Task 6
Collocation Mining of Zinj Dataset

Data Preprocessing

- The initial format of the data was quite difficult to process into a viable format. The Keyhole Markup Language (KML) file provided is an XML notation used explicitly for geographic visualization of “placemarks” and annotations on said placemarks. The process followed for initial processing and extraction of the relevant data was to (a) first parse the XML using a specifically designed XML parser, (b) filter the relevant data from the XML nodes, and finally (c) output the data in a flat file format with which further processing could proceed.
- For the parsing of the data, each node name was read and compared against a known list of valid nodes. If the node was flagged for extraction, the value was printed to file. Both inner and outer boundaries of the placemarks were contained in “coordinates” nodes, but the inner boundaries were not needed. To account for this, extra checks had to be implemented specifically for this.
- Another issue encountered was the “nature” of the buildings or essentially the class of the buildings. This data was enclosed in small blocks of HTML within XML comment or “CDATA” nodes. To accommodate this, first, the parsing of CDATA had to be enabled within the XML parser library’s settings. Then, when a comment node was encountered, the value of the CDATA node was captured and required further processing as it was an incomplete block of HTML. A second instance of the XML parser was required to process this data separately and, because the parser required a single parent node, the HTML had to be manually enclosed in an artificial element, in this case <parent></parent>. Once the HTML block was able to be loaded into the XML parser, the parent’s descendants could be extracted as strings, the “nature” element could be searched for, and the value of that element could be written to file. From this point, the data, now extracted into Zinj.txt such that each single piece of data was on a new line, could be further processed with ease.
- From the text file generated above, we parse line wise and create a data frame with columns <Zone, Id, Type, Latitude, Longitude>. The Latitude and Longitude are extracted by forming a polygon from of the building coordinates and computing their CENTROIDS. The centroids are thus a single point representation of each building (Lat, Long). The dataset this generated can be seen in Figure 1 below.

```
In [5]: df_obj.head()
```

Out[5]:

	Zone	Id	Type	Lat	Long
0	Zone_1	2667	single_house	48.614921	7.737321
1	Zone_1	4053	single_house	48.633323	7.728775
2	Zone_1	3999	single_house	48.633106	7.728857
3	Zone_1	2743	garage	48.616600	7.729351
4	Zone_1	3915	single_house	48.632716	7.729757

Figure 1. Dataframe created from Zinj Dataset

The dataset contains data from 3 different zones. The scatterplot of the computed centroids of all 3 zones is depicted in Figure 2.

Figure 2. Visualization of newly created centroids of Zinj Dataset

Randomization:

In our experiments and Analysis, we require a set of randomized points to compare the collocations of Zinj dataset with. In order to compute the randomized set to conserve the proportion of house/building types within each zone, we use R to sample() the data within Zone 1, Zone 2 and Zone 3 separately by shuffling the list of Types, then putting them together to form a randomized set for each zone. These randomized points of 3 zones are put together to form a single randomized dataset. The head of the dataset can be found in Figure 3. and Figure 4. visualizes the randomized dataset.

```
> head(random_dataset)
    Zone    Id         Type      Lat     Long
1 Zone_1 2667 single_house 48.61492 7.737321
2 Zone_1 4053 single_house 48.63332 7.728775
3 Zone_1 3999 single_house 48.63311 7.728857
4 Zone_1 2743 single_house 48.61660 7.729351
5 Zone_1 3915 single_house 48.63272 7.729757
6 Zone_1 2695 single_house 48.61558 7.731594
```

Figure 3. Randomized Dataset

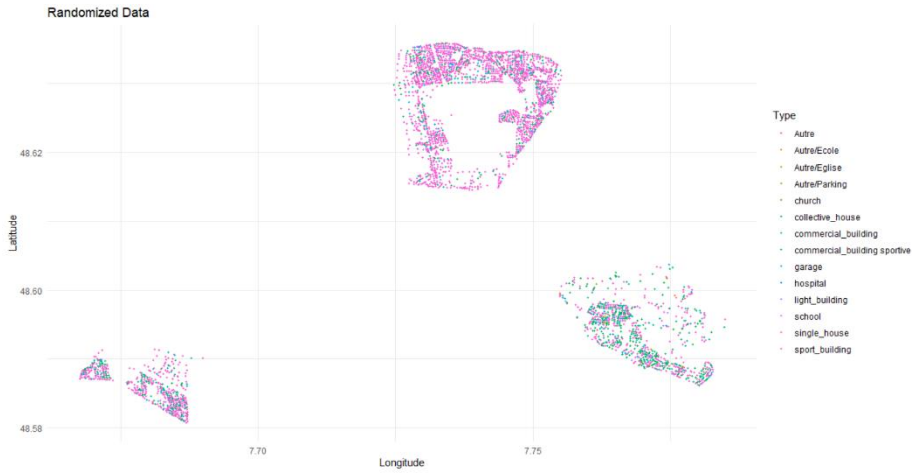


Figure 4. Visualization of Randomized Dataset

Implementation:

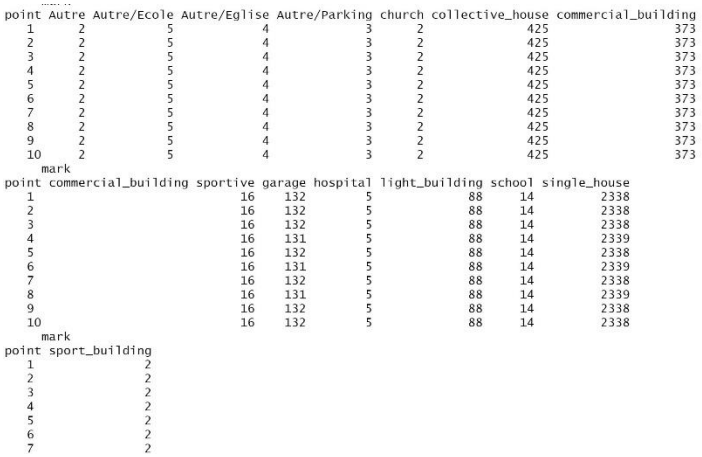
1. Generation and Mapping of Point pattern

The minimum and maximum values were calculated for the longitudes and latitudes vectors. Longitudes would be considered as the X axis and Latitudes as Y axis. To generate and map the points from the longitude and latitude pairs, min and maximum limits are necessary. This helps in determining the boundaries for the point pattern box. The point pattern can be mapped using rectangular box, convex hull for a set of points, elliptical etc. We have used rectangular box for mapping the point pattern. This defines the study region for the data tasks.

For generating point pattern, ppp function is used from Spatstat library. The point pattern is generated by passing Longitude as X vector [3410 * 1, Single column vector], Latitude as Y vector, [3410 * 1, Single column vector]. Also, for each pair of X and Y values, marks are attached which denotes the labels for the type of observation. This mapping creates a categorical study region in X, Y plane in which every point has X, Y coordinate and class label. It should be noted that since we are not considering the elevational values for the observations the pattern generated is purely 2-dimensional pattern. This is called multitype point pattern in spatial terms.

2. Exploratory Data Analysis

For each point in point pattern, marktable function calculates the neighboring marks within a given radius. It compiles the contingency table of the marks within a given radius of each data point.



point	Autre	Autre/Ecole	Autre/Eglise	Autre/Parking	church	collective_house	commercial_building
1	2	5	4	3	2	425	373
2	2	5	4	3	2	425	373
3	2	5	4	3	2	425	373
4	2	5	4	3	2	425	373
5	2	5	4	3	2	425	373
6	2	5	4	3	2	425	373
7	2	5	4	3	2	425	373
8	2	5	4	3	2	425	373
9	2	5	4	3	2	425	373
10	2	5	4	3	2	425	373

mark

point	commercial_building	sportive	garage	hospital	light_building	school	single_house
1		16	132	5	88	14	2338
2		16	132	5	88	14	2338
3		16	132	5	88	14	2338
4		16	131	5	88	14	2339
5		16	132	5	88	14	2338
6		16	131	5	88	14	2339
7		16	132	5	88	14	2338
8		16	131	5	88	14	2339
9		16	132	5	88	14	2338
10		16	132	5	88	14	2338

mark

point	sport_building
1	2
2	2
3	2
4	2
5	2
6	2
7	2

Figure 5. Contingency Table

- From the contingency table in Figure 5, we clearly see the marks pattern for given specified radius (0.1). No of commercial buildings, single houses, collective houses are considerably high a compare to other observations. This insight is practically comparable with real world. No of single houses is maximum in this case. Therefore, one can assume that the data mainly comes from the residential area or a mix of residential and commercial area. If the data would have been from industrial or commercial area then the observations for commercial buildings, garages would have been more than single houses.
- It is also found that the dataset has observations which aren't part of either primary building types or secondary building types. In further investigation of the data collocation, these observations would not be considered.

3. Quadrat Analysis

In quadrat analysis, the study region gets divided equally into several rectangles which are called as quadrats. No of observations or points are calculated for each rectangle. This helps in understanding of the gird pattern of the data. It can also help in checking if the data is centralized in certain rectangles than others etc. This technique is an elementary technique for analyzing spatial point patterns.

	x			
y	[7.67,7.7)	[7.7,7.73)	[7.73,7.76)	[7.76,7.78]
[48.62,48.64]	0	16	1476	0
[48.61,48.62)	0	0	551	0
[48.6,48.61)	0	0	0	5
[48.59,48.6)	0	0	7	546
[48.58,48.59)	532	0	0	277

Figure 6. Result of Quadrat Analysis

Table 6 shows the result of the qaudratount function. As we can see, the study region has been divided into 5*4 grid (specified by the user) and observations are counted for each grid thus formed. It can be clearly observed from the graph that the data is highly clustered for some regions since we are having low to 0 counts for other regions. To be specific, the data at 2 lower right grids, lower left corner and at the upper region for 3rd column seems to have maximum observations. The points X and Y represents longitude and latitude respectively for the study

region under consideration. This result from Quadrat Analysis can be corroborated from the density plot of the point pattern which can be seen in Figure 7.

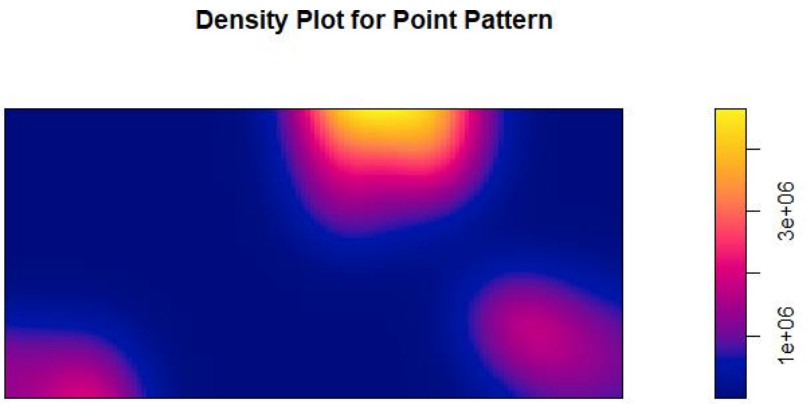


Figure 7. Point Pattern Density Plot

Task A:

To check whether buildings are clustered or randomly distributed we have leveraged the use of Ripley's reduced second moment function. This function calculates the expected number of additional random points within distance r for a random point X in the study region. With this statistic, we are trying to summarize the aspects of inter point dependency or clustering. In simpler words, this would try to capture the relationship between random points for different radius. If plotted on the graph we should be able to see how clustering or dispersion changes at different neighboring radius.

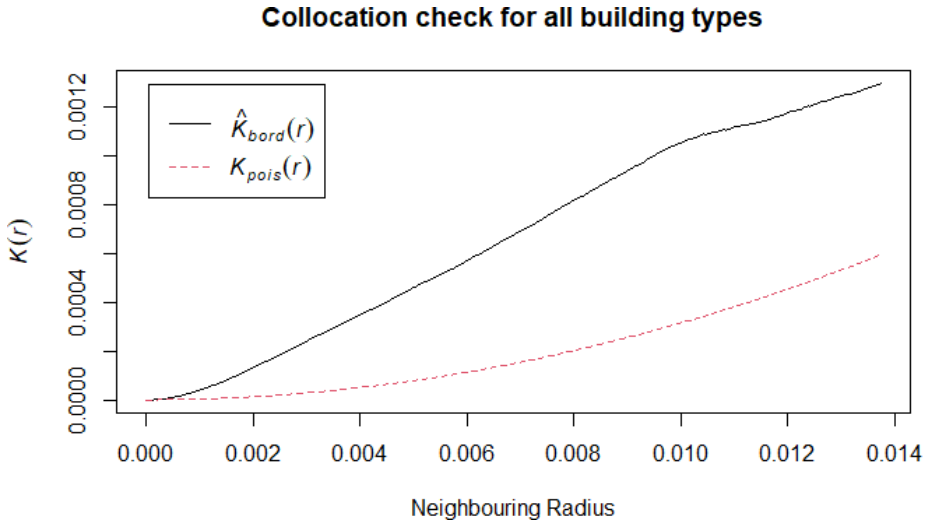


Figure 8. Collocation Check for all building types

For inferential purpose, the K estimate for observed spatial pattern is compared with the theoretical Poisson process. Theoretically, The Poisson process is completely random in nature for which $K(r) = \pi r^{2.}$ As we can see from the graph, the red curve represents the poison process or theoretical process whereas black curve represents the observed spatial pattern for the area

under consideration. It is clearly visible from the graph that the observed spatial curve deviates from theoretical curve by considerable amount in positive direction This represents spatial clustering or collocation. More specifically the observed pattern deviates with strong positive slope till radius = 0.010 and then the slope starts decreasing. Therefore, the rate of change until 0.010 neighboring radius is high as compared to the neighboring radius beyond 0.010.
The Ripley's K function for observed spatial pattern is calculated using border edge correction. Edge correction is used to minimize the effects of points on the border. If otherwise calculated without using edge correction would mislead results by affecting overall statistic. Border points generally has many points to their one side and very low no of points or none in other side. This may adversely affect the whole function calculation. This is the main motivation behind edge correction method implied in this graph.

Task B:

1. Auto Collocation for Commercial Buildings

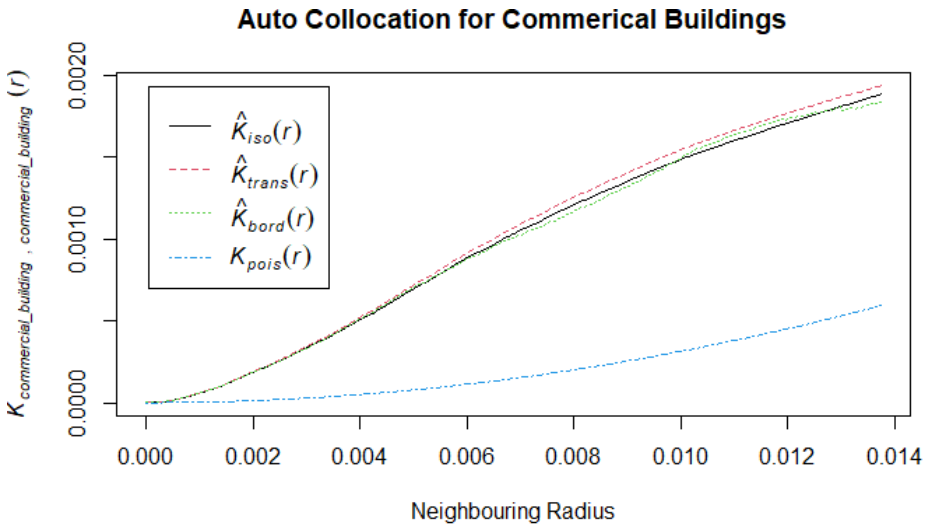


Figure B1.

As we can see in the plot from Figure B1 the Ripley isotopic correction estimate (K_{iso}) closely tracks along with the k_{trans} and k_{bord} staying fully within the envelope and having great precision at the beginning (0.0 -0.5) and gradually begins to deviate away downwards which shows that the k estimate is not so good an estimate for the data set and may continue to deviate further as the radius increases thereby decreasing the k's quality which is as a result of the sample data changes with respect to distance. The K_{pois} is below the k_{iso} indicating that the graph is more clustered than a random distribution.

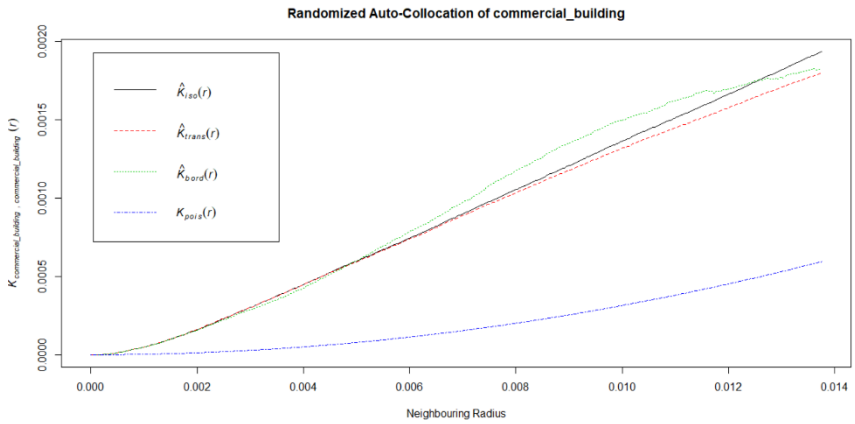


Figure B2.

In Figure B2 the k_{iso} goes above the envelope as variance increases. At the beginning (0.0 - 0.65) the observed k tracks the expected k correctly but begins to deviate as radius increases and fully exits the envelope at 0.0128 which could indicate a more clustered data set but still not a great k prediction.

2. Auto Collocation for Collective Houses

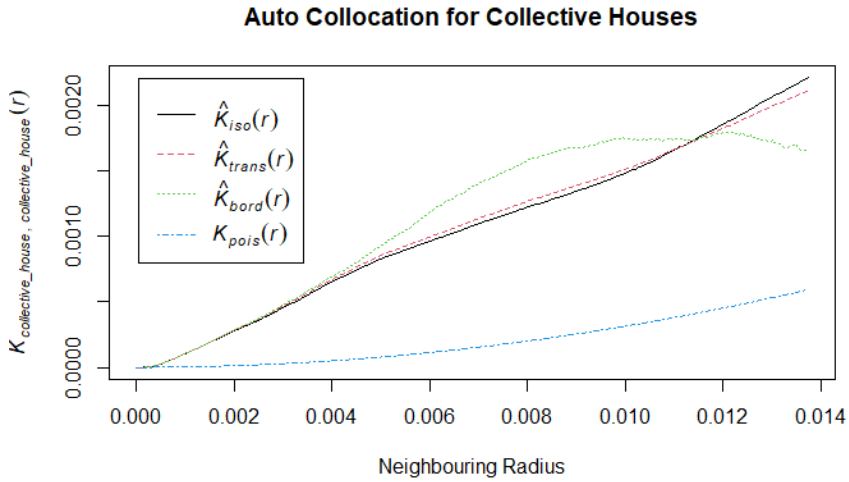


Figure B3.

In Figure B3 we see a better tracking from the observed k as it deviates a little over distance but starts to lose quality at 0.012 and exits the envelop also. The k value here is higher than the upper confidence envelope and can indicate that spatial clustering is statistically significant.

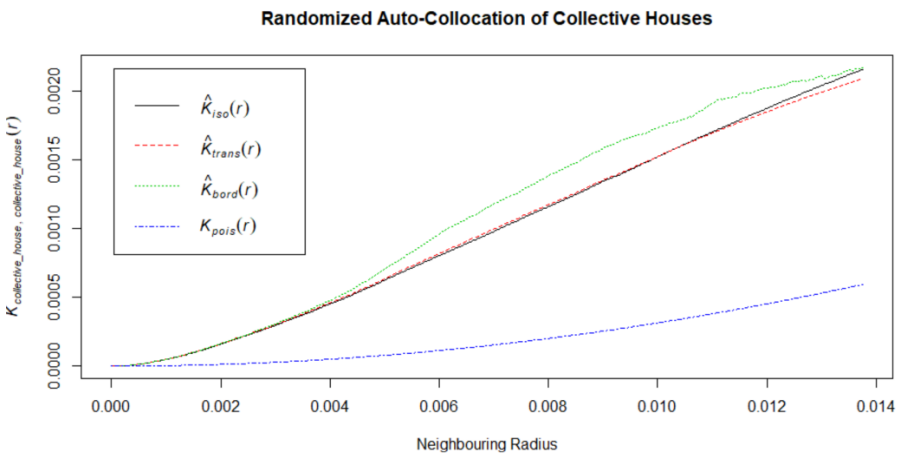


Figure B4.

In Figure B4, we do not really see much impact as regard tracking with respect to Figure B3 but the major difference here is that the observed k value and expected k both remain within the envelope.

Task C: Bivariate Spatial Analysis

1. Commercial Buildings and Light Buildings

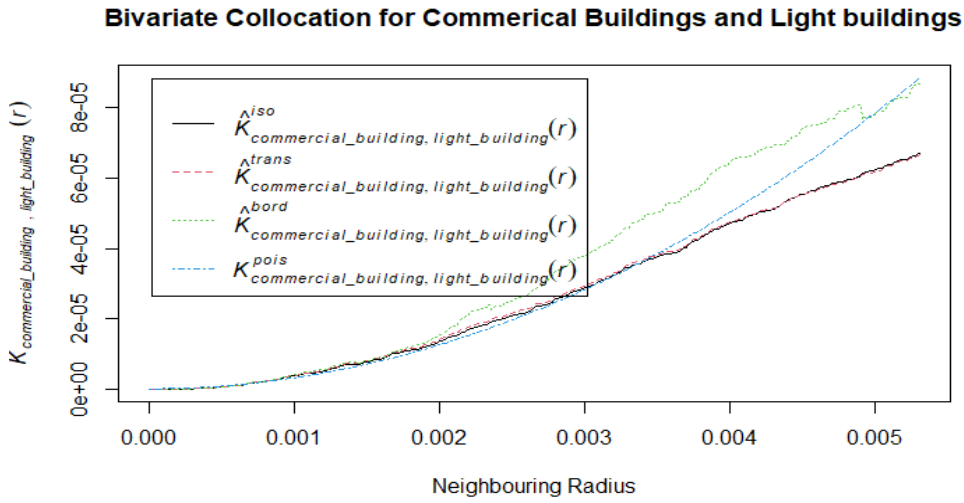


Figure C1

Figure C1 clearly shows that there is a random relation between the commercial and light buildings till the radius is about 0.002. After this point, when radius is increased, there is a positive collocation for only the border-corrected estimate and the other two remain random. After the radius value goes above 0.003 units, the collocation for translation and isotropic corrected estimate become negative which tells us that they are dispersed. If the radius becomes further large above 0.005 units, all three distributions will give a dispersed output.

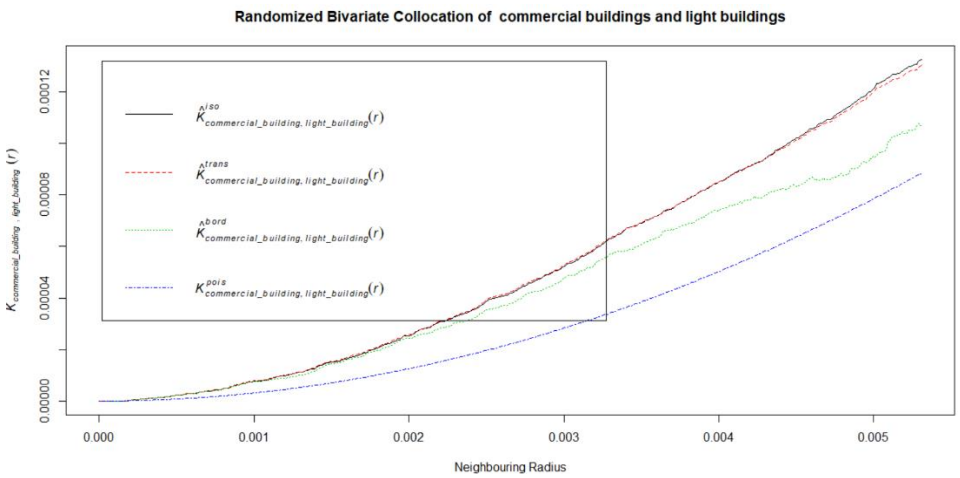


Figure C2

Figure C2 for random data clearly shows that there is a positive collocation for all the estimates from the very start of the data. This is not like the interpretation above for non-randomized data. Hence, we reject this null hypothesis and accept the other hypothesis.

Thus, we can say that there is a positive collocation for only border corrected estimate above the radius of 0.002 for Commercial buildings and light houses.

Normalization:

It should be noted that there are considerably fewer Light Buildings to other building types and the produced analyses should not be relied upon.

2. Commercial Buildings and Single Houses

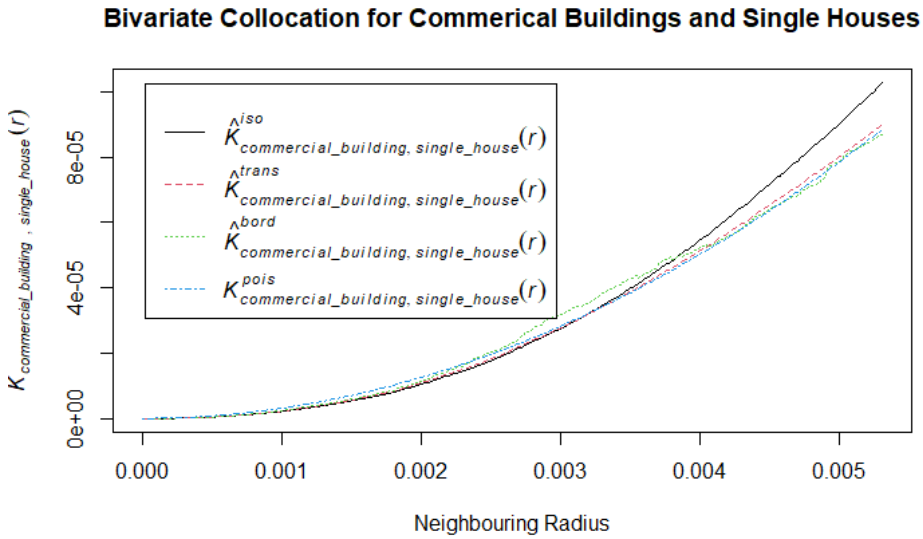


Figure C3

Figure C3 shows that there is a random relation between the commercial buildings and single houses till the radius is about 0.0035. This is because the lines are closely located to the poison estimate

and are neither collocated nor anti-collocated. After this point, when radius is increased, there is a positive collocation for only the isotropic-corrected estimate and the other two remain random.

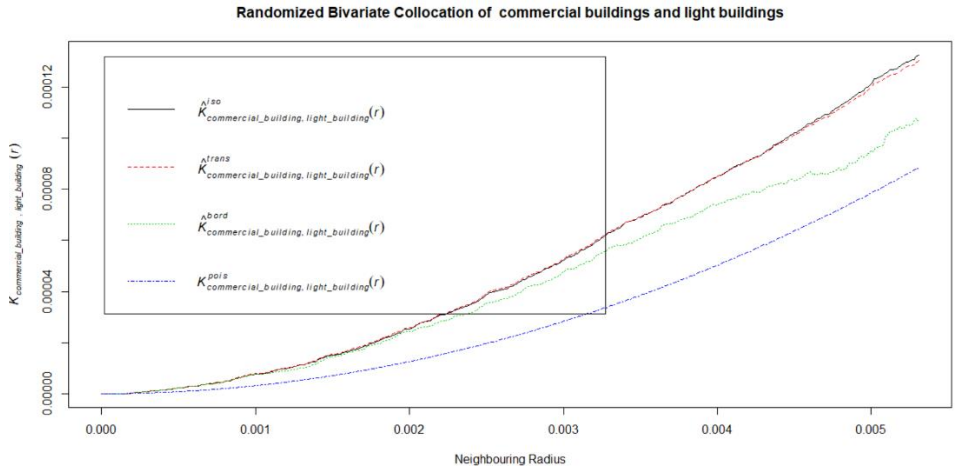


Figure C4

Figure C4 for random data clearly shows that there is a positive collocation for all the estimates from the very start of the data. This is not like the interpretation above for non-randomized data. Hence, we reject this null hypothesis and accept the other hypothesis.

Thus, we can say that there is a positive collocation for only isotropic-corrected estimate above the radius of 0.0035 for Commercial buildings and single houses.

3. Commercial Buildings and Garages

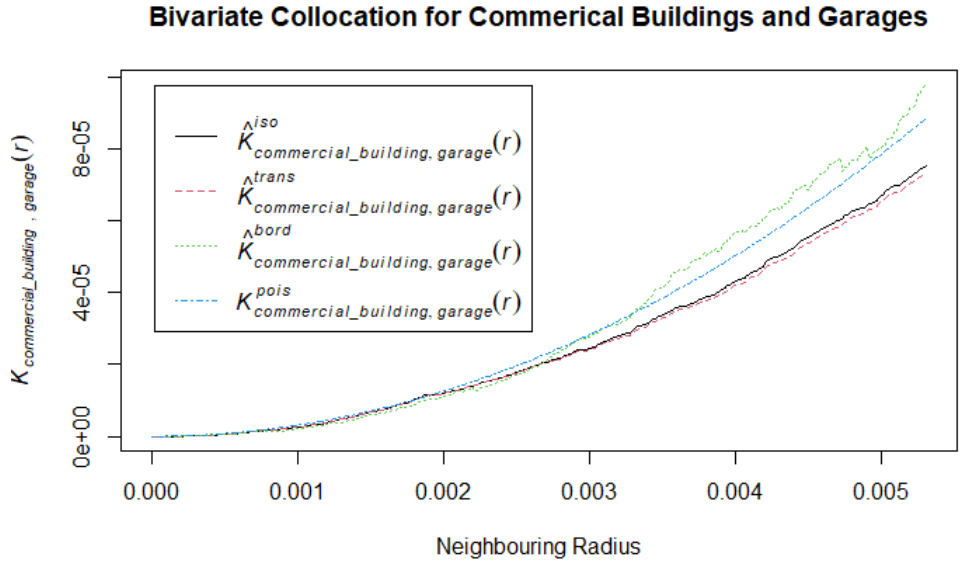


Figure C5

Figure C5 clearly shows that there is a random relation between the commercial and garages till the radius is about 0.002. After this point, when radius is increased, there is a negative collocation for translation and isotropic corrected estimate and only the border-corrected estimate remains random. After the radius value goes above 0.0035 units, the collocation for translation and isotropic corrected estimate remain negative which tells us that they are dispersed, and border-corrected estimate becomes positively collocated which means they are located closely.

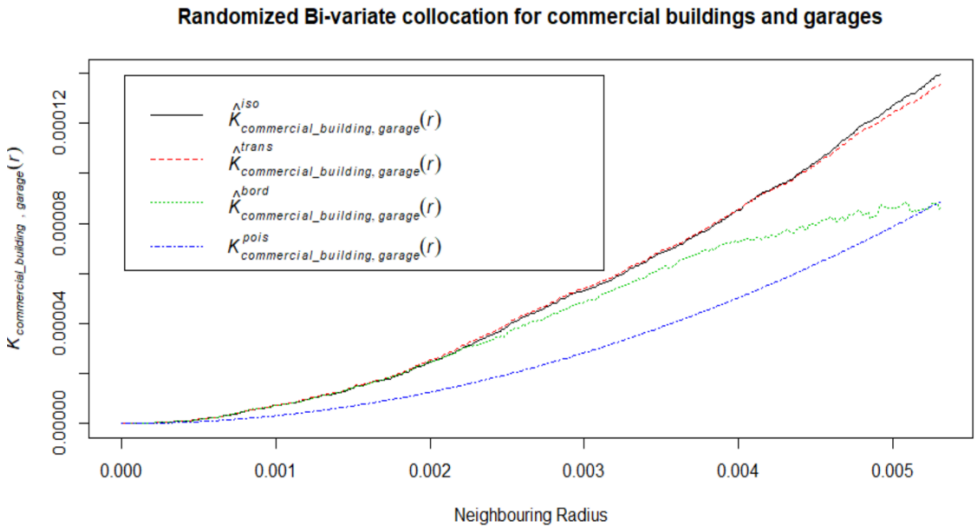


Figure C6

Figure C6 for random data clearly shows that there is a positive collocation for all the estimates from the very start of the data. This is not like the interpretation above for non-randomized data. Hence, we reject this null hypothesis and accept the other hypothesis.

Thus, we can say that there is a negative collocation for translation and isotropic corrected estimate and for border-corrected estimate there is a positive collocation above the radius of 0.0035 for Commercial buildings and garages.

4. Commercial Buildings and Schools

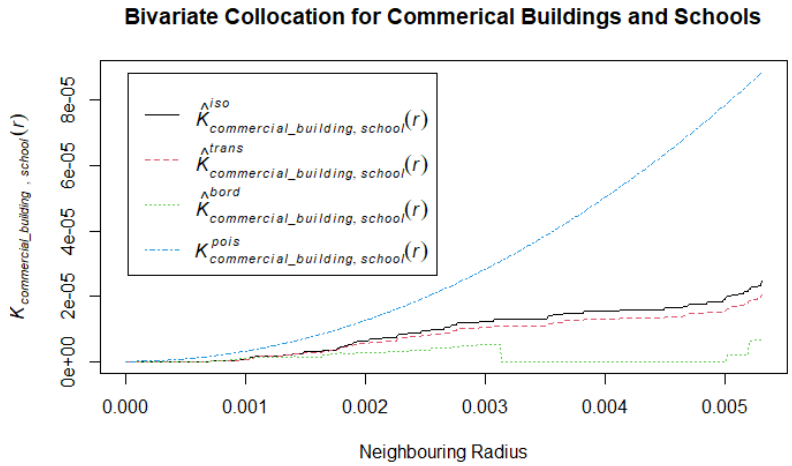


Figure C7

All the estimates have a negative collocation for commercial buildings and schools.

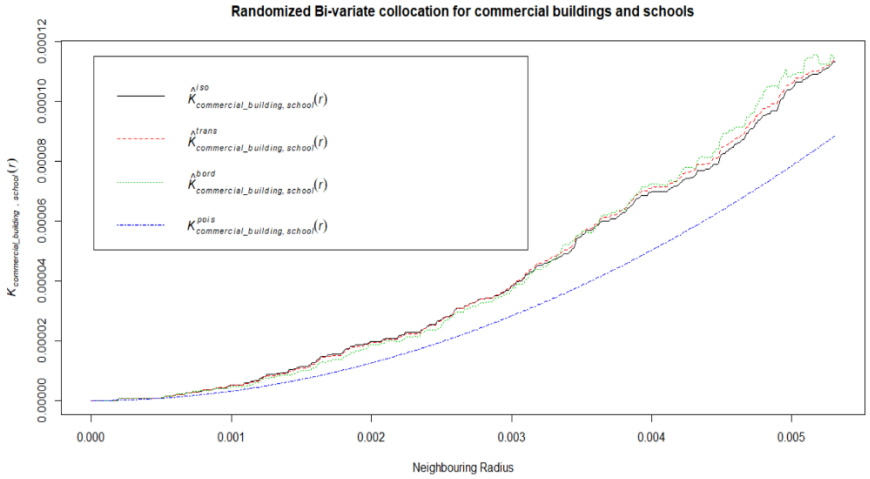


Figure C8

All the estimates in random data have positive collocation for commercial buildings and schools. This is not like the interpretation above for non-randomized data. Hence, we reject this null hypothesis and accept the other hypothesis.

Thus, we can say that there is a negative collocation for translation, isotropic corrected estimate and border-corrected estimate for Commercial buildings and garages.

Normalization:

It should be noted that there are considerably fewer Schools to other building types and the produced analyses should not be relied upon.

5. Collective Houses and Light Buildings

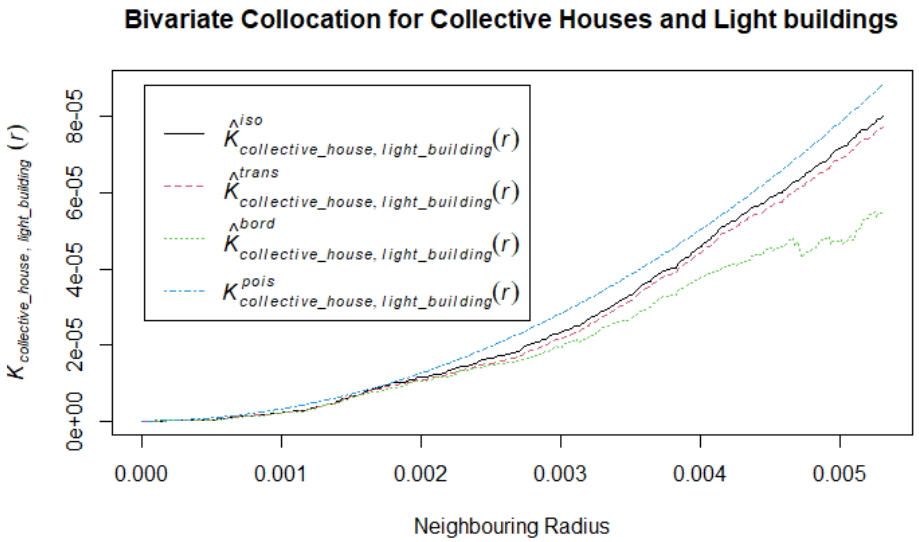


Figure C9

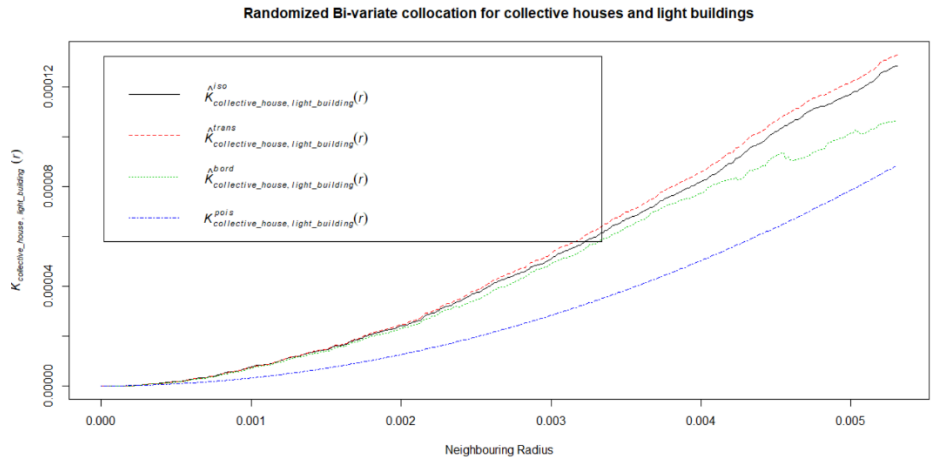


Figure C10

From Figure C9 and Figure C10, we can see that the randomized data clearly shows that the collective houses and light buildings are collocated. Three edge correction methods are almost similar till the radius = 0.002, but afterwards starts deviating. For the non-randomized data however, the curves show that these two building types are anti collocated or dispersed. For all edge correction methods. This means that the null hypothesis gets rejected since the alternative hypothesis shows different relationship than complete spatial randomness. Therefore, we can assess that the buildings must be anti-collocated for the study region under consideration. The buildings don't show high dispersion unlike graph no. 4 but buildings are said to be dispersed by moderate magnitude.

Normalization:

It should be noted that there are considerably fewer Light Buildings to other building types and the produced analyses should not be relied upon.

6. Collective Houses and Single Houses

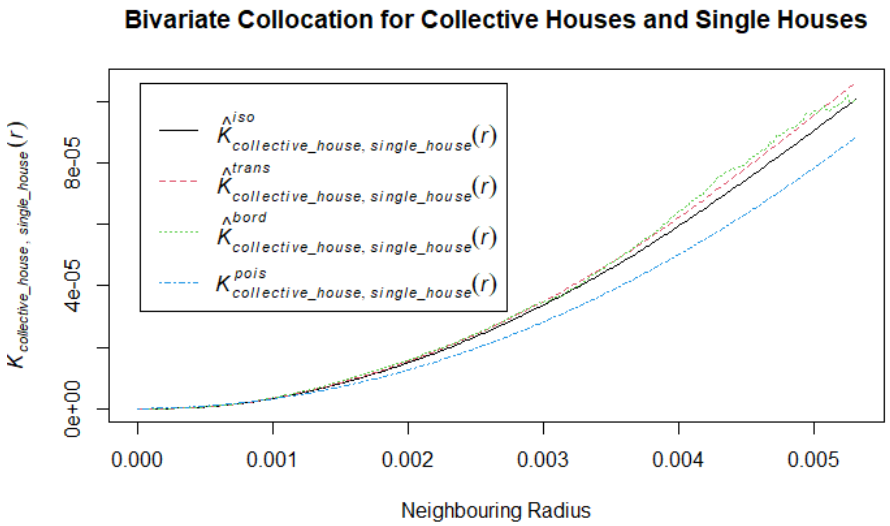


Figure C11

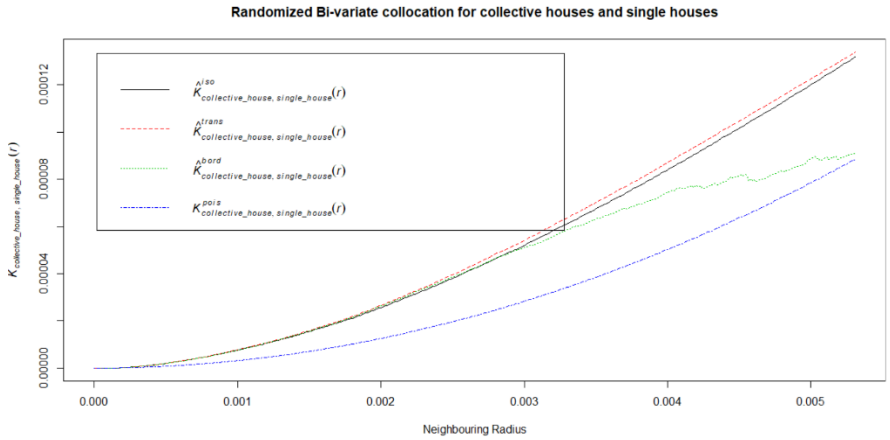


Figure C12

Figure C11 shown above present slightly higher K values than the K Poisson curve; however, Figure C12 shows a greater separation between the K iso and K trans functions and the K Poisson curve. The K bord function climbs steadily with the other two K functions but starts to fall near the 0.004 radius, unlike the nonrandomized bivariate collocation graph where the K bord follows the other two functions. Therefore, we may reject the null hypotheses. This indicates a slight collocation correlation between the collective houses and single houses.

7. Collective Houses and Garages

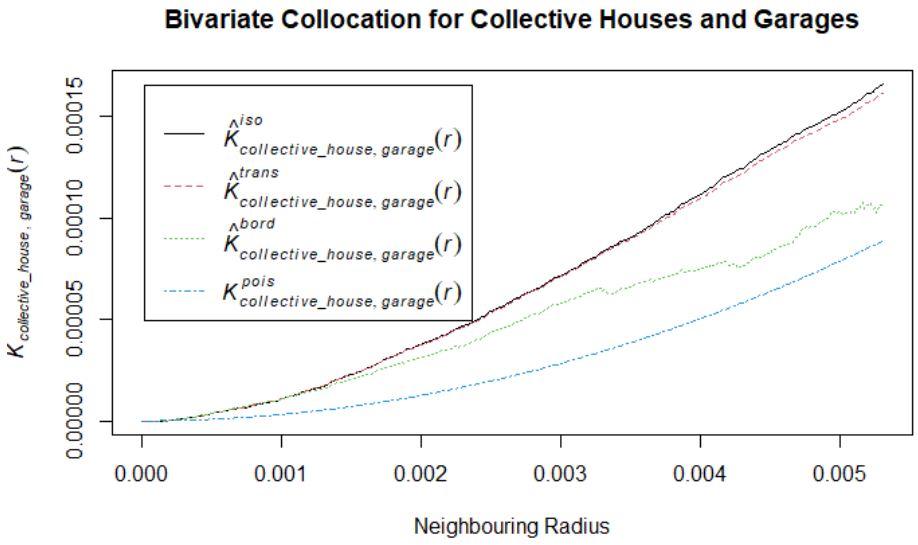


Figure C13

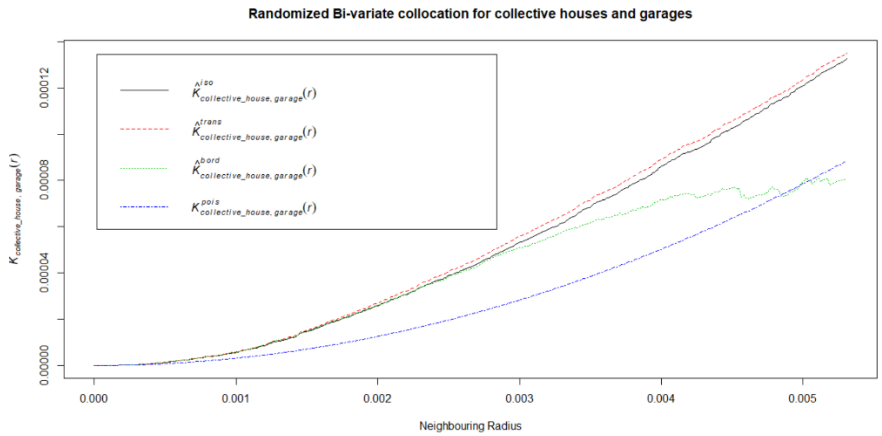


Figure C14

Figure C13 shown above present slightly higher K values than the K Poisson curve. Figure C14 also shows the separation between the K functions and the K Poisson curve at roughly the same distance. Both the K bord functions climb steadily with the other two K functions but starts to fall near the 0.003 radius. Therefore, we must accept the null hypotheses which indicates a no collocation correlations between the collective houses and garages.

8. Collective Houses and Schools

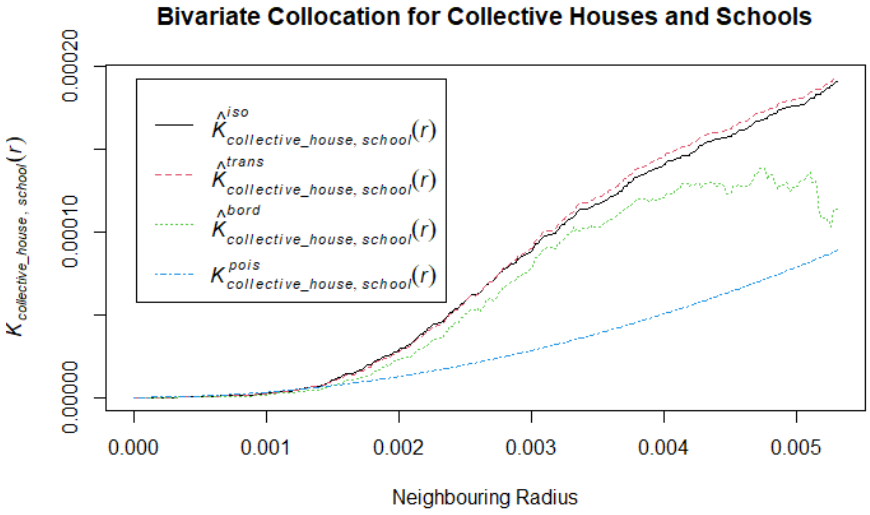


Figure C15

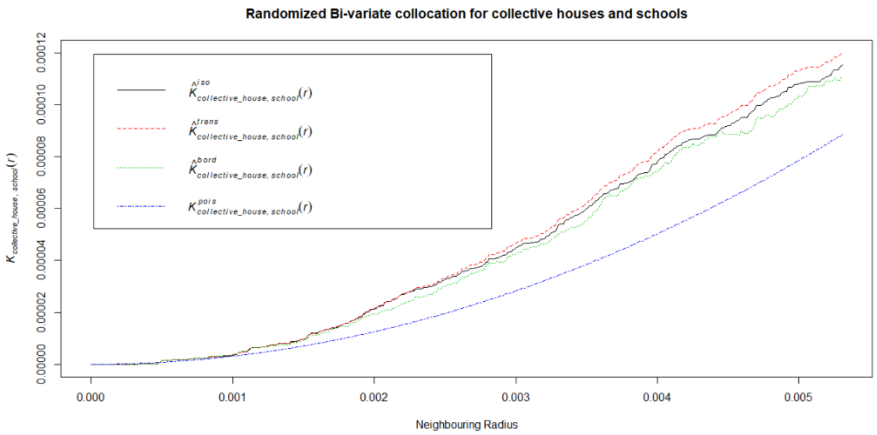


Figure C16

The nonrandomized collocation functions shown above present significantly higher K values than the K Poisson curve; however, the randomized collocation graph shows only a slight separation between the K functions and the K Poisson curve. The K bord function climbs steadily with the other two K functions but starts to fall near the 0.004 radius, unlike the nonrandomized bivariate collocation graph where the K bord follows the other two functions. Therefore, we may reject the null hypotheses. This indicates a strong collocation correlation between the collective houses and schools.

Normalization:

It should be noted that there are considerably fewer Schools to other building types and the produced analyses should not be relied upon.

9. Collective Houses and Commercial Buildings

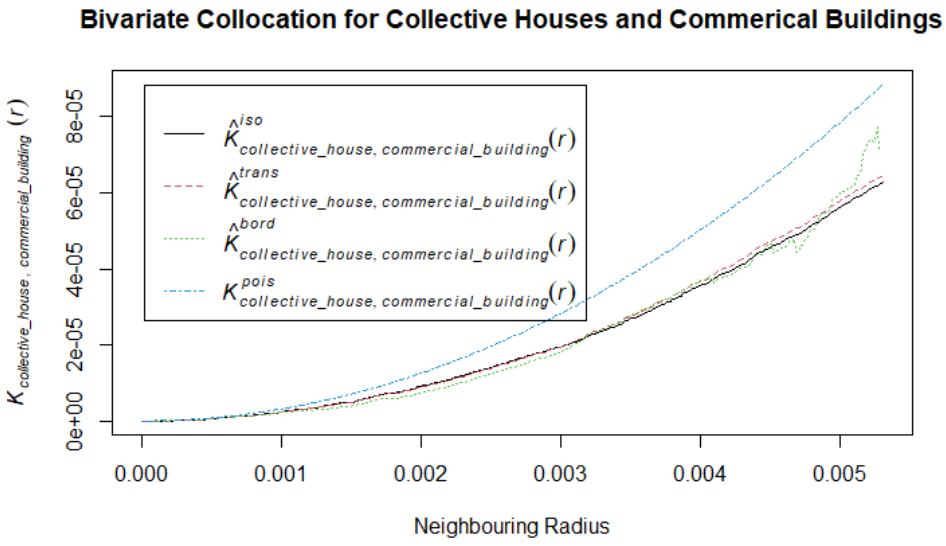


Figure C17

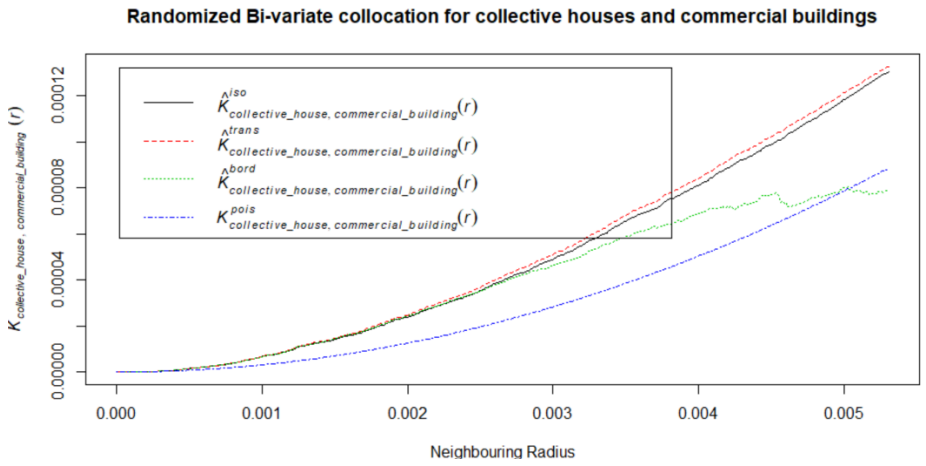


Figure C18

Figure C17 shown above present slightly lower K values than the K Poisson curve; however, Figure C18 shows slightly higher K values than the K Poisson curve. The K bord function climbs steadily with the other two K functions throughout the nonrandomized collocation graph, but the nonrandomized bivariate collocation graph presents the K bord follow the other two functions only up to roughly 0.003 radius. Therefore, we may reject the null hypotheses. This indicates a strong anti-collocation between the collective houses and commercial buildings.

Conclusion

Overall, we see that the Primary buildings such as Commercial Buildings and Collective houses seem to show collocation (positive spatial clustering) with itself. However, Commercial buildings and Collective houses are anti-collocated. In simpler terms, distribution of Commercial buildings seems to be in clusters and distribution of Collective houses also seems to be in clusters, but Commercial buildings and Collective houses are not clustered together. When bivariate spatial analysis is done on Primary building types of Zone 1 which consists of most of the data, some interesting observations were made. For all building types we have observed high interpoint interaction of stochastic dependence (the buildings are found to be highly clustered).

The most important observations are that Schools are highly anti-collocated with Commercial Buildings but collocated with Collective Houses. Light buildings are slightly anti-collocated with Collective houses as well as Commercial Buildings. Single Buildings also seem to be quite collocated with Collective houses but seem to be anti-collocated with Commercial Buildings. Garages on the other hand are anti-collocated with Commercial buildings but quite collocated with Collective houses. The null hypotheses for Auto Collocation of Commercial Buildings and Collective Houses were not rejected and both the building types are clustered throughout the study region.

Spatial Relationships found between the building types seem to be very apt when compared to practical life.

Despite the results, because there are considerably fewer Light Houses and Schools compared to other building types, the produced analyses should not be relied upon.

References

1. https://wiki.landscapetoolbox.org/doku.php/spatial_analysis_methods:ripley_s_k_and_pair_correlation_function
2. https://cran.r-project.org/web/packages/spatstat/spatstat.pdf
3. https://mgimond.github.io/Spatial/point-pattern-analysis.html
4. https://training.fws.gov/courses/references/tutorials/geospatial/CSP7304/documents/PointPatterTutorial.pdf
5. https://www.seas.upenn.edu/~ese502/NOTEBOOK/Part_I/3_Testing_Spatial_Randomness.pdf

Software Description:

Data Preprocessing:

- Language: C#
- System.Xml and System.Xml.Linq: XML parsing libraries

CSV Data Formatting and Polygon/Centroids Calculations:

- Language: Python
- Pandas: Dataframe handling and manipulation
- Shapely: Creating Polygons and computing centroids

Spatial Data Clustering:

- Language: R
- Spatstat: Spatial Analysis including Ripley's K function, Spatial Exploratory Data Analysis.
- Dplyr: Dataframe processing and manipulation
- Ggplot: Data visualization