Lab Exercise #7

This week, we will explore different cluster detection methods in R, mainly scan-based methods including GAM, Kulldorff's spatial scan statistics, and Besag and Newell's method.

We will be still using the North Carolina birth data. This data set is for the 100 counties of North Carolina, and includes counts of numbers of live births (also non-white live births) and numbers of sudden infant deaths, for the July 1, 1974 to June 30, 1978 and July 1, 1979 to June 30, 1984 periods. Variables include BIR74 (total number of live births during 1974-1978), SID74 (numbers of sudden infant deaths during 1974-1978), NWBIR74 (number of non-white live births during 1974-1978), SIDR74 (sudden infant deaths rate during 1974-1978).

```
#R libraries we'll use
install.packages("epitools")
install.packages("DCluster")
library(epitools)
library(DCluster)
library(maptools)
library(spdep)

#import data
data(nc.sids) #we are importing the sids data as a system dataset from the loaded package
#Geographical analysis machine
#create a data frame for the observed number of SID cases
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. Radius is the radius of the circle of every scan,
#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)

##########
results of opgam function include five variables (x and y representing the coordinate of the
center of the cluster; statistic value; cluster representing whether it is a cluster or not; and a p
value.
##########
#Plot centroids
plot(sids$x, sids$y, xlab="Easting", ylab="Northing")
#Plot points marked as clusters
points(sidsgam$x, sidsgam$y, col="red", pch="*")
```

Q1. Take a screenshot of the results. Describe the pattern. (10 points)

Q2. Change the alpha (Significance level of the tests performed) to 0.05 and rerun the GAM. Take a screenshot of the results. Describe the differences between it and pattern from Q1. (10 points)

```
# 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
############
kn.iscluster is called from opgam to calculate Kulldorff and Nagarwalla's statistic. Fractpop is
The maximum fraction of the total population used when creating the balls. Model used to
generate random observation can be 'permutation', 'multinomial', 'poisson' or 'negbin'; R is the
number of bootstrap replicates to generate; mle are parameters need by the bootstrap
procedure
############
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")
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)
```

Q3. Take a screenshot of the results. Describe the pattern. (10 points)

Q4. Change the maximum fraction of the total population to 0.5 and rerun the test. Take a screenshot of the results. Describe the differences between it and pattern from Q3. (15 points)

```
###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)
#######
```
It calls the bn.iscluster to calculate the Besag and Newell's Statistic. The centroids is used as grid; the size of the cluster is 20; 100 simulations are performed; Poisson is the model used in the simulations to generate the Observations; Significance level is 0.05, even though multiple tests are made.
```
#########
bnresults<-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(bnresults$x, bnresults$y, col="red", pch=19)
```

Q5. Take a screenshot of the results. Describe the pattern. (10 points)

Q6. Change the k (e.g. 10, 30) and rerun the Besag and Newell's test. Take a screen shot of each plot. Compare these patterns against that from Q3. (15 points)

```
#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)
```

Q7. What is the pattern of this county? (10 points)

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. (20 points)