

Project 2

Problem 3

Neumann-Scott Event RF

In the Neumann-Scott Event RF points are assumed distributed around some mother locations. The number of mother nodes often follow a poisson distribution. With intensity λ_m , the number of points (children) around the mother node is then assume to follow a distribution according to the count pdf $p(k)$ and the intensity $p(\mathbf{x}|\mathbf{x}_j^M)$. The points are assumed independent, around a mother node. The intensity pdf is usually gaussian, with center in the mother node.

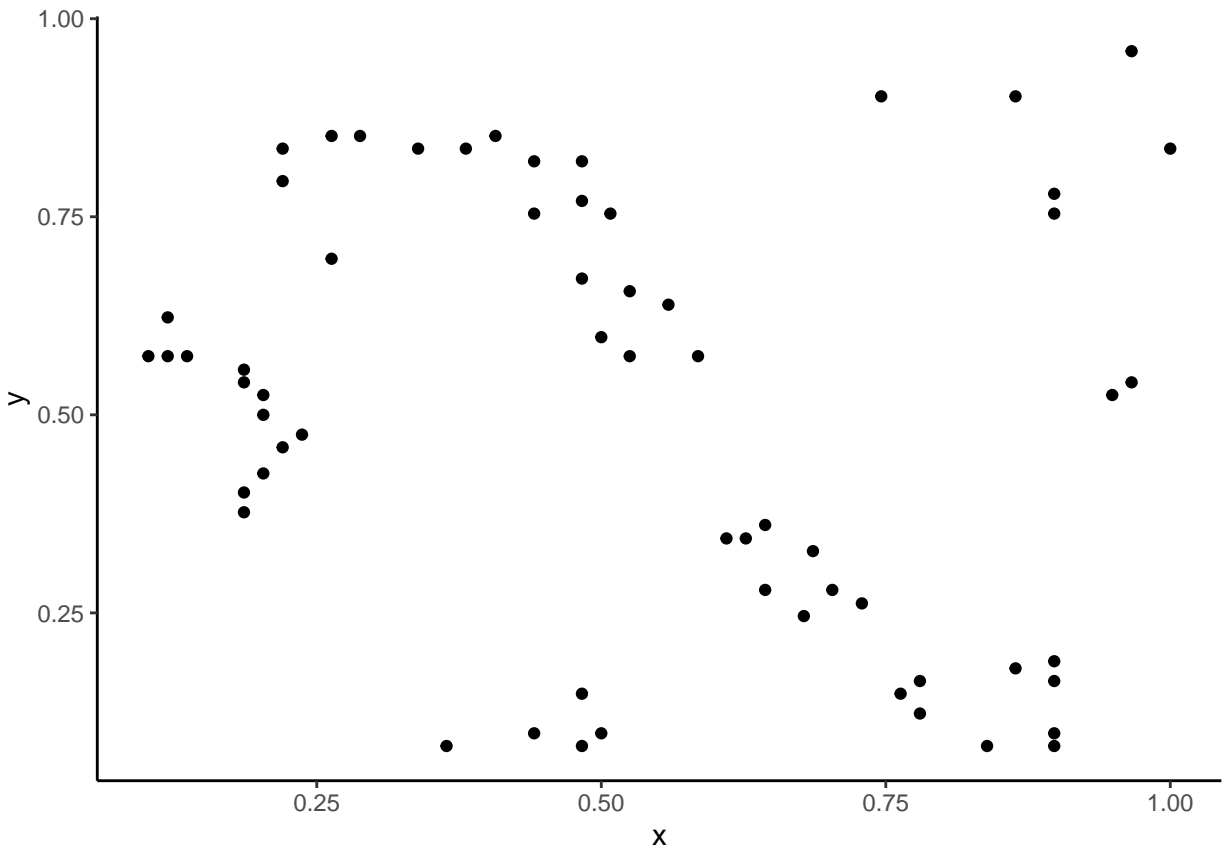
A possible border problem when dealing with a finite event space is that the intensity function, especially the gaussian, might generate points outside of the event space. A possibility is to use a function that enforces border conditions, i.e. the torus border condntions, when generating points.

Investigating the redwood data

We first plot the data:

```
redwood <- read.csv("redwood.dat", header = F, sep = " ")
redwood <- as.data.frame(redwood)
colnames(redwood) <- c("x", "y")

ggplot(data = redwood) + geom_point(aes(x,y)) + theme_classic()
```

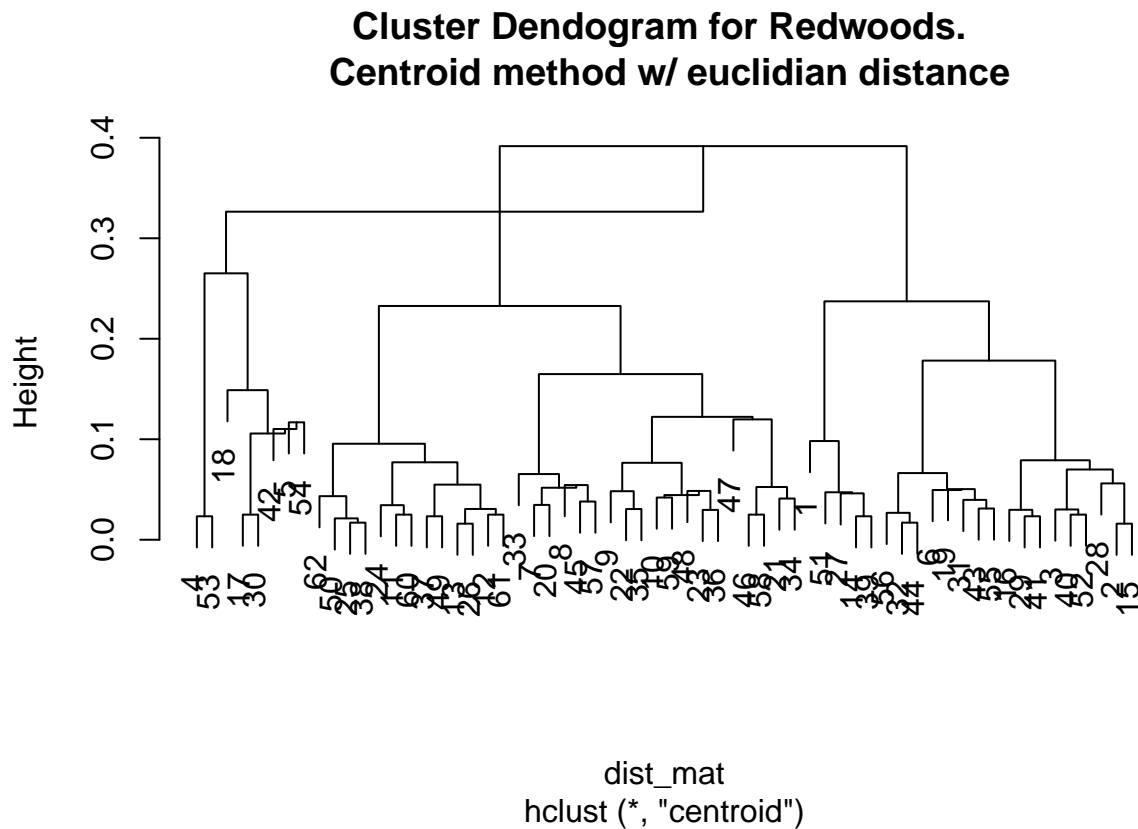


There seems to be some clusters of redwood. Want to fit an empirical Neumann Scott Event-ref.

From the data there seems to be 8 mother nodes, with clusters of redwoods around.

To avoid having to classify each point by hand, we use hierarchical clustering with L2 distance to find the cluster and their centers. We use centroid clustering, and will use the centroids as estimation of position to the mother node. We have already counted 8 mother nodes, so we will use that estimate.

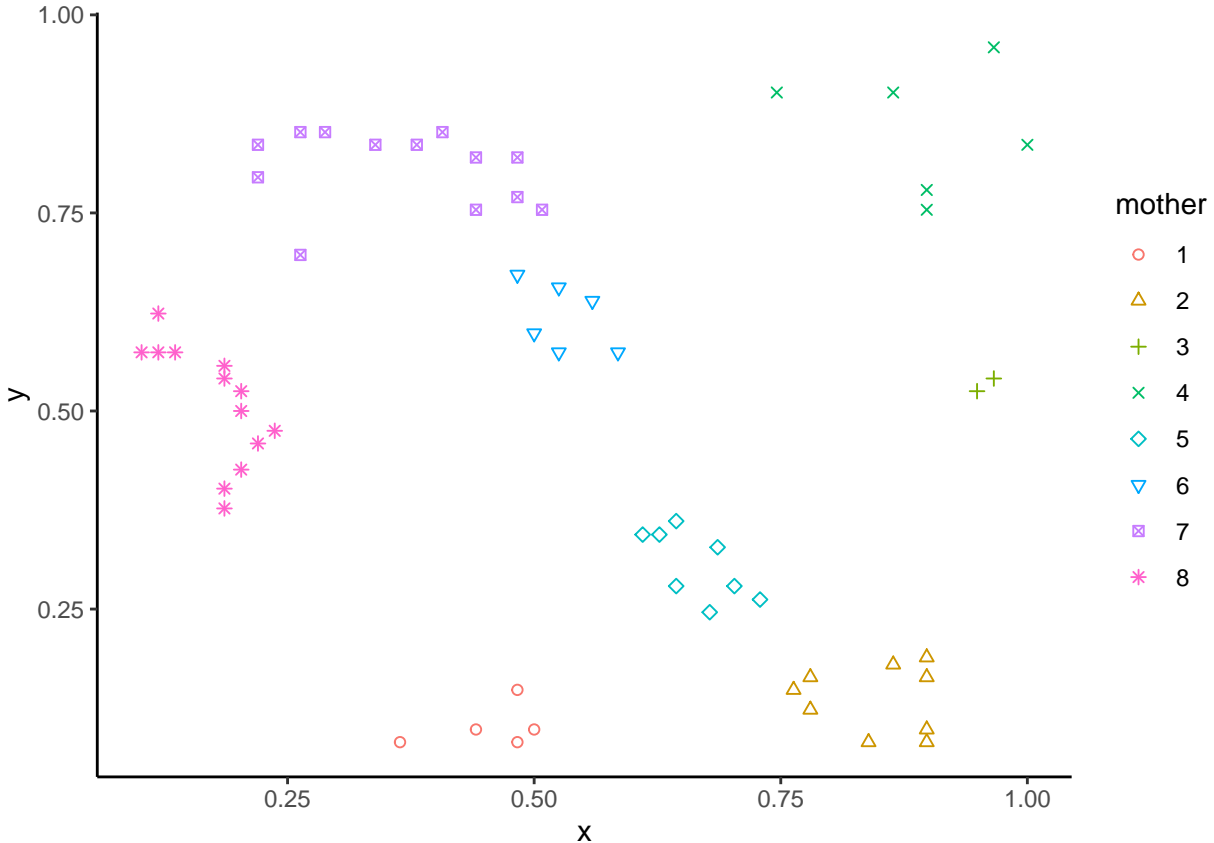
```
# Doing hierarchical clustering with centroid and euclidean
dist_mat <- dist(redwood, method = 'euclidean') # Calculate distances
hclust_centroid <- hclust(dist_mat, method = 'centroid') # Use centroid as mother node
plot(hclust_centroid, main = "Cluster Dendrogram for Redwoods. \n Centroid method w/ euclidian distance")
```



The cluster dendrogram is illustrated in Figure ??

```
cut_centroid <- cutree(hclust_centroid, k = 8) # Find node of elements when we have 7 groups
redwood$mother = as.factor(cut_centroid)

ggplot(redwood) + geom_point(aes(x= x, y = y, color = mother, shape = mother)) + theme_classic() + scale_x_continuous(limits = c(0, 100))
```



Classifying by cluster we get the following plot. See ??

We know find the position of the centroids. (As hclus does not give us that we do it manually)

```
# Finding centers and count for each mother
mothers <- redwood %>%
  group_by(mother) %>%
  summarise(x = mean(x),
            y = mean(y),
            n = n()) %>%
  as.data.frame()
```

We also want to estimate the standard deviation. If we assume that each cluster has independent standard deviation. And the each mother estimation is unbiased. We get: (as deviation in x-coordinate and y-coordinate are independent and equally normally distributed)

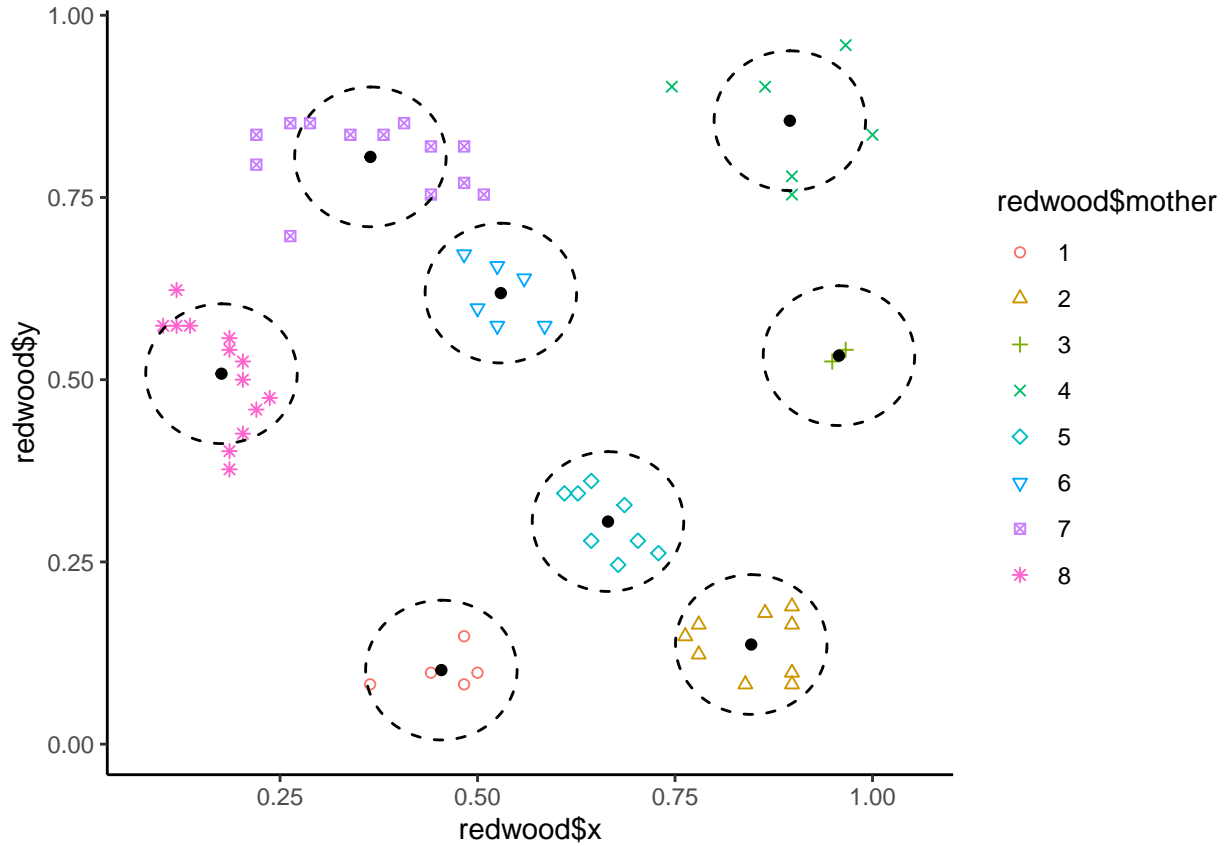
```
# Estimating standard dev
redwood_w_mother <- left_join(redwood, mothers, by = c("mother"))
colnames(redwood_w_mother) <- c("x", "y", "mother", "x.m", "y.m", "n")
redwood_w_mother$dist_mother <- sqrt((redwood_w_mother$x - redwood_w_mother$x.m)^2 + (redwood_w_mother$y - redwood_w_mother$y.m)^2)
redwood_w_mother$x.dist <- redwood_w_mother$x - redwood_w_mother$x.m
redwood_w_mother$y.dist <- redwood_w_mother$y - redwood_w_mother$y.m

std2 <- 1/(length(redwood_w_mother$x.dist) + length(redwood_w_mother$y.dist) - 1) * (sum(redwood_w_mother$dist_mother^2))
q95 <- qnorm(0.95, mean = 0, sd = sqrt(std2))
```

If we assume that they a common standard deviation we get: 0.0033968

Plotting the redwood forest, with the mother nodes we get:

```
ggplot() +
  geom_point(aes(x= redwood$x, y = redwood$y, color = redwood$mother, shape = redwood$mother)) +
  theme_classic() +
  scale_shape_manual(values=1:nlevels(redwood$mother)) +
  geom_point(aes(x = mothers$x, y = mothers$y)) +
  geom_circle(aes(r = rep(q95, length(mothers$y)), y0 = mothers$y, x0 = mothers$x), linetype = "dashed")
```



TODO: Q are mother nodes a part of the count ? TODO: Q what is the TORUS-3d algorithm not in my version of the note (p. 131). Is it to project outliers onto some torus? Then shifting them around some

Create a grid of 100 points and estimate the J-function.

$$J(t) = \frac{\mathbf{E}(k_{B_{\mathbf{x}_0}(t)} - 1)}{|B_{\mathbf{x}_0}(t) \cap D|} \quad (1)$$