Technical Working Paper:
Guide to parallel and combined analysis of the ONS LS, SLS and NILS

Harriet Young

July 2009

Contents

1. introduction –

Background
The Office for National Statistics Longitudinal Study (ONS LS), covering England and Wales was established in the mid 1970s and now contains data from up to four Censuses, together with information on vital events currently (July 2009) available from the 1971 Census until the end of 2006. In 2007 the Scottish Longitudinal Study (SLS) and the Northern Ireland Longitudinal Study (NILS) were both launched; both, like the ONS LS, include information from Census and vital registration sources, although there are a number of differences in linked data available which are referred to below. This means that it is now possible to carry out UK-wide analysis using Census based record linkage studies.

The aim of this document is to provide a guide for those who wish to carry out analysis of two or three of these datasets together. The information presented is based on experience of carrying out analysis on the three studies together as part of an ESRC Census Programme funded project. The next section briefly presents similarities and differences between the three datasets. Subsequent sections document legal and logistical issues in carrying out combined analyses; options for project design; procedures for the application process; and issues in data preparation and analysis.

2. Similarities and differences between the studies

The ONS LS, the NILS and the SLS are all Census based record linkage studies. Samples are initially drawn from census data based on a number of birthdays in the year. After the initial census, data from subsequent census points and vital events data are linked in for each individual. There are some variations in data available in each study. (see Table 1), but all three include vital events data (taken from registration data) on births to sample mothers, death of the longitudinal study (LS) member and their spouse, and international migration data. (Note that LS migration data are known to be very incomplete.) Samples are maintained by the addition of new births and immigrants born on LS birthdays.

There are a number of differences between the three datasets, and in the table below we present characteristics of each dataset and information about the vital events data available for each.

Table 1: Characteristics and vital event data for the ONS LS, SLS and NILS

	ONS LS	SLS	NILS
Percent of population in sample	1%	5.3%	28%
Method of sampling	4 birth dates in year	20 birth dates in year	104 birth dates in year
Approximate sample size in 2001	Approx 540,000	Approx 274,000	Approx 500,000
Census data available	1971, 1981,1991, 2001	1991, 2001	2001
Vital registration data			
New births	Y	Y	Y
Births to sample mother	Y	Y	Y

Death of spouse	Y	Y	Y
Death	Y	Y	Y
Embarkation from country	Y	Y	Y
Immigration and re-entry to country	Y	Y	Y
Internal migration (moves between health areas)	Included in LS in early 1970s only	Y	Y
Births to sample father	Included in LS in early 1970s only	Y	Y
Cancer registrations	Y	Y	-
Hospital admission & discharge	-	Y	-
Marriage event data	-	Y (from 1991)	Y (from 2004)
Valuations and lands agency data	-	-	Y

This shows variations in data captured for each study. In addition there are some differences in census questions and the coding of census data which meant that variables included in the three studies are in a few cases slightly different. More detail on variable and coding differences may be found in the combined data thesaurus, which contains information on 65 key variables (http://www.celsius.lshtm.ac.uk/ukanal.html). Information is also available in the data dictionaries for each study, available on the CeLSIUS, the Longitudinal Studies Centre Scotland and the NILS Research Support Unit websites respectively. (Details of support unit contact information and websites are listed in Appendix 1.)

3. Data access and confidentiality guidelines

For each study, access to anonymised individual level data is only possible in the respective statistical office safe setting. Permission to release data outputs from the safe settings is governed by disclosure control policies, and outputs must be cleared for release to the researcher by the statistical office concerned. For all three datasets, it is possible to release tabulations and regression output that meet minimum cell count guidelines, as outlined in Table 2 below.

Table 2: Minimum cell counts possible for release of data from the ONS LS, SLS and NILS as at May 2009

Outputs	LS	SLS	NILS
Intermediate outputs (for use by research team only)	2 or 3 *	3	n/a
Final outputs (for presentations or papers)	3	3	10 **

* The minimum cell count for release to researchers in England & Wales is 2. For researchers elsewhere, the minimum cell count is 3.

** This is a guideline and lower cell counts may be allowed on application.

The ONS LS disclosure control policy allows for the release of aggregated (weighted count) datasets[1] that meet the minimum cell count guidelines. For the NILS and SLS, at the time of our project, there was no permission to release aggregated datasets, although it is possible to apply for permission to use these, for consideration on a case by case basis. The variation in the disclosure control policy between statistical offices reflects differing sampling fractions for the respective studies. For example in Northern Ireland, with a sampling fraction of 28 per cent, there is a much higher chance that a sample unique will also be a population unique, leading to a more conservative disclosure control policy for the NILS than for the ONS LS and SLS.

4. Project design

There are three possible methods of analysing the ONS LS, NILS and SLS together. The first and most straight-forward method involves parallel, but separate analyses of each dataset and comparisons of the results. The second method involves making aggregated count datasets (that meet the disclosure control guidelines mentioned above) for each Longitudinal Study, and collating these for combined analysis. A specific application must be made to the relevant statistical offices in order to obtain aggregated datasets. The third method is to combine subsets of individual level data from the ONS LS, SLS and NILS. It is not possible to carry out this latter method at present due to the data confidentiality policies of all three studies. However, ONS, GROS and NISRA are together considering the feasibility of carrying out combined analysis of the three LSs, so it may be possible in the future. In the remainder of this document ‘parallel analysis’ will refer to the first method and ‘combined analysis’ to the second.

5. The application process

Parallel analysis

For separate parallel analysis, the application process is straightforward. Application for use of each dataset must be made separately to the respective statistical offices via the relevant user support service. The steps involved are outlined on the LS websites.

Combined aggregated analysis

As discussed in section 3. above, GROS and NISRA disclosure control guidelines do not currently allow for release of aggregated datasets to researchers. Researchers wishing to prepare and combine aggregated datasets need to apply to ONS, GROS and NISRA via the relevant user support service for consideration on a case by case basis. It may be the case (as with our project) that this will be approved on the proviso that dataset combination and subsequent analysis takes places in a statistical office safe setting at NISRA, ONS or GROS. Any tabulations or regression output from the appended aggregated dataset would have to be cleared by all three statistical offices for release to the researcher.

NISRA disclosure control guidelines stipulate a minimum cell count of 10 for release of any output from the safe setting, compared with 3 for the SLS and ONS LS. For our aggregated datasets, we wanted equivalent minimum cell counts for all three studies, and so applied to NISRA for permission to have a minimum cell count of 3 for release of

[1] In an aggregated weighted count dataset, each row of data consists of a count of people with a particular set of characteristics, e.g. a count of 4 individuals in a dataset with the characteristics of woman, aged 35-49, good self-rated health, no long term illness, married, car, no education.

these datasets. This was only granted because the data was being sent to a statistical office safe setting, and for all final outputs from ONS to the researcher, the minimum cell count remained at 10.

6. Data preparation

This section outlines the data preparation process that we went through for our project.

Parallel analysis

For parallel analysis, we followed the usual procedures for each statistical office for dataset preparation. We sent our specifications and programme files to the academic support units for each LS and they prepared the datasets, and produced results which were cleared and sent to the researcher.

Combined analysis

Preparation of combined analysis datasets was more complex and we had to go through a number of stages before the data was ready for analysis. Our goal was to develop aggregated datasets with identical specifications and variables for each study. These had to be produced separately in the three safe settings, and each aggregated dataset was restricted to a count of 3 or more individuals for each row of aggregated data.

We started with initial exploration of ONS LS data. We then travelled to GROS in Edinburgh to develop and produce aggregated clearable datasets for the SLS. We set the specifications for all datasets using the SLS because the sample is the smallest, making it likely to be the most restrictive dataset. We then used this data specification template to produce NILS datasets at NISRA in Belfast and ONS LS datasets at ONS in London. To avoid the necessity for further travel to NISRA and GROS, user support officers for the NILS and SLS were instructed on how to make future datasets if necessary. Once datasets were finalised, they were cleared and released by NISRA and SLS and sent to ONS. This whole process took a number of weeks and involved travel to GROS and NISRA for three and two days respectively plus initial exploration at ONS and time to arrange clearance of datasets.

Initially, when we simply aggregated individual level datasets, we obtained hundreds of cell counts of one and two. We therefore used three methods of data manipulation to obtain the final aggregated datasets:

1. Instead of creating one aggregated dataset for each Longitudinal Study, we created separate datasets for the different sections of analysis. For example, we made separate datasets for the self-rated health and limiting long term illness outcome variables. This meant that we had fewer variables in each dataset and could maximise the number of categories for the variables we did include. Note that we also had to create separate datasets for mortality outcome analysis where we needed counts of death (rather than a count of individuals) and a sum of person years.

2. We reduced the number of variable categories to ensure that the majority of rows contained counts of three or more. For example, we aggregated single year of age into ten or twenty year age groups, and created binary variables from three or four category variables. For aggregated datasets with rare outcomes (e.g. mortality outcome for those aged under 75), where a count of one or two deaths for a particular set of characteristics was more likely, we had to combine four socio-economic variables into a summary score

because the dataset would not support more detail without large numbers of counts of 1 or 2. We tested different variable and variable category combinations, and the final choice was based on a combination of initial findings of parallel analysis, the testing of different combinations of variables and categories in the data development stage and on what was practically possible.

3. We included some counts of one and two in the initial versions of the aggregated datasets, and manipulated the data manually to remove these. In any one dataset, we had between zero and approximately ten cells that needed this manual manipulation. For each row with a count of one or two, we changed one variable characteristic so that this row was subsumed into another one. Once we completed this process, we created regression models comparing the unchanged and the changed dataset to check that results did not differ, and found that they did not. These changes are therefore likely to have minimal or no effect on final results.

7. Data analysis

In descriptive analysis, it is necessary to weight data to produce summary statistics for two or more surveys combined, because of the different sampling fractions in the three studies. We derived country weights by dividing the proportion of the UK population in country x by the proportion of the sample population in country x, and then applied these as population weights in STATA using the (pweight) command, for example, '*tab health gender [pw=countryweight], m row* '.

For more information on comparisons of parallel and combined analysis methods and the strengths and weaknesses of each, see our forthcoming paper in *Population Trends*, and a powerpoint presentation available on the CeLSIUS web site (http://www.celsius.lshtm.ac.uk/documents/presents/BSPS_presentation_May09_Harriet_Young.pdf).

Acknowledgements

This document was written by Harriet Young on behalf of the project team, Emily Grundy, Paul Boyle, Dermot O'Reilly. We would like to thank Rachel Stuchbury as author of the data thesaurus, and the user support offices (CeLSIUS, LSCS, NILS RSU) and the statistical offices (ONS, GROS and NISRA) for all their support.

Appendix 1
Contact information

ONS LS

ONS LS data is held at the Office for National Statistics (ONS) in Titchfield, and is also available for access via a Virtual Microdata Laboratory facility in London. The support unit for academics wishing to use the ONS LS is the Centre for Longitudinal Study Information and User Support (CeLSIUS), based at the London School of Hygiene and Tropical Medicine.
Contact details:
www.celsius.lshtm.ac.uk
Celsius@lshtm.ac.uk
+44 (0)20 7299 4634

SLS

SLS data is held at the General Register Office for Scotland (GROS) in Edinburgh. The support unit for academics wishing to use the SLS is the Longitudinal Studies Centre – Scotland (LSCS), based at the University of St Andrews.
Contact details:
www.lscs.ac.uk
lscs@st-andrews.ac.uk
+44 (0)1334 463 992

NILS

NILS data is held at the Northern Ireland Statistics and Research Agency (NISRA) in Belfast. The academic support unit for the NILS is the NILS Research Support Unit (NILS RSU).
Contact details:
www.qub.ac.uk/research-centres/NILSResearchSupportUnit/
nils-rsu@qub.ac.uk
+44 (0)28 9082 8210