Mapping in R: Practical Exercise

Adapted from material by Daniel Gardiner and Amy Mikhail

Thursday, 14 November 2019

Introduction

There has been an outbreak of STEC O157 involving 5 cases that visited a particular farm in the South West. An initial assessment of exposure data from their STEC enhanced surveillance questionnaires showed that all five cases reported drinking raw (unpasteurised) milk.

Whole genome sequencing results confirm that the five cases are genetically clustered; in addition another 4 cases are identified in the same genetic cluster that didn't initially report visiting the implicated farm or drinking raw milk. When the four new cases were re-interviewed, they also reported drinking raw milk that was ultimately traced back to the same farm. A sample of milk from the farm was tested and was also found to be positive for the outbreak strain.

The outbreak control team has identified a further 36 cases that are also closely related genetically to the outbreak strain; however it isn't possible to re-interview them all as some of the cases had onset dates of up to four years ago. The OCT is concerned that there might be an additional source of transmission for this strain and would like to know if the additional cases cluster geographically and how the distribution of cases has changed over time.

Aim

In order to support the outbreak investigation we want to describe the original and additional cases in terms of time, place and person. We hope to achieve this using maps created within R.

Objective

To adapt a pre-existing R script containing mapping code in order to map outbreak cases.

Questions to consider for the OCT:

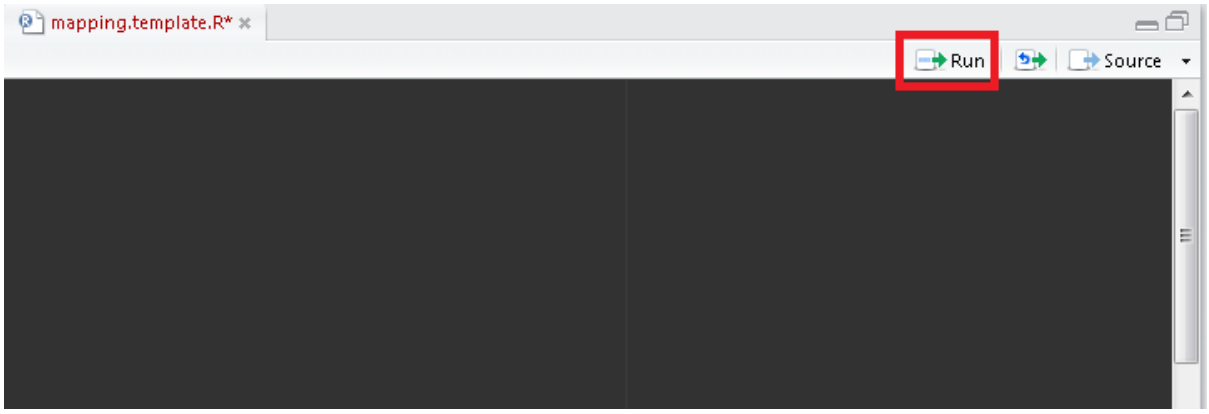
1. Are all the cases geographically clustered (by residence)?
2. Does mapping the case exposure postcodes change this picture?
3. Does stratifying / filtering the cases according to the epi definition clarify this picture?
4. Is there any difference in geographic distribution between:
    - Cases exposed to raw drinking milk vs. unexposed cases
    - Cases from the RDM clade vs. those that are not
    - Cases from different outbreaks
    - Cases with onset in different years
    - Cases with different age or sex profiles

Initial setup

Within the folder you have been provided with there is an R script titled **mapping.template.R**, this is a template R script which produces a series of maps from a dummy dataset. In this practical you will make small adaptations to this script in order to produce maps for the VTEC O157 outbreak described above. There are also three subfolders (data, R scripts, shapefiles) – these can be ignored for the purposes of the practical.

First let's check **mapping.template.R** runs properly as is.

Open **mapping.template.R**, Press 'ctrl + A' to highlight all code and press the run button to run the script (Note: using 'ctrl + enter' is the same as pressing the run button).



If an error/warning message occurs this could be due to (1) the version of R you are using (2) the version of the packages you are using (3) the location of your files (4) not having SQL server native client 11.0 installed (4) your ODBC link to the PHEGISDB (4) permission to the PHEGISDB.

Task 1

Read through the **mapping.template.R** script to gain an understanding of what the code is doing. See Appendix 1 for summary of code for each map.

We now want to start adapting the mapping template script to produce maps for our outbreak.

Task 2

Import the case study data (CaseStudy_RDM_anon_data.csv)

- **Hint**: replace "dummy.mapping.data.csv" in the read.csv() function on line 15

Look at what columns are included in the case study data

- **Hint**: use the colnames() head() or View() function

Task 3

Convert postcodes into coordinates and PHE geographies in order to plot outbreak cases.

- **Hint**: use the get.geog() function on line 35 and alter the pcodes argument, decide which postcode column to use in the case study data (see appendix 2 for a list of key variables and variable names)

Task 4

Produce a plot of cases by gender to see the overall distribution of cases in the UK using ggmap. Adapt section 02. of the R script to do this

- **Hint**: try increasing and decreasing the manual zoom for the boundary box to the map (adjust the "f" argument in the ggmap::make_bbox function) until the zoom level best fits the distribution of the cases.

Task 5

Produce a plot of cases using a shapefile. Run section 03. to do this. Change the size of the points

- **Hint**: use the size argument within the geom_point() function

Task 6

Plot cases grouped by infection status (primary, co-primary or secondary). Adapt section 04. to do this.

- **Hint**: change the group argument in the ggplot() function
- **Hint**: the scale_colour_manual() function is used to specify colours, in the template script only two colours are specified, either add more colours to allow for more groups "#08519C", "#000000" or remove this line and default colours will be used

Task 7

Case clustering is hard to visualise when plotted using a spot map due to overlap. Plot cases using a heat map in order to visualise where cases are most concentrated. Run section 05. to do this. Play with the arguments to change the look of the map.

- **Hint**: to re-run previously executed code use 'ctrl + shift + P'
- **Hint**: adjust the bins argument within the stat_density2d() function
- **Hint**: adjust the low and high arguments within the scale_fill_gradient() function
- **Hint**: adjust the range argument within the scale_alpha() function
- **Hint**: add a colour argument to the geom_density2d() function e.g. colour = "red"
- **Hint**: add a size argument to the geom_density2d() function e.g. size = 0.5

Task 8

Use a choropleth map to visualise case density by PHE Centre. Run section 06. of the R script to do this. Adjust the arguments to change the look of the map.

- **Hint**: adjust the colour and size arguments in the geom_path() function
- **Hint**: adjust the palette and breaks arguments within the scale_fill_distiller () function

Task 9

View cases stratified by year to visualise case incidence over time. Run section 07. to do this. Then view cases by month in order to visualise seasonality. Adapt section 07. to do this.

- **Hint**: use the month column
- **Hint**: adjust the ncol argument in the facet_wrap() function

Task 10

Plot cases on an interactive map. Run section 08. to do this. Then convert markers into clustered markers

- **Hint**: update the popup argument on lines 280 - 282
- **Hint**: to convert into a clustered map use clusterOptions = markerClusterOptions()

Appendix 1 – description of data operations in the R script

01. Read in (import) your data and map cases to geographic coordinates:

a. Set working directory
b. Read in data
c. Remove the lab.postcode field
d. Look at imported data
e. Load MRAtools function to convert postcodes into coordinates/PHE geography
f. Apply MRAtools function

02. Now let's try plotting a map - using google maps as a backdrop for our cases

a. Load necessary packages (ggmap, ggplot2)
b. Create a boundary box around the map using longitude and latitude to define the size of the box, then decide on the zoom level (f) to determine how much of the map outside the boundary box will be visible.
c. Download a map from google maps that is centred on the boundary box of the cases.
d. Plot the google map in the background.
e. Plot the cases onto the map, stratifying by sex, using ggplot2 to add this layer to the map, identify the variable to stratify by and identify the colours to use when plotting the cases of each sex.
f. View the map to see if it is ok or needs any adjustments
g. Save the map as a .pdf file

03. Another point map - this time with shapefiles:

a. Load necessary packages: maptools
b. Import shapefile
c. Fortify shapefile (i.e. convert the shape into a format useful for plotting)
d. Add a dummy column titled group onto the data
e. Plot the fortified shapefile using longitude for x-axis and latitude for y-axis, and using the group column to so that polygons are correctly connected
f. Specify to use geom_polygon to plot the polygons
g. Use coord_fixed to fix a specified ratio between x-axis and y-axis
h. Use theme_nothing to remove the legend
i. Use geom_point to plot cases in the data data.frame using easting on x-axis and northing on y-axis

04. Another point map - stratified by sex:

a. Same as in 03. except adding a colour aesthetic to the geom_point function to indicate points are to be coloured by sex

05. Where are the cases most concentrated? Let's see with a heat map:

a. Same as in 03. except using geom_density2d, stat_density2d, scale_fill_gradient, scale_alpha functions to add heat map layer
b. geom_density2d adds contours onto map
c. stat_density2d adds density colouring
d. scale_fill_gradient specifies the density of the colour gradient
e. scale_alpha specifies the transparency of the colour gradient

06. Viewing case density by PHE Centre: a choropleth map

a. Load necessary packages: maptools
b. Import shapefile
c. Fortify shapefile (i.e. convert the shape into a format useful for plotting)
d. Tabulate cases by PHE geography (centre)
e. Merge counts of cases by PHE geography onto fortified shapefile
f. Plot the fortified shapefile using longitude for x-axis and latitude for y-axis, and using the group column to so that polygons are correctly connected
g. Specify to use geom_polygon to plot the polygons with fill argument defined as the column containing the count of cases for each PHE centre
h. Use coord_fixed to fix a specified ratio between x-axis and y-axis
i. Use theme_nothing to remove the legend
j. Use guides to order legend colours from highest down to lowest
k. Use scale_fill_distiller to set colours

07. Time series maps - show cases by year of onset:

a. Same as 03. except use facet_wrap to stratify by year

08. Interactive map:

a. Create an empty leaflet map, add Open Street Map tiles, add markers using longitude/latitude and popup text

Appendix 2 – description of key variables

Variable name	Description
RDM clade (rdm.clade)	Indicates if the case is a member of the phylogenetic clade (group of genetically closely related isolates) associated with the raw drinking milk outbreak. 1 = yes; 0 = no.
Barton farm (barton.farm)	Indicates if the case reported visiting the farm with milk contaminated with the outbreak strain.
Any outbreak (any.outbreak)	Indicates if the case was associated with any outbreak (yes) or a sporadic case (no).
Case type (case.type)	Name of the outbreak (if case is associated with an outbreak) or sporadic.
Year (year)	Year of onset (or specimen date if asymptomatic)
Month (month)	Month of onset (or specimen date if asymptomatic)
Age (age)	Age of the case in years
Postcode (postcode)	Postcode of residence
UK travel exposure postcode (uk.travel.exposure.postcode)	As described (if case travelled in UK) or postcode of residence (if no travel)
Exposure postcode (exposure.postcode)	Postcode of location where the case was exposed to raw drinking milk (if known) or postcode of residence (if not)
Gender (gender)	Sex of the case (F = female; M = male)
Microbiology case definition (microdef)	Confirmed (outbreak strain in faeces); Probable (faeces negative but seropositive for O157 and epilinked to a confirmed case); Suspect (no lab results but epilinked to a confirmed case)
Epidemiological definition (epidef)	Primary (index case in the household) Co – primary (same onset date as the index case in the household) Secondary (onset date after the household index case) Unsure (no onset date as asymptomatic)
Unpasteurised milk (unpast)	Indicates if the case reported consuming unpasteurised milk on their enhanced surveillance questionnaire

Appendix 3 – spatial analysis approach

Background

It appears that a 50-SNP whole genome sequencing cluster around the outbreak strain is associated with the South West of England. When all cases within this cluster are plotted, their distribution is wider than the SW with cases appearing in every PHE Centre. No cases appear around the farm linked to the outbreak. The outbreak was spatially complex; the following account details the approach taken to improve the accuracy of the spatial analysis.

- **Cases are not resident near the farm linked to the outbreak:**

Cases linked to the outbreak received milk via a delivery service or ordered the product over the internet. Sporadic cases may have travelled to the SW region, been linked to outbreaks in the region or may have eaten food produced in the area.

- **Information needed to produce a more accurate map linked to exposure:**

Locations and dates of travel within the UK, locations of settings where cases linked to outbreaks were thought to be exposed to infection and dates of these outbreaks. Details of secondary and asymptomatic cases is also important.

- **Dealing with cases that are also linked to other outbreaks:**

Cases associated with a particular outbreak are not independent. Ignoring this lack of independence may result in an underestimation of variability, increase the probability of rejecting a null hypothesis or result in incorrect estimates of measures of association.

- **Controlling for the spatial effects of cases linked to more than one outbreak:**

Possible techniques for controlling for clustering may include the random selection of one case from each outbreak to remove the clustering effect by design rather than through a statistical technique. Analytical weights could also be used to down-weight the impact of multiple cases from a single outbreak.

Enhanced surveillance allows for the identification of travel destinations in the UK.

- **During this outbreak investigation, the factors above were controlled as follows:**
    - Use of primary cases only where cases were linked to household outbreaks (two or more cases at the same postcode).
    - Single location of exposure for cases linked to outbreaks.
    - Location of travel to the SW area in the 10 days before onset.

- **Further analysis with SatScan to explore the strength of the spatial relationship with the farm/South West Region:**
    - Adjusting maximum cluster size and population at risk.
    - Using the Bernoulli model for a case-control approach.
    - Using space time models.