**Analysis of Spatial Clustering of   
Diabetes-Related Hospitalizations in Colorado**

Adelle Price

University of Colorado Denver

MATH 6384

**Introduction**

Approximately 6.9% of adults currently living within Colorado are diagnosed with diabetes.5 It is estimated that diagnosed diabetes costs around $3.6 billion to treat in Colorado each year.5 To aide in the goal of reducing the number of adults living with diabetes and the associated costs, this study aims to analyze spatial clustering of diabetes-related hospitalizations within census tracts of Colorado. Identifying these clusters – if they exist – will allow more precise allocation of the resources contributing to diabetes prevention.

**Data**

All data used in this project was sourced from the Colorado Department of Public Health and Environment(CDPHE) via their publicly available open data catalog. Geographic coordinates denoting the border of each census tract was provided in each dataset.

*Diabetes Hospitalization Rate (Census Tracts) data2*

This dataset contained the Age-Adjusted Colorado Census Tract Rate of Diabetes-Related Hospital Discharges (2013-2017) and Inpatient Hospitalizations per 100,000 persons1. The rates were calculated using the billing addresses of discharged individuals with the 2013-2017 population estimates from the American Community Survey. The data is a subset of the Colorado Hospital Association’s Hospital Discharge Dataset which is published annually by the CDPHE. This data contained 1,249 census tract level observations. There were 11 observations removed due to missing data.

*Population Density (Census Tracts) data4*

This dataset contained select census tract level population density estimates within Colorado from the 2013-2017 American Community Survey1. This dataset contained 1,242 census tract level population density observations.

The two datasets were merged by Census Tract FIPS (Federal Information Processing System) Codes. The merged dataset contained 1,238 census tract level observations and all were used for analysis. The number of cases per census tract was calculated by multiplying the reported rate of diabetes-related hospitalizations by the reported population density in each given tract.

**Methods**

Geographic coordinates denoting a point within each census tract (the centroid) were calculated using the *sf* package (written in R) and were used for all statistical inference. The following regional count data methods were used detect clusters or clustering of diabetes-related hospitalizations within 1,238 Colorado census tracts:

*Turnbull et al.’s Cluster Evaluation Permutation Procedure1*

This method creates windows from each census tract’s centroid with a fixed number of persons at risk, *n\**. To account for multiple comparisons, the test statistic is taken as the maximum number of cases observed across all windows in the study area (Colorado). The statistical significance of this test statistic is evaluated by simulating regional count data under the Constant Risk Hypothesis(CRH) and calculating the corresponding Monte Carlo p-value.

For this method, tests using *n\** = 30,000, *n\** = 150,000, and *n\* =* 27,1084.4were used for analysis.

*Spatial Scan Statistic for Heterogeneous Poisson Processes1*

This method evaluates radii (and the cases and population of each census tract if the respective census tract’s centroid is contained within the circle) from distances ranging from the intercentroid distance to a user-defined upper bound (50% of the width of the study area in this analysis) and calculates the test statistic with the following formula:

Diagram

Description automatically generated,

where denotes window , with . is the number of cases in the window, is the expected number of cases in the window, is the number of cases outside the window, is the expected number of cases outside the window.

The expected number of cases inside each window was calculated with the following formula:

To account for multiple comparisons, the maximum test statistic, , across all windows is used for inference. The statistical significance of this test statistic is evaluated by simulating regional count data under the Constant Risk Hypothesis(CRH) and calculating the corresponding Monte Carlo p-value.

For analysis with this method, population upper bounds of 10% and 50% were used.

*Tango’s Index1*

This method compares the proportions of cases in each census tract to the proportions expected under the Constant Risk Hypothesis. The following formula is used to calculate Tango’s Index:

A white paper with black writing

Description automatically generated with low confidence

where denotes spatially defined weights indicating how close regions and are, is the number of cases in a given census tract divided by the total number of cases in the study area, is the number of persons in a given census tract divided by the total number of persons in the study area, and . Note that ; where is a dependence scaling parameter and is the distance between census tract centroids.

In this particular version of Tango’s Index, the index is shown as a sum of two parts with the first measuring the goodness-of-fit in each census tract (clustering within census tracts), and the second measuring spatial similarity between two regions (clustering between census tracts).

To assess the statistical significance of the observed Tango’s Index, simulations are completed with the total number of cases per census tract fixed at and the probability of each case falling in a respective census tract equal to for . A Monte Carlo approach is then applied to calculate the p-value of the observed Tango’s Index.

Analysis using Tango’s Index was completed with equal to the minimum intercentroid distance (394.9 units) as well as equal to the maximum intercentroid distance (55137.9 units).

**Data Exploration**

Chart

Description automatically generatedA)

Figure 1 Legend:

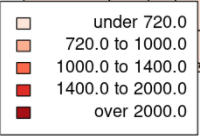


Figure 1:

**Choropleth maps showing the Age-Adjusted Colorado Census Tract Rate of Diabetes-Related Hospitalizations (per 100,000 persons) in Colorado census tracts.** A) Choropleth map of the rate within census tracts for entire state of Colorado. High rates of diabetes-related hospitalizations are shown along the midline of the map and on the right side of the map. B) Choropleth map of the rate within census tracts for the Denver metro area (zoom-in view of A). This map shows high rates of diabetes-related hospitalizations concentrated within the right side of the metro area, just to the right of the metro area, and several high rate tracts just above the metro area. Importantly, a large census tract with missing data is shown in white. C) Choropleth map showing the rate within Colorado Springs and Pueblo (zoom-in view of A). High rates of diabetes-related hospitalizations are seen in almost all census tracts within Pueblo, just to the right of Pueblo, concentrated in the lower area of Colorado Springs, and just to the right of Colorado Springs. There is also a census tract with missing data shown in white.

Map

Description automatically generatedB)

A picture containing diagram

Description automatically generatedC)

To ensure each census tract’s centroid was calculated to be within its corresponding census tract, region centroid’s were plotted at their geographic coordinates and superimposed on a map of the 1,238 Colorado census tracts.

Chart, scatter chart

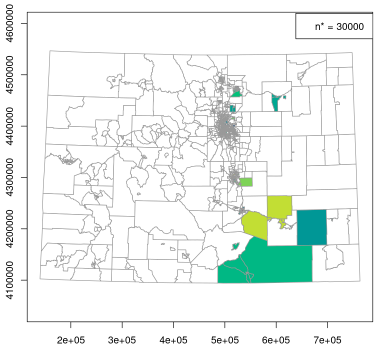
Description automatically generated

Figure 2:

**Census tracts within Colorado plotted in gray with their respective centroid shown as a red dot.** The large census tracts clearly display a single centroid in this map, however the smaller census tracts and the location of their centroids are not discernable in this figure. Due to time constraints, the calculated position of each centroid within the smaller regions not seen here was assumed to be accurate.

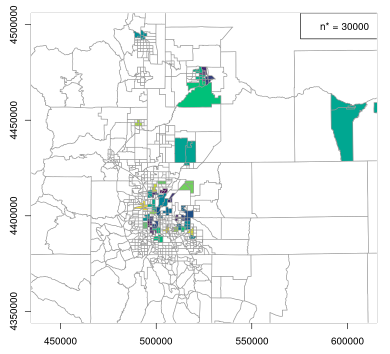
**Results**

Figure 3: **Plotted results of Turnbull et al.’s Cluster Evaluation Permutation Procedure with n\* = 30,000. This analysis produced 41 statistically significant clusters with a p-value .005.** A) A map of census tracts within the entire state of Colorado with statistically significant clusters shown in color (1 color indicates one cluster). B) Census tracts within the Denver metro area with statistically significant clusters shown in color (zoom-in view of A). C) Census tracts within the Colorado Springs and Pueblo area with statistically significant clusters shown in color (zoom-in view of A).

**A)**

Interpretation of A):

There are several clusters seen in the lower right of Colorado. One cluster in blue containing only one census tracts, and two others in dark and light green containing multiple clusters.



**B)**

Interpretation of B:

There are numerous clusters scattered primarily on the right hand side of the Denver metro area, one cluster seen in Fort Collins, and several clusters seen between the Denver metro area and Fort Collins.

Diagram, map

Description automatically generated

**C)**

Interpretation of C:

There are three clusters seen within Pueblo, multiple clusters seen primarily in the lower part of Colorado Springs, and one large green cluster directly to the right of Colorado Springs.

Figure 4: **Plotted results of Turnbull et al.’s Cluster Evaluation Permutation Procedure with n\* = 150,000. This analysis produced 6 statistically significant clusters with a p-value .005.** A) A map of census tracts within the entire state of Colorado with statistically significant clusters shown in color (1 color indicates one cluster). B) Census tracts within the Denver metro area with statistically significant clusters shown in color (zoom-in view of A). C) Census tracts within the Colorado Springs and Pueblo area with statistically significant clusters shown in color (zoom-in view of A).

Diagram

Description automatically generated

**A)**

Interpretation of A):

There is one large cluster shown in dark blue (containing Pueblo) within the lower right area of Colorado. There is also one large cluster shown in green just above the Denver metro area.

Map

Description automatically generated**B)**

Interpretation of B):

There are four clusters seen within the Denver metro area; two on the right side of the metro area in blue and yellow, and one in the upper right of the metro area in dark purple, as well as one in the center of the metro area in light blue.

Map

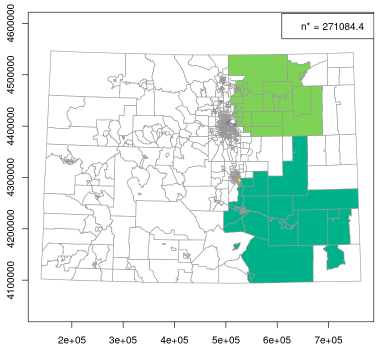
Description automatically generated with low confidence

**C)**

Interpretation of C):

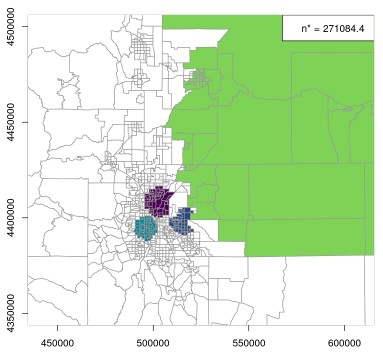
There is one large cluster seen in the lower right side of Colorado Springs.

Figure 5: **Plotted results of Turnbull et al.’s Cluster Evaluation Permutation Procedure with n\* = 271084.4 (Here n\* is equal to 5% of the total population in the study area). This analysis produced 6 statistically significant clusters with a p-value .005.** A) A map of census tracts within the entire state of Colorado with statistically significant clusters shown in color (1 color indicates one cluster). B) Census tracts within the Denver metro area with statistically significant clusters shown in color (zoom-in view of A). C) Census tracts within the Colorado Springs and Pueblo area with statistically significant clusters shown in color (zoom-in view of A).

**A)**

Interpretation of C):

There are two significantly large clusters observed: one in the lower right side of Colorado and the second in the upper right side of Colorado.



**B)**

Interpretation of C):

There are three distinct clusters seen in the Denver metro area. One cluster in the upper middle shown in purple, a second shown in dark blue in the right side of the Denver metro area, and third shown in light blue in the center of the Denver metro area.

Map

Description automatically generated

**C)**

Interpretation of C):

There is one large cluster seen in the lower right side of Colorado Springs.

Figure 6: **Plotted results of the Spatial Scan Statistic for heterogeneous Poisson processes with a population upper bound of 10%. This analysis produced 32 statistically significant clusters with a p-value .005.** A) A map of census tracts within the entire state of Colorado with statistically significant clusters shown in color (1 color indicates one cluster). B) Census tracts within the Denver metro area with statistically significant clusters shown in color (zoom-in view of A). C) Census tracts within the Colorado Springs and Pueblo area with statistically significant clusters shown in color (zoom-in view of A).

1. Diagram

   Description automatically generated

Interpretation of A):

There is one large cluster of census tracts seen in dark blue in the lower right area of Colorado as well as one smaller cluster seen in light blue with several census tracts in the bottom middle of Colorado.

**Diagram, map

Description automatically generatedB)**

Interpretation of B):

There is one large cluster shown in dark purple covering the right side of the Denver metro area, along with multiple small clusters scattered around the large purple cluster. There are six clusters scattered above the Denver metro area that contain more rural census tracts.

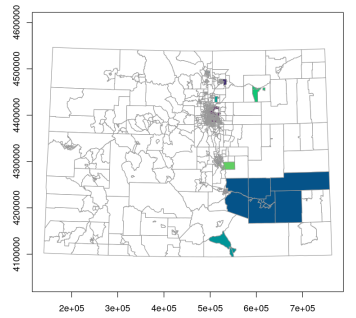
**Map

Description automatically generatedC)**

Interpretation of C):

Pueblo is seen as one large cluster in dark purple. There is one large dark blue cluster seen in the lower right side of Colorado Springs. Three smaller clusters are seen in the upper right, upper left, and lower middle of Colorado Springs, as well as one large light green cluster just to the right of Colorado Springs.

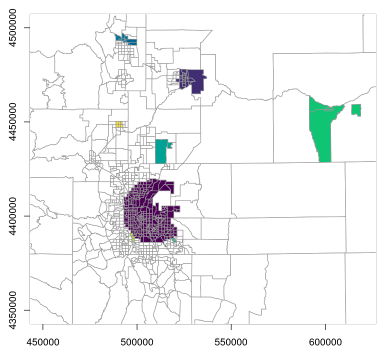
Figure 7: **Plotted results of the Spatial Scan Statistic for heterogeneous Poisson processes with a population upper bound of 50%. This analysis produced 18 statistically significant clusters with a p-value .005.** A) A map of census tracts within the entire state of Colorado with statistically significant clusters shown in color (1 color indicates one cluster). B) Census tracts within the Denver metro area with statistically significant clusters shown in color (zoom-in view of A). C) Census tracts within the Colorado Springs and Pueblo area with statistically significant clusters shown in color (zoom-in view of A).



**A)**

Interpretation of A):

There is one large cluster of census tracts seen in dark blue in the lower right area of Colorado as well as one smaller cluster seen in light blue with several census tracts in the bottom middle of Colorado.

 **B)**

Interpretation of B):

There is one large cluster shown in dark purple covering the right side of the Denver metro area, along with two small clusters (in yellow and green) just below. There are five clusters scattered above the Denver metro area that contain more rural census tracts.

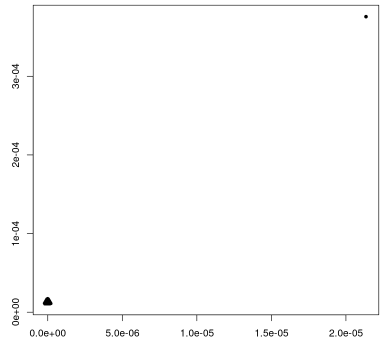
Map

Description automatically generated**C)**

Interpretation of C):

Pueblo is seen as one large cluster in dark purple. There is one large dark blue cluster seen in the lower right side of Colorado Springs, two smaller clusters in the upper right and upper left of Colorado Springs, as well as one large light green cluster just to the right of Colorado Springs.

Figure 8: **Plotted results of the goodness of fit portion versus the spatial autocorrelation portion of the observed Tango’s Index (shown as a circular dot) as well as the simulated Tango’s Indices (shown as triangles). Note: the goodness of fit portion of the observed Tango’s Index is equivalent using both values of as in both instances.** A) Tango’s index calculated in both observed and simulated data using = 394.9; which yielded a Monte Carlo p-value of .002. B) Tango’s index calculated in both observed and simulated data using = 55137.9; which yielded a Monte Carlo p-value of .002.



**A)**

Interpretation of A):

The weight matrix (calculated with = 394.9) used in this calculation of the Tango’s Index of observed and simulated data indicates very weak spatial correlation between census tracts. It is observed that even when considering the spatial correlation between census tracts to be weak, there are extreme differences in the spatial autocorrelation as well as the goodness of fit portion of Tango’s Index of the observed data compared to the simulated data. This shows there is significant clustering between census tracts in the study area as well as clustering within census tracts in the study area.

Goodness of Fit

Spatial Autocorrelation

Chart

Description automatically generated**B)**

Interpretation of B):

The weight matrix (calculated with= 55137.9) used in this calculation of the Tango’s Index of observed and simulated data indicates very strong spatial correlation between census tracts. It is observed that even when considering the spatial correlation between census tracts to be strong, there are extreme differences in the spatial autocorrelation as well as the goodness of fit portion of Tango’s Index of the observed data compared to the simulated data. However, as expected, the difference in the spatial autocorrelation portion is larger here than in A). But similarly to A), this shows there is significant clustering between census tracts in the study area as well as clustering within census tracts in the study area.

Goodness of Fit

Spatial Autocorrelation

**Conclusion**

Based on the results of the various analyses, we conclude there is statistically significant clustering of diabetes-related hospitalizations within Colorado census tracts. Though each method - and parameter choices within each method - yields slightly different results, the following Colorado counties are consistently observed as part of statistically significant clusters: Las Animas County, Weld County, Pueblo County, El Paso County, Denver County, Arapahoe County, Jefferson County, Adams County, Boulder County (Longmont area), Larimer County (upper right of Fort Collins), Otero County, Crowley County, and Bent County. For more precise results, it would be prudent to create a list of all census tracts that were shown to be part of statistically significant clusters – however in the interest of time and precision, that is not included in this report.

The first policy recommendation as a result of this analysis includes advising Primary Care Physicians working within Colorado’s clustered census tracts to bring new focus to diabetes prevention and education. This includes- but is not limited to- testing for pre-diabetes more frequently, educating patients around nutrition and balanced life-styles, as well as preparing newly diabetic patients to manage their disease effectively.

The second policy recommendation involves organizing publicly available classes such as the Diabetes Self-Management Education and Support (DSMES)2 to be offered specifically within Colorado’s clustered census tracts. This could involve both classes offered in a community center setting as well as classes held within individuals’ homes.

If completed, future modifications of this study would involve using counties or neighborhoods within Colorado as the regions of interest. Because a high number of census tracts were used for the analysis discussed in this report, large numbers of statistically significant clusters were often reported by the various methods. It may allow for more precise results and this policy recommendations if the regions of interest within Colorado were larger.

**References**

1. Joshua French, *Chapter 7 Regional-Count-Data Notes*, University of Colorado Denver, MATH 6384.

2. *Diabetes self-management education and support.* Department of Public Health & Environment. (n.d.). Retrieved December 1, 2021, from https://cdphe.colorado.gov/DSMES-Toolkit.

3. *Diabetes hospitalization rate (census tracts).* CDPHE Open Data. (n.d.). Retrieved November 30, 2021, from https://data-cdphe.opendata.arcgis.com/datasets/CDPHE::diabetes-hospitalization-rate-census-tracts.

4. *Population density (census tracts).* CDPHE Open Data. (n.d.). Retrieved December 8, 2021, from https://data-cdphe.opendata.arcgis.com/datasets/2128d5e4260a47c28b3fd124f79008a1\_0/about.

5. https://diabetes.org/sites/default/files/2021-10/ADV\_2021\_State\_Fact\_sheets\_Colorado.pdf

Appendix

#load diabetes hospitalization rate data shape <-sf::st\_read("Diabetes\_Hospitalization\_Rate\_(Census\_Tracts)/Diabetes\_Hospitalization\_Rate\_(Census\_Tracts).shp") suppressPackageStartupMessages(library(GISTools))

population < -sf::st\_read("Population\_Density\_(Census\_Tracts)/Population\_Density\_(Census\_Tracts).shp")

plot(population$geometry, border="grey60")

select\_shape <- na.omit(shape)

choropleth(select\_shape$geometry, select\_shape$DIABETES\_A, axes = TRUE) shades = auto.shading(select\_shape$DIABETES\_A) choro.legend(px='bottomleft', sh=shades, fmt="%4.1f")

title(main = "Diabetes-Related Hospitalization Rate in Colorado Census Tracts")

choropleth(select\_shape$geometry, select\_shape$DIABETES\_A, axes = TRUE, xlim = range(-104.5,-105.5), ylim = c(39.4,40.4)) shades = auto.shading(select\_shape$DIABETES\_A)

choropleth(select\_shape$geometry, select\_shape$DIABETES\_A, axes = TRUE, xlim = range(-104.5,-105.5), ylim = c(38,39)) shades = auto.shading(select\_shape$DIABETES\_A)

choropleth(select\_shape$geometry, select\_shape$DIABETES\_A, axes = TRUE, xlim = range(-108, -109), ylim = c(38,40)) shades = auto.shading(select\_shape$DIABETES\_A)

#merge diabetes data with population data

diabetes\_pop <- merge(as.data.frame(select\_shape), as.data.frame(population), by.x = "TRACT\_FIPS", by.y = "Tract\_FIPS") diabetes\_pop$geometry.y <- NULL

library(sf) newshape <- sf::st\_sf(diabetes\_pop)

#make geography non-scalar newshape <- st\_transform(newshape, "+init=epsg:32613") p2 <- st\_point\_on\_surface(newshape)

newshape <- as.data.frame(newshape) newshape$counts = (newshape$DIABETES\_A/100000)\*newshape$Population

p2 <- as.data.frame(p2)

newshape$centroids <- p2$geometry.x library(tidyverse) separated\_coord <- newshape %>% mutate(latitude = unlist(map(newshape$centroids,1)), longitude = unlist(map(newshape$centroids,2)))

newshape2 <- sf::st\_sf(separated\_coord) plot(newshape2$geometry.x, border="grey60") points(newshape2$latitude, newshape2$longitude, col = 'red', pch = 19)

suppressPackageStartupMessages(library(smerc)) suppressPackageStartupMessages(library(spdep))

sum(newshape2$Population)

coords = as.data.frame(cbind("x" = newshape2$latitude, "y" = newshape2$longitude)) cases = (newshape2$counts)

cepp16000 = cepp.test(coords = coords, cases = cases, pop = newshape2$Population, nstar = 16000, alpha = 0.01)

cepp20000 = cepp.test(coords = coords, cases = cases, pop = newshape2$Population, nstar = 20000, alpha = 0.01) cepp30000 = cepp.test(coords = coords, cases = cases, pop = newshape2$Population, nstar = 30000, alpha = 0.01)

summary(cepp16000)

summary(cepp20000)

summary(cepp30000)

cepp150000 = cepp.test(coords = coords, cases = cases, pop = newshape2$Population, nstar = 150000, alpha = 0.01)

summary(cepp150000)

plot(newshape2$geometry.x, border = "grey60", axes = TRUE, col = color.clusters(cepp150000))

plot(newshape2$geometry.x, border = "grey60", axes = TRUE, col = color.clusters(cepp150000), xlim = range(450000:600000), ylim = c(4350000, 4500000))

plot(newshape2$geometry.x, border = "grey60", axes = TRUE, col = color.clusters(cepp150000), xlim = range(450000:600000), ylim = c(4350000, 4500000))

cepp.05 = cepp.test(coords = coords, cases = cases, pop = newshape2$Population, nstar = .05\*sum(newshape2$Population), alpha = 0.01) summary(cepp.05)

plot(newshape2$geometry.x, border = "grey60", axes = TRUE, col = color.clusters(cepp.05))

plot(newshape2$geometry.x, border = "grey60", axes = TRUE, col = color.clusters(cepp.05), xlim = range(450000:600000), ylim = c(4350000, 4500000))

coords = as.data.frame(cbind("x" = newshape2$latitude, "y" = newshape2$longitude))

cases = (newshape2$counts) population = newshape2$Population # expected number of cases

e = sum(cases)/sum(population) \* population # apply circular scan method scan\_.5 = scan.test(coords = coords, cases = cases, pop = population, ex = e, ubpop = 0.5, nsim = 999, alpha = 0.005) summary(scan\_.5)

# apply circular scan method scan\_.1 = scan.test(coords = coords, cases = cases, pop = population, ex = e, ubpop = 0.1, nsim = 999, alpha = 0.005) summary(scan\_.1)

mycol = grDevices::hcl.colors(18)

plot(newshape2$geometry.x, border="grey60", axes=TRUE, col = color.clusters(scan\_.5, col = mycol))

title(main = "Plot of clusters with population upper bound .5") plot(newshape2$geometry.x, border = "grey60", col = color.clusters(scan\_.5, col = mycol), xlim = range(450050:620000), ylim = c(4350000, 4500000)) title(main = "Plot of clusters with population upper bound .5") plot(newshape2$geometry.x, border = "grey60", col = color.clusters(scan\_.5, col = mycol), xlim = range(450000:600000), ylim = c(4350000, 4500000)) title(main = "Plot of clusters with population upper bound .5")

mycol = grDevices::hcl.colors(32)

# color.clusters(scan, col = mycol) colors the 3 clusters using the desired clusters

plot(newshape2$geometry.x, border="grey60", axes=TRUE, col = color.clusters(scan\_.1, col = mycol)) title(main = "Plot of clusters with population upper bound .1")

plot(newshape2$geometry.x, border = "grey60", col = color.clusters(scan\_.1, col = mycol), xlim = range(450050:620000), ylim = c(4350000, 4500000)) title(main = "Plot of clusters with population upper bound .1") plot(newshape2$geometry.x, border = "grey60", col = color.clusters(scan\_.1, col = mycol), xlim = range(450000:600000), ylim = c(4250000, 4320000)) title(main = "Plot of clusters with population upper bound .1")

coords = as.matrix(cbind("east" = newshape2$longitude, "north" = newshape2$latitude))

cases = (newshape2$counts) pop = newshape2$Population

# Find distance matrix d = as.matrix(dist(coords))

ds <- apply(d, 1, sort) range(ds[2,])

coords = as.matrix(cbind("east" = newshape2$longitude, "north" = newshape2$latitude))

cases = (newshape2$counts) pop = newshape2$Population ################################################## # Exponential decay weight matrix # use different kappas in defining weights

w.1 <- dweights(coords, kappa = 394.9)

w7 <- dweights(coords, kappa = 55137.9)

(tango\_394.9 <- tango.test(cases, pop, w.1, nsim = 499))

(tango\_55137.9 <- tango.test(cases, pop, w7, nsim = 499))

gof <- c(tango\_394.9$gof,tango\_394.9$gof)

sa <- c(tango\_55137.9$sa,tango\_55137.9$sa) plot(tango\_394.9) title(main = "plot of Tango Statistics, Null and Observed, k = 394.9") plot(tango\_55137.9) title(main = "plot of Tango Statistics, Null and Observed, k = 55137.9")