A Synthetic Longitudinal Study for the United Kingdom

Adam Dennett[*], Paul Norman[†], Nicola Shelton[‡], Rachel Stuchbury[§]

Research Material

Abstract

BACKGROUND

In the United Kingdom, there exist three Census-based longitudinal datasets, known collectively as the Longitudinal Studies. The England and Wales Longitudinal Study (LS) is a 1% sample of the population of England and Wales It started with a sample from the 1971 Census and links the records of individuals (and other members of the sample member's household) from the last five censuses. There are around 500,000 individuals present at each census, although not all of these will be linked for the full time-series. The LS also records key life events such as births, deaths, marriages and cancer registrations. Similar datasets of linked Census records exist for Scotland and Northern Ireland, although these data have a shorter time-series (currently linked back to 1991) and larger sample fractions (which constitute around 5% and 28% of their populations respectively at each Census). Whilst immensely valuable datasets for demographic research in the UK, all of the Longitudinal Studies are under-used when compared to other Census data products. Part of the reason for this is the restricted access researchers have to the microdata due to the potentially disclosive detail contained within. Consequently, in order to introduce potential researchers to the data and increase the user-base, a synthetic general-use version of the Longitudinal Studies is proposed.

OBJECTIVE

This paper details a simple and reproducible method for generating a general use synthetic Longitudinal Study-like dataset from pre-existing Census microdata and non-disclosive outputs from the real Longitudinal datasets, using the England and Wales Longitudinal Study as the case example.

[*] Centre for Advanced Spatial Analysis, University College London – a.dennett@ucl.ac.uk

[†] School of Geography, University of Leeds – p.d.norman@leeds.ac.uk

[‡] Department of Epidemiology and Public Health, University College London – n.shelton@ucl.ac.uk

[§] Department of Epidemiology and Public Health, University College London – r.stuchbury@ucl.ac.uk

METHODS

The new dataset will be known as the synthetic LS 'spine' dataset as it will only include transitions of key demographic variables included in the national LSs. It is generated using the 2011 England and Wales Teaching SAR (Samples of Anonymised Records) dataset, available from the Office for National Statistics and a series of 2011 back to 2001 transitional probabilities taken from the England and Wales LS. A series of algorithms, written in R, are used to firstly estimate the numbers of individuals in particular age groups undergoing each longitudinal state transition and then allocate transitions to the appropriate number of pre-exiting individuals in the SAR micro-dataset, resulting in a new, plausible, LS-like dataset.

RESULTS

The England and Wales synthetic LS spine dataset augments the original 'Teaching SAR' dataset by, firstly, estimating single year of age values from broad age groups contained in the original. New ten year age groups are re-estimated from the new single year of age variables and ten-year transitions between 2011 and 2001 for categories of General Health (15 transitions), Marital Status (25 transitions), Religion (5 transitions), Approximate Social Grade (16 transitions) as well as estimates of the number of live births to females over the ten year period (4 categories) and a variable estimating those individuals who may have died over the period, based on age and general health status.

CONCLUSION

The method detailed here using the England and Wales case can be used to apply longitudinal transitions to all similar micro-datasets in the UK, and indeed elsewhere, where there might be a need to introduce people to the longitudinal microdata and its unique temporal transitions for individuals, but where access to these data are frequently restricted due to the need to protect the confidentiality of individuals in the study.

The constraints imposed our own access to the data from the original LS dataset, in terms of being unable to remove small cell counts (those less than 10), from the secure microdata laboratory, mean that we are unable to account for some of the more nuanced interactions between variables – for example the interactions between general health and marital status as well as age. However, having chosen age as the key interacting variable with all others, we believe that we are able to generate a plausible longitudinal population relatively cost (manpower costs in terms of programming and computational costs in terms of processing time) effectively.

The synthetic LS dataset for England and Wales which we have devised can be used by academics to train students in longitudinal methods and for researchers wishing to familiarise themselves with a range of variables in the census-based LSs. Any outputs which people generate will, in the main, provide a similar picture to results obtained using the original LS data but do not themselves have any research utility nor represent individuals in reality.

1 Introduction

The United Kingdom is home to three Census-based longitudinal datasets, known collectively as the Longitudinal Studies. In England and Wales, the ONS Longitudinal Study (LS) comprises around a 1% sample of Census records (some 500,000 people at each Census), linking the full decennial Census returns of individuals (and members of their household) back to 1971. As well ask linking Census variables, key life events such as births (to members of the study), deaths and cancer registrations between censuses are also included. The Scottish and Northern Irish Longitudinal Studies (SLS and NILS respectively) have a shorter time-series than the LS (currently linked back to 1991) but have larger sample fractions (comprising around 5% and 28% of their populations respectively). Both the SLS and NILS feature linkage to additional health data not currently available in the LS. Collectively, however, these datasets comprise some of the richest demographic data available for research in the UK – and indeed the world.

However, usage of the LSs is relatively low compared to other Census data products, with active projects using the data counted in double figures on an annual basis, whereas downloads of some of the more readily accessible census data products counted in the tens and even hundreds of thousands. Part of this under-usage it is almost certainly down to access restrictions; and part must be attributed to the nature and complexity of the data; but these two factors combined mean that it is very rare for undergraduates and those making their early steps in social science research careers to have been exposed to Longitudinal Study data at all.

To address this problem, the Synthetic Data for the Longitudinal Studies (SYLLS) project has been funded by the Economic and Social Research Council, with two main aims: the first to introduce new scholars to Longitudinal Study-like data earlier in their careers by creating new, general usage synthetic teaching datasets, enabling plausible longitudinal analysis of LS-like microdata to be carried out for certain key variables. The teaching datasets are not designed with accuracy in mind (although counts of individuals undergoing particular transitions should not be wildly inaccurate), but rather as a pedagogic tool to introduce novices to longitudinal data. The second; to devise a new modelling methodology to allow requests for particular customised data requests from the Longitudinal Studies to be synthesised, so that microdata which look very much like the original (with accurate frequency distributions and modelling coefficients) can be released from the data custodians for use outside of the safe settings (where all research must presently be carried out), without fear of data disclosure and the compromise of confidentiality. The second of these aims has been achieved through the synthpop software developed in the R language and is reported elsewhere (Nowok, Raab, and Dibben 2014; Raab, Nowok, and Dibben 2015). This paper reports on the methodology employed to achieve the first of these aims – the creation of general usage teaching datasets.

While general usage synthetic data will be created for each of the national Longitudinal Studies, here we report only on the first dataset – that generated for the England and Wales ONS LS. The methodology, however, will be identical for the ONS LS, SLS and NILS. Henceforth we will refer to this new synthetic ONS LS dataset as the ‘Synthetic Spine’ data.

2 Input Data

The core data used to create the new England and Wales LS Synthetic Spine is the Office for National Statistics Microdata Teaching File[5]. This data file is freely available under the Open Government Licence and comprises a random 1% sample of the full 2011 Census output database for England and Wales. While derived from the more detailed Samples of Anonymised Records (SAR) data, the variables within the Teaching File have been aggregated to broad categories to minimise the likelihood of disclosive detail being released, with record swapping and mixing also used to add further uncertainty to unusual records. The Microdata Teaching File contains records for 569,741 individuals across 17 variables.

Using the Microdata Teaching File has a number of advantages for this project. Firstly, the 1% sample size is the same as that for the 2011 LS sample, meaning that realistic numbers of individuals will be undergoing longitudinal transitions. Secondly, being available under the Open Government Licence means that the newly derived LS spine dataset should also be available under the same terms, as long as any data extracted from the original LS dataset are also cleared for public release.

In order to estimate the numbers of individuals undergoing the various transitions we wished to capture, data from the England and Wales ONS Longitudinal Study[6] are also used. The data are for 4 key longitudinal state transitions from 2011 back to 2001, chosen for three main reasons:

1) The variables are of key research interest to social scientists (examples of recent longitudinal research using each of the variables selected for inclusion are given before the definition of each, below).
2) Equivalent or similar variables exist in both 2011 and 2001 so that a meaningful transition can be generated
3) Variable disaggregation does not lead to an overabundance of table cell values under 10 (the threshold for final output public access clearance for LS data).

Transitional matrices consisting of counts of individuals undergoing each transition of interest in 9 ten-year age groups (age in 2011 with groups 10-19, 20-21, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99) are generated for each variable of interest within the England and Wales LS. Age group 0-9 was not included as these individuals would not have been alive in 2001. Additionally, an age group consisting of those older than 100 years was not included as very few individuals in the dataset are aged 100 and over.

2.1 Transitional Variables Included

General Health

A number of health-related research projected have been undertaken using the national LSs in recent years, including Boyle (2010) and Dykstra et al. (2009).

The General Health variable has 5 categories in 2011:

1. Very good health

[5] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/census-microdata/microdata-teaching-file/index.html

[6] http://www.ucl.ac.uk/celsius

2. Good health
3. Fair health
4. Bad health
5. Very bad health

And 3 categories in 2001:

1. Good health
2. Fair health
3. Bad health

The transition matrix included all ($5 \times 3 = 15$) transitions between the 2011 and 2001 variables.

Marital Status

Marital Status was the focus of longitudinal research carried out by Feng et al. (2010).

The Marital Status variable has been aggregated to 5 categories in 2011 which are comparable to the 2001 categories. These are:

1. Single (never married or never registered a same-sex civil partnership)
2. Married or in a registered same-sex civil partnership
3. Separated but still legally married or separated but still legally in a same-sex civil partnership
4. Divorced or formerly in a same-sex civil partnership which is now legally dissolved
5. Widowed or surviving partner from a same-sex civil partnership

Whilst it should not be possible for transitions between every state in 2001 and every state in 2011 to occur (for example, it should not be possible for someone to be married in 2001 and single in 2011 given the separated, divorced or widowed category), in practice, errors in coding or census form completion by individuals means that improbable transitions can occur. Therefore, a full $5 \times 5 = 25$ matrix of transitions between 2011 and 2001 is used.

Religion

Religion has been the focus of a number of pieces of longitudinal research including Platt et al. (2014) and Simpson, Jivraj, and Warren (2014)

While Religion has 9 categories (6 main religions in the UK, plus no religion, other religion and not stated) in the Microdata Teaching File dataset, a full $9 \times 9 = 81$ interaction matrix resulted in far too many small cell counts where, for example, transitions between the Muslim and Jewish faith are very small. Therefore we aggregated religion to the following transition matrix:

2001	2011
No Religion	No Religion
No Religion	Religion
Religion	Same Religion
Religion	Different Religion
Religion	No Religion

Approximated Social Grade

Social status and grade in another popular topic of longitudinal research, recent work including that by Champion, Coombes, and Gordon (2014) and Dini (2010).

Approximated Social Grade (derived from occupation, employment status qualification, tenure and whether working full or part-time variables[7]) has 4 categories in the Microdata Teaching File dataset:

1. AB – Upper Middle Class and Middle Class, Higher & intermediate managerial, administrative, professional occupations
2. C1 – Lower Middle Class; Supervisory or clerical and junior managerial, administrative or professional
3. C2 – Skilled Working Class; Skilled manual workers
4. DE – Working Class and non-working; Semi and unskilled manual workers, casual or lowest grade workers, pensioners, and others who depend on the welfare state for their income

Whilst these codes exist in both the 2011 and 2001 censuses, the variables and algorithms used to generate the social grade classification differ slightly, so the variables are not directly comparable. However, given that we are modelling state transitions, this does not matter a great deal. Consequently we use a full $4 \times 4 = 16$ matrix of transitions between 2011 and 2001.

Live Births

Births to LS members allow for the estimation of fertility, with a number of pieces of research focusing on fertility using LS data, including Grundy (2009) and Robards, Berrington, and Hinde (2011).

The births transition differs from other variables in that it is a motherhood transition and therefore only applies to the females in the dataset who give birth during the 10 year period. LS members who are already mothers but do not give birth during the transition period are not included. Live births to LS women between 2001 and 2011 are simply 4 categories of counts of 0, 1, 2 or 3 and more births.

Deaths

Mortality is a recurring theme in research using the longitudinal studies, with many studies using mortality as a key outcome. Work by Scott and Timæus (2013) is just one example of this.

Counts of those who died between 2001 and 2011 by age, but also general health status are included in order that a death estimate variable could be generated. The matrix consisted of the dichotomous alive/dead in 2011 variable by the three 2001 general health categories by the same 9 age groups as for all other variables.

[7] The model used to allocated Approximated Social Grade by the Market Research Society Census and Demographics group, with full details of their methodology available here: https://www.mrs.org.uk/pdf/Social%20Grade%20Allocation%20for%202011%20Census.pdf

2.2 Transitional variables not included

Of course an argument could be made for the inclusion of a number of other transitional variables, particularly migration (with longitudinal research such as that by Champion (2012) and Riva, Curtis, and Norman (2011)) and residence type. Time constraints relating to the delivery of the final data, however, have meant that for the present iteration, transitions are limited to those variables outlined above. Future

3 Method

The method we employ is, at its core, a simple one-dimensional proportional fitting exercise making it somewhat more straightforward than the multi-dimensional iterative proportional fitting first proposed by Deming and Stephan (1940) and exemplified in a social science context by Norman (1999) and Simpson and Tranmer (2005). It has been necessary to avoid multi-dimensional variable interactions due to the small cell counts that would occur in the transition matrices extracted from the raw Longitudinal Study data and the very practical constraint that we would not have been able to remove these small cell counts from the virtual microdata laboratory due to their disclosive potential. We could, of course, have estimated these variable interaction transitions, but given the pedagogic purpose of the dataset and the time constraints imposed by earlier unforeseen problems in the project, simple transitions are preferred. While the proportional fitting element of the exercise is relatively straightforward, a challenge still remains in allocating longitudinal transitions, for the variables described above, to appropriate individuals within the 2011 Microdata Teaching File. In order to do this, we have chosen age as our constraining variable – all transitions will be accurate when aggregated to age, although not necessarily when aggregated to another variable such as geographic region. The rationale for selecting age rather than any other common predictor of transitions is that it is almost certainly a better predictor of transitions between one variable state and another, than any other single variable. Furthermore, by constraining to age, we are likely to capture covariate interactions (such as transitions to ill health or marital status) in the process, resulting in (as will be shown at the end of this paper) plausible multi-variate transitions (marriage transitions by health status, for example) without attempting to achieve these explicitly in the estimation process. In addition, by using age in 10 year age groups, the task of estimating transitions over a 10 year period is made much simpler.

3.1 Estimating 10 year age groups

The first problem to overcome is that in the SAR Microdata Teaching File, age is recorded for 8 uneven age groups (indices indicative of those used in the data):

1. 0-15
2. 16-24
3. 25-34
4. 35-44
5. 45-54
6. 55-64
7. 65-74
8. 75 and over

These groups need to be re-estimated so that we have 11 even 10 year groups:

0. 0-9
1. 10-19
2. 20-29
3. 30-39
4. 40-49
5. 50-59
6. 60-69
7. 70-79
8. 80-89
9. 90-99
10. 100+

The rationale for this is simple – in the lowest age group, those aged 0-5 would still be in the same group 10 years later, thus making the 10 year transition difficult to calculate. Where all age groups are 10 years, then calculating a 10 year transition becomes more straightforward. To carry out the re-estimation to new groups, the single year of age for each person in each original age group needs to be estimated before they can then be allocated a new broad age group. To estimate the single year of age for each of the 569,741 individuals in the dataset, we use data on single year of age for each UK region from the 2011 Census aggregate tables[8]. These Census tables can be aggregated into any age group required and the relative proportions each single age comprises in each group calculated. In doing this, single year of age counts are disaggregated by region as well. This is needed due to the large differences in the proportion of the population in each age group in London compared to all other regions in England and Wales.

The total number of individuals of single year of age a in region r will be a fraction of the total number of individuals in age group A in region r:

$$a^r \in A^r$$

Such that:

$$\sum_{a^r=1}^{n} = A^r$$

and

$$\sum_{a^r=1}^{n} \frac{a^r}{A^r} = 1$$

By calculating all proportions of $\frac{a^r}{A^r}$ for each age group A^r using the Census aggregate tables single year of age file, it is possible to decompose and re-estimate age group data as required. The method and R code to achieve this is detailed below in seven stages:

[8] Office for National Statistics, 2011 Census: Aggregate data (England and Wales) [computer file]. UK Data Service Census Support. Downloaded from: http://infuse.mimas.ac.uk . This data is licensed under the terms of the Open Government Licence [http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2]

Stage 1 – Calculate the numbers of people in each age group for each region in the Microdata Teaching File

First load into memory (‘library’) some packages we will use for data manipulation and then read in the data:

```
library(plyr)
library(reshape2)

SAR2011<-read.csv("../SAR Data/rft-teaching-file/2011 Census
Microdata Teaching File.csv")
```

Calculate the numbers of people in each age group for each region in the SAR Data and store in a new data frame – this will be used in a later stage in the process.

```
SARAgeTotals<-ddply(SAR2011, .(Region, Age), nrow)

#cast back into a square data frame matrix

SARAgeTotalsDF<-dcast(SARAgeTotals,Age~Region)
```

Table 1 – Example of the SARAgeTotalsDF data frame

Age group	**E12000001**	**E12000002**	**E12000003**	**E12000004**	**…**
1	4771	13266	9939	8474	…
2	3425	9339	7164	5819	…
3	3181	8999	6729	5309	…
4	3414	9567	7313	6195	…
5	3847	9862	7346	6608	…
6	3361	8569	6255	5648	…
…	…	…	…	…	…

Stage 2 – Read in single year of age data and aggregate to the 8 age groups in the Microdata Teaching File

```
SingleYearAge<-read.csv("../SAR Data/AgeByRegion/SingleYearAge.csv")
```

Table 2 – Example of the first rows and columns in the SingleYearAge data frame

Single Year of Age	**E12000001**	**E12000002**	**E12000003**	**E12000004**	**E12000005**
Age under 1	30606	87370	65950	54274	71547
Age 1	30202	86767	66738	54017	70778
Age 2	29686	86133	65854	54168	71508
Age 3	29934	86913	65663	54533	71329
Age 4	29415	84908	64242	53182	69639
Age 5	29041	83620	63577	52394	69048
Age 6	27545	79699	60059	49603	66434
Age 7	27103	77789	59182	49537	65326
Age 8	26256	75408	57364	47884	63113

Create a vector containing the break values for each of the 8 uneven age groups and then use this vector to add the age group value to the date file in a new column called ‘AgeGrp’

```
AgeGrpVector<-c(16,25,35,45,55,65,75,101)

#Use the following loop to the add values for the uneven age
#groups to the single year of age file

i=0
j=1
counter=0
for (i in 0:nrow(SingleYearAge)){
  if (counter<=AgeGrpVector[j]){
    SingleYearAge[counter,"AgeGrp"]=j
    counter=counter+1
  } else {
    j=j+1
    next
  }
}
```

Now calculate the total number of people in each of the 8 age groups by aggregating the SingleYearAge data frame into a new AgeGroups data frame using the new AgeGrp variable:

```
AgeGroups<-ddply(SingleYearAge,.(AgeGrp),summarise,E12000001=sum(E12000001),E12000002=sum(E12000002),E12000003=sum(E12000003),E12000004=sum(E12000004),E12000005=sum(E12000005),E12000006=sum(E12000006),E12000007=sum(E12000007),E12000008=sum(E12000008),E12000009=sum(E12000009),W92000004=sum(W92000004))
```

Table 3 – Example of the AgeGroups data frame

AgeGrp	E12000001	E12000002	E12000003	E12000004	…
1	462437	1324548	997792	838455	…
2	322208	858593	665550	547411	…
3	315455	896267	668632	546384	…
4	340381	964851	720793	627735	…
5	377399	987491	725788	638957	…
6	329521	849272	630607	561332	…
…	…	…	…	…	…

Stage 3 – Calculate the proportion of each age group that each single year of age comprises

Create two copies of the SingleYearAge data frame – one reordered so that the ‘AgeGroup’ variable is the new first column and a copy of this to put the new proportions into.

```
SingleYearProps2<-SingleYearAge[c(12,2,3,4,5,6,7,8,9,10,11)]
SingleYearEmpty<-SingleYearAge[c(12,2,3,4,5,6,7,8,9,10,11)]
```

To calculate the proportion of each age group that each single year of age comprises, loop through using the ‘AgeGroup’ column to calculate each proportion in turn.

```
i=1
j=1
counter=1
for (i in 1:nrow(SingleYearProps2)){
  if (SingleYearProps2[i,1]==AgeGroups[j,1]){
    SingleYearEmpty[i,2:11]<-
SingleYearProps2[i,2:11]/AgeGroups[j,2:11]
    print(paste0("i=",i))
  }else{
    SingleYearEmpty[i,2:11]<-
SingleYearProps2[i,2:11]/AgeGroups[j+1,2:11]
    j=j+1
    print(paste0("j=",j))
    #i=i-1
    #print(paste0("i=",i))
    next
  }
}
```

Add the single year of age variable back in and re-order the data frames

```
SingleYearEmpty$row.names<-SingleYearAge$X
SingleYearProps3<-SingleYearEmpty[c(12,1,2,3,4,5,6,7,8,9,10,11)]
```

Stage 4 – Using the proportions of each age group that each single year of age comprises from the aggregate Census data, calculate how many individuals in the Microdata Teaching File should have each single year of age.

Create some copies of the single year of age data frames to store the new data in and then fill them with new data. Integers are needed but rounding may cause errors, so create two data frames, one rounded and one not – these will be used to assess the level of error in the next stage.

```
#create some dummy data frames to store data in

NumbersInAge<-SingleYearProps3[,-1]
NumbersInAgeRound<-SingleYearProps3[,-1]
Temp<-NumbersInAge
Temp<-NumbersInAgeRound

#now fill this data frame with estimated numbers of people in each
#age group, rounded

i=1
j=1
counter=1
for (i in 1:nrow(Temp)){
  if(Temp[i,1]==SARAgeTotalsDF[j,1]){
    NumbersInAgeRound[i,2:11]<-
round(Temp[i,2:11]*SARAgeTotalsDF[j,2:11],0)
    print(paste0("i=",i))
  }else{
    NumbersInAgeRound[i,2:11]<-
round(Temp[i,2:11]*SARAgeTotalsDF[j+1,2:11],0)
    j=j+1
    print(paste0("j=",j))
    next
  }
}

#now fill another data frame will estimated numbers of people in
#each age group, not rounded

i=1
j=1
counter=1
for (i in 1:nrow(Temp)){
  if(Temp[i,1]==SARAgeTotalsDF[j,1]){
    NumbersInAge[i,2:11]<-Temp[i,2:11]*SARAgeTotalsDF[j,2:11]
    print(paste0("i=",i))
  }else{
    NumbersInAge[i,2:11]<-Temp[i,2:11]*SARAgeTotalsDF[j+1,2:11]
    j=j+1
    print(paste0("j=",j))
    next
  }
}

#add an ID column

NumbersInAgeRound$ID<-seq(0,nrow(NumbersInAgeRound)-1)
NumbersInAge$ID<-seq(0,nrow(NumbersInAge)-1)
```

Stage 5 – Compare the rounded to the un-rounded file by generating an error matrix. Use this error matrix to then adjust the numbers in the data frame containing the rounded estimates for the number of individuals in the Microdata Teaching File at each age group

```
#check that everything adds up

NumbAgeAgg<-
ddply(NumbersInAge,.(AgeGrp),summarise,E12000001=sum(E12000001),E120
00002=sum(E12000002),E12000003=sum(E12000003),E12000004=sum(E1200000
4),E12000005=sum(E12000005),E12000006=sum(E12000006),E12000007=sum(E
12000007),E12000008=sum(E12000008),E12000009=sum(E12000009),W9200000
4=sum(W92000004))

NumbAgeRoundAgg<-
ddply(NumbersInAgeRound,.(AgeGrp),summarise,E12000001=sum(E12000001)
,E12000002=sum(E12000002),E12000003=sum(E12000003),E12000004=sum(E12
000004),E12000005=sum(E12000005),E12000006=sum(E12000006),E12000007=
sum(E12000007),E12000008=sum(E12000008),E12000009=sum(E12000009),W92
000004=sum(W92000004))

#it will not add up due to rounding errors, so generate a rounding
#error matrix
```

```
ErrorMatrix<-NumbAgeAgg-NumbAgeRoundAgg

#use the rounding error matrix to add or remove values from
#the last single age in each age group using the AgeGrpVector from
#before - crude, but it at least gets us to kind of where we want to
#be...

AgeGrpVector<-c(16,25,35,45,55,65,75,101)

i=1
j=1
k=1
for (i in 1:length(AgeGrpVector)){
  k=AgeGrpVector[i]
  NumbersInAgeRound[k,2:11]<-
NumbersInAgeRound[k,2:11]+ErrorMatrix[j,2:11]
  j=j+1
}
```

Table 4 – Example of the NumbersInAgeRound data frame

Age	**AgeGrp**	**E12000001**	**E12000002**	**E12000003**	**E12000004**	**…**
1	**1**	316	875	657	549	…
2	**1**	312	869	665	546	…
3	**1**	306	863	656	547	…
4	**1**	309	870	654	551	…
5	**1**	303	850	640	537	…
…	**…**	…	…	…	…	…

Stage 6 – Subset the Microdata Teaching File data frame (SAR2011) into separate regions and then add single year of age values to each individual using an algorithm that simply counts to each value in the NumbersInAgeRound data frame, adding values until the number is reached and then moves on to the new value. When complete, recombine everything back into a single data frame.

```
#first create a region list from the unique values in SAR2011

regionlist<-as.vector(unique(SAR2011$Region))

#this loop then subsets the data frame by regions and stores each
#subset in a new data frame and stores each data frame in a list

i=1
dflist<-list()
for (i in 1:length(regionlist)){
  region<-regionlist[i]
  print(region)
  #assign each dataframe to a list
  dflist[[i]]<-
assign(paste0("SARsub_",region),subset(SAR2011,SAR2011$Region==regio
n))
}

#cycle through the dataframes in list and add single year of age
#values in the appropriate places

dataframe=1
for (dataframe in 1:length(dflist)){
  dfname<-regionlist[dataframe]
  df<-as.data.frame(dflist[dataframe])
  i=1
  j=1
  counter=1
  for (i in 1:nrow(df)){
    if (counter<=NumbersInAgeRound[j,dfname]){
      df[i,"AgeSingle"]<-NumbersInAgeRound[j,"ID"]
      counter=counter+1
    }else{
      df[i,"AgeSingle"]<-NumbersInAgeRound[j+1,"ID"]
      counter=2
      j=j+1
    }
  }
  print(paste0("Finished estimation for region ",dfname))
  assign(paste0("SARsub",dfname),df)
}

#recombine everything back into the original data frame
SAR2011<-
rbind(SARsubE12000001,SARsubE12000002,SARsubE12000003,SARsubE1200000
4,SARsubE12000005,SARsubE12000006,SARsubE12000007,SARsubE12000008,SA
RsubE12000009,SARsubW92000004)
```

Stage 7 – Now a new single year of age has been estimated for each individual in the Microdata Teaching File, it is an elementary task to recode these single years of age into new, 10 year age groups.

An efficient way to recode single year of age into 10 year groups is using a custom function:

```
newvar<-0
recode_agegroups<-function(original_age_variable){
  newvar[original_age_variable >=100]<-10
  newvar[original_age_variable >=90 & original_age_variable <=99]<-9
  newvar[original_age_variable >=80 & original_age_variable <=89]<-8
  newvar[original_age_variable >=70 & original_age_variable <=79]<-7
  newvar[original_age_variable >=60 & original_age_variable <=69]<-6
  newvar[original_age_variable >=50 & original_age_variable <=59]<-5
  newvar[original_age_variable >=40 & original_age_variable <=49]<-4
  newvar[original_age_variable >=30 & original_age_variable <=39]<-3
  newvar[original_age_variable >=20 & original_age_variable <=29]<-2
  newvar[original_age_variable >=10 & original_age_variable <=19]<-1
  newvar[original_age_variable >=0 & original_age_variable <=9]<-0
  return(newvar)
}
```

This function can now be used to generate the new variable:

```
SAR2011$Age10YrGrp<-recode_agegroups(SAR2011$AgeSingle)
```

3.1.1 Drawbacks of the 10 year age estimation process

The 10 year age estimation process will not generate absolutely accurate estimates. The main source of error is that in the first original uneven age group – 0-15, there are some individuals that should, in the main, not have their single year of age estimated as below the age of 5 – these are students. In the original Microdata Teaching File, there is a binary variable for student: 1=Student, 2=Non-Student. In this context a student is anyone who is either a school child or full time student. Given that children do not start school until the age of 5 in England and Wales and all children between the ages of 5 and 16 should be attending school, assuming an even spread of ages across the age group, we would expect that for the age group 0-15, roughly one third should be recorded as non-students and two-thirds as students. In reality this is not entirely the case as some young children attending nursery or private boarding schools will be included as students, however the majority will not. The algorithm used to estimate single year of age did not select out the under-5s and classify as non-students, so around 23,000 individuals (around ¼ of the age group) may be misclassified. However, in the final release of the data, the single year of age variable is removed in order that the perception of potentially disclosive data being released is avoided. In this situation where only the new ten year age group remains in the data, then the errors associated with misallocation of students to the youngest single years of age are minimised through the aggregation.

3.2 Estimating individuals undergoing longitudinal transitions for each of the variables – a general methodology

The estimation of each longitudinal variable transition is carried out in almost exactly the same way (minor variations will be detailed later). Below the series of stages in this process are detailed using Approximated Social Grade as the exemplar.

Stage 1 – Generating Transitional Probability Matrices

Transitional matrices of the same format are generated for each variable of interest from ONS Longitudinal Study, broadly comparable to the example table below (Table 5) which shows the transitional counts for the Approximated Social Grade variable.

Table 5 – Transitional counts between states of Approximated Social Grade, 2011-2001

2011	2001	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	**1**	0	200	4968	8608	7590	5805	3329	1092	0
1	**2**	0	1033	5456	4954	4200	3682	1775	417	0
1	**3**	0	183	808	903	695	505	309	87	0
1	**4**	0	3003	2250	1089	839	1068	854	232	0
2	**1**	11	236	2769	5450	5062	4128	2152	725	0
2	**2**	29	1865	8969	11866	10954	9233	5668	2090	0
2	**3**	10	422	1625	1850	1541	1449	843	218	0
2	**4**	60	4732	4159	3722	2828	2785	2227	821	0
3	**1**	0	83	669	1303	1260	999	441	114	0
3	**2**	0	556	2031	3295	3054	2884	1525	399	0

3	**3**	0	805	3107	5318	5225	4029	2478	605	0
3	**4**	0	2105	3464	4933	4255	3934	2294	591	0
4	**1**	0	95	538	1003	1076	1439	943	313	0
4	**2**	10	690	2162	3072	3115	3725	2309	616	0
4	**3**	0	411	1403	2281	2452	2844	2262	715	0
4	**4**	11	3379	6723	10162	9864	10433	9347	3531	11

Source: ONS Longitudinal Study

As 2011 is our base population, transitional probabilities are calculated from the counts of transitions with each 2001 state calculated as a proportion of the corresponding 2011 state in turn. Table 6 exemplifies this more clearly:

Table 6 – Transitional probabilities for Approximated Social Grade, 2011-2001

2011	2001	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	**1**	0	0.045	0.368	0.553	0.570	0.525	0.531	0.597	0
1	**2**	0	0.234	0.405	0.319	0.315	0.333	0.283	0.228	0
1	**3**	0	0.041	0.060	0.058	0.052	0.046	0.049	0.048	0
1	**4**	0	0.680	0.167	0.070	0.063	0.097	0.136	0.127	0
2	**1**	0.1	0.033	0.158	0.238	0.248	0.235	0.198	0.188	0
2	**2**	0.264	0.257	0.512	0.518	0.537	0.525	0.520	0.542	0
2	**3**	0.091	0.058	0.093	0.081	0.076	0.082	0.077	0.057	0
2	**4**	0.545	0.652	0.237	0.163	0.139	0.158	0.204	0.213	0
3	**1**	0	0.023	0.072	0.088	0.091	0.084	0.065	0.067	0
3	**2**	0	0.157	0.219	0.222	0.221	0.243	0.226	0.233	0
3	**3**	0	0.227	0.335	0.358	0.379	0.340	0.368	0.354	0
3	**4**	0	0.593	0.374	0.332	0.308	0.332	0.340	0.346	0
4	**1**	0	0.021	0.050	0.061	0.065	0.078	0.063	0.060	0
4	**2**	0.476	0.151	0.200	0.186	0.189	0.202	0.155	0.119	0
4	**3**	0	0.090	0.130	0.138	0.149	0.154	0.152	0.138	0
4	**4**	0.524	0.739	0.621	0.615	0.598	0.566	0.629	0.682	1

Source: ONS Longitudinal Study

Taking the first row of Table 5 (Transitions between social grade 1 (AB) in 2011 and social grade 1 in 2001), we can observe that at age group 20-29 (2011 age group), 200 individuals in the LS underwent that transition. Table 6 shows that this is a proportion of 0.045 (4.5%) of all people of social grade 1 at age group 20-29 in 2011 (200/(200+1033+183+3003)=0.045). For each 2011 variable, all 2001 state proportions at each age group will sum to 1. Similar transitional probability tables are generated for each of the variables outlined in section 2.

Stage 2 – Apply transitional probabilities to Microdata Teaching File data to create estimates of the total number of people undergoing each transition

The following code could be generalised to work for each variable in exactly the same way and indeed was used virtually identically for the estimation of each variable in the new synthetic dataset.

Firstly read in the transitional probability table (Table 6). For Social Grade, this is known as ‘Social.csv’. Then calculate some totals for social grade from the Microdata Teaching File

```
####################################################################
####
#Social Grade
```

```
#before going any further create a pseuso ID column for reordering
#data later on in the estimation process...

SAR2011$Pseudo_ID<-seq(1,nrow(SAR2011))

##*notes
#For social grade we have some difficult transitions in the 10-19 &
#90-99 age groups - therefore I've borrowed the transitional probs
#from the neighbouring age groups - 10-19 is the same as 20-29 and
#90-99 is the same as 80-89...

social_trans <-read.csv("Social.csv", header=TRUE)
social_trans_totals <- social_trans
social_trans_totals_round <- social_trans

##calculate the social status totals
SARSocialTotals<-ddply(SAR2011, .(Age10YrGrp,
Approximated.Social.Grade), nrow)

##cast back into a square matrix
SARSocialTotalsDF<-
dcast(SARSocialTotals,Approximated.Social.Grade~Age10YrGrp)
SARSocialTotalsDF<-SARSocialTotalsDF[2:5,c(1,3:11)]
```

Next estimate the total number of people (un-rounded and rounded) undergoing each transition, using the transition matrix and calculate the error between the rounded and the un-rounded estimates.

```
#calculate the numbers of people undergoing each social transition

i=1
k=1
for(i in 1:nrow(social_trans)){
  k<-social_trans[i,1]
  social_trans_totals[i,3:11]<-
social_trans[i,3:11]*SARSocialTotalsDF[k,2:10]
}

i=1
k=1
for(i in 1:nrow(social_trans)){
  k<-social_trans[i,1]
  social_trans_totals_round[i,3:11]<-
round(social_trans[i,3:11]*SARSocialTotalsDF[k,2:10],0)
}

#aggregate back into a matrix the same shape as SARSocialTotals
social_trans_agg<-
ddply(social_trans_totals_round,.(Social_2011),summarise,X1=sum(X1),
X2=sum(X2),X3=sum(X3),X4=sum(X4),X5=sum(X5),X6=sum(X6),X7=sum(X7),X8
=sum(X8),X9=sum(X9))


######now generate an error matrix
ErrorMatrix<-SARSocialTotalsDF-social_trans_agg
```

Next use the error matrix to alter the totals in the rounded estimate so that everything adds up to the expected total. To do this, first create a vector with the break values for each new 2011 group in the transitional probabilities file.

```
SocialGrpVector<-c(1,5,9,13)

#now use the error matrix to update social_trans_totals_round
i=1
j=1
k=1
for (i in 1:length(SocialGrpVector)){
  k=SocialGrpVector[i]
  social_trans_totals_round[k,3:11]<-
social_trans_totals_round[k,3:11]+ErrorMatrix[j,2:10]
  j=j+1
}

#check that you've updated correctly - error matrix should be full
of zeros
social_trans_agg1<-
ddply(social_trans_totals_round,.(Social_2011),summarise,X1=sum(X1),
X2=sum(X2),X3=sum(X3),X4=sum(X4),X5=sum(X5),X6=sum(X6),X7=sum(X7),X8
=sum(X8),X9=sum(X9))
ErrorMatrix<-SARSocialTotalsDF-social_trans_agg1

#if there is a little error in there - edit manually.
social_trans_totals_round<-edit(social_trans_totals_round)
```

Stage 3 – Use the estimates of the total number of people undergoing each transition to update the Microdata Teaching File with expected transitions for the correct number of people.

The next stage is important as it both allows us to deal with missing data (-9 values) in the original file and makes sure that the estimates do not have strange regional allocations. In the original Microdata Teaching File, data are arranged in regional order, and so any algorithmic allocation of individuals undergoing each transition row-by-row in the data (as used here) would result in odd regional distributions. By splitting the Microdata Teaching File into 10 year age groups to estimate each decennial transition in turn and then randomising the cases in the subset, we can protect against this error. The randomisation in this instance is carried out using the `sample()` base function in R.

```
#first create an age group list from the unique values in SAR2011

AgeGrpList<-as.vector(unique(SAR2011$Age10YrGrp))

#divide SAR2011 up into age groups, randomise and put -9s at the top
i=1
#create an empty list to store all of the elements of the data frame
#in
AgeGrpDFList<-list()

for (i in 1:length(AgeGrpList)){
  AgeGrp<-AgeGrpList[i]
  print(AgeGrp)
  #create the age group subset
  temp_df<-subset(SAR2011,SAR2011$Age10YrGrp==AgeGrp)
  #ramdomise the order of the subset
  temp_df<-temp_df[sample(1:nrow(temp_df)),]
  #now put the missing data (-9) values at the top of the file
  temp_df<-arrange(temp_df,Approximated.Social.Grade)
  AgeGrpDFList[[i]]<-assign(paste0("Age10YrGrp_",AgeGrp),temp_df)
}
```

Now the data have been subset by age group and randomised, the next stage is to use the estimates of the total number of people undergoing each transition to update the subsets with transitional values. The algorithm devised is sensitive to zeros contained in the transitional estimates, therefore to deal with zeros a custom function is implemented within the algorithm to identify rows where zeros may exist and inform the algorithm to skip to the next non-zero value. The algorithm is also designed to ignore the first and last age group subsets. The first age group subset is ignored as those aged 0-9 in 2011 will not have been born in 2001. The final subset is ignored as the numbers of individuals aged over 100 in 2011 that have undergone a marital, health, religious or social grade transition is deemed negligible.

```
################################################################
#a function to create a vector identifying the rows of the matrix
#with non-zeros data in

indexVectorCreator<-function(transitionMatrix,column){
  index<-vector()
  i=1
  for (i in 1:nrow(transitionMatrix)){
    if (transitionMatrix[i,column]!=0){
      index[[i]]<-i
    } else {
      next
    }
  }
  index<-index[!is.na(index)]
  return(index)
}
#################################################################
#Estimation loop
############

dataframe=1
for (dataframe in 3:length(AgeGrpDFList)-1){
  #create a little index vector so that the programme knows to
  #ignore 0s in the loop
  dfname<-AgeGrpList[dataframe]
  df<-as.data.frame(AgeGrpDFList[dataframe])
  i=1
  j=1
  counter=1
  for (i in 1:nrow(df)){
    if (df[i,"Approximated.Social.Grade"]==-9){
      df[i,"Social2001"]<--9
      print(i)
      next
    } else {
      print(paste0("Sub-dataframe ",j," group number
",indexcounter," row ",counter))
      index<-indexVectorCreator(social_trans_totals_round,dfname+2)
      indexcounter<-index[[j]]
      if
(counter<=social_trans_totals_round[indexcounter,dfname+2]){
        df[i,"Social2001"]<-
social_trans_totals_round[indexcounter,"Social_2001"]
        counter=counter+1
      } else {
        indexcounter<-index[[j+1]]
        df[i,"Social2001"]<-
social_trans_totals_round[indexcounter,"Social_2001"]
        counter=2
        j=j+1
      }
    }
  }
  print(paste0("Finished estimation for social group ",dfname))
  assign(paste0("AgeSocialSub",dfname),df)
}
```

After the estimation loop has finished, create missing data values for the age 0-9 and 100+ age groups, recombine these with the newly estimated data, and rearrange back into the original order using the pseudo ID column.

```
# create missing data values in first and last age groups
Age10YrGrp_0$Social2001<--7
Age10YrGrp_10$Social2001<--7

#recombine everything back into a single data frame
SAR2011<-
rbind(Age10YrGrp_0,AgeSocialSub1,AgeSocialSub2,AgeSocialSub3,AgeSoci
alSub4,AgeSocialSub5,AgeSocialSub6,AgeSocialSub7,AgeSocialSub8,AgeSo
cialSub9,Age10YrGrp_10)

#reorder by pesudo_id
SAR2011<-arrange(SAR2011,Pseudo_ID)

write.csv(SAR2011,"SAR2011.csv")
```

3.3 Special Estimation Cases

While the methodology described above in section 3.2 is used almost identically for each variable in the synthetic data set, there are some special cases where the method varied slightly. The variation occurs in the data preparation stage, so will be outlined briefly here.

3.3.1 Religion

Religion posed a unique problem as the variable itself was recoded from the original 9 category variable in 2011 to a binary Religion/No Religion variable. This in turn led to the possibility that individuals could transition between these states, but also from religion to religion where the religion stayed the same and religion to religion where the religion is different.

These transitions are generated from the original LS data, but as the synthetic dataset creates estimates moving backwards from 2011 to 2001, a multi-staged process was required.

Stage 1 – Recode the 2011 Religion variable into a binary religion variable

```
newvar<-0
recode_religion<-function(original_religion_variable){
  newvar[original_religion_variable ==-9]<--9
  newvar[original_religion_variable ==1]<-1
  newvar[original_religion_variable ==2]<-2
  newvar[original_religion_variable ==3]<-2
  newvar[original_religion_variable ==4]<-2
  newvar[original_religion_variable ==5]<-2
  newvar[original_religion_variable ==6]<-2
  newvar[original_religion_variable ==7]<-2
  newvar[original_religion_variable ==8]<-2
  newvar[original_religion_variable ==9]<-2
  return(newvar)
}

#now create a new variable in the SAR2011 dataset which is the re-
coded
#religion data
SAR2011$ReligionBinary<-recode_religion(SAR2011$Religion)
```

Stage 2 – Estimate the number of people by religion in 2011 who had a different religion in 2001 from the LS transitional probabilities.

To carry out this estimation, firstly create a table with the sum of each binary religion variable for each age group in the data set, then divide the religious group into same religion and different religion using proportions derived from the LS data. Use this new table of non-religious, religious (same religion) and religious (different religion) (Table 7) to

Table 7 – Ternary religion 2011 estimated from LS proportions

Religion_2011	ReligionBinary	10-19	20-29	30-39	40-49	…
-9	**-9**	2972	3083	99	10	…
1	**1**	22084	27185	23847	21393	…
2	**2**	45693	49485	50187	60424	…
3	**2**	280	597	553	534	…

If we call Table 7 ‘relig_recode’ we can use the following code to create a new ternary religion variable for 2011.

```
dataframe=1
for (dataframe in 3:length(AgeGrpDFList)-1){
  dfname<-AgeGrpList[dataframe]
  df<-as.data.frame(AgeGrpDFList[dataframe])
  i=1
  j=3
  counter=1
  for (i in 1:nrow(df)){
    if (df[i,"ReligionBinary"]==-9){
      df[i,"ReligionTernary"]<--9
      #print(i)
      next
    } else if (df[i,"ReligionBinary"]==1){
      df[i,"ReligionTernary"]<-1
      #print(i)
      next
    } else {
      print(paste0("starting next loop - counter is: ",counter))
      if (counter<=relig_recode[j,dfname+2]){
        df[i,"ReligionTernary"]<-relig_recode[j,"Religion_2011"]
        counter=counter+1
      } else {
        df[i,"ReligionTernary"]<-relig_recode[j+1,"Religion_2011"]
        counter=2
        j=j+1
      }
    }
  }
  print(paste0("Finished estimation for ReligionTernary ",dfname))
  assign(paste0("ReligSub",dfname),df)
}

#create some missing data values for the first and last age groups
Age10YrGrp_0$ReligionTernary<--7
Age10YrGrp_10$ReligionTernary<--7

#recombine everything back into a single dataframe
SAR2011<-
rbind(Age10YrGrp_0,ReligSub1,ReligSub2,ReligSub3,ReligSub4,ReligSub5
,ReligSub6,ReligSub7,ReligSub8,ReligSub9,Age10YrGrp_10)

#reorder by pesudo_id
SAR2011<-arrange(SAR2011,Pseudo_ID)
```

Having recoded 2011 religion into a ternary value, we can then use the method described in Section 3.2 to finish the estimation process.

3.4 Births and Deaths

The estimation of births and deaths in the data is somewhat of a special case, but again, uses very similar programmatic loops to those described above to randomise and then apportion transitions to the correct numbers of individuals at each age group.

In order to estimate births, we simply count the number of live births (0, 1, 2, 3 or more) to LS members between 2001 and 2011 for females in each age group. We then generate probabilities as before, apply these to our SAR data (female population only, of course) to calculate the number of individuals who should transition and then, as before, randomise and update the microdata records.

With the death transition, we are able to incorporate an interaction with health and age in 2001 without small numbers becoming an issue. Clearly age interacting with general health is likely to result in more reliable death transitions than age interacting with any of our other chosen variables. Probabilities of death between 2001 and 2011 are calculated for each age group and each general health category, e.g. Good Health 2001 to Alive 2011, Good Health 2001 to dead 2011, etc. and then applied to populations in these sub-categories to estimate individuals undergoing each transition, before randomisation of individuals and the assignment to the transition to individuals using the loops described before.

4 Results

The completed synthetic spine for England and Wales can now be downloaded from https://dl.dropboxusercontent.com/u/8649795/NewSpine.zip.

As well as the 17 original variables from the 2011 SAR teaching file, there exist 10 synthetic variables – 5 longitudinal transitions from 2001 (age, religion, health, marital status, approximate social grade, births to mothers and deaths);1 re-estimation of the original age group variable into a new 10 year-age group variable; and 2 re-estimations of the 2011 religion variable into binary (religious/non-religious) and ternary (same religion, different religion, religion to no religion) groupings.

Returning to the original aims of this project, we set out to produce a general usage synthetic dataset which can be used to introduce novices to longitudinal data. As such, our concern was with producing something which was usable and contained broadly plausible transitions and that it could be produced relatively quickly (after earlier unforeseen problems in the project). Our choice of age as the constraining variable for the transitions that occurred between other variable states between 2001 and 2011 appears to have been a sensible one, as due to age being collinear with a number of other variables such as marital status and health, we have actually captured some of these multi-variate interactions in the process. To explain consider the following example using Marital Status Transitions:

Figure 1 – Transitions between Marital Status, 2001 to 2011, all ages, SYLLS Synthetic Estimates

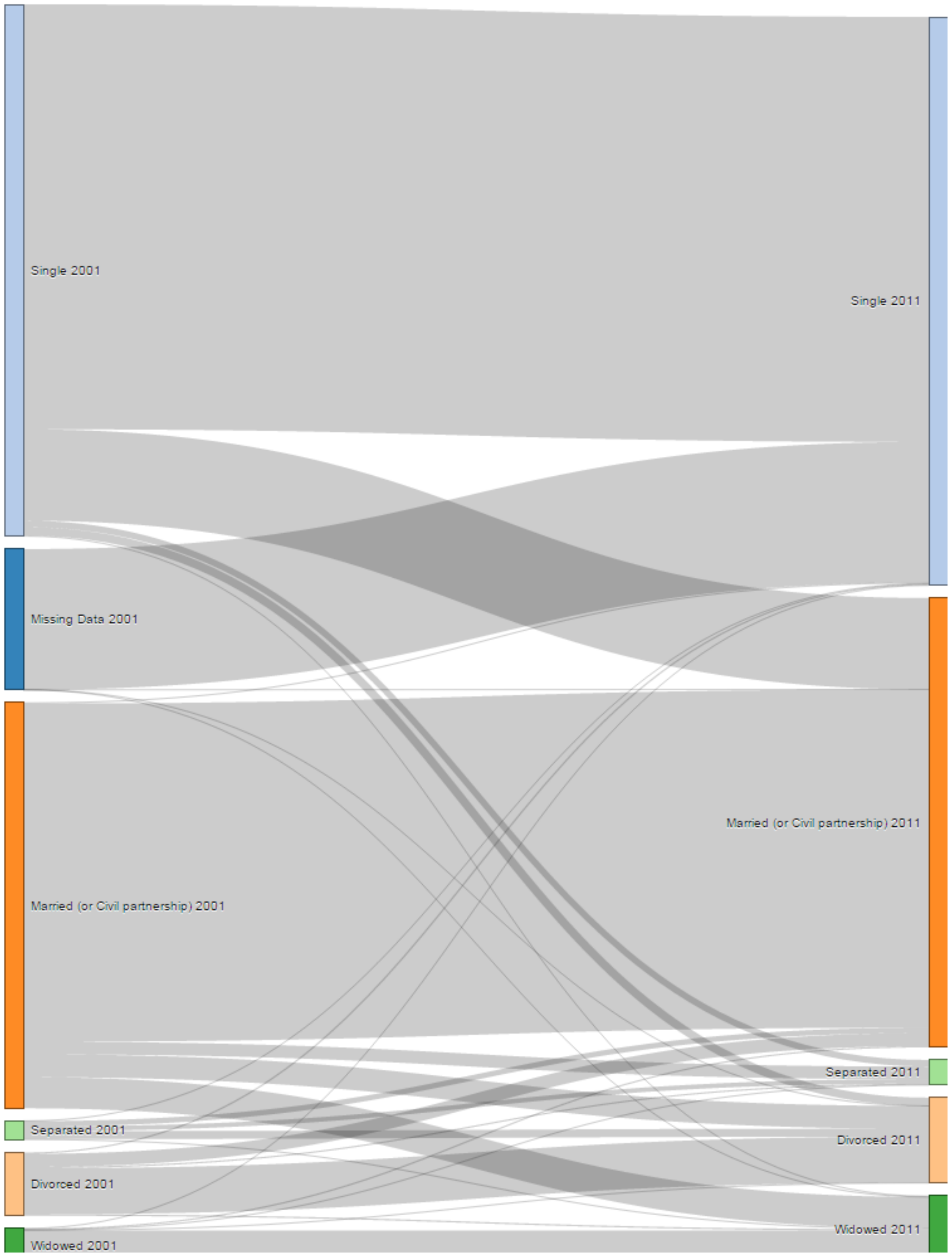
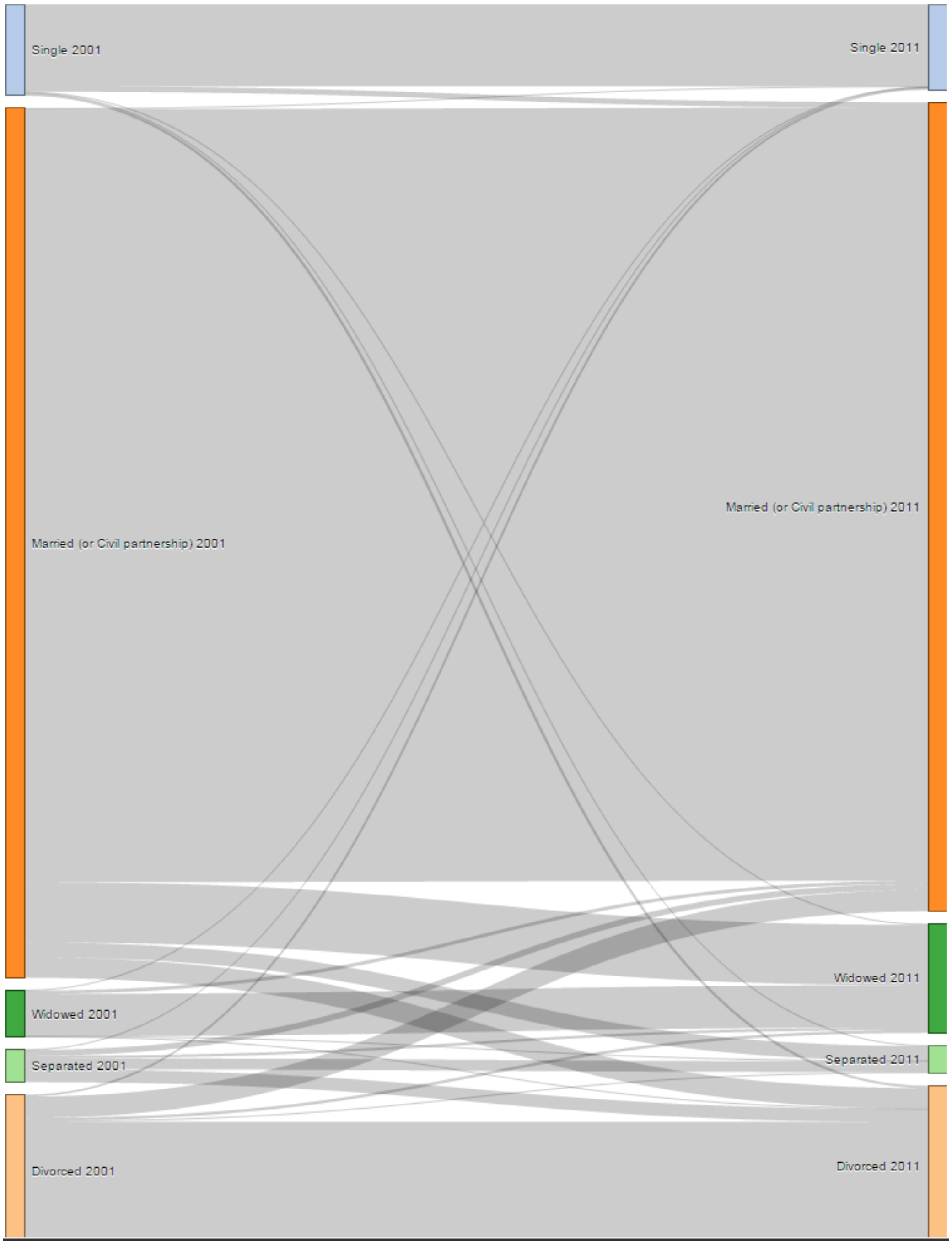


Figure 2 – Transitions between Marital Status, 2001 to 2011, Age 60-69 (in 2001), SYLLS Synthetic Estimates



Figures 1 and 2 show transitions between martial states between 2001 and 2011 in the synthetic spine dataset. While Figure 1 represents all ages Figure 2 examines these transitions for those aged 60-69 in 2001. As we might expect, a much larger proportion of those aged 60-69 in 2001 are married compared to the overall figure. Equally, the proportion of those transitioning from married to widowed is far more pronounced in this age group. Now we would also expect that for those transitioning from married to widowed, this transition might be more likely where the deceased partner suffered from poor health at the beginning of the period. It might be an incorrect assumption to make, but if we assume for the time being that the poor health of one partner may also be experienced by the other if they experience the same environmental and socio-economic conditions which could give rise to poor health, then we might expect a greater proportion of those with poor health in 2001 transitioning from married to widowed than those with good health. Examining Table 8 below, we can see that this is indeed the case, with 12.7% of those with bad or very bad health in 2001 transitioning from married to widowed, whereas only 5.7% of those with good health underwent the same transition.

Table 8 – Proportions of those with a given marital state and health status in 2001 transitioning to another martial state in 2011.

Health2001			**Marital 2011**					**Total**
			Single	**Married**	**Separated**	**Divorced**	**Widowed**	
Good / V Good	**Marital 2001**	**Single**	79.7	17.5	1.2	1.5	0.2	100.0
		Married	0.3	84.9	3.3	5.9	5.7	100.0
		Separated	2.0	31.1	21.5	41.3	4.1	100.0
		Divorced	2.3	23.6	2.2	70.1	1.8	100.0
		Widowed	1.3	5.0	0.5	1.2	92.1	100.0
	Total		45.5	40.9	2.3	7.2	4.1	100.0
Fair	**Marital 2001**	**Single**	77.2	18.8	1.5	2.2	0.4	100.0
		Married	0.2	81.2	2.8	4.9	11.0	100.0
		Separated	1.6	23.7	25.9	42.5	6.3	100.0
		Divorced	1.9	17.0	1.9	76.4	2.8	100.0
		Widowed	0.6	3.1	0.3	0.9	95.0	100.0
	Total		25.8	49.3	2.6	10.3	12.0	100.0
Bad / V Bad	**Marital 2001**	**Single**	79.3	15.5	1.8	2.7	0.7	100.0
		Married	0.2	78.7	3.0	5.4	12.7	100.0
		Separated	1.3	17.0	32.2	41.7	7.8	100.0
		Divorced	2.0	12.4	2.0	80.5	3.1	100.0
		Widowed	0.7	2.6	0.5	1.1	95.1	100.0
	Total		19.0	49.3	3.2	13.2	15.3	100.0

Of course, these assumptions may be incorrect, but the synthetic data at least captures a plausible multi-variable transition thanks to age acting as the main constraint in the estimation process. Scrutinising the table further, other plausible interactions can be identified, such as the slightly lower likelihood of single people with poor health transitioning to married than those with good health undergoing the same transition.

5 Conclusions

This paper has detailed a generalisable method for applying longitudinal transitions to pre-existing microdata in order to produce a universal, publicly available teaching dataset enabling new researchers to familiarise themselves with the unique properties of longitudinal microdata outside of restrictive secure data laboratories. The intention is that datasets such as the Synthetic Spine will be used widely in introductory teaching classes.

There are various improvements that can be made to this dataset, such as the inclusion of additional variable transitions (region and a 10 year inter-region migration transition would be particularly useful) or a longer time series (going back to the 1991 Census or even further). But as this paper has demonstrated, the simplicity of the method and the algorithms for updating individual microdata cases means that the Synthetic Spine can be easily extended in the future. Furthermore, once similar transitional matrices have been produced from the Scottish and Northern Irish Longitudinal Studies, the same techniques can be applied to the SAR teaching files for Northern Ireland and Scotland generating comparable Synthetic Spine data for these countries and completing a full UK Synthetic Longitudinal Spine dataset.

Acknowledgements

The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is supported by the ESRC Census of Population Programme (Award Ref: ES/K000365/1). The authors alone are responsible for the interpretation of the data.

This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

References

Boyle, P. 2010. Pilot study commissioned by Scottish Collaboration for Public Health Research and Policy.

Champion, T. 2012. Testing the return migration element of the ‘escalator region’ model: an analysis of migration into and out of south-east England, 1966–2001. *Cambridge Journal of Regions, Economy and Society* 5 (2):255-270.

Champion, T., M. Coombes, and I. Gordon. 2014. How Far do England's Second-Order Cities Emulate London as Human-Capital ‘Escalators’? *Population, Space and Place* 20 (5):421-433.

Deming, W., and F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 11:427-444.

Dini, E. 2010. Older workers’ withdrawal from the labour market 1991 to 2007: impact of socio-demographic characteristics, health and household circumstances. *Population Trends* 142 (Winter):52-77.

Dykstra, P. L., E. Grundy, T. Fokkema, J. de Jong Gierverld, G. Ploubidis, S. Read, and C. Tomassini. *Health and well-being at older ages: the interlinkage with family life histories, gender and national contexts* 2009 [cited. Available from http://depot.knaw.nl/7596/.

Feng, Z., P. Boyle, M. van Ham, and G. Raab. 2010. Neighbourhoods and the creation, stability and success of mixed ethnic unions.

Grundy, E. 2009. Women's fertility and mortality in late mid life: A comparison of three contemporary populations. *American Journal of Human Biology* 21 (4):541-547.

Norman, P. 1999. Putting iterative proportional fitting on the researcher’s desk. In *Working Paper 99/3*. Leeds: School of Geography, University of Leeds.

Nowok, B., G. Raab, and C. Dibben. 2014. synthpop: Bespoke Creation of Synthetic Data in R.

Platt, L., C. Zuccotti, E. Kaufmann, G. Catney, and G. Harris. 2014. Ethnic and religious change in Britain, 1991-2011. In *Census Linkage Launch Event*. Church House, Westminster, London.

Raab, G., B. Nowok, and C. Dibben. 2015. A simplified approach to generating synthetic data for disclosure control. arXiv:1409.0217 [stat.ME].

Riva, M., S. Curtis, and P. Norman. 2011. Residential mobility within England and urban–rural inequalities in mortality. *Social Science & Medicine* 73 (12):1698-1706.

Robards, J. W., A. M. Berrington, and A. Hinde. 2011. Estimating fertility rates using the ONS Longitudinal Study – what difference does the inclusion of non-continually resident members make? *Population Trends* 144 (Summer):33-47.

Scott, A. P., and I. M. Timæus. 2013. Mortality differentials 1991−2005 by self-reported ethnicity: findings from the ONS Longitudinal Study. *Journal of Epidemiology and Community Health*.

Simpson, L., S. Jivraj, and J. Warren. 2014. The stability of ethnic group and religion in the Censuses of England and Wales 2001-2011. Manchester: CoDE Working Paper.

Simpson, L., and M. Tranmer. 2005. Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software*. *The Professional Geographer* 57 (2):222-234.