Spatial Statistics Lab 7

Onyedikachi J Okeke

2024-03-02

Load the library.
```
library(sp)
library(spatstat)
library(sf)
library(spatstat.geom)
library(ctv)
library(terra)
library(spdep)
library(rgdal)
library(rgdal)
library(terra)
library(RColorBrewer)
library(classInt)
library(epitools)
library(DCluster)
```

Library
```
## Loading required package: spatstat.data
## Loading required package: spatstat.geom
## spatstat.geom 3.2-8
## Loading required package: spatstat.random
## spatstat.random 3.2-2
## Loading required package: spatstat.explore
## Loading required package: nlme
## spatstat.explore 3.2-5
## Loading required package: spatstat.model
## Loading required package: rpart
## spatstat.model 3.2-8
## Loading required package: spatstat.linnet
## spatstat.linnet 3.1-3
##
## spatstat 3.0-7
## For an introduction to spatstat, type 'beginner'
## Linking to GEOS 3.11.2, GDAL 3.7.2, PROJ 9.3.0; sf_use_s2() is TRUE
## terra 1.7.65
##
## Attaching package: 'terra'
## The following objects are masked from 'package:spatstat.geom':
##
##     area, delaunay, is.empty, rescale, rotate, shift, where.max,
##     where.min
## Loading required package: spData
```

```
## Please note that rgdal will be retired during October 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
## See https://r-spatial.org/r/2023/05/15/evolution4.html and
https://github.com/r-spatial/evolution
## rgdal: version: 1.6-7, (SVN revision 1203)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.6.2, released 2023/01/02
## Path to GDAL shared files: C:/Users/GIS/AppData/Local/R/win-
library/4.3/rgdal/gdal
##  GDAL does not use iconv for recoding strings.
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 9.2.0, March 1st, 2023, [PJ_VERSION: 920]
## Path to PROJ shared files: C:\Program
Files\PostgreSQL\14\share\contrib\postgis-3.2\proj
## PROJ CDN enabled: FALSE
## Linking to sp version:2.1-0
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp
or rgdal.
##
## Attaching package: 'rgdal'
## The following object is masked from 'package:terra':
##
##     project
## Loading required package: boot
##
## Attaching package: 'boot'
## The following object is masked from 'package:spatstat.explore':
##
##     envelope
## Loading required package: MASS
##
## Attaching package: 'MASS'
## The following object is masked from 'package:terra':
##
##     area
## The following object is masked from 'package:spatstat.geom':
##
##     area
```

Q1 Loading Sids data

```
data(nc.sids)
sids<-data.frame(Observed=nc.sids$SID74)
#combine the expected number of SID cases to the data frame
sids<-cbind(sids,
Expected=nc.sids$BIR74*sum(nc.sids$SID74)/sum(nc.sids$BIR74))
#combine x, y coordinates to the data frame
sids<-cbind(sids, x=nc.sids$x, y=nc.sids$y)
```

GAM using the centroids of the areas in data

```r
#step is the step of the grid, alpha is the Significance level of the tests
performed.
sidsgam<-opgam(data=sids, radius=30, step=10, alpha=.002)

#Plot centroids
plot(sids$x, sids$y, xlab="Easting", ylab="Northing")
#Plot points marked as clusters
points(sidsgam$x, sidsgam$y, col="red", pch="*")
title("Alpha level = 0.002")

#Change the alpha (Significance level of the tests performed) to 0.05 and
rerun the GAM
sidsgam05<-opgam(data=sids, radius=30, step=10, alpha=.05)

#Plot centroids
plot(sids$x, sids$y, xlab="Easting", ylab="Northing")
#Plot points marked as clusters
points(sidsgam05$x, sidsgam05$y, col="red", pch="*")
title("Alpha level = 0.05"
```
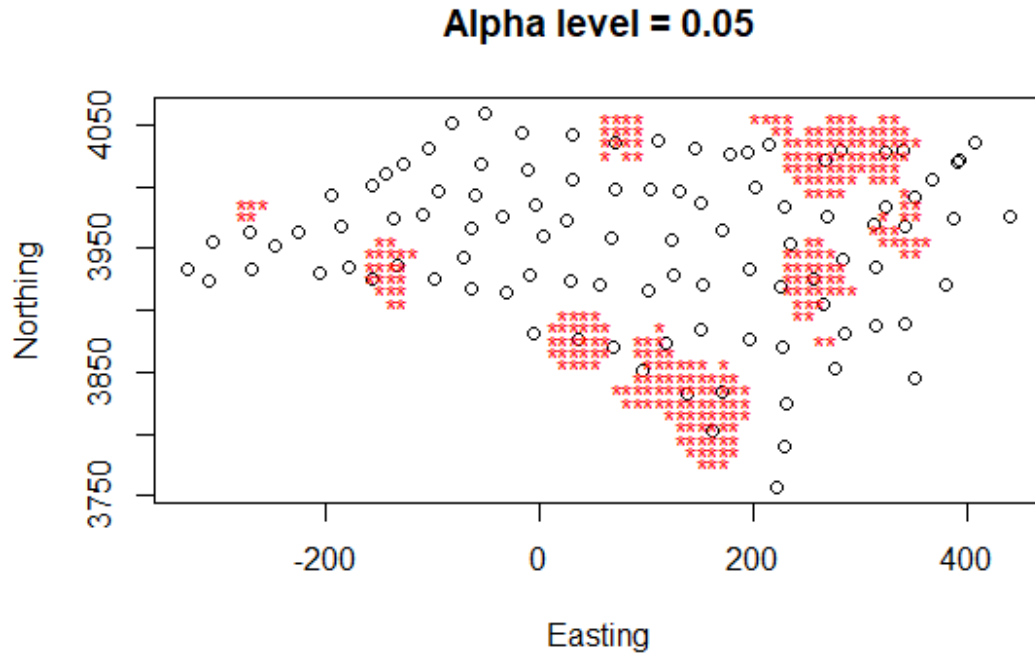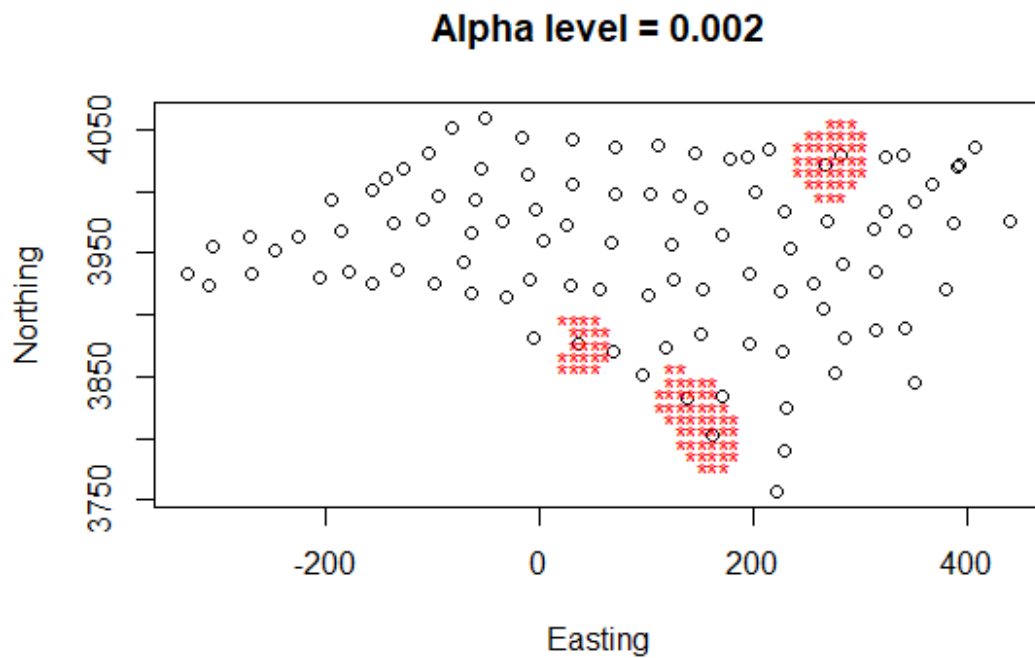
Alpha level = 0.002

Alpha level = 0.05

Q1. Describe the pattern Q2. Change the alpha (Significance level of the tests performed) to 0.05 and rerun the GAM. Describe the differences between it and pattern from Q1.

The difference between the two images lies in the alpha level used for the significance of the tests performed to identify these clusters.

In the first image, with an alpha level of 0.002, fewer points are marked as clusters. This is because a lower alpha level means that the criterion for identifying a cluster is more stringent, so only the strongest clusters are identified. The clusters appear to be more concentrated and localized.

In the second image, with an alpha level of 0.05, more points are marked as clusters. A higher alpha level decreases the stringency of the cluster identification criterion, leading to more points being identified as part of clusters. This results in a larger number of clusters that are more spread out across the grid.

The pattern of the cluster's changes significantly between the two alpha levels. With a more stringent alpha level (0.002), the clusters appear to be smaller and more discrete. When the alpha level is less stringent (0.05), the clusters become larger and more numerous, indicating a more liberal interpretation of what constitutes a cluster. This reflects a fundamental trade-off in statistical testing: a lower alpha level reduces the chance of a Type I error (false positive), but increases the chance of a Type II error (false negative), and vice versa.

Kulldorff's spatial scan statistics

```
sids<-data.frame(Observed=nc.sids$SID74)
sids<-cbind(sids,
Expected=nc.sids$BIR74*sum(nc.sids$SID74)/sum(nc.sids$BIR74))
#combine risk population, and x, y coordinates to the data frame
sids<-cbind(sids, Population=nc.sids$BIR74, x=nc.sids$x, y=nc.sids$y)
# Kulldorff's spatial scan statistics method over the centroids
# there are different models and here we use negative binomial
mle<-calculate.mle(sids, model="negbin") # Calculate parameters in sampling
procedures

knresults<-opgam(data=sids, thegrid=sids[,c("x","y")], alpha=.05,
iscluster=kn.iscluster, fractpop=.15, R=99, model="negbin", mle=mle)
#Plot all centroids and significant ones in red

plot(sids$x, sids$y, main="Kulldorff and Nagarwalla's method, Fractpop =
0.15")
points(knresults$x, knresults$y, col="red", pch=19)

#Plot the cluster with the highest likelihood ratio test in green
clusters<-get.knclusters(sids, knresults)
idx<-which.max(knresults$statistic)
points(sids$x[clusters[[idx]]], sids$y[clusters[[idx]]], col="green", pch=19)

#Change the maximum fraction of the total population to 0.5 and rerun the
test
knresults05<-opgam(data=sids, thegrid=sids[,c("x","y")], alpha=.05,
iscluster=kn.iscluster, fractpop=.5, R=99, model="negbin", mle=mle)
```

```
#Plot all centroids and significant ones in red
plot(sids$x, sids$y, main="Kulldorff and Nagarwalla's method, Fractpop =
0.5")
points(knresults05$x, knresults05$y, col="red", pch=19)

#Plot the cluster with the highest likelihood ratio test in green
clusters<-get.knclusters(sids, knresults05)
idx<-which.max(knresults05$statistic)
points(sids$x[clusters[[idx]]], sids$y[clusters[[idx]]], col="green", pch=19)
```
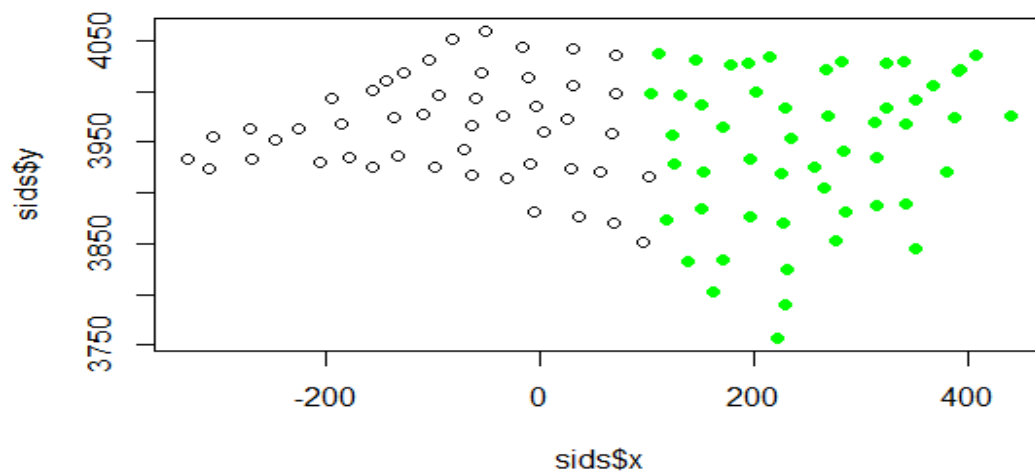
Result

Kulldorff and Nagarwalla's method, Fractpop = 0.15
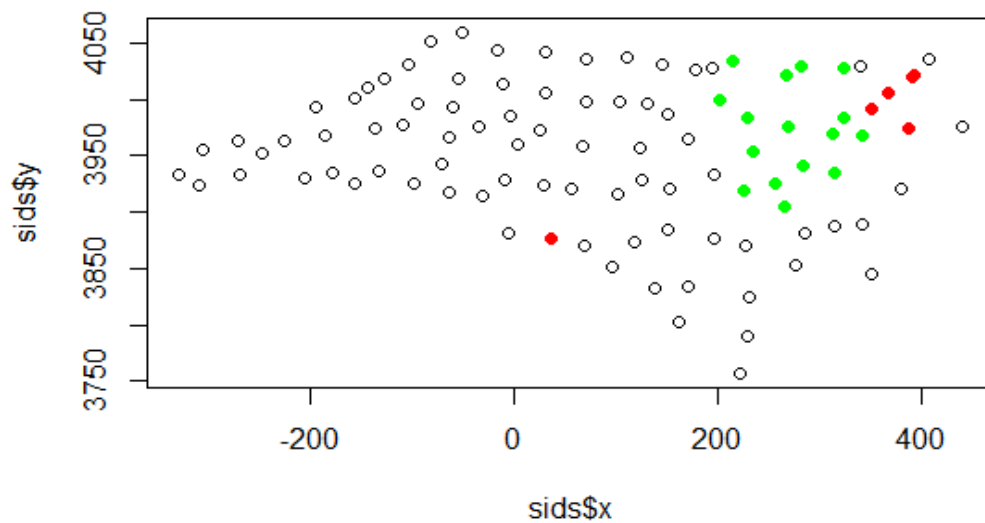
Kulldorff and Nagarwalla's method, Fractpop = 0.5

Q3. Describe the pattern.

The test identifies clusters of points that are statistically significant, with red points indicating significant individual points and green points indicating the cluster with the highest likelihood ratio test.

Pattern of the Two Images:

- **First Image (Fractpop = 0.15):** There is a mix of red and green points. The green points, representing the most statistically significant cluster, are grouped together towards the upper right quadrant, suggesting a localized area of high significance. The red points are fewer and are scattered among the non-significant points (black points), indicating other areas with some level of significance but not as strong as the green cluster.

- **Second Image (Fractpop = 0.5):** All the significant points are colored green, indicating a single large cluster. This suggests that when the maximum fraction of the total population for the test is increased to 0.5, the method detects a more extensive cluster of significance that encompasses a larger area and includes more points.

Q4. Describe the differences between it and pattern from Q3.

1. **Identification of Clusters:** In the Q3, clusters were identified with red asterisks, and the significance was adjusted by the alpha level. In these images, clusters are identified by red and green points, and the significance is adjusted by the "Fractpop" parameter, which seems to influence the size and identification of clusters.

2. **Fractpop Parameter:** The difference in the "Fractpop" parameter between the two images here shows a clear change in the pattern of clustering. With a lower "Fractpop" value, individual significant points and a most significant cluster are identified. With a higher "Fractpop" value, a larger, more inclusive cluster is identified. This suggests that the "Fractpop" parameter controls the maximum proportion of the population that can be included in a cluster, which in turn affects the sensitivity of the cluster detection.

In summary, the "Fractpop" parameter has a substantial impact on the detection of clusters, with higher values leading to larger, more inclusive clusters. This contrasts with the alpha level adjustments in the GAM method, which seemed to control the stringency of cluster identification rather than the size of the clusters.

Besag and Newell's Statistic for Spatial Clustering

```
sids<-data.frame(Observed=nc.sids$SID74)
sids<-cbind(sids,
Expected=nc.sids$BIR74*sum(nc.sids$SID74)/sum(nc.sids$BIR74))
sids<-cbind(sids, x=nc.sids$x, y=nc.sids$y)

#Where K = 10
bnresults10<-opgam(sids, thegrid=sids[,c("x","y")], alpha=.05,
iscluster=bn.iscluster, set.idxorder=TRUE, k=10, model="poisson",
R=100, mle=calculate.mle(sids) )
#Plot all the centroids
```

```r
plot(sids$x, sids$y)
#Plot significant centroids in red
points(bnresults10$x, bnresults10$y, col="red", pch=19)
title("Besag and Newell's test, K = 10")

#Where K = 20
bnresults20<-opgam(sids, thegrid=sids[,c("x","y")], alpha=.05,
iscluster=bn.iscluster, set.idxorder=TRUE, k=20, model="poisson",
R=100, mle=calculate.mle(sids) )
#Plot all the centroids
plot(sids$x, sids$y)
#Plot significant centroids in red
points(bnresults20$x, bnresults20$y, col="red", pch=19)
title("Besag and Newell's test, K = 20")

#Where K = 30
bnresults30<-opgam(sids, thegrid=sids[,c("x","y")], alpha=.05,
iscluster=bn.iscluster, set.idxorder=TRUE, k=30, model="poisson",
R=100, mle=calculate.mle(sids) )
#Plot all the centroids
plot(sids$x, sids$y)
#Plot significant centroids in red
points(bnresults30$x, bnresults30$y, col="red", pch=19)
title("Besag and Newell's test, K = 30")

#Where K = 40
bnresults40<-opgam(sids, thegrid=sids[,c("x","y")], alpha=.05,
iscluster=bn.iscluster, set.idxorder=TRUE, k=40, model="poisson",
R=100, mle=calculate.mle(sids) )
#Plot all the centroids
plot(sids$x, sids$y)
#Plot significant centroids in red
points(bnresults40$x, bnresults40$y, col="red", pch=19)
title("Besag and Newell's test, K = 40")
```
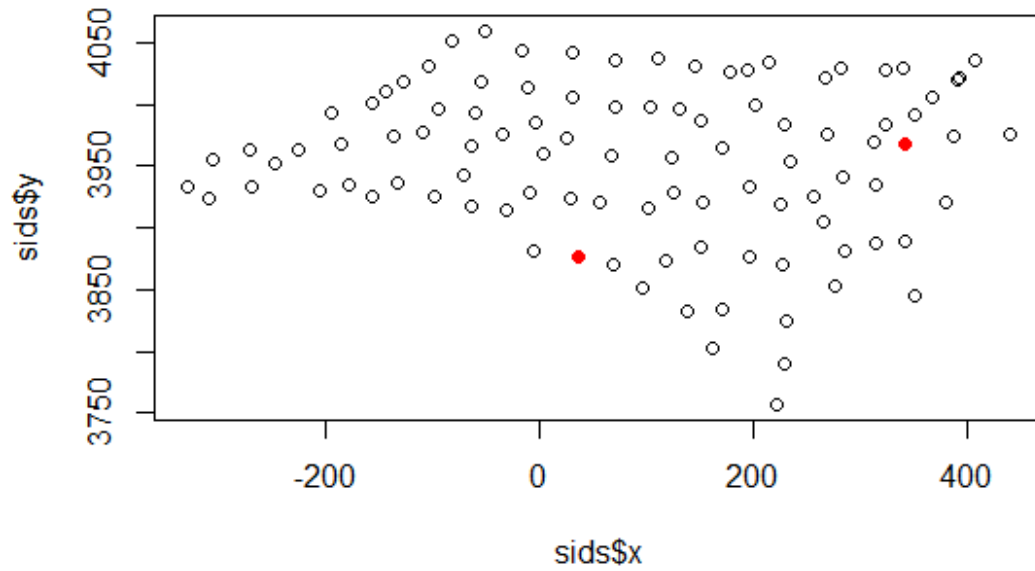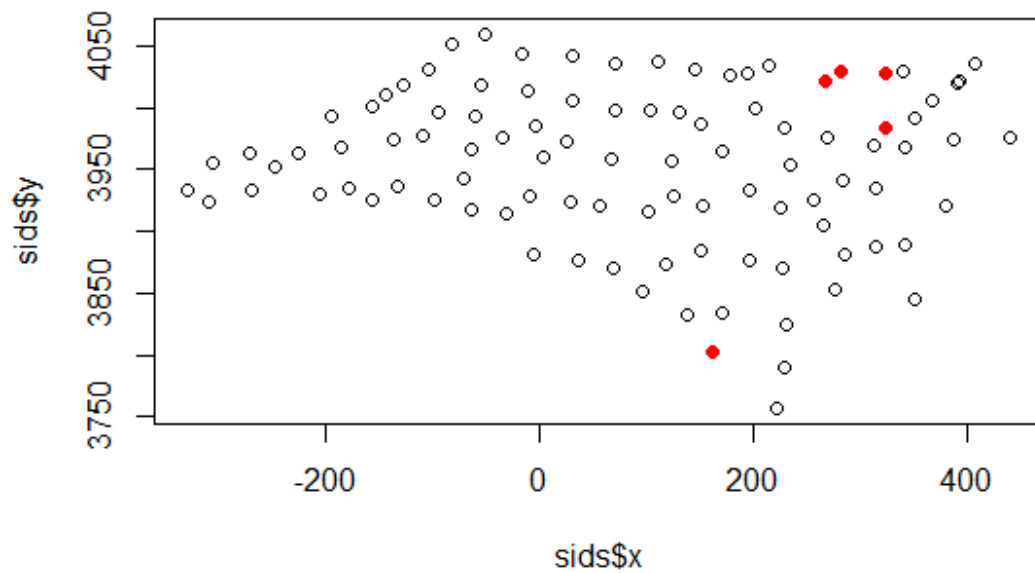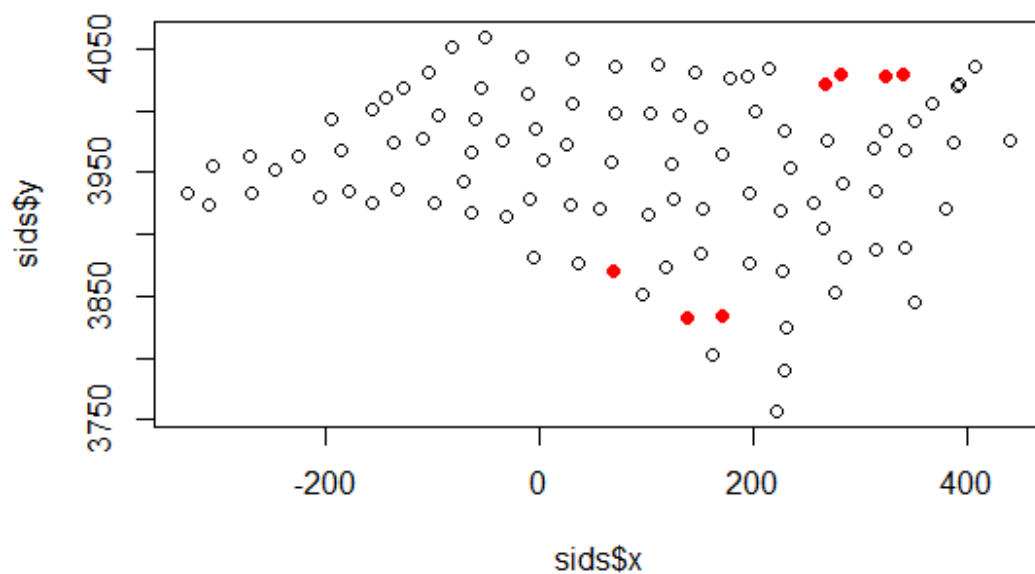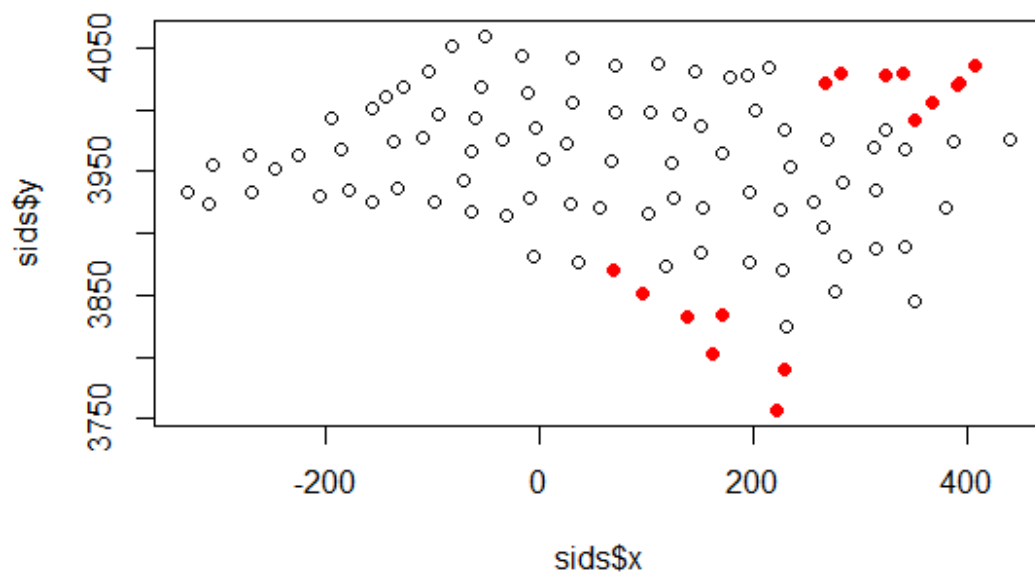
Besag and Newell's test, K = 10

Besag and Newell's test, K = 20

Besag and Newell's test, K = 30

Besag and Newell's test, K = 40

5. Describe the pattern.

1. **K = 10:** Shows very few red points, indicating significant clusters when considering the 10 nearest neighbors. The significant points are few and scattered, suggesting isolated pockets of significance.

2. **K = 20:** A few more red points appear compared to K = 10, indicating that expanding the neighborhood to 20 nearest neighbors identifies a few additional significant points.

3. **K = 30:** There's an increase in the number of red points, showing that considering 30 nearest neighbors leads to the identification of more significant clusters, which are still relatively scattered.

4. **K = 40:** This chart has the highest number of red points, which also tend to form a more coherent cluster pattern in the upper right quadrant of the plot, indicating a more substantial area of significance when the neighborhood includes 40 nearest neighbors.

Q6. Take a screen shot of each plot. Compare these patterns against that from Q3.

Compared to the patterns observed with Kulldorff and Nagarwalla's method (as asked in Q3), the Besag and Newell's test results show a different approach to clustering. Kulldorff and Nagarwalla's method identified clusters based on statistical likelihood and population data, which resulted in either distinct, localized clusters or a single extensive cluster depending on the "Fractpop" parameter.

In contrast, Besag and Newell's method, as evidenced by the increasing K values, does not seem to consider population data directly but rather defines clusters based on the spatial proximity of significant points. As K increases, the clusters become more prominent and more numerous, indicating that the method is sensitive to the spatial density of significant points. Unlike Kulldorff and Nagarwalla's method, which can identify a single large cluster, Besag and Newell's method identifies multiple clusters that become more interconnected as K increases.

Stone test

```
sids<-data.frame(Observed=nc.sids$SID74)
sids<-cbind(sids,
Expected=nc.sids$BIR74*sum(nc.sids$SID74)/sum(nc.sids$BIR74))
sids<-cbind(sids, x=nc.sids$x, y=nc.sids$y)
#Compute Stone's statistic around a county
region<-which(row.names(nc.sids)=="Robeson")
stone.stat(sids, region=region, lambda=1)
stone.test(Observed~offset(log(Expected)), sids, model="poisson", R=99,
region=region, lambda=1)
```

Result

```
##           region
## 2.078983 4.000000

## Stone's Test for raised incidence around locations
##
##   Type of boots.: parametric
##   Model used when sampling: Poisson
##   Number of simulations: 99
```

```
## Statistic: 2.078983
## p-value :  0.01
```

Q7. What is the pattern of this county?

Based on the provided results, Stone's statistic for the region "Robeson" is approximately 2.079 with a corresponding p-value of 0.01. This suggests that the observed number of events (e.g., disease cases, incidents, etc.) in the region is significantly higher than what would be expected based on the model used (Poisson distribution), after accounting for the expected number of events. The pattern indicates a potential cluster or hotspot of the observed events in the Robeson County area with a high level of statistical significance (since a p-value of 0.01 indicates a 1% probability that the observed pattern is due to random chance).

Q8. Compare results generated from the above three methods and those from local Getis G* and local moran's I (in Lab 6). Explain the differences if any.

Stone's test indicates significant incidence in Robeson compared to expectations from a Poisson model, without considering spatial distribution of neighbors. Besag and Newell's test identifies clusters by spatial proximity, increasing with neighbor count (K). Kulldorff and Nagarwalla's method assesses clusters through statistical likelihood and population, with results varying by "Fractpop" parameter. Getis G* and Moran's I also detect spatial patterns but differ in sensitivity to hotspots versus autocorrelation. The methods vary in defining clusters and accounting for spatial relationships or population data, affecting the interpretation of significance and spatial patterns. Stone's significant result suggests a cluster in Robeson, potentially aligning or contrasting with findings from the other methods.