Description of datazone linking process used to produce dissimilarity app data

Jon Minton

18 August 2015

# Introduction

The purpose of this document is to describe the contents of the 2001 and 2011 census tables used to produce dissimilarity estimates previously, and the lookup files and process used to link between them. This document will also look at the variables available for ethnic groupings, country of origin, and religion between censuses, in order to make judgements about the consistency and comparability over time.

# Loading, linking, and standardising the structure of 2001 and 2011 census tables

The R packages used are described and loaded in the code chunk below

rm(list=ls())  
  
  
  
require(plyr)

## Loading required package: plyr

require(stringr)

## Loading required package: stringr

require(tidyr)

## Loading required package: tidyr

require(dplyr)

## Loading required package: dplyr  
##   
## Attaching package: 'dplyr'  
##   
## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

require(repmis)

## Loading required package: repmis

require(ggplot2)

## Loading required package: ggplot2

This section will first load the source data tables from the 2001 census, then data from the 2011 data, followed by the lookup file used to link between them. The function source\_DropboxData from the repmis package is used to load these tables directly from Dropbox. The files themselves are on the USI Dropbox account, in the data subdirectory. The table inputs are then passed to the tbl\_df function within the dplyr package to improve their appearance when being displayed on the console. Doing this can also improve the speed of R as much of the underling dplyr code is written in C++. The dplyr piping function (%>% ) is used here as well to make the code easier to read.

## Part 1: Loading 2001 output area tables

###############  
  
# Load tables  
  
# 2001  
  
coo\_2001 <- source\_DropboxData(  
 file="Census\_\_2001\_\_KS005\_\_Country\_Of\_Origin.csv",  
 key="erj713wnp535q20"  
 ) %>% tbl\_df

## Downloading data from: https://dl.dropboxusercontent.com/s/erj713wnp535q20/Census\_\_2001\_\_KS005\_\_Country\_Of\_Origin.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## e6f40b202efaecadbfad77eb182a92f083313da1

eg\_2001 <- source\_DropboxData(  
 file="Census\_\_2001\_\_KS006\_\_Ethnic\_Group\_And\_Language.csv",  
 key="kfgj930jel9qep0"  
 ) %>% tbl\_df

## Downloading data from: https://dl.dropboxusercontent.com/s/kfgj930jel9qep0/Census\_\_2001\_\_KS006\_\_Ethnic\_Group\_And\_Language.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## ce83eda941bc7ad5d458661bd0bcda2f5df040fb

rel\_2001 <- source\_DropboxData(  
 file="Census\_\_2001\_\_KS007\_\_Religion.csv",  
 key="zfnhe8kwxnpupwa"  
 ) %>% tbl\_df

## Downloading data from: https://dl.dropboxusercontent.com/s/zfnhe8kwxnpupwa/Census\_\_2001\_\_KS007\_\_Religion.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## 7f3fcbfb126531eeeb7882bf50f7aaa5c43d7194

Three tables have now been loaded, using the following shorthand

|  |  |
| --- | --- |
| Shorthand | Description |
| coo | Country of Origin |
| eg | Ethnic group |
| rel | Religion |

Let's now loook at the first few lines of these three data tables.

coo\_2001

## Source: local data frame [42,604 x 9]  
##   
## Zone Code All\_People England Scotland Wales Northern\_Ireland  
## 1 60QA000001 58 8 47 0 0  
## 2 60QA000002 174 26 129 0 0  
## 3 60QA000003 76 6 62 0 0  
## 4 60QA000004 66 16 35 1 0  
## 5 60QA000005 82 11 29 0 0  
## 6 60QA000006 198 77 68 2 2  
## 7 60QA000007 73 16 42 0 0  
## 8 60QA000008 108 19 76 4 0  
## 9 60QA000009 153 23 117 5 1  
## 10 60QA000010 65 15 40 0 1  
## .. ... ... ... ... ... ...  
## Variables not shown: Rep\_Ireland (int), Other\_EU (int), Elsewhere (int)

eg\_2001

## Source: local data frame [42,604 x 18]  
##   
## Zone Code All\_People White\_Scottish Other\_White\_British White\_Irish  
## 1 60QA000001 58 46 11 0  
## 2 60QA000002 174 128 28 0  
## 3 60QA000003 76 62 6 0  
## 4 60QA000004 66 34 18 0  
## 5 60QA000005 82 35 10 2  
## 6 60QA000006 198 62 80 1  
## 7 60QA000007 73 45 16 0  
## 8 60QA000008 108 83 15 0  
## 9 60QA000009 153 121 28 1  
## 10 60QA000010 65 42 19 0  
## .. ... ... ... ... ...  
## Variables not shown: Other\_White (int), Indian (int), Pakistani (int),  
## Bangladeshi (int), Other\_South\_Asian (int), Chinese (int), Caribbean  
## (int), African (int), Black\_Scottish\_Or\_Other\_Black (int),  
## Any\_Mixed\_Background (int), Other\_Ethnic\_Group (int),  
## Gaelic\_Speaker\_And\_Born\_In\_Scotland (int),  
## Gaelic\_Speaker\_And\_Not\_Born\_In\_Scotland (int)

rel\_2001

## Source: local data frame [42,604 x 13]  
##   
## Zone Code All\_People Church\_Of\_Scotland Roman\_Catholic Other\_Christian  
## 1 60QA000001 58 28 1 7  
## 2 60QA000002 174 69 9 18  
## 3 60QA000003 76 28 5 1  
## 4 60QA000004 66 16 7 23  
## 5 60QA000005 82 21 22 20  
## 6 60QA000006 198 29 13 88  
## 7 60QA000007 73 24 10 18  
## 8 60QA000008 108 48 6 19  
## 9 60QA000009 153 79 3 20  
## 10 60QA000010 65 24 2 9  
## .. ... ... ... ... ...  
## Variables not shown: Buddhist (int), Hindu (int), Jewish (int), Muslim  
## (int), Sikh (int), Another\_Religion (int), None (int), Not\_Announced  
## (int)

The variable Zone Code refers to 2001 output area codes. These need to be merged to 2001 datazones.

The subsequent columns (relabeled by me beforehand) contain different ways of subdividing populations within the output areas.

## Part 2: loading areal links

Two link files are used to link the 2001 output areas to 2001 datazones.

* OUTPUT\_AREA\_2001\_Lookup : *Original source still to be determined*
* Census\_2011\_Lookup\_\_OA\_TO\_HIGHER\_AREAS : Sourced from the NRS. Further information is available [here](http://www.nrscotland.gov.uk/files/geography/2011-census/geog-2011-census-index-file-specification.pdf)

The link files, which I call areal\_links\_pt1 and areal\_links\_pt2, are loaded as follows

area\_links\_pt1 <- source\_DropboxData(  
 file="OUTPUT\_AREA\_2001\_LOOKUP.csv",  
 key="39wszvlpxy4qvpf"  
 ) %>% tbl\_df

## Downloading data from: https://dl.dropboxusercontent.com/s/39wszvlpxy4qvpf/OUTPUT\_AREA\_2001\_LOOKUP.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## 876596b9ace0c1af53027cf40749c09e4afbc119

area\_links\_pt2 <- source\_DropboxData(  
 file="Census\_2011\_Lookup\_\_OA\_TO\_HIGHER\_AREAS.csv",  
 key="95x5ozuw0c6xgxk"  
 ) %>% tbl\_df

## Downloading data from: https://dl.dropboxusercontent.com/s/95x5ozuw0c6xgxk/Census\_2011\_Lookup\_\_OA\_TO\_HIGHER\_AREAS.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## cf7850c5b957af36f753b18310e77c73d8026165

## Warning in fread(data, sep = sep, header = header, ...): Bumped column 12  
## to type character on data row 223, field contains '"S21000003"'. Coercing  
## previously read values in this column from logical, integer or numeric  
## back to character which may not be lossless; e.g., if '00' and '000'  
## occurred before they will now be just '0', and there may be  
## inconsistencies with treatment of ',,' and ',NA,' too (if they occurred in  
## this column before the bump). If this matters please rerun and set  
## 'colClasses' to 'character' for this column. Please note that column type  
## detection uses the first 5 rows, the middle 5 rows and the last 5 rows, so  
## hopefully this message should be very rare. If reporting to  
## datatable-help, please rerun and include the output from verbose=TRUE.

area\_links <- area\_links\_pt1 %>%   
 inner\_join(area\_links\_pt2) %>%   
 select(  
 OutputArea2011Code,  
 NRSoldOutputArea2001Code,  
 OutputArea2001Code,  
 Datazone2001Code,  
 Easting,  
 Northing  
 )

## Joining by: "OutputArea2001Code"

Again, let's look at the first few rows of each file.

area\_links\_pt1

## Source: local data frame [42,604 x 2]  
##   
## OutputArea2001Code NRSoldOutputArea2001Code  
## 1 S00000001 60QA000001  
## 2 S00000002 60QA000002  
## 3 S00000003 60QA000003  
## 4 S00000004 60QA000004  
## 5 S00000005 60QA000005  
## 6 S00000006 60QA000006  
## 7 S00000007 60QA000007  
## 8 S00000008 60QA000008  
## 9 S00000009 60QA000009  
## 10 S00000010 60QA000010  
## .. ... ...

area\_links\_pt2

## Source: local data frame [46,351 x 34]  
##   
## OutputArea2011Code LAU2011Level2Code Islands2011Code Settlement2010Code  
## 1 S00093944 S31000001 S20000477  
## 2 S00132898 S31000378 S20000110  
## 3 S00095247 S31000216 S20000349  
## 4 S00119131 S31001050 S20000286  
## 5 S00097920 S31000496 S20000309  
## 6 S00095782 S31000223 S20000447  
## 7 S00120583 S31000184 S20000461  
## 8 S00097516 S31000471 S20000298  
## 9 S00119061 S31001029 S20000234  
## 10 S00128942 S31001215 086 S20000294  
## .. ... ... ... ...  
## Variables not shown: Locality2010Code (chr), CouncilArea2011Code (chr),  
## LocalCharacteristicSector2011Code (chr),  
## DetailedCharacteristicSector2011Code (chr), PostcodeSector2011 (chr),  
## HealthBoardArea2006Code (chr), CivilParish1930Code (chr),  
## NationalPark2010Code (chr), StrategicDevelopmentPlanningArea2008Code  
## (chr), CommunityHealthPartnership2012Code (chr),  
## CommunityHealthPartnershipSubAreas2011Code (chr),  
## EnterpriseRegion2008Code (chr), ScottishParliamentaryRegion1999Code  
## (chr), UKParliamentaryConstituency2005Code (chr),  
## ScottishParliamentaryConstituency1999Code (chr),  
## ScottishParliamentaryRegion2011Code (chr),  
## ScottishParliamentaryConstituency2011Code (chr), ElectoralWard2007Code  
## (chr), Datazone2001Code (chr), IntermediateZone2001Code (chr),  
## ScottishIndexofMultipleDeprivation2012Code (chr),  
## UrbanRural8Fold2011\_12Code (chr), UrbanRural6Fold2011\_12Code (chr),  
## OutputArea2001Code (chr), OutputArea1991Code (chr), MasterPostcode  
## (chr), Easting (chr), Northing (chr), Hectarage (dbl), SquareKilometres  
## (dbl)

area\_links

## Source: local data frame [46,351 x 6]  
##   
## OutputArea2011Code NRSoldOutputArea2001Code OutputArea2001Code  
## 1 S00088957 60QA000002 S00000002  
## 2 S00088958 60QA000002 S00000002  
## 3 S00088956 60QA000002 S00000002  
## 4 S00088959 60QA000003 S00000003  
## 5 S00088960 60QA000005 S00000005  
## 6 S00088961 60QA000006 S00000006  
## 7 S00088962 60QA000007 S00000007  
## 8 S00088963 60QA000008 S00000008  
## 9 S00088964 60QA000008 S00000008  
## 10 S00088965 60QA000009 S00000009  
## .. ... ... ...  
## Variables not shown: Datazone2001Code (chr), Easting (chr), Northing (chr)

rm(area\_links\_pt1, area\_links\_pt2)

The first of these tables is used to link between the old NRS output codes, prefixed with 60QA, to the newer 2001 (UK standardised?) output area codes prefixed S00. The second of these tables then allows these figures to be linked to a wide range of higher geographies, including to 2001 datazones. **However 2011 datazones are not available within this file.**

The following section of code renames some of the variables in datasets to make them consistent

coo\_2001 <- coo\_2001 %>% rename(zonecode = `Zone Code`)  
eg\_2001 <- eg\_2001 %>% rename(zonecode= `Zone Code`)  
rel\_2001 <- rel\_2001 %>% rename(zonecode= `Zone Code`)  
  
area\_links <- area\_links %>%   
 rename(  
 zonecode=NRSoldOutputArea2001Code,  
 datazone=Datazone2001Code  
 )  
  
coo\_2001 <- coo\_2001 %>% inner\_join(area\_links) %>% arrange(datazone)

## Joining by: "zonecode"

eg\_2001 <- eg\_2001 %>% inner\_join(area\_links) %>% arrange(datazone)

## Joining by: "zonecode"

rel\_2001 <- rel\_2001 %>% inner\_join(area\_links) %>% arrange(datazone)

## Joining by: "zonecode"

names(coo\_2001) <- tolower(names(coo\_2001))  
names(eg\_2001) <- tolower(names(eg\_2001))  
names(rel\_2001) <- tolower(names(rel\_2001))  
  
  
coo\_2001

## Source: local data frame [46,351 x 14]  
##   
## zonecode all\_people england scotland wales northern\_ireland  
## 1 60QA000008 108 19 76 4 0  
## 2 60QA000008 108 19 76 4 0  
## 3 60QA001254 135 23 102 1 0  
## 4 60QA001254 135 23 102 1 0  
## 5 60QA001256 87 14 63 2 0  
## 6 60QA001258 93 4 82 1 2  
## 7 60QA001343 92 6 78 0 0  
## 8 60QA001805 102 10 82 2 0  
## 9 60QA001807 148 28 104 1 1  
## 10 60QA001780 163 13 145 0 0  
## .. ... ... ... ... ... ...  
## Variables not shown: rep\_ireland (int), other\_eu (int), elsewhere (int),  
## outputarea2011code (chr), outputarea2001code (chr), datazone (chr),  
## easting (chr), northing (chr)

eg\_2001

## Source: local data frame [46,351 x 23]  
##   
## zonecode all\_people white\_scottish other\_white\_british white\_irish  
## 1 60QA000008 108 83 15 0  
## 2 60QA000008 108 83 15 0  
## 3 60QA001254 135 109 18 1  
## 4 60QA001254 135 109 18 1  
## 5 60QA001256 87 69 11 0  
## 6 60QA001258 93 82 7 0  
## 7 60QA001343 92 78 7 0  
## 8 60QA001805 102 87 12 0  
## 9 60QA001807 148 106 28 1  
## 10 60QA001780 163 150 9 0  
## .. ... ... ... ... ...  
## Variables not shown: other\_white (int), indian (int), pakistani (int),  
## bangladeshi (int), other\_south\_asian (int), chinese (int), caribbean  
## (int), african (int), black\_scottish\_or\_other\_black (int),  
## any\_mixed\_background (int), other\_ethnic\_group (int),  
## gaelic\_speaker\_and\_born\_in\_scotland (int),  
## gaelic\_speaker\_and\_not\_born\_in\_scotland (int), outputarea2011code (chr),  
## outputarea2001code (chr), datazone (chr), easting (chr), northing (chr)

rel\_2001

## Source: local data frame [46,351 x 18]  
##   
## zonecode all\_people church\_of\_scotland roman\_catholic other\_christian  
## 1 60QA000008 108 48 6 19  
## 2 60QA000008 108 48 6 19  
## 3 60QA001254 135 48 6 23  
## 4 60QA001254 135 48 6 23  
## 5 60QA001256 87 23 1 9  
## 6 60QA001258 93 43 2 14  
## 7 60QA001343 92 33 10 7  
## 8 60QA001805 102 45 4 9  
## 9 60QA001807 148 66 5 9  
## 10 60QA001780 163 56 10 7  
## .. ... ... ... ... ...  
## Variables not shown: buddhist (int), hindu (int), jewish (int), muslim  
## (int), sikh (int), another\_religion (int), none (int), not\_announced  
## (int), outputarea2011code (chr), outputarea2001code (chr), datazone  
## (chr), easting (chr), northing (chr)

coo\_2001 <- coo\_2001 %>%   
 select(  
 -zonecode, -outputarea2011code,  
 -outputarea2001code, -easting, -northing  
 ) %>%  
 gather(  
 type,  
 count,  
 -datazone  
 ) %>%   
 mutate(year=2001) %>%  
 arrange(datazone, year)  
  
eg\_2001 <- eg\_2001 %>%  
 select(  
 -zonecode, -outputarea2001code,  
 -outputarea2011code, -easting, -northing  
 ) %>%  
 gather(  
 type, count,  
 -datazone  
 ) %>%   
 mutate(year=2001) %>%   
 arrange(datazone, year)  
   
rel\_2001 <- rel\_2001 %>%  
 select(  
 -zonecode, -outputarea2001code,  
 -outputarea2011code, -easting, -northing  
 ) %>%  
 gather(  
 type, count,  
 -datazone  
 ) %>%  
 mutate(year=2001) %>%  
 arrange(datazone, year)  
  
coo\_2001

## Source: local data frame [370,808 x 4]  
##   
## datazone type count year  
## 1 S01000001 all\_people 108 2001  
## 2 S01000001 all\_people 108 2001  
## 3 S01000001 all\_people 135 2001  
## 4 S01000001 all\_people 135 2001  
## 5 S01000001 all\_people 87 2001  
## 6 S01000001 all\_people 93 2001  
## 7 S01000001 all\_people 92 2001  
## 8 S01000001 all\_people 102 2001  
## 9 S01000001 all\_people 148 2001  
## 10 S01000001 england 19 2001  
## .. ... ... ... ...

eg\_2001

## Source: local data frame [787,967 x 4]  
##   
## datazone type count year  
## 1 S01000001 all\_people 108 2001  
## 2 S01000001 all\_people 108 2001  
## 3 S01000001 all\_people 135 2001  
## 4 S01000001 all\_people 135 2001  
## 5 S01000001 all\_people 87 2001  
## 6 S01000001 all\_people 93 2001  
## 7 S01000001 all\_people 92 2001  
## 8 S01000001 all\_people 102 2001  
## 9 S01000001 all\_people 148 2001  
## 10 S01000001 white\_scottish 83 2001  
## .. ... ... ... ...

rel\_2001

## Source: local data frame [556,212 x 4]  
##   
## datazone type count year  
## 1 S01000001 all\_people 108 2001  
## 2 S01000001 all\_people 108 2001  
## 3 S01000001 all\_people 135 2001  
## 4 S01000001 all\_people 135 2001  
## 5 S01000001 all\_people 87 2001  
## 6 S01000001 all\_people 93 2001  
## 7 S01000001 all\_people 92 2001  
## 8 S01000001 all\_people 102 2001  
## 9 S01000001 all\_people 148 2001  
## 10 S01000001 church\_of\_scotland 48 2001  
## .. ... ... ... ...

The final chunk of code aggregates counts by datazone, as well as removes gaelic/non gaelic sub groups within the ethnic group files to avoid double counting of populations.

rel\_2001 <- rel\_2001 %>% group\_by(datazone, year, type) %>%   
 summarise(count=sum(count)) %>%  
 select(datazone, year, type, count)  
  
coo\_2001 <- coo\_2001 %>% group\_by(datazone, year, type) %>%  
 summarise(count = sum(count)) %>%  
 select(datazone, year, type, count)  
  
eg\_2001 <- eg\_2001 %>% group\_by(datazone, year, type) %>%  
 summarise(count = sum(count)) %>%  
 select(datazone, year, type, count)  
  
# Ethnic group counts do not add up correctly as categories on gaelic included  
  
eg\_2001 <- eg\_2001 %>%  
 filter(type!="gaelic\_speaker\_and\_born\_in\_scotland" & type !="gaelic\_speaker\_and\_not\_born\_in\_scotland")  
  
  
rel\_2001

## Source: local data frame [78,000 x 4]  
## Groups: datazone, year  
##   
## datazone year type count  
## 1 S01000001 2001 all\_people 1008  
## 2 S01000001 2001 church\_of\_scotland 402  
## 3 S01000001 2001 roman\_catholic 46  
## 4 S01000001 2001 other\_christian 132  
## 5 S01000001 2001 buddhist 0  
## 6 S01000001 2001 hindu 0  
## 7 S01000001 2001 jewish 0  
## 8 S01000001 2001 muslim 4  
## 9 S01000001 2001 sikh 0  
## 10 S01000001 2001 another\_religion 4  
## .. ... ... ... ...

coo\_2001

## Source: local data frame [52,000 x 4]  
## Groups: datazone, year  
##   
## datazone year type count  
## 1 S01000001 2001 all\_people 1008  
## 2 S01000001 2001 england 146  
## 3 S01000001 2001 scotland 765  
## 4 S01000001 2001 wales 16  
## 5 S01000001 2001 northern\_ireland 3  
## 6 S01000001 2001 rep\_ireland 3  
## 7 S01000001 2001 other\_eu 22  
## 8 S01000001 2001 elsewhere 53  
## 9 S01000002 2001 all\_people 687  
## 10 S01000002 2001 england 66  
## .. ... ... ... ...

eg\_2001

## Source: local data frame [97,500 x 4]  
## Groups: datazone, year  
##   
## datazone year type count  
## 1 S01000001 2001 all\_people 1008  
## 2 S01000001 2001 white\_scottish 806  
## 3 S01000001 2001 other\_white\_british 131  
## 4 S01000001 2001 white\_irish 3  
## 5 S01000001 2001 other\_white 56  
## 6 S01000001 2001 indian 0  
## 7 S01000001 2001 pakistani 0  
## 8 S01000001 2001 bangladeshi 2  
## 9 S01000001 2001 other\_south\_asian 2  
## 10 S01000001 2001 chinese 2  
## .. ... ... ... ...

## 2011 Tables

The code belows does much the same as was done with the 2001 tables. However, the census tables are already available within **2001** datazone format and so the additional linking is not needed

# 2011  
  
# populations  
  
  
# ethnic group 2a ks201sc  
eg\_2011 <- source\_DropboxData(  
 file="KS201SC.csv",  
 key="adxgd6edmvgyk82"  
 ) %>% tbl\_df()

## Downloading data from: https://dl.dropboxusercontent.com/s/adxgd6edmvgyk82/KS201SC.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## 603ac6f02d89c40f4002cad60997ea07c1a05ef5

# country of origin 2a ks 204sc  
coo\_2011 <- source\_DropboxData(  
 file="KS204SC.csv",  
 key="xrs0mxq9alumojs"  
 ) %>% tbl\_df()

## Downloading data from: https://dl.dropboxusercontent.com/s/xrs0mxq9alumojs/KS204SC.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## c32168bc6e3b9cd79469f3f1e54d880ce1593536

# religion 2a 209scb  
  
rel\_2011 <- source\_DropboxData(  
 file="KS209SCb.csv",  
 key="aejzg3hbu443pxl"  
) %>% tbl\_df()

## Downloading data from: https://dl.dropboxusercontent.com/s/aejzg3hbu443pxl/KS209SCb.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## 31a29eca41fb170378878428aef2e5418df83b1a

names(eg\_2011) <- names(eg\_2011) %>% tolower()  
names(coo\_2011) <- names(coo\_2011) %>% tolower()  
names(rel\_2011) <- names(rel\_2011) %>% tolower()  
  
  
eg\_2011

## Source: local data frame [6,501 x 26]  
##   
## v1 all people white white: scottish white: other british  
## 1 S92000003 5,295,403 5,084,407 4,445,678 417,109  
## 2 S01000001 872 844 659 131  
## 3 S01000002 678 635 576 36  
## 4 S01000003 788 755 698 33  
## 5 S01000004 612 582 533 40  
## 6 S01000005 643 610 505 78  
## 7 S01000006 644 628 558 48  
## 8 S01000007 751 726 611 89  
## 9 S01000008 580 549 488 51  
## 10 S01000009 681 674 636 25  
## .. ... ... ... ... ...  
## Variables not shown: white: irish (chr), white: gypsy/traveller (chr),  
## white: polish (chr), white: other white (chr), mixed or multiple ethnic  
## groups (chr), asian, asian scottish or asian british (chr), asian, asian  
## scottish or asian british: pakistani, pakistani scottish or pakistani  
## british (chr), asian, asian scottish or asian british: indian, indian  
## scottish or indian british (chr), asian, asian scottish or asian  
## british: bangladeshi, bangladeshi scottish or bangladeshi british (chr),  
## asian, asian scottish or asian british: chinese, chinese scottish or  
## chinese british (chr), asian, asian scottish or asian british: other  
## asian (chr), african (chr), african: african, african scottish or  
## african british (chr), african: other african (chr), caribbean or black  
## (chr), caribbean or black: caribbean, caribbean scottish or caribbean  
## british (chr), caribbean or black: black, black scottish or black  
## british (chr), caribbean or black: other caribbean or black (chr), other  
## ethnic groups (chr), other ethnic groups: arab, arab scottish or arab  
## british (chr), other ethnic groups: other ethnic group (chr)

coo\_2011

## Source: local data frame [6,501 x 10]  
##   
## v1 all people england northern ireland scotland wales  
## 1 S92000003 5295403 459486 36655 4411884 17381  
## 2 S01000001 872 129 6 648 3  
## 3 S01000002 678 41 2 581 0  
## 4 S01000003 788 41 3 694 2  
## 5 S01000004 612 50 4 527 1  
## 6 S01000005 643 79 1 506 2  
## 7 S01000006 644 54 2 550 5  
## 8 S01000007 751 90 2 598 4  
## 9 S01000008 580 51 0 491 1  
## 10 S01000009 681 34 0 622 0  
## .. ... ... ... ... ... ...  
## Variables not shown: republic of ireland (int), other eu: member countries  
## in march 2001 (1) (int), other eu: accession countries april 2001 to  
## march 2011 (int), other countries (int)

rel\_2011

## Source: local data frame [6,501 x 13]  
##   
## v1 all people church of scotland roman catholic other christian  
## 1 S92000003 5,295,403 1,717,871 841,053 291,275  
## 2 S01000001 872 269 46 74  
## 3 S01000002 678 171 36 37  
## 4 S01000003 788 161 53 22  
## 5 S01000004 612 174 33 27  
## 6 S01000005 643 201 30 49  
## 7 S01000006 644 269 22 31  
## 8 S01000007 751 292 33 56  
## 9 S01000008 580 206 18 43  
## 10 S01000009 681 207 20 26  
## .. ... ... ... ... ...  
## Variables not shown: buddhist (chr), hindu (chr), jewish (chr), muslim  
## (chr), sikh (chr), other religion (chr), no religion (chr), religion not  
## stated (chr)

# filter away to total (the first v1)  
  
eg\_2011 <- eg\_2011 %>% rename(datazone=v1) %>%   
 filter( datazone!="S92000003") %>%   
 gather(  
 type,  
 count,  
 -datazone  
 ) %>% mutate(year=2011) %>%  
 mutate(  
 count=str\_replace(count, "-", "0"),  
 count=str\_replace(count, ",", ""),  
 count=as.numeric(as.character(count))  
 ) %>%  
 select(datazone, year, type, count)  
  
  
coo\_2011 <- coo\_2011 %>% rename(datazone=v1) %>%  
 filter(datazone!="S92000003") %>%   
 gather(  
 type,  
 count,  
 -datazone  
 ) %>% mutate(year=2011) %>%  
 mutate(  
 count=str\_replace(count, "-", "0"),  
 count=str\_replace(count, ",", ""),  
 count=as.numeric(as.character(count))  
 ) %>%  
 select(datazone, year, type, count)  
  
  
rel\_2011 <- rel\_2011 %>% rename(datazone=v1) %>%  
 filter(datazone!="S92000003") %>%   
 gather(  
 type,  
 count,  
 -datazone  
 ) %>% mutate(year=2011) %>%  
 mutate(  
 count=str\_replace(count, "-", "0"),  
 count=str\_replace(count, ",", ""),  
 count=as.numeric(as.character(count))  
 ) %>%  
 select(datazone, year, type, count)  
  
  
  
eg\_2011

## Source: local data frame [162,500 x 4]  
##   
## datazone year type count  
## 1 S01000001 2011 all people 872  
## 2 S01000002 2011 all people 678  
## 3 S01000003 2011 all people 788  
## 4 S01000004 2011 all people 612  
## 5 S01000005 2011 all people 643  
## 6 S01000006 2011 all people 644  
## 7 S01000007 2011 all people 751  
## 8 S01000008 2011 all people 580  
## 9 S01000009 2011 all people 681  
## 10 S01000010 2011 all people 514  
## .. ... ... ... ...

coo\_2011

## Source: local data frame [58,500 x 4]  
##   
## datazone year type count  
## 1 S01000001 2011 all people 872  
## 2 S01000002 2011 all people 678  
## 3 S01000003 2011 all people 788  
## 4 S01000004 2011 all people 612  
## 5 S01000005 2011 all people 643  
## 6 S01000006 2011 all people 644  
## 7 S01000007 2011 all people 751  
## 8 S01000008 2011 all people 580  
## 9 S01000009 2011 all people 681  
## 10 S01000010 2011 all people 514  
## .. ... ... ... ...

rel\_2011

## Source: local data frame [78,000 x 4]  
##   
## datazone year type count  
## 1 S01000001 2011 all people 872  
## 2 S01000002 2011 all people 678  
## 3 S01000003 2011 all people 788  
## 4 S01000004 2011 all people 612  
## 5 S01000005 2011 all people 643  
## 6 S01000006 2011 all people 644  
## 7 S01000007 2011 all people 751  
## 8 S01000008 2011 all people 580  
## 9 S01000009 2011 all people 681  
## 10 S01000010 2011 all people 514  
## .. ... ... ... ...

## Section summary

This section has shown how the 2001 and 2011 censuses can be linked to 2001 datazones. At the time the data processing was performed 2011 datazones were not constructed. However the 2011 areal lookup file is able to link between many areal unit types used in different censuses, as it works with the smallest areal unit types.

So far, although the 2001 and 2011 census tables have been put into a structurally similar format, they categories within the three pairs of tables have not been reduced and mapped to a single, compatible schema. The next stage will describe this process in more detail.

# Linking categories between 2001 and 2011 census tables

Let's first summarise the different 'types' available in each of the six tables.

eg <- eg\_2001 %>% bind\_rows(eg\_2011)

## Warning in rbind\_all(list(x, ...)): Unequal factor levels: coercing to  
## character

coo <- coo\_2001 %>% bind\_rows(coo\_2011)

## Warning in rbind\_all(list(x, ...)): Unequal factor levels: coercing to  
## character

rel <- rel\_2001 %>% bind\_rows(rel\_2011)

## Warning in rbind\_all(list(x, ...)): Unequal factor levels: coercing to  
## character

eg %>% xtabs( ~ type + year, data = .)

## year  
## type 2001  
## african 6500  
## african: african, african scottish or african british 0  
## african: other african 0  
## all people 0  
## all\_people 6500  
## any\_mixed\_background 6500  
## asian, asian scottish or asian british 0  
## asian, asian scottish or asian british: bangladeshi, bangladeshi scottish or bangladeshi british 0  
## asian, asian scottish or asian british: chinese, chinese scottish or chinese british 0  
## asian, asian scottish or asian british: indian, indian scottish or indian british 0  
## asian, asian scottish or asian british: other asian 0  
## asian, asian scottish or asian british: pakistani, pakistani scottish or pakistani british 0  
## bangladeshi 6500  
## black\_scottish\_or\_other\_black 6500  
## caribbean 6500  
## caribbean or black 0  
## caribbean or black: black, black scottish or black british 0  
## caribbean or black: caribbean, caribbean scottish or caribbean british 0  
## caribbean or black: other caribbean or black 0  
## chinese 6500  
## indian 6500  
## mixed or multiple ethnic groups 0  
## other ethnic groups 0  
## other ethnic groups: arab, arab scottish or arab british 0  
## other ethnic groups: other ethnic group 0  
## other\_ethnic\_group 6500  
## other\_south\_asian 6500  
## other\_white 6500  
## other\_white\_british 6500  
## pakistani 6500  
## white 0  
## white: gypsy/traveller 0  
## white: irish 0  
## white: other british 0  
## white: other white 0  
## white: polish 0  
## white: scottish 0  
## white\_irish 6500  
## white\_scottish 6500  
## year  
## type 2011  
## african 6500  
## african: african, african scottish or african british 6500  
## african: other african 6500  
## all people 6500  
## all\_people 0  
## any\_mixed\_background 0  
## asian, asian scottish or asian british 6500  
## asian, asian scottish or asian british: bangladeshi, bangladeshi scottish or bangladeshi british 6500  
## asian, asian scottish or asian british: chinese, chinese scottish or chinese british 6500  
## asian, asian scottish or asian british: indian, indian scottish or indian british 6500  
## asian, asian scottish or asian british: other asian 6500  
## asian, asian scottish or asian british: pakistani, pakistani scottish or pakistani british 6500  
## bangladeshi 0  
## black\_scottish\_or\_other\_black 0  
## caribbean 0  
## caribbean or black 6500  
## caribbean or black: black, black scottish or black british 6500  
## caribbean or black: caribbean, caribbean scottish or caribbean british 6500  
## caribbean or black: other caribbean or black 6500  
## chinese 0  
## indian 0  
## mixed or multiple ethnic groups 6500  
## other ethnic groups 6500  
## other ethnic groups: arab, arab scottish or arab british 6500  
## other ethnic groups: other ethnic group 6500  
## other\_ethnic\_group 0  
## other\_south\_asian 0  
## other\_white 0  
## other\_white\_british 0  
## pakistani 0  
## white 6500  
## white: gypsy/traveller 6500  
## white: irish 6500  
## white: other british 6500  
## white: other white 6500  
## white: polish 6500  
## white: scottish 6500  
## white\_irish 0  
## white\_scottish 0

coo %>% xtabs( ~ type + year, data = .)

## year  
## type 2001 2011  
## all people 0 6500  
## all\_people 6500 0  
## elsewhere 6500 0  
## england 6500 6500  
## northern ireland 0 6500  
## northern\_ireland 6500 0  
## other countries 0 6500  
## other eu: accession countries april 2001 to march 2011 0 6500  
## other eu: member countries in march 2001 (1) 0 6500  
## other\_eu 6500 0  
## rep\_ireland 6500 0  
## republic of ireland 0 6500  
## scotland 6500 6500  
## wales 6500 6500

rel %>% xtabs( ~ type + year, data = .)

## year  
## type 2001 2011  
## all people 0 6500  
## all\_people 6500 0  
## another\_religion 6500 0  
## buddhist 6500 6500  
## church of scotland 0 6500  
## church\_of\_scotland 6500 0  
## hindu 6500 6500  
## jewish 6500 6500  
## muslim 6500 6500  
## no religion 0 6500  
## none 6500 0  
## not\_announced 6500 0  
## other christian 0 6500  
## other religion 0 6500  
## other\_christian 6500 0  
## religion not stated 0 6500  
## roman catholic 0 6500  
## roman\_catholic 6500 0  
## sikh 6500 6500

We can see that the categories are much more consistent between 2001 and 2011 for religion and country of origin than for ethnicity, and also that there are a lot of ethnicity categories and sub-categories. How these categories should be collapsed and mapped to a single schema is a researcher judgement. The dissimilarity app allowed researchers to select the categories they wanted to combine together themselves.