.) *The Census Data System.* John Wiley, Chichester, pp 269-280.

Denhman, C. and Rhind, D. (1983) The 1981 Census and its results. In Rhind, D. (ed.) *A Census User's Handbook.* Methuen, London, pp. 265-286.

Duke-Williams, O. (2009) The geographies of student migration in the UK. *Environment and Planning A,* 41(8): 1826-1848.

Duke-Williams, O. and Stillwell, J. (2007) Investigating the potential effects of small cell adjustment on interaction data from the 2001 Census. *Environment and Planning A*, 39(5):1079-1100.

Martin, D., Cockings, S. and Harfoot, A. (2013) Development of a geographical framework for census workplace data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2): 585-602.

ONS (Undated a) Census 1911-2011. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/census-history/census-1911-2011/index.html.

ONS (Undated b) 2011 Census Origin-Destination Data User Guide. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/release-plans-for-2011-census-statistics/subsequent-releases-of-specialist-products/flow-data/origin-destination-data--user-guide.pdf.

ONS (2012) 2011 Census: Methods and Quality Report Confidentiality Protection Provided by Statistical Disclosure Control. Available at: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/statistical-disclosure-control-methods/confidentiality-protection-provided-by-statistical-disclosure-control.pdf.

ONS (2014) Workplace Zones: A new geography for workplace statistics. Available at: http://webarchive.nationalarchives.gov.uk/20160105160709/https://geoportal.statistics.gov.uk/Docs/An_overview_of_workplace_zones_for_workplace_statistics_V1.01.zip.

NRS (2007) 2006 Census Test Form. Available at: http://www.scotlandscensus.gov.uk/documents/preparation/2006-census-test-form1.pdf.

Rees, P. (1977) The measurement of migration, from census data and other sources. *Environment and Planning A*, 9: 247-272.

Rees, P.H., Thomas, F. and Duke-Williams, O.W. (2002) Migration data from the census. In Rees, P., Martin, D. and Williamson, P. (eds.) *The Census Data System*. Wiley, Chichester, pp.245-267.

Rhind, D. (1984) The SASPAC story. *BURISA*, 60: 8-10.

Stillwell, J. and Duke-Williams, O., 2003. A new web-based interface to British census of population origin–destination statistics. *Environment and Planning A*, *35*(1):113-132.

Stillwell, J. and Duke-Williams, O. (2007) Understanding the 2001 UK census migration and commuting data: the effect of small cell adjustment and problems of comparison with 1991. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2): 425-445.

Townsend, A.R., Blakemore, M.J. and Nelson, R. (1987) The NOMIS data base: Availability and uses for geographers. *Area*, 19: 43-50.

United Nations (1998) Recommendations on Statistics of International Migration, Revision 1[J3]