# Sample Answer Lab01: Evaluation of central tendencies, directional data and point densities

**Handed out:** Wednesday, January 30, 2019

**Return date:** Wednesday, February 13, 2019, at the beginning of the class.

**Grading:** This lab counts 12 % towards your final grade

### Task 1: Cluster identification [1 points]

Open the file **ClusteredPoints.dbf** to identify bivariate normal distributed clusters is this dataset with the function **Mclust( )**.

* 1. Estimate the optimal number of clusters in this dataset [0.25 points]

> p.Mclust <- Mclust(ClusterP, G=c(3,4,5), modelNames="VVV")

> summary(p.Mclust)

----------------------------------------------------

Gaussian finite mixture model fitted by EM algorithm

----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and

orientation) model with 4 components:

log.likelihood n df BIC ICL

-734.7653 200 23 -1591.392 -1608.453

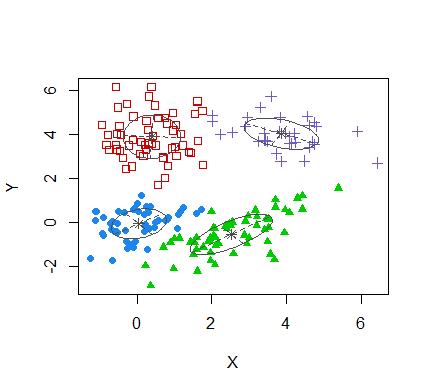
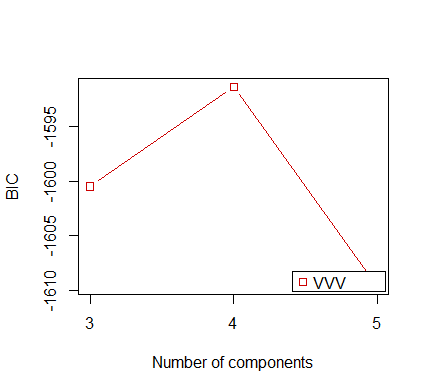
Clustering table:

1 2 3 4

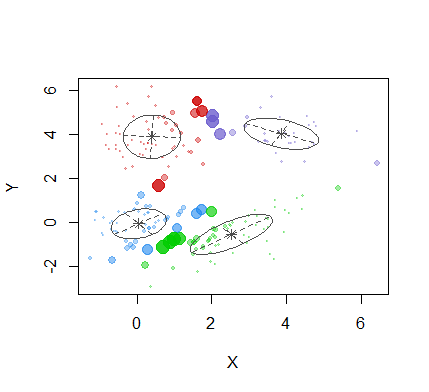
50 58 59 33

The estimated optimal number of clusters is 4.

* 1. Generate a plot of the estimated classification [0.25 points]



* 1. Generate a plot of the estimated uncertainty of assigning a point to a particular cluster and interpret the plot. [0.25 points]



* 1. Report in a table the estimate number of points in a cluster and the cluster centroids. [0.25 points]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster No. | 1 | 2 | 3 | 4 |
| No. of points | 50p | 58 | 59 | 33 |
| Cluster centroids | X 0.049  Y -0.046 | X 0.399  Y 3.913 | X 2.529  Y -0.528 | X 3.868  Y 4.047 |

### Task 2: Calculation of City Centroids/Medians [2 points]

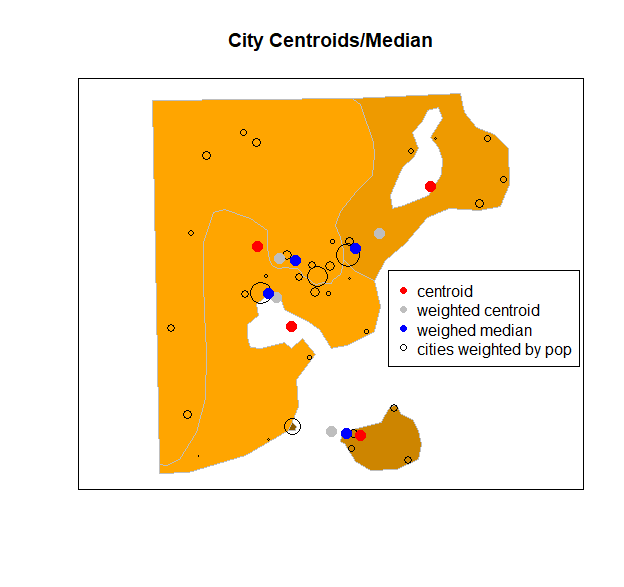
Open the map in the file **MyStudyArea.zip**.

Task 2.1: Calculate the simple city centroids in each region. [0.25 points]

Task 2.2: Calculate the population weighted city centroids and each region. [0.25 points]

Task 2.3: Calculate the population weighted Euclidian city medians in each region. [0.25 points]

Task 2.4: Map all three calculate centroids/medians jointly into a map with a proper map legend. Discuss which centroids or the median is most appropriate to show the center of the population distribution in a region. [0.5 points]



Weighted median is the most appropriate to show the center of population distribution within a region. The left two regions have concave shapes, the centroids are outside of the region, the region on the right has a hole inside, the centroid is also outside of the area. Since population distributions within a region are highly positively skewed, the weighted centroids are greatly impacted by the skewness, for example, the weighted centroid is outside of the bottom region. The island region has a small bridge head on the mainland. Therefore, the weighted centroid is pulled outside the study area.

The weighted median points are all within their region, and they best summarized the centers of population distributions within each region.

Task 2.5: The -script **EuclidianMedian.r** (see Lecture01) can be substantially optimized. [0.75 points]

Use the user-defined function **euclidMedian( )** in the script in combination with the  optimization function **nlm( )** to find the population weighted Euclidian Median. Show that part of your script which uses this optimization approach.

euclidMedian <- function(xy){

## Function that evaluates the cost at any point XY given

## the global vector argments: x, y, and w

cost <- sum(w \* sqrt( (x-xy[1])^2 + (y-xy[2])^2 ))

return(cost)

} #end::euclidMedian

**Med.nlm<- nlm(euclidMedian, c=(5,5))**

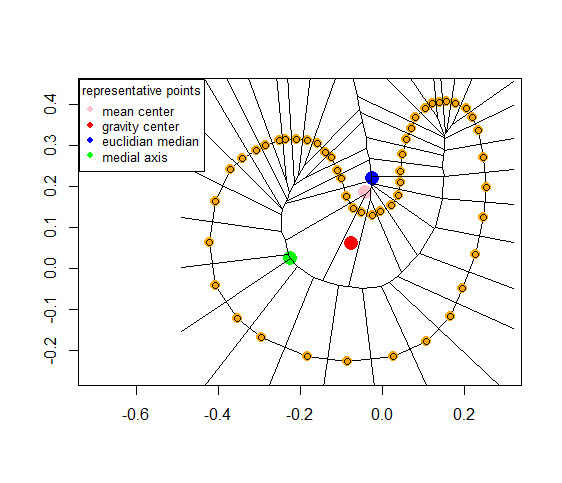
cat("x-Median = ",round(nlmOpti$estimate[1],3) ,"\n",

"y-Median = ",round(nlmOpti$estimate[2],3) ,"\n",

"Min. Cost= ",nlmOpti$minimum,"\n")

### Task 3: Polygon Centers [2 points]

Open the shape file **ShapeKindey.shp** which outlines the boundary of a polygon. You will identify several more or less representative points for this polygons. In task 3.1 to 3.3 plot these points into the polygon and make also sure to highlight the boundary shape points. No need to your code.



Note: the orange points depict the shape points of the polygon. These are used to evaluate the different polygon centers. Each shape point has an equal weight.

Task 3.1: Add the ***arithmetic mean center*** to the map of your polygon. Does this point reasonably well represent the center of the polygon? [0.5 points]

Outside of the polygon.

Task 3.2: Add the ***center of gravity*** to the map of the polygon. Does this point reasonably well represent the center of the polygon? [0.5 points]

Inside the polygon but pulled toward the region of high density shape points. This point is calculated by default for shape-files. It can be extracted from a polygon with command **coordinates( )** .

Task 3.3: Add the ***Euclidian median*** to the map of the polygon. Does this point reasonably well represent the center of the polygon? [0.5 points]

Outside of the polygon. For this application of the Euclidian median each shape-point has an equal weight of one.

Task 3.4: Add the ***medial axis*** point to the map of the polygon. Does this point reasonably well represent the center of the polygon? [0.5 points]

Inside the polygon and furthest away from the boundary edge.

### Task 4: Directional Data Analysis [3 points]

The file **CarCrashDirection.dbf** holds information about the time and direction a car was travelling, when an accident occurred. An insurance analyst wants to know if, dependent on the time of day, accidents happen at predominate travelling directions.

The analyst breaks the time of the crash down into a factor with four categories (a) morning rush hour, (b) day time, (c) evening rush hour, and (d) night time.

It is known that irrespectively of the time of day an equal number of cars is travelling in any direction, that is, the overall travel direction of all trips is uniformly distributed at any time.

Note: You need to use the function **circular( )** to notify  that the accident directions are reported in degrees. You may also use a geographic template.

Task 4.1: Why is it important for the analysis of the accident directions that the overall travel direction of all trips at any time is uniformly distributed? [1 point]

If the general traffic at any time would have a directional bias and assuming that the traffic accidents are proportional to the overall traffic, then we would expect to see coincidentally more accidents where more traffic takes place. On the other hand, if the general traffic follows a uniform directional pattern than any excessive direction of accidents points at a directional accident bias. Accident directions are a subset of the overall travel directions.

Task 4.2: Generate 4 rose diagrams for accident directions at each time of day and calculate the average direction and mean resultant length . Interpret the plots. [1 point]

|  |  |
| --- | --- |
|  |  |
| N:259, average: 95.27  Rho: 0.360, *p*-val: 0 | N:155, average: 18.90  Rho: 0.018, p-val: 0.949 |
|  |  |
| N:341, average: -87.59  Rho: 0.324, p-val: 0 | N:99, average: -61.86  Rho: 0.101, p-val: 0.367 |

The four plots indicate that the car accidents in the day and night follow an uniform distribution, however during the morning rush hours, we can observe more car accidents in the eastern direction and during the evening rush hours, more accidents can be observed in the western direction.

The -values for the day and night plots are close to 0, which also suggest an approximately uniform distribution. For the two plots on the right side, their -values are 0.360 and 0.324 respectively, which means their distributions are less likely to be uniform. Moreover, their direction means suggest that morning rush hours have more accidents on the eastern direction and evening rush hours have more accidents on the western direction. This may be related to the blinding effects of the sun being low on the horizon.

Task 4.3: Perform Rayleigh tests for each time period to test whether the accidents at a given time period occur uniformly distributed or at a predominate direction. Draw your conclusions with regards to your visual interpretation in task 4.2. [1 point]

See 𝑝-values in above plot. The morning and evening accident distributions have a significant directional bias, excessive crashes in the morning for vehicles driving into the rising easterly sun and in the evening for vehicles driving into the westerly setting sun.

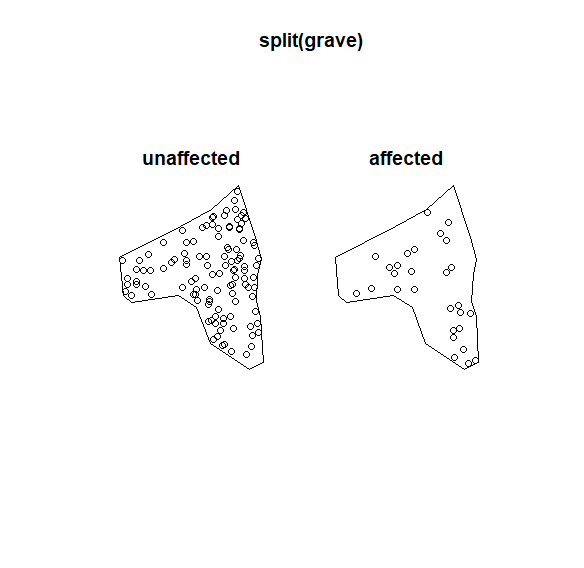
### Task 5: Kernel Densities and Relative Risk Ratios [4 points]

You will be using the medieval grave site example discussed in Waller & Gotway (2004). *Applied Spatial Statistics for Public Health Data*, which is available to read online at UTD’s library. The example is discussed on pp 134–136 and 167‒171. The marked point locations and the study area boundary are available in the  library **smacpod** and can be attached to your session with the command **data("grave", package="smacpod")**. Grave sites of persons with tooth defects are the cases.

Task 5.1: ***Map*** both the point locations of cases and controls in their study area. Describe the observed pattern. By ***visual inspection*** can you tell whether at all locations the number of cases is proportional to the number of controls? [0.5 points]

> data("grave", package="smacpod")

> plot(split(grave))



Both case and control points are well spread out in the study region, we can observe slight local clustering of affected grave sites. By visual inspection, it looks the number of case is not proportional to the number of controls at all locations. For example, on the north of the study area, it clusters some control points, however the case points do not show a cluster pattern at local level.

Task 5.2: ***Find*** the reasonable bandwidths of the cases and controls so that their kernel density maps describe the spatial pattern of the point locations in a meaningful way. [1.5 point]

***Compare*** your choosen density maps to the ones, which are slightly over-smoothing and under-smoothing the point pattern densities of both the cases and controls locations.

***Show*** all map pattern (optimal bandwidth, over- and under-smoothed) and ***justify*** your selection of the bandwidths for the cases and controls.

In general one should use a higher bandwidth for sparser point pattern, in order to avoid sub-regions with a density of zero. However, in relative risk ratio analysis the denominator of the ratio (the control) should be somewhat smoother. The there are no sub-regions with extremely low densities, which would make the ratio extremely large.

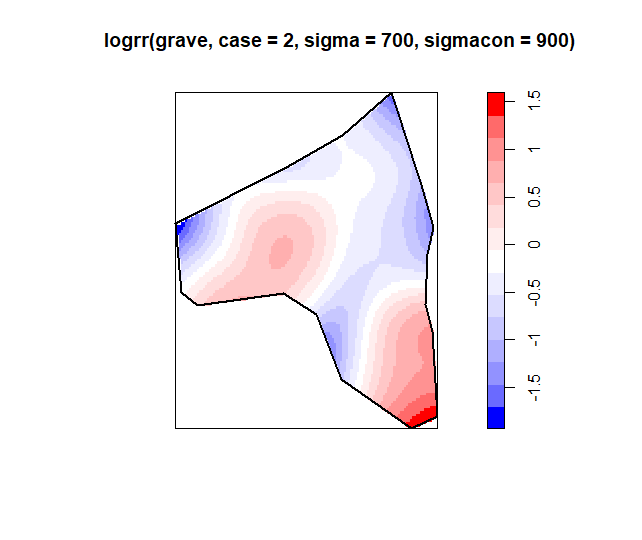
|  |  |  |
| --- | --- | --- |
|  | **Controls (n=113)** | **Cases (n=30)** |
| Under-smoothing leading to a granular map density |  |  |
| Reasonable. Local concentrations of case and controls as well as spars areas are well depicted. |  |  |
| Over-smoothing. The density surface becomes too homogeneous polishing away local properties of the point distributions. |  |  |

The top two kernel density maps are little spiky, they capture mainly individual point location but have not yet form a collective of points within a cluster. The middle two maps estimate the overall spatial pattern well, so they provide more reasonable representations of the spatial patterns with clusters of points and sub-regions of sparse distributions. The bottom two maps are slightly over-smoothed because they tend to blend distinct point clusters together.

Task 5.3: ***Generate*** the relative risk surface (i.e., log-risk surface) of case density over control density. You may what to further ***experiment*** with the bandwidths of the cases and controls to obtain a meaningful relative risk surface. [1 point]

Why is it advisable to make the bandwidth of the controls larger than the bandwidth of the cases?

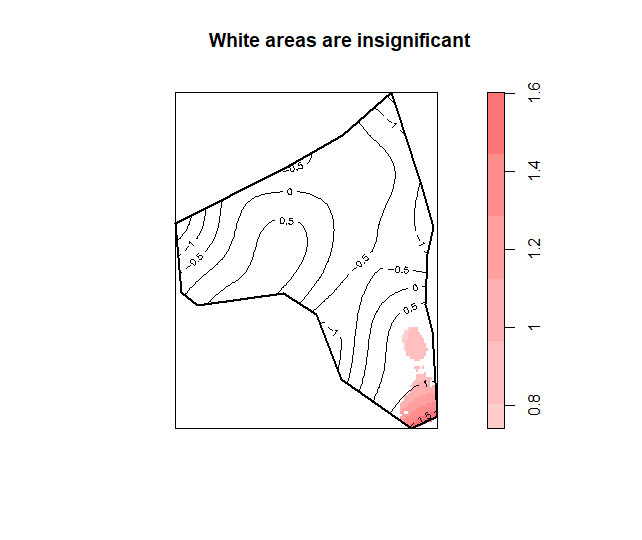
Interpret the observed map pattern. Which regions have a higher relative density of case burial sites and which regions have a smaller relative density? Does this match your visual inspection in task 5.1?



Because risk surfaces of the density is calculating the log relative risk ratio with control density pattern as the denominator, the relative risk will be undefined if the control density is close to zero, so it is advisable to over-smooth the control density by selecting a larger bandwidth. Meanwhile, in this case, there are more control points than the case points, so it is better to choose a larger bandwidth for control to capture the overall spatial pattern.

The risk surface map above indicates that the west and south of the study area has relatively higher risk, while moving towards to the north, the risk gets lower. These are consistent with what we can observe from the point map.

Task 5. 4: ***Evaluate*** in a one-sided test at an error probability of which regions have a significantly higher relative density of cases. Base your evaluations on 999 simulation runs. ***Show*** the obtained map and add the outline of the study area as reference frame to it. [1 point]



The null hypothesis of this test is that the relative risk is greater than 0, the p value is 0.01, so we should reject the null hypothesis and tentatively assume that the relative risk is significantly less than 0.

The significant map above shows most of the region appears insignificant when we assume the risk is greater than 0, it indicates that almost the whole study area has a relatively low risk except the southern sub-region, which might have a slightly higher risk.