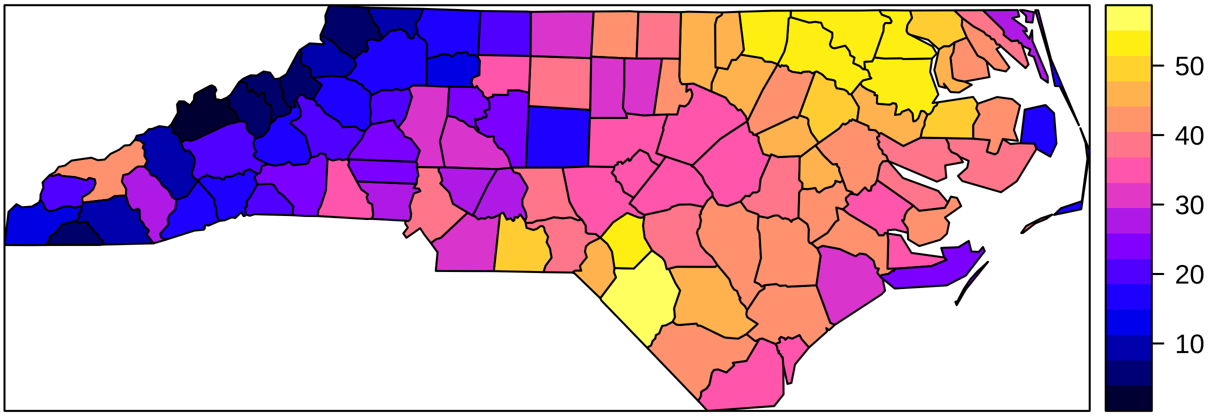
Quan_Tech_Areal_Data_Exam

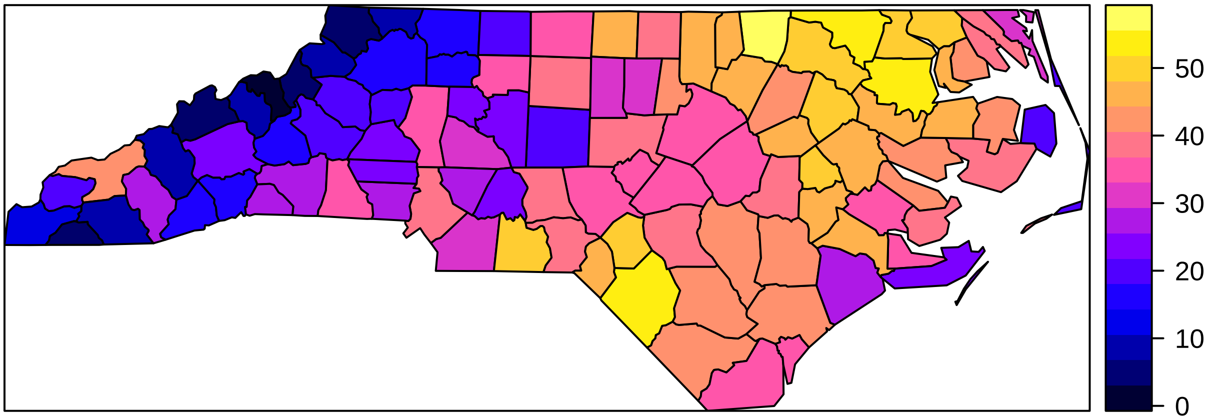
Auriel Fournier

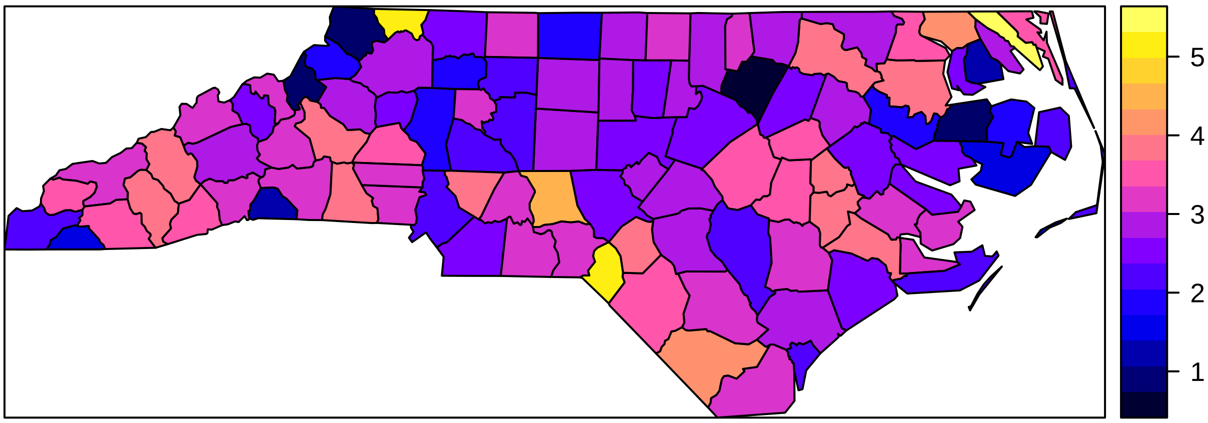
Monday, March 16, 2015

```
library(ggplot2)
library(maptools)
library(spdep)
library(xtable)
library(sp)
```

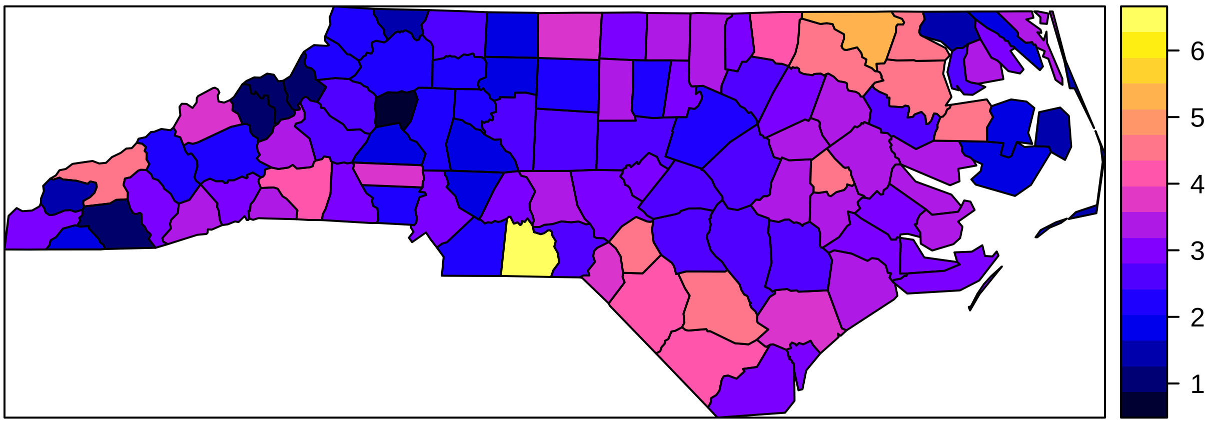
```
nc.sids <- readShapePoly(system.file("etc/shapes/sids.shp", package="spdep")[1], ID="FIPSNO", proj4strin
rn <- sapply(slot(nc.sids, "polygons"), function(x) slot(x, "ID"))
ncCC89_nb <- read.gal(system.file("etc/weights/ncCC89.gal", package="spdep")[1], region.id=rn)
ncCC85_nb <- read.gal(system.file("etc/weights/ncCR85.gal", package="spdep")[1], region.id=rn)

nc.sids$sid74ft = sqrt(1000)*(sqrt(nc.sids$SID79/nc.sids$BIR79) +sqrt((nc.sids$SID79+1)/nc.sids$BIR79))

nc.sids$nwbir74ft = sqrt(1000)*(sqrt(nc.sids$NWBIR79/nc.sids$BIR79)+sqrt((nc.sids$NWBIR79+1)/nc.sids$BII

nc.sids$sid79ft = sqrt(1000)*(sqrt(nc.sids$SID74/nc.sids$BIR74) +sqrt((nc.sids$SID74+1)/nc.sids$BIR74))

nc.sids$nwbir79ft = sqrt(1000)*(sqrt(nc.sids$NWBIR74/nc.sids$BIR74)+sqrt((nc.sids$NWBIR74+1)/nc.sids$BII
```




0
10
20
30
40
50

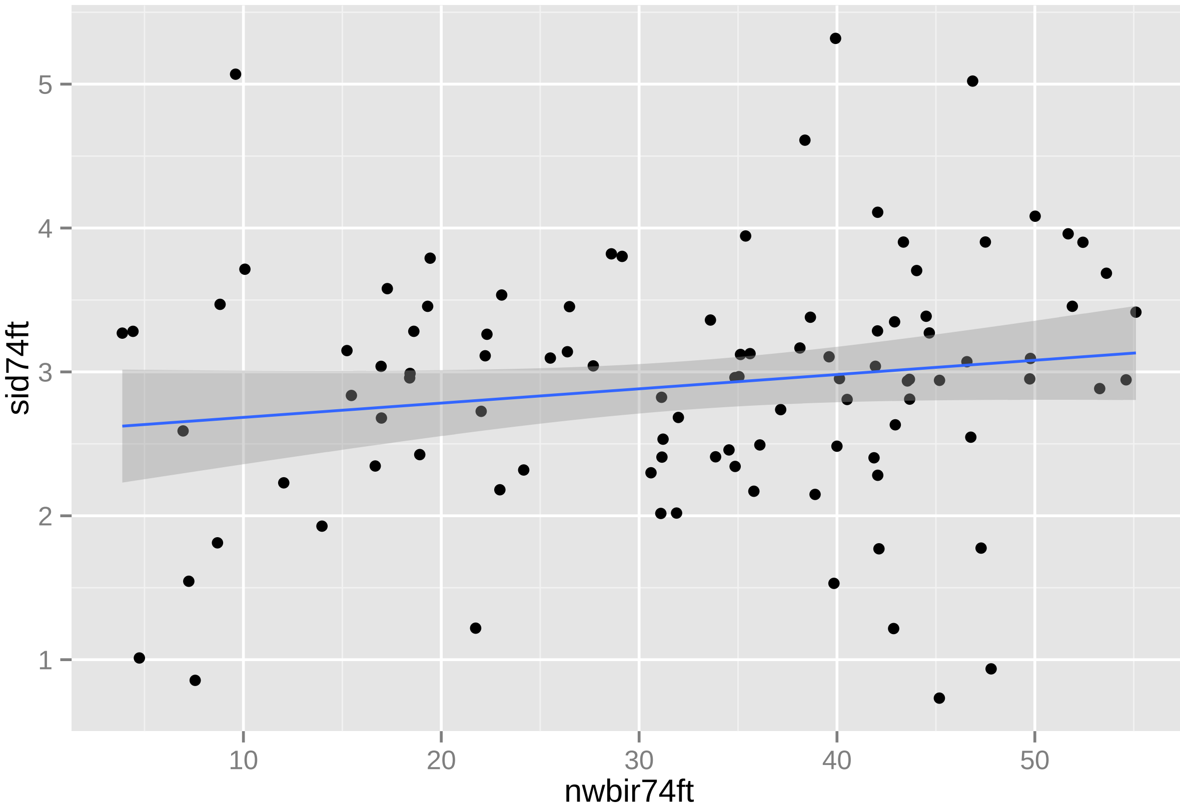


1
2
3
4
5

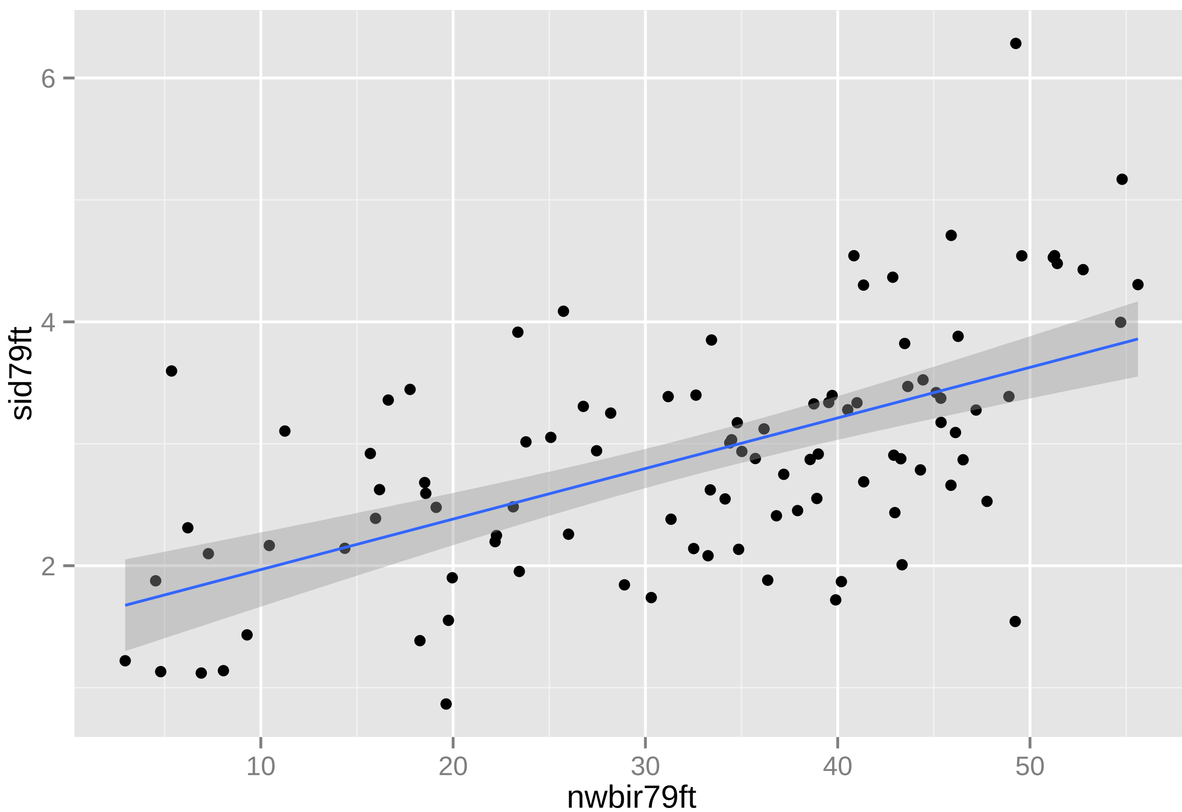


First I'd like to examine any relationship between these two variables without taking into account spatial variation. I'll do this by running a linear model of non-white births vs sids.

```
##
## Call:
## lm(formula = nwbir74ft ~ sid74ft, data = nc.sids)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.547 -11.057   2.296  10.874  22.012
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.123      4.833   5.198  1.1e-06 ***
## sid74ft        2.539      1.595   1.592    0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.62 on 98 degrees of freedom
## Multiple R-squared:  0.02521,    Adjusted R-squared:  0.01526
## F-statistic: 2.534 on 1 and 98 DF,  p-value: 0.1146
```




```
##
## Call:
## lm(formula = nwbir79ft ~ sid79ft, data = nc.sids)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.975  -8.848   3.054   7.730  27.899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.564      3.543   2.417   0.0175 *
## sid79ft        8.275      1.157   7.155 1.53e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.23 on 98 degrees of freedom
## Multiple R-squared:  0.3431, Adjusted R-squared:  0.3364
## F-statistic: 51.19 on 1 and 98 DF,  p-value: 1.533e-10
```



So in 1974 there is a significant relationship between the two, but not in 1979. In both cases the R is low (1974 it is .3431) suggesting that even when it is significant there might be other factors at play.

So lets start to examine things spatially.

```
col.qb <- nb2listw(poly2nb(nc.sids, queen=T),style="B")
col.rw <- nb2listw(poly2nb(nc.sids, queen=F),style="W")
col.rb <- nb2listw(poly2nb(nc.sids, queen=F),style="B")
coords <- coordinates(nc.sids)
nc.d2 <- nb2listw(dnearneigh(coords, d1=0, d=2))
nc.d1 <- nb2listw(dnearneigh(coords, d1=0, d=1))
knn1 <- nb2listw(knn2nb(knearneigh(coords, k=1)))
knn2 <- nb2listw(knn2nb(knearneigh(coords, k=2)))
knn3 <- nb2listw(knn2nb(knearneigh(coords, k=3)))
knn4 <- nb2listw(knn2nb(knearneigh(coords, k=4)))
knn5 <- nb2listw(knn2nb(knearneigh(coords, k=5)))
knn6 <- nb2listw(knn2nb(knearneigh(coords, k=6)))
knn7 <- nb2listw(knn2nb(knearneigh(coords, k=7)))
knn8 <- nb2listw(knn2nb(knearneigh(coords, k=8)))
knn9 <- nb2listw(knn2nb(knearneigh(coords, k=9)))
knn10 <- nb2listw(knn2nb(knearneigh(coords, k=10)))
knn20 <- nb2listw(knn2nb(knearneigh(coords, k=20)))
knn30 <- nb2listw(knn2nb(knearneigh(coords, k=30)))
```

1974

```
##      Names Observed Moran's I  Expectation     Variance
## 4   nc.d2          0.5265788 -0.010020618 0.0003137188
## 17  knn30          0.5627748 -0.009980457 0.0004085248
## 16  knn20          0.6179374 -0.010045452 0.0007102911
## 5   nd.d1          0.6343964 -0.010062152 0.0011981431
## 1  col.qb          0.6745137 -0.011239995 0.0038299674
## 3  col.rb          0.6815243 -0.011507868 0.0040892687
## 15  knn10          0.6841310 -0.010328769 0.0016391118
## 14   knn9          0.6944897 -0.010445549 0.0018442993
## 10   knn5          0.7007027 -0.011230610 0.0034341086
## 11   knn6          0.7037435 -0.010891495 0.0028563442
## 13   knn8          0.7043062 -0.010546185 0.0020993996
## 12   knn7          0.7097198 -0.010886501 0.0024147708
## 9    knn4          0.7263582 -0.010964629 0.0043598222
## 8    knn3          0.7403804 -0.011312599 0.0059041044
## 6    knn1          0.7416131 -0.012597687 0.0148741947
## 2  col.rw          0.7421808 -0.011725012 0.0044682526
## 7    knn2          0.7702663 -0.012006722 0.0085670587
```

The 1974 Moran's Is are from .42 to .55, which is pretty uninformative since they are all so similar. They are all positive, which means some clumping. The rook contiguity row normalized model and 2 nearest neighbors are the most clustered.

1979

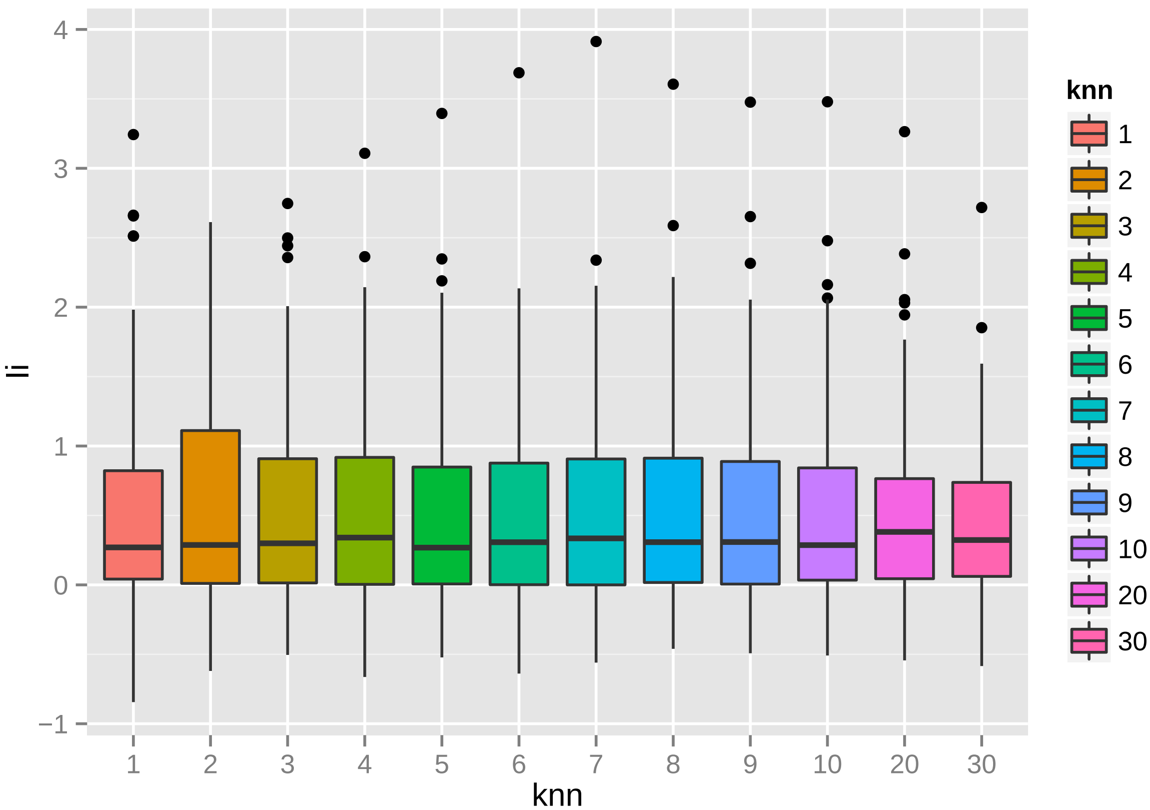
```
##      Names Observed Moran's I Expectation     Variance
## 4   nc.d2          0.4295463 -0.01102507 0.0002763484
## 17  knn30          0.4346759 -0.01138426 0.0003611534
## 6    knn1          0.4781801 -0.01451673 0.0148945768
## 3  col.rb          0.4934032 -0.01273419 0.0040261304
## 16  knn20          0.4948162 -0.01173555 0.0006619085
## 1  col.qb          0.4955004 -0.01246842 0.0037833654
## 5   nd.d1          0.5104433 -0.01191525 0.0011523389
## 9    knn4          0.5212247 -0.01267656 0.0043188931
## 11   knn6          0.5212701 -0.01239808 0.0028066664
## 10   knn5          0.5237032 -0.01233118 0.0033861981
## 15  knn10          0.5261396 -0.01204413 0.0015861664
## 14   knn9          0.5264196 -0.01205059 0.0017925279
## 12   knn7          0.5328092 -0.01224801 0.0023609457
## 13   knn8          0.5345926 -0.01229564 0.0020478836
## 8    knn3          0.5350312 -0.01272175 0.0058789835
## 7    knn2          0.5385829 -0.01340789 0.0085426953
## 2  col.rw          0.5538677 -0.01276221 0.0044131882
```

The 1979 Moran's Is are from .52 to .77, which is more informative then the 1974 data since they are all closer to 1. They are all positive, which means some clumping. The same top models (those with the highest Moran's I) are here as there were for the 1974 dataset.

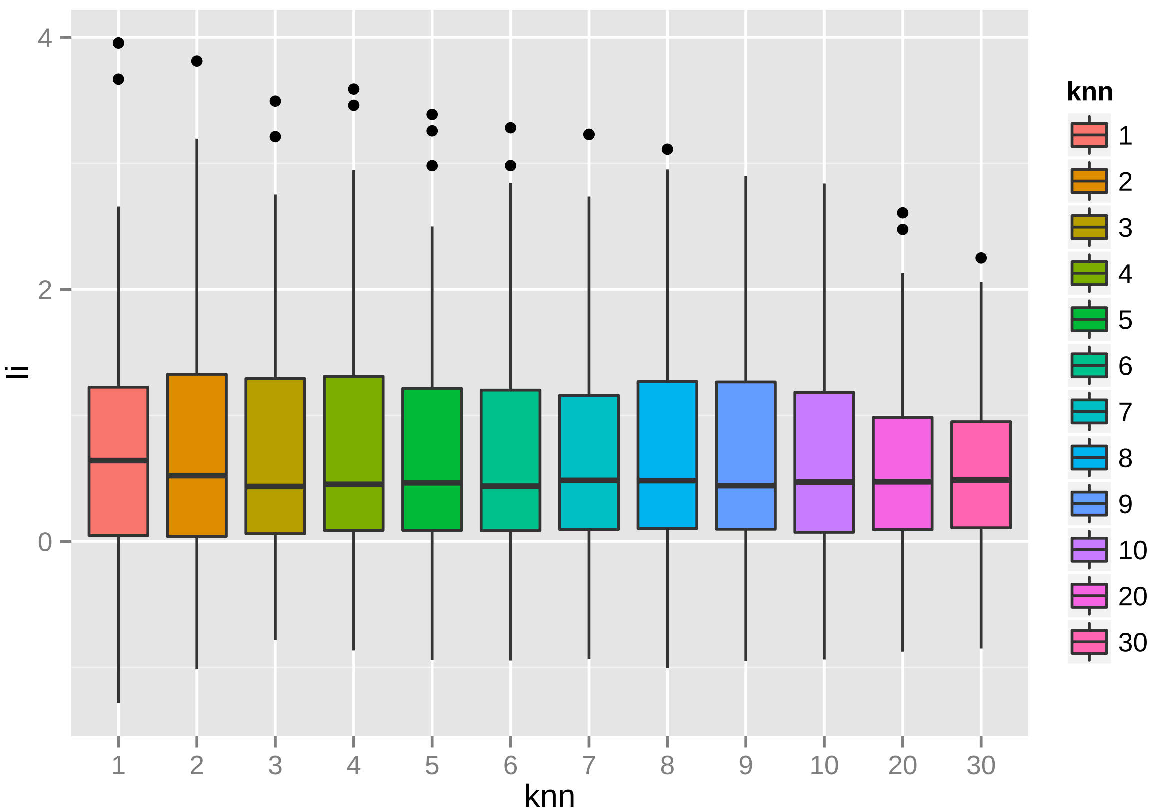
This suggests that the clustering is happening at the small in both situations.

I made the following two graphs after reading section 7.5 in O'Sullivan and Unwin (2010) and seeing Figure 7.10 which showed a graphical way of seeing how many neighbors are an appropriate scale. I thought this might help in my own analysis but as you can see instead of drop off after 3 neighbors as is seen in O'Sullivan and Unwin (2010) it is straight line, with a fairly high spread of values and several outliers. So this was not informative for the number of neighbors to cut it off at. The fact that the values are roughly the same at each number of neighbors suggests that there may not be clumping, unless the clumping is larger then 10. Originally I picked 1-10 numbers and there was no difference, so then I added 20 and 30 to see if the trend continued, and as you can see, it does. though data is tighter. I feel that cutting off the number of neighbors at 30 is appropriate since that is ~30% of the data.

1979



1974



Discussion

Bibliography

O'Sullivan, David., and David J. Unwin. 2010. *Geographic Information Analysis*. Wiley.