Tree Point Process For   
The City of Portland, Maine

By

Luke Kaim

If questions please email [Lucas.kaim@maine.edu](mailto:Lucas.kaim@maine.edu)

# Introduction (objectives and hypotheses)

# The objectives of this project were to better understand a tree point process for the City of Portland, Maine. One of the questions that I was interested in looking at in more detail was if it was possible to help the City of Portland, Maine know specific geographic locations to plant more trees. I worked for the City of Portland this summer and helped them with their inventory. I wanted to see if it was possible to use some of the spatial analysis techniques that we have learned this semester to suggest locations that should be targeted for increased planting. Can spatial analysis be used by the City of Portland to suggest places to plant more trees? Spatial analysis techniques that we have learned can be used to help suggest planting locations. The other topic that I was interested in look at was how fairly the city allocates city resources across space and across socioeconomic levels. This topic could suggest how fairly the City assets are allocated. This led to this question; is there a positive correlation between income and tree intensity? There would be a high correlation between income and tree intensity. If people make more money I reasoned that they would have more say or power in the City and as a result have a higher tree density in area where there is a higher socioeconomic level. I looked at a point process of trees for the City of Portland, Maine. My work assumes that the tree survey was fully enumerated. This was an assumption that does not hold true, but for some of the statistical tests one needs to use the population. There is sampling bias in the survey, but that will be covered later in the paper.

# Deception of data

The city of Portland was gracious enough to give me almost all of the City’s GIS data. I did delete a few points because there spatial location looked questionable to me. There object ID numbers are 17584, 18384, and 20008. I also deleted all points if the field stated the tree were dead because the City is going to remove those trees anyways. This work is to locate areas where more trees should be planted so I am treating the dead trees as not being there. There was another field for maintenance and in that field there was a remove ID. I did not delete those points because I did not know the unique id for remove. This could be done for future work. There was a field called stump which delineated if there was a tree stump or not. I did not delete stump; though I could have. This could be done for future work. I used the GIS Shapefile titled zoning because when I dissolved that file I was left with only one polygon. When I tried to do the same thing with the parcel file there were topological errors. Also the parcel file is not closed all the way around which presented an issue and would have been very hard to correct given the time that I had. I did join road ways to parcels and for the most part that looks to be a reasonable parcel file, but there were still a few issues even when I took the union of the two files. I did clip all data to Portland proper. For this analysis I used nearest neighbor distances and having water between the islands would give incorrect nearest neighbor distances. For future work I could look at the islands within Portland.

I also needed to obtain Census tiger files because I was interested in income and population count. I obtained the Census data from the US Census Bureau. I also needed socioeconomic level per tract; that I got from the 5 year American Community Survey. I joined the tract to the income table. This then gave me the polygon with population and median income.

# Similar studies

For the analysis of this point process I used most of the methods that Bailey and Gatrell discuss in Interactive Spatial Data Analysis. I used the CrimestatIII reading to understand what the kernel density estimation should be divided by. I thought population could work, but they suggest parcel centroid. One of the articles that was very similar to my topic is titled Ecological Information from Spatial Patterns of Plants: Insights from Point Process Theory. This was written by Law, R. et al. This article used kernel density to look at tree density. I also looked into how interpolation methods could be used to understand forest distribution. The title of this article is Geostatistical Evaluation of Natural Tree Regeneration of a Disturbed Forest by J. G. F. Garnica et al. This article looked at using Inverse distance weighting to estimate forest distribution. Because I was using kernel density heavily in my analysis I also wanted to look at papers specific to kernel density. I looked at Estimating Probability Surfaces for Geographical Point Data: An Adaptive Kernel Algorithm by Brunsdon, C. I also looked at Splancs: Spatial Point Pattern Analysis Code in S-Plus by Rowlingson, B. S., and P. J. Diggle to get a better understanding of the methods used for point processes.

# Methods

# I was interested in understanding the tree point process at a higher level. One of the first ways of achieving this goal is to simply plot the data. I plotted the data and saw that there were more points in Southern Portland then in Northern Portland. Next I used kernel density estimation to look at a first order trend. A first order trend can be thought of in terms of the intensity or mean changing over space. The intensity is higher in Southern Portland. This would indicate that there is a first order trend in the data. A second order trend can be thought of as how much spatial dependency there is in the data. Second order effects look at local or at a small scale. Second order effects happen because of spatial correlation of the data. Trees in a forest have a second order effect because ones trees offspring will be close to their parents. There are different statistical techniques to tease out if there is a first order trend or a second order trend or a combination of the two. With the City of Portland data we see both a first and second order effect. This leads me to believe that we are seeing a combination of first and second order effects.

# Kernel Density

Kernel density looks at first order effects within a point population. Kernel density estimation is a way to visually see if the variation is different at different scales. This is done by using bandwidth. Of course depending on how large or small ones bandwidth is will affect the intensity estimation. The intensity is calculated using a moving window. Depending on how many points fall within the window gives one the intensity.

One issue with this is that changing bandwidth changes the intensity estimation. There are ways to find the best bandwidth to use. One way is to plot the K function to zero where the highest point on the graph tells us the maximum amount of clustering. Then one needs to look at the distance that that maximum clustering occurs at to tell the optimal bandwidth. In R there is also a command called bw.diggle which does a similar process and tries to minimize mean error. Once the optimal bandwidth is chosen then one can plot the kernel density. The kernel density can help us understand the intensity or mean over space.

Nearest neighbor distances

Nearest neighbor distances is a way of characterizing the second order effect in the data. Nearest neighbor works by looking at inter-event distances. Nearest neighbor uses the distance between points to calculate if there is spatial dependency in the data. If the distance is small this would be a good indication that there is a second order effect. This can be taken a step farther, however, and one could look at Ghat and Fhat. Ghat is a measure of event-event distance. This can be plotted to help us understand the phenomenon. Fhat is a measure similar to Ghat, but it measures the distance from a random point to the closest event. Ghat and Fhat typically have very similar plots. Envelopes can be added to test against completely spatially random (CSR).

# K function

# The K function is used to look at second order effect over larger range of scales. This can only be used if there is no first order effect in the data and the process is thought to be isotropic, meaning intensity is constant over direction. This might be an issue with using the k function for my data because there is a first order trend. The K function can be thought of as the events within distance h of an arbitrary event.

# Clark-Evans Test

# The Clark-Evans test is a test designed to look at if the point pattern is completely spatially random (CSR). The null hypothesis for this test is that the point pattern is completely spatially random the alternative hypothesis is that there is clustering in the data. From the Clark-Evans test a Z score can be calculated and this will tell how likely one can expect to see a more extreme value than the one that was calculated.

# Correlation and regression

# Correlation is a measure of how similar two variables are. I thought there could be a high positive correlation between income and tree intensity. One way to test this is to plot both variables on opposite axes and then see if there is a correlation. Regression is a mathematical formula that tries to draw a linear line of best fit between the N variables. The smaller the sum of the residual distance the higher the Rsquared or the better relationship.

# Results

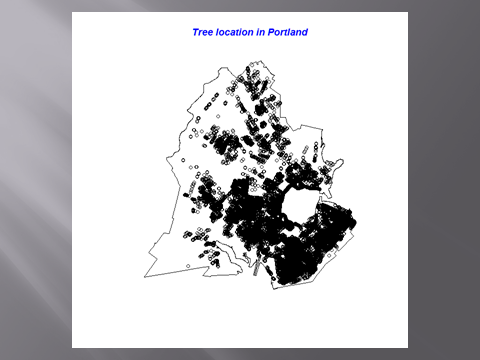


Figure 1: This is a point map of trees in Portland. This is used to visualize the data.

Looking at this map we see that there is a higher intensity of points (trees) in Southern Portland then in Northern Portland. It is hard to see second order effects in this plot. One of the reasons that the intensity is higher in Southern Portland then in Northern Portland is because in Northern Portland homeowners are responsible for their own trees. This is known as set back planting. The tree survey was only interested in city planted trees. As a result the intensity is higher in the south then the north.

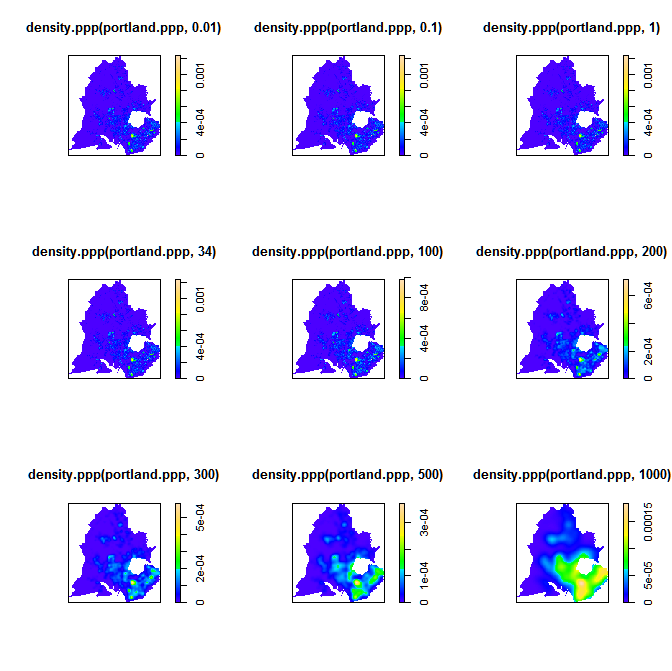


Figure 2: These are plots using kernel density estimation for trees in Portland. We see higher intensity in the south.

This plot shows one of the issues with kernel density estimation, mainly that depending on the bandwidth there can be different results. Figure 2 does show a first order effect though.

|  |  |
| --- | --- |
|  |  |

Figure 3: This is the Clark-Evans test. We also see the mean nearest neighbor distances. I also used bw.diggle to calculate the distance at which the highest amount of clustering occurs. The distance at which there is the highest cluster for trees is 34.14. The distance at which the parcel centroid has the highest clustering is 82.18. This could be an issue later on because I chose a distance of 35 for my optimal bandwidth for kernel density. Looking at the Clark-Evans test we see significant clustering. Having a Z score value of -149.85 means that we can reject the null hypothesis of CSR and accept that the data is clustered. This does not tell us which type of clustering, only that the data is clustered. This is not that surprising when thinking about city planted trees. If there is a city tree it would make sense that there would be another city tree close by.

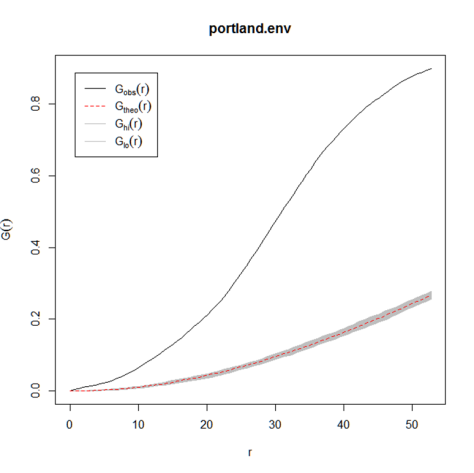


Figure 4: This is Ghat with an envelope to test against CSR. Looking at this plot we can safely say that there is second order effect in the data. This also is in agreement with the Clarke-Evans Test.

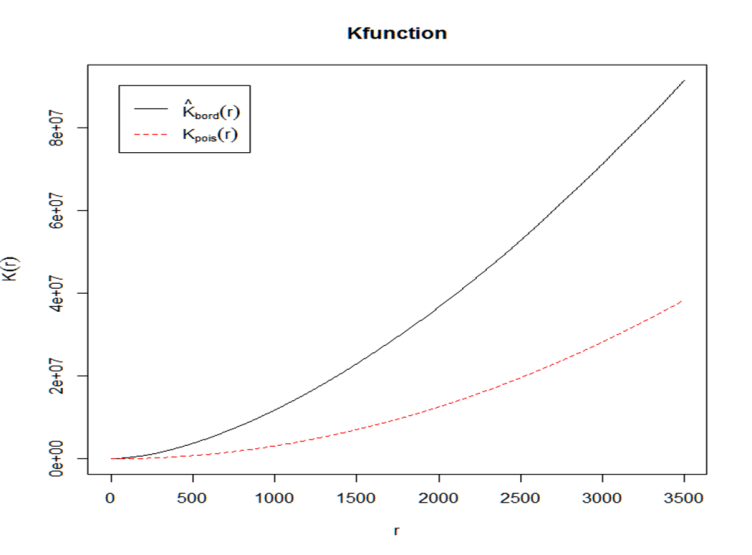


Figure 5: This is the K function for the point pattern. We see that there is second order clustering. One issue with my data is that it is not isotropic so this might be making the K functions higher than it would otherwise be. Clearly though there is a first order effect and a second order effect looking at all of the previous plots.

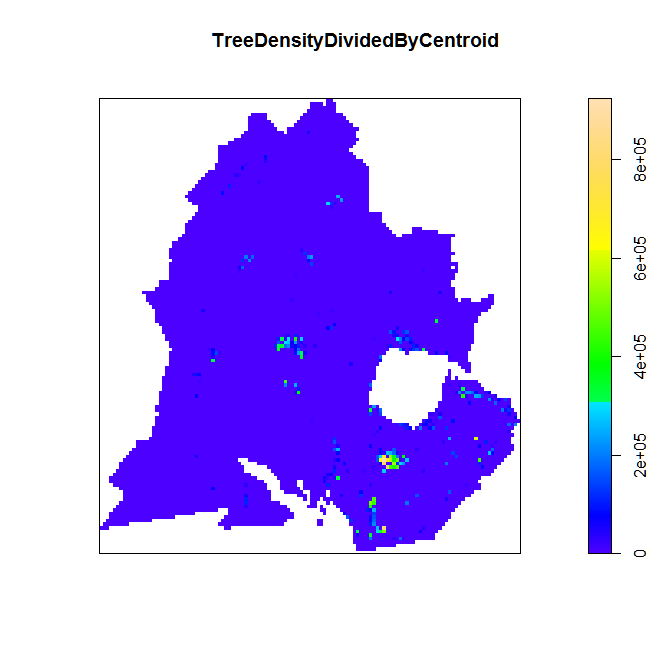


Figure 6: This is a map of highest intensity using kernel density estimation. The tree density image is divided by parcel centroids image to give a comparison. We see Deering Oak Park and Evergreen Cemetery with the highest intensity.

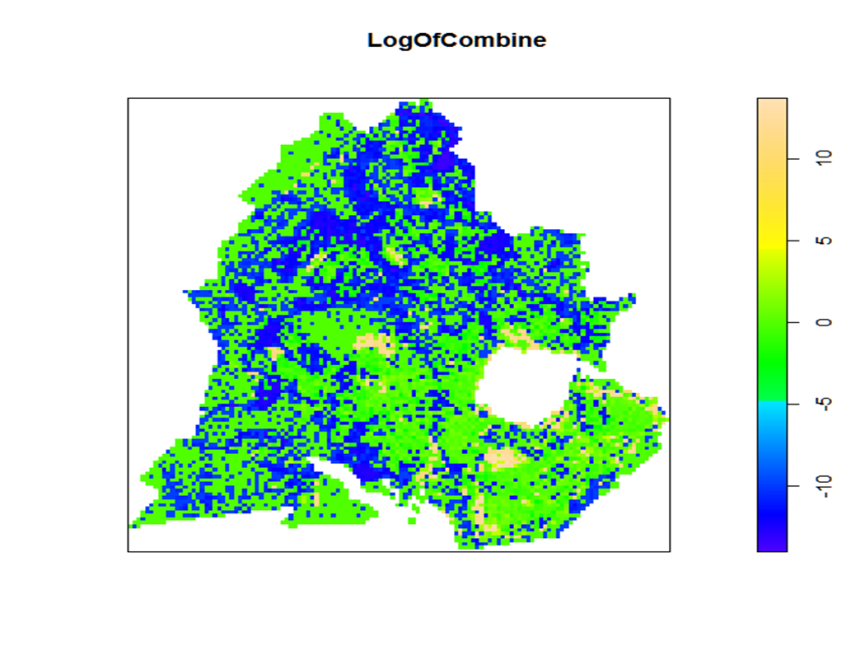


Figure 7: This is the log of the previous plot. This is to centralize the data around 0. Green and blue show where more trees could be planted. Tan depicts where there is high intensity.

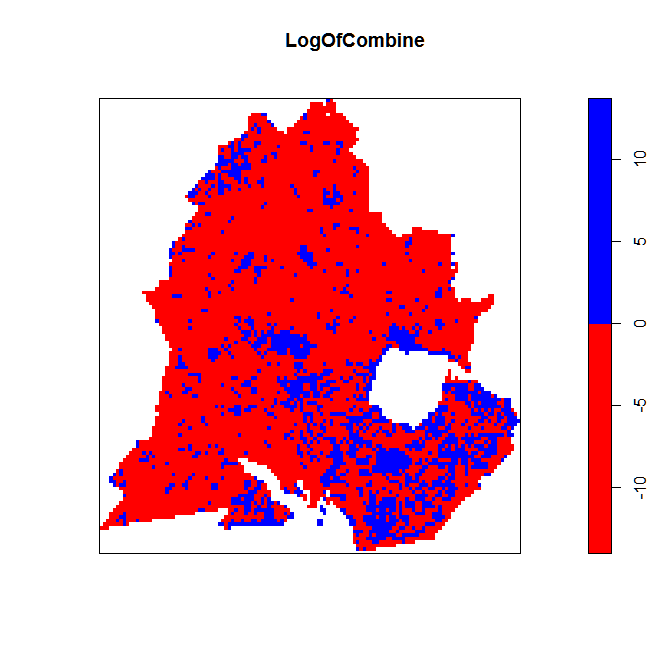


Figure 8: This is a reclassified map to show where more trees can be planted. Blue shows high tree intensity. Red shows where more trees can be planted.

I did the same type of analysis in R and in ArcGIS 10.1. I used the spatial analysis extension. This was done to compare the outputs and to validate my findings.  


Figure 9: This is kernel density estimation using ArcGIS. We see Deering Oak Park and Evergreen Cemetery have the highest intensity again. Light area show where there is lower intensity or less trees.

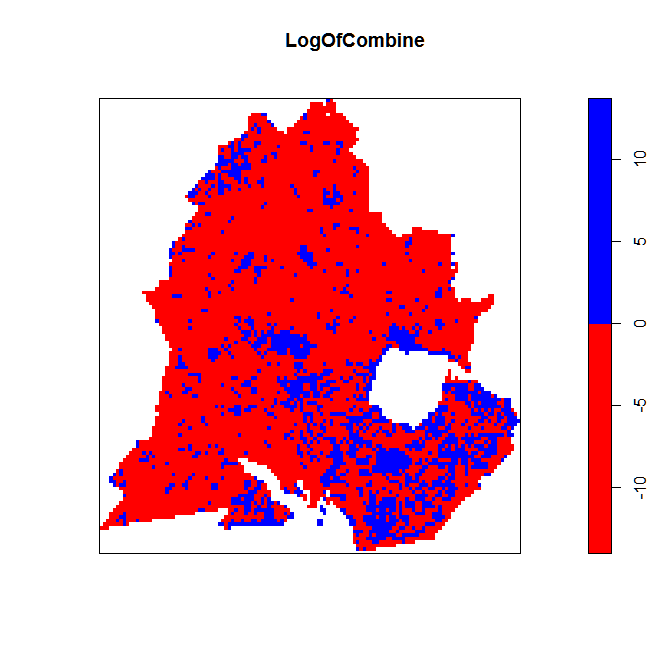


Figure 10: This is the log of the previous plot. This was done to centralize the data around 0. Black in this case is where more trees can be planted. White is where there is high density. It looks like the raster does not go all the way to the boundary of Portland. That is not an error. This is where we see sampling bias. There were not enough points in that region to calculate kernel density.

# Second question

Now we will move on to testing if there is a correlation between income and tree density. I used a slightly different approach than before. I started by visualizing the data.

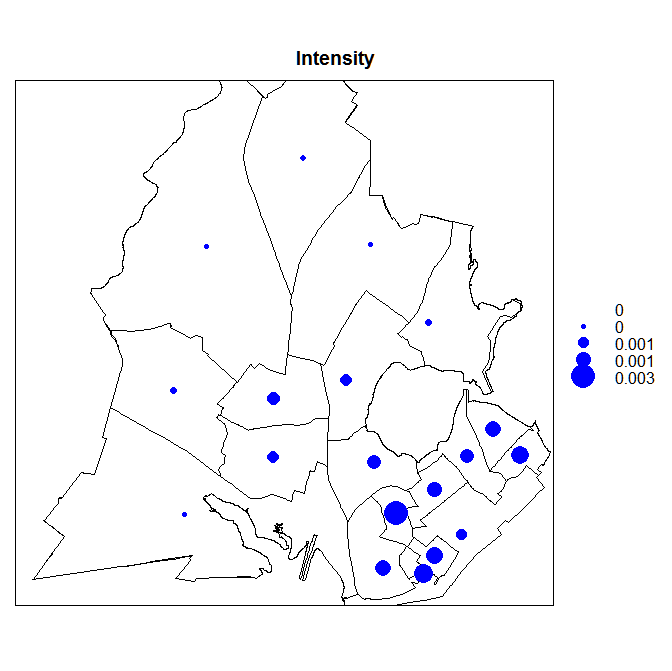


Figure 11: This is a graduated symbol map of tree intensity per census tract. We see higher intensity in Southern Portland then in Northern Portland. This is very similar to the findings previously.

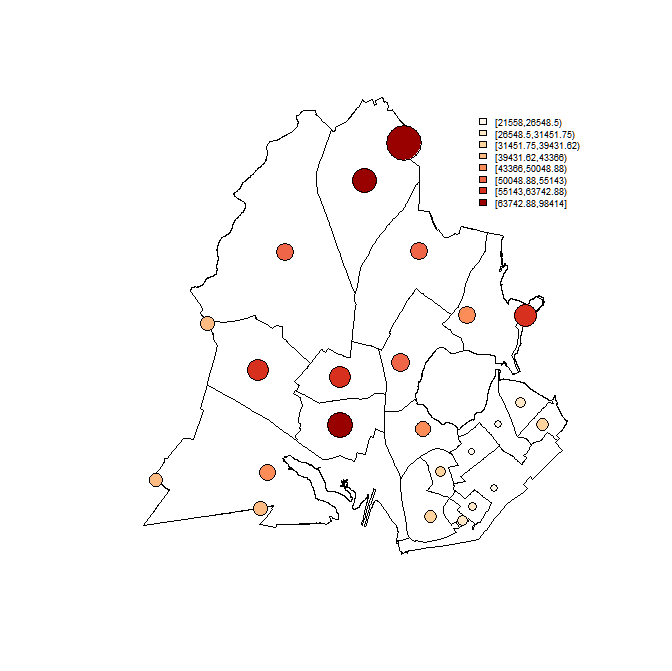


Figure 12: This is a graduated symbol map of income per census tract. We see higher median income in Northern Portland then in Southern Portland. I thought that higher income would be in Southern Portland, but I was incorrect in my assumption.

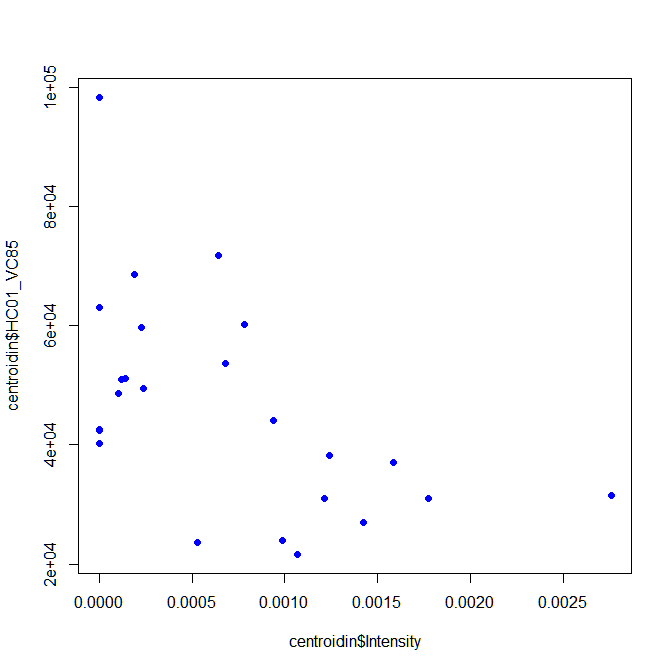
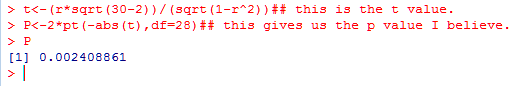


Figure 13: This is a scatter plot between income and intensity. We see a negative relationship. We also see non-constant variance in this plot.

This plot suggests that my hypothesis, that as income goes up so does tree intensity was refuted.



This is the P value for this correlation. Because the P value is less than .05 we reject the null hypothesis in favor of the alternative hypothesis. The null hypothesis is H0: ρ = 0 and the alternative is H1: ρ ≠ 0. Because the P value is less than .05 this is telling us that the correlation is greater than we would expect by chance.

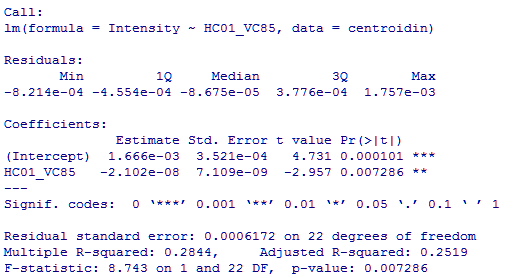


Figure 14: I took this a step farther and wanted to see how well income could predict tree intensity. The R squared is .28. This seems to be a very large variable to explain tree density in Portland. I did predict there would be a positive relationship between income and tree intensity. That turned out to be incorrect; however, it seems that income is still a strong variable to predict tree intensity.

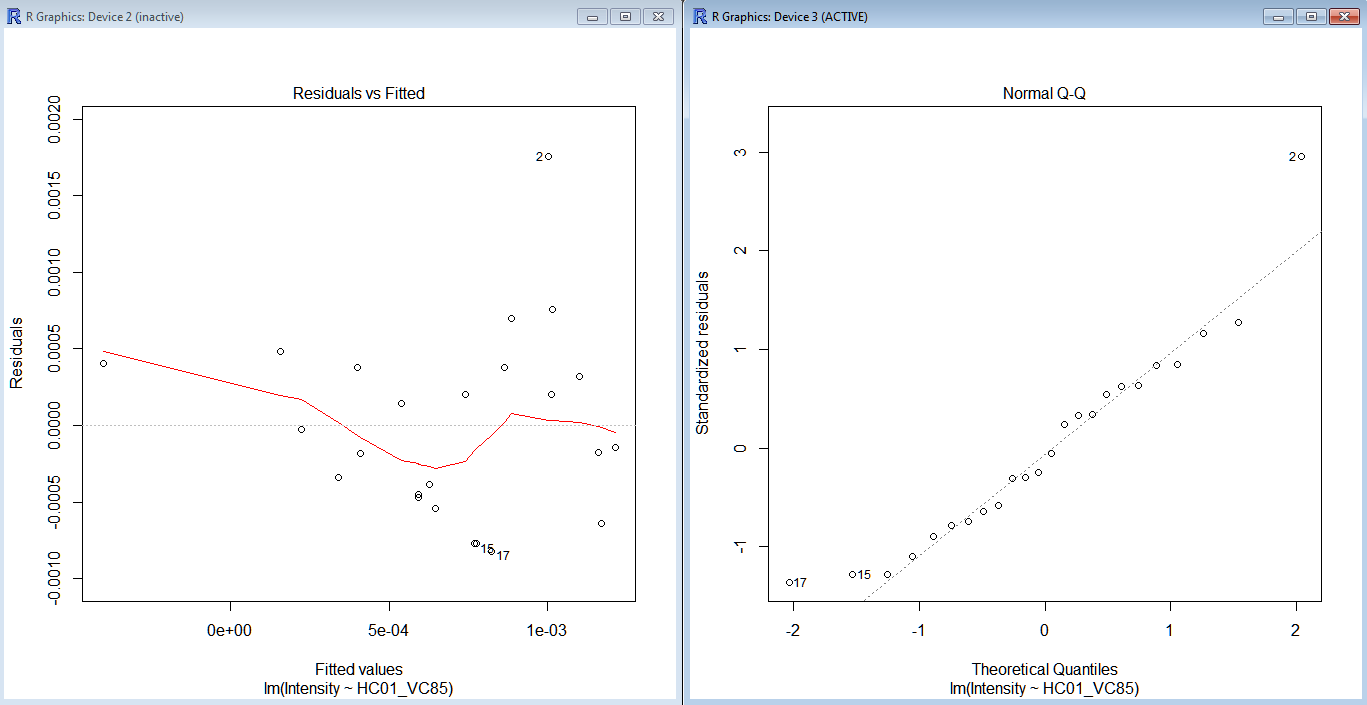


Figure 15: This is a plot showing the residuals vs fitted and a QQ plot.

Looking at these plots we see there is over prediction on the low end of the graph. In the middle of the graph there is under prediction. At the high end of the graph it looks to be constant. There is non-constant variance over the plot however. Looking at the QQ plot we see short and long tails. This suggests positive skew in the data. With income data that is not that surprising. 

Figure 16: This is a plot of the residuals. Darker colors indicate over estimation. Light values represent under estimation.

Discussion

Looking at all the data we see that indeed there is a first and second order effect. More than likely it is a combination of the two. Looking at the results it is possible to map locations that could have more city trees. These results are useful because it can help the City with planting in the future. It was possible to take the statistical techniques we have learned over the semester and apply them to the tree distribution to suggest places were more trees could be planted. My other hypothesis was incorrect and there is a negative correlation between income and tree intensity. Where the blue arrows are located in figure 10 show locations where more planting would be beneficial. This analysis suggests that there are areas that have lower intensity. This investigation also locates areas that the City could plant more trees in.

Future work could include deleting stumps and trees that are marked for removal out of the point pattern. I know that completely spatially random is not an appropriate model to use. Future work could include finding a better model to represent this data. I looked at tree intensity compared to income, but I think another interesting analysis that could be done is tree intensity to human population.

# I would like to take this time and thank Dr. Kate Beard for help with some of the technical issues with this analysis.

# Bibliography

Bailey, T. C., and A. C. Gatrell. Interactive Spatial Data Analysis. Vol. 413. Longman Scientific & Technical Essex, 1995. Print.

Brunsdon, C. “Estimating Probability Surfaces for Geographical Point Data: An Adaptive Kernel Algorithm.” Computers & Geosciences 21.7 (1995): 877–894. Print.

CrimeStat III. “Chapter 8 Kernel Density.” Web. 20 Dec. 2012.

J. G. F. Garnica, D. A. M. Gonzalez, and J. D. B. Solorio. “Geostatistical Evaluation of Natural Tree Regeneration of a Disturbed Forest.” Notes (2006): n. pag. Print.

Law, R. et al. “Ecological Information from Spatial Patterns of Plants: Insights from Point Process Theory.” Journal of Ecology 97.4 (2009): 616–628. Print.

Rowlingson, B. S., and P. J. Diggle. “Splancs: Spatial Point Pattern Analysis Code in S-Plus.” Computers & Geosciences 19.5 (1993): 627–655. Print.Biblogragragphy

# Appenix 1 Tree density

# Luke Kaim

#Final project

# tree density

setwd('C:/Users/Luke Kaim/Documents/University of Maine/Fall\_2012/Spatial Analysis/Project/treesPortlantProper')

library(gstat)

library(classInt)

library(RColorBrewer)

library(lattice)

library(maptools)

library(spatstat)

library(splancs)

library(raster)

centroid<-readShapePoints('centroid.shp',proj4string=CRS("+proj=utm +north +zone=19 +datum=NAD83"))

trees<-readShapePoints('porltandtrees.shp',proj4string=CRS("+proj=utm +north +zone=19 +datum=NAD83"))

portland<-readShapePoly('portlandpoper1.shp',proj4string=CRS("+proj=utm +north + zone=19 +datum=NAD83"))

trees.ppp<- as.ppp(trees)

plot(portland)

points(trees.ppp)

title(main="Tree location in Portland", col.main="Blue", font.main=4)

centroid.ppp<-as.ppp(centroid)

plot(portland)

points(centroid.ppp)

#class(centroid.ppp)

summary(trees.ppp)

#Read in a polygon shape file

#portlandpoly=readShapePoly("PortlandPoly.shp")

#Get the polygon cords: and get a length - 1 value

xy<-portland@polygons[[1]]@Polygons[[1]]@coords

n <- length(xy[,1]) - 1

x <- rev(xy[,1][1:n])

y <- rev(xy[,2][1:n])

p.list <- list(x=x,y=y)

W <- owin(poly=p.list)

portland.ppp <- ppp(trees$POINT\_X, trees$POINT\_Y, window=W)

plot(portland.ppp)

cent.ppp<- ppp(centroid$POINT\_X, centroid$POINT\_Y, window=W)

summary(portland.ppp)

par(mfrow=c(3,3))

plot(density.ppp(portland.ppp, .01))

plot(density.ppp(portland.ppp, .1))

plot(density.ppp(portland.ppp, 1))

plot(density.ppp(portland.ppp, 34))

plot(density.ppp(portland.ppp, 100))

plot(density.ppp(portland.ppp, 200))

plot(density.ppp(portland.ppp, 300))

plot(density.ppp(portland.ppp, 500))

plot(density.ppp(portland.ppp, 1000))

dev.new()

TreeDensity<-(density.ppp(portland.ppp, 35))

plot(TreeDensity)

summary(TreeDensity)

dev.new()

CentroidDensity<-(density.ppp(cent.ppp, 35))

plot(CentroidDensity)

summary(CentroidDensity)

dev.new()

ScaleT<-eval.im(TreeDensity\*1000000)

GetRidOfZeroT<-eval.im(ScaleT+.001)

plot(GetRidOfZeroT)

ScaleC<-eval.im(CentroidDensity\*1000000)

plot(ScaleC)

GetRidOfZeroC<-eval.im(ScaleC+.001)

plot(GetRidOfZeroC)

TreeDensityDividedByCentroid<-eval.im(GetRidOfZeroT/GetRidOfZeroC)

LogOfCombine<-eval.im(log(TreeDensityDividedByCentroid))

plot(TreeDensityDividedByCentroid)

plot(LogOfCombine)

summary(LogOfCombine)

colormap<-colourmap(c("red","blue"), breaks=c(-14.0256180083768,0,13.7379362070014))

plot(LogOfCombine,col=colormap)

treesprop.nndist<- nndist(portland.ppp)

treesprop.nndist

summary(treesprop.nndist)

var<-var(treesprop.nndist)

var

#Square feet

AREAFT<-area.owin(W)

AREAFT

points<-18187

AveragedIntensity<-18187/AREAFT

mean<-35.16

EW<- 1/(2\*sqrt(AveragedIntensity))

EW

varW<-(4-pi)/(4\*points\*AveragedIntensity\*pi)

varW

z<-(mean-EW)/sqrt(varW)

z

dev.new()

portland.env<-envelope(portland.ppp,fun=Gest, nsim=99)

portland.env

plot(portland.env)

help(envelope)

dev.new()

Kfunction<-Kest(trees.ppp)

plot(Kfunction)

A<-bw.diggle(trees.ppp)

A

B<-bw.diggle(centroid.ppp)

B

C<-bw.diggle(portland.ppp)

C

# Appendix 2 Income

#Income

setwd('C:/Users/Luke Kaim/Documents/University of Maine/Fall\_2012/Spatial Analysis/Project/income')

library(gstat)

library(classInt)

library(RColorBrewer)

library(lattice)

library(maptools)

library(spatstat)

library(splancs)

library(raster)

Tract<-readShapePoly('TractJoinJoin.shp',proj4string=CRS("+proj=utm +north + zone=19 +datum=NAD83"))

centroidin<-readShapePoints('centroidin.shp',proj4string=CRS("+proj=utm +north + zone=19 +datum=NAD83"))

plot(centroidin)

summary(Tract)

plot(Tract)

bubble(centroidin,'Intensity',col='blue',pch=16,sp.layout=list(sp.polygons, Tract, col='black'))

plotvar <- centroidin$Intensity

nclr=8 #sets number of class levels

plotclr <- brewer.pal(nclr,"OrRd")# sets a color palette to orange-red with 8 levels

class <-classIntervals(plotvar, nclr, style ="quantile") #sets class intervals as quantiles

colcode <- findColours(class, plotclr) # links color palette with class intervals

max.symbol.size=5

min.symbol.size=1

symbol.size <- ((plotvar-min(plotvar))/ (max(plotvar)-min(plotvar))\*(max.symbol.size-min.symbol.size)+min.symbol.size)

plot(Tract) #plots boundary of california

points(centroidin, pch=16, col=colcode, cex=symbol.size)

points(centroidin, cex=symbol.size)

legend(locator(1), legend=names(attr(colcode, "table")), fill=attr(colcode, "palette"), cex=0.6, bty="n")

plotvar <- centroidin$HC01\_VC85

nclr=8 #sets number of class levels

plotclr <- brewer.pal(nclr,"OrRd")# sets a color palette to orange-red with 8 levels

class <-classIntervals(plotvar, nclr, style ="quantile") #sets class intervals as quantiles

colcode <- findColours(class, plotclr) # links color palette with class intervals

max.symbol.size=5

min.symbol.size=1

symbol.size <- ((plotvar-min(plotvar))/ (max(plotvar)-min(plotvar))\*(max.symbol.size-min.symbol.size)+min.symbol.size)

plot(Tract) #plots boundary of california

points(centroidin, pch=16, col=colcode, cex=symbol.size)

points(centroidin, cex=symbol.size)

legend(locator(1), legend=names(attr(colcode, "table")), fill=attr(colcode, "palette"), cex=0.6, bty="n")

plot(centroidin$Intensity, centroidin$HC01\_VC85, pch=16, col=4)

r<-cor(Tract$HC01\_VC85, Tract$Intensity)

r

t<-(r\*sqrt(30-2))/(sqrt(1-r^2))## this is the t value.

P<-2\*pt(-abs(t),df=28)## this gives us the p value I believe.

P

lm.precip<-lm(Intensity~HC01\_VC85 , data=centroidin)

summary(lm.precip)

plot.lm(lm.precip, which=1)

plot.lm(lm.precip, which=2)

resids <-lm.precip$res

newtest<-spCbind(centroidin, resids)

plotvar <-newtest$resids

nclr <- 8 plotclr <- brewer.pal(nclr,"RdBu")

class <-classIntervals(plotvar, nclr, style ="sd")

colcode <- findColours(class, plotclr)

plot(Tract)

points(newtest, pch=16, col=colcode, cex=2)

legend(locator(1), legend=names(attr(colcode, "table")),fill=attr(colcode, "palette"), cex=0.6, bty="n")

plotclr <- plotclr[nclr:1] # reorder colors