Master Universitario en Ciencia de Datos

Practical Application 3

Spatial Statistics

El Abbassi Widad

January 14th,2020

I. Introduction

This article is revolving around explanatory data analysis of spatial statistics using the spatstat package available in R. To do so, we will start by studying the intensity, then the Density based analysis, spacing and finally the distance based analysis

II. Problem description

This article's goal is to use the different functions available for explanatory data analysis in order to investigate the intensity of a point process, the CSR condition, and the dependence between points of the data to finally have a better understanding of the underline spatial process.

III. Methodology

1. Analysis of the Data

This dataset is a SpatialPointsDataFrame representing the location of Starbucks in Massachusetts. The data objects consist of three spatial data layers:

- **starbucks:** A ppp point layer of Starbucks stores in Massachusetts;
- **ma:** An owin polygon layer of Massachusetts boundaries;
- **pop:** An im raster layer of population density distribution.

IV. Results

1.Density Based Analysis:

As a first step in this work, we will begin by loading the Starbucks Dataset:

```
> str(starbucks)
List of 5
 $ window    :List of 4
  ..$ type  : chr "rectangle"
  ..$ xrange: num [1:2] 648032 917741
  ..$ yrange: num [1:2] 4609785 4748107
  ..$ units :List of 3
  .. ..$ singular  : chr "unit"
  .. ..$ plural    : chr "units"
  .. ..$ multiplier: num 1
  .. ..- attr(*, "class")= chr "unitname"
  ..- attr(*, "class")= chr "owin"
 $ n         : int 171
 $ x         : num [1:171] 917741 911147 902987 876188 875868 ...
 $ y         : num [1:171] 4637151 4628510 4628982 4616741 4616719 ...
 $ markformat: chr "none"
 - attr(*, "class")= chr "ppp"
```

After this, we will bind the Massachusetts boundary polygon to the Starbucks point feature object using the Window and plot it to ensure that it was properly defined:

The length unit of this data is in meters which is not suitable for this scale of analysis. Therefore, we will rescale the spatial objects to kilometer in order to obtain more readable results as shown below:

```
> summary(starbucks.km)
Planar point pattern:  171 points
Average intensity 0.008268627 points per square km

Coordinates are given to 4 decimal places

Window: polygonal boundary
single connected closed polygon with 128 vertices
enclosing rectangle: [623.1572, 921.9238] x [4602.709, 4756.659] km
                     (298.8 x 154 km)
Window area = 20680.6 square km
Unit of length: 1 km
Fraction of frame area: 0.45
```

Now, how is the intensity? , is it uniform or does it vary from location to another,

```
> intensity(starbucks.km)
[1] 0.008268627
```

✓ **Investigating intensity of a point process**

The classic approach is to define manually the number :

```
Q<- quadratcount(starbucks.km, nx= 4, ny=5)
```

But the problem in this case will be that the choice of quadrat numbers and quadrat shape will influence the measure of local density.

⇨ For Quadrat counting tests for CSR (**quadrat.test**) :

H0 : homogeneous Poisson process (CSR)

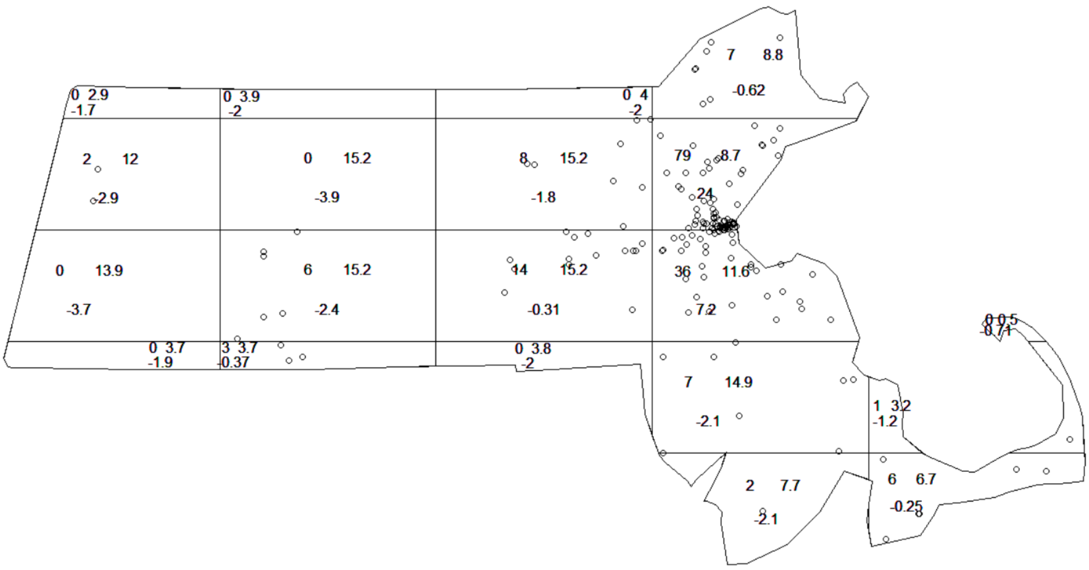
Ha : inhomogeneous Poisson process. Under H0,

Result : The p-value is way too small, H0 is rejected.

```
> quadrat.test(starbucks.km)

        Chi-squared test of CSR using quadrat counts

data:  starbucks.km
X2 = 695.5, df = 19, p-value < 2.2e-16
alternative hypothesis: two.sided

Quadrats: 20 tiles (irregular windows)
```

To be honest, I passed a quite fair amount of time trying manually to find some x,y parameters that could give me as a result a homogeneous Poisson process, but in all the cases, the obtained result is always p**-value** < **2.2e-16.**

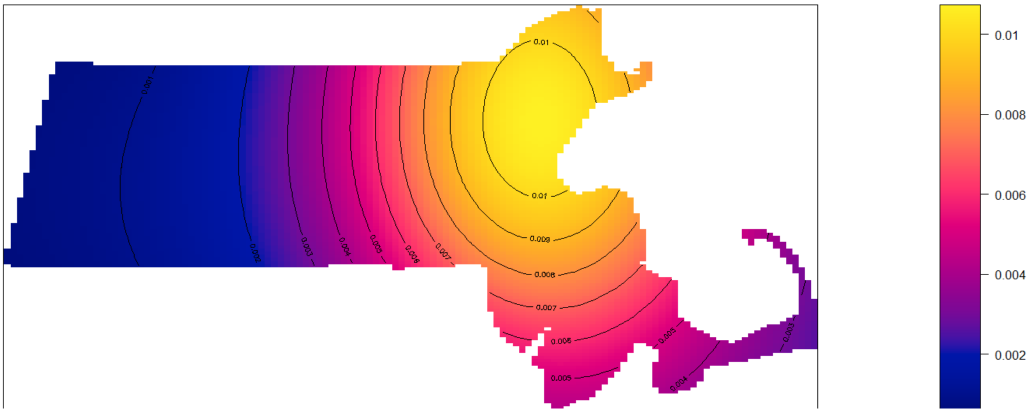
Plot of quadrat.test : we can see below that the gap between the observed counts and the expected counts are quite large in some specific ares, which can also be concluded using the Pearson residuals to see if whether the observed frequency in some cell is higher/lower than the fitted value.



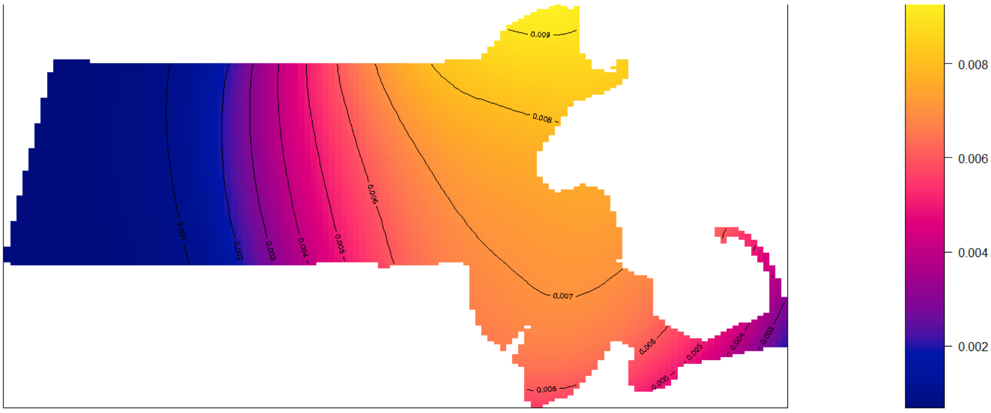
Note : Another alternative will be the use of population density raster to define non-uniform quadrats.

- ✓ **Kernel Density**

After this, we can compute the kernel intensity estimate of the point pattern with sitting a bandwidth of 50 km and a Gaussian Kernel.



And if we try : kernel="epanechnikov", we get:

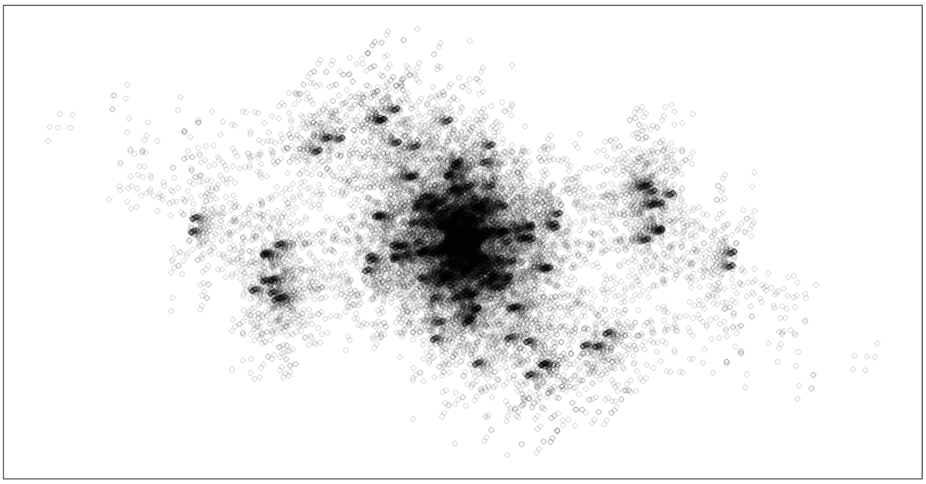


Although, the Gaussian kernel is the one that is widely use due to its infinitely many (nonzero) unlike the Epanechnikov kernel that has only 3 derivatives before it's identically zero.

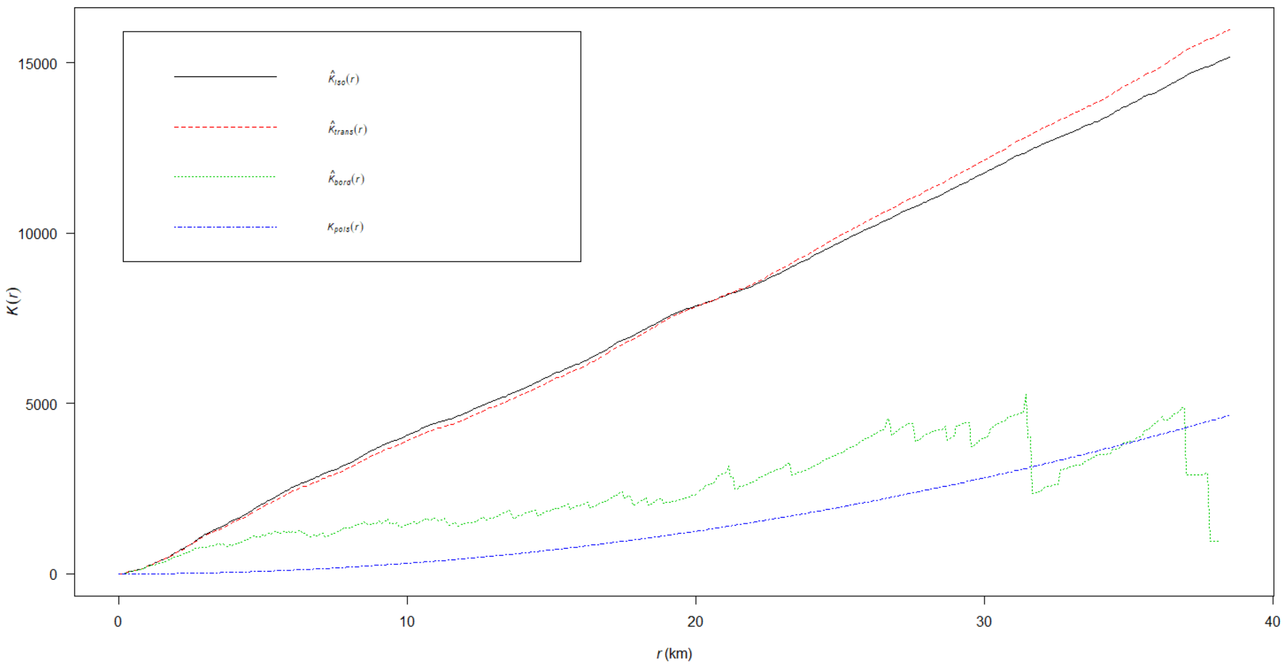
2. Distance Based Analysis:

- ✓ **Correlation**

For this part, fryplot will be a good first step to assess dependence:



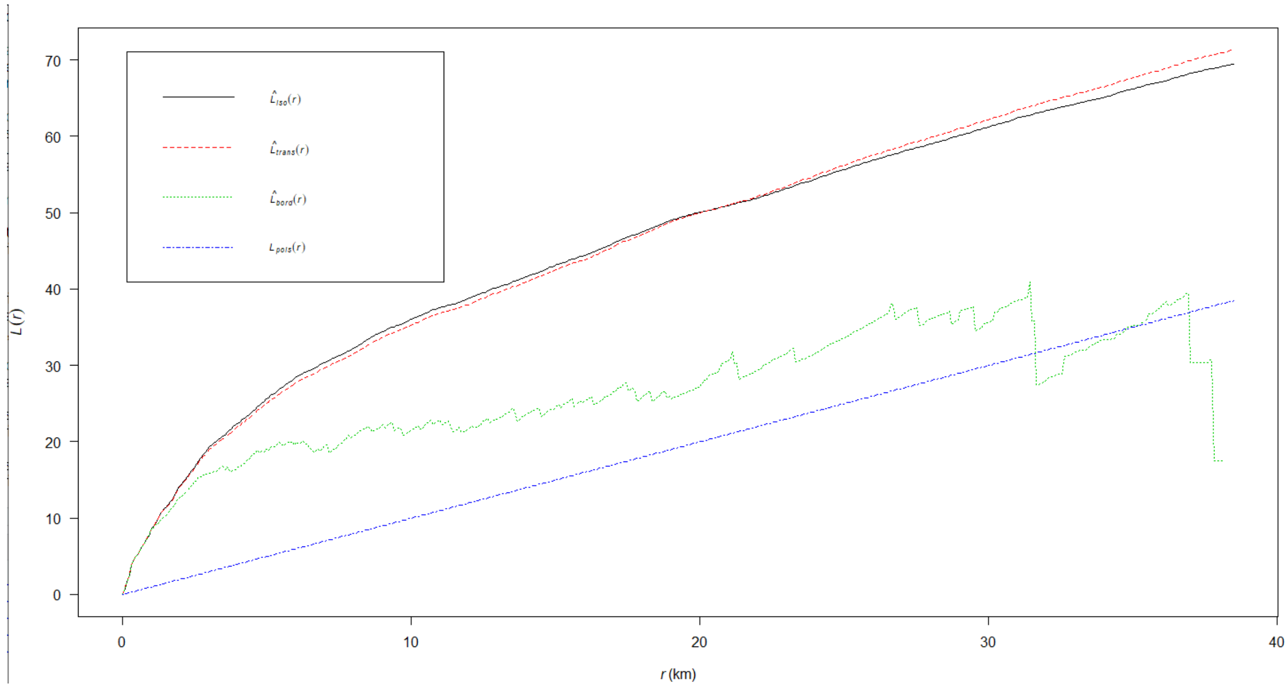
After this, will go directly to Ripley's K function and plot it as shown below:



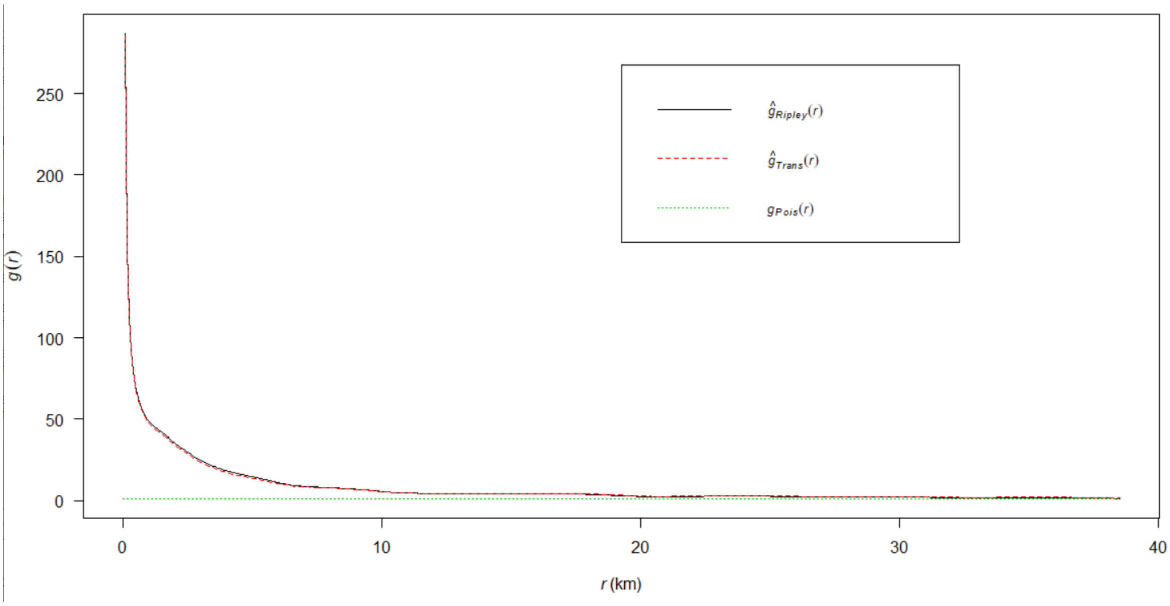
The plot returns different estimates of K depending on the edge correction chosen.

- **Kpois** represents the theoretical K function under the null hypothesis that the points are completely randomly distributed (CSR/IRP) and since most of them are at top of the theoretical line for most r (<u>except the isotropic-corrected estimate of K(r) which falls of directly after 30 km </u>) , they suggest that the pattern is not regular.

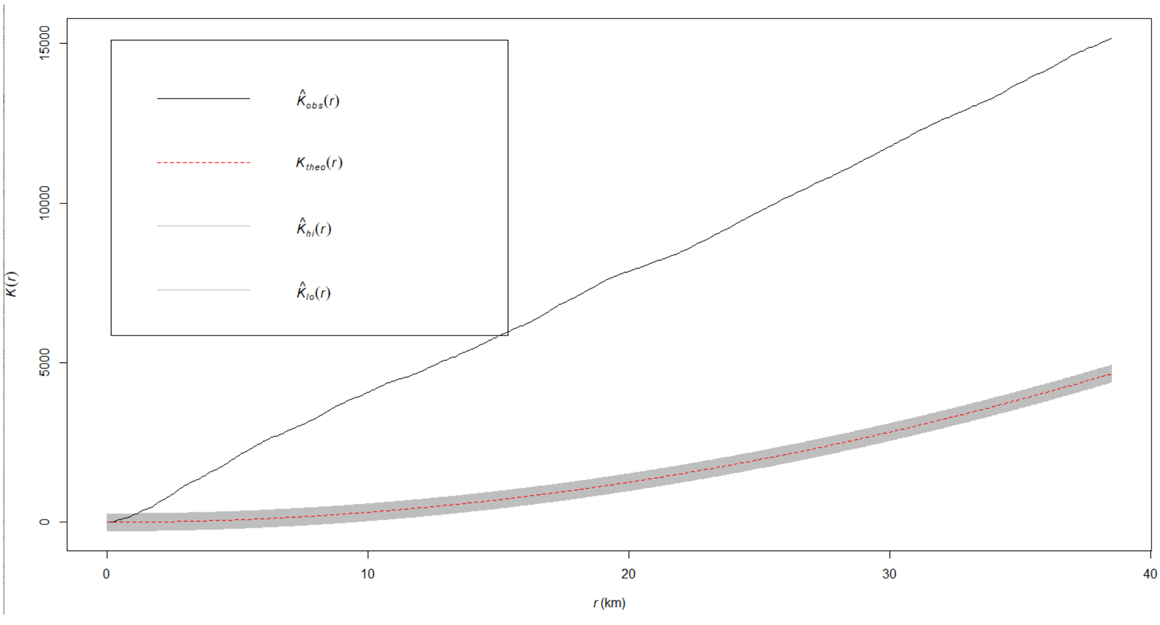
We can also use the L-function which transforms the theoretical K-function of the CSR to the straight line as shown below



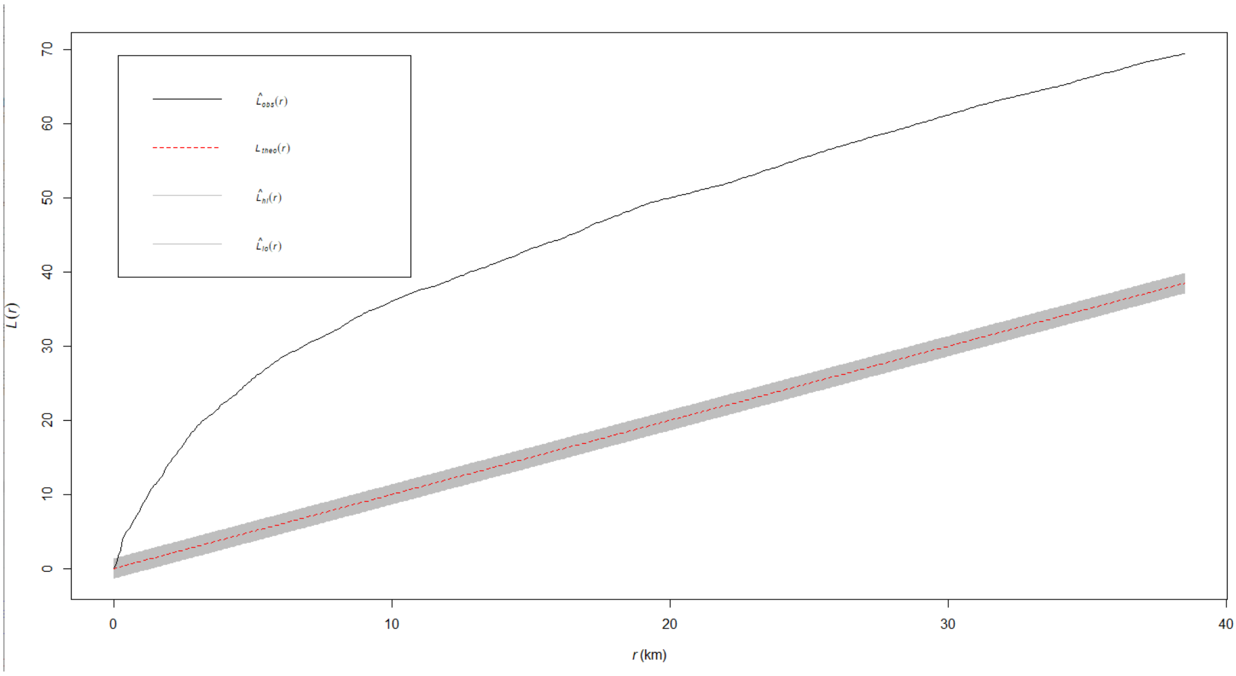
The pair correlation function is also used for comparison: so if the observed g is greater than gpois we can expect more clustering than expected and where the observed g is less than gpois we can expect more dispersion than expected. [Poisson process. g(r) > 1 suggests clustering; g(r) < 1 suggests regular pattern].



Another alternative will be the hypothesis testing with simulation envelopes, here we used the global envelopes.

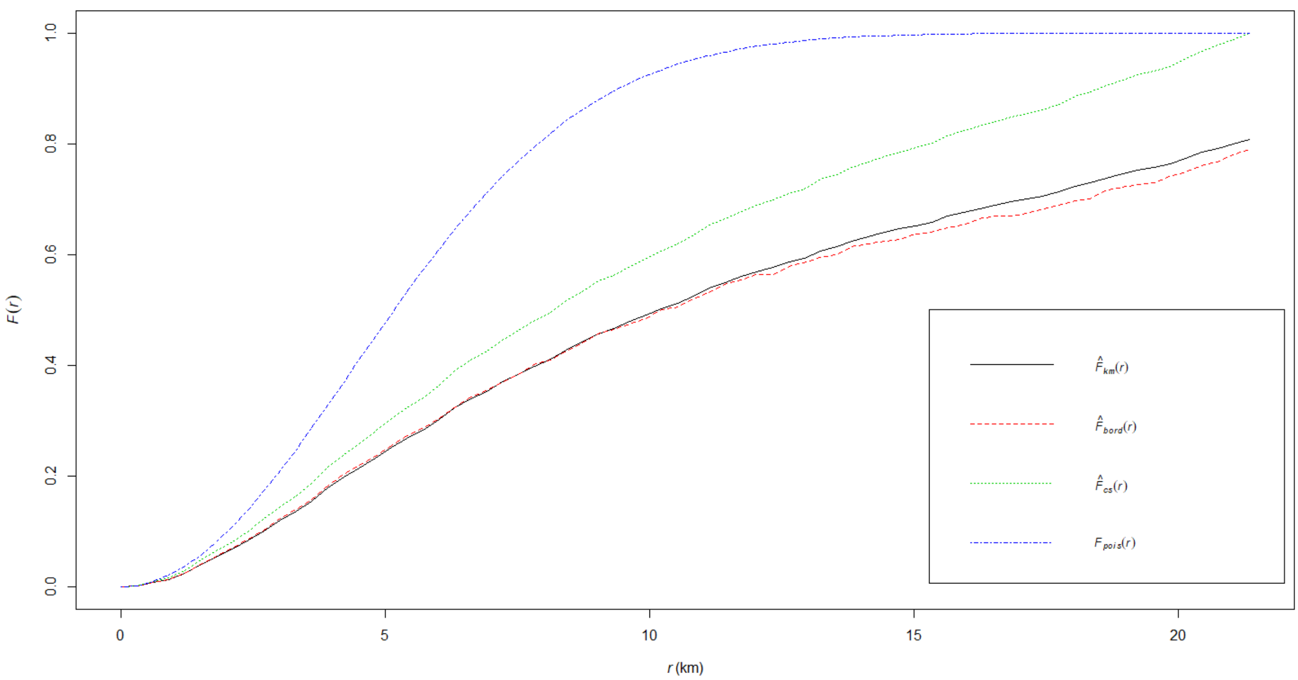


We can get a better results if we opt for L function, but in our case not much difference was observed. Also, We can interpret pointwise envelopes as significance tests if we choose the distance r beforehand.

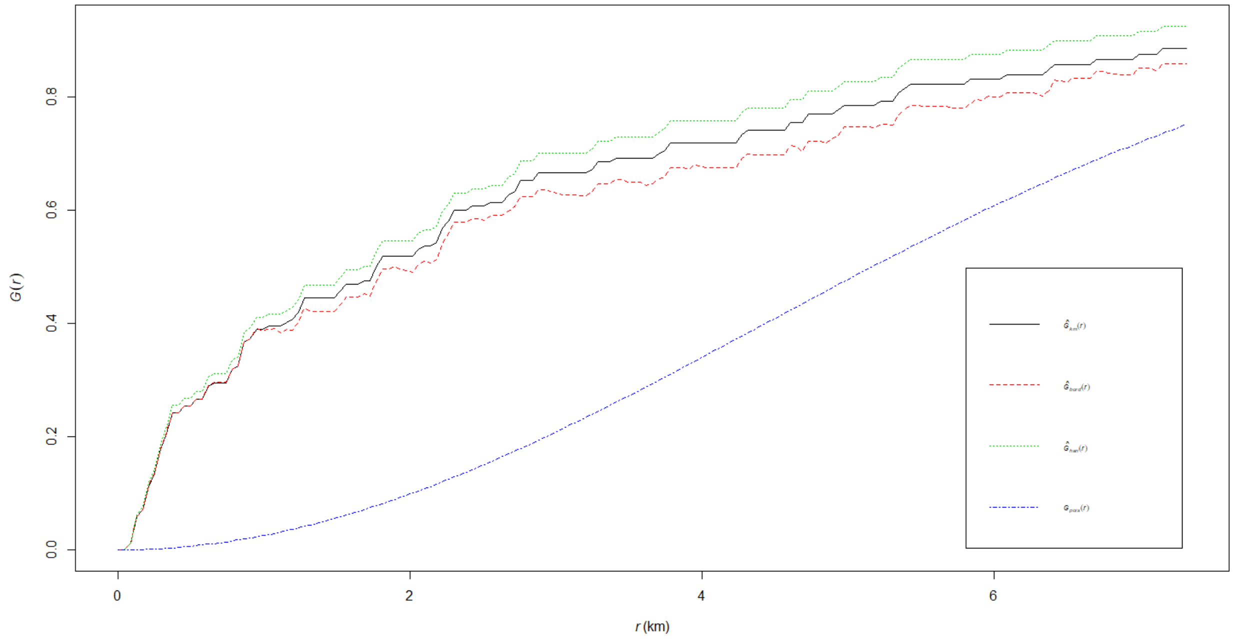


- ✓ **Spacing :**

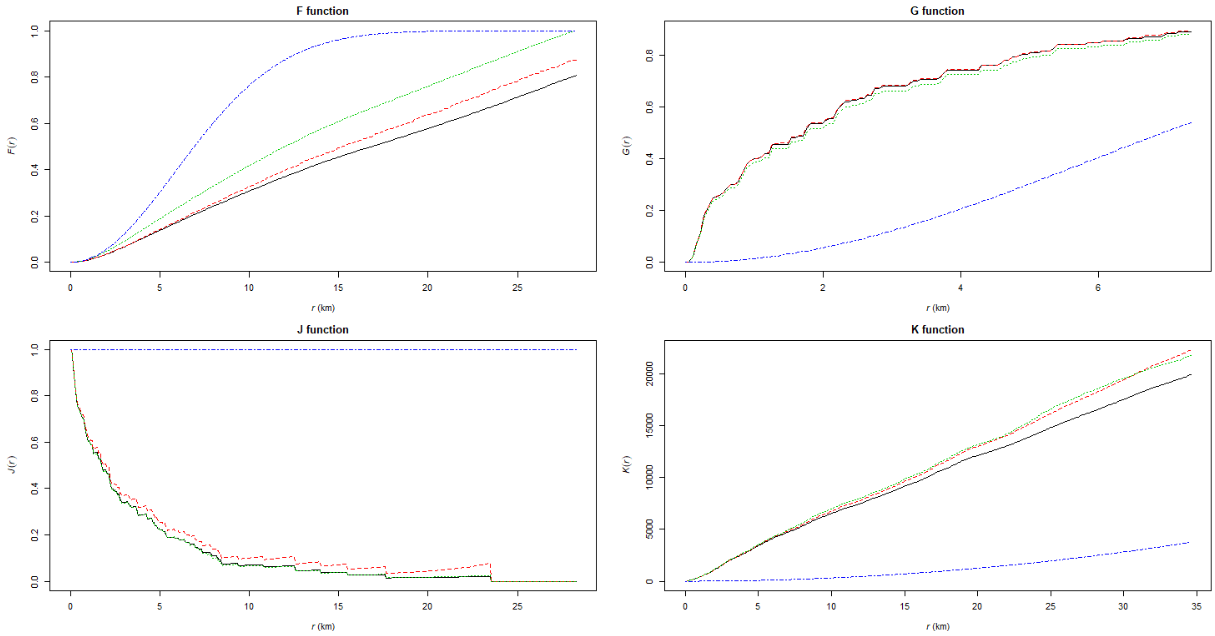
Using Fest() computes an estimate of the empty space function, the results show that it is more of a clustered pattern situation.



Event to event- or Gest() function is also analogous to Fest(), but its interpretation is the same as Kest. In our case a clustered pattern.



And if we desire to have a complete view, we can compute a summary for all four functions :



V. Discussions:

In this work, we carried an inclusive explanatory data analysis to better understand the spatial point pattern of Starbucks stores in Massachusetts, starting from the inspection of the data till measuring the spacing.

We've also seen that the results of correlation plots, and their interpretation will be incomplete if we didn't bind it with measuring the shortest distances to provide more complementary information and which was also, done and proved by many reference manuals I've came crossed.

This analysis work is just a simple warmup towards fitting a realistic model to this point pattern, which in this case will have to take into account the covariates and the existence of those marks before beginning the study.

Conclusion

To sum up, we must state that this process allowed to evaluate and test the different methods and summary functions such as K, F, G, and J functions. Most of the reference manuals and researches recommend to perform all of them to see the difference in the point pattern with different estimators instead of choosing just one to work with to inspect the interpoint interaction.

References

[1]https://moodle.upm.es/titulaciones/oficiales/pluginfile.php/2575283/mod_resource/content/1/spatstat.pdf

[2] https://rdrr.io/cran/GmAMisc/f/README.md

[3] https://cran.r-project.org/web/packages/spatstat/vignettes/replicated.pdf

[4] https://rdrr.io/cran/GmAMisc/f/README.md

[5] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.8464&rep=rep1&type=pdf

[6]https://training.fws.gov/courses/references/tutorials/geospatial/CSP7304/documents/PointPatterTutorial.pdf