ukdataservice.ac.uk

Mapping 2011 Census Microdata using R

UK Data Service

Author: Patricio Troncoso and Jo Wathan
Updated: n/a
Version: 1
Date: 21/06/2017

We are happy for our materials to be used and copied but request that users should:

- link to our original materials instead of re-mounting our materials on your website
- cite this as an original source as follows:

Troncoso, Patricio and Wathan, Jo (2017). Guide to mapping 2011 Census Microdata using R. UK Data Service, University of Manchester.

Contents

List of Figures

List of Tables

Scope

This guide aims to show the strength of using Census Microdata for a variety of research purposes, via a working example taken from real-life research. In this guide, we assume that you are familiar with Census Microdata, mapping, and statistical software. Further information on these topics can be found in the resources section.

1. Producing maps from individual census data in R

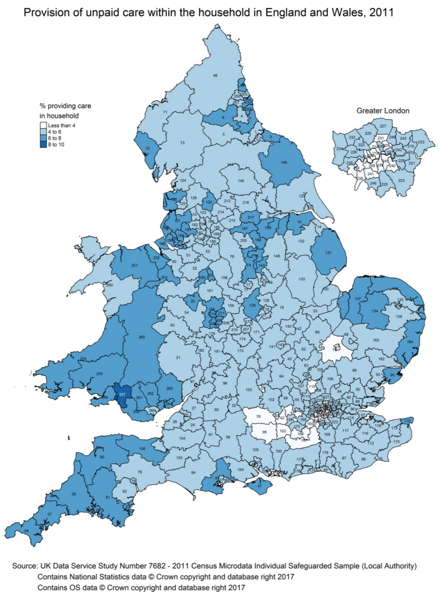
This guide aims to show the strength of using Census Microdata for a variety of research purposes, via a worked example taken from real-life research. In this guide we assume that you have some familiarity with microdata, mapping, and statistical software. Further information on these topics can be found in the resources section.

1.1. The worked example – the geography of unpaid care

One of the main advantages of using Census Microdata is that users can derive bespoke variables unavailable in Census Tables. In this guide we will explore the provision of unpaid care within the household across the grouped Local Authorities in England and Wales. This level of detail about unpaid care is not available in Census Tables, see for instance: http://infusecp.mimas.ac.uk/.

The example to be illustrated follows the work of Norman and Purdam (2013). The authors use the Small Area Microdata from the 2001 Census Samples of Anonymised Records (SARs) to examine geographic and socio-demographic variations in unpaid caring across England and Wales, comparing individuals who provide unpaid care within and outside their household.

Figure 1: Thematic map of England and Wales to be produced in this guide


Provision of unpaid care within the household in England and Wales, 2011
% providing care in household
Less than 4
4 to 6
6 to 8
8 to 10
Greater London

Source: UK Data Service Study Number 7682 - 2011 Census Microdata Individual Safeguarded Sample (Local Authority)
Contains National Statistics data © Crown copyright and database right 2017
Contains OS data © Crown copyright and database right 2017

We will similarly be using Census Microdata from the 2011 Census to produce a map of unpaid carers who are likely to be providing unpaid care of a household member, by identifying those who:

a) provide unpaid care for someone else,

b) are usual residents in a household with at least one person suffering from a long-standing illness or disability, and

c) are not the sole disabled/long-standing ill person in the household.

To do this we will draw upon the following three separate pre-existing variables: Provision of unpaid care, Number of individuals in the household with long-standing illness or disability and Long-term health problem. We will use the "popbasesec" variable to limit our analysis to usual residents and will use an ID variable ("caseno") to count the number of records in each area.

This guide provides detailed instructions on how to derive this bespoke variable, how to aggregate it to the grouped LA level and how to combine it with boundary data to produce a map similar to figure 1.

2. What you will need to create your map

This section identifies the resources you will need in order to produce a map like the one above and suggests some related guides which may be of use.

2.1. Useful guides

This guide assumes some familiarity with mapping, R and Census Microdata. If you would like to know more about any of these concepts you may find the following resources useful in combination with this guide:

- A more detailed guide to Census Microdata can be found on the UK Data Service website: https://census.ukdataservice.ac.uk/use-data/guides/microdata
- A more detailed explanation on boundary data is available here: https://census.ukdataservice.ac.uk/use-data/guides/boundary-data
- For users unfamiliar with R, an introduction covering a range of topics is also available from the UK Data Service website: https://www.ukdataservice.ac.uk/media/398726/usingr.pdf
- Furthermore, users might be interested in the UK Data Service webinar "Putting data on maps" for further examples of mapping data. A recording, the slides, and syntax files are available here: https://www.ukdataservice.ac.uk/news-and-events/eventsitem/?id=4971

2.2. Data

You will need a copy of the microdata file and appropriate boundary files.

2.2.1. Accessing the microdata

The worked example uses 2011 Census Microdata Individual Safeguarded Sample (Local Authority): England and Wales (Office for National Statistics, 2015), which is available from the UK Data Service: http://doi.org/10.5255/UKDA-SN-7682-1. The R commands used here can also be downloaded from this page (see syntax at bottom).

The data do require registration as they are not open data. You will therefore be prompted to sign in when you download these. If you are not registered to use the service, you will be redirected to a registration page which requires you to agree to some terms and conditions of use. Although this does take a couple of minutes, it is an important safeguard which enables you to access useful geographical detail which would otherwise be unavailable.

Download the data as a ".csv" file. You will obtain a zipped file which will need to be expanded into a directory of your choosing in which you will be working. We will refer to this as 'your directory'.

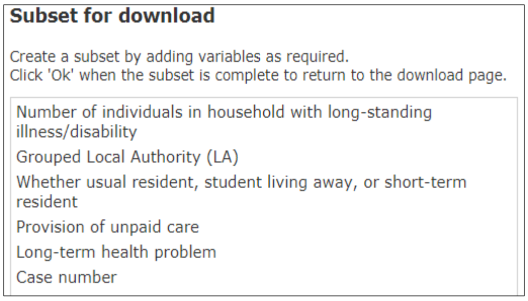
Optional: tip for users with limited memory

The data file is over 500MB in size and may be too large for some machines to read into memory. If you find yourself in this position you may prefer to download only a subset of the data. This can be done by following these steps:

1) Click on the Access online (✓ Access online) link in the dataset's catalogue entry instead of download/order.

2) Inside the Nesstar tool click on the disk (💾) icon to design your download.

3) Select your download format as Comma Separated Values (CSV) and click subset to select only those variables that you need:



Subset for download

Create a subset by adding variables as required.
Click 'Ok' when the subset is complete to return to the download page.

Number of individuals in household with long-standing illness/disability
Grouped Local Authority (LA)
Whether usual resident, student living away, or short-term resident
Provision of unpaid care
Long-term health problem
Case number

Variables are listed in folders on the left hand side, from where they can be selected for inclusion. Number of individuals in the household with long-standing illness/disability can be found in

the household folder.
Case number can be found in the administrative folder.
The other variables can be found in the person folder.

4) Click on Select, and OK to obtain a zip file.

Note that the variable names will be the same as if you downloaded the whole CSV file. However, the dataset name will be different from that specified below so you should rename your file, or modify your R commands, as appropriate. A video tutorial on the use of Nesstar is available at: https://www.ukdataservice.ac.uk/get-data/explore-online/nesstar/nesstar

2.2.2. Obtaining the boundary data

Boundary data which allows you to map to the grouped Local Authority areas contained in this Census Microdata file can be downloaded from the UK Data Service at https://borders.ukdataservice.ac.uk/bds.html

To download the English/Welsh boundary data:

1) Set the dropdown menus to:
    a) Country: England
    b) Geography: Census
    c) Dates: 2011 and later
2) Click on "Find"
3) Boundaries listed as: "English and Welsh Census Microdata Local Authority Groupings, 2011"
4) Scroll down and click on "Extract Boundary Data"
5) Download these features in Shapefile format as a zip file.

The boundary data will be a zip folder containing four files named "ew_groupedla_2011", followed by the extensions ".dbf", ".prj", ".shp" and ".shx". There is also a file containing the Terms and Conditions of use ("TermsAndConditions.html"). This zip folder can also be referred to as a "shapefile". A shapefile (data format developed by ESRI-Environmental Systems Research Institute) is not actually one file but a collection of files with a common prefix, which are stored in the same directory. These files store geometric location and associated attribute information of the areas. The shapefile refers to these areas as "shapes", which can be points, lines and/or polygons. For example, a city can be a point with specified coordinates which is located in a particular region that is drawn as a series of lines (which are in themselves series of points) forming a polygon.

2.3. Software requirements

This guide uses RStudio (RStudio Team, 2016), version 1.0.136, along with R (R Core Team, 2017), version 3.4.0. Furthermore, this guide requires the use of the following R packages: "tmap" (Tennekes, 2017), "rgdal" (Bivand, Keitt, & Rowlingson, 2017), "dplyr" (Wickham & Francois, 2016) and "data.table" (Dowle & Srinivasan, 2017), which can be installed from the R console in the standard way:

```
install.packages("rgdal")
install.packages("tmap")
install.packages("dplyr")
install.packages("data.table")
```

3. Set up and data manipulation

Firstly, standard practice for using R would be to specify a working directory where boundary data and census microdata are stored. This is done in the following way:

```
setwd("yourdirectory")
```

3.1. Read in the boundary data

Next, the boundary data need to be read into R, for which the rgdal package is used. The necessary R commands are given below with annotations.

R command listed in left column explanation in right column

`library(rgdal)`	This calls the rgdal package to read the shapefile.
`dsnengw <- "C:/yourdirectory "`	This specifies the location of the unzipped shapefile (folder). Change as appropriate.

`engw <- readOGR(dsn = dsnengw,`
`               layer = "ew_groupedla_2011")`
	This creates a new R object containing the information of the shapefile of grouped LAs in England and Wales.

The resulting R object "engw" is a "Large Spatial Polygons Data Frame" containing details grouped into five slots prefixed by "@", out of which we will focus on three:

a) data: the attribute table.

b) polygons: the coordinates of the boundaries of the areas, and

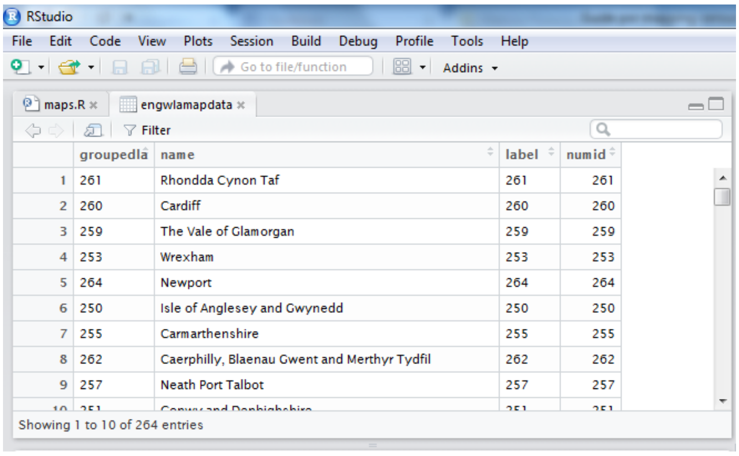
c) bbox: the coordinates of the bounding box.

The attribute table ("@data") can be manipulated to incorporate any information of interest about the areas. It can be manipulated as a standard R object of class data frame or data table. In this case, the package "data.table" is called to export the attribute table from the spatial polygons data frame.

`library(data.table)`	This calls the data.table package.
`engwlamapdata <- data.table(engw@data)`	This creates a new data frame using the attribute table from the shapefile.

`engwlamapdata$numid <- as.numeric(paste(`
`         engwlamapdata$label))`
	This creates a new numeric identifier for the areas. This step is taken to convert the shapefile's area ids from factor to numeric, as working with factor variables can be problematic.

The resulting R data frame will look like the following image:



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
maps.R
engwlamapdata
Filter

	groupedla	name	label	numid
1	261	Rhondda Cynon Taf	261	261
2	260	Cardiff	260	260
3	259	The Vale of Glamorgan	259	259
4	253	Wrexham	253	253
5	264	Newport	264	264
6	250	Isle of Anglesey and Gwynedd	250	250
7	255	Carmarthenshire	255	255
8	262	Caerphilly, Blaenau Gwent and Merthyr Tydfil	262	262
9	257	Neath Port Talbot	257	257
10	251	Conwy and Denbighshire	251	251

Showing 1 to 10 of 264 entries

As with any other data frame in R, it can be manipulated if and as needed. For this guide, we will add columns for each area to represent aggregate statistics for the provision of unpaid care.

3.2. Manipulating the Census Microdata

We will define in-household carers as usual residents who indicated that they provide care for someone else and who are not themselves the sole disabled person within a household containing at least one person suffering from a long-standing illness or disability. This has been derived using three variables from the 2011 Census Microdata Safeguarded Local Authority file:

Table 1: Variables used in this guide from the 2011 Census Microdata Safeguarded Local Authority file

Variable name	*Description*	*Coding*
CARER	Provision of unpaid care	1. No 2. Yes, 1-19 hours 3. Yes, 20-49 hours 4. Yes, 50+ hours
ILLHUK11G	Number of individuals in household with long-standing illness/disability	0. No one in household with long-standing illness/disability 1. 1 household members with long-standing illness/disability 2. 2 or more household members with long-standing illness/disability
DISABILITY	Long-term health problem	1. Day-to-day activities limited a lot 2. Day-to-day activities limited a little 3. Day-to-day activities not limited

Then, to derive the indicator of in-household carers, we assign a code 1 to all those persons who reported having provided care for someone else (codes 2, 3 and 4 of CARER), who reported 1 or more household members with a long-standing illness or disability (codes 1 and 2 of ILLHUK11G), and who did not report long-term health problems (code 3 of DISABILITY). Also, those persons who do report a long-term health problem (codes 1 and 2 of DISABILITY) are considered in-household carers if they reported having provided care for someone else (codes 2, 3 and 4 of CARER) and they reported 2 or more household members with long-standing illness or disability (code 2 of ILLHUK11G). All other cases are considered either carers outside the household or non-carers. This can be summarised as follows:

CARER_INSIDE = **1** if (CARER > 1 & ILLHUK11G > 0 & DISABILITY = 3)
or
(CARER > 1 & ILLHUK11G > 1 & DISABILITY < 3);
otherwise,

CARER_INSIDE = 0

To derive this indicator of provision of unpaid care within the household, we need to read the Census Microdata file into R and do some manipulations at the individual level. The code to be implemented is as follows:

```
censusdata <- read.csv("recodev12.csv",
                        header = T)
```

This creates a new data frame in R called "censusdata", reading in the comma separated file called "recodev12.csv" (the safeguarded Census microdata file).

```
censusdata$carerin <- ifelse(
                   censusdata$carer > 1 &
                   censusdata$illhuk11g > 0 &
                   censusdata$disability == 3 |
                   censusdata$carer > 1 &
                   censusdata$illhuk11g > 1 &
                   censusdata$disability < 3,
                   1, 0)
```

This creates a new variable indicating whether a person provides unpaid care within their own household.

Once carers within the household are identified, individual data can be aggregated at the grouped LA level. To do this, the package "dplyr" is called.

```
library(dplyr)
```

This calls the dplyr package

```
censusladata <- censusdata %>%
  filter(popbasesec == 1) %>%
  group_by(la_group) %>%
  summarise(totcare = sum(carer > 1,
                        na.rm = TRUE),
            carein = sum(carerin,
                       na.rm = TRUE),
            pop = length(unique(caseno)),
            careinpc = carein/pop*100)
```

This creates a new data frame in R called "censusladata", containing the grouped LA codes (la_group), the total number of carers (totcare), the number of carers inside the household (carein), the total number of inhabitants in the grouped LA (pop) and the percentage of carers inside the household with respect to the total number of inhabitants of the grouped LA (careinpc)

Notes: 2011 Census Microdata need to be filtered by usual residence to prevent double counting students. By selecting those who take the value 1 for popbasesec, we select residents only, excluding short-term residents. Indentation is not required; it is only displayed as such to improve readability.

```
engwlamapdata2 <- left_join(engwlamapdata,
                          censusladata,
                   by = c("numid" =
                          "la_group"))
```

This joins the newly aggregated data with the attribute table extracted previously from the shapefile.

```
engw@data <- engwlamapdata2
```

This replaces the attribute table of the shapefile with the newly created attribute table including the information on unpaid care provision.

4. Thematic Maps

The thematic maps can now be drawn calling the “tmap” package. An empty map showing only the borders and the codes of the grouped LAs can be produced in the following way:

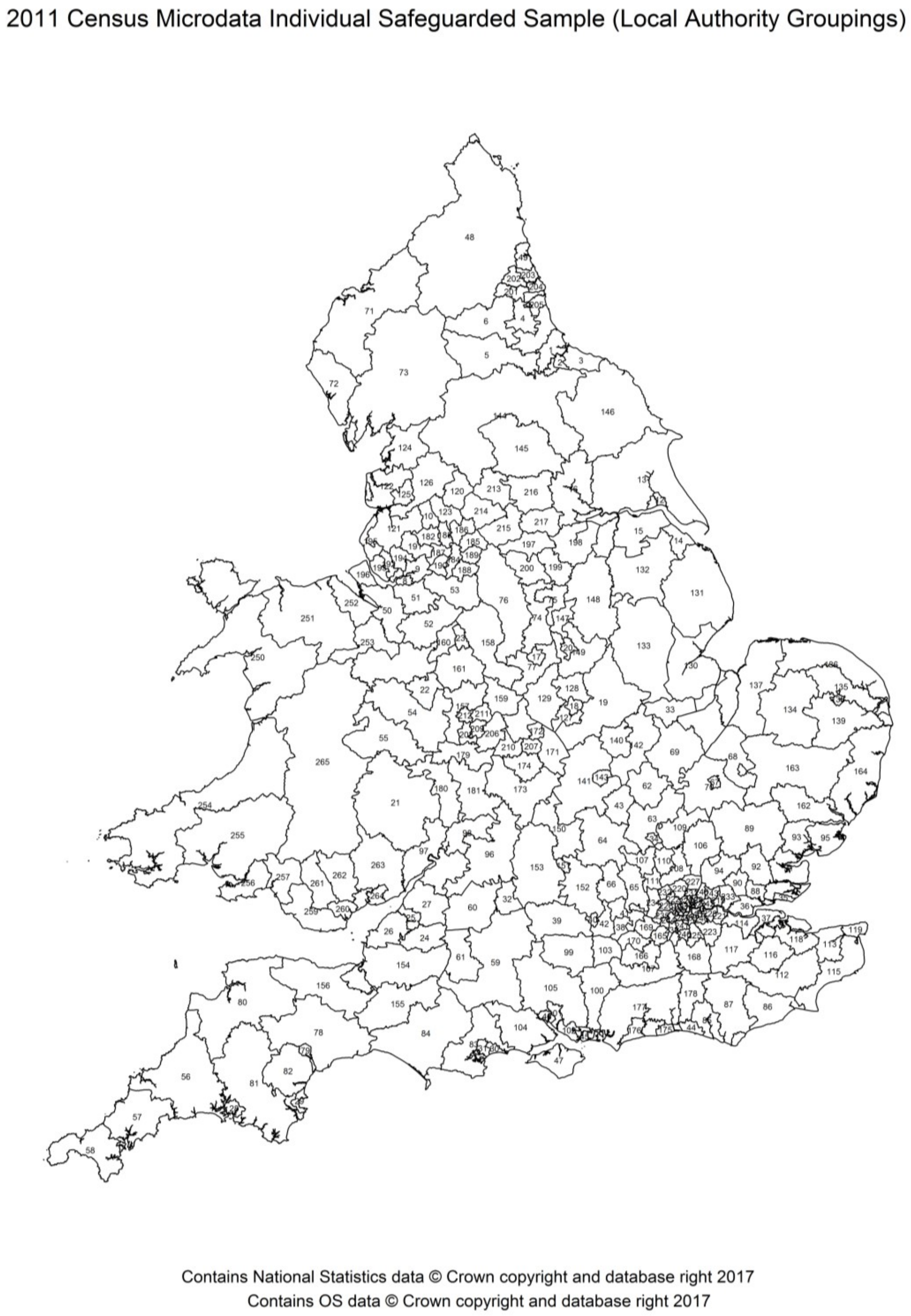
```
library(tmap)

EngWmap <- tm_shape(engw)  +
 tm_borders(col = "black") +
 tm_text("numid",
          size = 0.5) +
 tm_layout(title =
     "2011 Census Microdata Individual Safeguarded Sample (Local Authority Groupings)",
      title.size = 1.5,
      title.position = c("center", "top"),
      frame = FALSE,
      bg.color = "white",
      inner.margins = c(0.05, 0.05, 0.05, 0.05),
      outer.margins = 0.1,
      asp = 0) +
 tm_credits(
           "Contains National Statistics data © Crown copyright and database right 2017
            Contains OS data © Crown copyright and database right 2017",
            position = c("center", "bottom"), size = 1)

png("EngWmap.png", width = 11.69, height = 16.53, units = "in", res = 600)
EngWmap
dev.off()
```

The command “png” is used to save the resulting map as a png image file, which will be stored in the working directory.

IMPORTANT NOTE: The plot viewer tab in RStudio resizes plot elements to fit its current dimensions. Given that users may use diversely sized screens and can change the relative size of RStudio’s screen tiles, images displayed in the plot tab are very likely to be distorted. This is why it is not recommended to use the function “copy to clipboard”, but the standard plot saving R functions, such as “png()” or “pdf()”. To view the map displayed correctly, go to your working directory and open the saved image file.

The empty map of England and Wales should look as follows:

Figure 2: Grouped Local Authorities, England and Wales, 2011



2011 Census Microdata Individual Safeguarded Sample (Local Authority Groupings)

Contains National Statistics data © Crown copyright and database right 2017
Contains OS data © Crown copyright and database right 2017

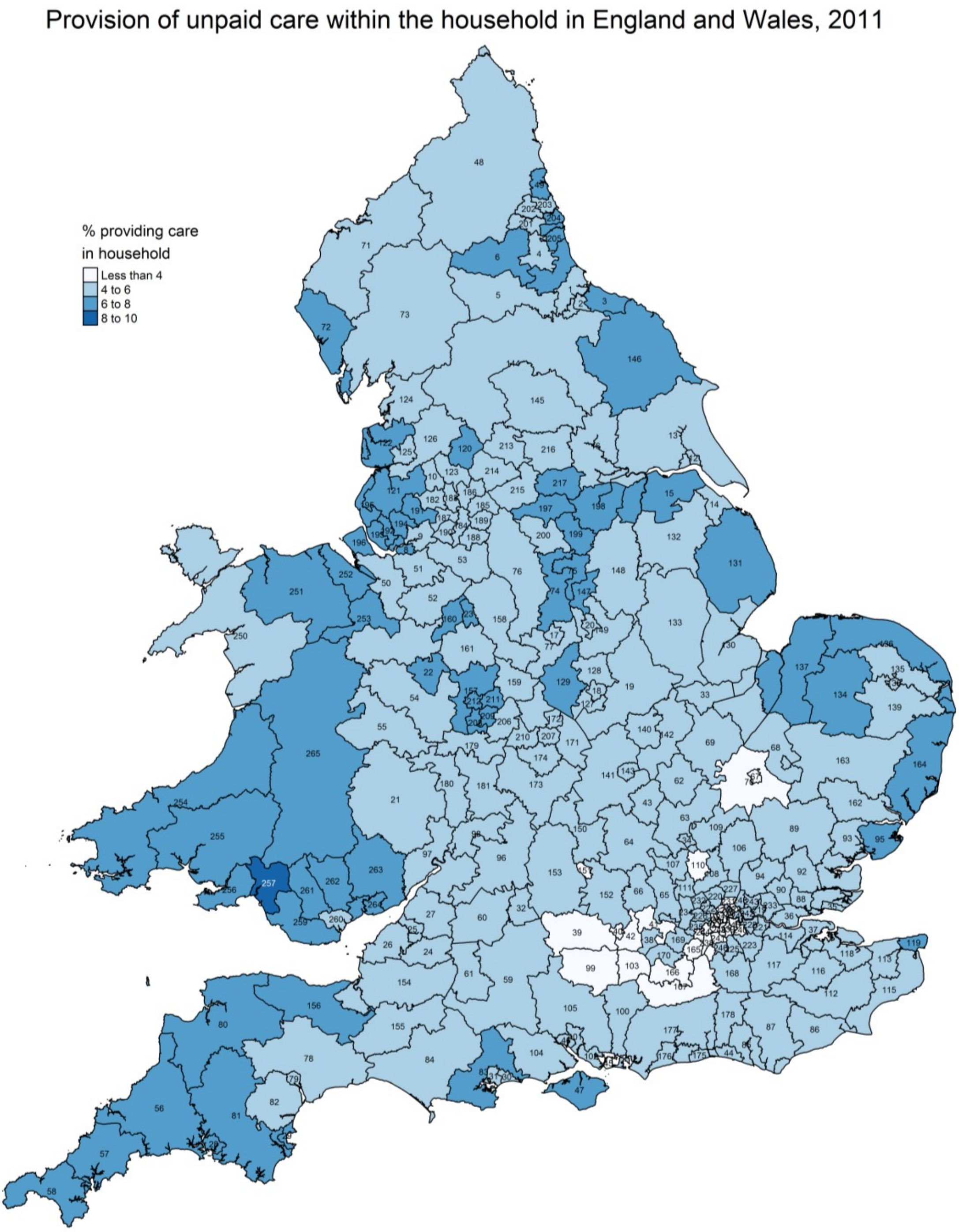
This image has been cropped to fit the page so margins can differ. Also, the layer "tm_credits" might need adjustment; "\n" can be used to add extra lines of text. After producing the empty map it is possible to add further layers to display information about the areas. For this, the "tm_polygons" layer is added to the code in the following way:

```
carermapin1 <- tm_shape(engw)  +
  tm_polygons("careinpc",
              textNA = "No data",
              title = "% providing care\nin household",
              title.size = 0.7,
              palette = "Blues",
              position = c("left", "top"),
              border.col = "black",
              breaks = c(-Inf, 4, 6, 8, 10)) +
  tm_text("numid",
          size = 0.5) +
  tm_layout(title =
           "Provision of unpaid care within the household in England and Wales, 2011",
            title.size = 1.5,
            title.position = c("center", "top"),
            frame = FALSE,
            bg.color = "white",
            inner.margins = c(0.05, 0, 0, 0),
            outer.margins = 0.1,
            asp = 0,
            legend.position = c(0.15, 0.75),
            legend.just = c("center", "bottom")) +
 tm_credits(
          "Source: UK Data Service Study Number 7682 - 2011 Census Microdata
           Individual Safeguarded Sample (Local Authority)
           Contains National Statistics data © Crown copyright and database right 2017
           Contains OS data © Crown copyright and database right 2017",
           position = c("center", "bottom"),
                        size = 1)

png("carermapin1.png", width = 11.69, height = 16.53, units = "in", res = 600)
carermapin1
dev.off()
```

This will produce a choropleth map in shades of blue with breaks as defined in the code. Breaks can be modified to add further categories but this guide uses a limited number for clarity. Colour palettes can be modified manually by using Hexadecimal RGB codes or simply use the predefined palettes; examples are: palette = "Greens" or palette = "RdYlGn" to produce a map with colour codes going from red to green.

The resulting map is displayed below:

Figure 3: Thematic map of England and Wales



Source: UK Data Service Study Number 7682 - 2011 Census Microdata Individual Safeguarded Sample (Local Authority)
Contains National Statistics data © Crown copyright and database right 2017
Contains OS data © Crown copyright and database right 2017

Some areas might be difficult to see in this map. For example, Greater London is barely distinguishable. Zooming in some areas might be worthwhile. The following code provides an example of how this can be done.

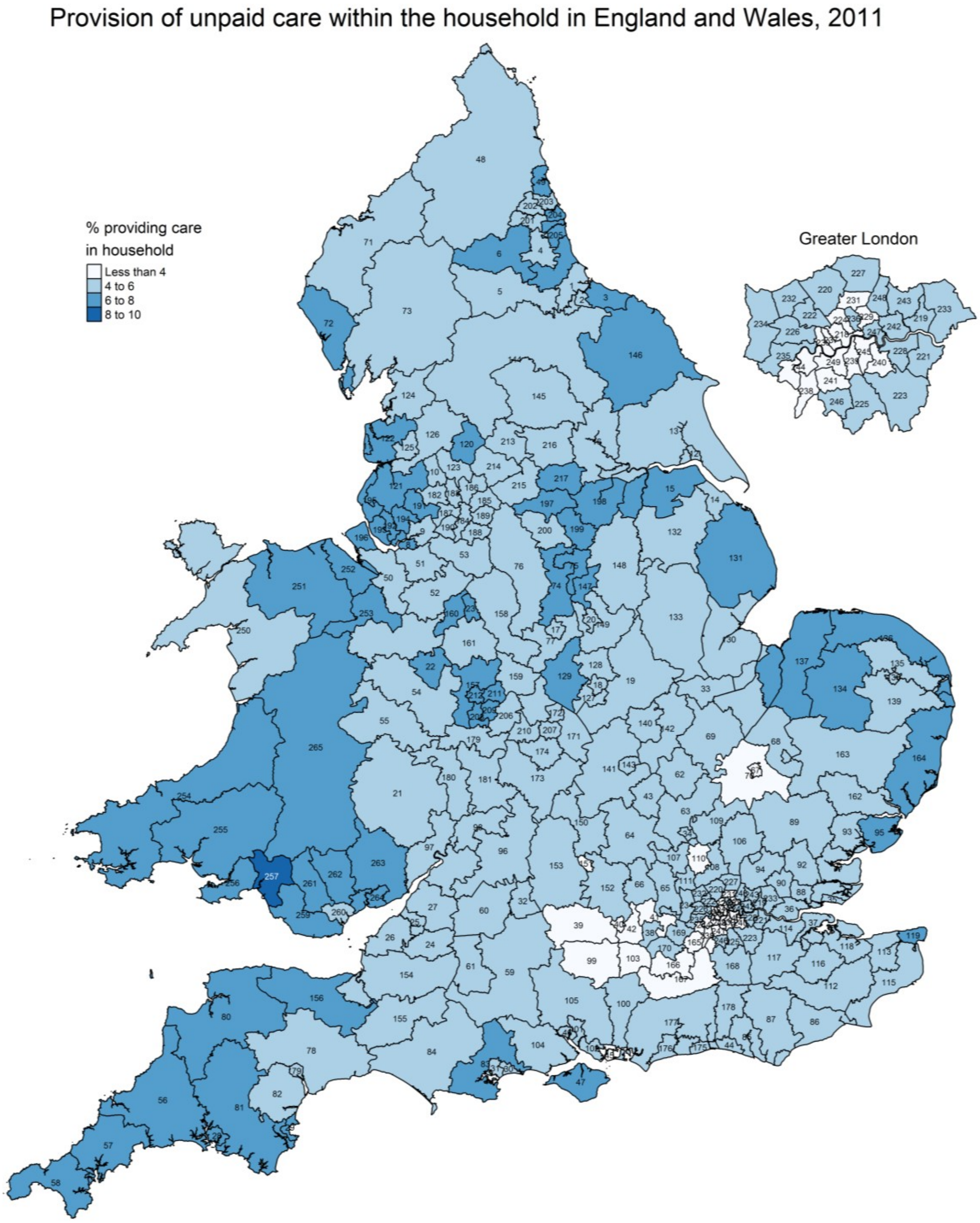
```
london <- engw[engw$numid > 217 & engw$numid < 250, ]

carermapin2 <- tm_shape(london) +
  tm_polygons("careinpc",
              textNA = "No data",
              title = "Proportion\nproviding care\nin household",
              title.size = 0.7,
              palette = "Blues",
              position = c ("left", "top"),
              border.col = "black",
              breaks = c(-Inf, 4, 6, 8, 10)) +
  tm_text("numid",
            col = "black",
            size = 0.5) +
  tm_layout(title = "Greater London",
            title.size = 1,
            title.position = c("center", "top"),
            frame = FALSE,
            bg.color = "white",
            inner.margins = c(0, 0, 0.18, 0),
            outer.margins = c(0, 0, 0, 0),
            asp = 0,
            legend.show = F)

library(grid)

png("carermapin2.png", width = 11.69, height = 16.53, units = "in", res = 600)
carermapin1
print(carermapin2,
      vp = viewport(x = 0.82,
                  y = 0.7,
                  width = 0.22,
                  height = 0.13))
dev.off()
```

The first line of code creates a subset of the shapefile representing Greater London. By using the grouped LA codes, any subset of interest can be created. The second part creates a map for Greater London only, using the newly created subset. Afterwards, the base package "grid" is called (there is no need to install as it is part of R base), to allow for the map object "carermapin2" (Greater London) to be printed in the previously created map "carermapin1" (England and Wales).

Another way of zooming in a particular zone is to use the option "bbox" after "tm_shape". For instance, the following layer can be added to "carermapin1" (previous map code): tm_shape(engw, bbox = "Greater London"). This will create a bounding box for Greater London by running an Open Street Map search query (for more details, see documentation for "tmap").

The resulting map from the code above will be as below:

Figure 4: Thematic map of England and Wales, including a zoomed image of Greater London


Provision of unpaid care within the household in England and Wales, 2011
% providing care in household
Less than 4
4 to 6
6 to 8
8 to 10
Greater London

Source: UK Data Service Study Number 7682 - 2011 Census Microdata Individual Safeguarded Sample (Local Authority)
Contains National Statistics data © Crown copyright and database right 2017
Contains OS data © Crown copyright and database right 2017

Further maps can be created using other aggregated statistics. This guide provides the R code for the following aggregated statistics: a) unpaid carers within their own household as a percentage of the total population of the grouped LA, and b) the total number of unpaid carers. Needless to say, this approach is not limited to this phenomenon and the code provided can be modified to suit other research questions.

5. References

Bivand, R., Keitt, T., & Rowlingson, B. (2017). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.2-7. Retrieved from https://cran.r-project.org/package=rgdal

Dowle, M., & Srinivasan, A. (2017). data.table: Extension of `data.frame`. R package version 1.10.4. Retrieved from https://cran.r-project.org/package=data.table

Norman, P., & Purdam, K. (2013). Unpaid caring within and outside the carer's home in England and Wales. Population, Space and Place, 39(1), 15–31.

Office for National Statistics. (2011). 2011 Census: boundary data (England and Wales) [data collection]. UK Data Service. SN:5819 UKBORDERS: Digitised Boundary Data, 1840- and Postcode Directories, 1980-. Retrieved from https://discover.ukdataservice.ac.uk/catalogue/?sn=5819&type=Data catalogue

Office for National Statistics. (2015). 2011 Census Microdata Individual Safeguarded Sample (Local Authority): England and Wales. [data collection]. UK Data Service. SN: 7682. Retrieved from http://doi.org/10.5255/UKDA-SN-7682-1

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

RStudio Team. (2016). RStudio: Integrated Development for R. Boston, USA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Tennekes, M. (2017). tmap: Thematic Maps. R package version 1.10. Retrieved from https://cran.r-project.org/package=tmap

Wickham, H., & Francois, R. (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. Retrieved from https://cran.r-project.org/package=dplyr

21 June 2017

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service delivers quality social and economic data resources for researchers, teachers and policymakers.

UK Data Census Support

UK Data Service

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL