Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. http://www.jstatsoft.org/

DClusterm: Model-based detection of disease clusters

Virgilio Gómez-Rubio
Universidad de Castilla-La Mancha

Paula Moraga-Serrano
London School of Hygiene & Tropical Medicine

Abstract

Keywords: disease cluster, spatial statistics, R.

1. Introduction

Kulldorff (1997) proposes a test for detecting disease clusters which will find the most likely cluster. This is called the Spatial Scan Statistic and the significance of the test is found via a Monte Carlo test. The test statistic is based on a likelihood ratio test for the following test:

$$\begin{aligned} H_0: &\quad \theta_z = \theta_{\overline{z}} \\ H_1: &\quad \theta_z > \theta_{\overline{z}} \end{aligned}$$

Here, z represents a cluster (i.e., a set of contiguous areas), θ_z the relative risk in the cluster and $\theta_{\overline{z}}$ the relative risk outside the cluster. Many different clusters are tested in turn. The most likely cluster is the one with the highest value of the test statistic. Then a Monte Carlo test is used to compute the p-value of the most likely cluster.

2. Generalised Linear Models for cluster detection

Jung (2009); Zhang and Lin (2009) show that the test statistic for a given cluster is equivalent to fitting a Generalised Linear Model using a cluster variable as a predictor. This cluster variable is a dummy variable which is 1 for the areas in the cluster and 0 for the areas outside the cluster.

Firstly, given that we are using GLM's we could include covariates in the model. For example, for a Poisson model with expected counts E_i we could have:

$$O_i \sim Po(E_i \theta_i)$$

$$\log(\theta_i) = \log(E_i) + \alpha + \beta x_i$$

Fitting this model will provide estimates $\hat{\alpha}$ and $\hat{\beta}$. This will account for the (spatial) effects of the covariates. In order to include the cluster variable the effects of the covariates will be keep fixed. Hence, the clusters covariates will be used in a model with fixed coefficients for the covariates:

$$\log(\theta_i) = \log(E_i) + \hat{\alpha} + \hat{\beta} x_i + \gamma CLUSTER_i$$

This means that the offset now is $\log(E_i) + \hat{\alpha} + \hat{\beta} x_i$. γ is a measure of the difference of the risk in the cluster. We are only interested in cluster whose coefficient is higher than 0 (i.e., increased risk).

Testing different clusters will produce many different cluster covariates. We can use model selection techniques to select the most important cluster in the area. In particular, the log-likelihood can be used to compare the model with the cluster variable to the null model (i.e., the one with the covariates only). Note that we are interested in clusters with a high risk, so that

2.1. Leukemia in upstate New York

The `NY8` dataset is avaialble in package `DClusterm` and it provides cases of leukemia in different census tracts in upstate New York. This data set has been analysed by several authors (Waller, Turnbull, Clark, and Nasca 1992; Waller and Gotway 2004).

The location of leukemia is thought to be linked to the use of Trichloroethene (TCE) by several companies in the area. Figure 1 shows the Standardised Mortality Ratios of the census tracts and the locations of the industries using TCE.

In order to measure exposure, the inverse of the distance to the nearest TCE site has been used (PEXPOSURE). In addition, two other socioeconomic covariates have been used: the percetage of people aged 65 or more (PCTAGE65P) and the percentage of people who own their home (PCTOWNHOME).

```
> library(DClusterm)
> library(snowfall)
> library(xts)
> data(NY8)
> NY8$Cases2<-round(NY8$Cases)
> NY8$Observed<-NY8$Cases2
> NY8$EXP<-NY8$POP8*sum(NY8$Cases2)/sum(NY8$POP8)
> NY8$Expected<-NY8$EXP
> NY8$SMR<-NY8$Cases2/NY8$EXP
> NY8$x<-coordinates(NY8)[,1]
```

```
> NY8$y<-coordinates(NY8)[,2]
> NY8st<-STFDF(as(NY8, "SpatialPolygons"), xts(1,as.Date("1972-01-01")),
+ NY8@data, endTime=as.POSIXct(strptime(c("1972-01-01"), "%Y-%m-%d"), tz = "GMT"))
>
```

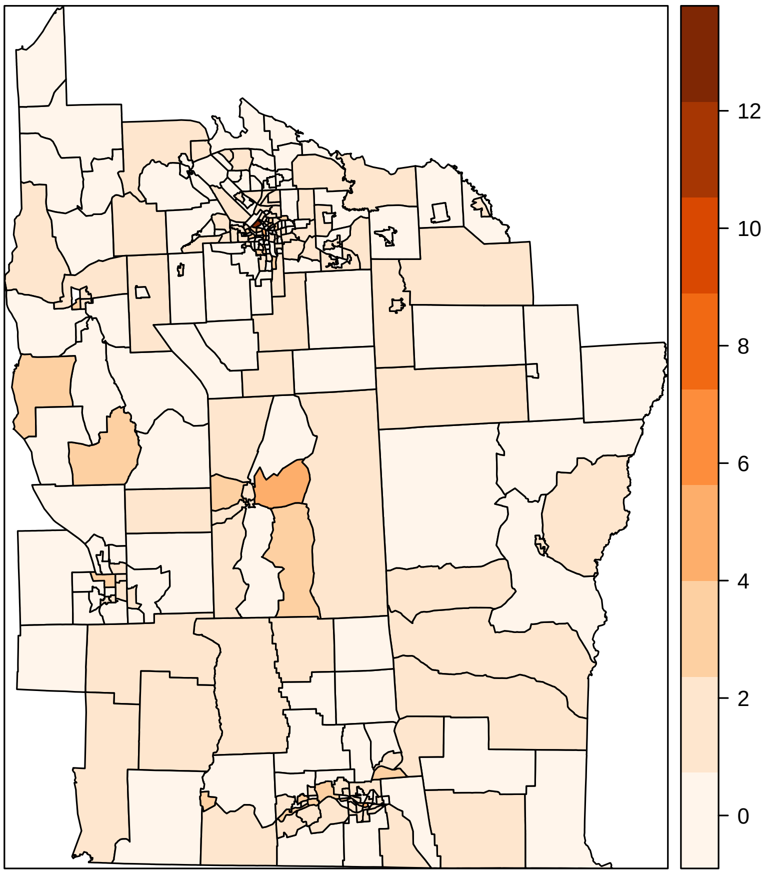

Figure 1: SMR of the incidence of Leukemia in upstate New York.

2.2. Cluster detection

Cluster detection with no covariates

First of all, a model with no covariates will be fitted and used as a starting point.

```
> m0<-glm(Cases2~offset(log(EXP))+1, family="poisson", data=NY8)
> idxcl<-c(120, 12, 89, 139, 146)
> cl0<-DetectClustersModel(NY8st, thegrid=as.data.frame(NY8)[idxcl,c("x", "y")],
+    fractpop=.15, alpha=0.05, radius=Inf, step=NULL,
+    typeCluster="S", R=NULL, numCPUS=2, model0=m0)
```

Below is a summary of the clusters detected with this method. The dates can be ignored as this is a purely spatial cluster.

```
> cl0
```

```
           x       y size      minDateCluster      maxDateCluster statistic
11  424728.9 4661404   39 1972-01-01 01:00:00 1972-01-01 01:00:00  8.044846
88  409430.4 4720092    9 1972-01-01 01:00:00 1972-01-01 01:00:00  6.967107
119 404710.7 4768346   24 1972-01-01 01:00:00 1972-01-01 01:00:00  3.254824
    cluster       pvalue
11     TRUE 0.0000604120
88     TRUE 0.0001893208
119    TRUE 0.0107290781
```

The centre of the clusters detected are shown in Figure 2.

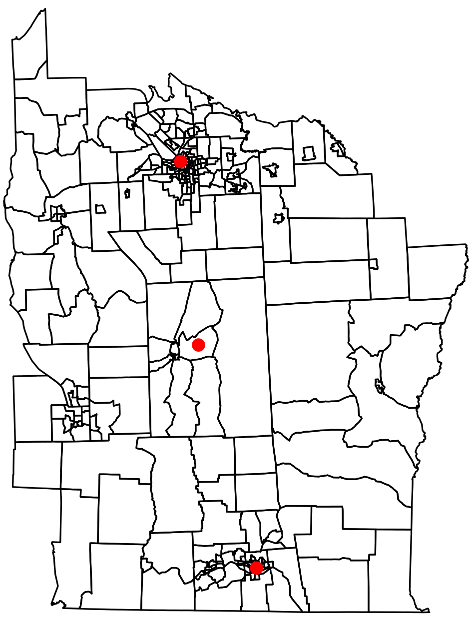

Figure 2: Clusters detected when no covariates are included in the model.

Cluster detection after adjusting for covariates

Similarly, clusters can be detected after adjusting for significant risk factors. First we will fit a GLM with the 3 covariates mentioned earlier. As it can be seen, all three are significant:

```
> m1<-glm(Cases2~offset(log(EXP))+PCTOWNHOME+PCTAGE65P+PEXPOSURE,
+    family="poisson", data=NY8)
> summary(m1)

Call:
glm(formula = Cases2 ~ offset(log(EXP)) + PCTOWNHOME + PCTAGE65P +
    PEXPOSURE, family = "poisson", data = NY8)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9099  -1.1294  -0.1768   0.6385   3.2426

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.65507    0.18550  -3.531 0.000413 ***
PCTOWNHOME  -0.36472    0.19316  -1.888 0.058998 .
PCTAGE65P    4.05031    0.60559   6.688 2.26e-11 ***
PEXPOSURE    0.15141    0.03165   4.784 1.72e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 459.05  on 280  degrees of freedom
Residual deviance: 384.01  on 277  degrees of freedom
AIC: 958.97

Number of Fisher Scoring iterations: 5
```

The cluster detection method is run as before, but now we use the previous model instead:

```
> cl1<-DetectClustersModel(NY8st, thegrid=as.data.frame(NY8)[idxcl,c("x", "y")],
+    fractpop=.15, alpha=.05,
+    typeCluster="S", R=NULL, numCPUS=2, model0=m1)

> cl1

           x       y size      minDateCluster      maxDateCluster statistic
88  409430.4 4720092    9 1972-01-01 01:00:00 1972-01-01 01:00:00  5.861204
119 404710.7 4768346   20 1972-01-01 01:00:00 1972-01-01 01:00:00  3.160591
    cluster       pvalue
88     TRUE 0.0006175202
119    TRUE 0.0119304026
```

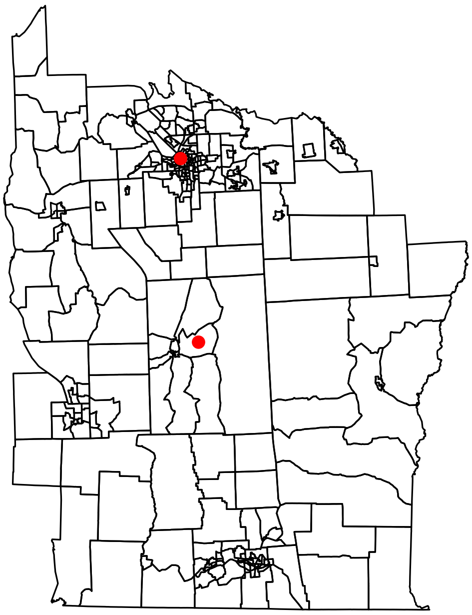

Figure 3: Clusters detected after adjusting for covariates.

Figure 3 shows the clusters detected after adjusting for covariates.

3. Spatio-temporal clusters

3.1. Brain Cancer in New Mexico

The `brainNM` data set contains yearly cases of brain cancer in New Mexico from 1973 to 1991 (inclusive). The data set has been taken from the SatScan website and the area boundaries from the U.S. Census Bureau. In addition, the location of Los Alamos National Laboratory has been included (from the Wikipedia). Inverse distance to this site can be used to test for increased risk in the areas around the Laboratory as no other covariates are available.

```
> library(DClusterm)
> #debug(DetectClustersModel)
> #debug(glmAndZIP.iscluster)
> #debug(CalcStatsAllClusters)
```

```
> library(snowfall)
> data(brainNM)
```

Expected counts have been obtained using age and sex standardisation over the whole period of time. Hence, yearly differences are likely to bee seen when plotting the data. The SMR's have been plotted in Figure 3.1.

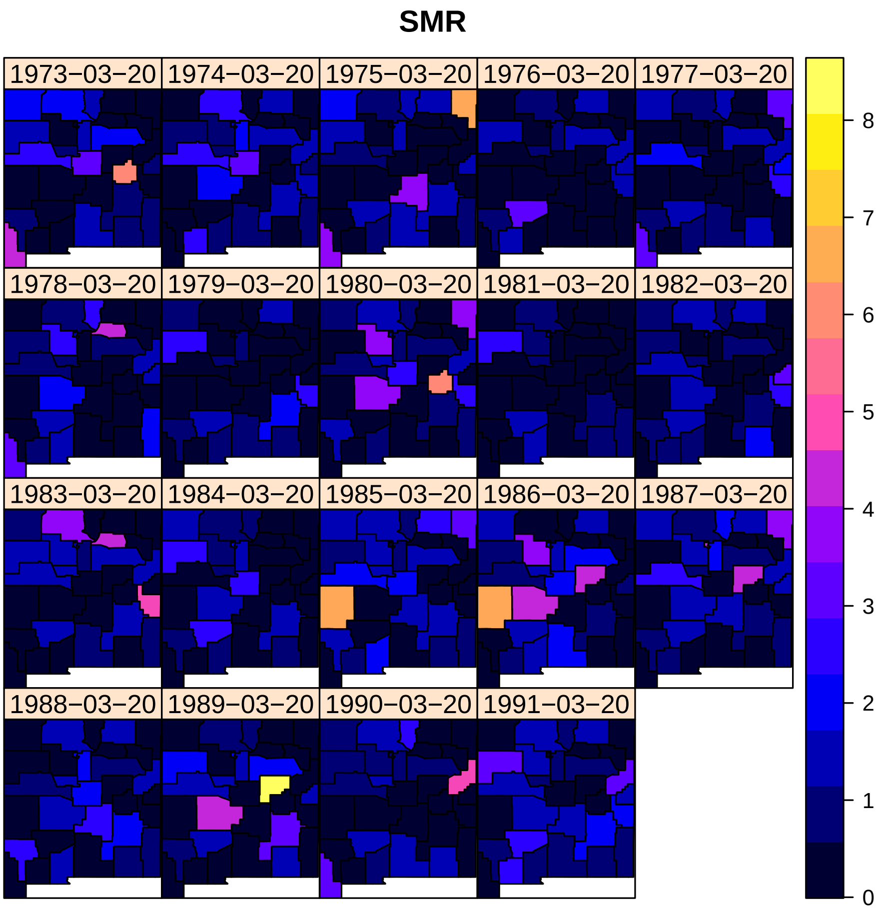

Figure 4: SMR of brain cancer in New Mexico.

3.2. Cluster detection

Cluster detection with no covariates

Similarly as in the spatial case, a GLM

```
> m0<-glm(Observed~offset(log(Expected))+1, family="poisson", data=brainst@data)
> summary(m0)

Call:
glm(formula = Observed ~ offset(log(Expected)) + 1, family = "poisson",
```

```
    data = brainst@data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4874  -0.9998  -0.4339   0.3773   3.1321

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.834e-16  2.917e-02       0        1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 631.64  on 607  degrees of freedom
Residual deviance: 631.64  on 607  degrees of freedom
AIC: 1585.6

Number of Fisher Scoring iterations: 5
```

```
> cl0<-DetectClustersModel(brainst, coordinates(brainst@sp),
+    minDateUser="1985-01-01", maxDateUser="1989-01-01",
+    fractpop=.15, alpha=0.05, typeCluster="ST", R=NULL, numCPUS=2, model0=m0)
```

```
> nrow(cl0)
```

```
[1] 180
```

```
> cl0[1:5,]
```

```
             x        y size      minDateCluster      maxDateCluster statistic
0286 -106.3073 35.86930    3 1986-03-20 01:00:00 1988-03-20 01:00:00  7.493492
0496 -105.9761 35.50684    2 1986-03-20 01:00:00 1988-03-20 01:00:00  6.438221
0531 -106.9303 34.00725    9 1985-03-20 01:00:00 1986-03-20 01:00:00  6.378992
0498 -105.9761 35.50684    2 1987-03-20 01:00:00 1988-03-20 01:00:00  6.331113
0288 -106.3073 35.86930    2 1987-03-20 01:00:00 1988-03-20 01:00:00  6.331113
     cluster       pvalue
0286    TRUE 0.0001082553
0496    TRUE 0.0003327442
0531    TRUE 0.0003544929
0498    TRUE 0.0003731179
0288    TRUE 0.0003731179
```

Cluster detection after adjusting for covariates

We will use the inverse of the distance to Los Alamos National Laboratory as a covariate.

```
> dst<-spDistsN1(coordinates(brainst@sp), losalamos, TRUE)
> nyears<-length(unique(brainst@data$Year))
```

```
> brainst@data$IDLANL<-rep(1/dst, nyears)
>

> m1<-glm(Observed~offset(log(Expected))+IDLANL,
+    family="poisson", data=brainst)
> summary(m1)

Call:
glm(formula = Observed ~ offset(log(Expected)) + IDLANL, family = "poisson",
    data = brainst)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4832  -0.9982  -0.4280   0.3775   3.1424

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.005721   0.029897  -0.191    0.848
IDLANL       0.338194   0.364900   0.927    0.354

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 631.64  on 607  degrees of freedom
Residual deviance: 630.84  on 606  degrees of freedom
AIC: 1586.8

Number of Fisher Scoring iterations: 5

> cl1<-DetectClustersModel(brainst, coordinates(brainst@sp), fractpop=.15,
+    alpha=0.05, minDateUser="1988-01-01", maxDateUser="1989-01-01",
+    typeCluster="ST", R=NULL, numCPUS=2, model0=m1)

> nrow(cl1)

[1] 6

> cl1[1:5,]

            x        y size      minDateCluster      maxDateCluster statistic
049 -105.9761 35.50684    2 1988-03-20 01:00:00 1988-03-20 01:00:00  2.433451
028 -106.3073 35.86930    2 1988-03-20 01:00:00 1988-03-20 01:00:00  2.433451
057 -105.8508 34.64048    2 1988-03-20 01:00:00 1988-03-20 01:00:00  2.431998
013 -106.8328 32.35265   17 1988-03-20 01:00:00 1988-03-20 01:00:00  2.010047
027 -105.4592 33.74524    3 1988-03-20 01:00:00 1988-03-20 01:00:00  2.007057
    cluster     pvalue
049    TRUE 0.02737662
```

```
028     TRUE 0.02737662
057     TRUE 0.02742274
013     TRUE 0.04496121
027     TRUE 0.04512090
```

We can easily display the most significant cluster as follows:

```
> stcl<-get.stclusters(brainst, cl0)
> brainst$CLUSTER<-0
> brainst$CLUSTER[stcl[[1]]]<-1

> print(stplot(brainst[,,"CLUSTER"], at=c(0, 0.5, 1.5)))
```

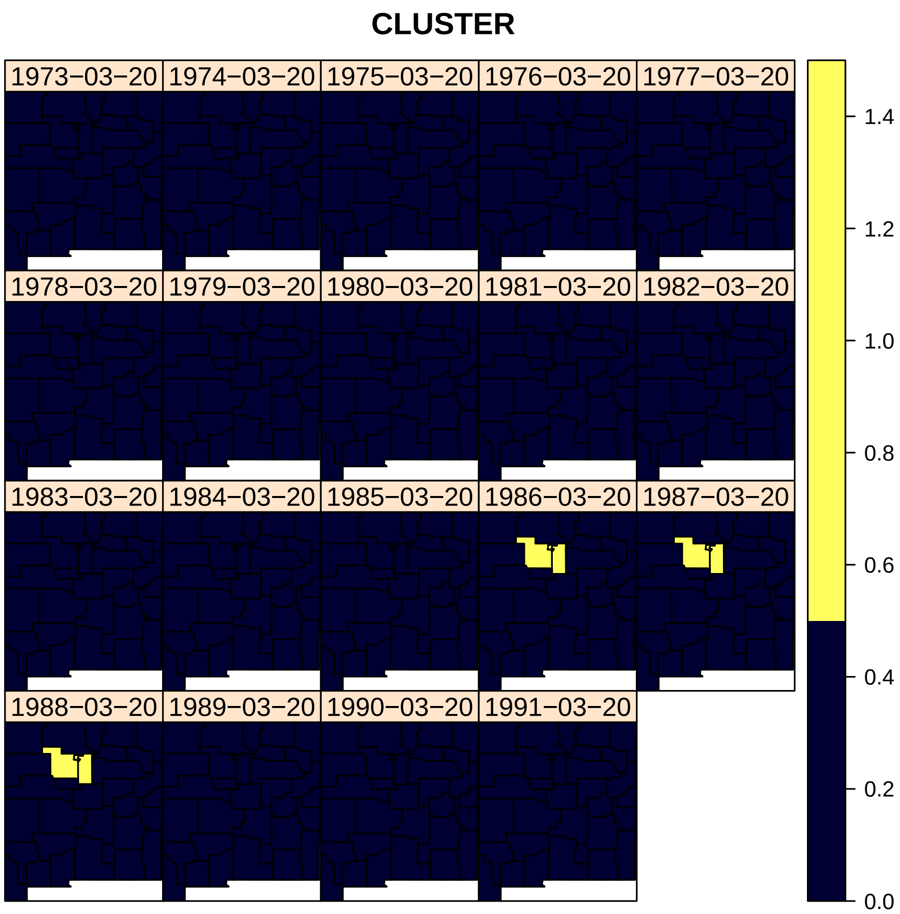

Figure 5: Spatio-temporal cluster of brain cancer detected in New Mexico.

4. Zero-inflated models for cluster detection

Gómez-Rubio and López-Quílez (2010) extend this method to account for zero-inflation. In this case the observed number of cases come from a mixture distribution:

$$Pr(O_i = n_i) = \begin{cases} \pi_i + (1-\pi_i)Po(0|\theta_i E_i) & n_i = 0 \\ (1-\pi_i)Po(n_i|\theta_i E_i) & n_i = 1, 2, \dots \end{cases}$$

The relative risk θ_i can be modelled using a log-linear model to depend on some relevant risk factors. Also, it is common that all π_i's are taken equal to a single value π.

4.1. Brain Cancer in Navarre (Spain)

Ugarte, Ibáñez, and Militino (2006) analyse the incidence of brain cancer in Navarre (Spain). The aggregation level is the health district. Figure 4.1 shows the SMR. As it can be seen there are many areas where the SMR is zero because there are no cases in those areas. Ugarte, Ibáñez, and Militino (2004) also tested for positive zero-inflation of these data compared to a Poisson distribution. The method implemented in this package is similar to the one used in Gómez-Rubio and López-Quílez (2010) for the detection of disease clusters of rare diseases.

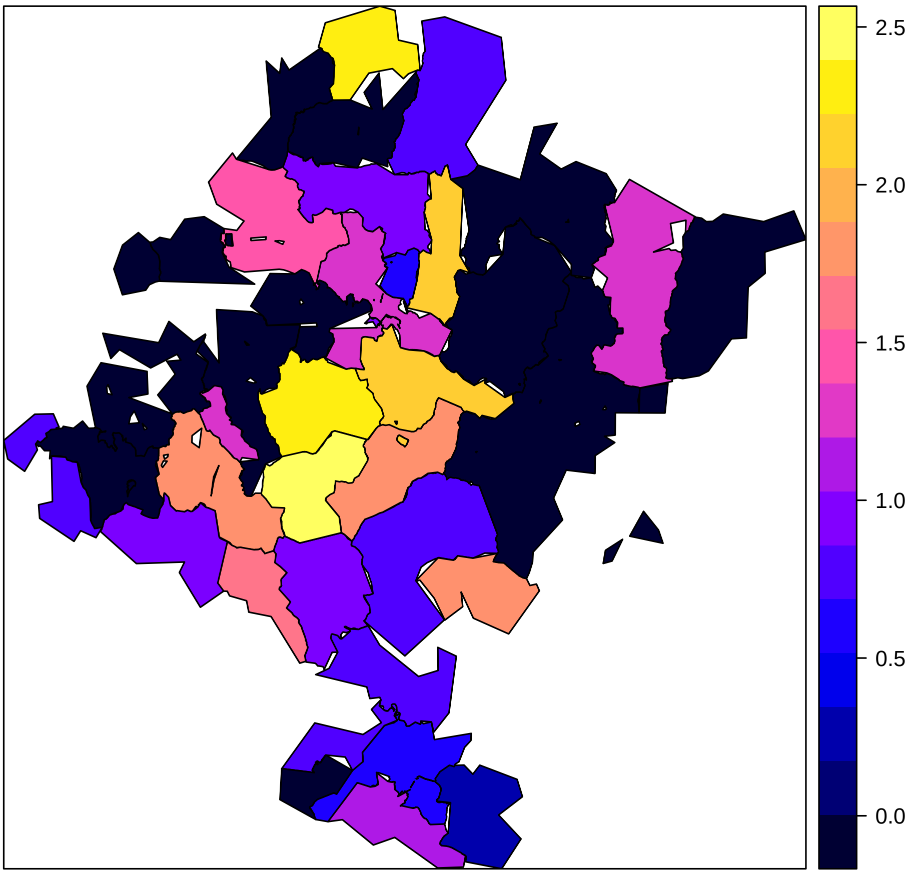

Figure 6: SMR of brain cancer in Navarre (Spain).

4.2. Cluster detection

Cluster detection with no covariates

Before starting our cluster detection methods, we will check the appropriateness of a Poisson GLM for this data. Fitting a log-linear model (with no covariates) gives the following model:

```
> m0<-glm(OBSERVED~ offset(log(EXPECTED))+1, family="poisson", data=brainnav)
> summary(m0)

Call:
glm(formula = OBSERVED ~ offset(log(EXPECTED)) + 1, family = "poisson",
    data = brainnav)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5227  -1.4783  -0.3203   0.7042   1.6393

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.752e-06  8.805e-02       0        1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 63.733  on 39  degrees of freedom
Residual deviance: 63.733  on 39  degrees of freedom
AIC: 145.02

Number of Fisher Scoring iterations: 5
```

Furthermore, a quasipoisson model has been fit in order to asses any extra-variation in the data:

```
> m0q<-glm(OBSERVED~ offset(log(EXPECTED))+1, family="quasipoisson",
+    data=brainnav)
> summary(m0q)

Call:
glm(formula = OBSERVED ~ offset(log(EXPECTED)) + 1, family = "quasipoisson",
    data = brainnav)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5227  -1.4783  -0.3203   0.7042   1.6393

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -7.752e-06  9.703e-02       0       1

(Dispersion parameter for quasipoisson family taken to be 1.214555)

    Null deviance: 63.733  on 39  degrees of freedom
Residual deviance: 63.733  on 39  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

The dispersion parameter in the previous model seems to be higher than 1, which may mean that the Poisson distribution is not appropriate.

For this reason, and following Ugarte *et al.* (2004), a zero-inflated Poisson model has been fit. Here is the resulting model:

```
> m0zip<-zeroinfl(OBSERVED ~ offset(log(EXPECTED))+1 | 1, data = brainnav,
+   dist="poisson", x=TRUE)
> summary(m0zip)

Call:
zeroinfl(formula = OBSERVED ~ offset(log(EXPECTED)) + 1 | 1, data = brainnav,
    dist = "poisson", x = TRUE)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.3585 -0.9137 -0.1378  0.7137  1.8091

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.09347    0.09459   0.988    0.323

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6158     0.6435  -2.511    0.012 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 9
Log-likelihood: -69.08 on 2 Df
```

Hence, the zero-inflated Poisson model will be used now to detect clusters of disease:

```
> brainnav$Expected<-brainnav$EXPECTED
> brainnavst<-STFDF(as(brainnav, "SpatialPolygons"),
+ xts(1,as.Date("1990-01-01")), brainnav@data,
+ endTime=as.POSIXct(strptime(c("1990-01-01"), "%Y-%m-%d"), tz = "GMT"))
```

```
> cl0<-DetectClustersModel(brainnavst, coordinates(brainnav), fractpop=.25,
+    alpha=.05,
+    typeCluster="S", R=NULL, numCPUS=2, model0=m0zip)

Library spdep loaded.
Library splancs loaded.
Library spacetime loaded.
Library DCluster loaded.
Library pscl loaded.
Library DClusterm loaded.
[1] 1 1

> cl0

          x       y size      minDateCluster      maxDateCluster statistic
31 596886.8 4710520    4 1990-01-01 01:00:00 1990-01-01 01:00:00  2.520092
30 611795.5 4713762    3 1990-01-01 01:00:00 1990-01-01 01:00:00  2.016942
   cluster     pvalue
31    TRUE 0.02476587
30    TRUE 0.04459518
```

As it can be seen, two clusters (with a p-value lower than 0.05) are detected. However, they overlap and we will just consider the one with the lowest p-value, which is shown in Figure 4.2.1

```
> names(cl0)[3]<-"size"
> #brainnav$x<-coordinates(brainnav)[,1]
> #brainnav$y<-coordinates(brainnav)[,2]
> knbinary(brainnav, cl0)

   CL1 CL2
1    0   0
2    0   0
3    0   0
4    0   0
5    0   0
6    0   0
7    0   0
8    0   0
9    0   0
10   0   0
11   0   0
12   0   0
13   0   0
14   0   0
15   0   0
```

```
16   0   0
17   1   0
18   0   1
19   0   0
20   0   0
21   0   0
22   0   0
23   0   0
24   1   0
25   0   0
26   0   0
27   0   0
28   0   0
29   0   0
30   1   1
31   1   1
32   0   0
33   0   0
34   0   0
35   0   0
36   0   0
37   0   0
38   0   0
39   0   0
40   0   0
```

```
> brainnav$CLUSTER<-as.factor(knbinary(brainnav, cl0)[,1])
> levels(brainnav$CLUSTER) <- c("", "CLUSTER")
```

References

Gómez-Rubio V, López-Quílez A (2010). "Statistical methods for the geographical analysis of rare diseases." *Advances in experimental medicine and biology*, **686**, 151–171.

Jung I (2009). "A generalized linear models approach to spatial scan statistics for covariate adjustment." *Statistics in Medicine*, **28**(7), 1131–1143.

Kulldorff M (1997). "A Spatial Scan Statistic." *Communications in Statistics — Theory and Methods*, **26**(6), 1481–1496.

Ugarte MD, Ibáñez B, Militino AF (2004). "Testing for Poisson Zero Inflation in Disease Mapping." *Biometrical Journal*, **46**(5), 526–539.

Ugarte MD, Ibáñez B, Militino AF (2006). "Modelling risks in disease mapping." *Statistical Methods in Medical Research*, **15**, 21–35.

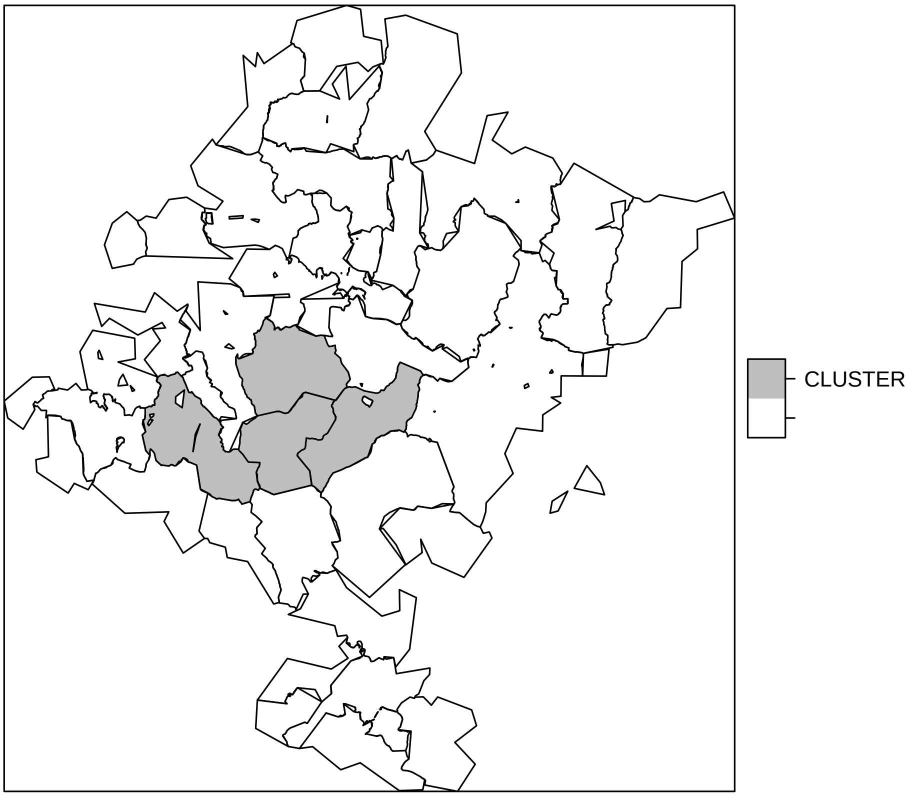

Figure 7: Cluster of brain cancer detected in Navarre (Spain).

Waller L, Turnbull B, Clark L, Nasca P (1992). "Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence in TCE-contamined dumpsites in upstate New York." *Environmetrics*, **3**, 281–300.

Waller LA, Gotway CA (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Hoboken, New Jersey.

Zhang T, Lin G (2009). "Spatial scan statistics in loglinear models." *Computational Statistics and Data Analysis*, **53**(8), 2851–2858.

Affiliation:

Virgilio Gómez-Rubio
Department of Mathematics
School of Industrial Engineering
University of Castilla-La Mancha
02071 Albacete, Spain

Paula Moraga-Serrano
London School of Hygiene & Tropical Medicine
Keppel Street
WC1E 7HT London, United Kingdom

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd