Case-Control Point Data Homework (First- and Second-Order Properties)

Instructions: Answer the following questions and write your answers in a word processor. Mathematical symbols should be written using the equation editor. Appropriate graphics should be included. The document may also be created in LaTeX, though this is NOT encouraged.

Problem 1

The `urkiola` data set in the **spatstat package** contains locations of birch (Betula celtiberica) and oak (Quercus robur) trees in a secondary wood in Urkiola Natural Park (Basque country, northern Spain). They are part of a more extensive dataset collected and analysed by Laskurain (2008). The coordinates of the trees are given in meters. Let the “oak” trees be the cases and “birch” trees be the controls.

```
        library(smacpod)
        library(spatstat)
```

```
Loading required package: spatstat.data

Loading required package: spatstat.univar

spatstat.univar 3.1-4

Loading required package: spatstat.geom

spatstat.geom 3.5-0

Loading required package: spatstat.random

spatstat.random 3.4-1

Loading required package: spatstat.explore

Loading required package: nlme

spatstat.explore 3.5-2

Loading required package: spatstat.model

Loading required package: rpart

spatstat.model 3.4-0

Loading required package: spatstat.linnet

spatstat.linnet 3.3-1
```

```
spatstat 3.4-0
For an introduction to spatstat, type 'beginner'
```

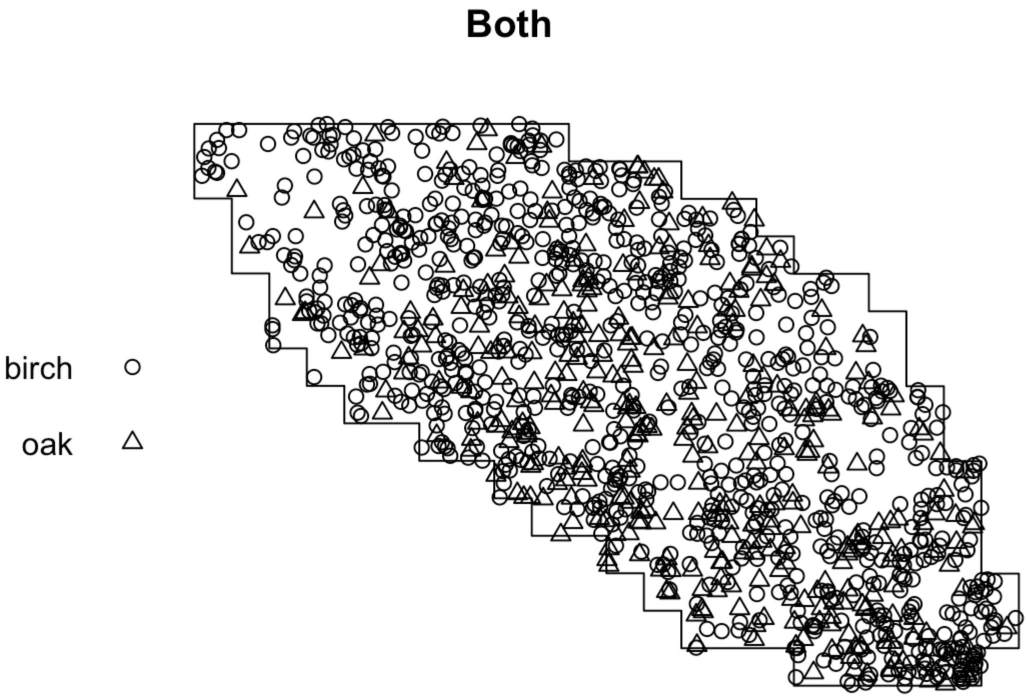
```
library(spatstat.random)
data(urkiola)
urkiola
```

```
Marked planar point pattern: 1245 points
Multitype, with levels = birch, oak
window: polygonal boundary
enclosing rectangle: [0.05, 219.95] x [0.05, 149.95] metres
```

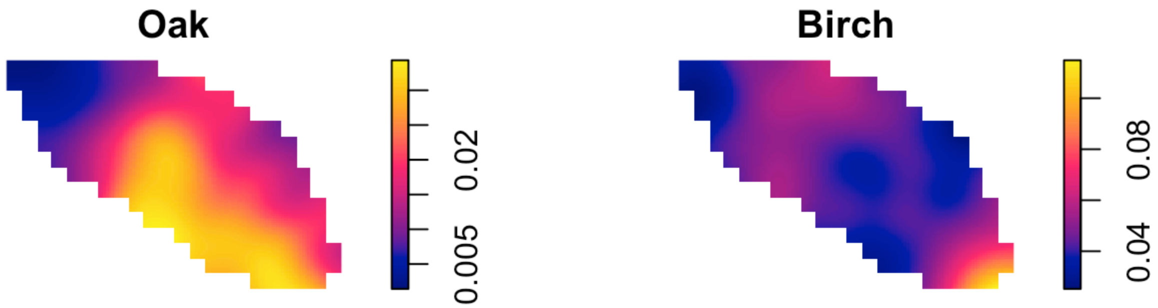
a.

Create a plot of the of the point pattern that distinguishes between the two types of trees. Do you notice any evidence of clusters? Explain.

```
plot(urkiola, main = "Both")
```



```
par(mfrow = c(1,2))
plot(density(urkiola[marks(urkiola)== "oak"], sigma = 14.5), main = "Oak")
plot(density(urkiola[marks(urkiola)== "birch"], sigma = 15), main = "Birch")
```



It seems like for oak there are more towards the bottom left while for birch we see events clustered in the southeast, however, this could be due to boundaries.

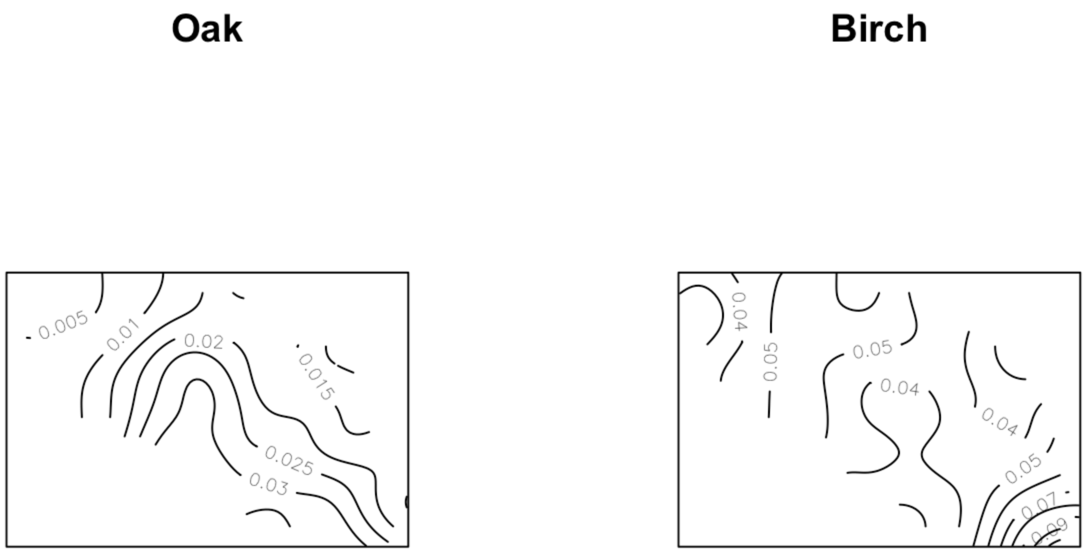
b.

Create contour plots of the spatial density function for the oak and birch trees. Use a bandwidth of 14.5 for the oak trees and 15 for the birch trees. Do any differences jump out to you?

```
par(mfrow = c(1,2))

contour(density(urkiola[marks(urkiola)== "oak"], sigma = 14.5), main = "Oak")

contour(density(urkiola[marks(urkiola)== "birch"], sigma = 15), main = "Birch")
```



Yes we seem to have a lot more oak trees in the middle compared to birch, basically they are complete opposite of each other except in the southwest. They do seem to have very different patterns to some degree looking at where the lines are.

c.

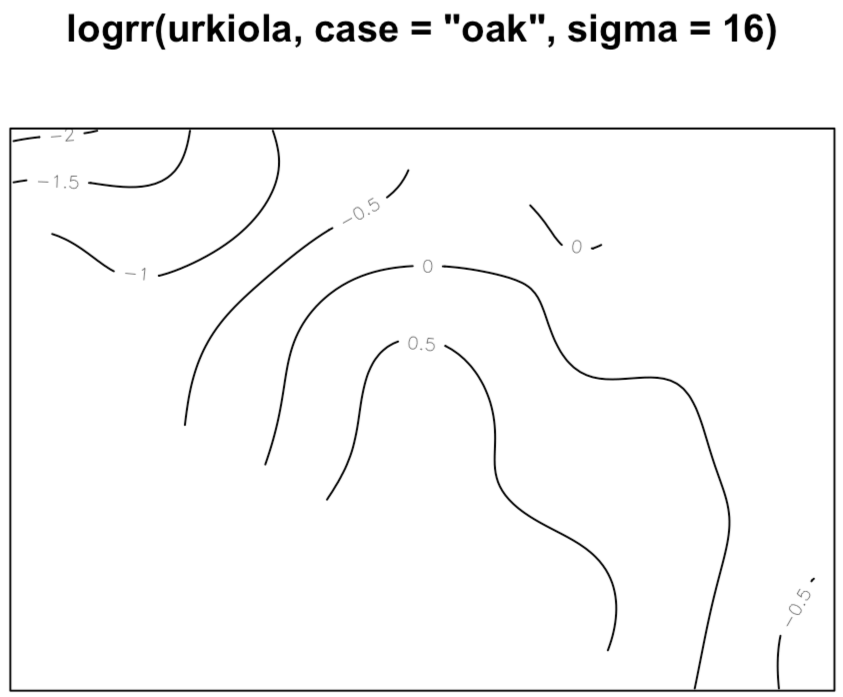
Estimate the log ratio of the spatial density of the oak trees relative to the birch trees ($r(s)$) using a bandwidth of 16. Create a contour plot of the log ratio. Which areas are most inconsistent with the belief that the spatial densities for the two types of trees are the same?

```
        logrr_ukiola <- logrr(urkiola, case = "oak", sigma = 16)
```

```
oak has been selected as the case group
```

```
        contour(logrr(urkiola, case = "oak", sigma = 16))
```

```
oak has been selected as the case group
```



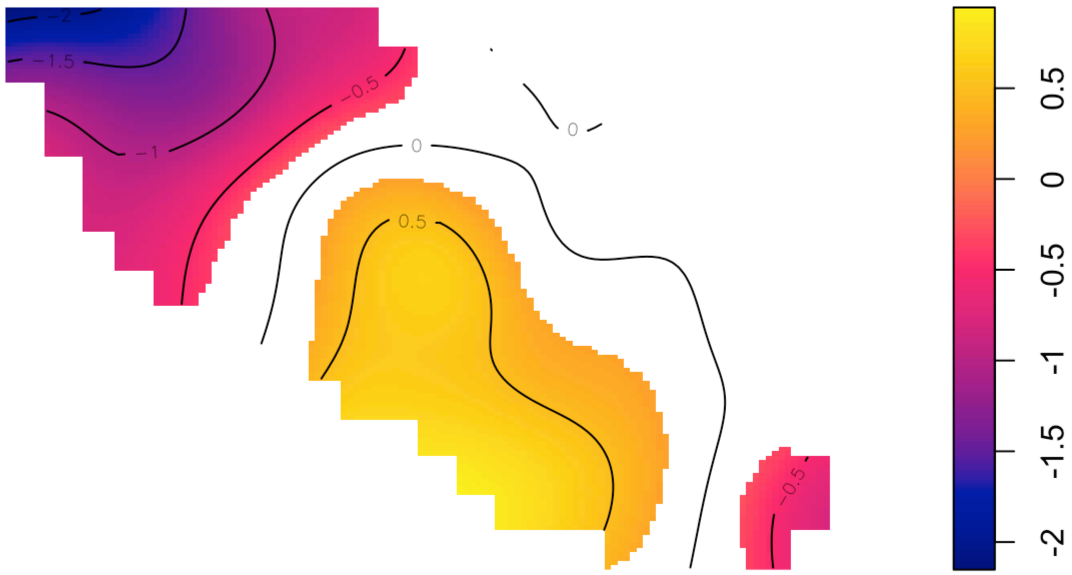
The areas that are the most inconsistent are exactly the same areas I was talking about. The south west where it is positive we are more likely to see oak trees and in the south east we are more likely to see birch. Now something I didn't catch before was the north west where we have a big negative value thus more likely to see birch than oak. If not exactly see the calculated densitys there are either bigger or smaller.

d.

Construct pointwise 95% tolerance envelopes for $r(s)$ using 499 data sets simulated under the random labeling hypothesis. Plot the regions above and below the tolerance envelopes in different colors. Overlay the contour plot of $\tilde{r}(s)$. What can you conclude?

```
if (file.exists("urkiola2.rda")) {
  load("urkiola2.rda")  # loads urkiola2 if it's inside the file
} else {
  urkiola2 <- logrr(urkiola, case = "oak", nsim = 499, level = 0.95)
  save(urkiola2, file = "urkiola2.rda")
}

plot(urkiola2)
```



It seems like there is some clustering in the reagions I talked about for oak trees in comparison to birch. Again, pretty much what I was able to see earlier but now a lot more defined. With these envelopes showing how different it is.

e.

Perform a global test of clustering using $\tilde{r}(s)$. Is there convincing evidence of clustering of one group relative to the other for at least one location in the study area?.

```
        logrr.test(logrr(urkiola, case = "oak", sigma = 16, nsim = 499))
```

```
oak has been selected as the case group


Kelsall and Diggle (1995) test for log relative risk

r(s) = ln[f(s)/g(s)]
f(s) = spatial density of cases at location s
g(s) = spatial density of controls at location s
case label:  oak
control label:  birch
```

```
null hypothesis: r(s) = 0 for all s in study area
alternative hypothesis: r(s) != 0 for at least one s in study area
test statistic: 7839.597
p-value: 0.002
nsim: 499
simulation procedure: random labeling
```

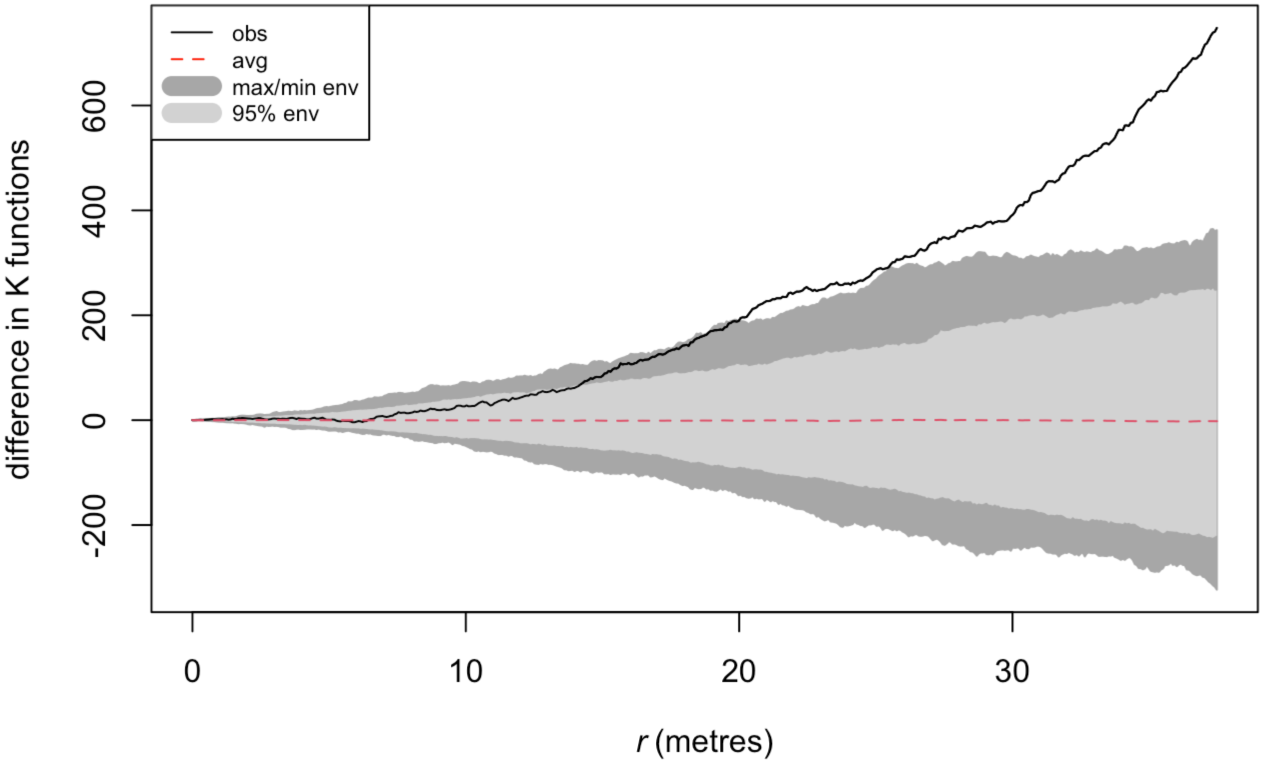
Yes, it seems like there is strong evidence of clustering under the study area since we see a p value of .002 which shows pretty strong evidence of clustering, again within the study area.

f.

Construct a plot for the difference in K functions between the oak trees and the birch trees.
Also include min/max envelopes for this difference using 499 simulated data sets under the random labeling hypothesis.
Also construct 95% pointwise tolerance envelopes.
Also include the mean difference from the simulations.
Provide a legend labeling these things.
Are there any spatial scales for which the oak trees are more clustered than the birch trees in comparison to what we expect under the random labeling hypothesis (or vice versa)? If so, what scales?

```r
nsim = 499
if (!file.exists("kdenv.rda")) {
  kdenv = kdest(urkiola, case = "oak", nsim = 499,
                level = 0.95)
  save(kdenv, file = "kdenv.rda")
}
load("kdenv.rda")

plot(kdenv, ylab = "difference in K functions")
legend("topleft",
        legend = c("obs", "avg", "max/min env", "95% env"),
        lty = c(1, 2, 1, 2),
       cex = .7,
        col = c("black", "red", "darkgrey", "lightgrey"),
        lwd = c(1, 1, 10, 10))
```



```
print(kdenv)
```

```
Envelopes for difference in estimated K functions

KD(r) = K_case(r) - K_control(r)
case label:  oak
control label:  birch
KD(r) computed for r betwen  0 and 37.475
number of simulations:  499
simulation procedure: random labeling
envelope level:  0.95
```

```
summary(kdenv)
```

```
KD(r) > upper envelope limit for the following r:
0.7319336
14.19951 to 37.475
```

Yes at 14.19 and 37.475 we see the upper envelope limit be reached. So basiaclly at a scale of 14.19 and bigger. Thus at that scale at least we see some more clustering done by oak in comparison to birch trees under the random labeling hypothesis.

g.

Perform a global test for clustering of the oak trees relative to the birch trees using the KD_+ statistic. Interpret the results.

```
kdplus.test(kdenv)
```

```
Diggle and Chetwynd (1991) test for difference in K functions

KD(r) = K_case(r) - K_control(r)
case label:  oak
control label:  birch

null hypothesis: KD(r) = 0 for all r between 0 and 37.475
alternative hypothesis: KD(r) > 0 for at least one r between 0 and 37.475
test statistic: 1533.721
p-value: 0.002
nsim: 499
simulation procedure: random labeling
```

Considering that the p value is so small we reject the null hypothesis that global there aren't at some scale some sort of clustering.

Problem 2

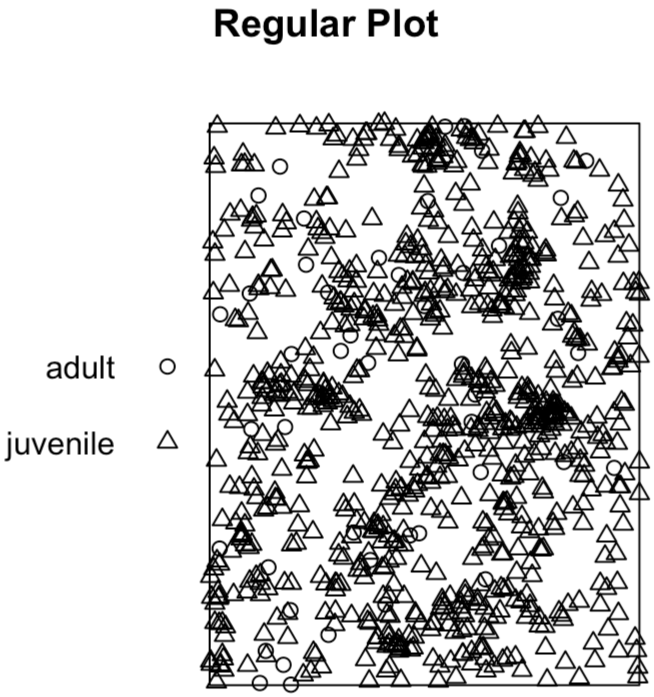
The `paracou` data set in the **spatstat** package contains data for Kimboto trees observed in Paracou, French Guiana. Let the juveniles be the controls and adults be the cases. Use a bandwidth of 40 when estimating the densities of the juveniles and 65 for the adults. Use a bandwidth of 52.5 when estimating the log ratio of spatial densities.

```
data("paracou")
```

a.

Create a plot of the of the point pattern that distinguishes between the two types of trees. Do you notice any evidence of clusters? Explain.

```
plot(paracou, main = "Regular Plot")
```

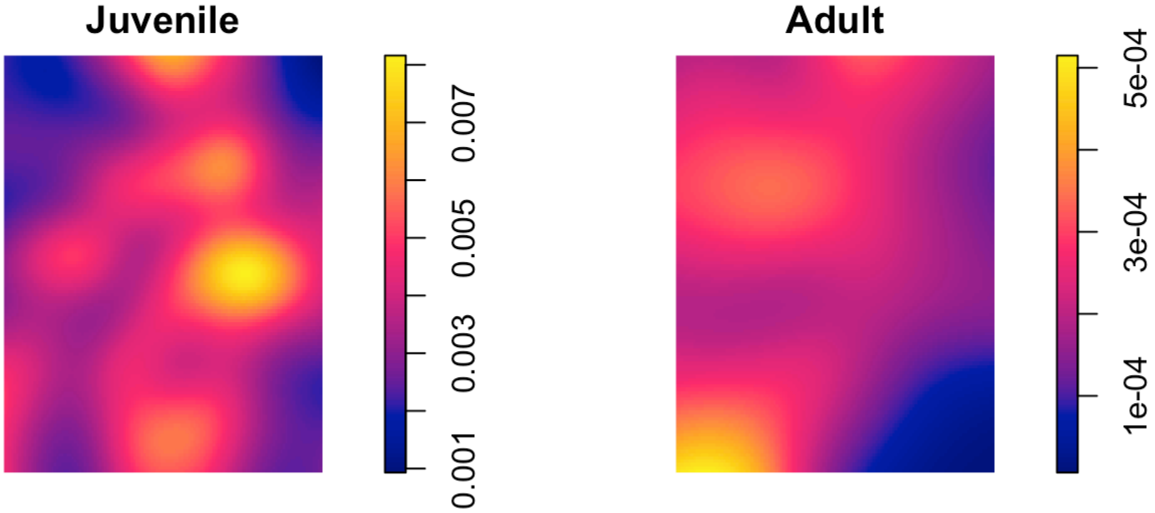


b.

Create contour plots of the spatial density function for the adult and juvenile trees. Use a bandwidth of 65 for the adult trees and 40 for the juvenile trees. Do any differences jump out to you?

```
        par(mfrow = c(1,2))

        plot(density(paracou[marks(paracou)== "juvenile"], sigma = 40), main = "Juvenile"
        plot(density(paracou[marks(paracou)== "adult"], sigma = 65), main = "Adult")
```



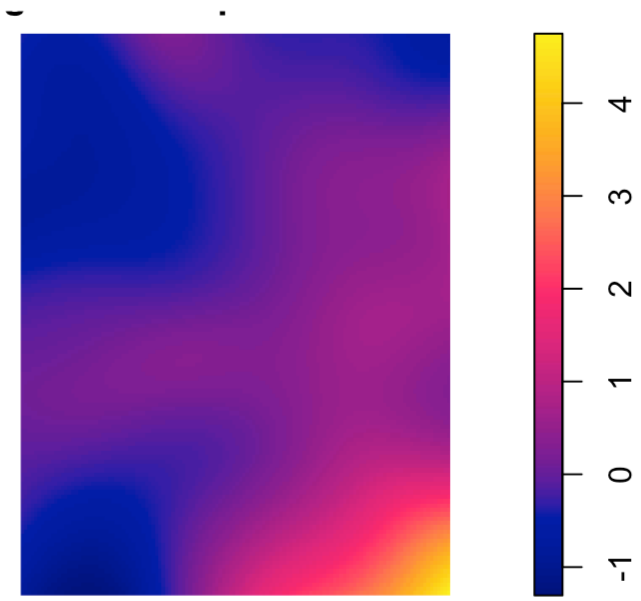
As we see there is high intensity for the juvenile where it is low for the adult population. The same thing can be said vice-versa.

c.

Estimate the log ratio of the spatial density of the adult trees relative to the juvenile trees $(r(s))$ using a bandwidth of 52.5. Create a contour plot of the log ratio. Which areas are most inconsistent with the belief that the spatial densities for the two types of trees are the same?

```
plot(logrr(paracou, case = "juvenile", sigma = 52.5), main = "Log Ratio of Spacia
```

```
juvenile has been selected as the case group
```

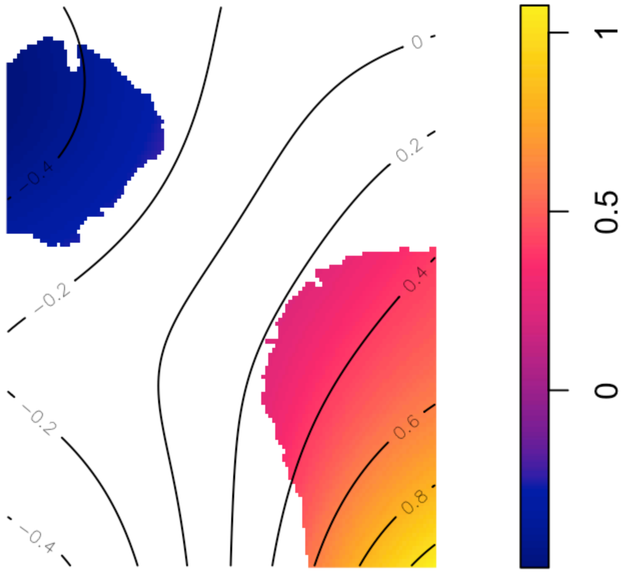


Alsmot everywhere, but especially in the bottom right corner and moving up, as well as the bottom left since they are both in different color spectrum.

d.

Construct pointwise 95% tolerance envelopes for $r(s)$ using 499 data sets simulated under the random labeling hypothesis. Plot the regions above and below the tolerance envelopes in different colors. Overlay the contour plot of $\tilde{r}(s)$). What can you conclude?

```
if (!file.exists("paracou2.rda")) {
  paracou2 = logrr(paracou, case = "juvenile", nsim = 499,
                   level = 0.95)
  save(paracou2, file = "paracou2.rda")
}
load("paracou2.rda")
plot(paracou2)
```



This shows that it is clearly in the bottom right where we see the most change and diviation from expected change. Thus we could say that there are more “cases” of juvenile trees in the bottom right in terms of clustering in comparison to our adult “cases”.

e.

Perform a global test of clustering using $\tilde{r}(s)$). Is there convincing evidence of clustering of one group relative to the other for at least one location in the study area?.

```
logrr.test(logrr(paracou, case = "juvenile", nsim = 499))
```

```
juvenile has been selected as the case group


Kelsall and Diggle (1995) test for log relative risk

r(s) = ln[f(s)/g(s)]
f(s) = spatial density of cases at location s
g(s) = spatial density of controls at location s
case label:  juvenile
control label:  adult
```

```
null hypothesis: r(s) = 0 for all s in study area
alternative hypothesis: r(s) != 0 for at least one s in study area
test statistic: 20469.2
p-value: 0.022
nsim: 499
simulation procedure: random labeling
```

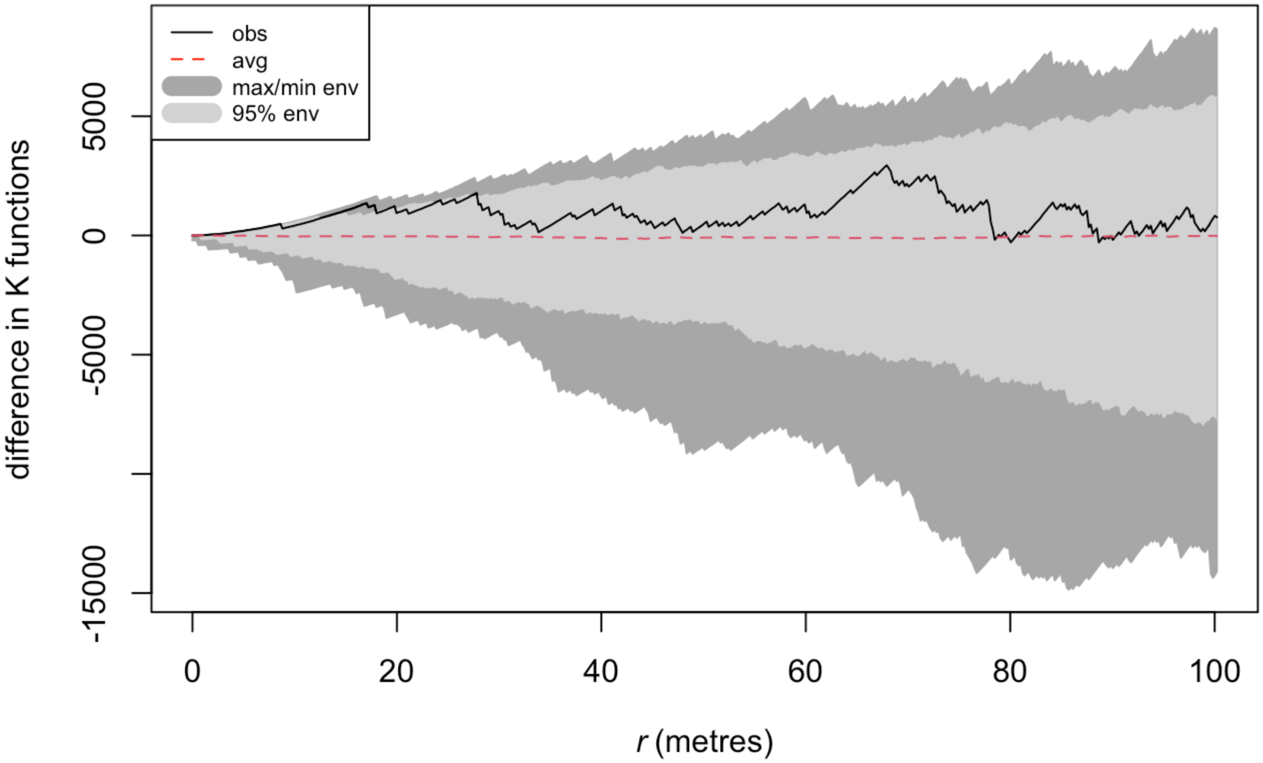
There is some evidence that at some level there is some clustering in at least one location of the study area. Since we reject the null hypothesis that for all the s in the study area our r(s) which is the ratio of intesnity functions is equal to zero.

f.

Construct a plot for the difference in K functions between the adult trees and the juvenile trees. Also include min/max envelopes for this difference using 499 simulated data sets under the random labeling hypothesis. Also construct 95% pointwise tolerance envelopes. Also include the mean difference from the simulations. Provide a legend labeling these things. Are there any spatial scales for which the adult trees are more clustered than the juvenile trees in comparison to what we expect under the random labeling hypothesis (or vice versa)? If so, what scales?

```r
        nsim = 499
        if (!file.exists("kkdenv.rda")) {
          kkdenv = kdest(paracou, case = "juvenile", nsim = 499,
                        level = 0.95)
          save(kkdenv, file = "kkdenv.rda")
        }
        load("kkdenv.rda")

        plot(kkdenv, ylab = "difference in K functions")
        legend("topleft",
                legend = c("obs", "avg", "max/min env", "95% env"),
               cex = .7,
                col = c("black", "red", "darkgrey", "lightgrey"),
                lwd = c(1, 1, 10, 10),
                lty = c(1, 2, 1, 2),
               )
```



As we can see it does seem like they are more clustered at every point than the average, however, notice that they do tend to be more clustered which is what we saw on those graphs, however, nothing above expectation according to random labeling hypothesis.

g.

Perform a global test for clustering of the adult trees relative to the juvenile trees using the KD_+ statistic. Interpret the results.

```
kdplus.test(kkdenv)
```

```
Diggle and Chetwynd (1991) test for difference in K functions

KD(r) = K_case(r) - K_control(r)
case label:  juvenile
control label:  adult

null hypothesis: KD(r) = 0 for all r between 0 and 100.2142
alternative hypothesis: KD(r) > 0 for at least one r between 0 and 100.2142
test statistic: 346.796
p-value: 0.152
```

```
nsim: 499
simulation procedure: random labeling
```

With this test we fail to reject the null hypothesis that there isn't any sort of global clustering. As we can see this matches up with our perception we saw with our envelopes since we don't really see our actual observed statistic leave this at any point. Given my bandwidth.

Problem 3

Let $\lambda_0(s)$ denote a control intensity function and $\lambda_1(s)$ denote a case intensity function defined over a study area D. Assume that $\lambda_1(s) = c\lambda_0(s)$ for all $s \in D$. Show that in this case, $r(s) = 0$ for all $s \in D$.

$$If\$\lambda_1(s) = c\lambda_0(s)\$thenlookingat \qquad \begin{aligned} r(s) &= \log \frac{f(s)}{g(s)} \\ &= \log \frac{\lambda_1(s)}{\lambda_0(s)} - \log \frac{\int \lambda_1}{\int \lambda_0} \\ &= \log \frac{c\lambda_0(s)}{\lambda_0(s)} - \log \frac{c \int \lambda_0}{\int \lambda_0} \\ &= \log(c) - \log(c) \\ &= 0 \end{aligned}$$

Problem 4

Let $\lambda_0(s)$ denote a control intensity function and $\lambda_1(s)$ denote a case intensity function defined over a study area D. Assume that $\lambda_1(s) = c\lambda_0(s)$ for all $s \in D$.. Show that $f(s) = g(s)$ for all $s \in D$, where f and g are the spatial densities of the cases and controls, respectively.

$$\begin{aligned} &\text{Assume } \lambda_1(s) = c\lambda_0(s). \\ &\text{From before we know } \log \frac{f(s)}{g(s)} = r(s) = 0. \\ &\text{Thus we must have } f(s) = g(s): \end{aligned}$$

$$\begin{aligned} r(s) &= \log \frac{f(s)}{f(s)} \\ &= \log 1 \\ &= 0. \end{aligned}$$