Areal aggregation on maps

18 marks

In this question, you will investigate the effect of various aggregation methods on some artificial data.

All data here will be in the form of an $m \times n$ rectangular array $\mathbf{X}$, say. The $[i, j]$ cell of $\mathbf{X}$ is some measurement on the $[i, j]$ township (or town).

Here $[i, j]$ determines the geographic location of the township (or town) $[i, j]$ – the row number gives the township's North-South geographic position (the higher the row number, the farther south the town/township lies) and the column number gives the township's West-East geographic position (the higher the column number, the farther east is the town/township).

A choropleth map display of the information in $\mathbf{X}$ will assume that every township/town has a rectangular shape and the same area.

For example, the number of houses in each township/town (all $28 = 4 \times 7$ of them) could be given by the matrix

$$\mathbf{X} = \begin{pmatrix} 2000 & 200 & 100 & 200 & 100 & 200 & 100 \\ 200 & 100 & 200 & 100 & 200 & 100 & 200 \\ 100 & 200 & 100 & 4000 & 100 & 200 & 100 \\ 100 & 200 & 100 & 200 & 100 & 200 & 1500 \end{pmatrix} = \texttt{NumHouses}$$

and displayed by a three level choropleth map

Number of Houses
South <-----------> North
West <-----------> East

where the colour saturation of the region for any township/town number is determined from the number of houses in that township/town.

To investigate the various effects of areal aggregation, a number of functions have been written for your use and may be found in the file `ArealAggregation.R` on the class web site. Note that this file requires some functionality from the `R` package called `RColorBrewer`, which you will need to install if you haven't already done so.

This file contains two important functions: `col_areas`, and `AggregateByID`. The first of these produces the choropleth (as above for the number of houses), the second constructs a new aggregation depending on which townships/towns belong to the same region.

Influenza. Suppose that there has been an outbreak of influenza in the region given by these townships/towns and that, in addition to the number of houses in each township/town, we also have the number of influenza cases there.

Load the data from the file `Data_a.R` to yield the two data matrices `NumHouses` and `NumCases`. The function `col_areas` can be used to produce the above choropleth for `NumHouses` by executing: `col_areas(NumHouses, main="Number of Houses")`

These data can be aggregated over different regions by specifying which townships/towns belong to which regions. This is accomplished by defining a matrix of region IDs (having the same dimensions as that of the townships/towns).

For example, the following matrix of region IDs

$$\texttt{RegionID1} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{pmatrix}$$

organizes the townships/towns into three East-West regions and is defined in the file `Data_a.R`. Aggregating the data at the level of these regions is had by executing

```
NumHouses1 <- AggregateByID(NumHouses, RegionID1)
```

for the number of houses; `NumCases1` would be similarly defined. The regional incidence of influenza would then be calculated as:

```
CasesPerHouse1 <- NumCases1/NumHouses1
```

a. **(2 marks)** Construct the choropleth for the ratio `CasesPerHouse1`. What does it indicate about the geographic distribution of influenza incidence?

b. **(3 marks)** A different definition of regions is given by `RegionID2`. Construct the appropriate choropleth for these regions and summarize what it indicates about the geographic distribution of influenza incidence?

c. **(3 marks)** A third definition of regions is given by `RegionID3`. This definition distinguishes population centres from the rural background. Construct the appropriate choropleth for these regions and summarize what it indicates about the geographic distribution of influenza incidence?

d. **(2 marks)** What do you conclude about aggregations by political regions as a means to understand the spatial distribution of influenza incidence? Be clear in your reasoning.

Load the data from the file `Data_b.R` to yield two new data matrices `NumHouses` and `NumCases`.

e. **(3 marks)** Construct the choropleth for the influenza incidence based on the townships/towns as regions. Summarize what this display indicates about the geographic distribution of influenza incidence?

f. **(3 marks)** The choropleth just constructed uses five-number summary to determine the colours of the townships/towns. In this way, each category has 25% of the data (i.e. boundaries are at the quartiles).

   The function `col_areas` also takes an argument `breaks` and so the same choropleth would be produced by

```
col_areas(NumCases/NumHouses, breaks=fivenum(NumCases/NumHouses),
          main="Influenza incidence")
```

   Rather than use the quartiles as boundaries, construct a choropleth for these data where the five break points (including the minimum and the maximum) are equally spaced from the minimum to the maximum. Summarize what this display indicates about the geographic distribution of influenza incidence?

g. **(2 marks)** What do you conclude about basing conclusions about the spatial distribution of influenza incidence based on data aggregated by the value of the incidence itself? Be clear in your reasoning.