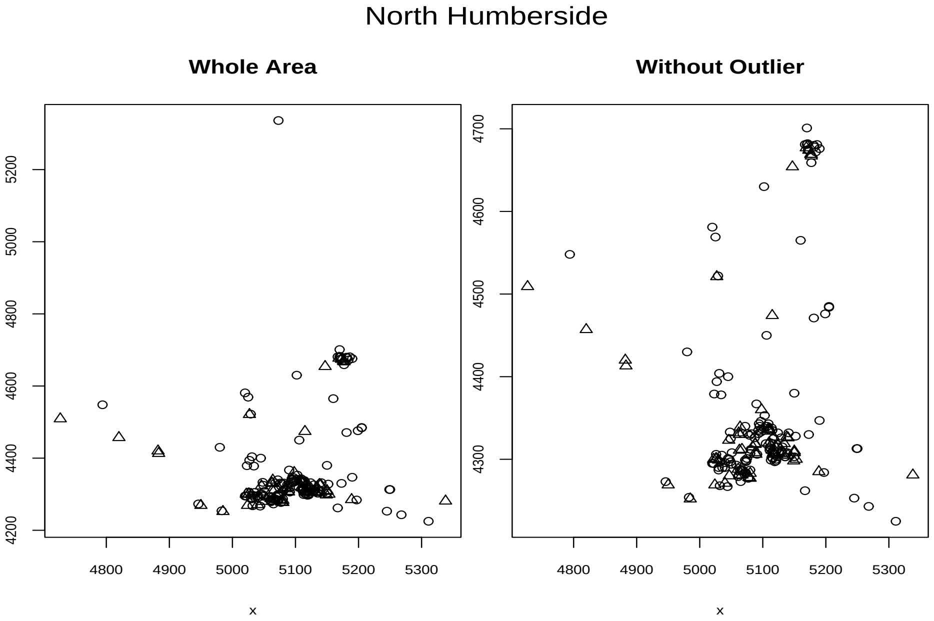
Bio284 Computer Lab 6
Spatial Cluster Detection for Binary Data

1 Goals for this lab:

1. Independent Binary Data
   (a) Spatial Scan Statistic
   (b) Cumulative Residual test
   (c) Apply to North Humberside Dataset
2. Matched Binary Data
   (a) Conditional Cumulative Residual test
   (b) Apply to Taiwan Petrochemical Study
3. ArcGIS to plot results



2 Independent Binary Data

1. Dataset

   (Included within the SatScan Package)

   Childhood Leukemia and Lymphoma Incidence in North Humberside

   This data set describes the spatial location of 62 cases of childhood leukemia and lymphoma in North Humberside between 1974 and 1986, as well as 141 controls.

   -Copy all NHumberside data on G:/shared into a directory

2. **Spatial Scan Statistic**

   We will run a cluster detection analysis on the North Humberside dataset using the Spatial Scan Statistic and the SatScan software.

   (a) Download the SatScan Software from www.satscan.us

   (Note: Documentation and downloads for SatScan at www.satscan.us)

   -Open SatScan from Start→ All Programs→ SatScan

   -Click on create a new session.

   (b) Conduct Analysis

   - *Input Tab*

     Click ⋯ by Case File space and go to where you saved the NHumberside files. You will automatically be looking for files of type ".cas". What this file type means is a text file with a column with ID of subject and a column with binary indicator of case (1) or control (0) and no headers.

     Click ⋯ by Control File space. You will automatically be looking for files of type ".ctl". What this file type means is a text file with a column with ID of subject and a column with binary indicator of case (0) or control (1) and no headers.

     Under time precision select, none, since we have no time component in our analysis.

     Click ⋯ by Coordinates File space. You will automatically be looking for files of type ".geo". What this file type means is a text file with a column with ID of subject, a column with x coordinates and a column with y coordinates, no header.

     Under coordinates select Cartesian.

   - *Analysis Tab*

     Under Probability model select Bernoulli.

   - *Output Tab*

     Select all outputs of type dbf since ArcGIS handles this readily. Select a file name for the results

   - *Execute Analysis*

     Click the yellow lightning rod to execute the code and perform the analysis.

(c) **Results**

```
                             -----------------------------

                                     SaTScan v5.0
                             -----------------------------


Program run on: Wed Dec 15 19:50:20 2004

Purely Spatial analysis scanning for clusters with high rates
using the Bernoulli model.
________________________________________________________________

SUMMARY OF DATA

Study period..........: 2000/1/1 - 2000/12/31 Number of
locations...: 191 Total population......: 203 Total
cases...........: 62
________________________________________________________________

MOST LIKELY CLUSTER

1.Location IDs included.: 19, 18, 14, 26
  Coordinates / radius..: (5026,4301) / 4.47
  Population............: 4
  Number of cases.......: 4
  Expected cases........: 1.22
  Observed / expected...: 3.274
  Log likelihood ratio..: 4.836516
  Monte Carlo rank......: 674/1000
  P-value...............: 0.674
```

Is there any clustering?

Where is the highest cluster located?

3. **Cumulative Geographic Residual**

For the cumulative geographic residual test we will be using the R package to conduct the analysis.

We will be using the file NHnumberside.txt for the dataset. It has a header with the variables ID, Y (indicator of case(1) and control(0)), x, and y.

```
NH<-read.table("C:/NHumberside.txt",header=T)

# Binary Outcome Variable
Y<-NH$Y

# Matrix of locations (x,y)
loc<-NH[,3:4]
```

The program is in a txt file called cumres-BinR.txt. It contains the main function, CumRes.Bern, and two corresponding plot function, plot.CR and plot.CRs.

```
source("C:/cumres-BinR.txt")
```

Lets run the first analysis. There is one main thing to think about when running the cumulative residual test.

Window Size : (b_1, b_2)

We will use (50,50) for now, but when running this method one should choose several window sizes and conduct a sensitivity analysis.

```
# Range of Study area
xrange<-c(min(loc$x),max(loc$x))
yrange<-c(min(loc$y),4710)

CRtest1<-CumRes.Bern(Y,loc,X=NULL,xarea=xrange,yarea=yrange,
                         b=c(50,50),jumps=c(25,25),1000)
```

RESULTS:

```
CRtest1[11:16]
$Sloc
[1] 0.2398025

$pval
         Z     D
[1,] 0.609 0.639
```

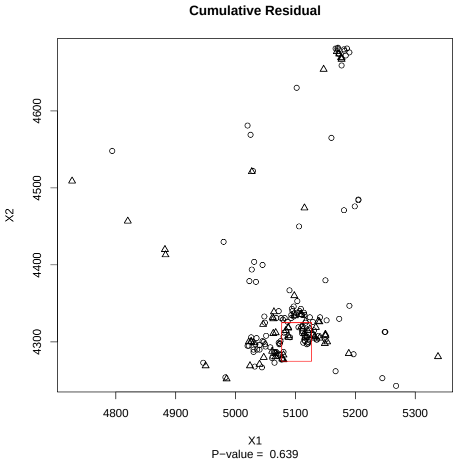
```
$maxloc
        x    y
[1,] 5127 4325

$crit
              Z         D
90%   0.4128042 0.4116202
95%   0.4680560 0.4696235
97.5% 0.5121604 0.5325975

$meanSlochat
             Z         D
[1,] 0.2829770 0.2848552

$quantSlochat
              Z         D
2.5%  0.1348415 0.1330336
5%    0.1534783 0.1494689
10%   0.1716653 0.1755866
50%   0.2654225 0.2684124
90%   0.4128042 0.4116202
95%   0.4680560 0.4696235
97.5% 0.5121604 0.5325975
```

(a) Is there significant clustering?

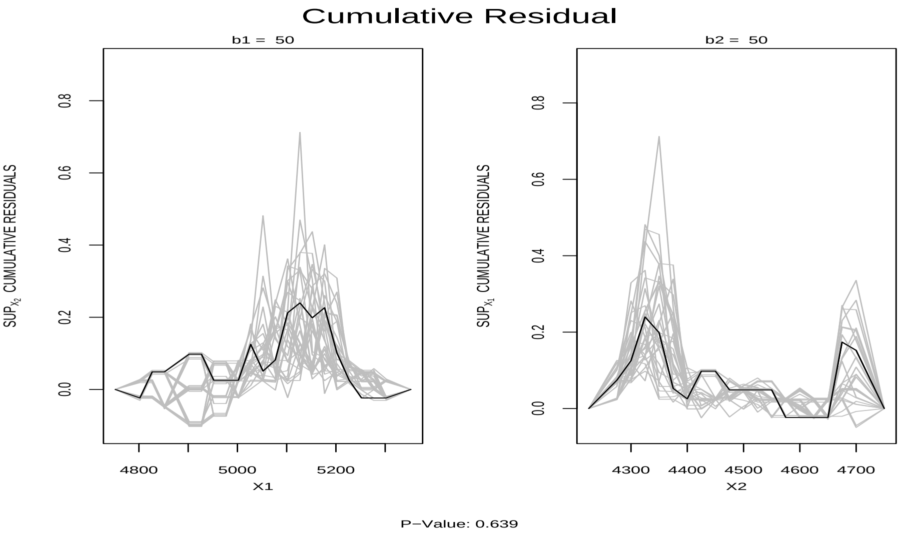
(b) Where is the highest cluster?

Plot the Results

```
plot.CR(CRtest1)
```



```
plot.CRs(CRtest1)
```



For input into ArcGIS:

```
NHarc<-data.frame(NH)
names(NHarc)<-c("ID","Y","xcoord","ycoord")
write.table(NHarc,"C:/NHarc.txt",row.names=F,sep=",")

# create a cluster area dataset maxloc<-CRtest1$maxloc
b<-CRtest1$b

x<-c(maxloc[1]-b[1],maxloc[1]-b[1],maxloc[1],maxloc[1])
y<-c(maxloc[2],maxloc[2]-b[2],maxloc[2]-b[2],maxloc[2])

write.table(cbind(x,y),"C:/NHcr.txt",row.names=F,sep=",")
```

3 Matched Binary Data

1. Dataset

   Childhood Leukemia and Petrochemical Exposure in Kaohsiung, Taiwan
   (P.I. David Christiani)

   This data set describes the spatial location of 121 cases of childhood leukemia in Kaohsiung as well as 287 controls.

   -Copy all petro data on G:/shared into a directory

2. Analysis

   We will be using the file petro.txt for the dataset. It has a header with the variables stratum, CASE, X.coord, Y.coord, SMOKE, and SUBSMOKE.

```
petro<-read.table("C:/petro.txt",header=T)
```

```
Y<-petro$CASE
x<-petro$X.coord
y<-petro$Y.coord
loc<-cbind(x,y)

stratum<-petro$stratum
X<-cbind(petro$SMOKE,petro$SUBSMOKE)
```

The program is in a txt file called cumres-MBinR.txt. It contains the main function CumRes.MBern.

```
source("C:/cumres-MBinR.txt")
```

We will use window size of (4800,9400) for now, but when running this method one should choose several window sizes and conduct a sensitivity analysis.

(a) WITHOUT COVARIATES

```
b<-c(4800,9400)
jumps<-c(1000,1000)

CRpetro1<-CumRes.MBern(Y=Y,loc=loc,X=NULL,b=b,stratum=stratum,jumps=jumps,
           nsims=1000,type="mixed")
```

RESULTS:

```
CRpetro1[12:17]
$Sloc [1] 0.939616

$pval
         Z     D
[1,] 0.028 0.045

$maxloc
          x       y
[1,] 179187 2511273

$crit
              Z         D
90%   0.7967008 0.8179731
95%   0.8761764 0.9090721
97.5% 0.9703929 1.0148213

$meanSlochat
            Z         D
[1,] 0.526095 0.5274363
```

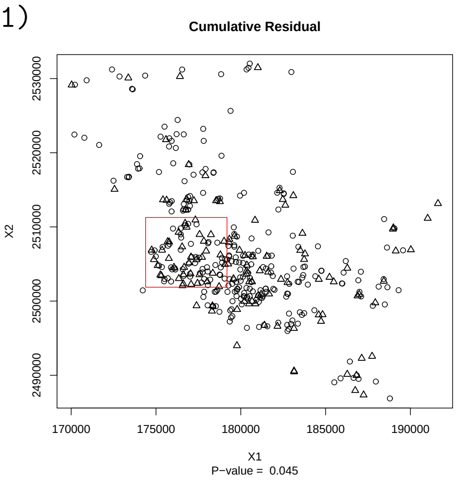
```
$quantSlochat
                Z         D
2.5%  0.2136219 0.1985431
5%    0.2425349 0.2396608
10%   0.2843336 0.2856456
50%   0.4982981 0.4992677
90%   0.7967008 0.8179731
95%   0.8761764 0.9090721
97.5% 0.9703929 1.0148213
```

i. Is there significant clustering?

ii. Where is the highest cluster?

Plot the Results

```
plot.CR(CRpetro1)
```



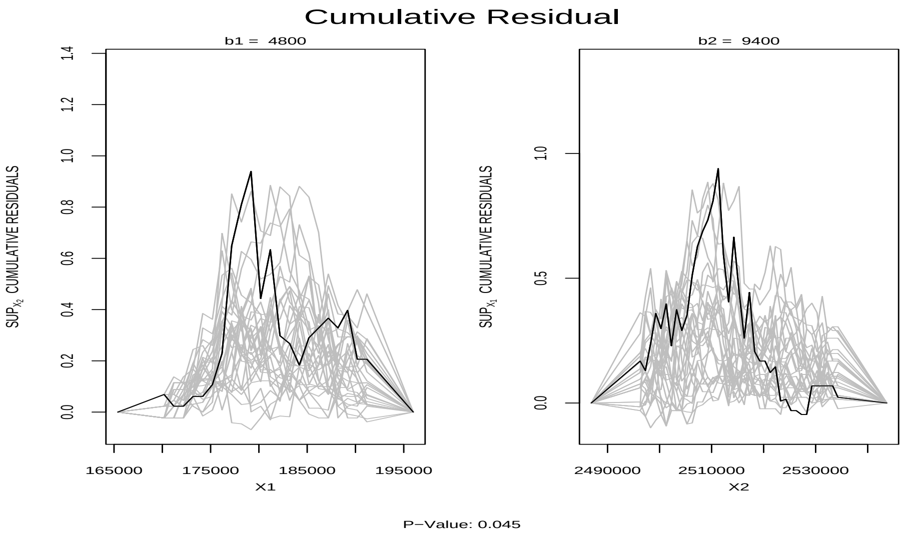
```
plot.CRs(CRpetro1)
```

(b) WITH SMOKING COVARIATES

```
b<-c(4800,9400)
jumps<-c(1000,1000)

CRpetro2<-CumRes.MBern(Y=Y,loc=loc,X=X,b=b,stratum=stratum,jumps=jumps,
             nsims=1000,type="mixed")
```

RESULTS:

```
CRpetro2[12:17]
$Sloc
[1] 0.9390835

$pval
         Z     D
```



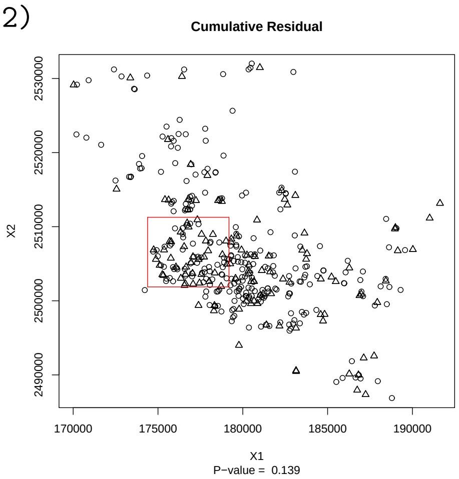
```
[1,] 0.148 0.139

$maxloc
          x       y
[1,] 179187 2511273

$crit
             Z        D
90%   1.017588 1.010507
95%   1.106199 1.138549
97.5% 1.213654 1.257334

$meanSlochat
            Z       D
[1,] 0.695705 0.69668

$quantSlochat
              Z         D
2.5%  0.3344124 0.3375272
5%    0.3746685 0.3760654
10%   0.4202907 0.4424850
50%   0.6664539 0.6676163
90%   1.0175876 1.0105074
95%   1.1061986 1.1385487
97.5% 1.2136542 1.2573343
```

i. Is there significant clustering?

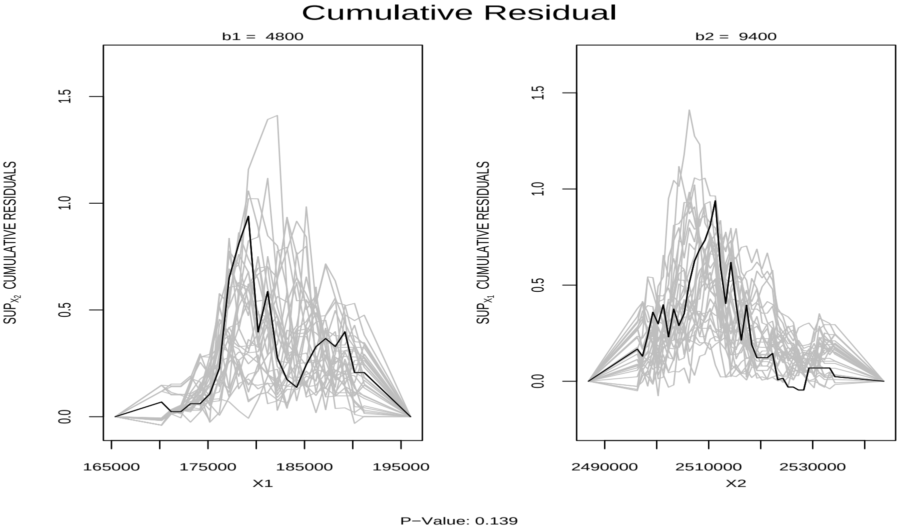
ii. Where is the highest cluster?

Plot the Results

```
plot.CR(CRpetro2)
```



```
plot.CRs(CRpetro2)
```



4 ArcGIS

We will plot the NHumberside data results for both the spatial scan statistic and the cumulative geographic residual test under the independence assumption.

1. Data Manipulation

   To get the data from SatScan into ArcGIS you need to open the SSresults.col.dbf in excel. The x,y, and radius columns are in text form instead of number format. Reformat these columns into number format by just selected the columns without the header and clicking on the question mark box. Also remove any ”» marks in the parts of the column names since ArcGIS cannot handle that type of naming. For our example delete all of the rows except for the first entry since we will only be looking at the first cluster.

2. Open ArcGIS and import data

   Addxy data NHarc.txt, SSResults.col.dbf, and NHcr.txt onto your layout.

   You will have a plot of all the cases and controls. You will also have a point for the center point of the circle for the Spatial Scan and will have four points depicting the highest square area from the cumulative residual results.

3. Spatial Scan buffer

   You need to add the buffer wizard to draw the circle. Go to Tools, customize, and click on the Commands tab. Scroll to tools and you will see the Buffer Wizard in the commands window. Drag and click it to the main map ArcGIS toolbar. Close the customize window.

   We need to put map units before using the buffer tool. Right click in map area and select Properties and click on the General tab. In the units section, click on kilometers and push okay.

   Open the attributes table of SSresults.col and highlight the first row. We need to select this row to create a buffer around this point.

   Now double click on the Buffer Wizard in the toolbar. Scroll to SSresults.col and you should see 1 for both number of features and number of features selected. Click next.

   At the specified distance click 4 kilometers since this is the radius of the file.

   Click next and rename your new file for a buffer. Click finish and a new shape file with your buffer will appear on the display. It is very small since it only had included four subjects.

4. Cumulative Residual

   Open attributes table of NHcr.txt. We will be using these points to draw a square. Click on the Editors tool and say start editing. (If the Editor tool is not displayed, right click in toolbar area, and select Editor).

   Click on the pencil icon. Then right-click in the map area and select Absolute X,Y. Enter the coordinates in the first line of NHcr.txt and press return. Right-click again, select Absolute X,Y, and enter coordinates in second line and press return. Continue for all four lines. Afterward, Right-click again and now click on Finish Sketch.

   Now you have both areas outlined for both methods. Click on Editor toolbar and say stop editing and save your results.

   Here is one way to create maps of these results. You can use all of the tools from previous labs to make things look better then what you have now.