Midterm-Project_ST5226

Name: Zhu Xu; User ID:E0337988; Student ID:A0191344H

13 October 2018

Sudden Infant Death Syndrome (SIDS) refers to the sudden unexplained death of a child less than one year of age. The data is contained in NC_4.rds. The SpatialPolygonsDataFrame nc contains counts of the number of live births (BIR74) and SIDS (SID74) cases in the 100 counties of North Carolina for the period July 1, 1974 to June 30, 1978.

Question-1:

Interpret the raw.rates vector that is being computed below:
> nc$raw.rates <- nc$SID74 / nc$BIR74 / 4 * 1e04

Solution-1:

```r
library(sp)
nc <- readRDS("C:/Users/MSI/Desktop/Notes/ST5226-Spatial Statistics/Mid-project/nc_4.rds")
#nc$SID74---the number of live births during 4 years.(SID74)
#nc$BIR74---the number of SIDS cases during 4 years.(BIR74)
```

nc$raw.rates <- nc$SID74/nc$BIR74/4*1e04
It means: $\frac{death}{birth} \times \frac{1}{4} \times 10^4$, i.e. the number of SIDS cases per year per 10^4 births.

Question-2:

Plot raw rates in chloropleth map. Explain why some smoothing of the rates is necessary.

Solution-2:

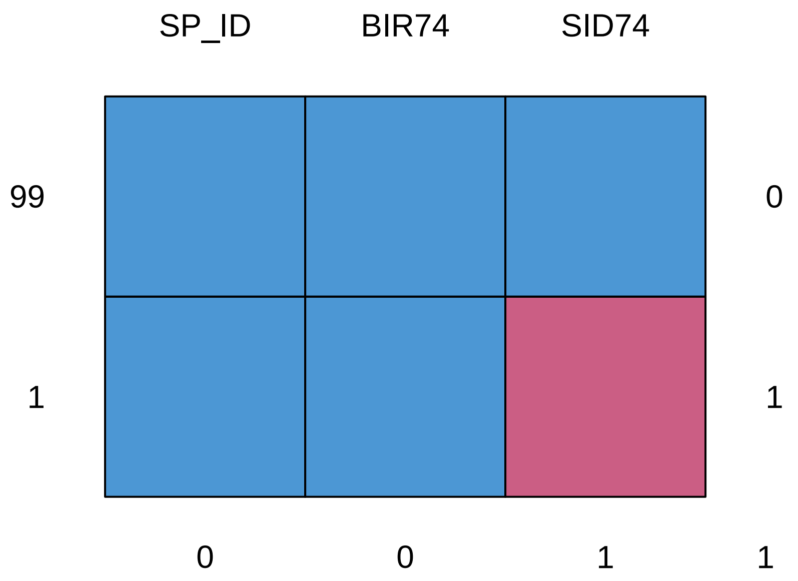
```r
library(mice)
str(nc, max.level = 2)
```

```
## Formal class 'SpatialPolygonsDataFrame' [package "sp"] with 5 slots
##   ..@ data       :'data.frame':  100 obs. of  23 variables:
##   .. ..- attr(*, "data_types")= chr [1:20] "C" "N" "N" "N" ...
##   ..@ polygons   :List of 100
##   ..@ plotOrder  : int [1:100] 82 24 78 9 92 71 10 51 31 7 ...
##   ..@ bbox       : num [1:2, 1:2] -84.3 33.9 -75.5 36.6
##   .. ..- attr(*, "dimnames")=List of 2
##   ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
```

```r
str(slot(nc,"data"), max.level = 1)
```

```
## 'data.frame':    100 obs. of  23 variables:
##  $ SP_ID    : Factor w/ 100 levels "37001","37003",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ CNTY_ID  : num  1904 1950 1827 2096 1825 ...
##  $ east     : num  278 179 183 240 164 138 406 411 321 353 ...
##  $ north    : num  151 142 182 75 176 154 118 148 53 6 ...
##  $ L_id     : num  1 2 1 3 1 1 2 1 4 4 ...
##  $ M_id     : num  3 2 2 2 2 2 4 4 3 3 ...
##  $ names    : Factor w/ 100 levels "Alamance","Alexander",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ AREA     : num  0.111 0.066 0.061 0.138 0.114 0.064 0.203 0.18 0.225 0.212 ...
##  $ PERIMETER: num  1.39 1.07 1.23 1.62 1.44 ...
##  $ CNTY_    : num  1904 1950 1827 2096 1825 ...
##  $ NAME     : Factor w/ 100 levels "Alamance","Alexander",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ FIPS     : Factor w/ 100 levels "37001","37003",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ FIPSNO   : num  37001 37003 37005 37007 37009 ...
##  $ CRESS_ID : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ BIR74    : num  4672 1333 487 1570 1091 ...
##  $ SID74    : num  13 0 0 NA 1 0 7 6 8 5 ...
##  $ NWBIR74  : num  1243 128 10 952 10 ...
##  $ BIR79    : num  5767 1683 542 1875 1364 ...
##  $ SID79    : num  11 2 3 4 0 0 4 5 5 6 ...
##  $ NWBIR79  : num  1397 150 12 1161 19 ...
##  $ raw.74   : num  5.57 0 0 19.11 1.83 ...
##  $ EB       : num  5.02 2.68 3.41 9.68 3.39 ...
##  $ DA       : num  3.87 1.86 2.01 10.26 1.73 ...
##  - attr(*, "data_types")= chr  "C" "N" "N" "N" ...
```

```r
raw_data <- slot(nc, "data")[, c("SP_ID", "SID74", "BIR74")]
md.pattern(raw_data) # display missing-data patterns in raw_data
```

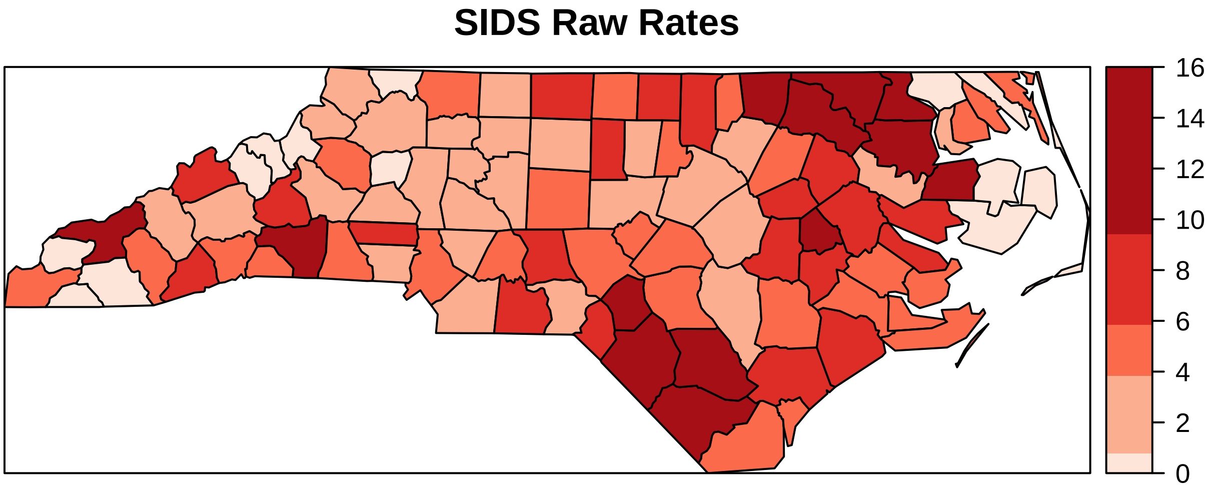


```
##    SP_ID BIR74 SID74  
## 99     1     1     1 0
## 1      1     1     0 1
##        0     0     1 1
```

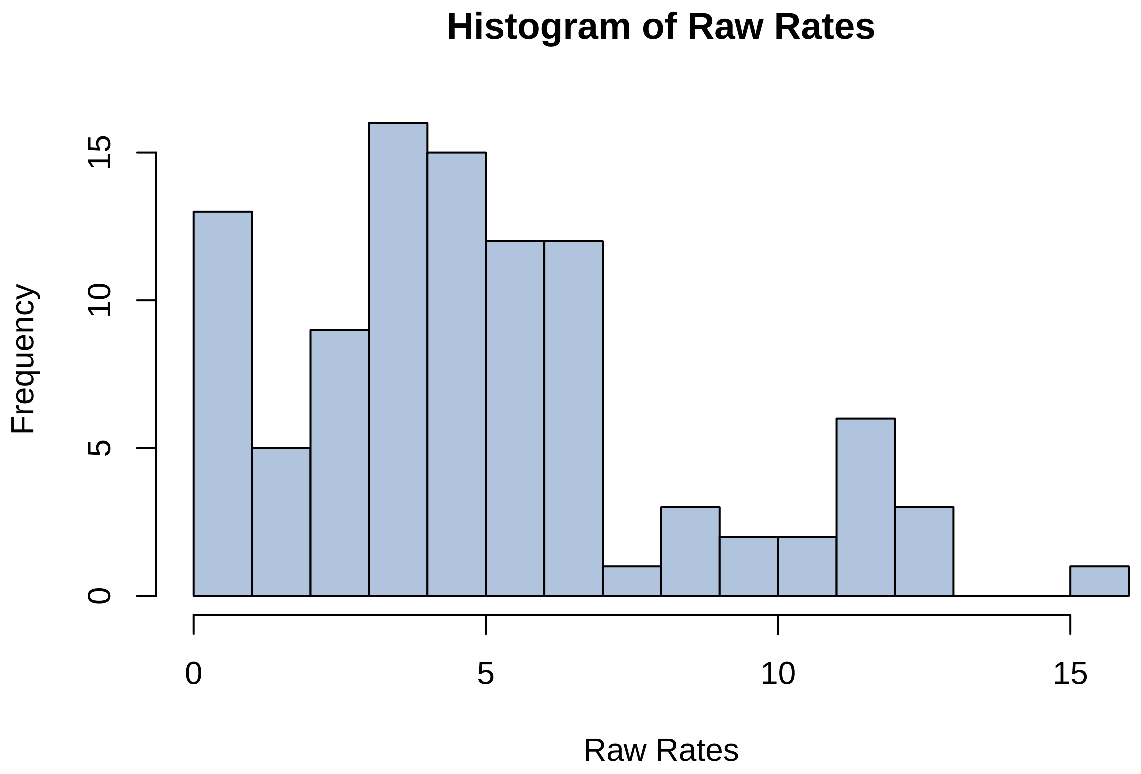
```r
library(Hmisc)
#locate the missing-data and fill in missing value with median
nc$SID74 <- impute(nc$SID74,median)  #nc[["SID74"]][4] <- impute(raw_data$SID74,median)
```

```r
nc_data <- slot(nc, "data")[, c("SP_ID", "SID74", "BIR74")]
nc$raw.rates <- nc$SID74/nc$BIR74/4*1e04 #compute the raw.rates
```

```r
library(RColorBrewer)
library(classInt)
pal <- brewer.pal(n=5, "Reds")
intervals <- classIntervals(nc$raw.rates, n=5, "fisher")
intervals$brks[6] <- 16
#make a chloropleth map of raw rates
figure_raw.rates <- spplot(
  nc, c("raw.rates"), col.regions=pal, at=intervals$brks, main="SIDS Raw Rates")
figure_raw.rates
```



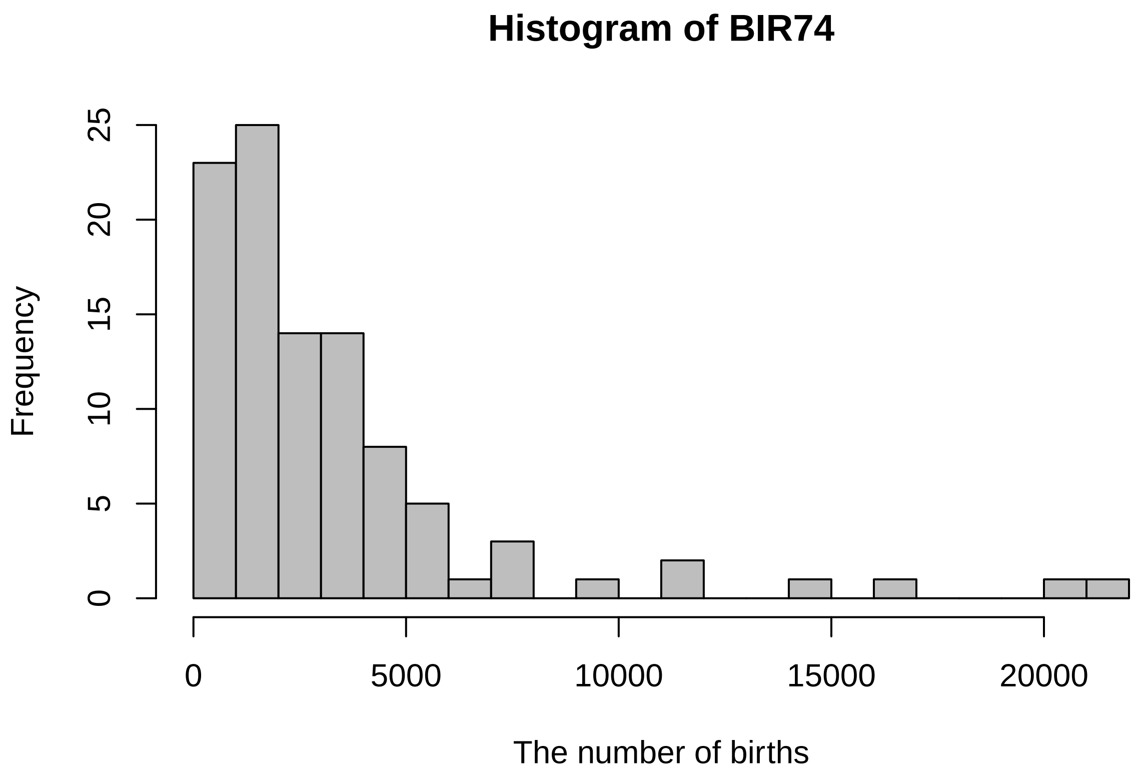
```r
hist(nc$raw.rates, breaks=20, xlab="Raw Rates",
     main="Histogram of Raw Rates", col="lightsteelblue")
```



```r
nc_data[nc_data$SID74 == 0, ]
```

```
##       SP_ID SID74 BIR74
## 37003 37003     0  1333
## 37005 37005     0   487
## 37011 37011     0   781
## 37029 37029     0   286
## 37043 37043     0   284
## 37055 37055     0   521
## 37073 37073     0   420
## 37075 37075     0   415
## 37095 37095     0   338
## 37113 37113     0   797
## 37121 37121     0   671
## 37177 37177     0   248
## 37199 37199     0   770
```

```
#make a histogram of the number of births
hist(nc$BIR74, breaks=20, xlab="The number of births",
     main="Histogram of BIR74", col="grey")
```



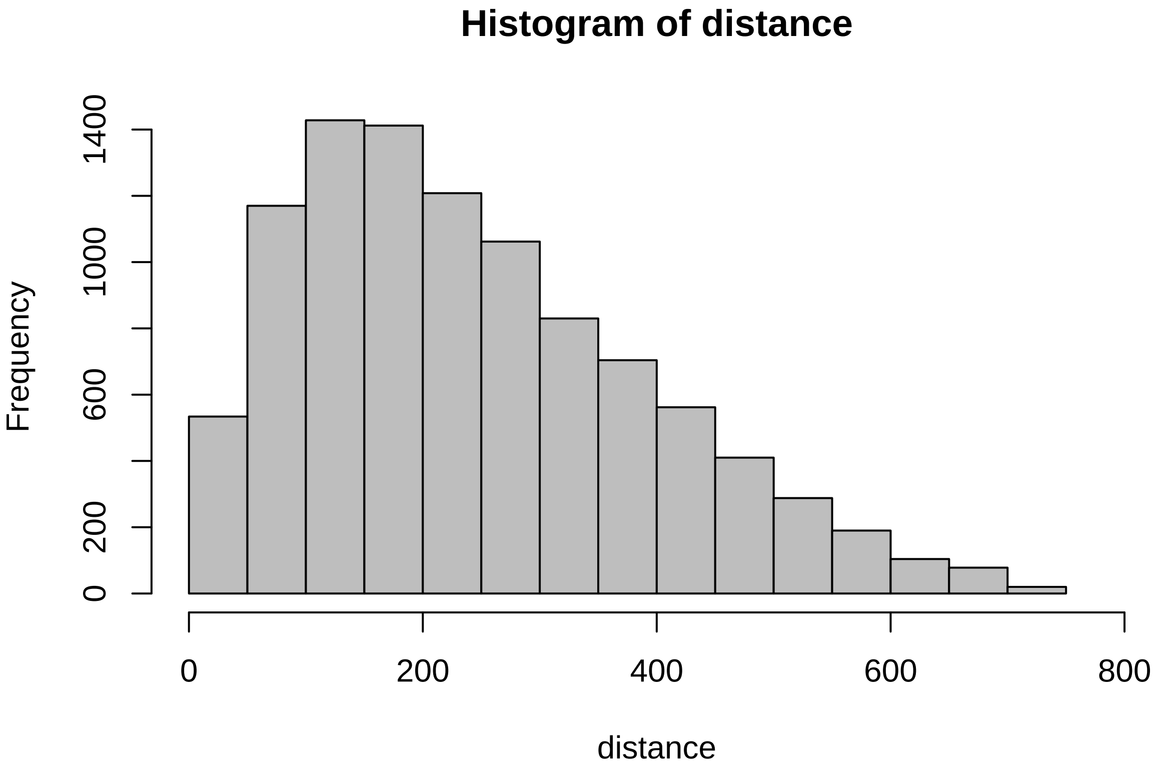
we can clearly see we are being impacted by the small number problem, i.e. there are regions with very small counts and small population sizes. Moreover the distribution of population sizes is quite skewed. So some of smoothing of the rates is necessary.

Question-3:

Consider smoothing the raw rates using the weighted average smoother we covered in class. Propose the weights that you think best for this data. Explain your result.

Solution-3:

```r
distance <- spDists(coordinates(nc), longlat=T)
hist(distance, breaks=20,xlim=c(0,800), col="grey")
```



So we choose the distance = 50 km

Method_1: $(d_{ij} = 50, w_{ij} = 1 \quad or \quad 0)$

Consider smoothing the raw rates using the expression below:

$$\hat{r_i} = \frac{\sum_{j=1}^{100} w_{ij} Y_j}{\sum_{j=1}^{100} w_{ij} n_j}$$

where

$\hat{r_i}$ is the smoothed rate for county i.

w_{ij} is the smoothing weight

$$w_{ij} = \begin{cases} 1, d_{ij} \leq 50 \\ 0, d_{ij} > 50 \end{cases}$$

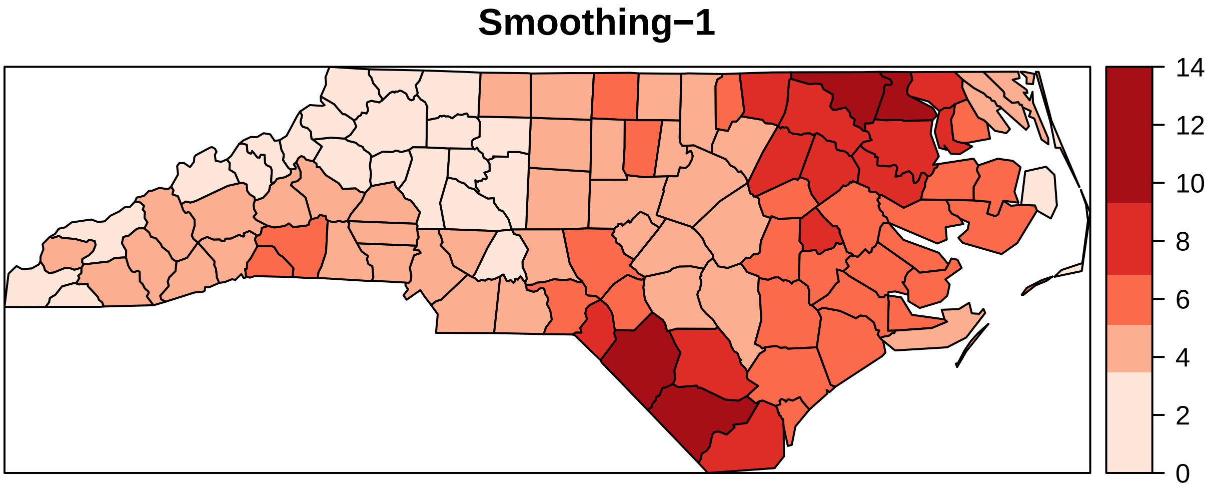
d_{ij} is the distance in kilometres between i and j.

Y_j is the number of SIDS cases in county j.

n_j is the number of births in county j.

```r
w_ij_1 <- ifelse(distance<=50, 1, 0)
nc$smoothed_rates_1 <- as.vector(w_ij_1 %*% nc$SID74 / w_ij_1 %*% nc$BIR74 /4 *1e04)
intervals_a <- classIntervals(nc$smoothed_rates_1, n=5, "fisher")
intervals_a$brks[6] <- 14
figure_smoothing_1 <- spplot(
  nc, "smoothed_rates_1", col.regions=pal, at=intervals_a$brks, main="Smoothing-1")
```

```
figure_smoothing_1
```



Method_2: $(d_{ij} = 50, w_{ij} = [1 - (\frac{d_{ij}}{50})^3]^3 \quad or \quad 0)$

Consider smoothing the raw rates using the expression below:

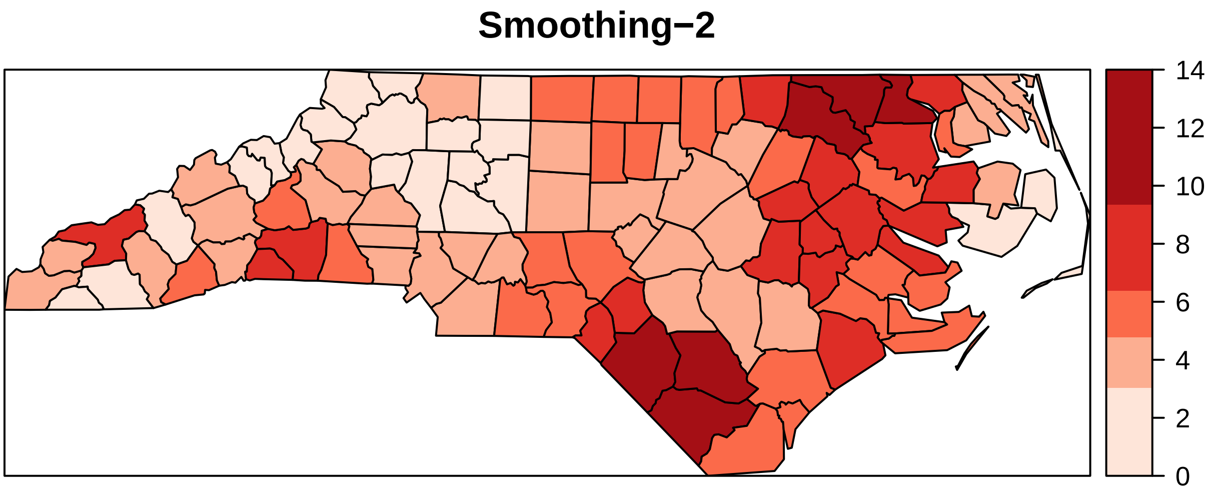
$$\hat{r_i} = \frac{\sum_{j=1}^{100} w_{ij} Y_j}{\sum_{j=1}^{100} w_{ij} n_j}$$

where

w_{ij} is the smoothing weight

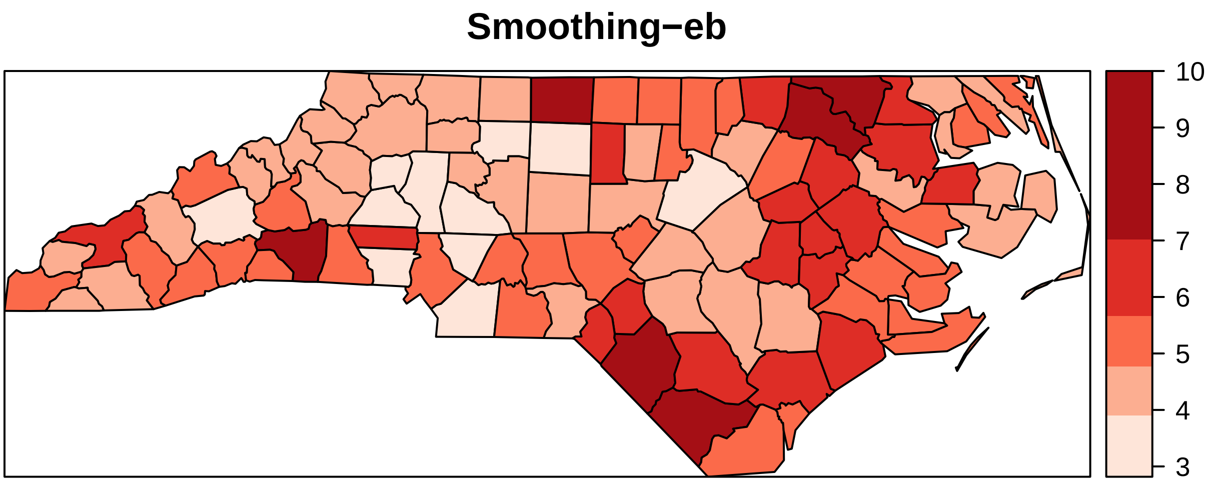
$$w_{ij} = \begin{cases} [1 - (\frac{d_{ij}}{50})^3]^3, d_{ij} \leq 50 \\ 0, \qquad\qquad\quad d_{ij} > 50 \end{cases}$$

```
w_ij_2 <- ifelse(distance<=50, (1-(distance/50)^3)^3, 0)
nc$smoothed_rates_2 <- as.vector(w_ij_2 %*% nc$SID74 / w_ij_2 %*% nc$BIR74 /4 *1e04)
intervals_b <- classIntervals(nc$smoothed_rates_2, n=5, "fisher")
intervals_b$brks[6] <- 14
figure_smoothing_2 <- spplot(
  nc, "smoothed_rates_2", col.regions=pal, at=intervals_b$brks, main="Smoothing-2")
figure_smoothing_2
```



Method_3:(Empirical Bayes)

```r
library(spdep)
library(rgdal)
nc$smoothed_rates_eb <- EBest(nc$SID74, nc$BIR74)$estmm /4 *1e04
intervals_eb <- classIntervals(nc$smoothed_rates_eb, n=5, "fisher")
intervals_eb$brks[6] <- 10
figure_smoothing_eb <- spplot(
  nc, "smoothed_rates_eb", col.regions=pal, at=intervals_eb$brks, main="Smoothing-eb")
figure_smoothing_eb
```

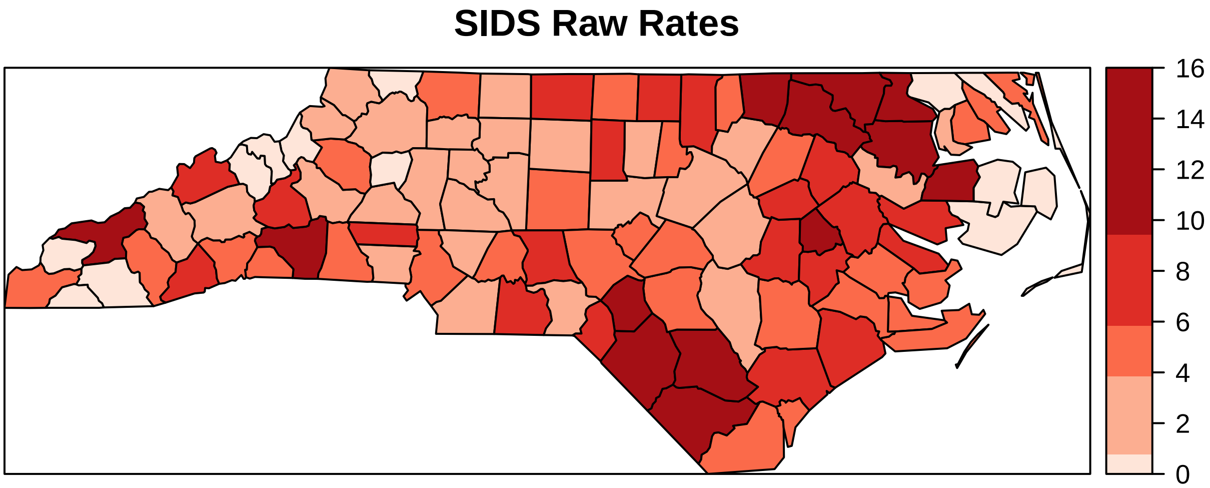


Question-4:

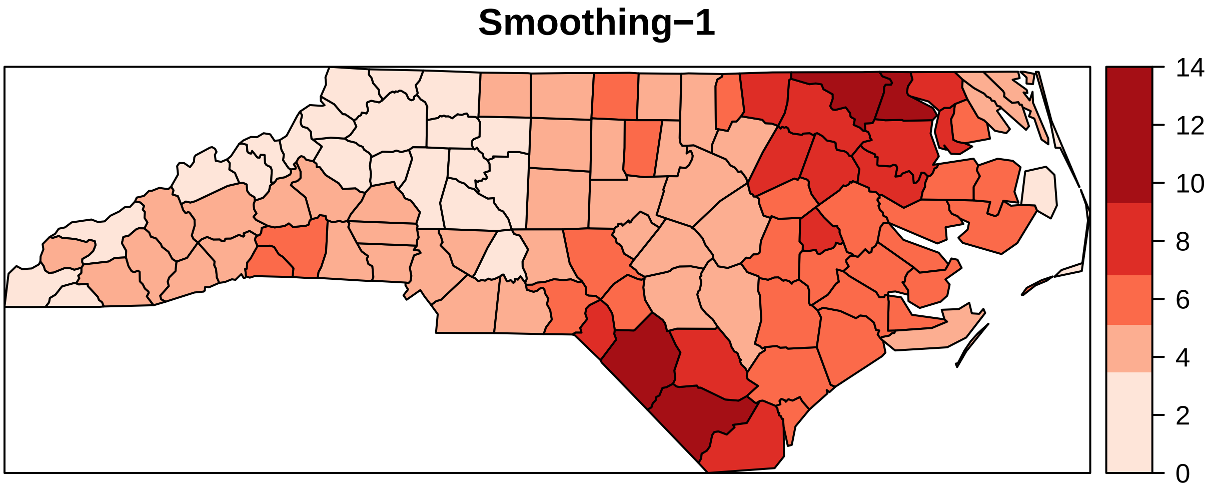
Compare your proposed smoother with the raw rates using chloropleth map. Interpret your discoveries.

Solution-4:

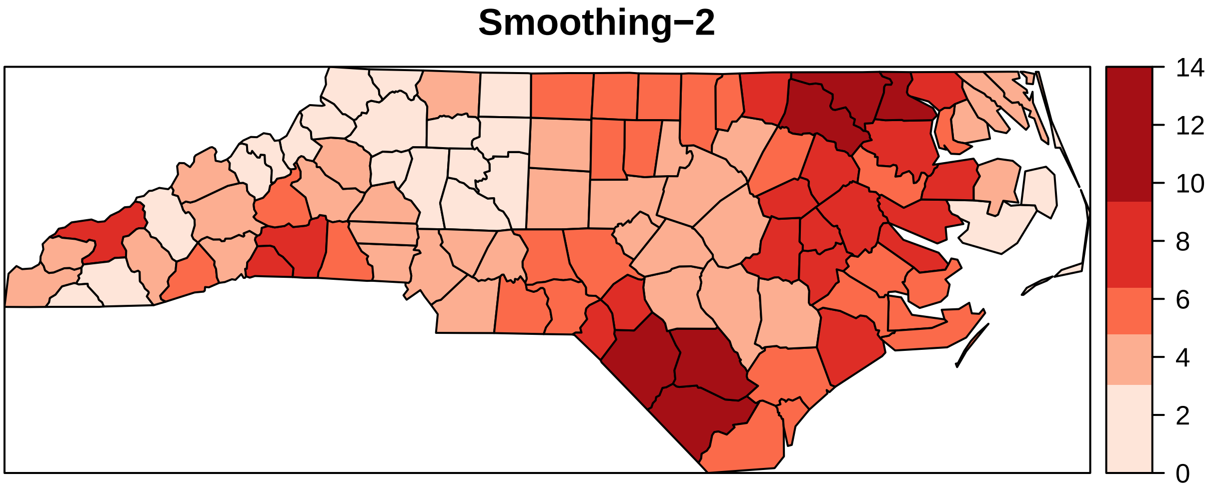
```
figure_raw.rates
```



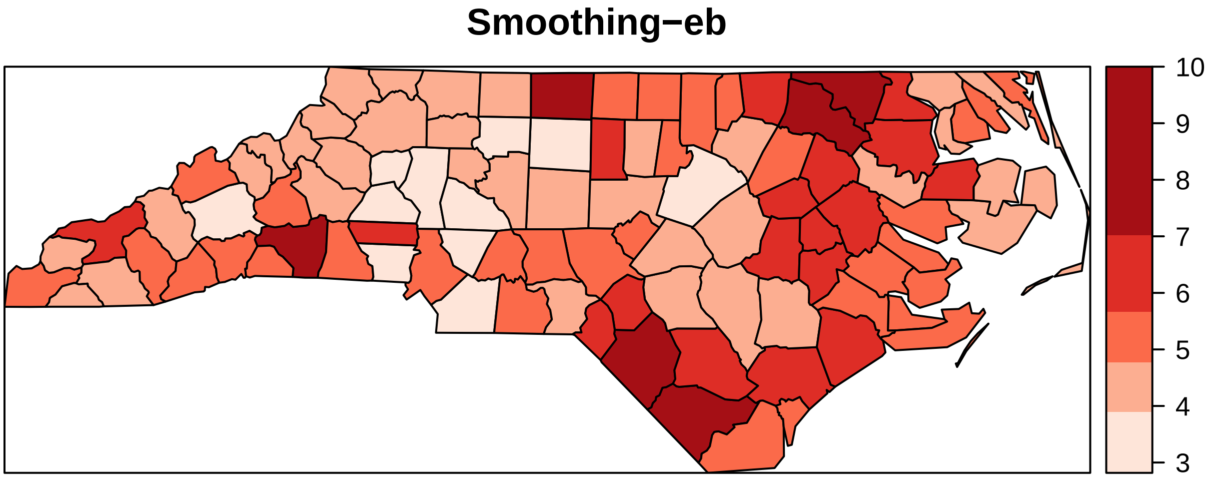
```
figure_smoothing_1
```



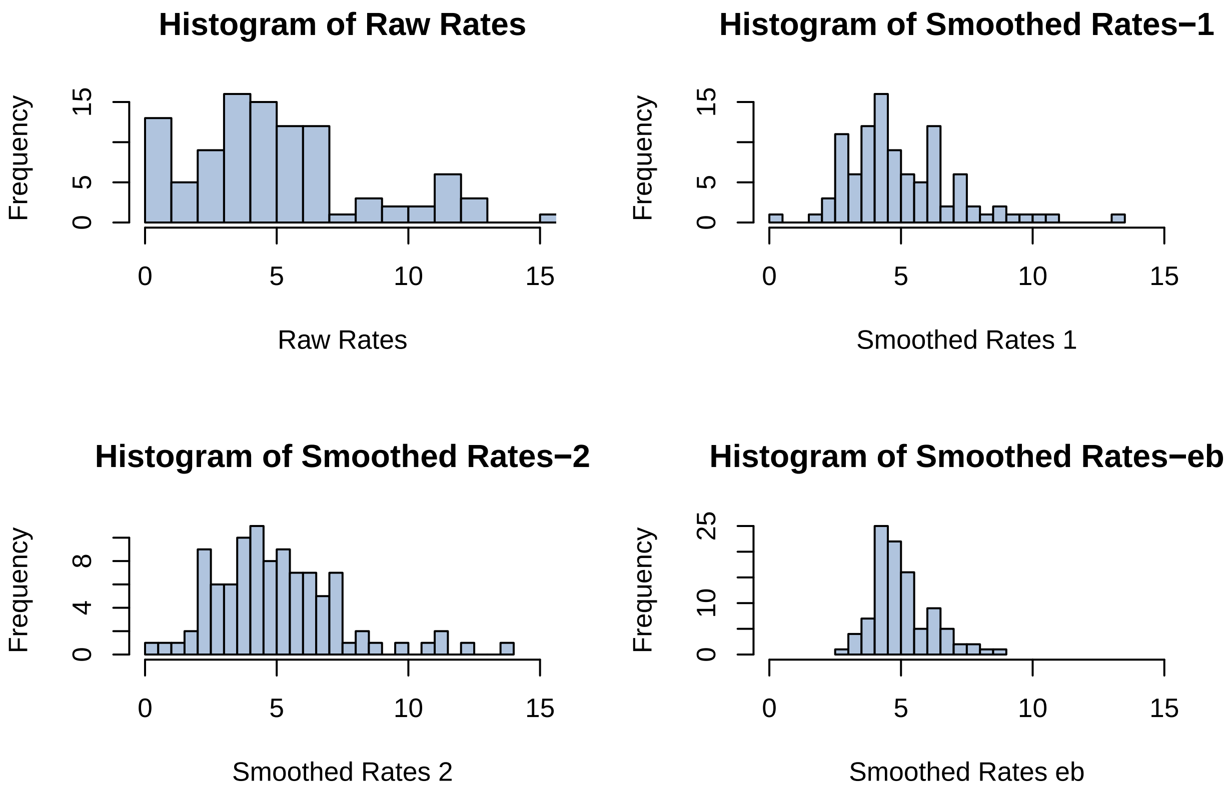
```
figure_smoothing_2
```



```
figure_smoothing_eb
```



```r
par(mfrow=c(2,2))
hist(nc$raw.rates, breaks=20, xlab="Raw Rates",
     xlim=c(0,15), main="Histogram of Raw Rates", col="lightsteelblue")
hist(nc$smoothed_rates_1, breaks=20, xlab="Smoothed Rates 1",
     xlim=c(0,15), main="Histogram of Smoothed Rates-1", col="lightsteelblue")
hist(nc$smoothed_rates_2, breaks=20, xlab="Smoothed Rates 2",
     xlim=c(0,15), main="Histogram of Smoothed Rates-2" , col="lightsteelblue")
hist(nc$smoothed_rates_eb, breaks=20, xlab="Smoothed Rates eb",
     xlim=c(0,15), main="Histogram of Smoothed Rates-eb", col="lightsteelblue")
```



We can assume that, $Y_j \sim Possion(\xi n_j)$ for all counties within 50 km.
Then:

$$Var(\hat{r_i}) = \frac{\sum w_{ij}^2 \xi n_j}{(\sum w_{ij} \xi n_j)^2} \le \frac{\sum w_{ij} \xi n_j}{(\sum w_{ij} \xi n_j)^2} = \frac{\xi}{\sum w_{ij} \xi n_j} \le \frac{\xi}{n_j} = Var(\frac{Y_i}{n_j})$$

It's clear that the smoothed map has a smaller range, and provides a clearer suggestion of regions that should be studied further.We can see the $var(\hat{r_{i2}})$ of method-2 $<$ the $var(\hat{r_{i1}})$ of method-1 via the histograms of smoothed rates, and there is an interesting note that using method-3(EB) we could get a "better" result, but actually the smoothed rates range is too small and method-3 is over-smoothing.So,the method-2, i.e.$w_{ij} = [1 - (\frac{d_{ij}}{50})^3]^3$ is more suitable.