Geospatial Analysis using R

Introduction to data-driven decision making

Ana J. Alegre, Cristian Silva. iGISc. November 4, 2021.

Agenda

Workshop day 2

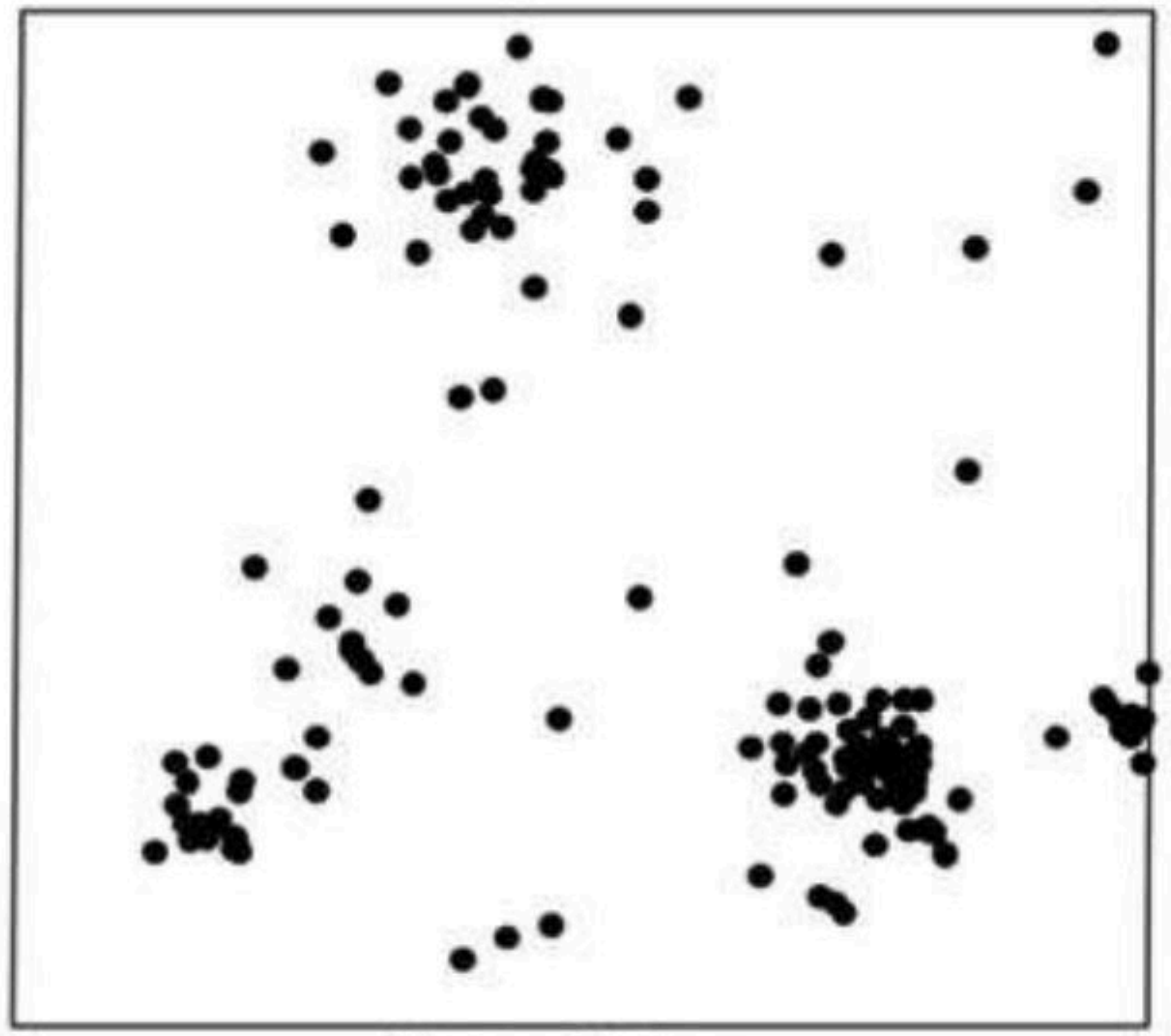
1. Q&A
2. Spatial point patterns
3. Hands-on session
4. Break
5. K-Means clustering
6. Hands-on session
7. Break
8. Data visualization and decision making
9. Hands-on session
10. Conclusions
11. Q&A

Questions?

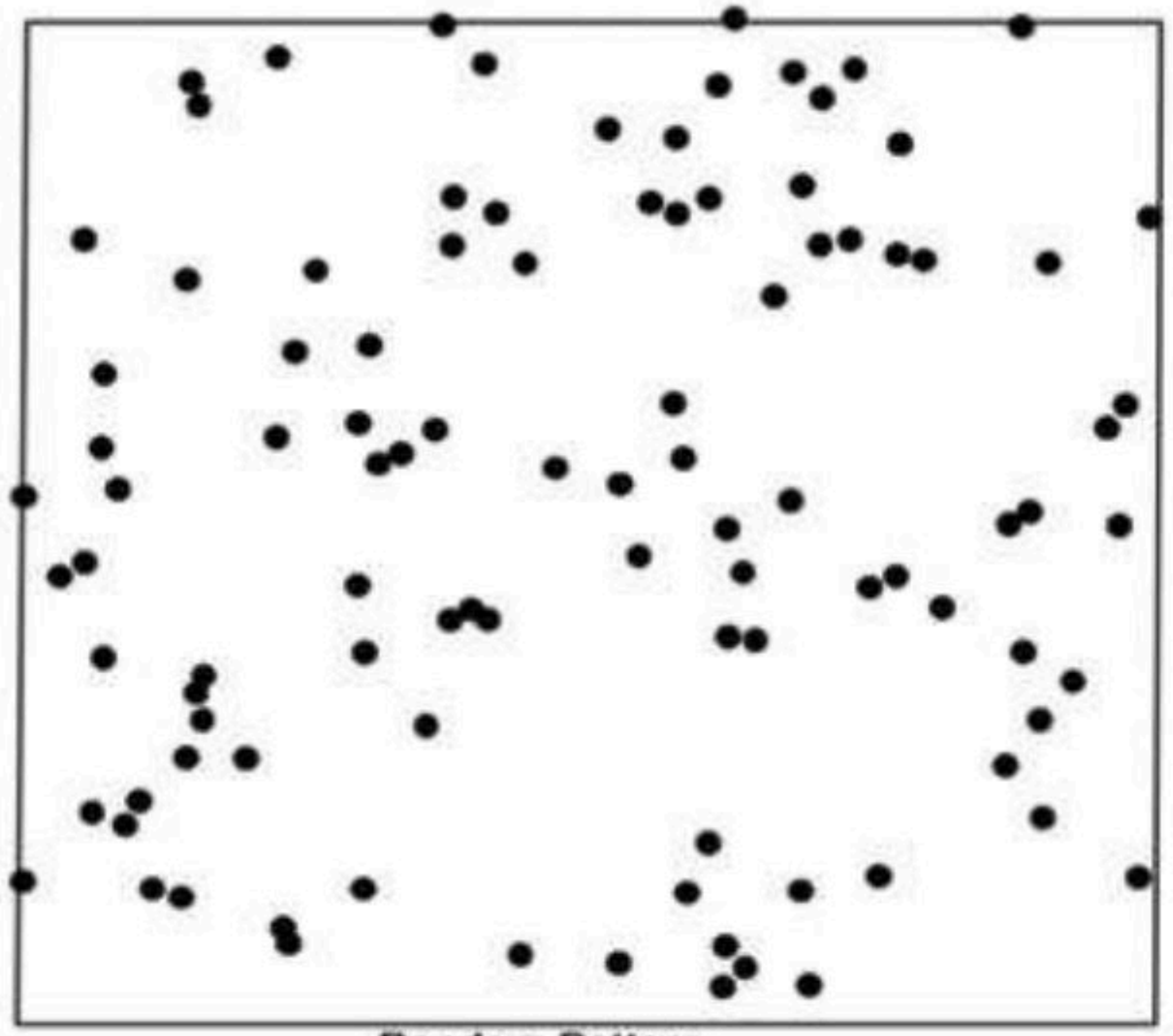
Spatial point patterns

Point Pattern Analysis

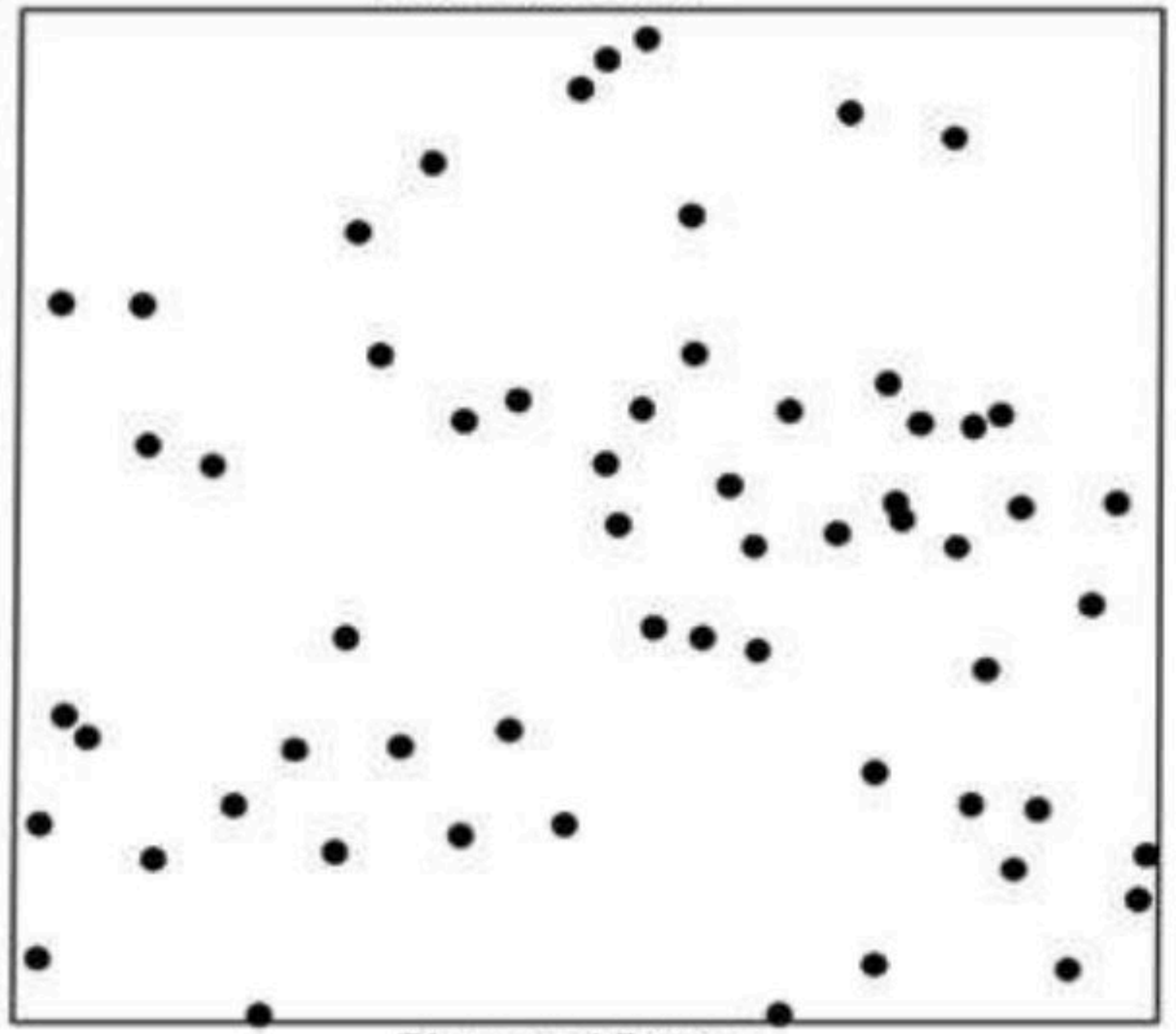
Three basic pattern structures exist:



Clustered Pattern



Random Pattern



Dispersed Pattern

Reference taken from Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

Exploring patterns, distributions, and trends

1. Quadrat Count

2. Nearest Neighbor Approach

3. K-Function Approach

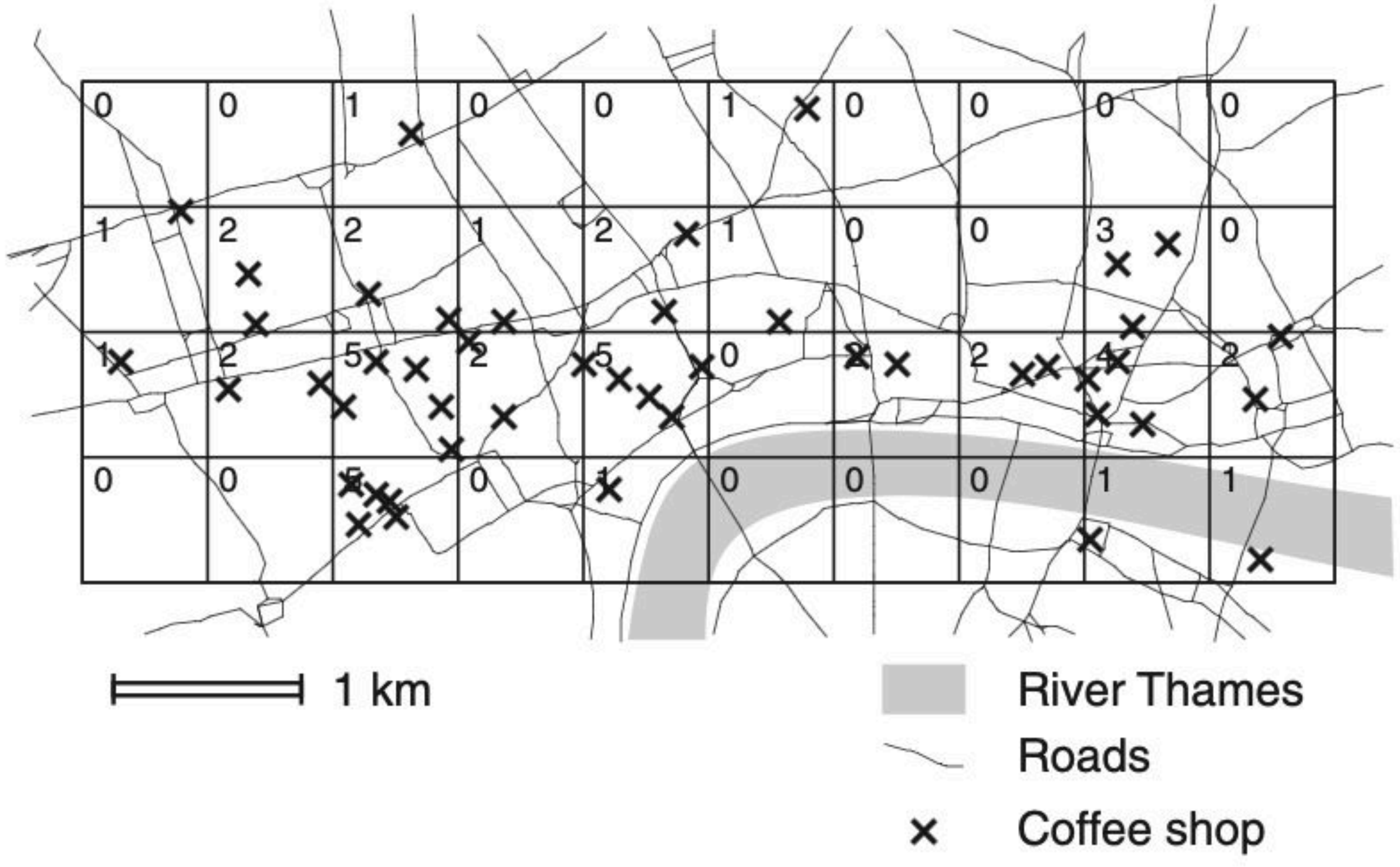
4. Kernel Estimation Approach

Reference taken from Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

Quadrat Count

Example: Coffee shops in central London



Quadrat Counts and Calculation of the Variance for Coffee Shop Pattern

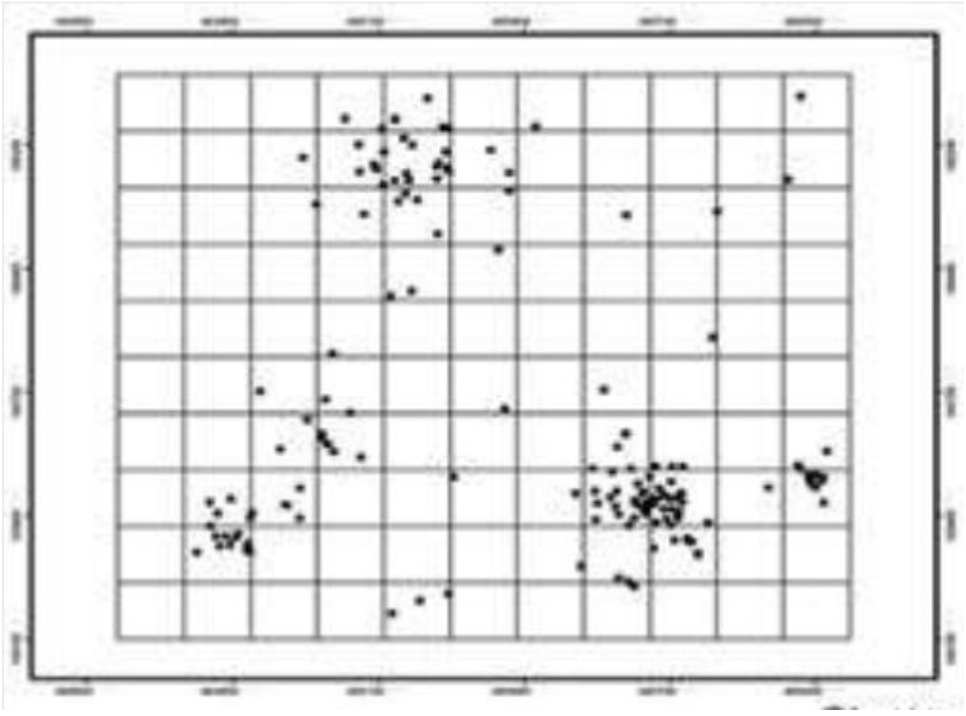
No. of events, K	*No. of quadrats, X*	$K-\mu$	$(K-\mu)^2$	$X(K-\mu)^2$
0	18	−1.175	1.380625	24.851250
1	9	−0.175	0.030625	0.275625
2	8	0.825	0.680625	5.445000
3	1	1.825	3.330625	3.330625
4	1	2.825	7.980625	7.980625
5	3	3.825	14.630625	43.891875
Totals	**40**			**85.775000**

Reference taken from O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons and, Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

Quadrat Count

Cluster Pattern

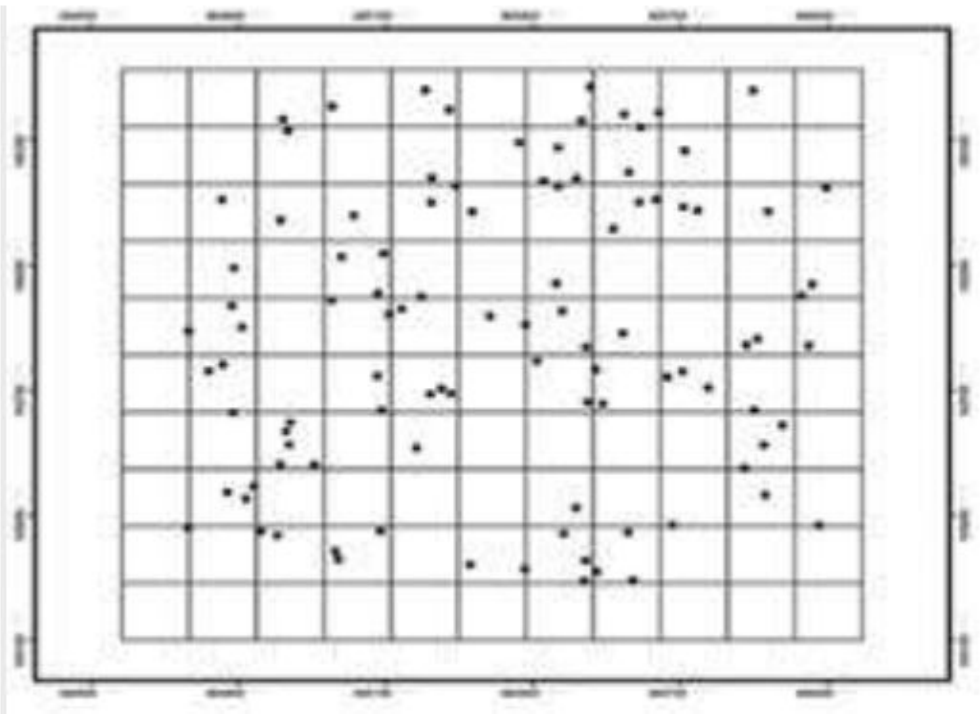


$$\bar{X} = \frac{\sum X_j f_j}{\sum f_j} = \frac{232.2}{110} = 2.1109$$

Chi-square statistic $\chi^2 = 2046.77/2.1109 = 969.61$

$P\text{-}value = 4.8248\text{E-}138$

Random Pattern

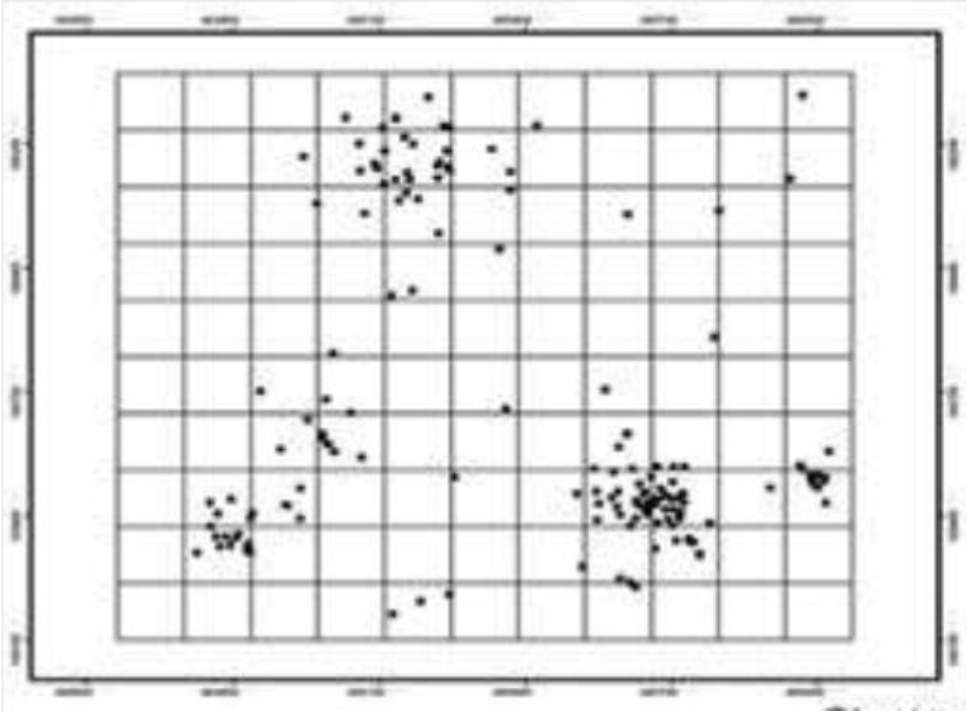


$$\bar{X} = \frac{\sum X_j f_j}{\sum f_j} = \frac{99}{110} = 0.90$$

Chi-square statistic $\chi^2 = 119.90/0.9 = 133.22$

$P\text{-}value = 0.057389389$

Dispersed Pattern



$$\bar{X} = \frac{\sum X_j f_j}{\sum f_j} = \frac{68.2}{110} = 0.62$$

Chi-square statistic $\chi^2 = 62.94/0.62 = 101.51$

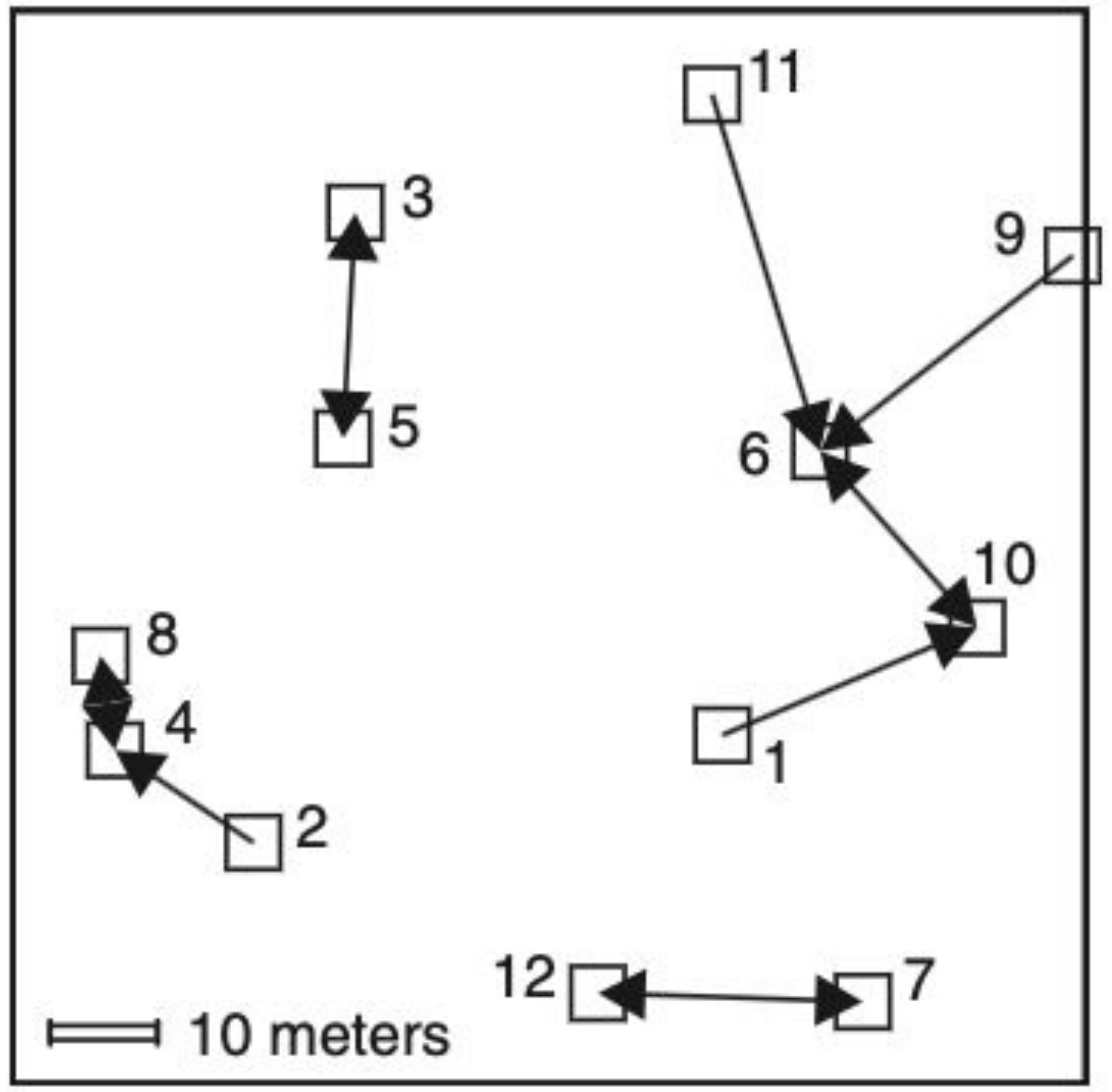
$P\text{-}value = 0.682243941$

Reference taken from O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons and, Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

Nearest Neighbor Approach

Example: Distance to the nearest neighbor for a small point pattern



Calculations for the Nearest-Neighbor Distance for the Point Pattern

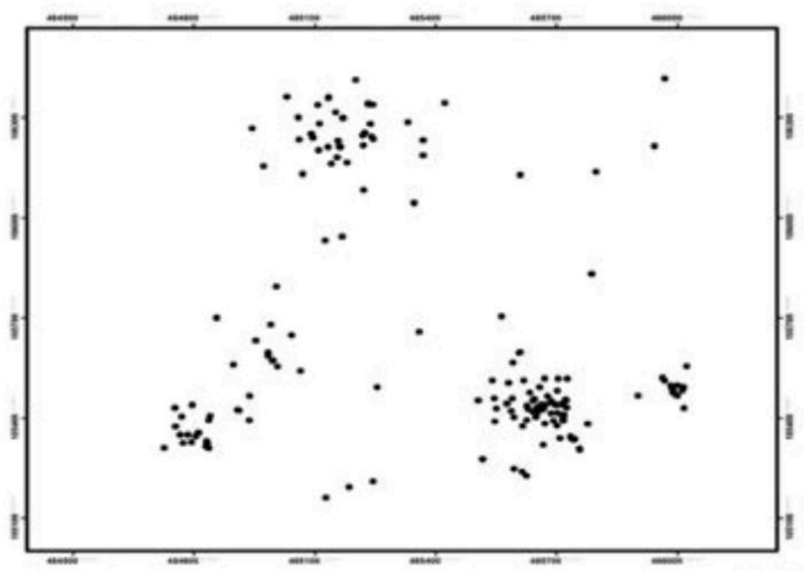
Point	*X*	*Y*	*Nearest neighbor*	D_{min}
1	66.22	32.54	10	25.59
2	22.52	22.39	4	15.64
3	31.01	81.21	5	21.11
4	9.47	31.02	8	9.00
5	30.78	60.10	3	21.14
6	75.21	58.93	10	21.94
7	79.26	7.68	12	24.81
8	8.23	39.93	4	9.00
9	98.73	77.17	6	29.76
10	89.78	42.53	6	21.94
11	65.19	92.08	6	34.63
12	54.46	8.48	7	24.81

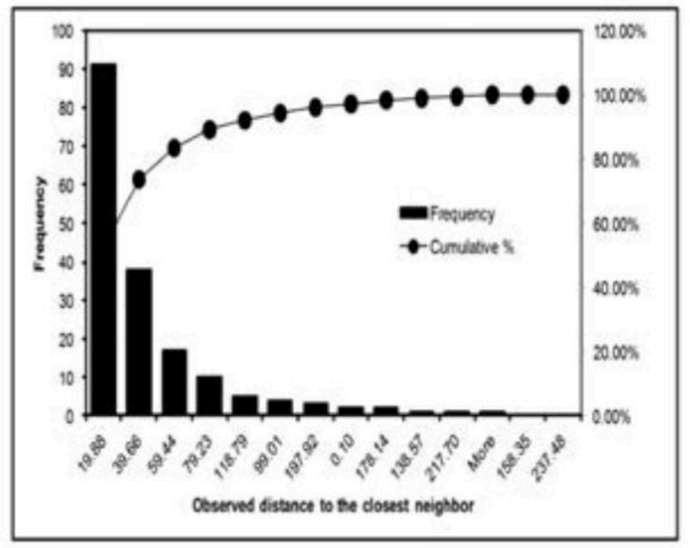
Reference taken from O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons and, Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

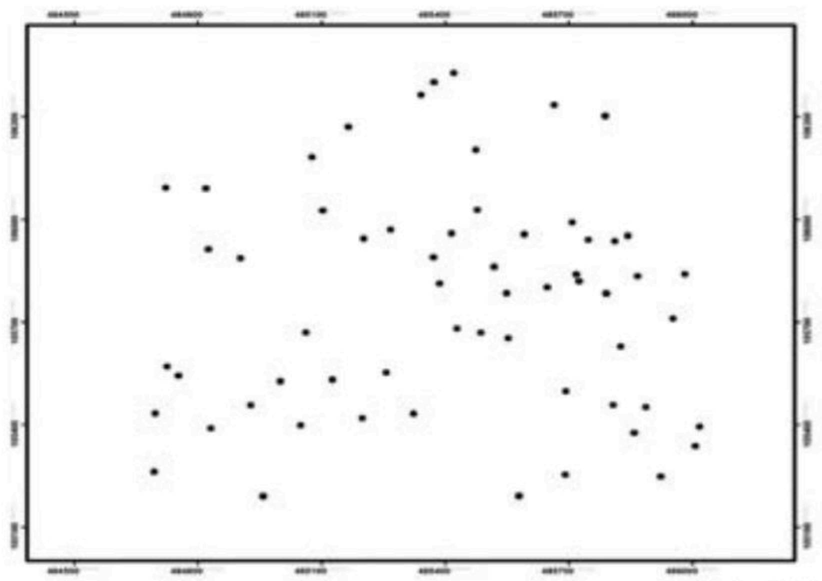
Nearest Neighbor Approach

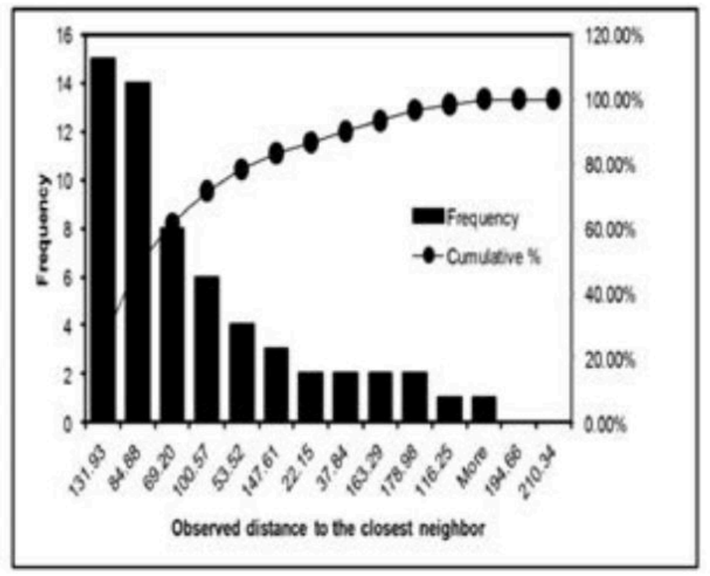
Cluster Pattern



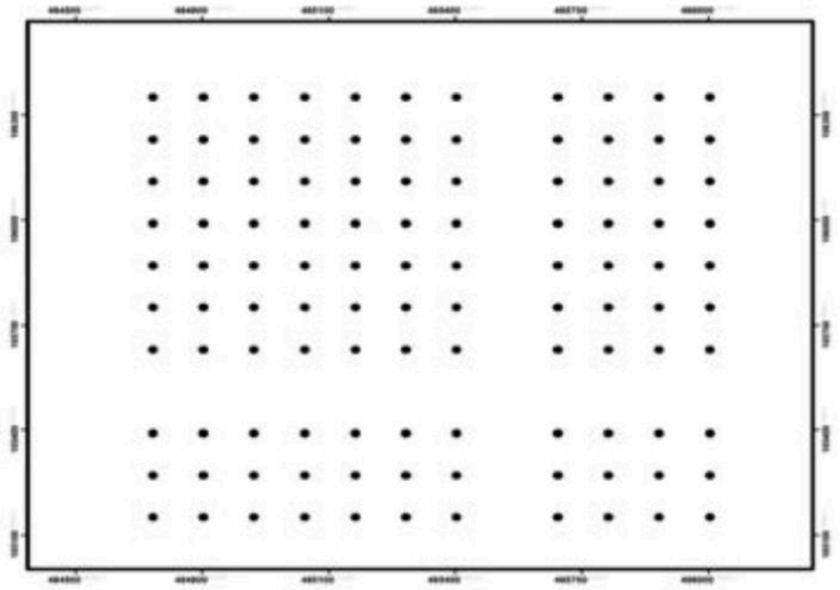


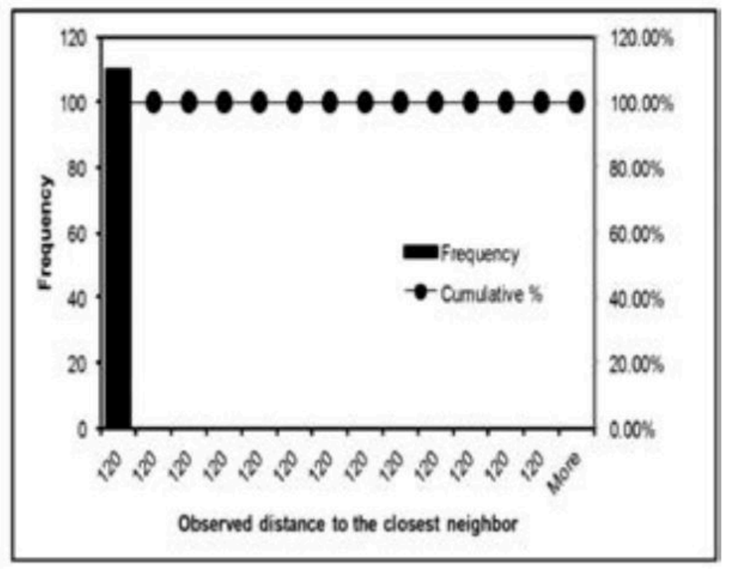
Random Pattern





Dispersed Pattern





Worktable for Nearest Neighbor Analysis for Potential Nesting Sites Showing Results for Three Basic Distributions

	Observed Mean Distance	Expected Mean Distance	Nearest Neighbor Ratio (*R*)	*z*-score	*p*-value
Clustered	31.671	54.30851	0.58316	−10.549206	0.00000
Dispersed	120	68.5	1.751825	15.08495	0.00000
Random	95.231	92.749438	1.026753	0.396444	0.691778

Reference taken from O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons and, Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

K-Function Approach

There are six major steps in conducting a *K*-function:

1. Determine/compare the observed and expected *K*. The observed *K* is obtained through the construction of a circle around each point event (*i*), counting the number of other events (*j*) within the radius (*h*) of the circle, and repeating the same process for all other events (*i*).

2. Next, determine the average number of events within successive distance bands. Find the overall point density for the study area. The observed *K* is the ratio of the numerator to the density of events. This can then be compared to the expected *K,* which is a random pattern, $K(h) = \pi h^2$.

3. Transform $K(h)$ estimates into a square root function to make it linear $L(d)$.

4. Determine the confidence envelope by estimating min $L(d)$ and max $L(d)$ values from several simulations at $\alpha = 0.05$ under the null hypothesis of random distribution.

5. Plot $L(d)$ estimates on a graph to reveal if any clustering occurs at certain distances.

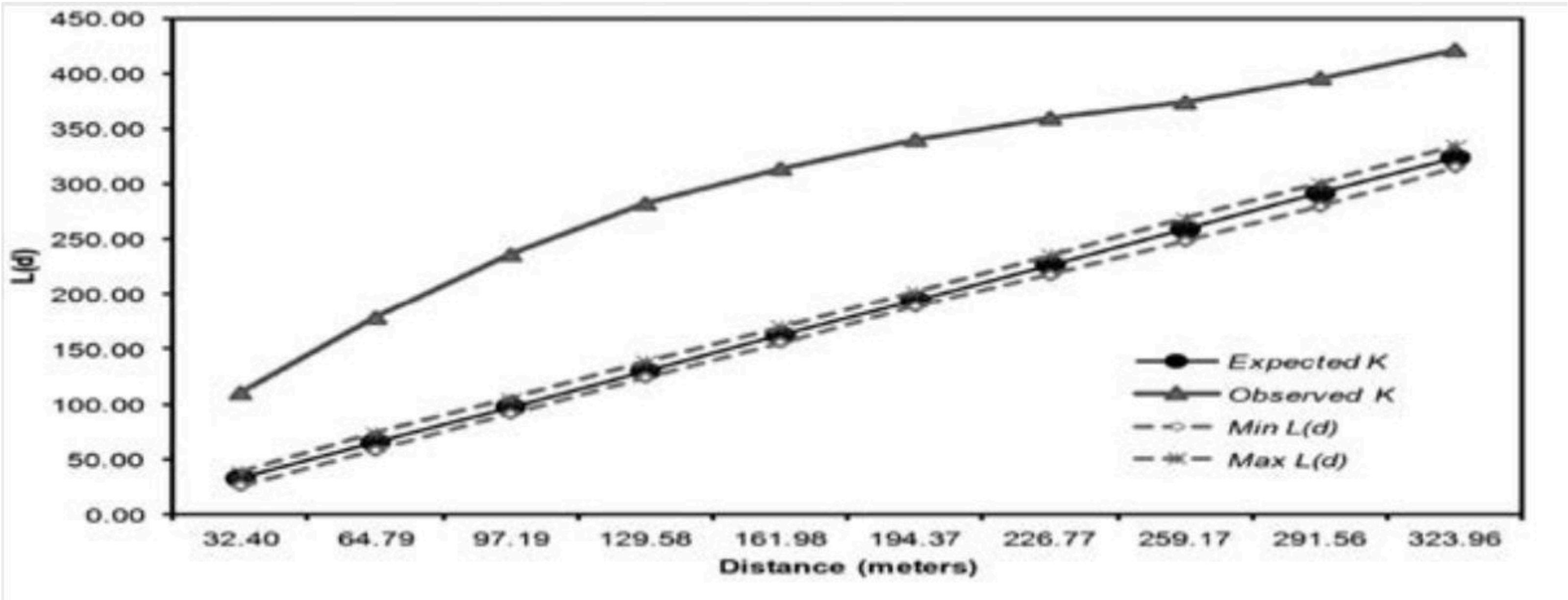
6. Interpret the results.

Reference taken from O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons and, Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

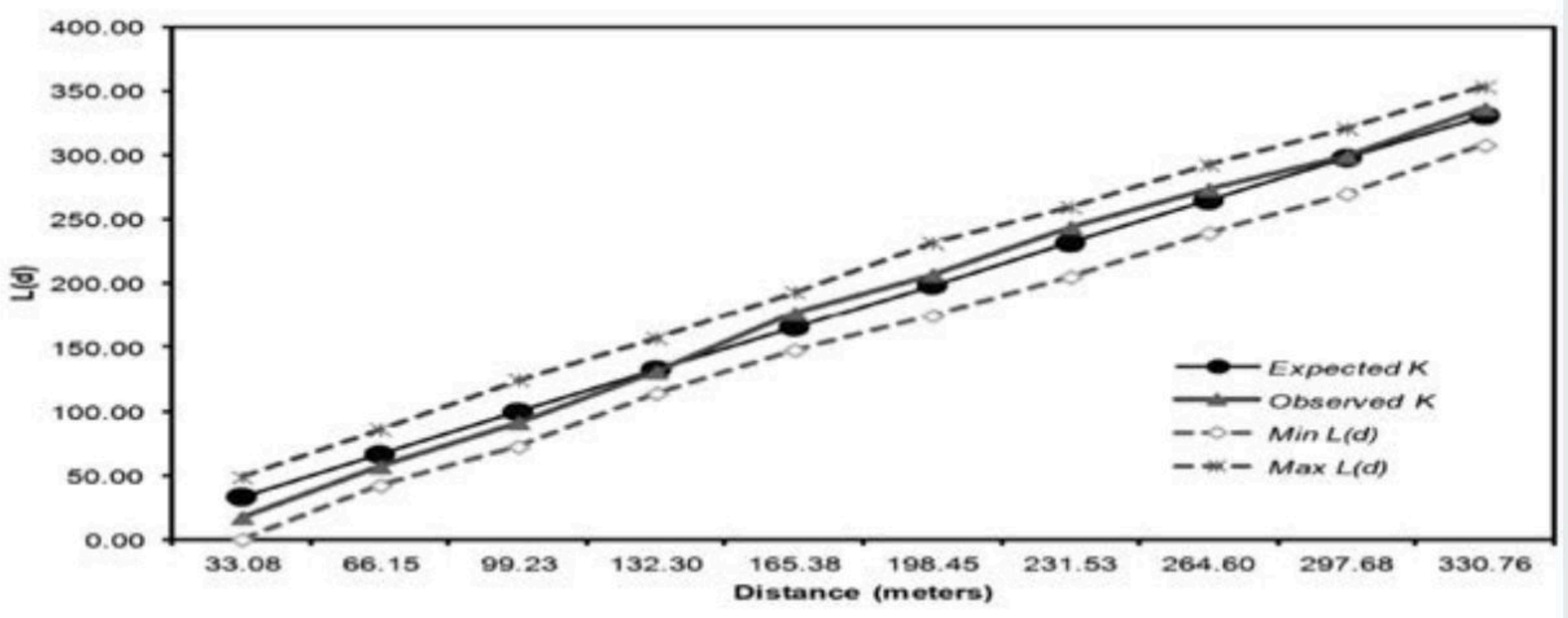
Point Pattern Analysis

K-Function Approach

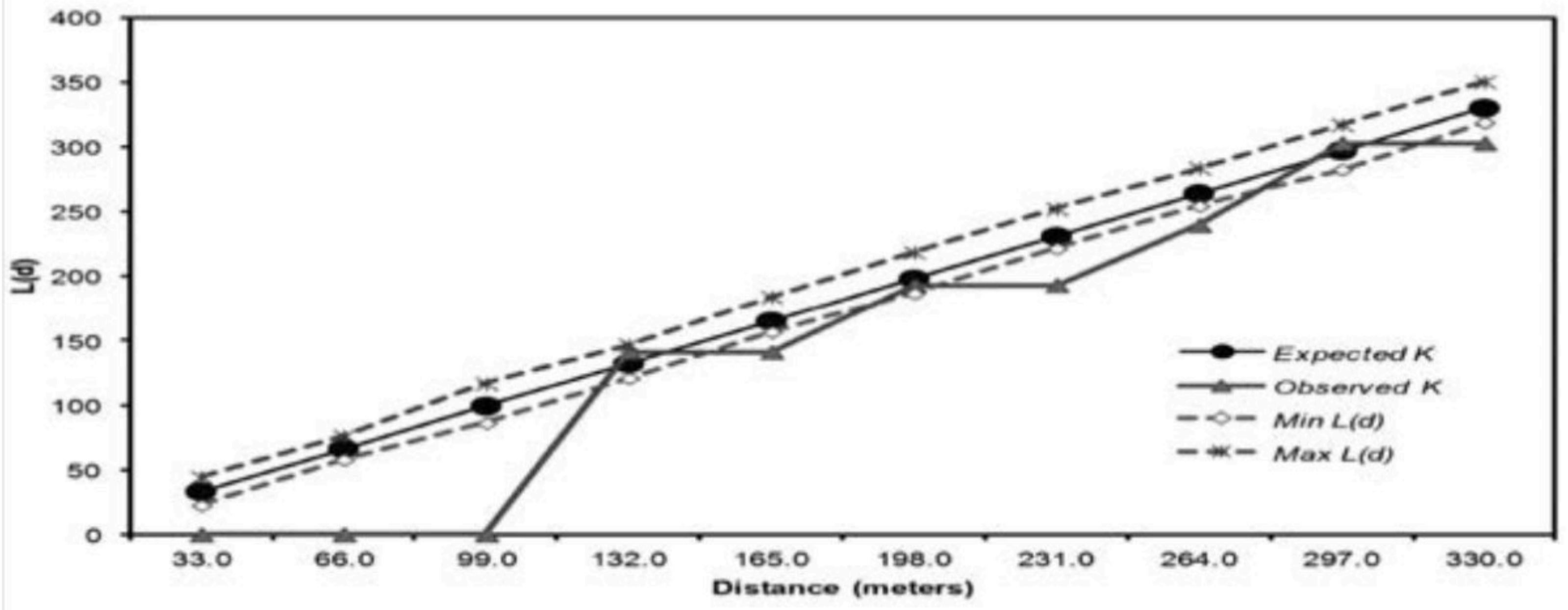
Cluster Distribution



Random Distribution



Dispersed Distribution



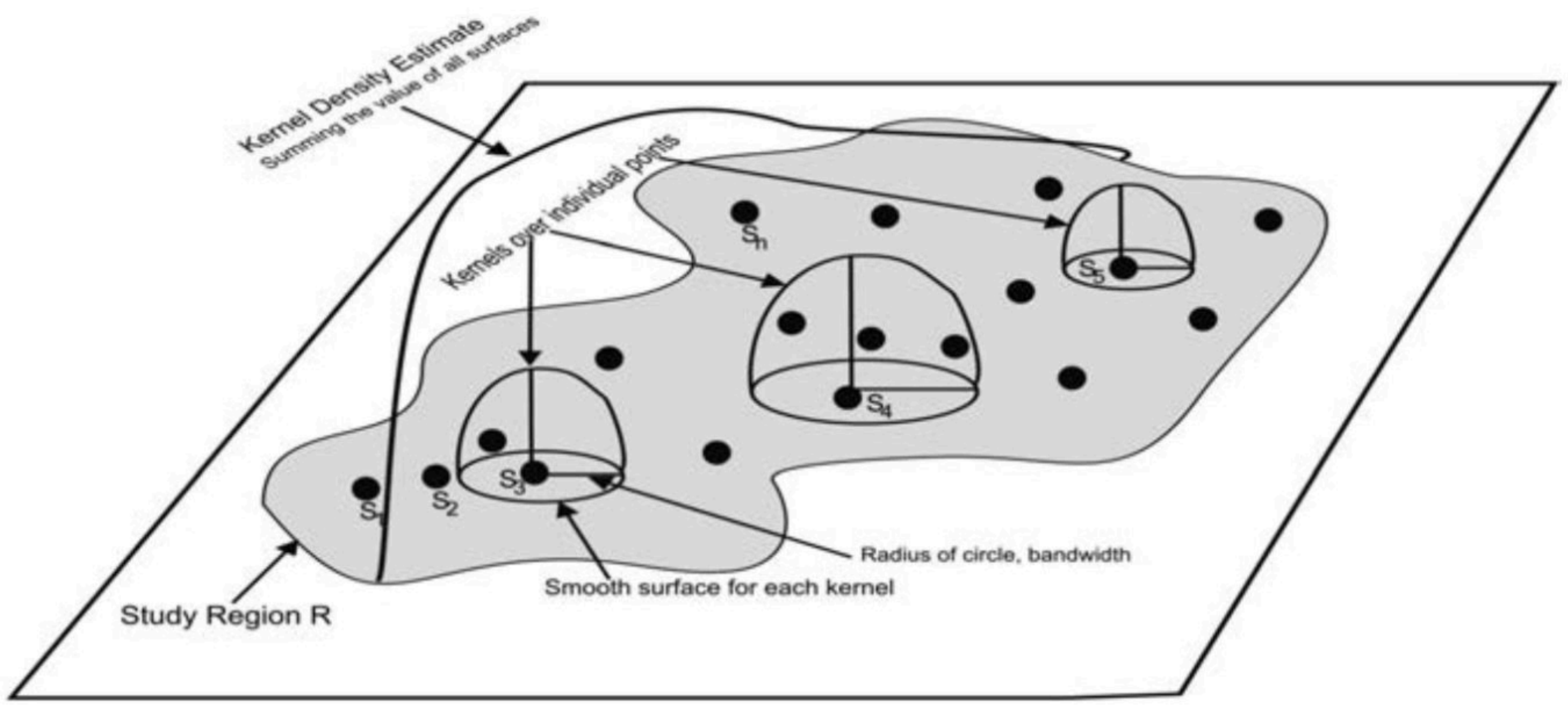
Plots of L(d) values for three dispersion patterns of an ecological study obtained from the K-function analysis. The findings were generated on the basis of 99 simulations under the null hypothesis of random distribution.

Reference taken from O'sullivan, D., & Unwin, D. (2003). *Geographic information analysis*. John Wiley & Sons and, Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Point Pattern Analysis

Kernel Estimation Approcah

The kernel estimation method applied to study region R



$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right)$$

τ is the bandwidth (a smoothing parameter, i.e., radius of the circle)

k(·) is the kernel

$s - s_i$ is the distance between two events (point s and s_i)

Reference taken from Oyana, T. J. (2020). Spatial Analysis with R: Statistics, Visualization, and Computational Methods. CRC press.

Hands-on session

Break

Clustering Analysis

Clustering Analysis

Objectives:

1. It seeks to solve a classification problem by discovering the number and composition of groups in a universe/sample of observations.

2. It uses methods of similarity/dissimilarity, or proximity, or distances between them.

3. Assign observations (e.g. places) to homogeneous groups in such observations according to variables of interest to form heterogeneous groups among them.

Reference taken from Clustering Analysis Class, Dr. Carlos Javier Vilalta P. (CentroGeo)

Clustering Analysis

The 2 most commonly used techniques are:

1. **K-means:** divides the sample and assigns observations to groups (K) based on the distance to their centroids until the minimum number in the sum of squares of these distances/errors (*minimum SSE*):

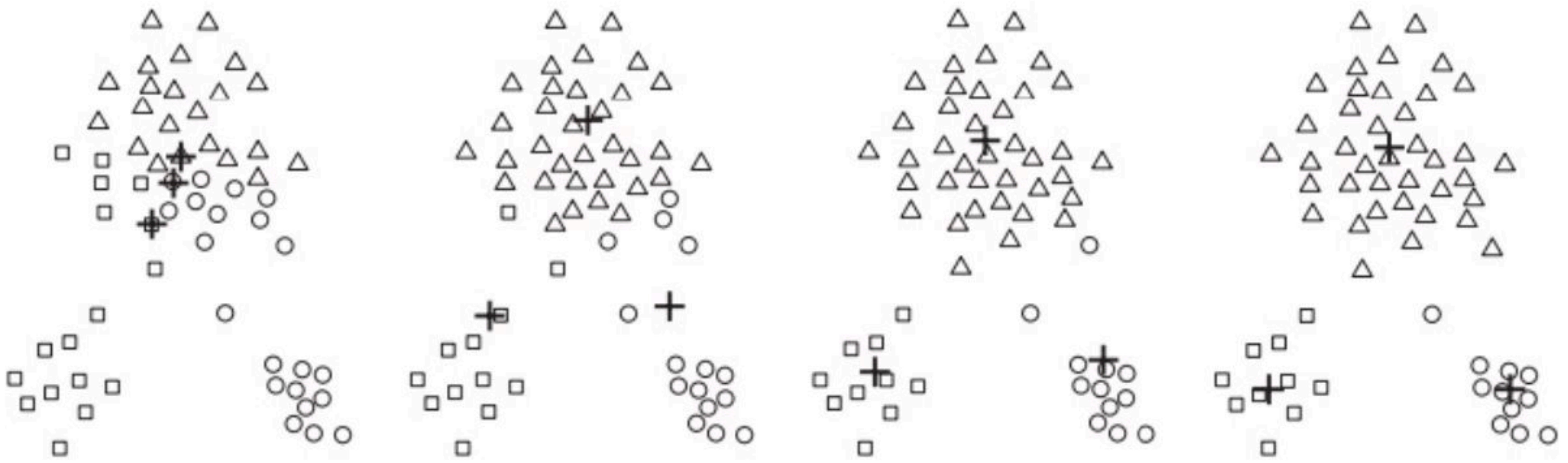
$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

2. **Hierarchical:** it starts by grouping the most similar observations and grouping these "groups" (k) into other groups until there is only one group or cluster.

Reference taken from Clustering Analysis Class, Dr. Carlos Javier Vilalta P. (CentroGeo)

Clustering Analysis

K-means



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

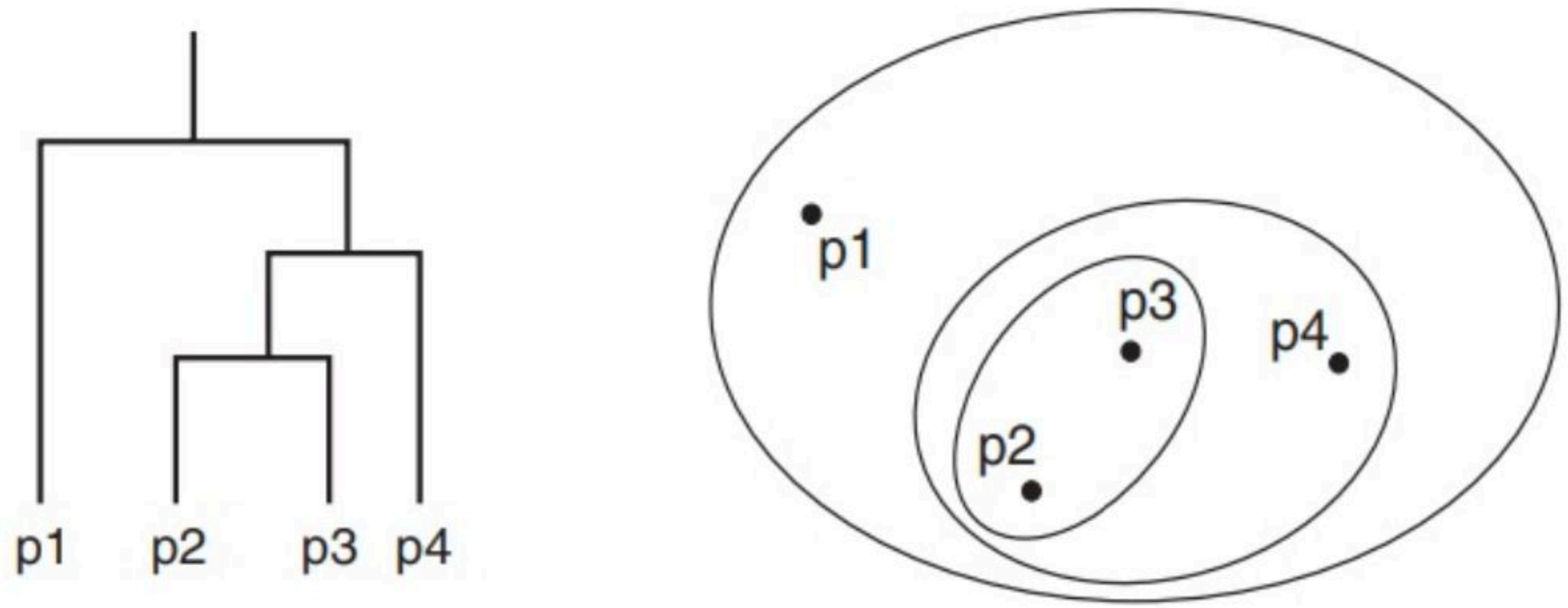
Fuente: University of Minnesota.

Start with 3 centroids (K = 3) assigning the closest observations and recalculating until the centroids do not change position.

Reference taken from Clustering Analysis Class, Dr. Carlos Javier Vilalta P. (CentroGeo)

Clustering Analysis

K-means



(a) Dendrogram. (b) Nested cluster diagram.

Fuente: University of Minnesota.

It starts by assigning the closest observations into groups and recalculating until there are no more groups (clusters) to "cluster" (conglomerate).

Reference taken from Clustering Analysis Class, Dr. Carlos Javier Vilalta P. (CentroGeo)

Spatial Clustering Analysis
Steps

1. Choose and standardize variables (Z)
2. Define neighbor matrix and weight nodes by dissimilarity
3. Plot and interpret results

Reference taken from Clustering Analysis Class, Dr. Carlos Javier Vilalta P. (CentroGeo)

Hands-on session

Break

Data visualization and decision making

Data visualization

Steps to create an effective visualization

Formulate the question

Gather the data

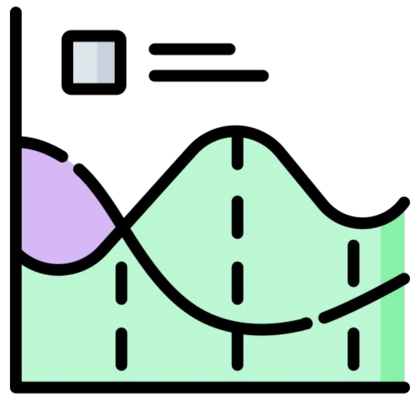
Apply a visual representation

Reference taken from Julie, S., & Noah, I. (2010). *Beautiful visualization.* O'Reilly
Icons made by srip from www.flaticon.com


COME AT ME BRO
I GOT DATA
Gotta Viz Em All
Most Featured Main Characters
Running Total of Plot Features
Top Movie Directors by Film
311 Service requests for graffiti removal - Chicago
Upto $2500 fine for a graffiti offence
Total requests
759,703
memegenerator.net

ggplot2

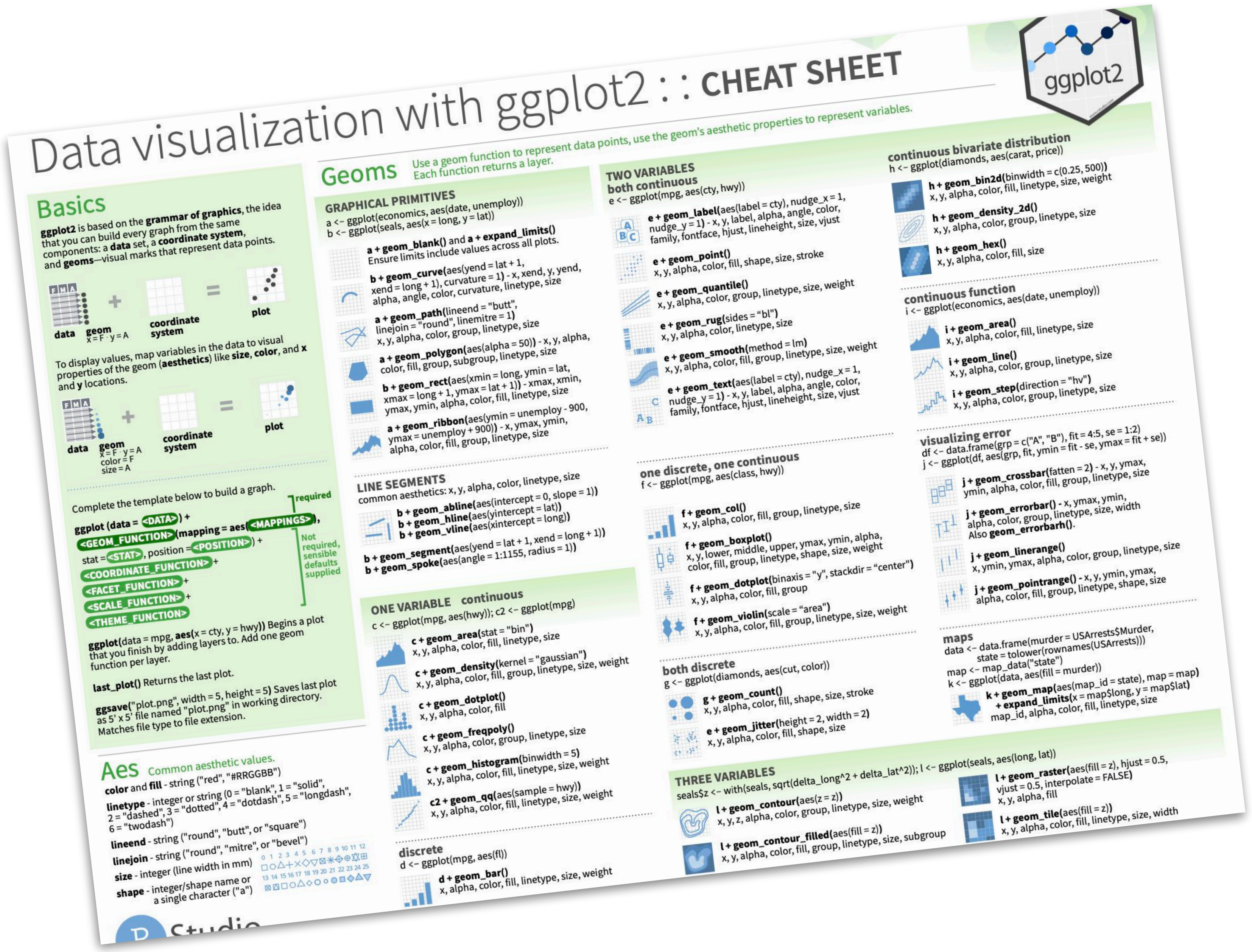
Visualizing data with graphs



R package used for producing statistical or data graphics.

Can create advanced and novel graphics tailored to specific problems using the *Grammar of Graphics,* instead using a set of predefined graphics.

Works iteratively adding layers of annotations and statistical summaries.



For further reference, check out **Data visualization with ggplot2 cheat sheet** included in the GitHub repository or RStudio Help Menu and **ggplot2: Elegant Graphics for Data Analysis** online book.

Icons made by Freepik from www.flaticon.com.

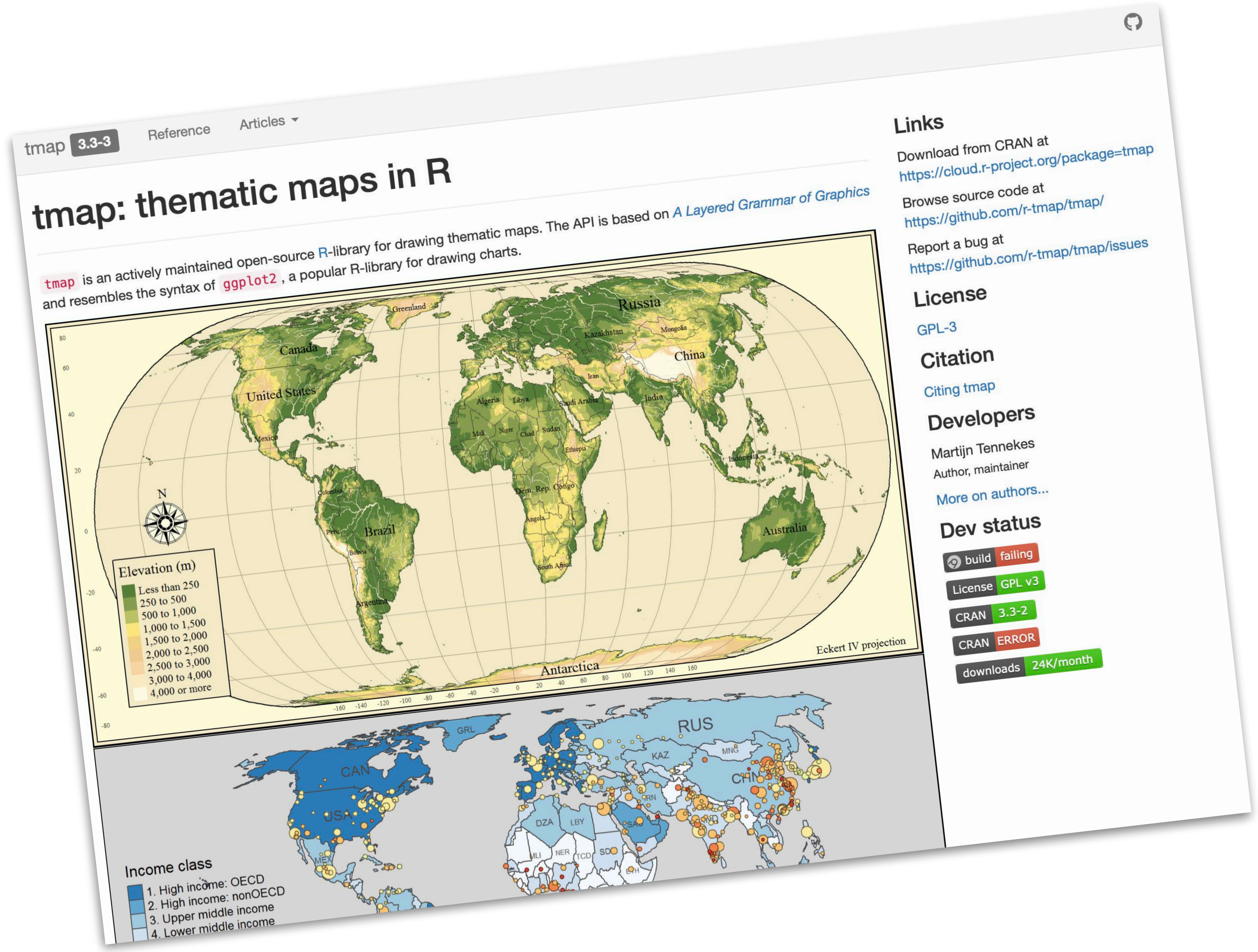
tmap

Creating thematic maps

R package used to produce high-quality maps for printing or interactive web maps in conjunction with other packages like *sf* or *raster*.

Uses a grammar similar to ggplot2's *Grammar of Graphics* to produce advanced maps suited to specific needs.

Works iteratively adding layers of geographic data, basemaps and annotations.



For further reference, check out **tmap package documentation** in https://r-tmap.github.io/tmap/ and Tennekes, M., 2018, tmap: Thematic Maps in R, Journal of Statistical Software, 84(6), 1-39.

Icons made by Freepik from www.flaticon.com.

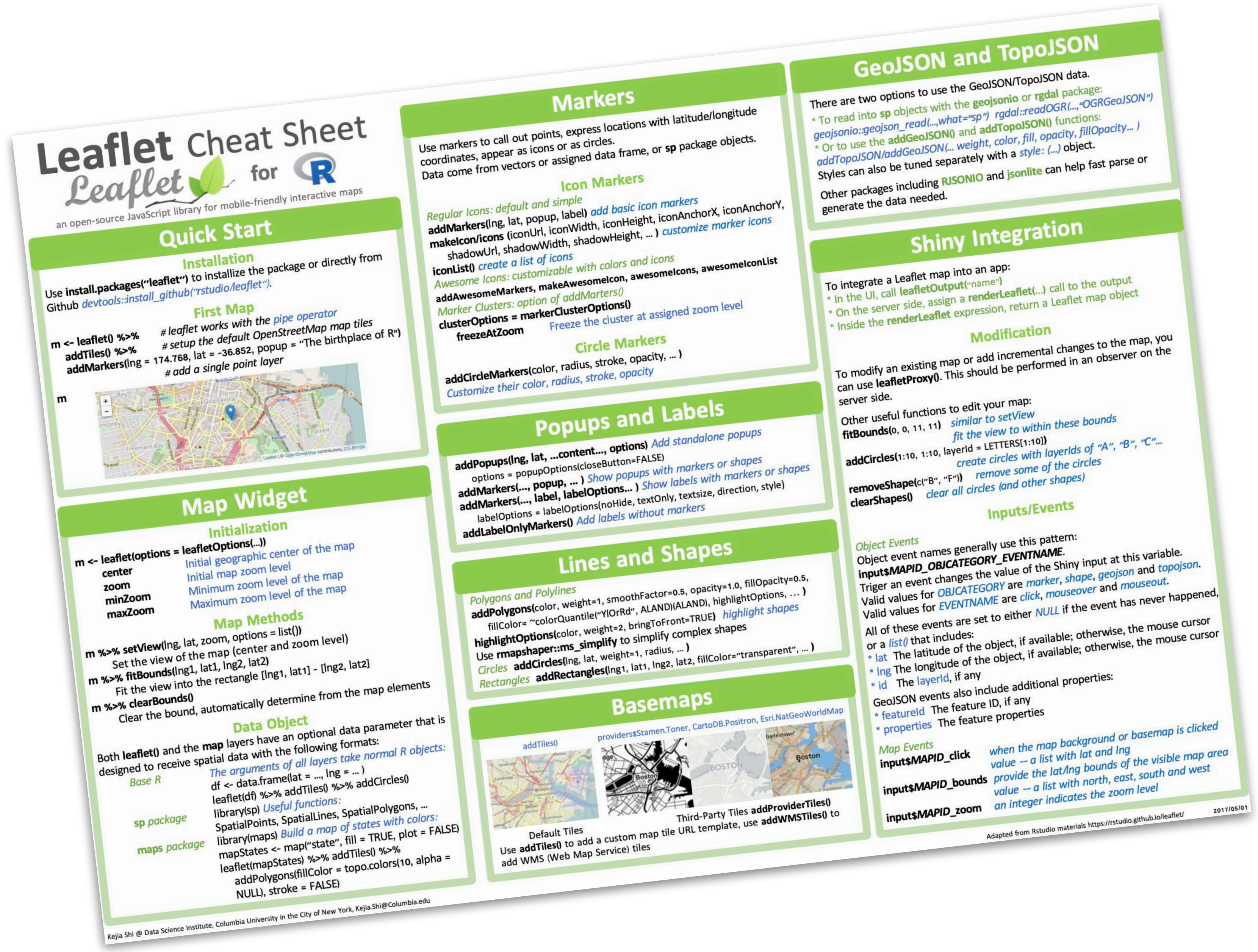
Leaflet

Interactive maps

R package used to produce interactive web maps in conjunction with other packages like *sf* or *raster*.

It has many advanced options to customize maps and it is designed to be used with pipes and integrates well with the Tidyverse.

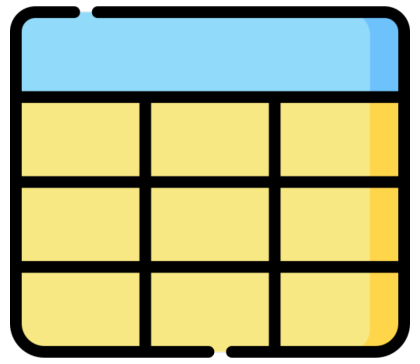
Works iteratively adding layers of geographic data, basemaps and annotations.



For further reference, check out **Leaflet for R cheat sheet** included in the GitHub repository.

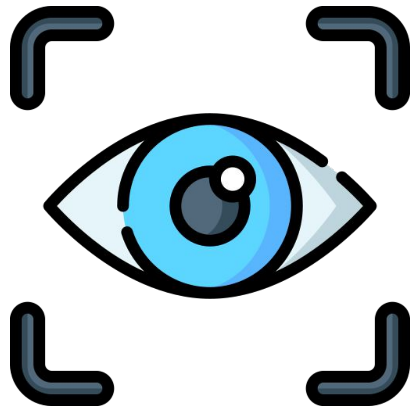
DT

Enhanced data tables



R package used to display matrices, data.frames or tibbles as interactive data tables with filtering, pagination, sorting features.

It is designed to be used with pipes and integrates well with the Tidyverse.



Allows visualizing tabular data with conditional formatting and visual aids.

For further reference, check out **DT package documentation** in https://rstudio.github.io/DT/

Hands-on session

Conclusions of workshop

Questions?

Contact us

Questions about the workshop?

Ana J. Alegre
jalegre@centrogeo.edu.mx

Cristian Silva
csilva@centrogeo.edu.mx

Thank you for attending this workshop!