Relaciones espaciales

Dr. Raimundo Sánchez

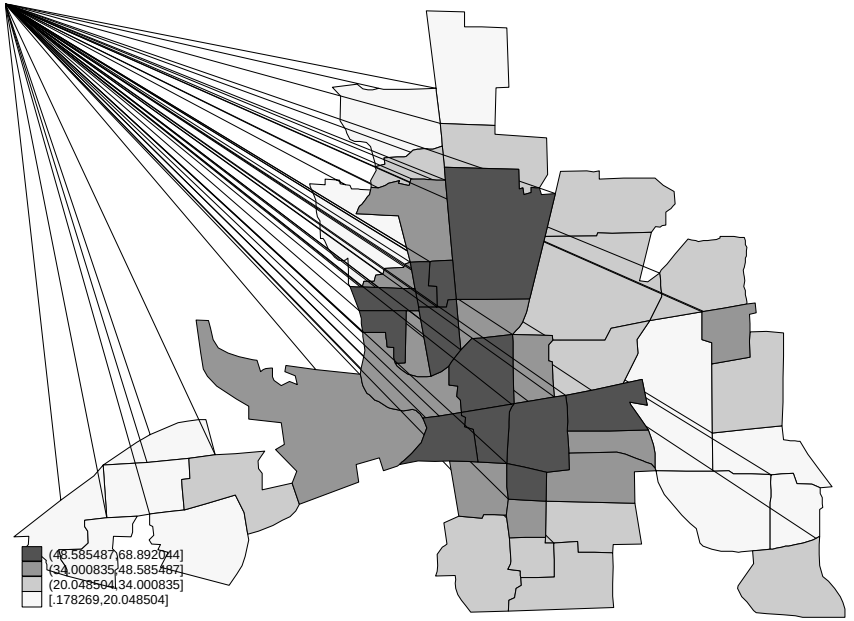
raimundo.sanchez@uai.cl

@raimun2

¿Relaciones espaciales?

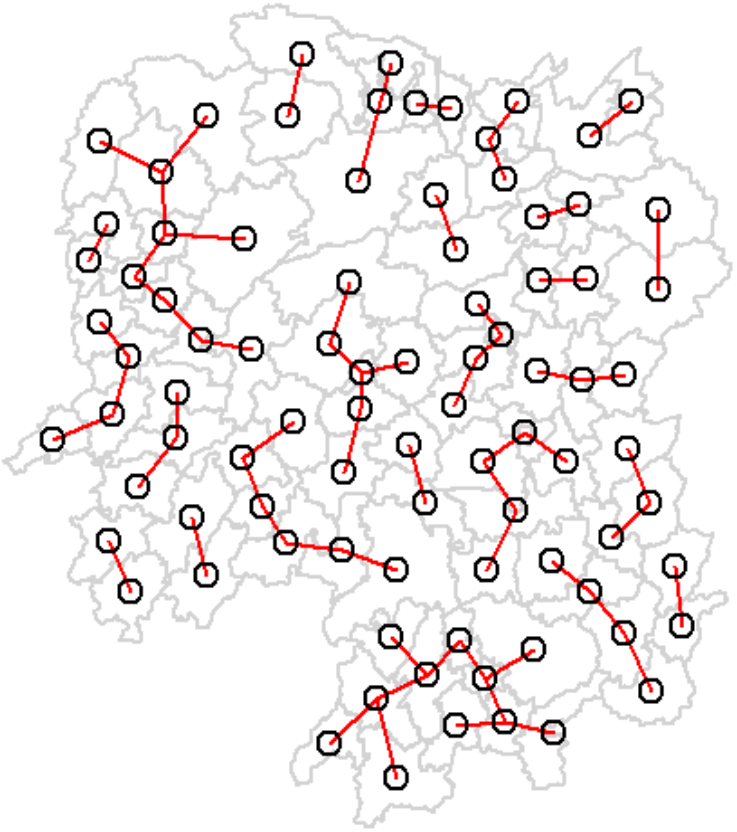
Precio de la vivienda

Crimen



Tipos de relaciones espaciales

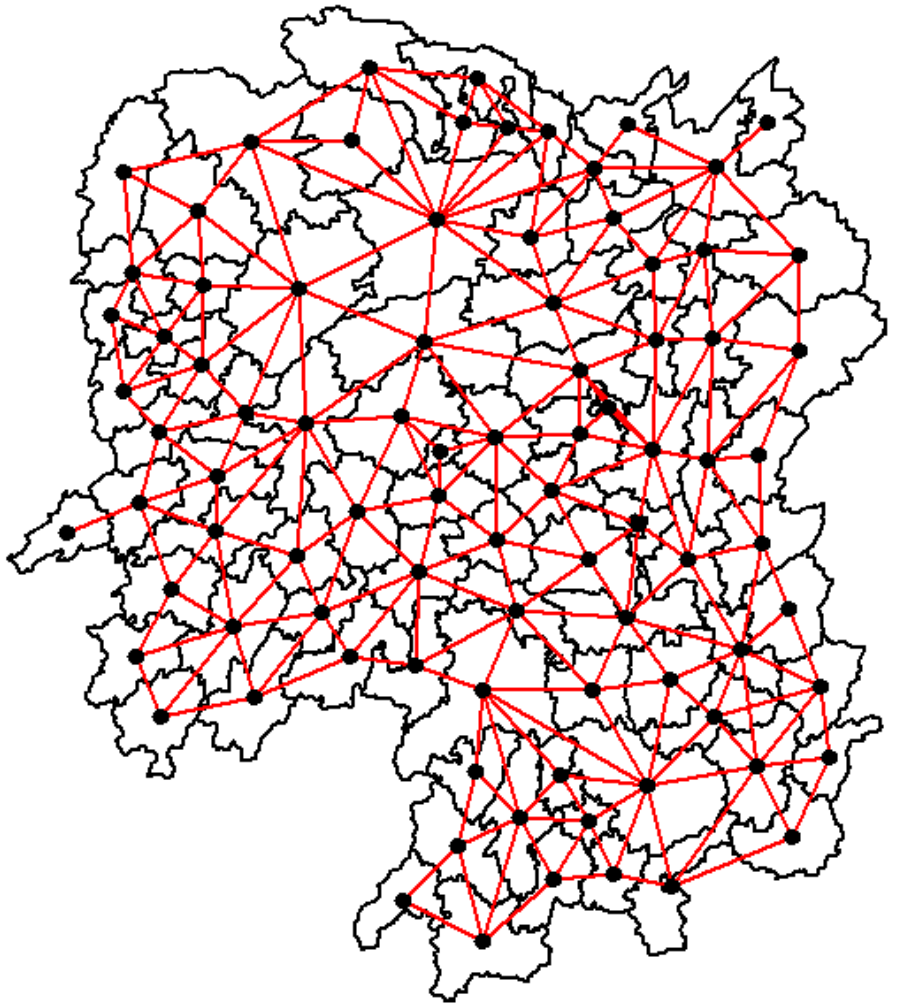
- Es la manera en que una observación espacial interactúa con el resto
- La principal razón de interacción se da con unidades vecinas
- Se representa matemáticamente a través de una matriz de relación o interacción
- Se han propuesto **distintas especificaciones**:
  - Contigüidad
  - Proximidad
  - Flujos



Contigüidad espacial

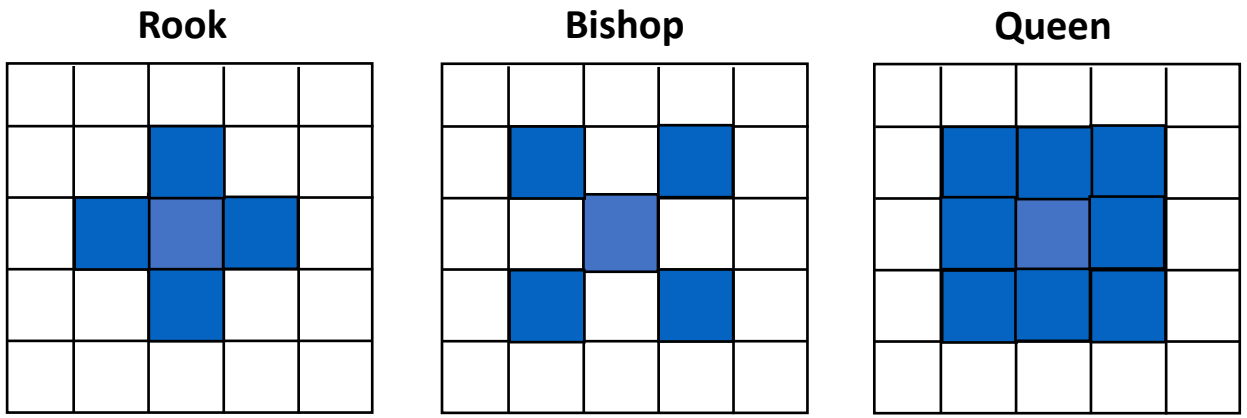
- La interacción espacial se refleja en base a la contigüidad, o mas bien si 2 unidades son vecinas entre si.
- La matriz de contigüidad es una matriz binaria, que representa 1 o 0, dependiendo si 2 unidades son vecinas.

$$w_{ij} = \begin{cases} 1 & \text{Si los sitios I y J están conectados} \\ 0 & \text{si no están conectados} \end{cases}$$



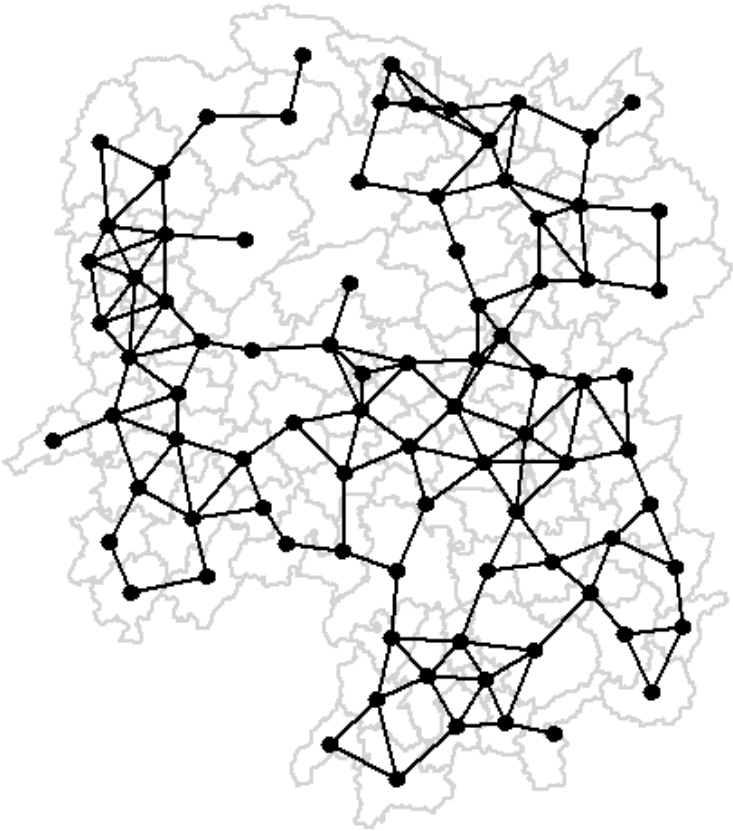
Definición de contigüidad

- La contigüidad se puede definir en base a 3 criterios principales
  - Criterio de torres: solo relaciones ortogonales
  - Criterio de Alfil: solo relaciones diagonales
  - Reina: tanto relaciones ortogonales como diagonales



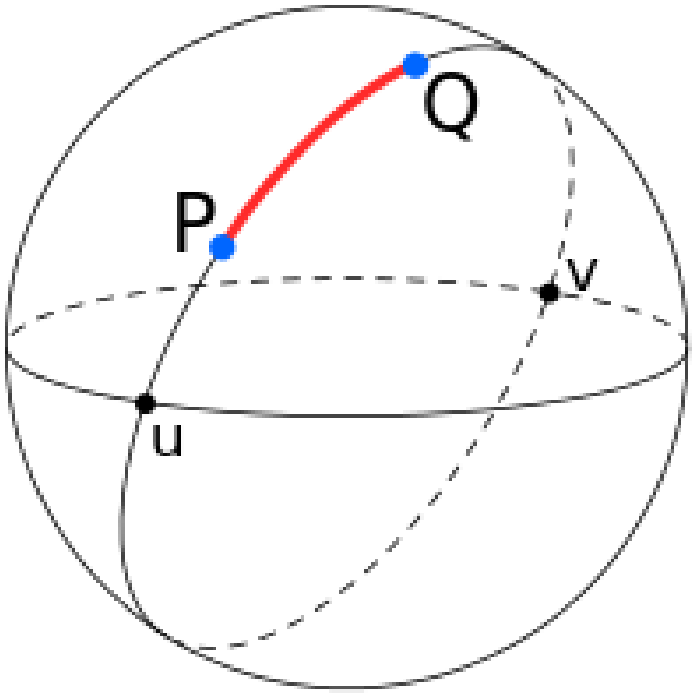
Proximidad

- Se pueden identificar vecinos que comparten bordes
- Otra alternativa es basarse en un radio de distancia, donde aquellos que caen en ese radio son vecinos.



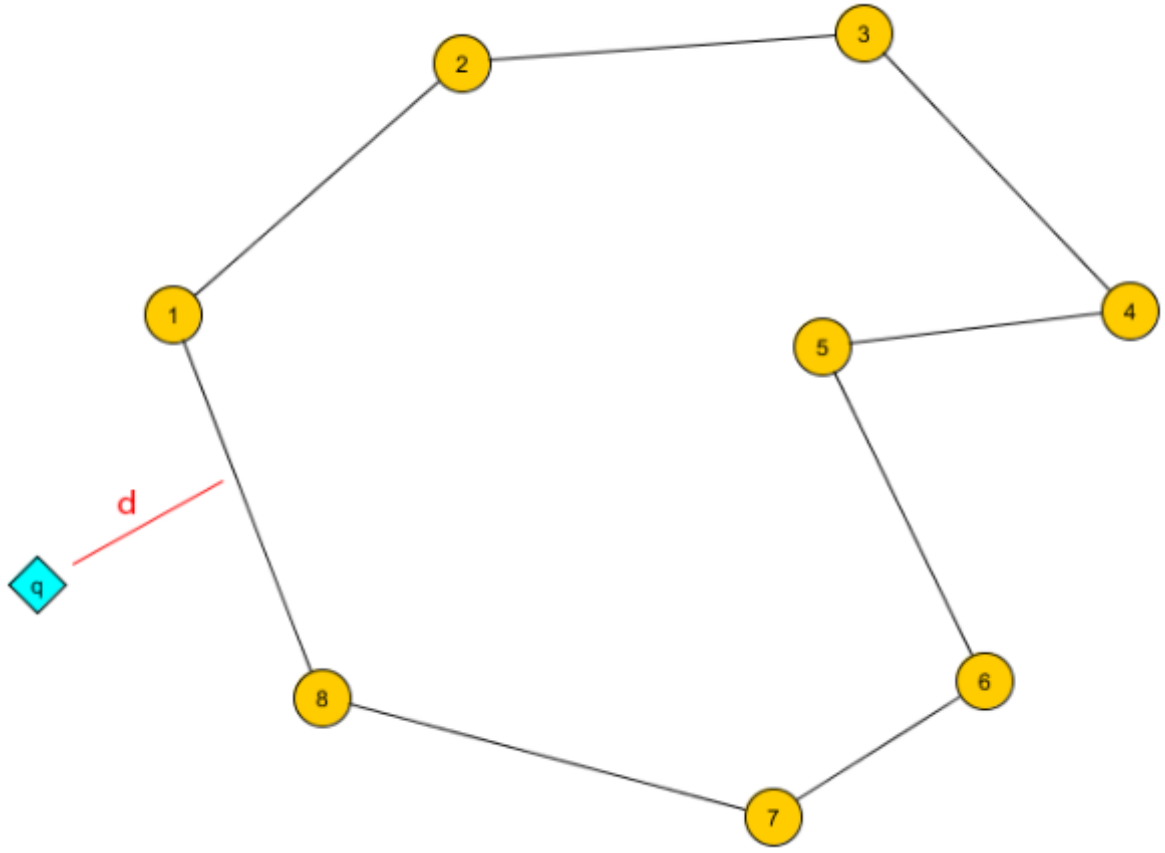
Distancias geográficas

- En el espacio las medidas de distancia tradicionales no son precisas
- Las medidas de distancia deben considerar la curvatura de la tierra, cuando la escala a analizar es relevante.
  - Distancia Haversine
  - Distancia Geodésica
  - Distancia ley de cosenos
  - Distancia de Vincenty esférica
  - Distancia de Vincenty elipsoide
  - Distancia de Meeus



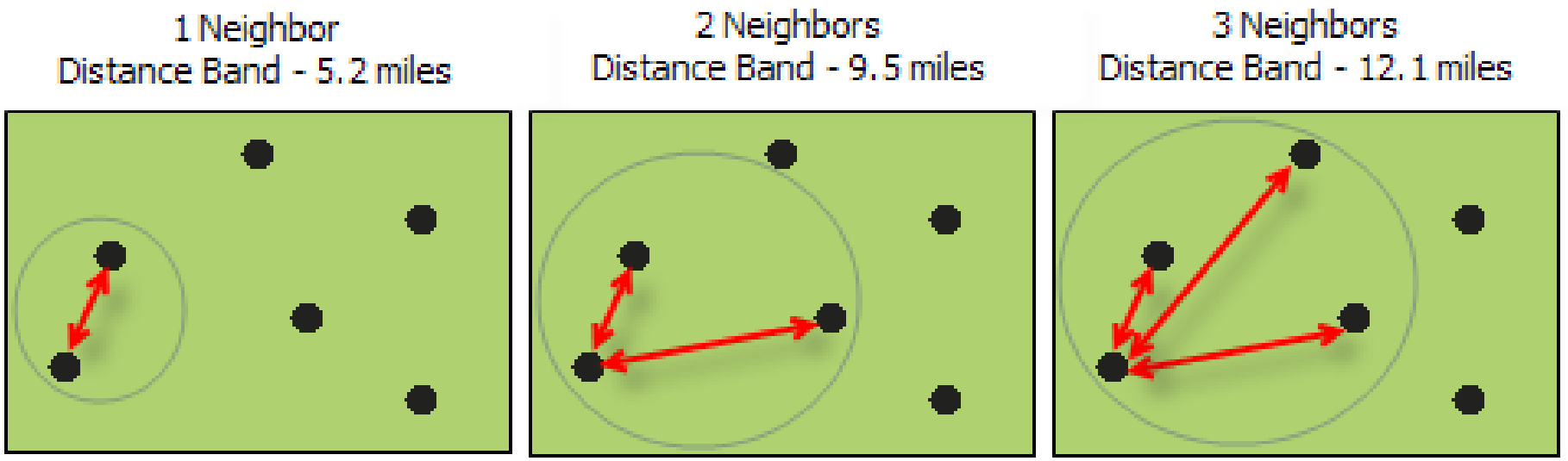
Distancias entre geometrías

- Distancias entre puntos y otras geometrías necesitan definiciones adicionales
- Dependiendo del objeto de estudio, puede ser necesario calcular la distancia mínima, media, máxima, u otra.
- Cada caso requiere un análisis ad-hoc
  - Punto a línea
  - Línea a línea
  - Punto a polígono
  - Línea a polígono
  - Polígono a polígono



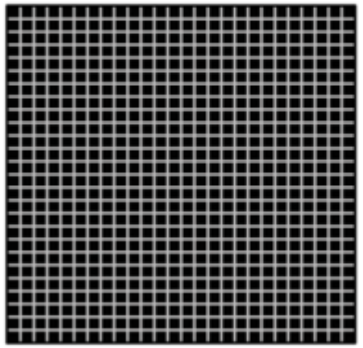
Matrices de distancias

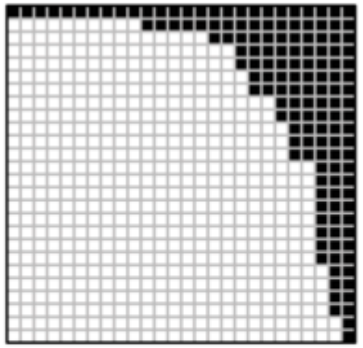
- Relación entre 2 entidades se basa en la distancia geográfica entre ellas
- La distancia se suele medir entre centroides, lo que puede ser problemático en polígonos irregulares
- Se puede medir la distancia mínima también
- Matriz de distancia contiene valores reales positivos

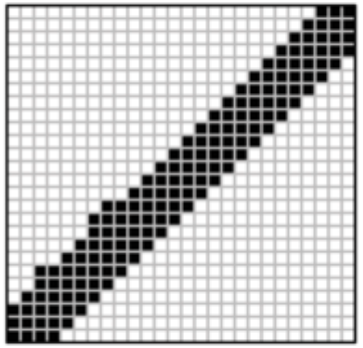


Flujos

- Se refiere a relaciones entre entidades espaciales que no se refieren explícitamente a su ubicación
- Algunos ejemplos son:
  - Migración entre entidades
  - Movilidad de personas
  - Movimiento de mercancías
  - Comercio
- Este tipo de relación entre entidades puede ser asimétrica
- Matriz de interacción contiene valores reales.

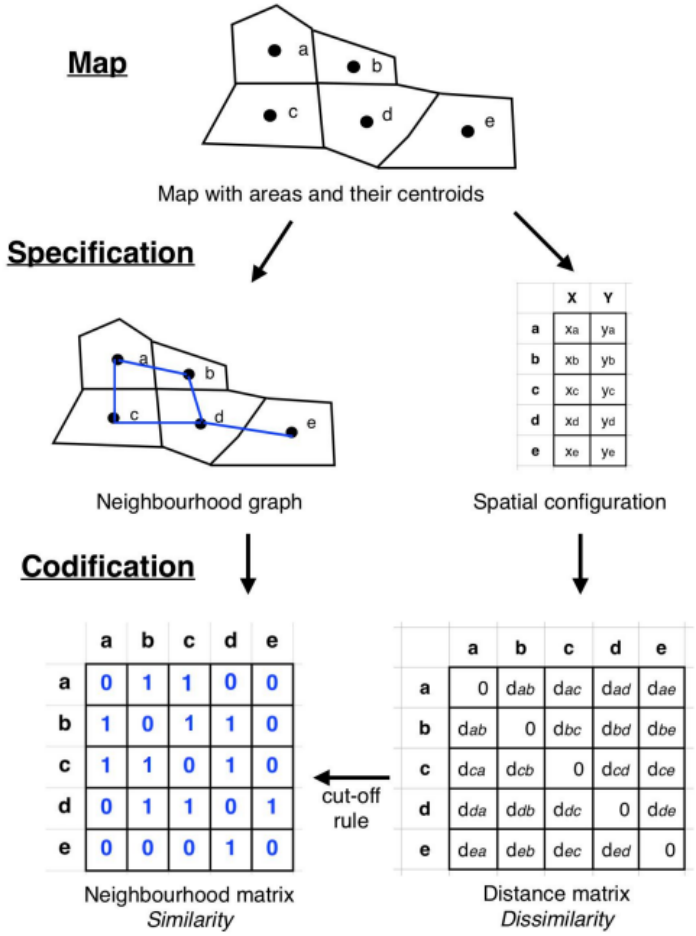






Matriz de pesos espaciales

- En el análisis estadístico tradicional se supone que cada una de las observaciones analizadas es independiente del resto
- En el territorio esto no siempre se cumple
- Para mitigar esto se incorpora en los análisis una matriz de pesos espaciales
- Estos pesos reflejan la fuerza de la interacción entre 2 entidades.
- Valores fluctúan entre 0 y 1
- Se puede construir a partir de las matrices de interacción:
  - Binaria
  - Inverso de la distancia
  - Variables de flujo normalizada



Relaciones espaciales

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

Autocorrelación espacial

Dr. Raimundo Sánchez

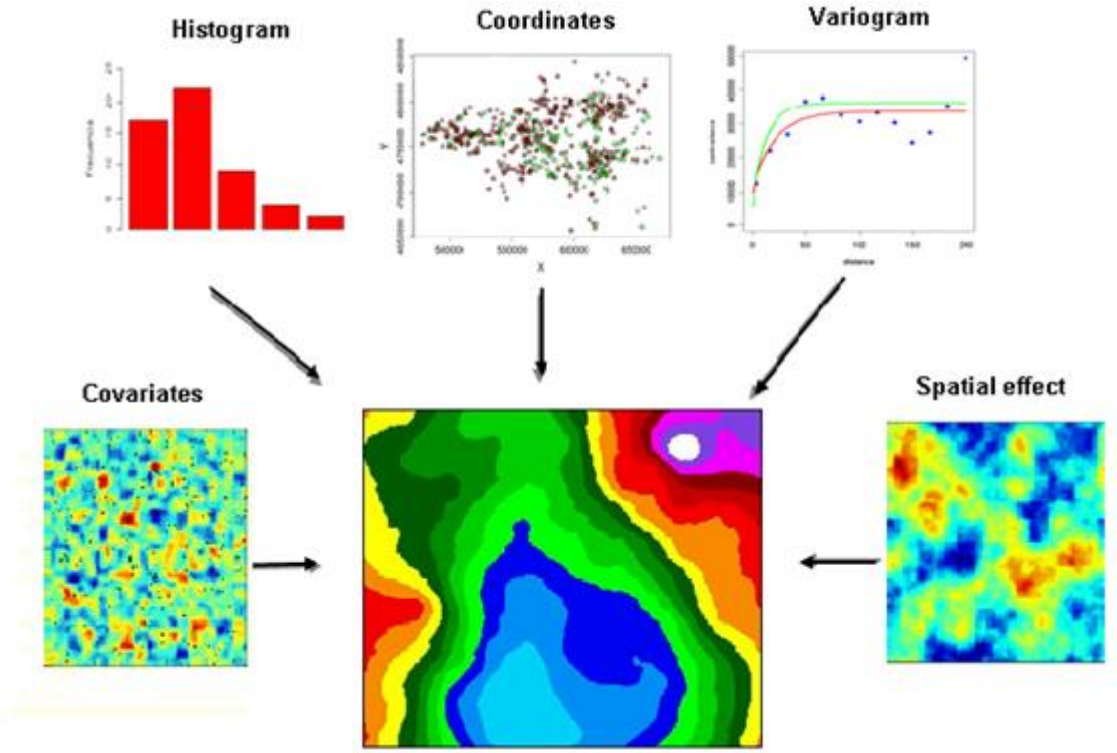
raimundo.sanchez@uai.cl

@raimun2

Geoestadística descriptiva

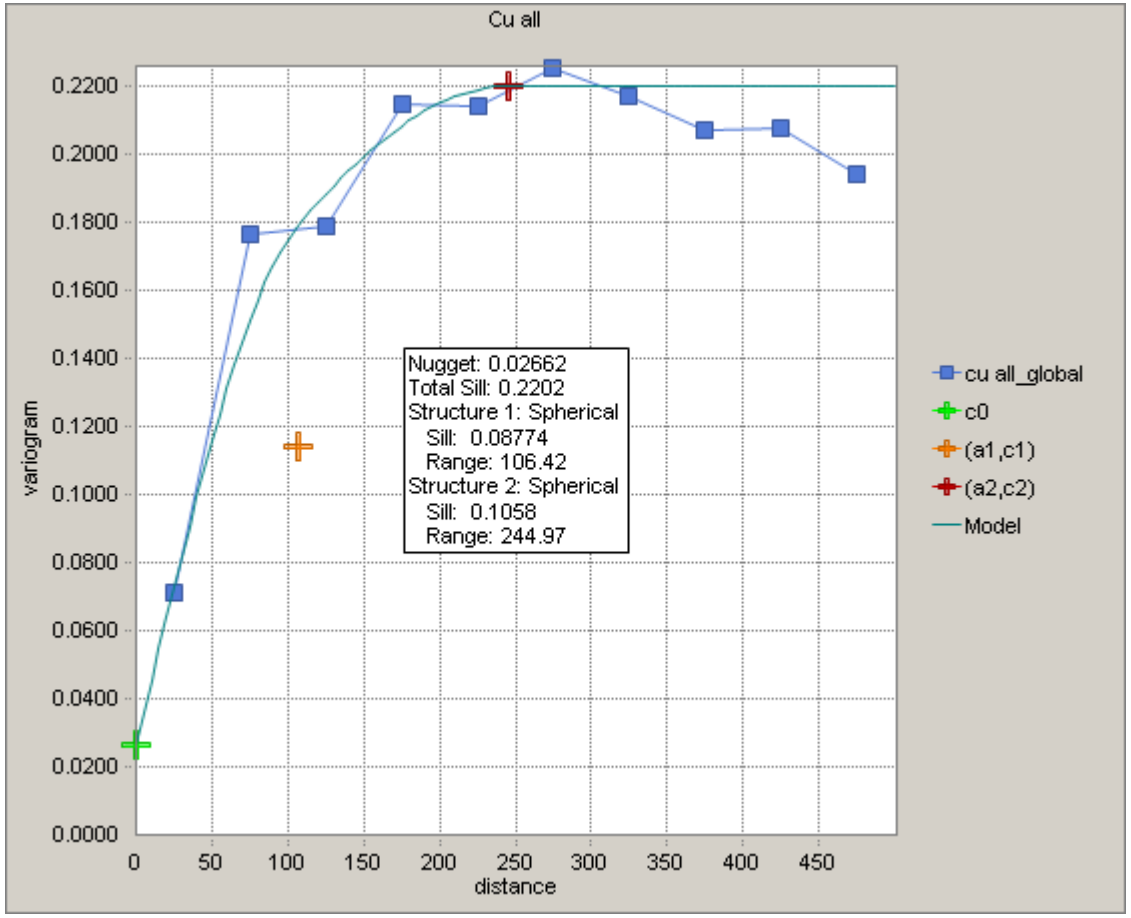
Los parámetros comúnmente utilizados para crear el resumen del comportamiento de la función geoespacial aleatoria:

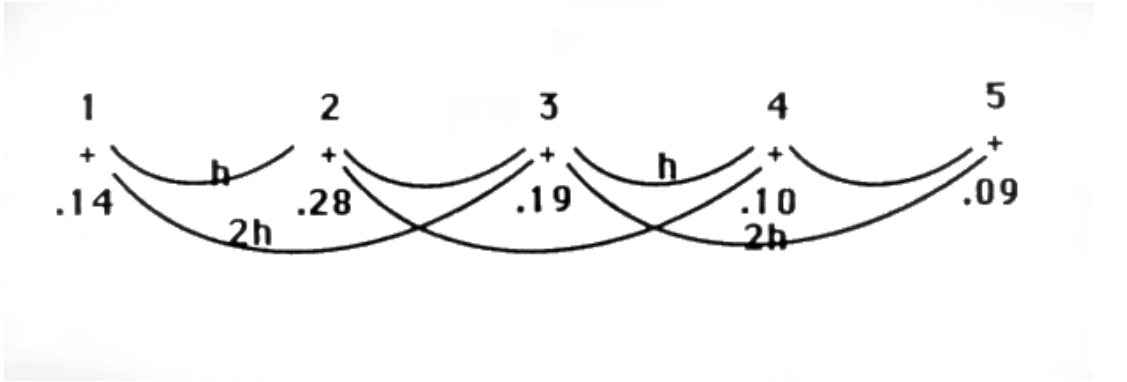
- Valor esperado
- Varianza
- Covarianza
- Variograma
- Correlograma



(semi) Variogramas

- Es una de las principales herramientas en geoestadística
- Mide la correlación entre muestras en el espacio
- Permite analizar la heterogeneidad y correlación espacial
- Se calcula en función de la distancia
- Dependiente de la distancia y dirección
- Como resultado se obtiene un vector que refleja la variación de una variable en el espacio.





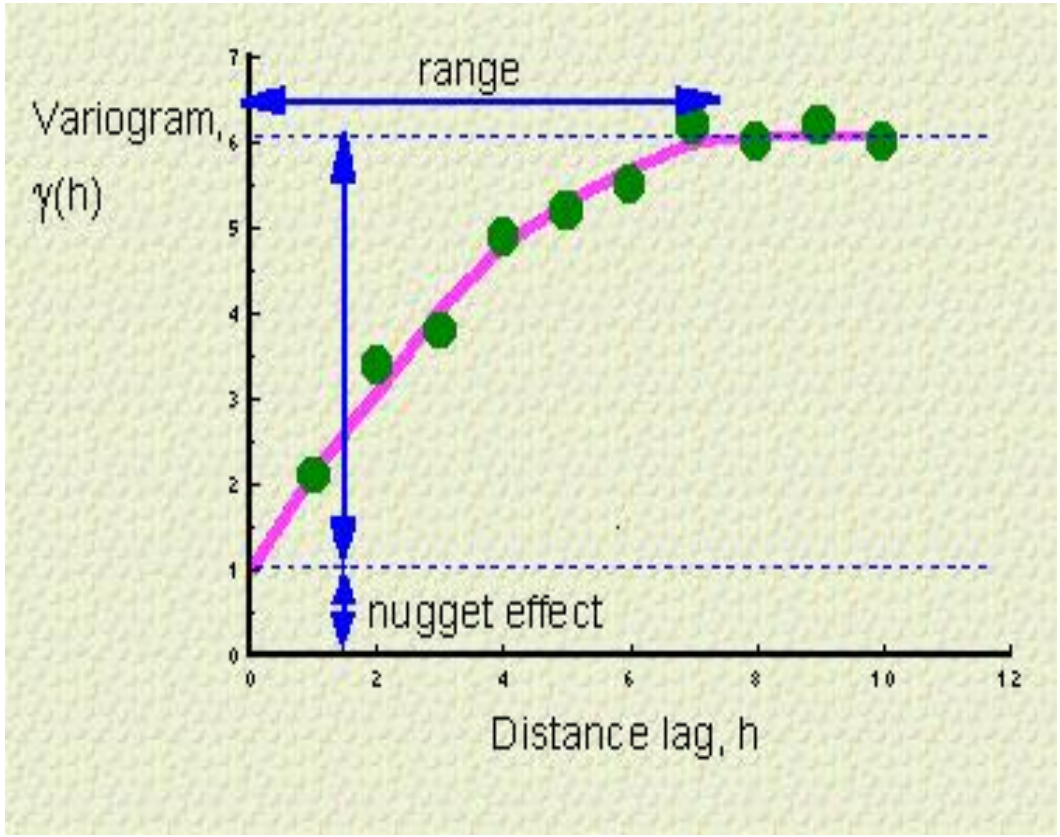
Variograma empírico

El calculo del variograma empírico esta definido por:

$$\gamma(h) = \frac{1}{2n}\sum[Z_i - Z_{i+h}]^2$$

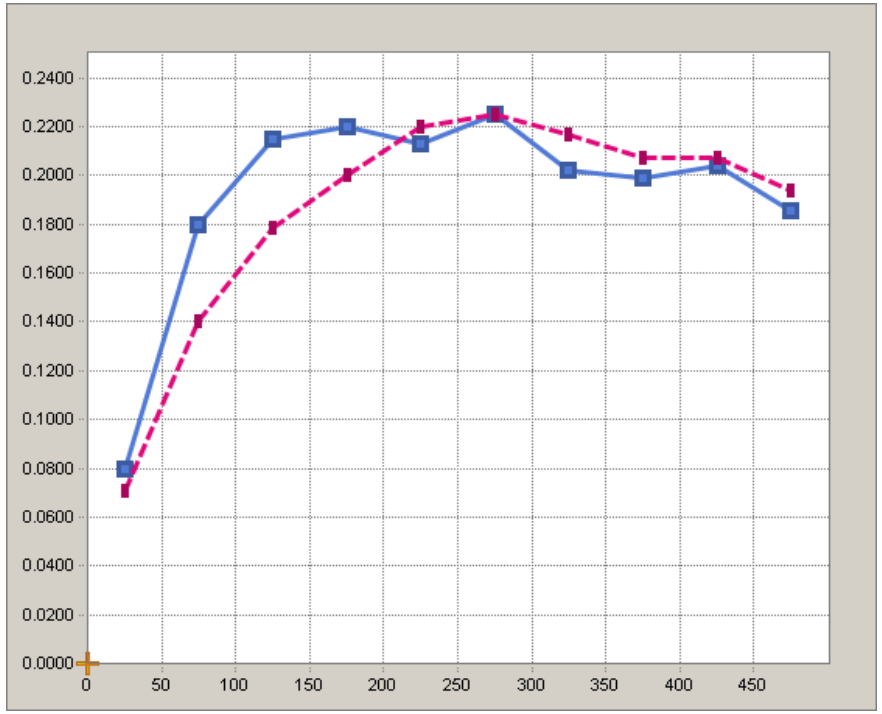
Al graficar estos valores se observan algunos parámetros empíricos:

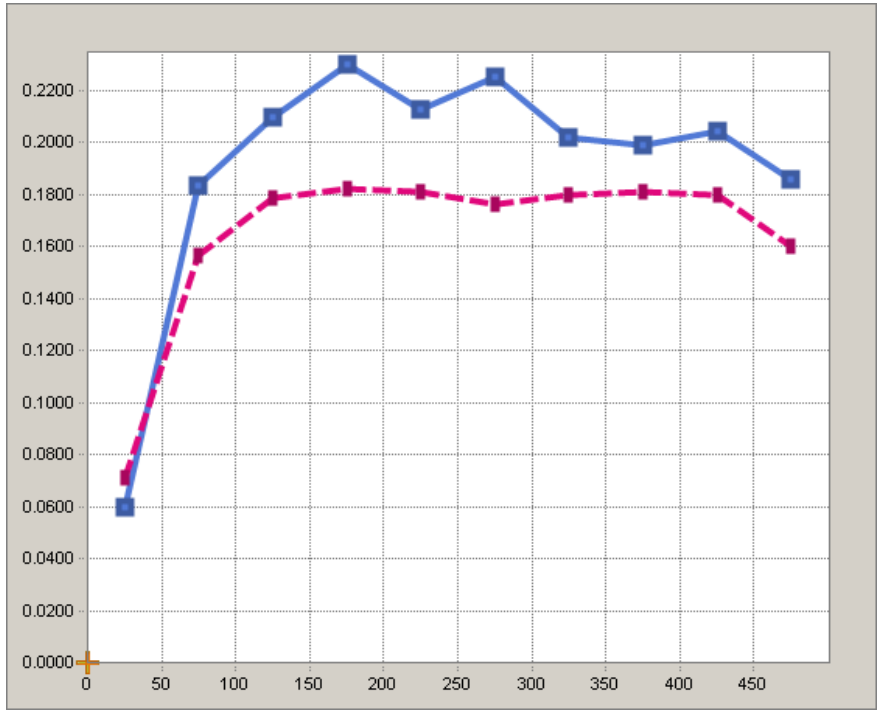
- Rango (range)
- Meseta o Umbral (sill)
- Intercepto o efecto pepita (Nugget Effect)



Anisotropía

- Es cuando las muestran no están homogéneamente distribuidas
- La distancia sobre las muestras son diferentes entre diferentes direcciones
- La mineralización, por ejemplo, puede ser mas continua en una dirección que en otra.
- El calculo del variograma empírico suele asumir isotropía
- Los variogramas se pueden calculan para diferentes direcciones.





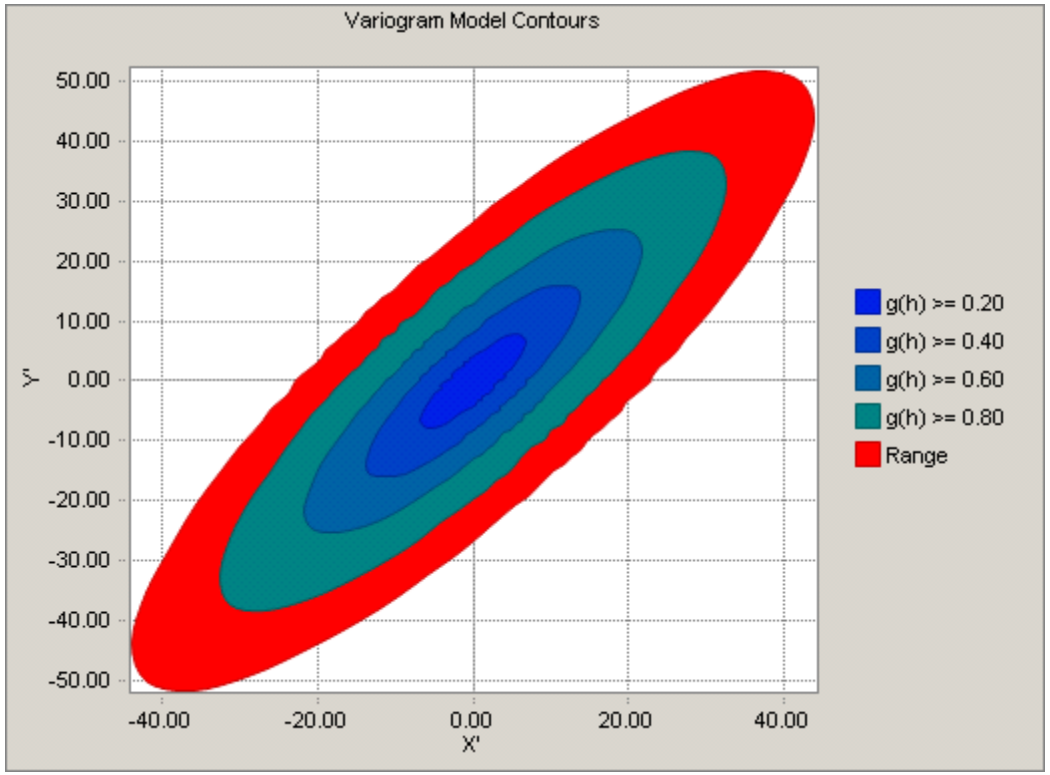
Variogramas teóricos

Lo que hacen estos modelos es ajustar una curva a los datos observados en el modelo empírico

Existen diferentes modelos que se pueden ajustar

- Esférico
- Lineal
- Exponencial
- Gaussiano

Estos modelos permiten hacer interpolaciones

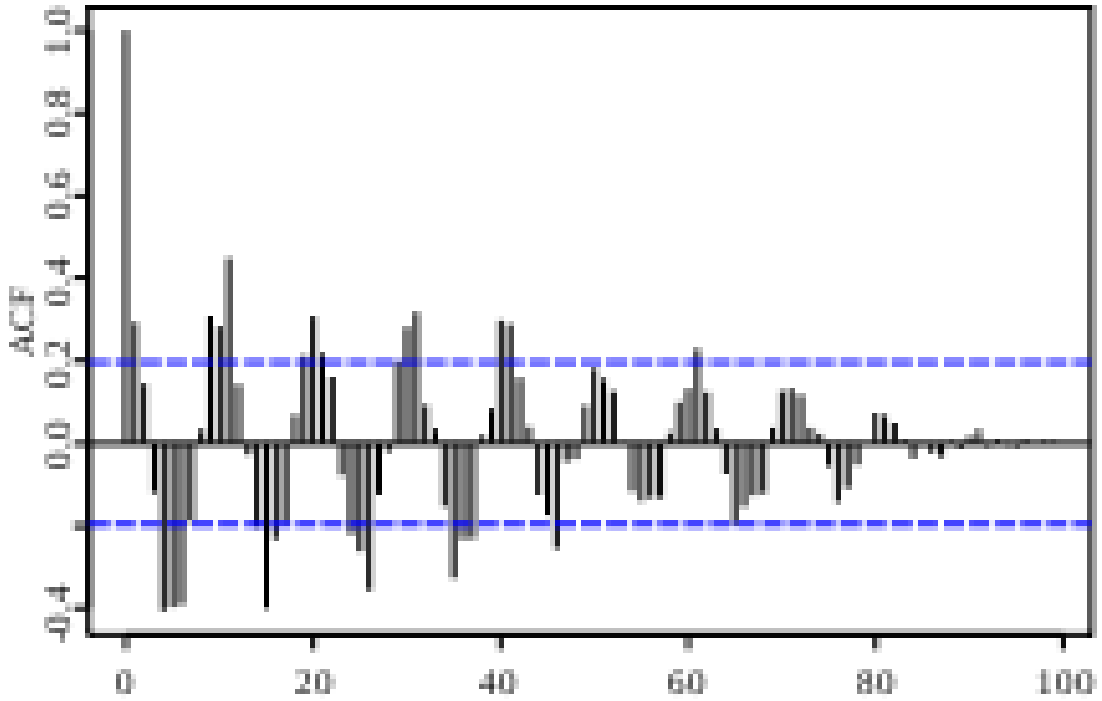


Correlogramas

- Usados para describir continuidad cruzada entre dos variables
- Permite comprender la fuerza de la relación entre 2 variables a medida que aumenta la distancia

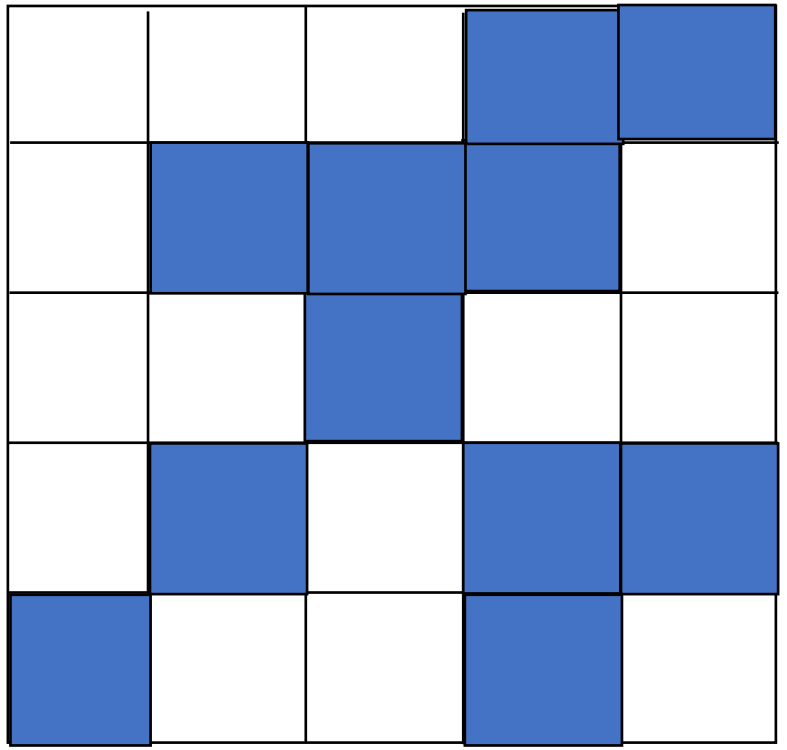
$$C(h) = \frac{1}{n}\frac{\sum(Z_i - Z_{i+h})\ (Y_i - Y_{i+h})}{\sigma_z \sigma_y}$$

- Al quitar el producto de las desviaciones estándar del denominador se obtiene un variograma de covarianzas



Autocorrelación espacial: I de Moran

- El I de Moran se comporta como el coeficiente de correlación de Pearson pero en el espacio
- Mide la probabilidad de que la concentración de una variable en el espacio no sea aleatoria
- Relacionada con la autoproducción o co-producción local de fenómenos en el espacio.
- Indicador Global de Moran: varía entre -1 (dispersión total) y 1 (autocorrelación total)



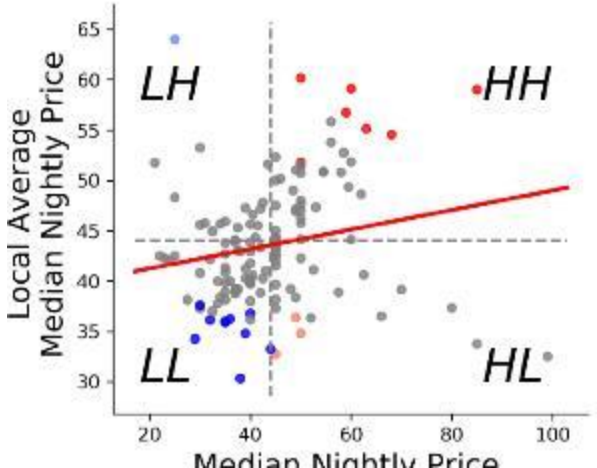
$$I = \frac{1}{p} \frac{\sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \qquad p = \sum_i \sum_j w_{ij} / n$$

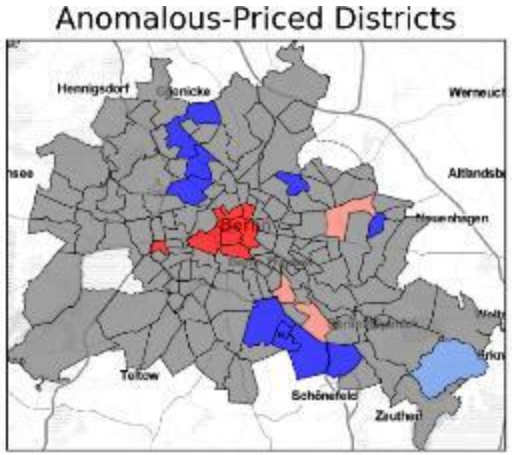
Moran I local

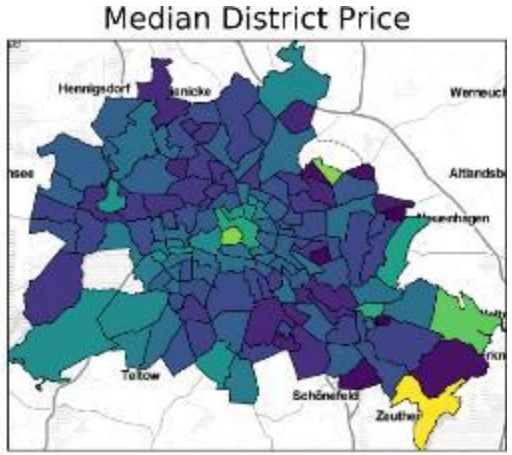
- Una versión local del I de Moran se puede escribir como

$$I_i = \frac{n}{(n-1)S^2}(z_i - \bar{z})\sum_{j=1}^{n} w_{ij}(z_j - \bar{z})$$

- La suma de la I de Moran local es proporcional al total I (escalado por W, los pesos totales):
- Calcula la autocorrelación entre cada unidad y las de su entorno
- Detecta regiones donde la autocorrelación es marcadamente diferente de otra área.
- Detecta clústeres espaciales locales (hotspots)



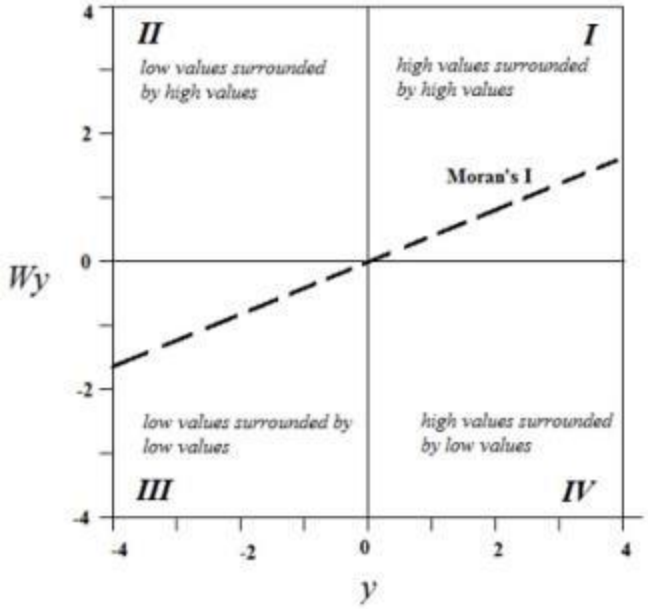




Análisis de autocorrelaciones

- El índice de Local Moran busca caracterizar el nivel y tipo de correlación que tiene un punto en el espacio con sus vecinos.
- Esta correlación se calcula en base a una variable (ej. Nivel de Educación),
- Si uno eventualmente quiere realizar el calculo en torno a 2 o mas variables debe primero consolidar estas en una vía
- El resultado es una de cinco categorías que cataloga la correlación, estas categorías son:
  - High-High (HH): Un punto con un valor alto en la variable rodeado de puntos que también tienen un nivel alto en la variable (Correlación Positiva)
  - Low-Low (LL)
  - High-Low (HL)
  - Low-High (LH)
  - No Significativo (NS): Un punto que no tiene una correlación significante con sus vecinos.
- Es importante analizar la robustez de los resultados a distintas especificaciones de la matriz.

Gráfico de Moran

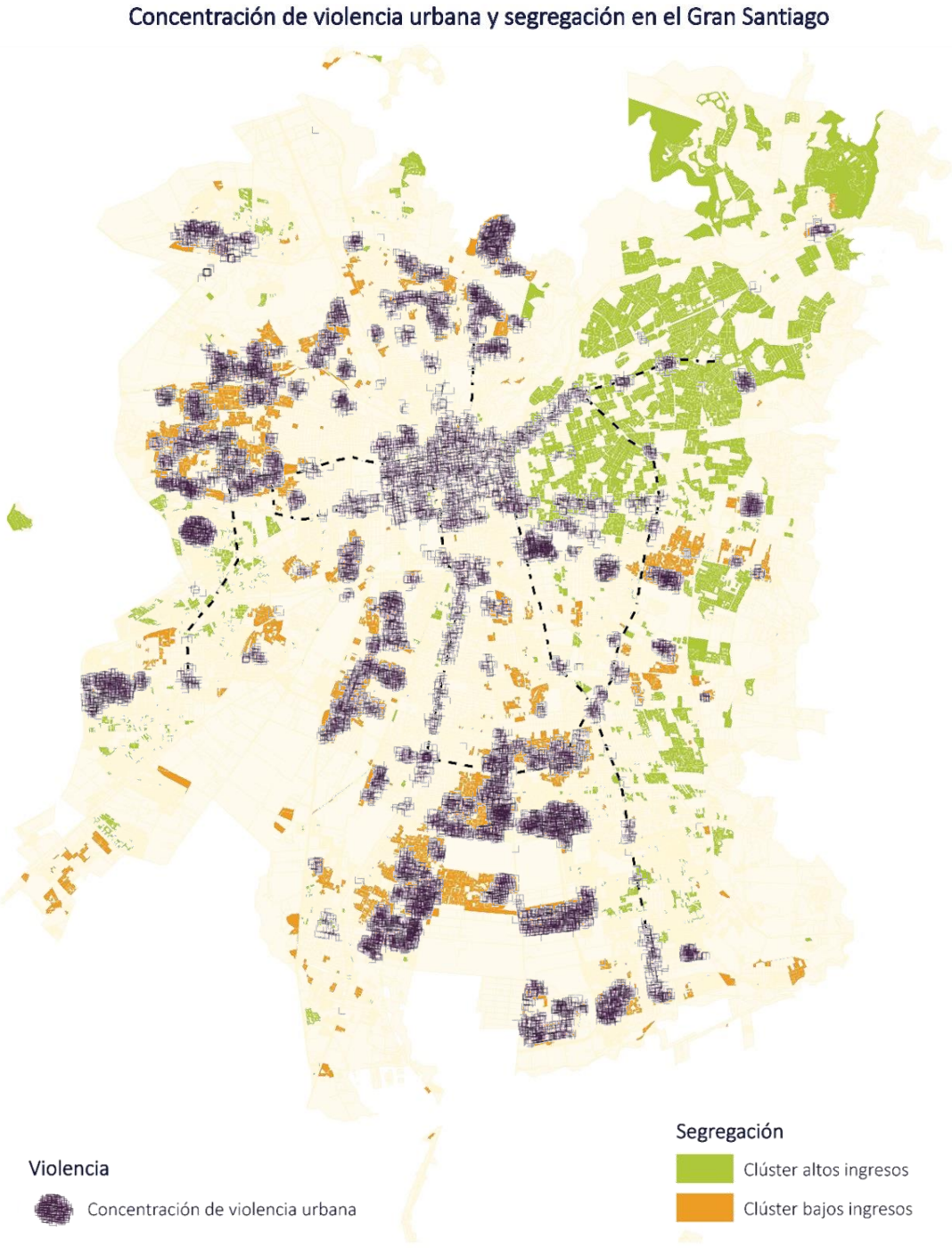


Ejemplo: segregación y delincuencia

- Índice Global de Moran: comparación entre ciudades

	Global Moran
Santiago	0.69
Buenos Aires	0.78
Lima	0.56
Ciudad de Mexico	0.67
Sao Paulo	0.74

- Índice Local de Moran: identificación de manzanas con autocorrelación de nivel socioeconómico estadísticamente significativa → identificación de zonas segregadas



Autocorrelación espacial

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

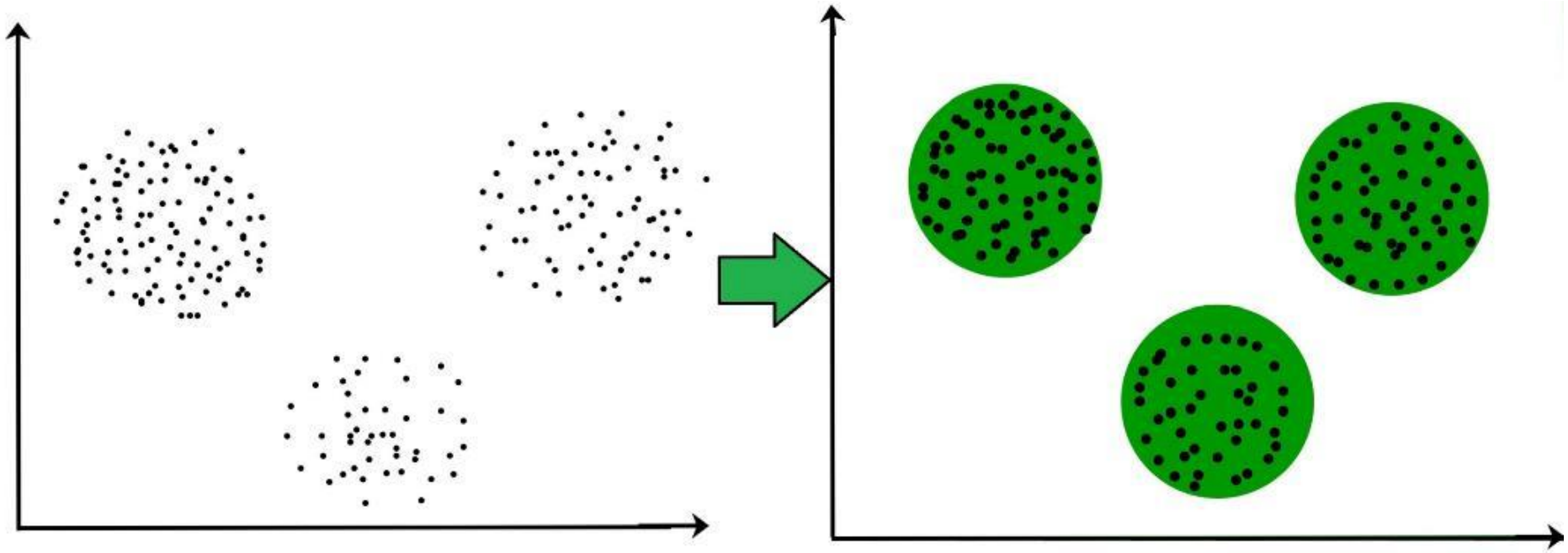
Clustering Espacial

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2

Clusters



Clusters

La noción de clúster puede ser ambigua

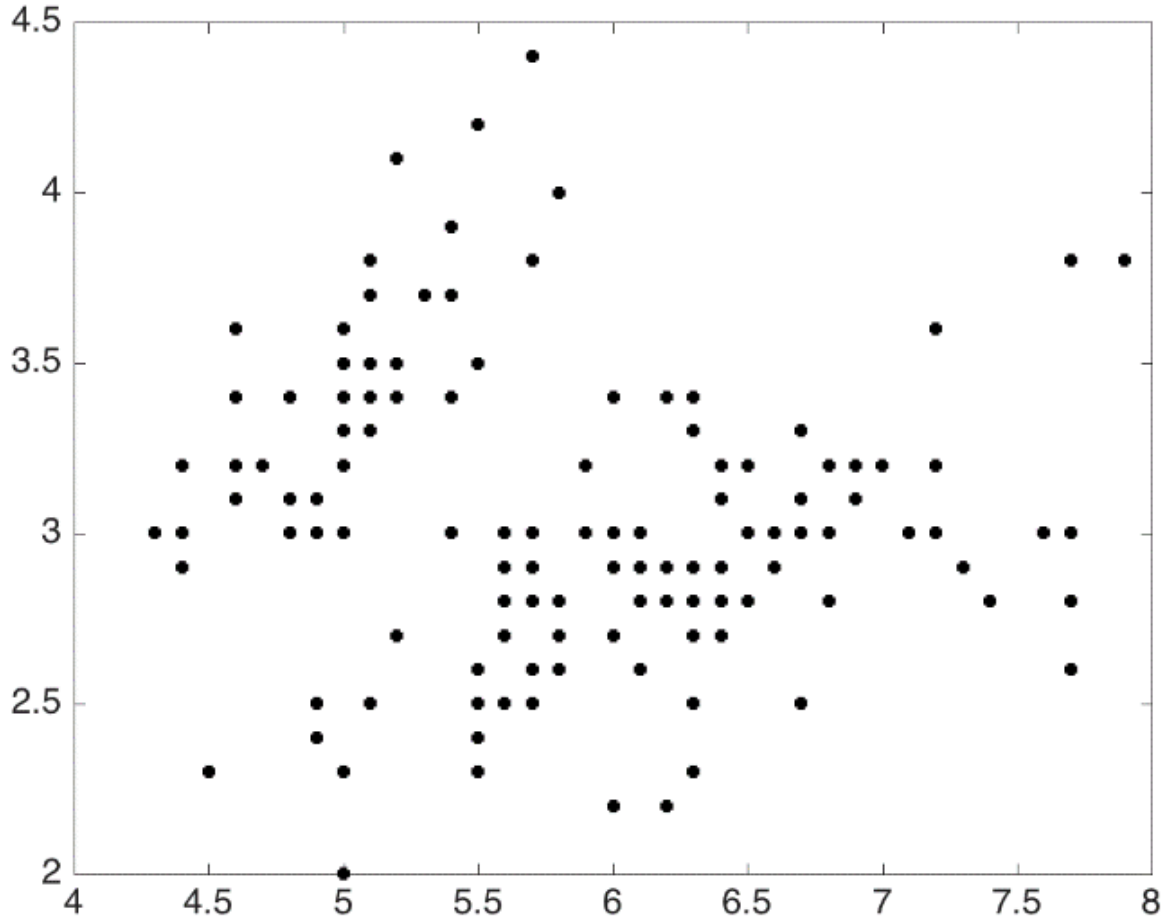
¿Cuántos clusters?

Seis grupos

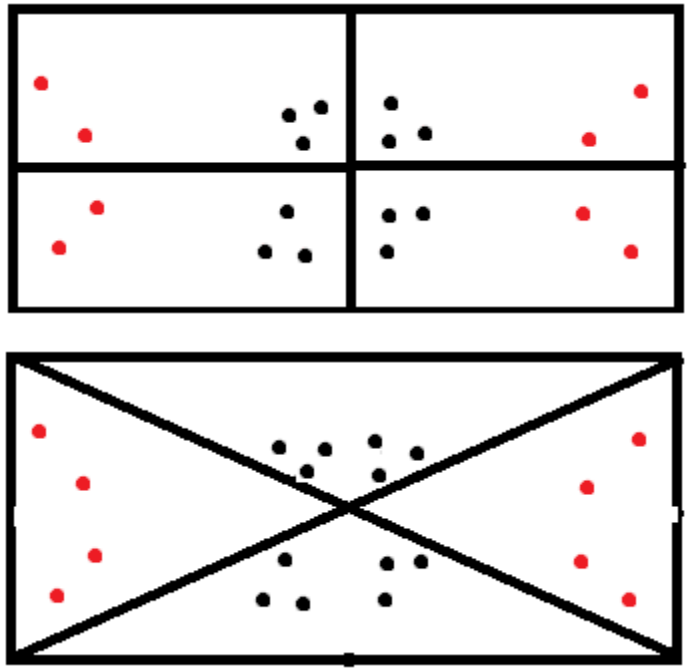
Dos grupos

Cuatro grupos

¿Cuántos clusters?

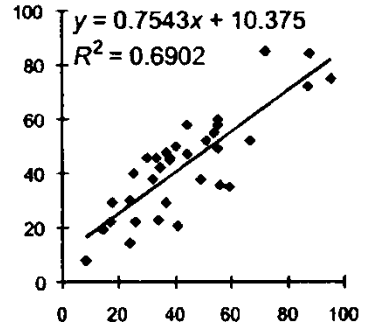


Modifiable area unit problem [MAUP]



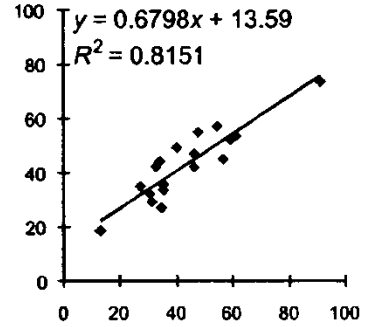
Efectos del clustering en la correlación
(escala y zonificación)
Gehlke & Biehl

Independent variable						Dependent variable					
87	95	72	37	44	24	72	75	85	29	58	30
40	55	55	38	88	34	50	60	49	46	84	23
41	30	26	35	38	24	21	46	22	42	45	14
14	56	37	34	8	18	19	36	48	23	8	29
49	44	51	67	17	37	38	47	52	52	22	48
55	25	33	32	59	54	58	40	46	38	35	55



Aggregation scheme 1

91	54.5	34	73.5	57	44
47.5	46.5	61	55	47.5	53.5
35.5	30.5	31	33.5	32	29.5
35	35.5	13	27.5	35.5	18.5
46.5	59	27	42.5	52	35
40	32.5	56.5	49	42	45



Aggregation scheme 2

63.5	75	63.5	37.5	66	29	61	67.5	67	37.5	71	26.5
27.5	43	31.5	34.5	23	21	20	41	35	32.5	26.5	21.5
52	34.5	42	49.5	38	45.5	48	43.5	49	45	28.5	51.5

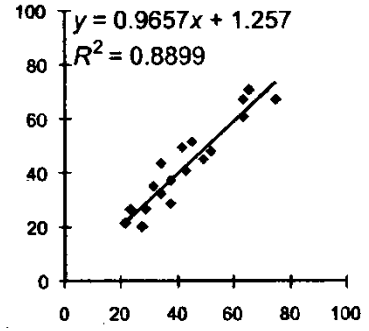


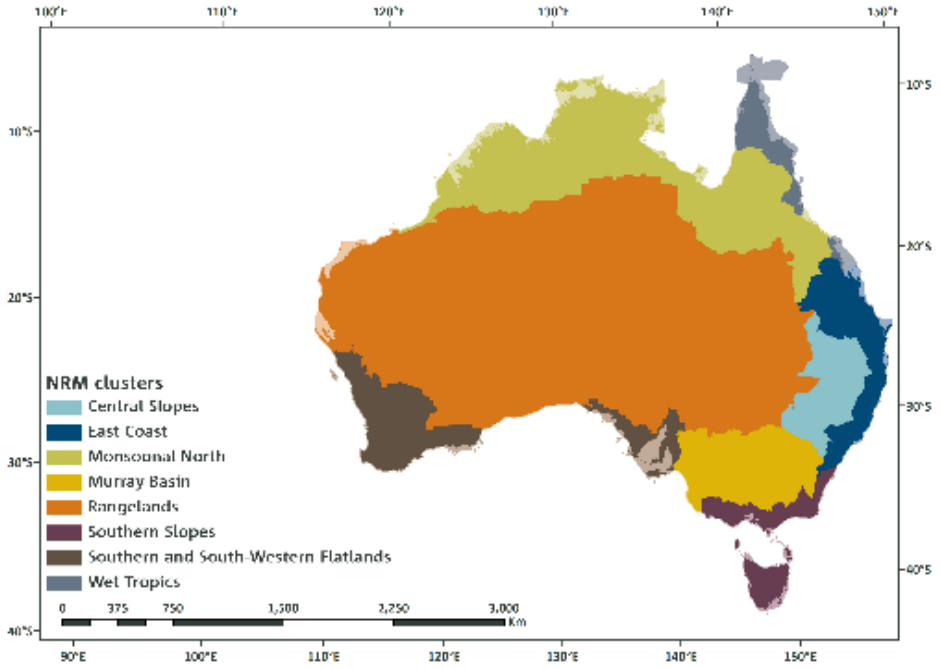
Figure 2.1 Modifiable areal unit problem.

Objetivos

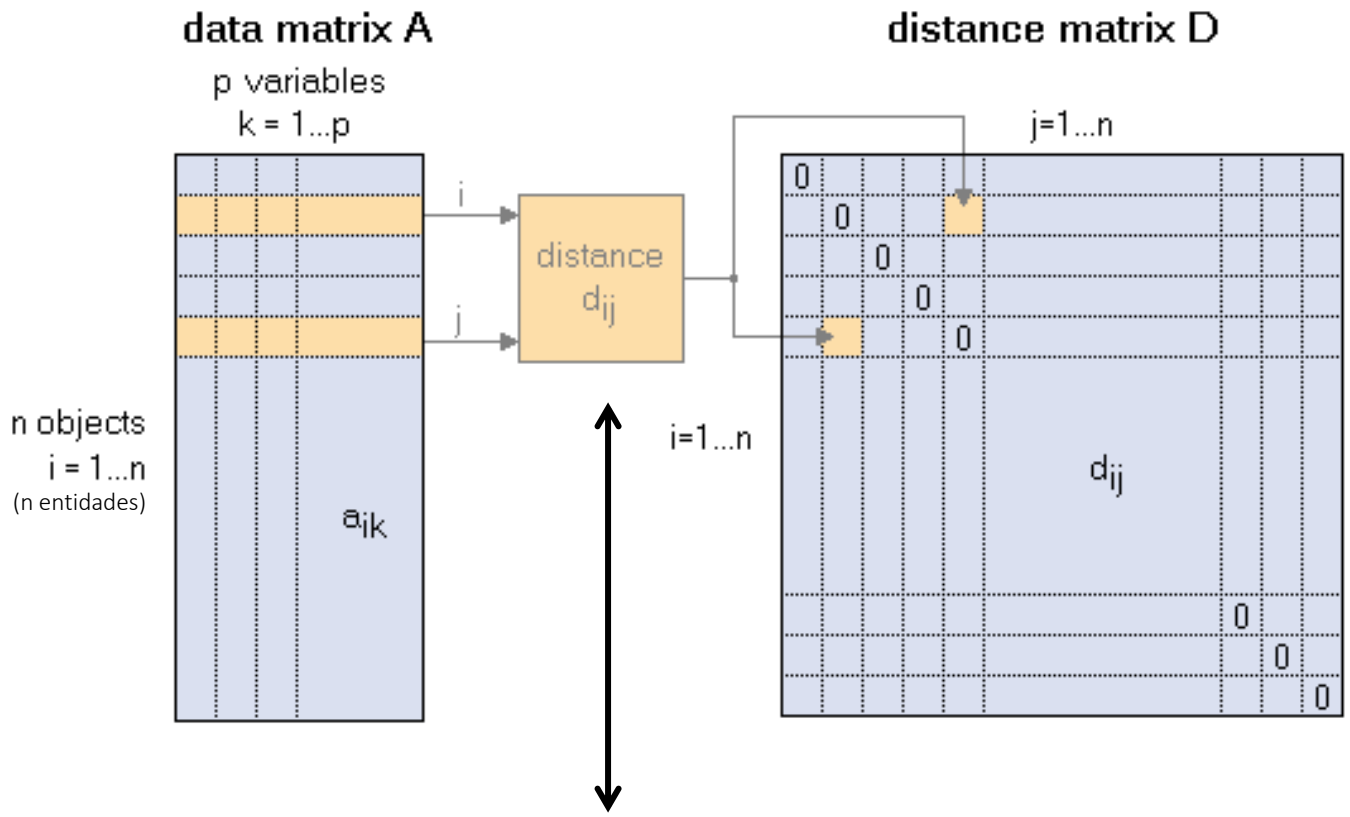
Encontrar **algorítmicamente** grupos de entidades tales que:

1. **La similitud intragrupo es alta**
2. **La similitud entre grupos es baja**
3. **Minimice el efecto del MAUP**

Las medidas de distancia y similitud son cruciales en este proceso



Distancias de atributos vs geográfica



(Euclidean, Manhattan, Chebyshev, Mahalanobis, Cosine, Haversine, etc.)

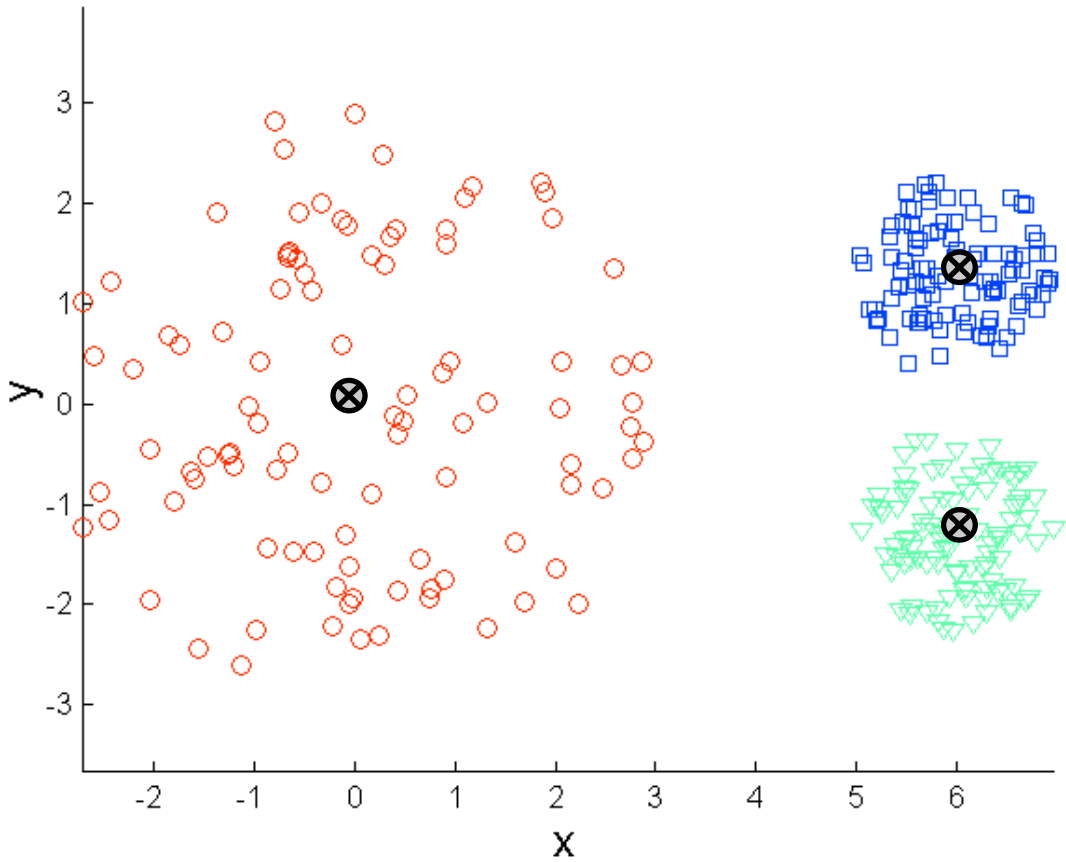
Enfoques al problema

Existen diferentes formas de abordarlo:

- Clusterizar atributos
- Clusterizar coordenadas
- Clusterizar combinación de atributos y coordenadas
- Clusterizar combinación de atributos y variable categórica territorial
- Clusterizar atributos sujeto a restricción de vecindad

Métodos basados en particiones

Los datos se separan en grupos, a los que pertenece cada punto **exclusivamente** a un solo grupo.

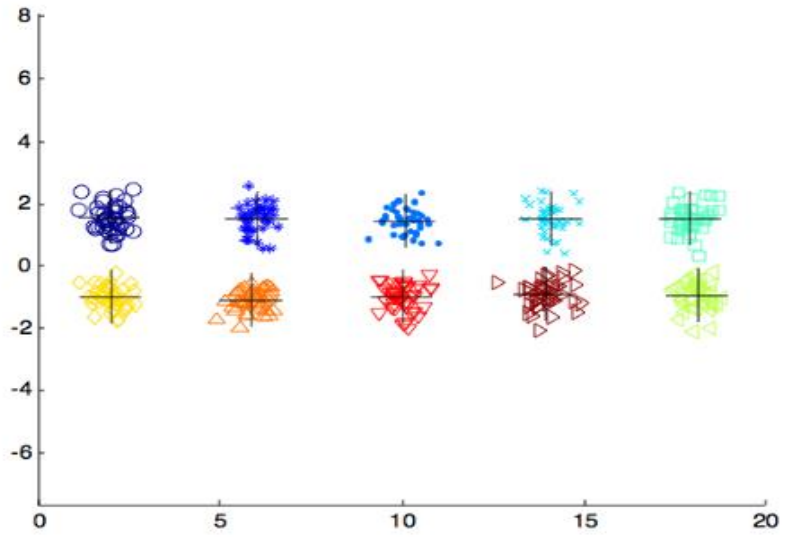


K-medias

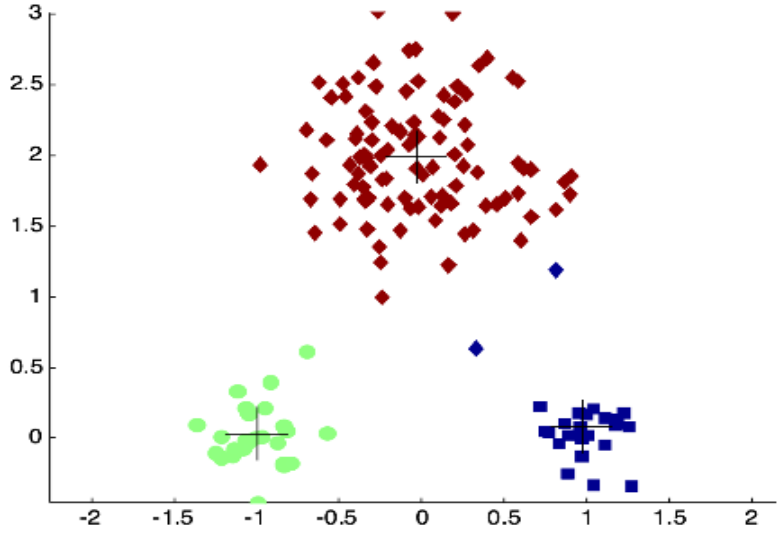
Uno de los algoritmos de clusters más simples.

Dado un número **K** de clusters (determinado por el usuario), cada **cluster** está asociado con un centroide y cada entidad se asigna al cluster con el centroide más cercano.

Variantes como K-medioides, o K-modas usan otros estadísticos como centroides



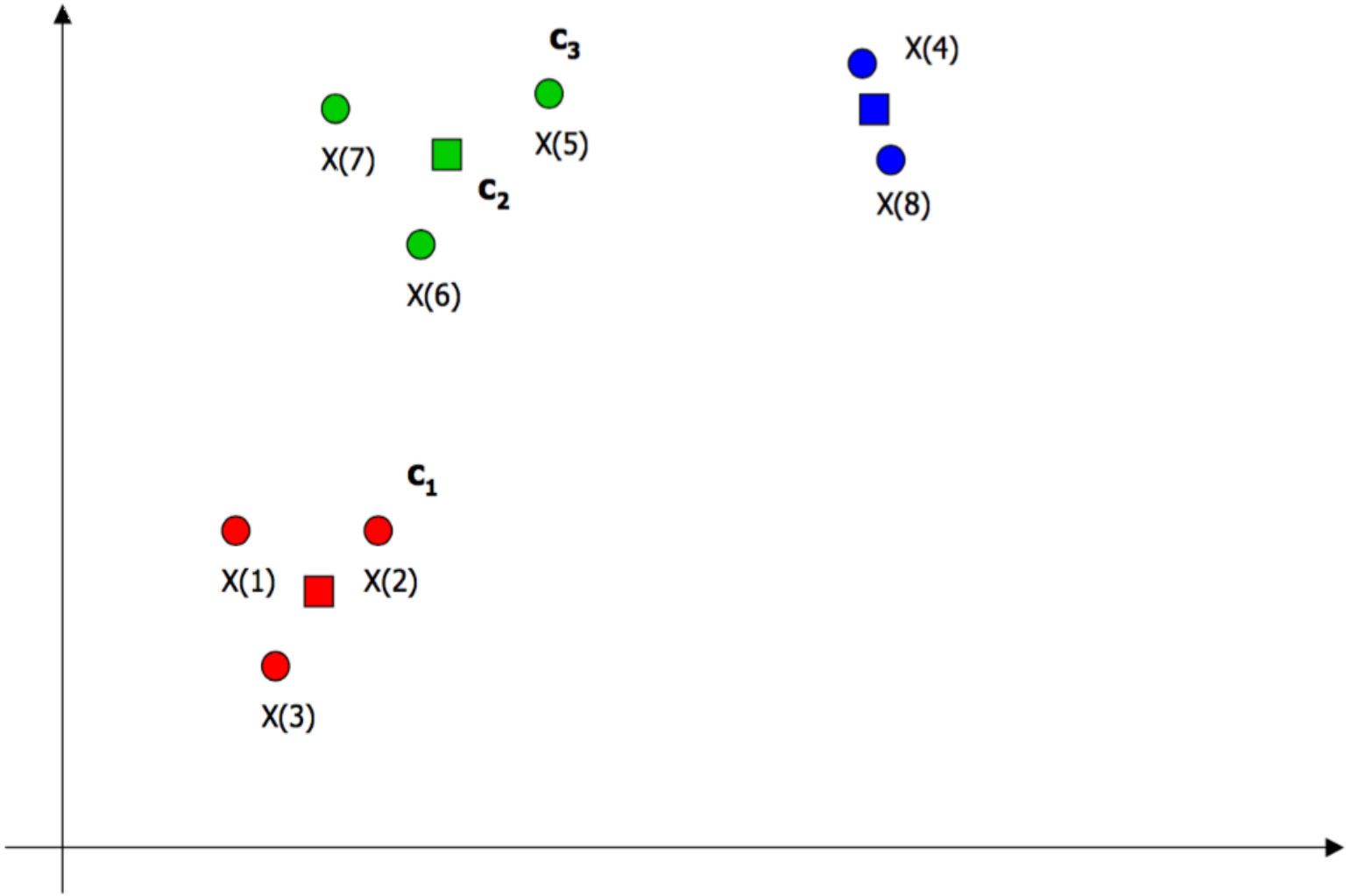
K = 10



K = 3

Algoritmo

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change



Fortalezas:

- Relativamente eficiente
- Encuentra grupos esféricos

Debilidades:

- Termina en el óptimo local
- Sensible a condiciones iniciales
- Aplicable solo cuando la media está definida (variables continuas)
- Necesita especificar K
- No funciona bien con grupos de diferente densidad
- Susceptible a valores atípicos

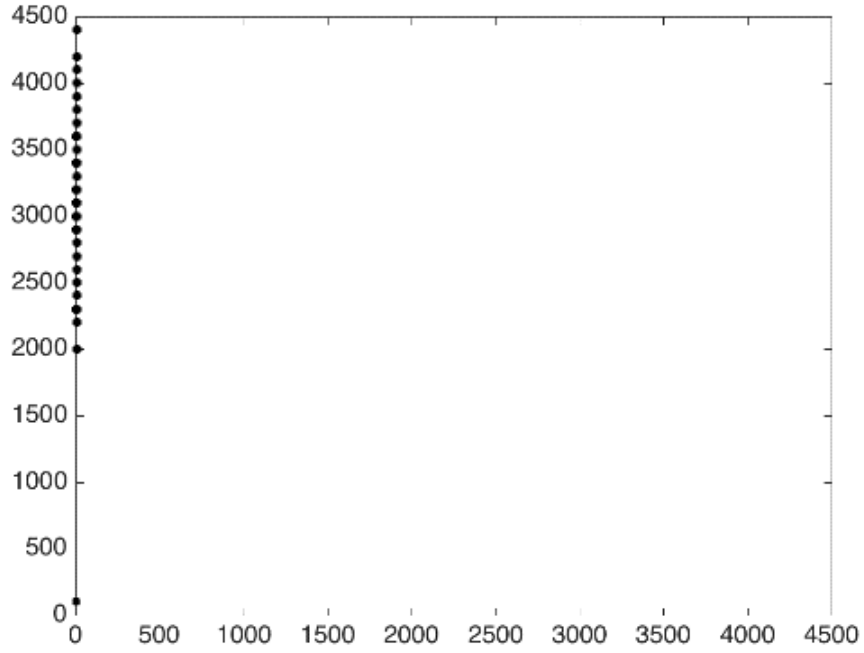
Pre procesamiento

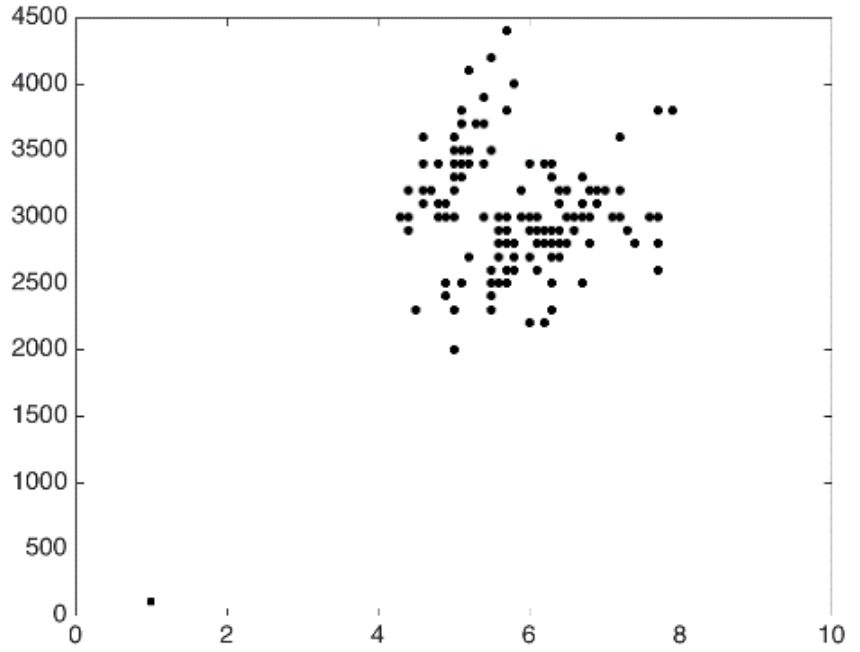
Los valores atípicos afectan el desempeño de la mayoría de los modelos.

La escala de los datos también puede jugar en contra

Por lo tanto, es necesario pre procesar los datos:

- Normalizar los datos
- Lidiar con valores atípicos





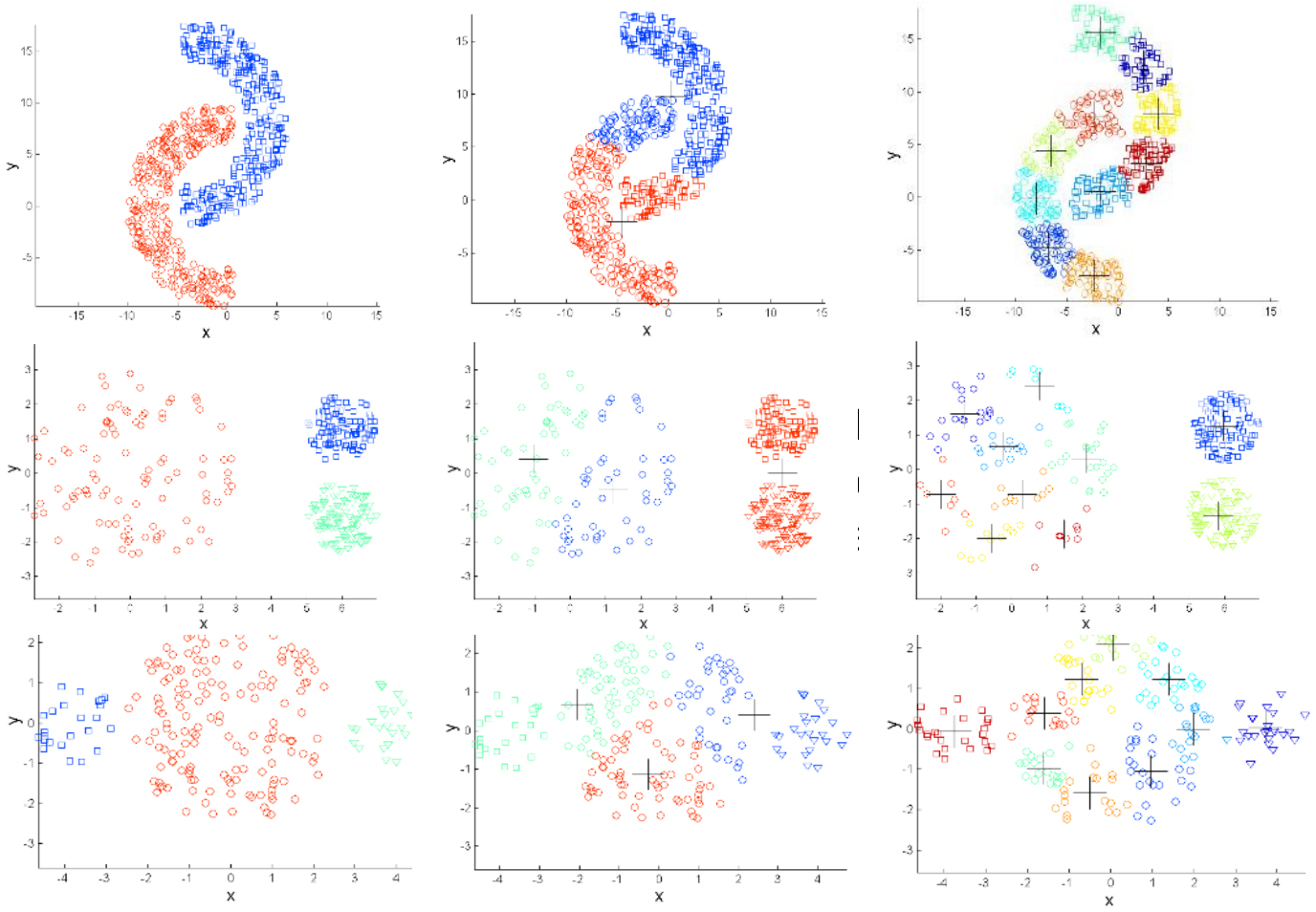
Post procesamiento

Algunos problemas se pueden resolver a través de fusión y división de clusters.

Métodos tradicionales generan clústeres esféricos

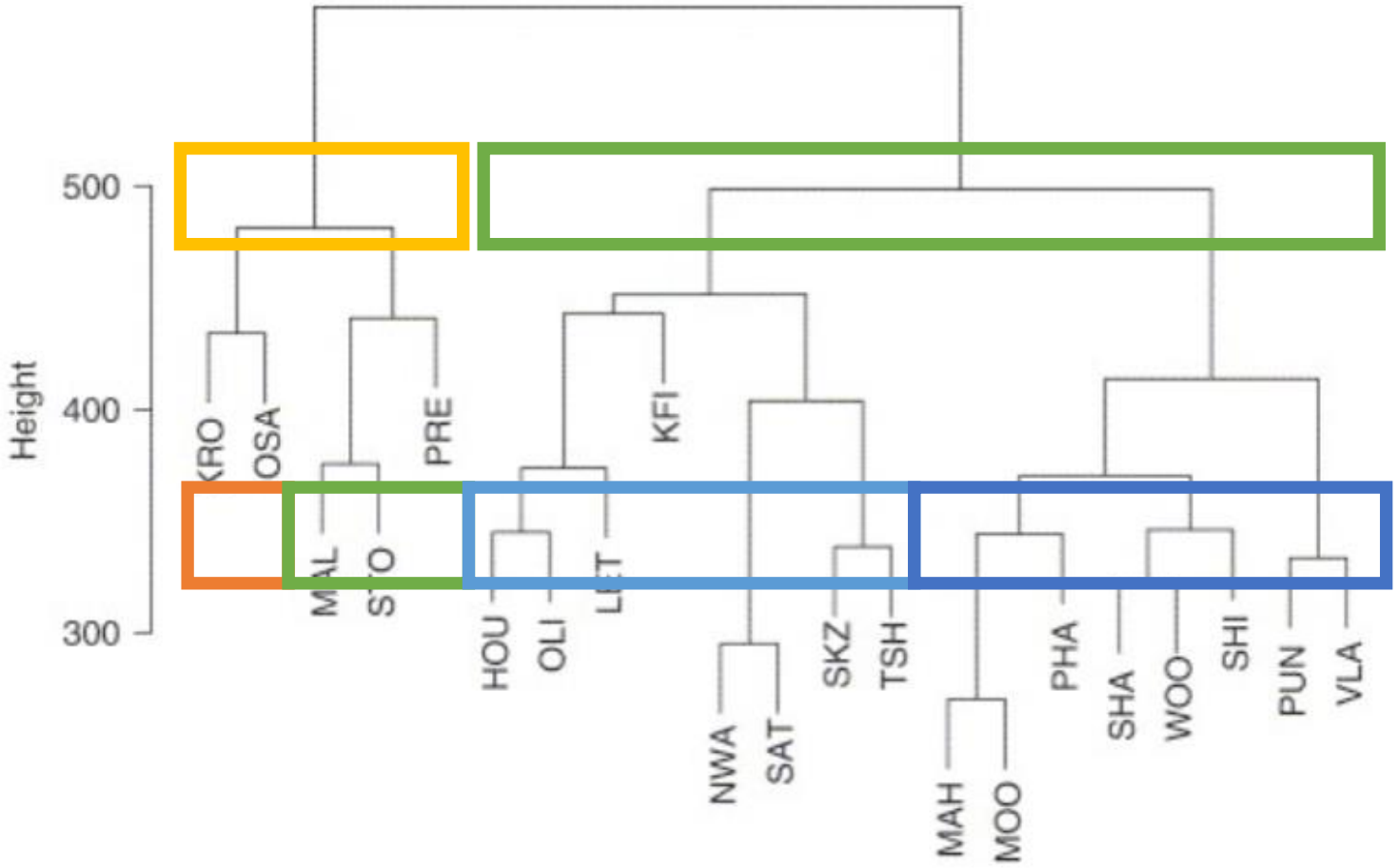
La fusión de clústeres mas pequeños podría mitigar este problema.

División de clústeres mas dispersos también mejora el desempeño



Métodos jerárquicos

Las entidades se agrupan en una jerarquía de clústeres anidados.



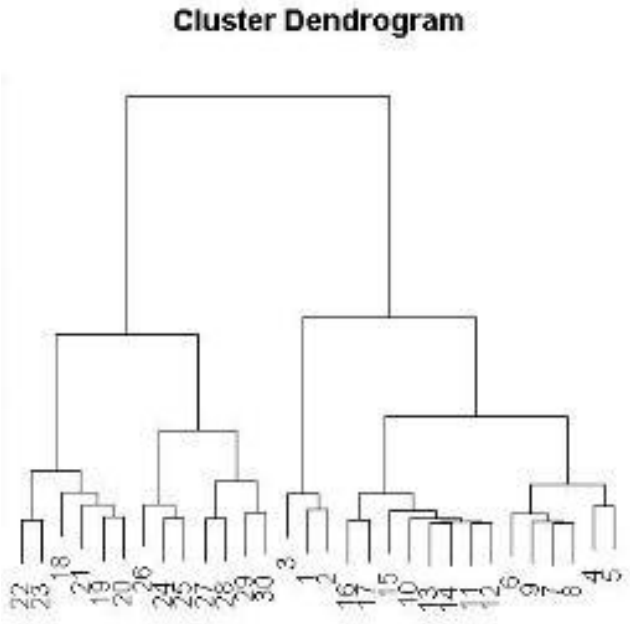
Clustering Jerárquico

El algoritmo básico para clustering aglomerativo es sencillo

1. Deje que cada punto de datos sea un clúster
2. Calcular la matriz de proximidad (matriz de distancia entre cada clúster)
3. Repetir hasta que sólo quede un solo clúster
    1. Fusionar los dos clústeres más cercanos
    2. Actualizar la matriz de proximidad

El paso clave es el cálculo de la proximidad de dos clústeres

Diferentes enfoques para definir la distancia entre clústeres distinguen los diferentes algoritmos



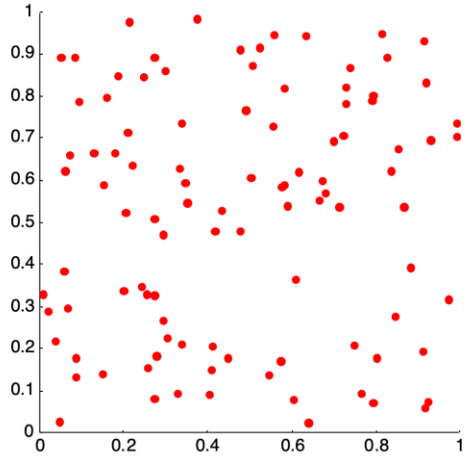
Evaluar la bondad de los clústeres resultantes

La evaluación de estructuras de clusters es la parte más difícil y frustrante del análisis de clústeres

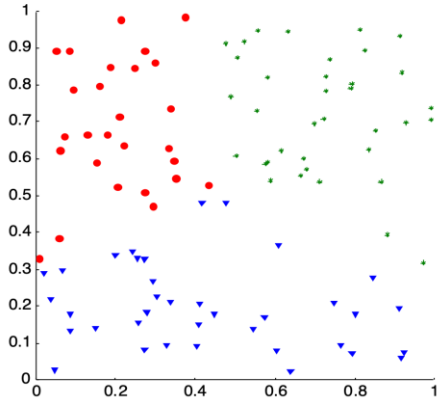
La calidad de un clúster es difícil de evaluar porque no conocemos los clústeres correctos

Hacerlo nos ayudará a:

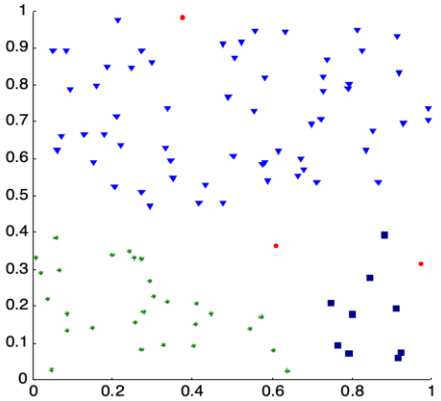
- Evitar encontrar patrones espurios
- Comparar algoritmos de agrupación en clústeres
- Comparar dos conjuntos de clústeres



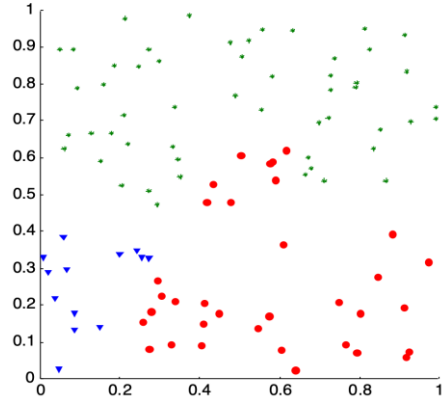
Random Points



K-means



DBSCAN

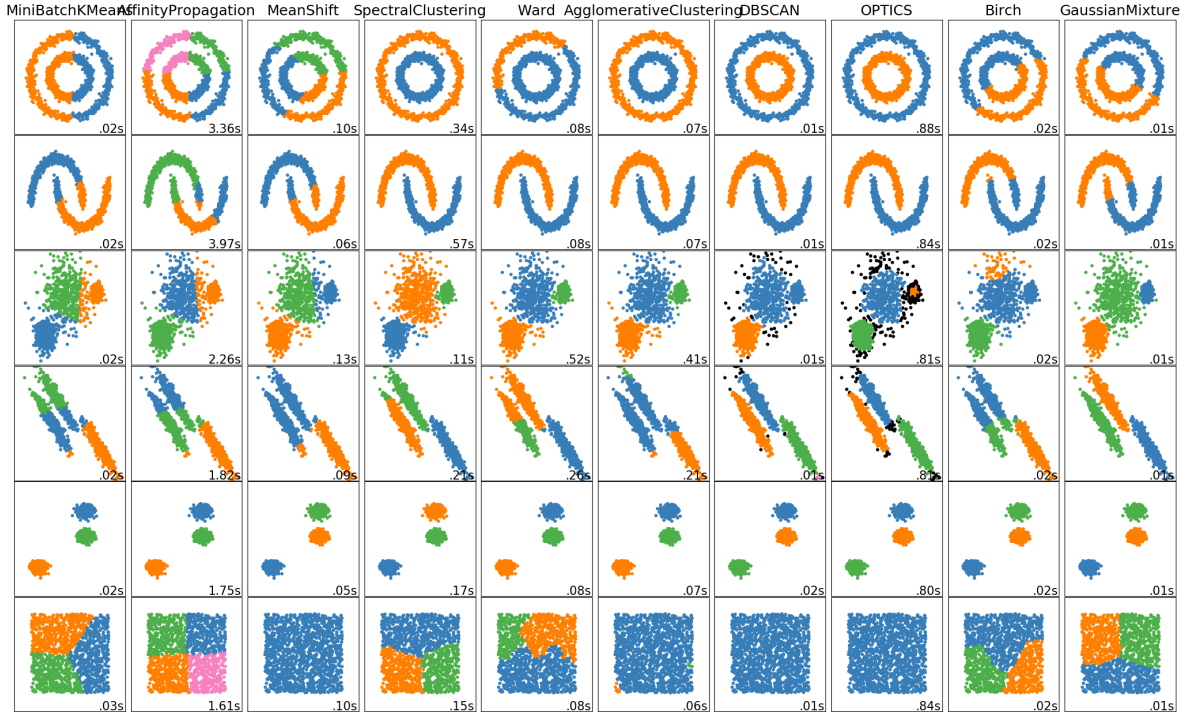


Complete Link

Evaluación de modelos no-supervisados

Existen diferentes métodos para evaluar clústeres:

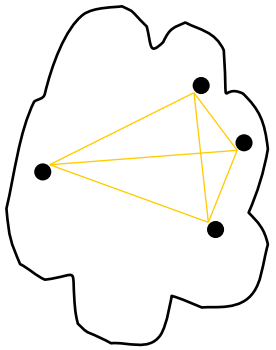
- Inspección visual basada en la matriz de proximidad
- Correlación entre similitud y resultados de agrupación en clústeres
- Estadístico de Hopkins
- Medidas internas: Coeficiente de cohesión, separación y silueta.



Coeficiente de cohesión

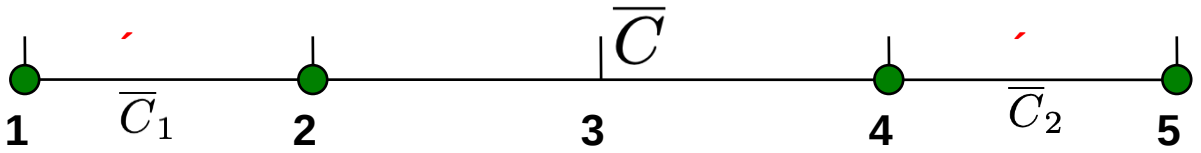
Mide cuán estrechamente relacionados están los objetos dentro de cada clúster.

Suma de errores cuadrados (SSE) es la suma de la distancia cuadrada de un punto al centroide de su clúster.



cohesion

$$SSE_{total} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \overline{C}_i)^2$$



$$\text{K=1} \Rightarrow SSE_{total} = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

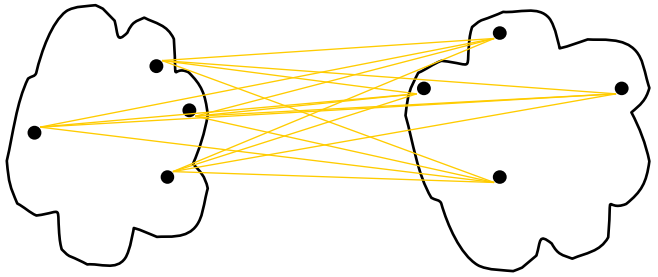
$$\text{K=2} \Rightarrow SSE_{total} = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

Coeficiente de separación

Mide cuán distinto es un clúster de los otros clusters.

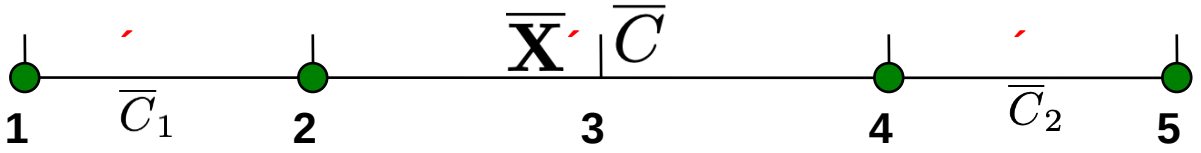
La suma de cuadrados intra grupo (SSB) es la suma de la distancia cuadrada de un centroide de clúster a la media general.

Minimizar la cohesión equivale a maximizar la separación.



separación

$$SSB_{total} = \sum_{k=1}^{K} |C_i|(\overline{C}_i - \overline{\mathbf{X}})^2$$



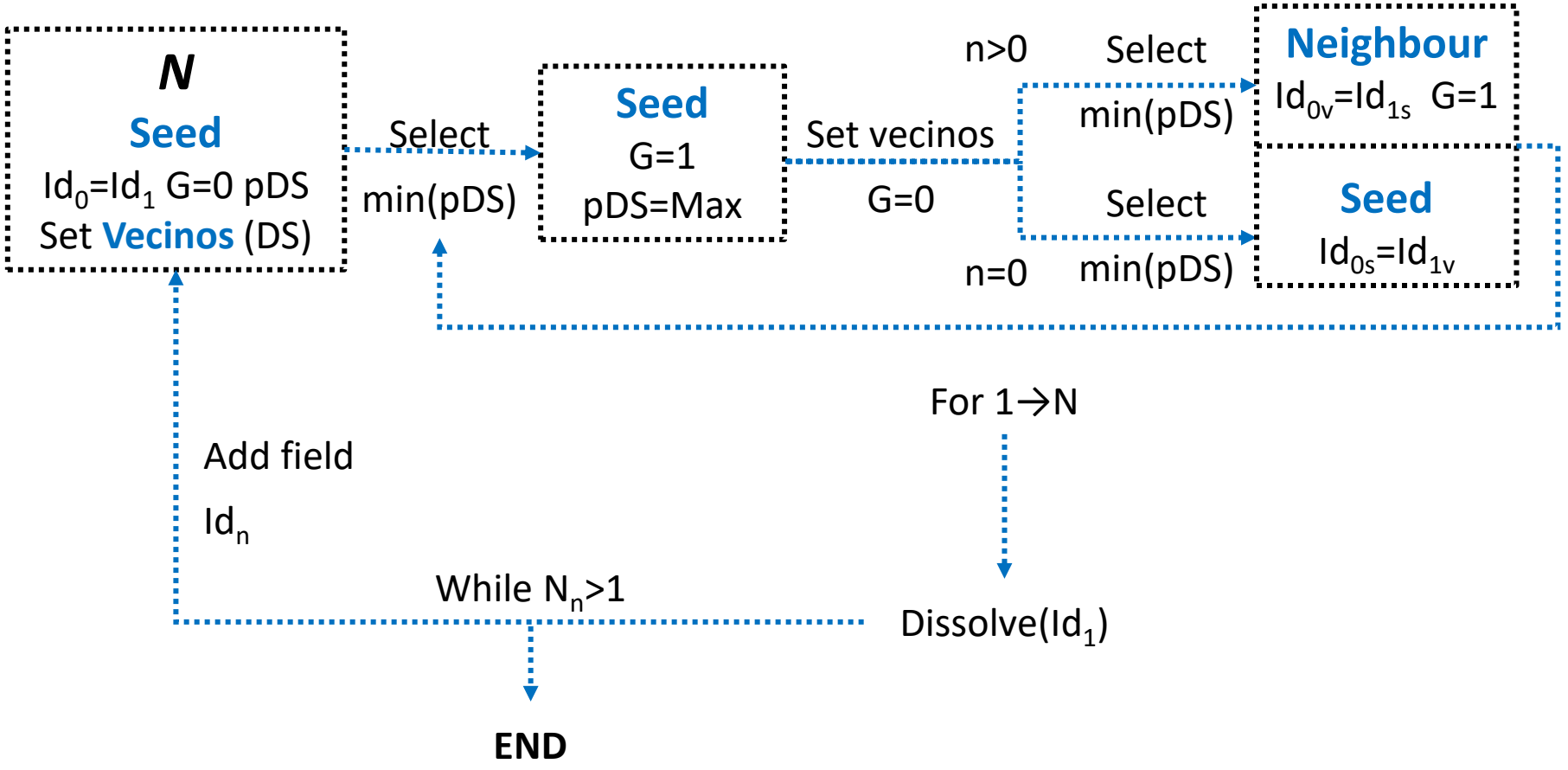
$$\text{K=1} \Rightarrow SSB_{total} = 4 * (3-3)^2 = 0$$

$$\text{K=2} \Rightarrow SSB_{total} = 2 * (1.5-3)^2 + 2 * (4.5-3)^2 = 9$$

Regionalización

- Regionalización es un caso particular de clustering de atributos, con restricción de contigüidad
  - Agregando
  - Fraccionamiento
- La regionalización jerárquica es la mas aceptada para tratar el efecto de escala MAUP

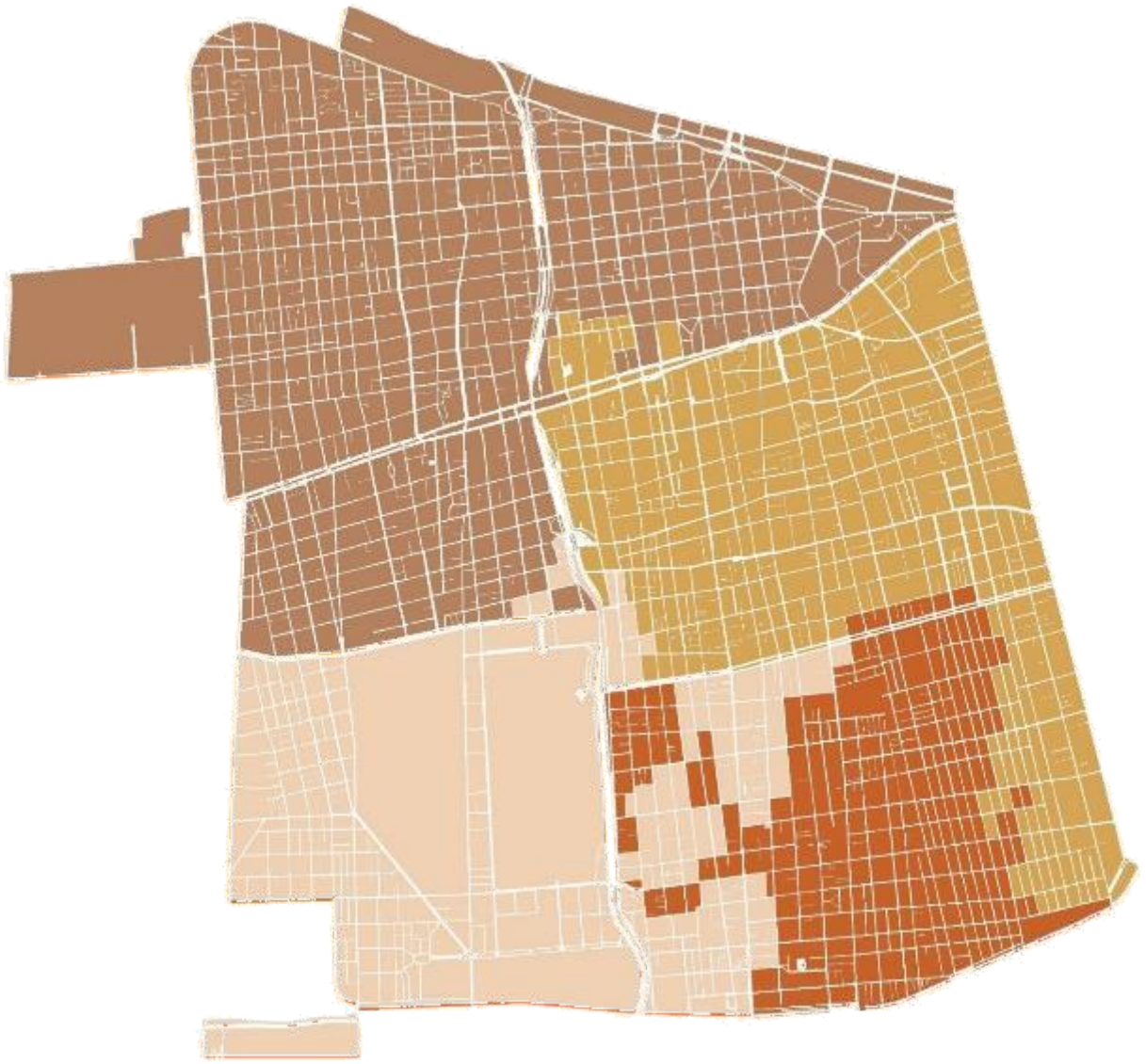
Algoritmo conceptual



Funcionamiento algoritmo

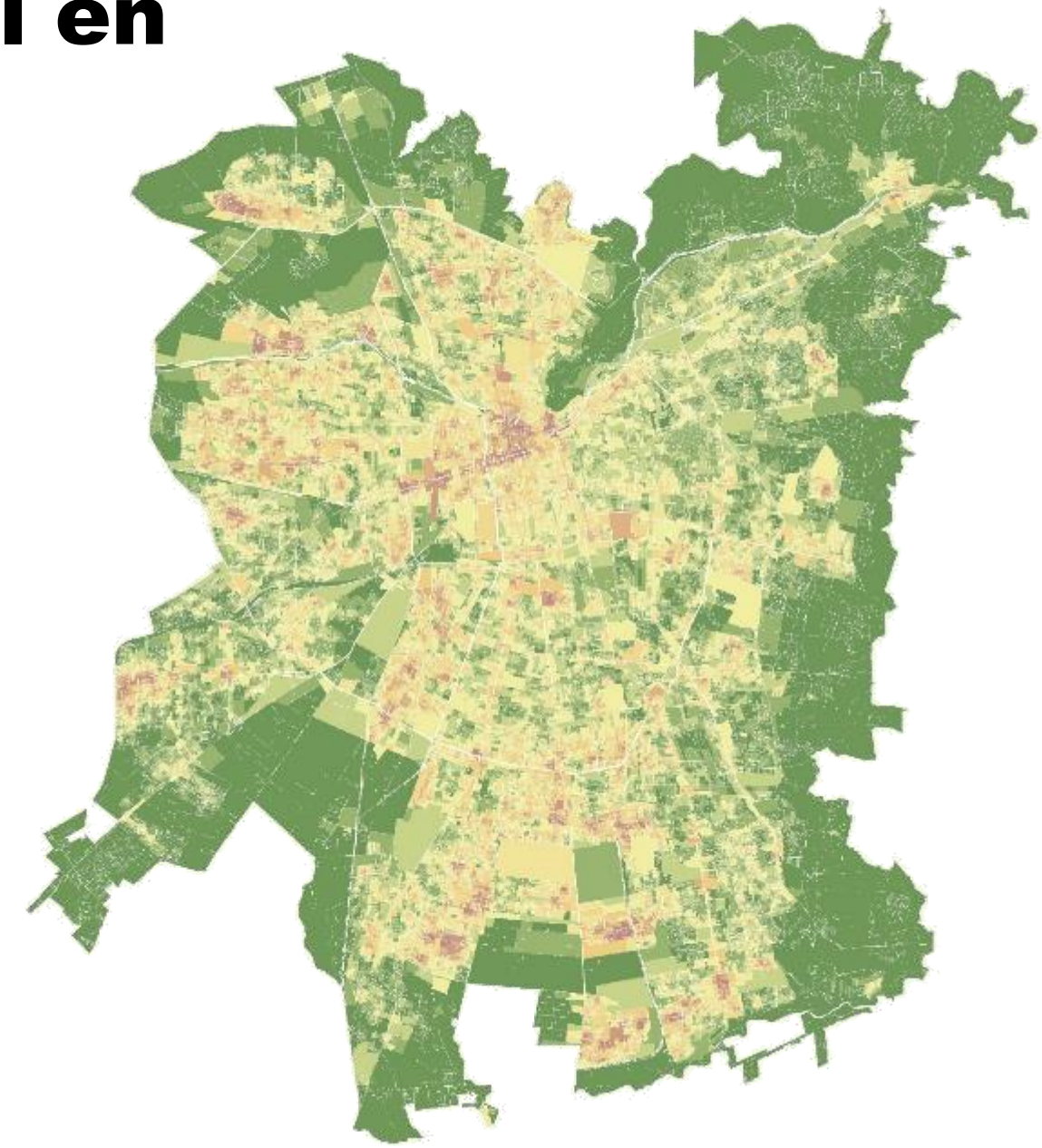
- En cada iteración, cada entidad se auto-organiza espacialmente según las reglas:
  - Identifica y evalúa vecinos.
  - Selecciona el vecino más parecido
  - Se fusiona

	AREA	POPULATION
Mean	4	1
Mín	1	1
Max	2	2
Var Coeff.	0,29	0,28



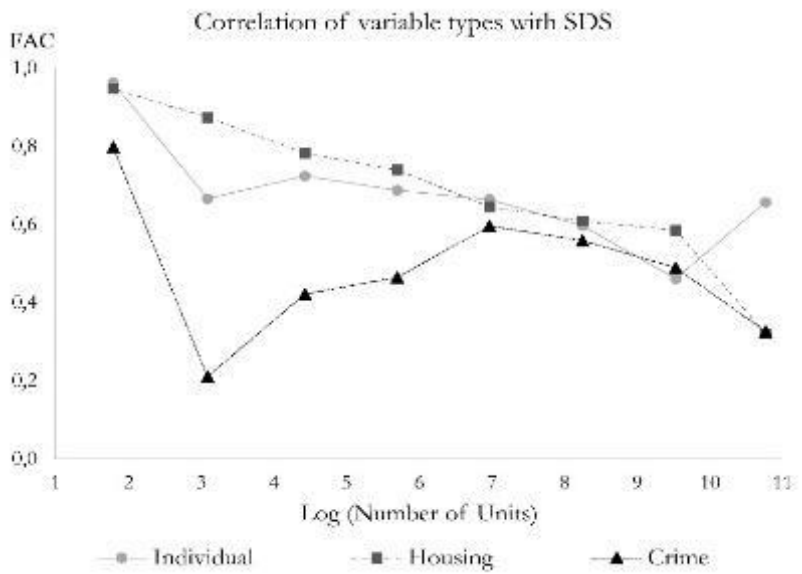
Caso de estudio: Vulnerabilidad social en Santiago

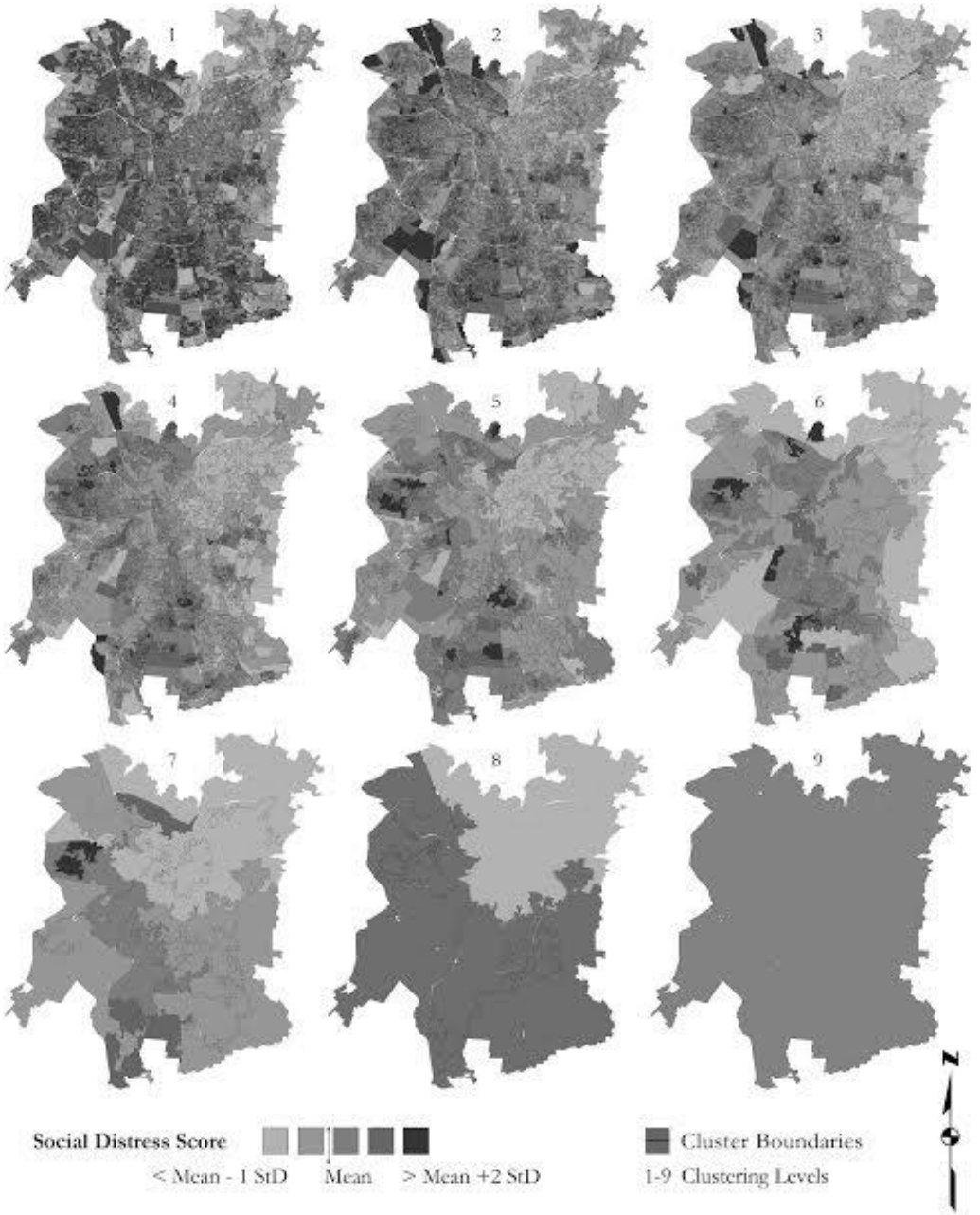
Variable	Description	Formula
Unemployment	Percentage of population willing to work but without employment	Unemployed/ Employed or willing to work
Dependence	Percentage of inactive or unemployed population	1- (Employed/Total population)
Uneducated	Inverse of education years for population older than 24 years	Population >24/ Sum of education years (>24)
Overcrowding	Average number of rooms for each inhabitant, calculated at household level	Mean (Rooms in residence/ Residents)
Precariousness	Percentage of shanty housing	Precarious accommodations/ Total accommodations
Insalubrity	Percentage of housing without formal sanitation systems	Insalubrious accommodations/ Total accommodations
High violence	Density of homicides, rapes and gravest injuries	High violence reports/ Area
Insurgence	Density of weapons-related offenses and aggressions to officers	Insurgence reports/ Area
Drugs	Density of drug-related crimes and offenses	Drug-related reports/ Area
Aggressions	Density of offenses against the person	Aggression reports/ Area



Zonas críticas en el Gran Santiago a múltiples escalas

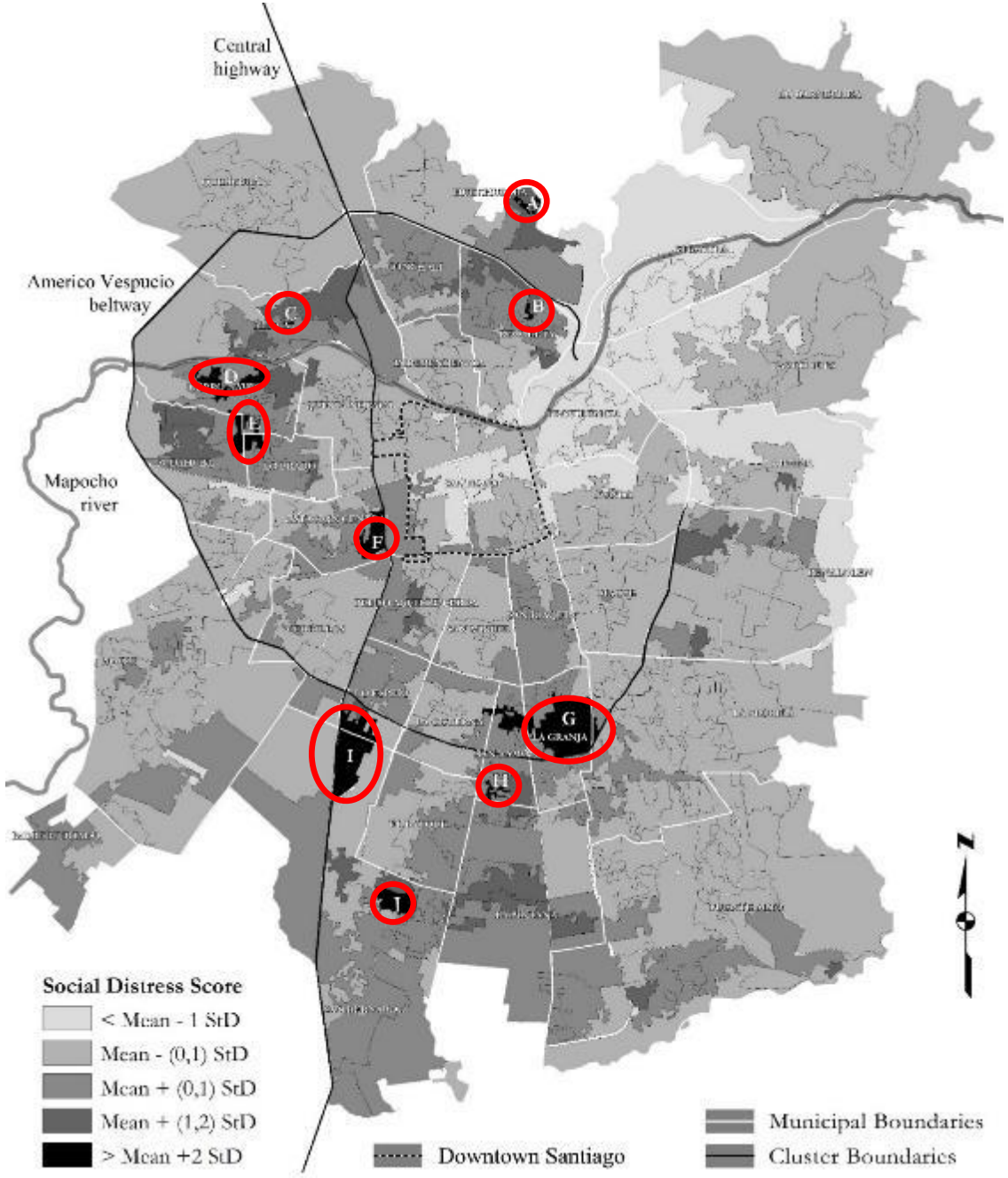
- Grupos altamente vulnerables en todas las escalas
- Las variables presentan correlaciones diferentes según la escala.





Análisis de vulnerabilidad

A	La Pincoya
B	Huamachuco
C	Resbalón
D	Barrancas
E	Araucanía y Nogales
F	Lo Espejo (Vespucio)
G	San Gregorio
H	La Pintana
I	Nueva Espejo
J	Portada



Clustering Espacial

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2