Universidad de los Andes

Análisis espacial de datos y sus aplicaciones en Python

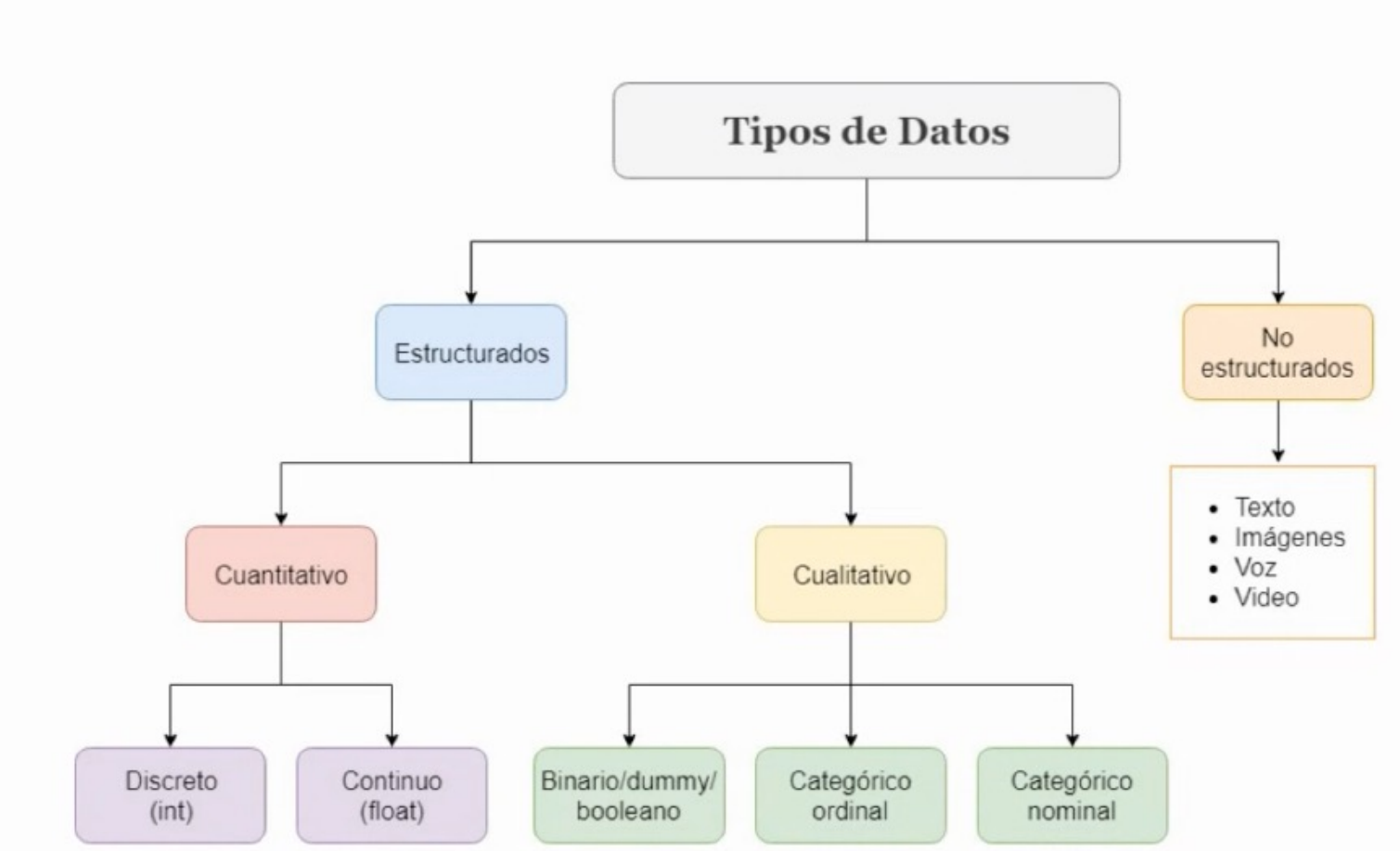
Profesor: Germán González

Sesión 3: Estadísticas descriptivas de geolocalización

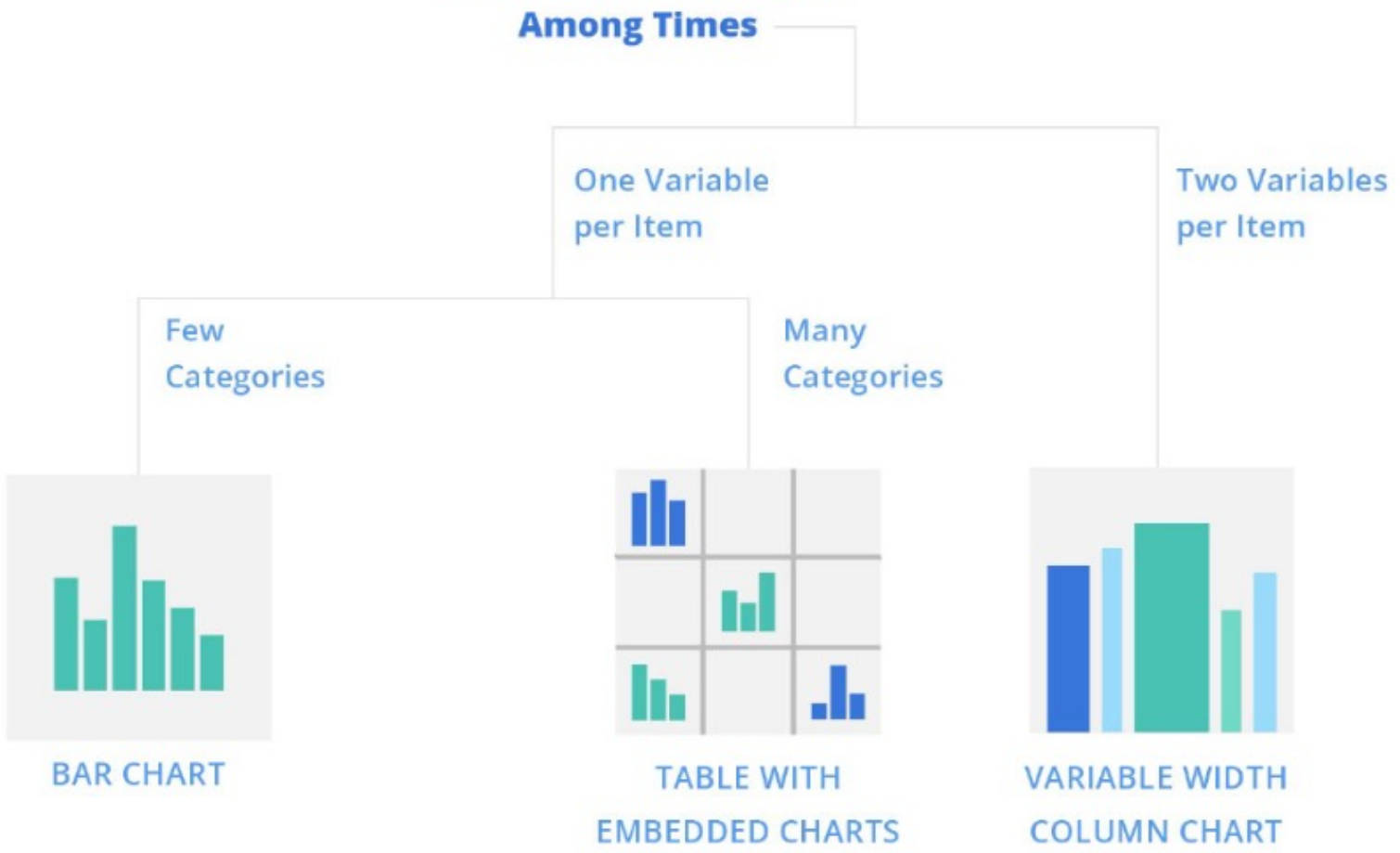
Índice

- Comparación estática
- Comparación a través del tiempo
- Composición estática
- Composición a través del tiempo
- Distribución

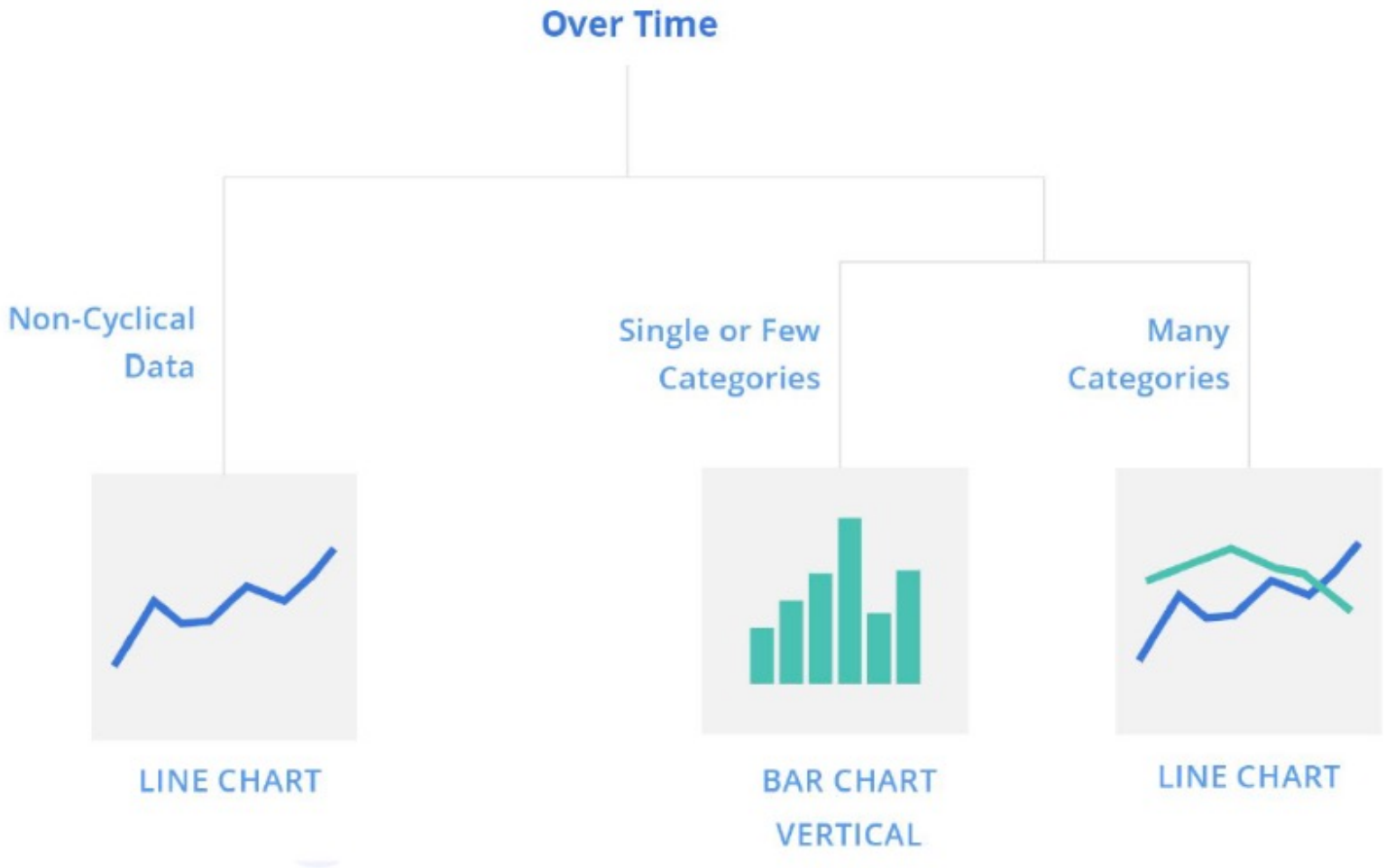
Tipos de datos



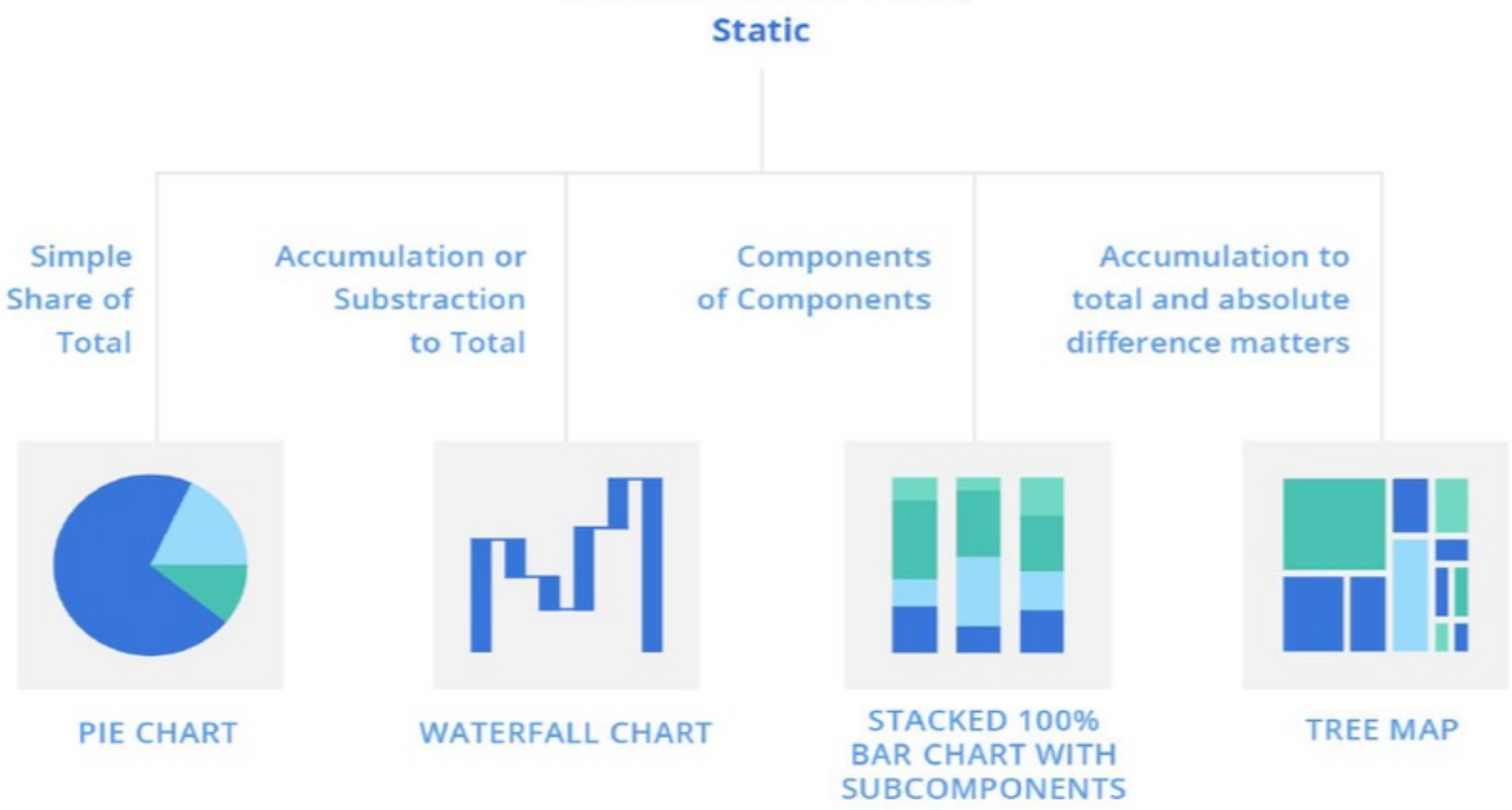
Visualización: Comparación estática



Visualización: Comparación a través del tiempo

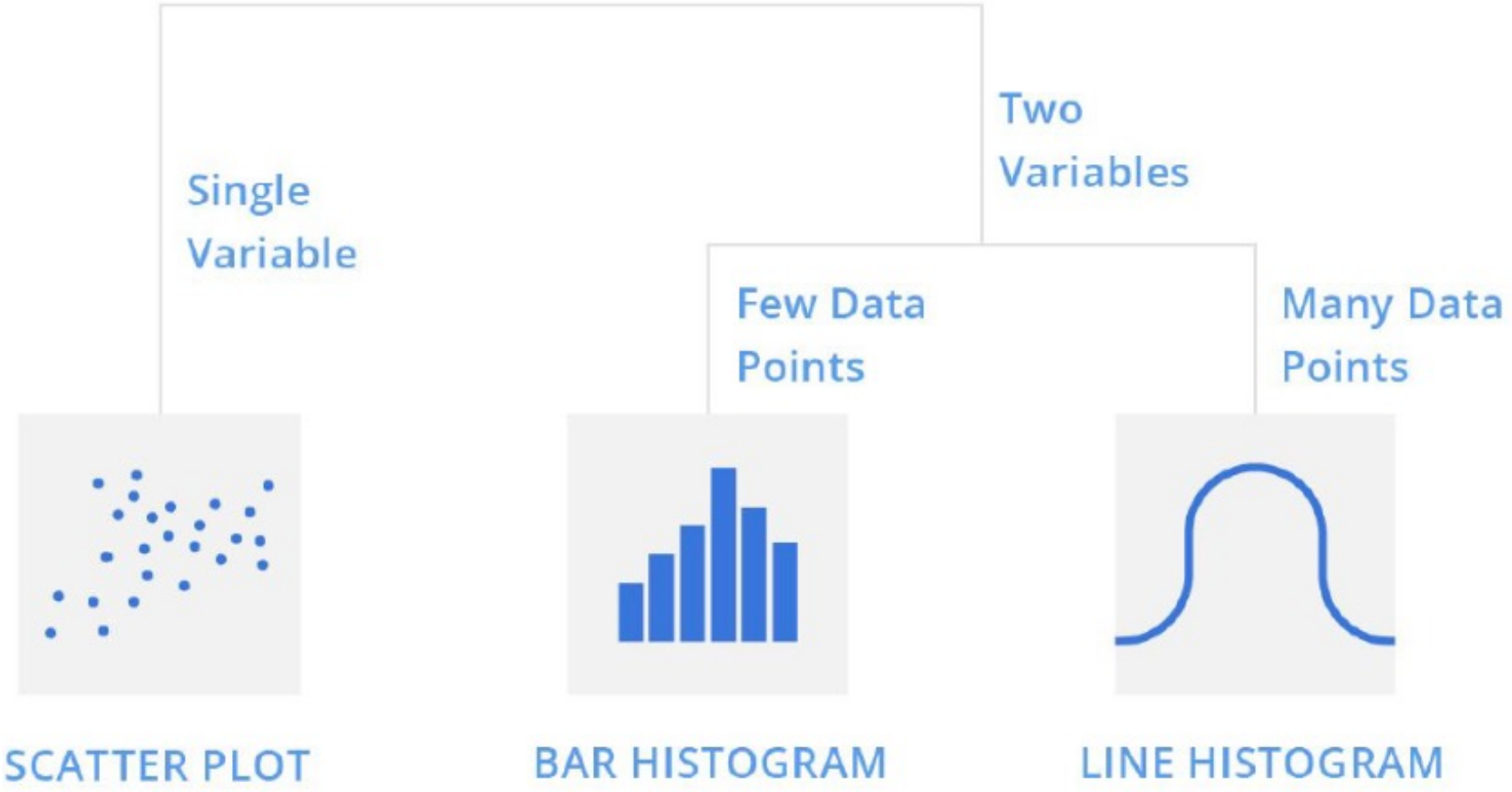


Visualización: Composición estática

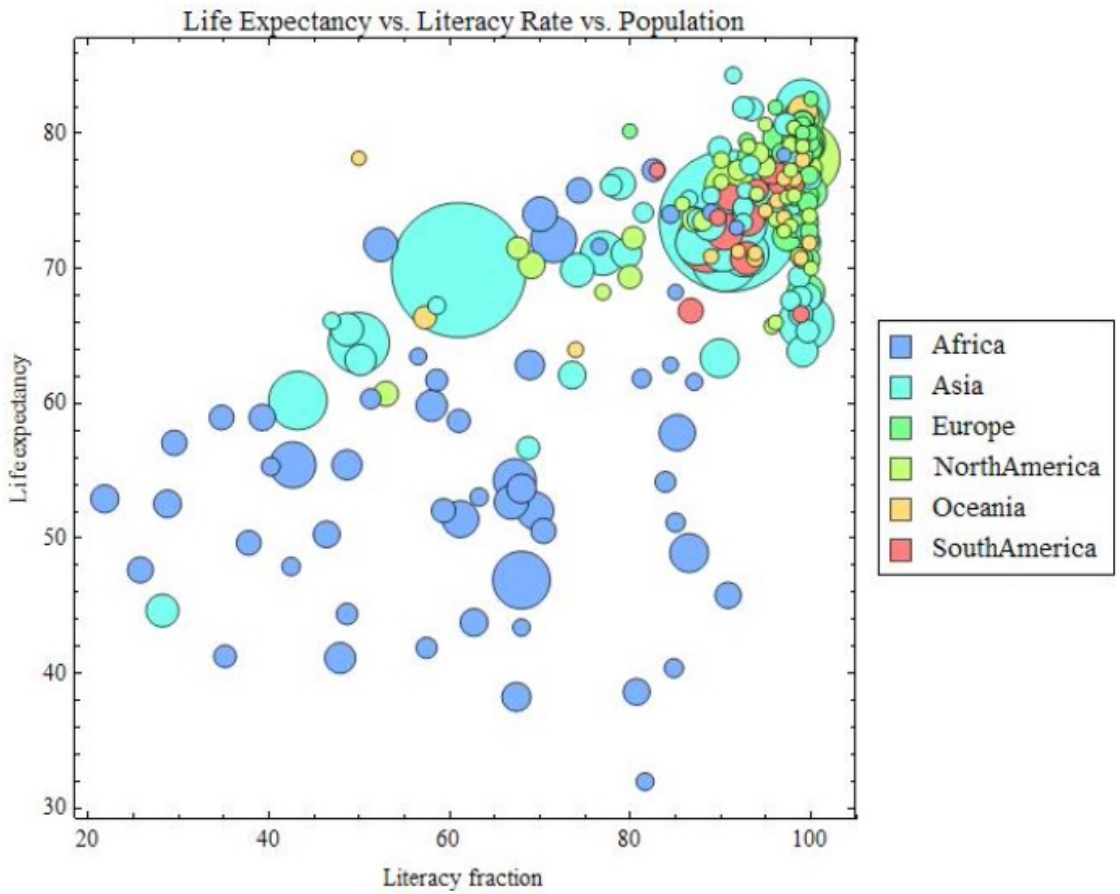


Visualización: Composición a través del tiempo

Distribuciones



Burbujas

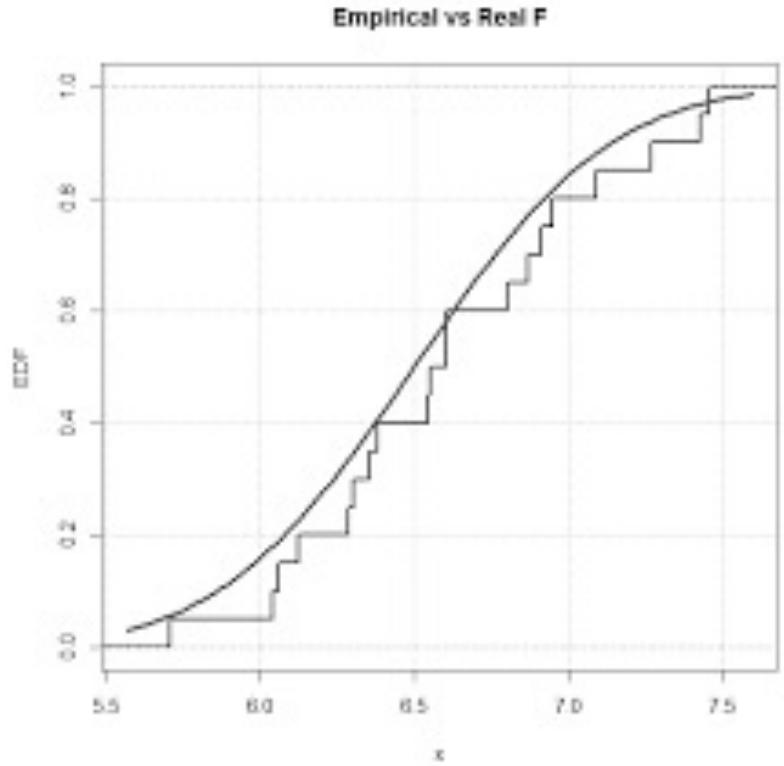

Life Expectancy vs. Literacy Rate vs. Population
Life expectancy
Literacy fraction
80
70
60
50
40
30
20
40
60
80
100
Africa
Asia
Europe
NorthAmerica
Oceania
SouthAmerica

Estimación de distribución empírica

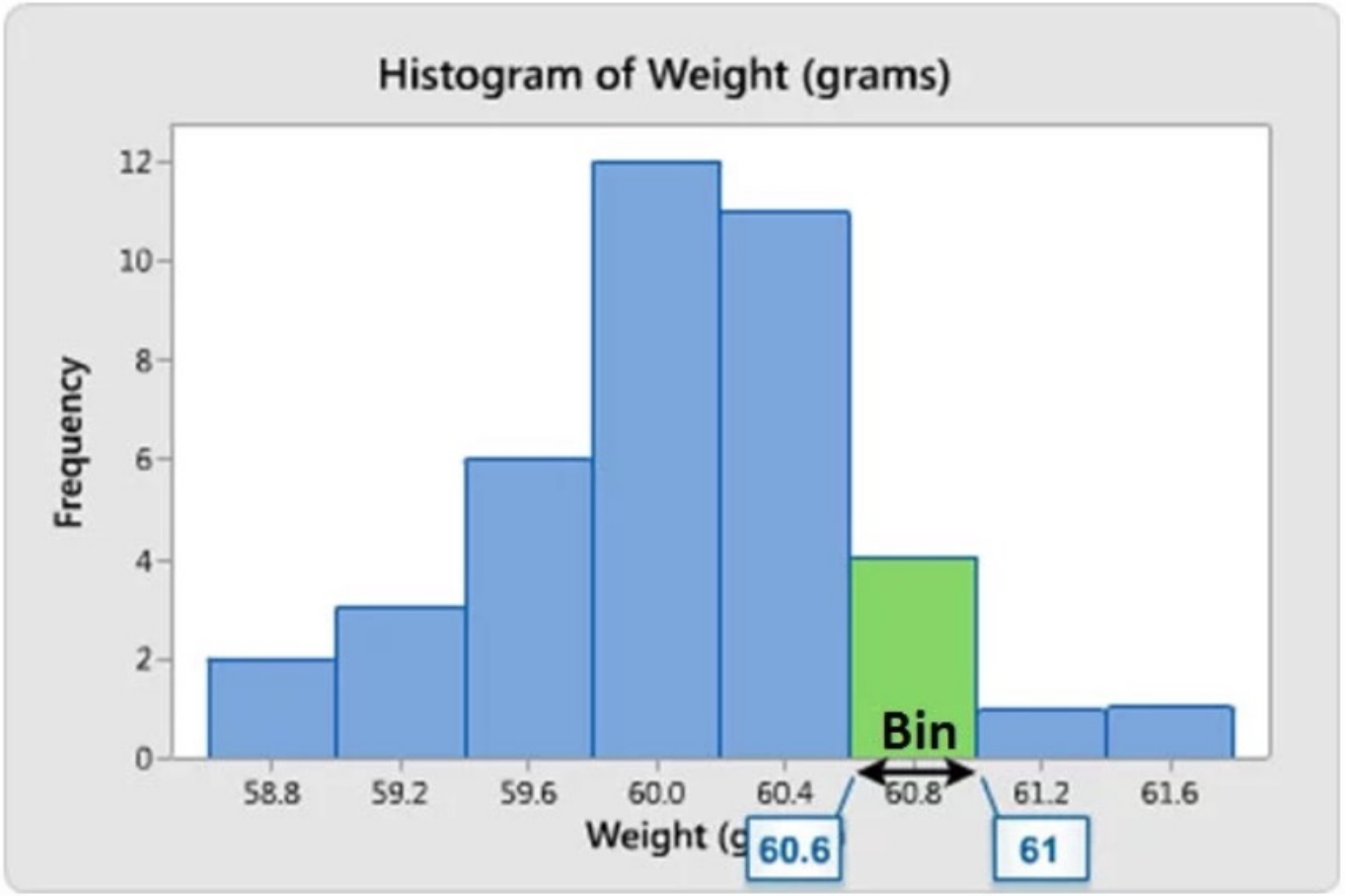
Es una herramienta muy poderosa que permite estudiar la distribución de una o varias variables.

Aproximación a las diferentes medidas de una distribución:

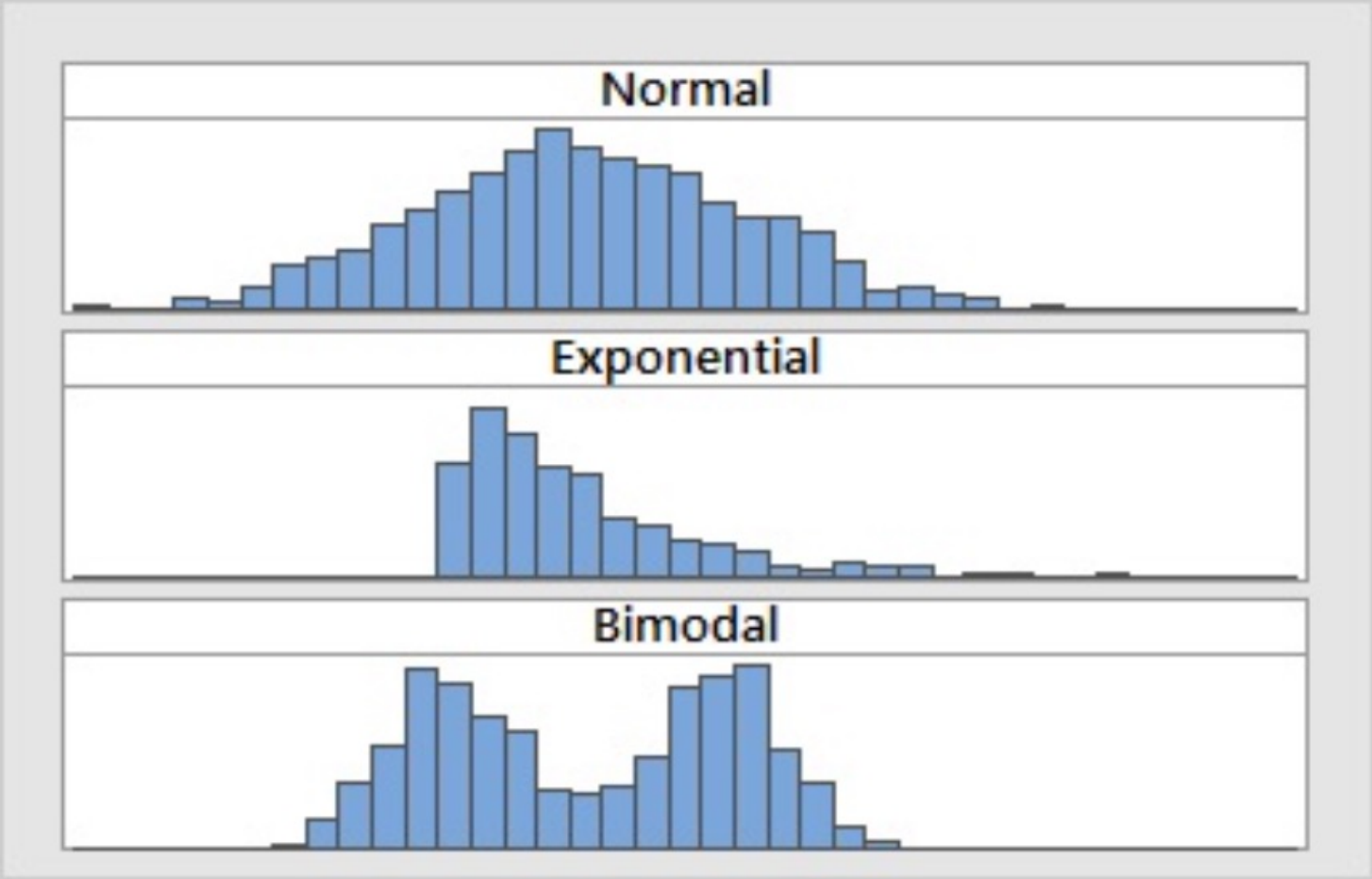
- Valores más y menos comunes.
- Función de probabilidad subyacente.
- Estimación de momentos de la distribución.



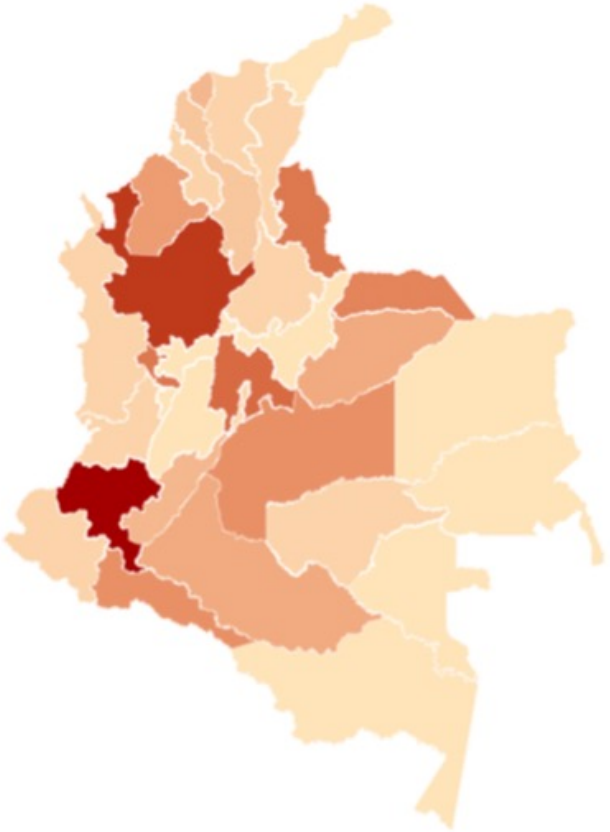
Histograma

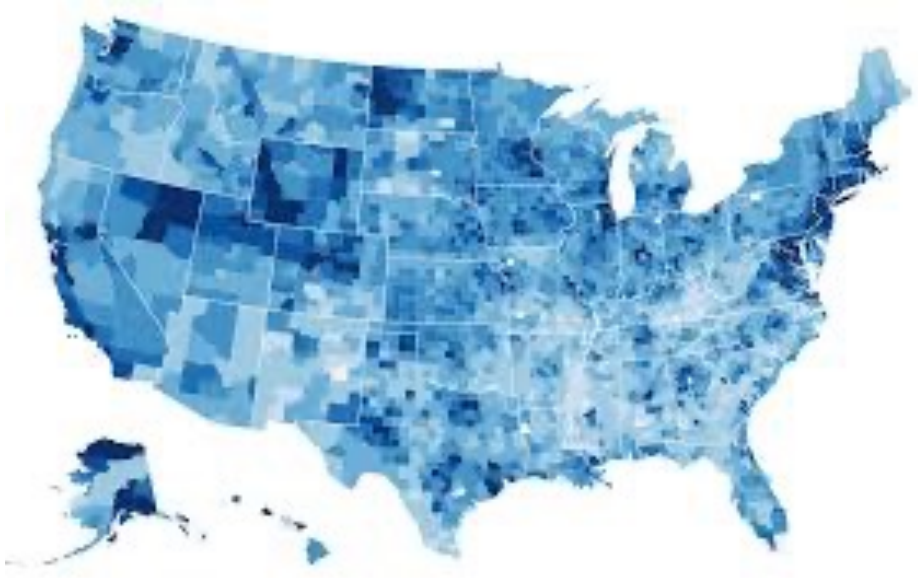

Histogram of Weight (grams)
Frequency
12
10
8
6
4
2
0
58.8
59.2
59.6
60.0
60.4
60.8
61.2
61.6
Bin
Weight (g
60.6
61

Función de probabilidad empírica



En datos espaciales: Mapa coroplético / conteo en grillas



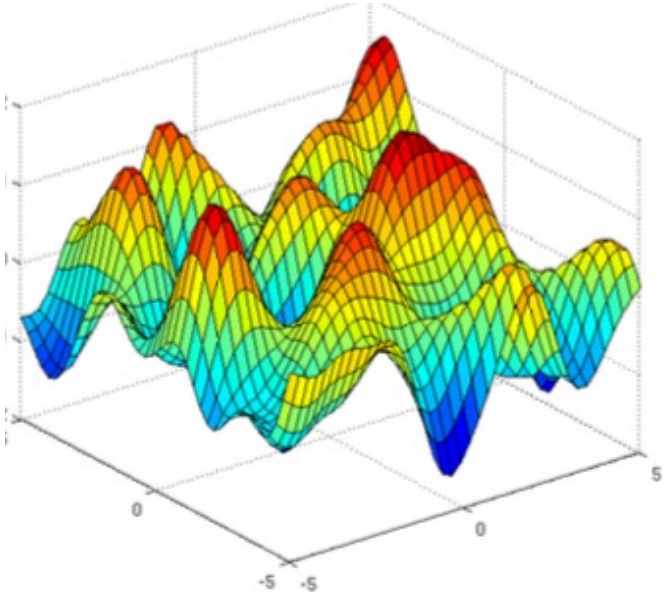


Estimación de Densidad de Kernel (KDE)

- El método de Estimación de Densidad de Kernel (KDE) generaliza la noción de histograma haciendo conteos sobre intervalos / celdas arbitrariamente pequeñas.

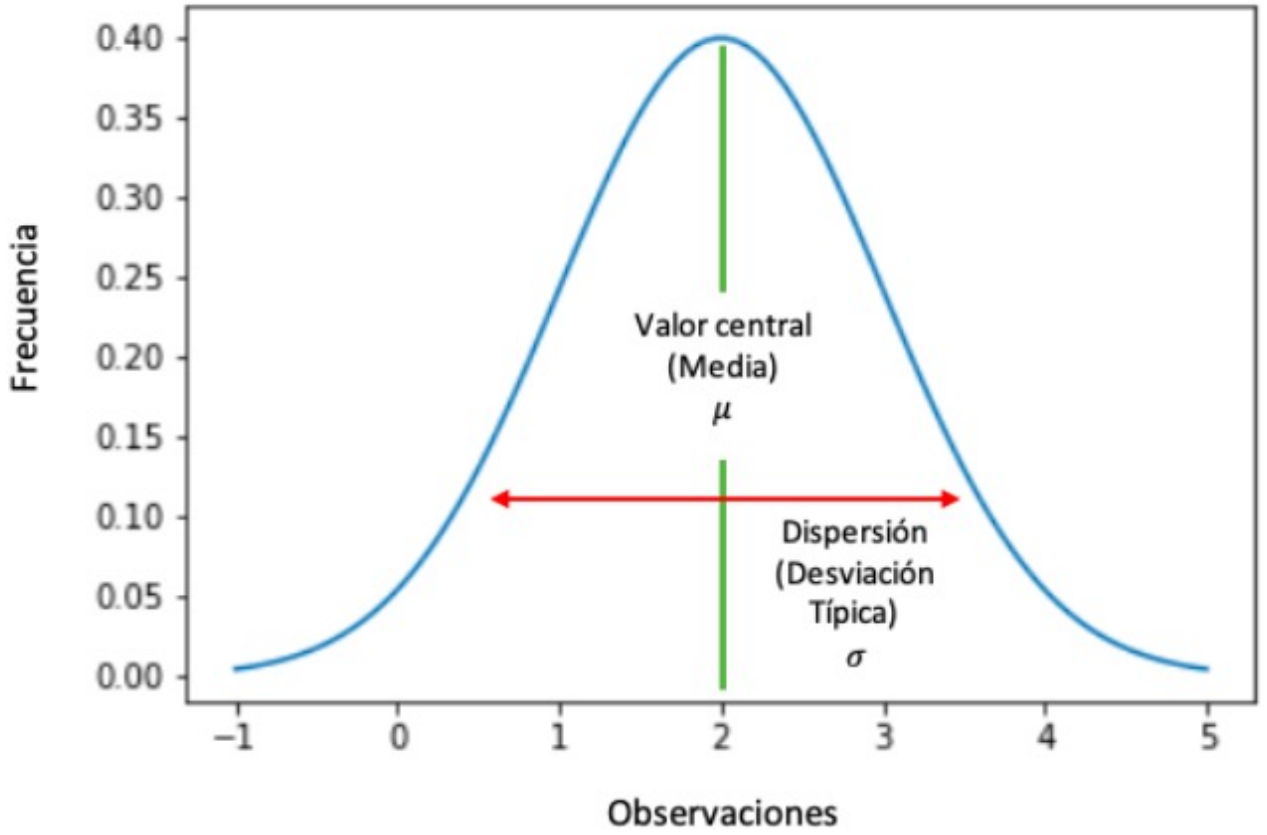
- Permite un ajuste más suave a la distribución de los datos dado que no se limita a regiones espaciales predefinidas.

- Clave contar con datos puntuales georreferenciados.

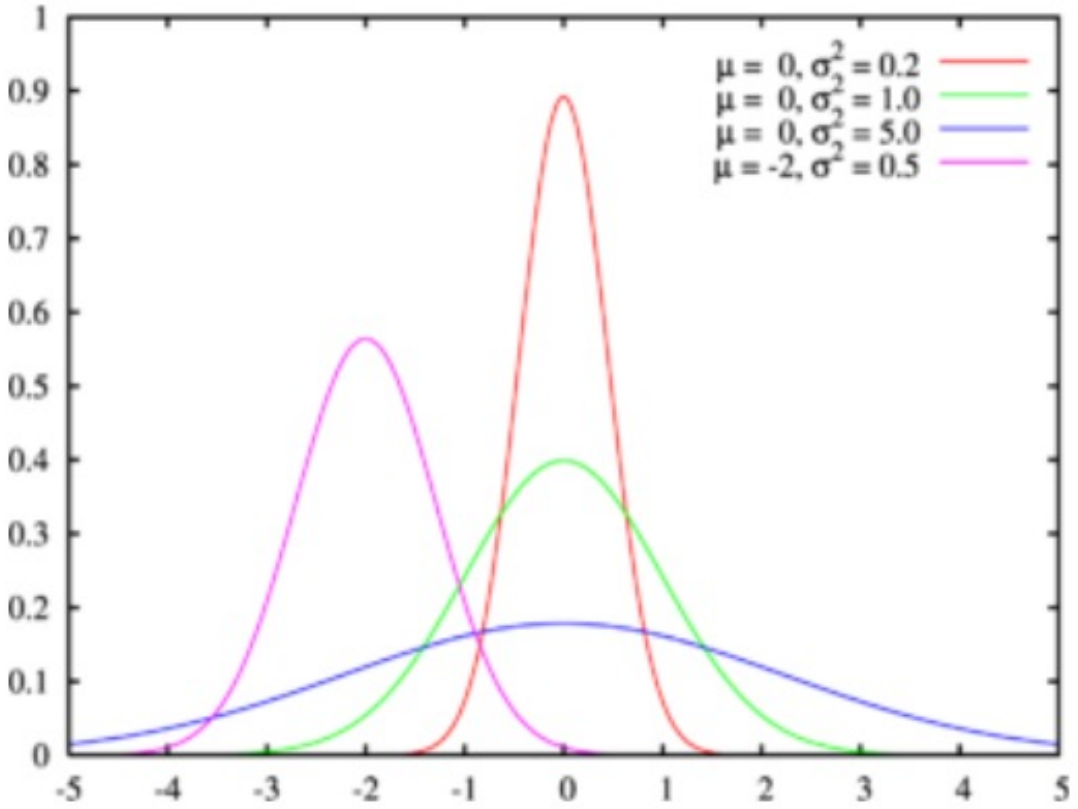


Distribución normal

La distribución normal o Gaussiana es una de las distribuciones de probabilidad más importantes en la estadística, utilizada para modelar muchos procesos físicos y sociales.

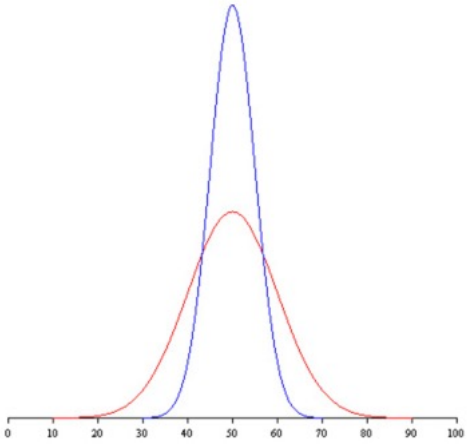


Efectos de media y varianza

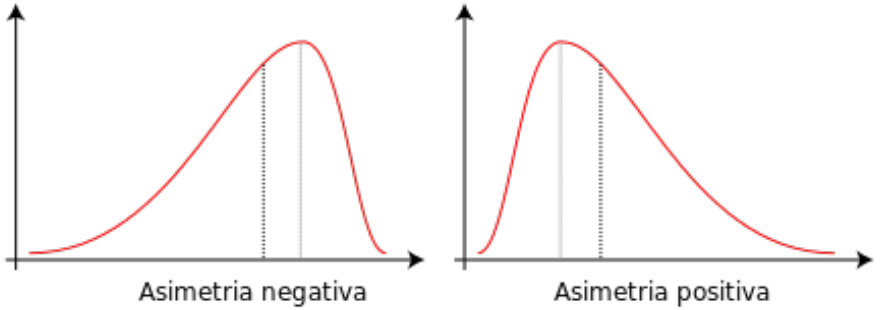


Distribución de los datos

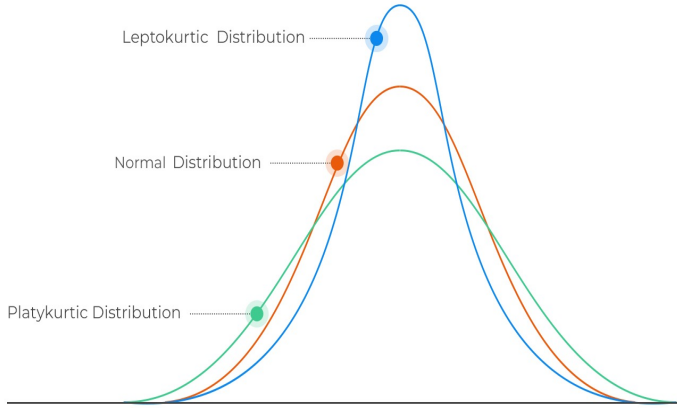
Media y varianza



Asimetria



curtosis



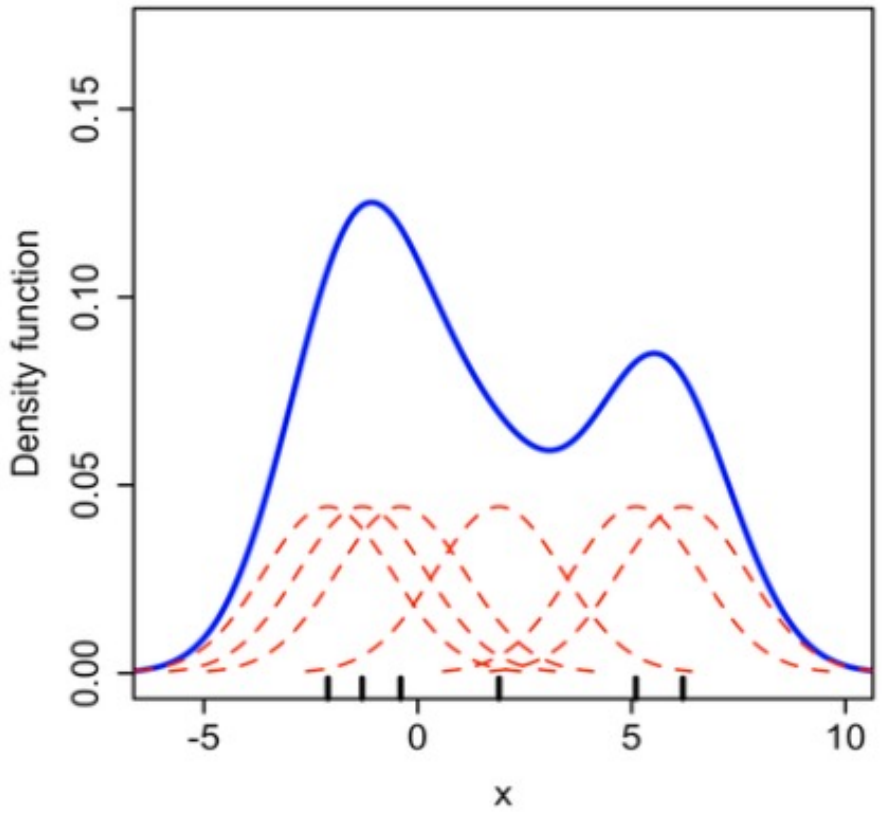
Estimación de Densidad de Kernel (KDE)

Se considera una función no negativa -el *kernel*- y un parámetro de suavizado denominado *bandwidth (h)*.

Cuando el kernel es una función gaussiana, cada observación es sustituida por una curva de este tipo centrada en dicho valor.

Se suman las curvas para obtener el valor de la densidad en cada punto.

La curva resultante se normaliza para que el área bajo ella sea igual a 1.



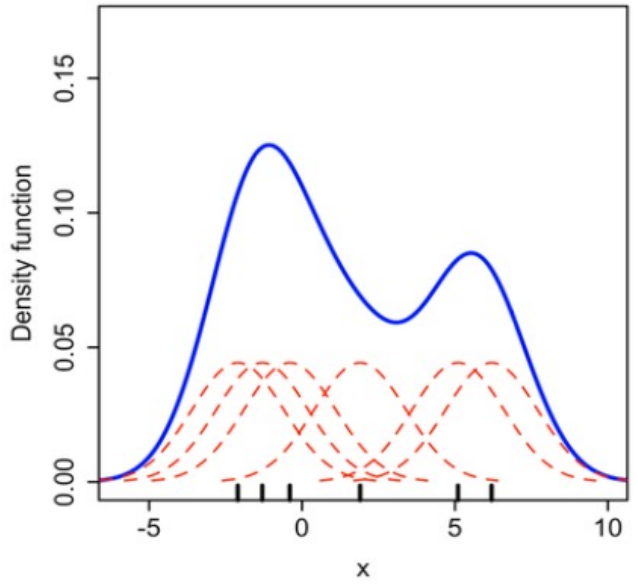
Estimación de Densidad de Kernel (KDE)

Dado un conjunto de datos

$$\{X_i\}_{i=1}^{n}$$

Se ubican las distribuciones Gaussianas sobre cada punto:

$$K_h(x, x_i) = \frac{1}{\sqrt{2\pi}h} exp\left[-\frac{(x - x_i)^2}{2h^2}\right],$$

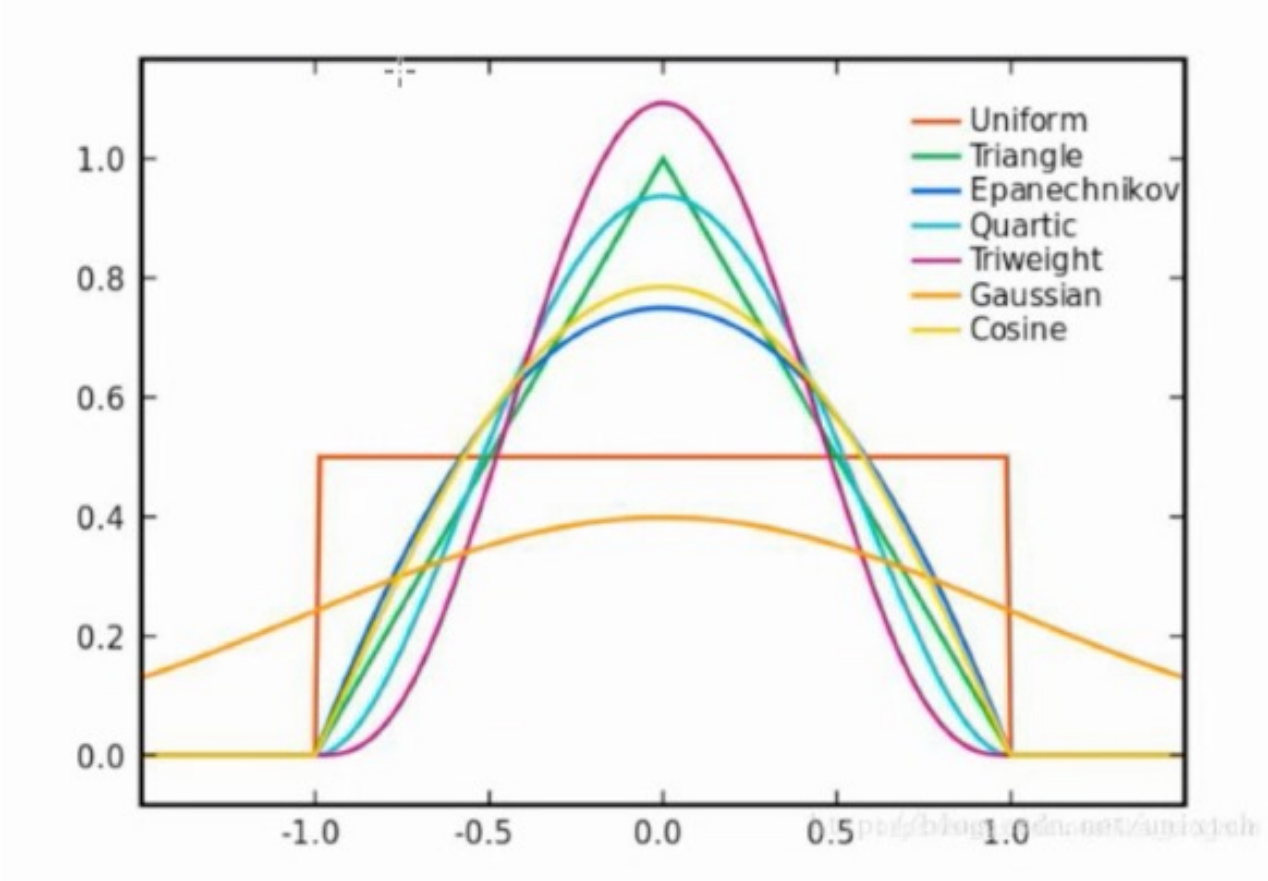


Luego, se suman (verticalmente) sobre todo el mapa

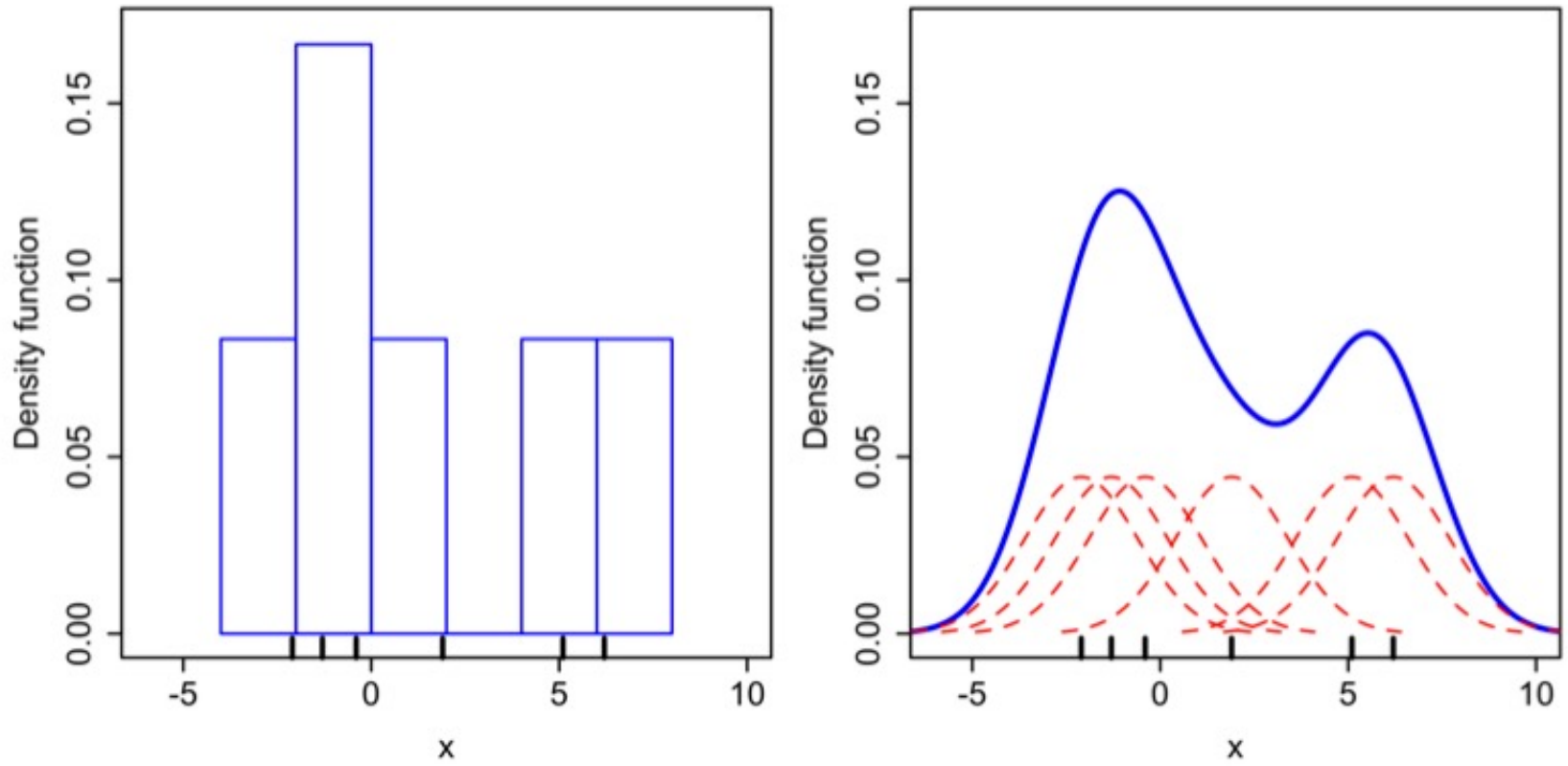
$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x, x_i).$$

El parámetro h de conoce como ancho de banda.

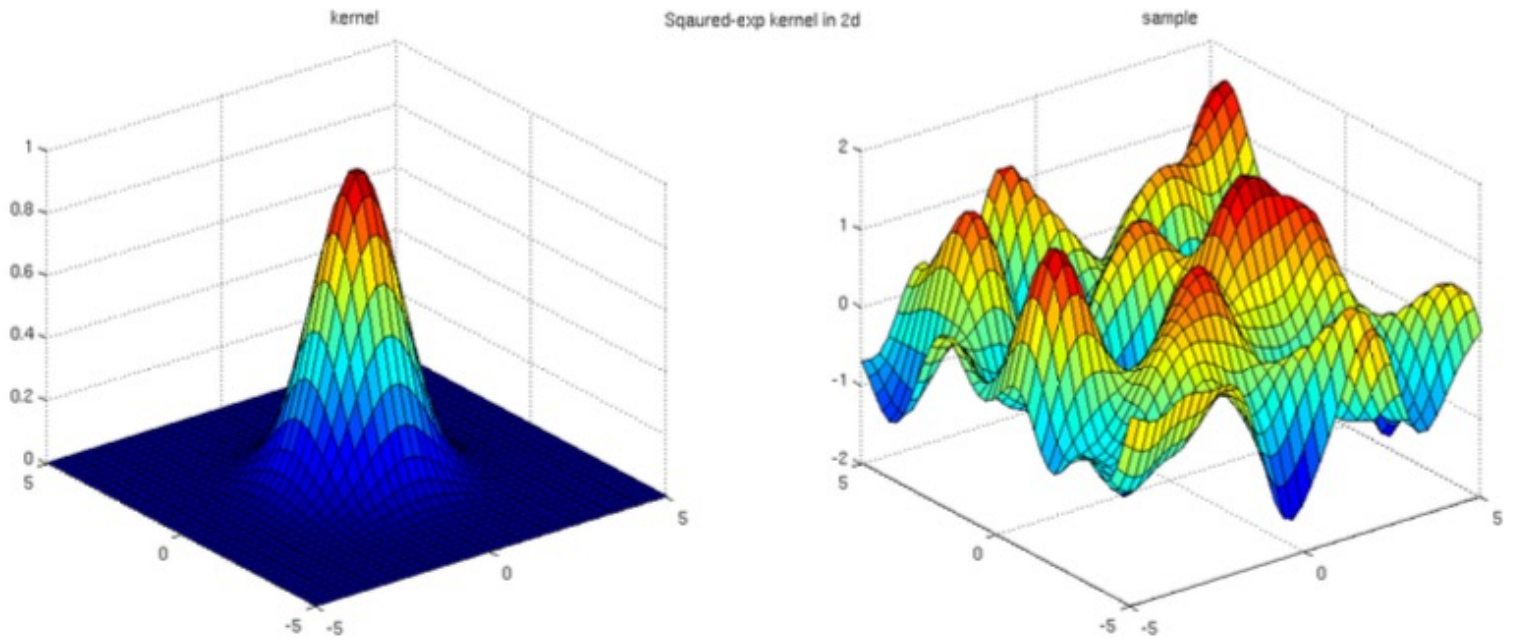
Estimación de Densidad de Kernel (KDE)



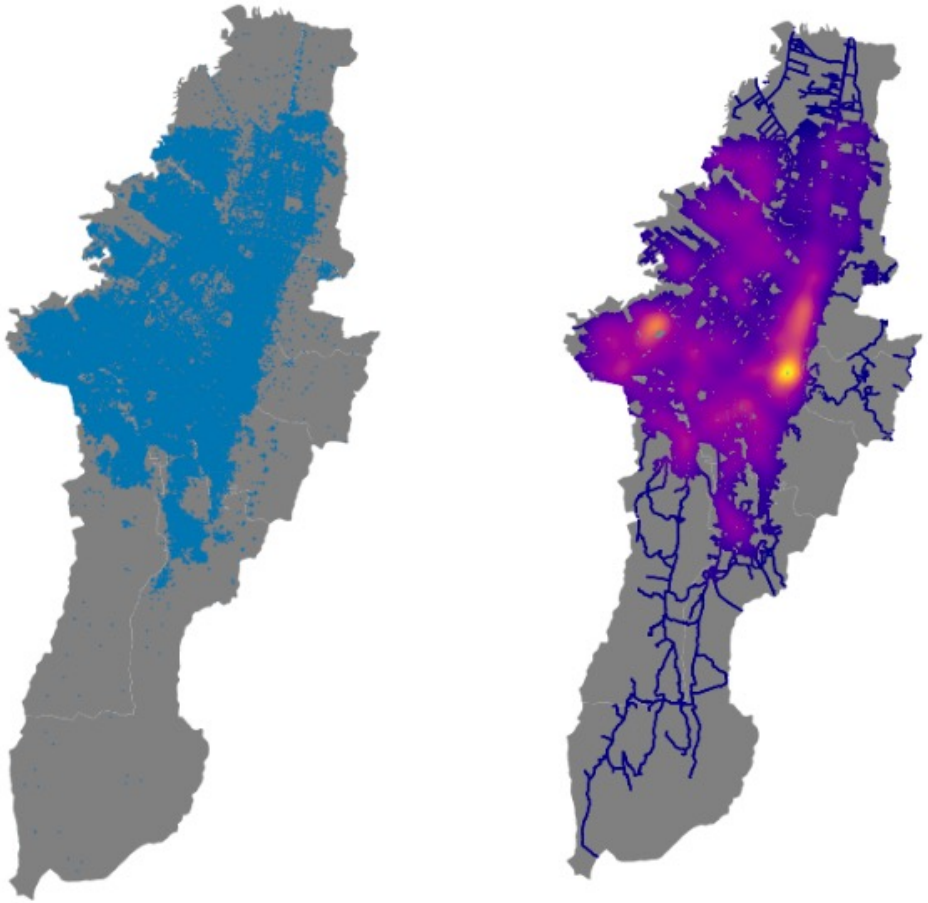
Método KDE vs Histograma



KDE en tres dimensiones: Heatmap



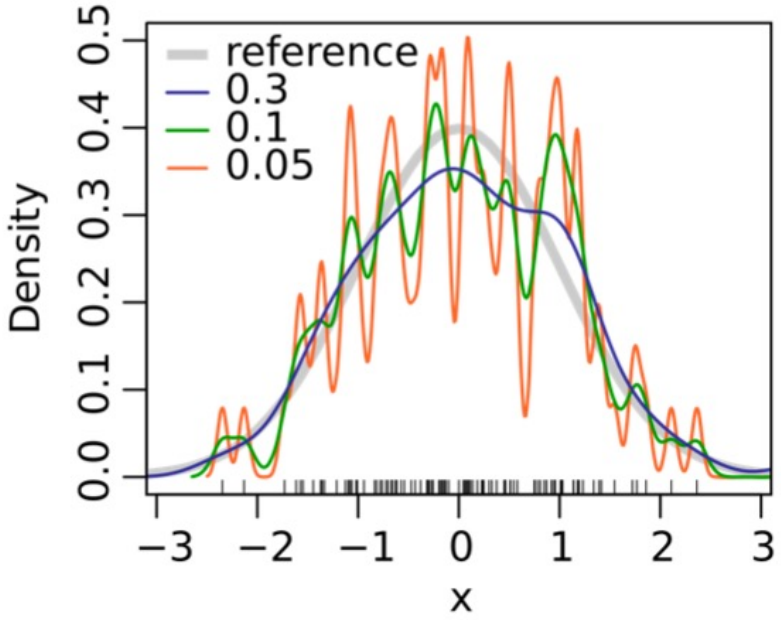
KDE Crimen en Bogotá



Estimación de Densidad de Kernel (KDE)

- **Definición**: que tanto permite que una observación me afecte las regiones vecinas
- El ancho de banda es equivalente a la varianza de las distribuciones Gaussianas ajustadas.
- En general, indica qué tanto se permite que la ocurrencia de un evento tenga efecto en lugares cercanos.
- Disyuntiva entre sesgo y varianza.

- **H grande**: celdas muy grandes (Bogotá)
- **H pequeño**: celda muy pequeñas

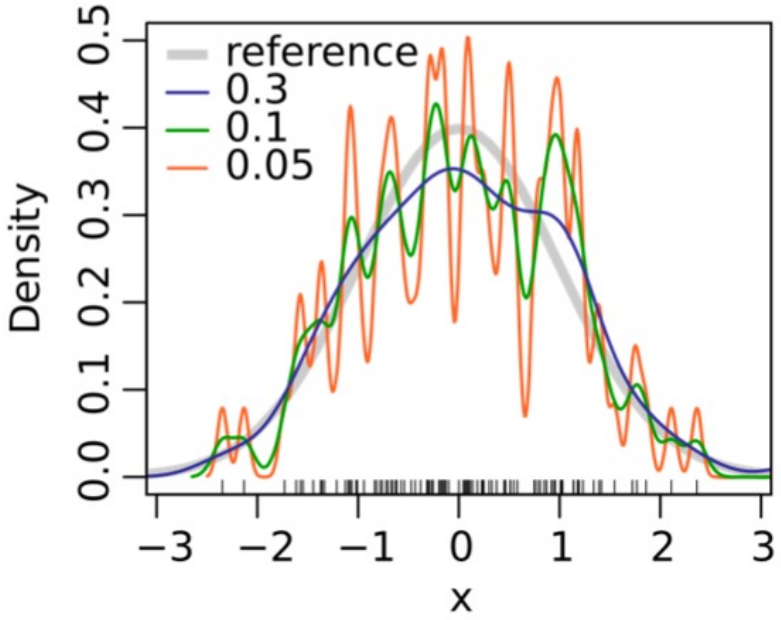


h grande sobre-suaviza / pequeño sub-suaviza la dsitribución.

Método KDE: Escoger el ancho de banda

$h = 1.06\sigma\hat{}n{-}1/5$ minimiza el Mean Integrated Squared Error.

Recomendado: Búsqueda sobre grilla para maximizar verosimilitud de observar los datos, más robusto si se acompaña de validación cruzada.



h grande sobre-suaviza / pequeño sub-suaviza la dsitribución.

Gracias