# Datos Atípicos

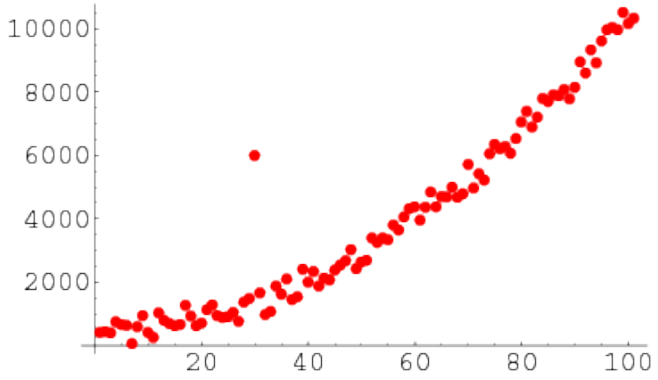
Dr. Raimundo Sánchez

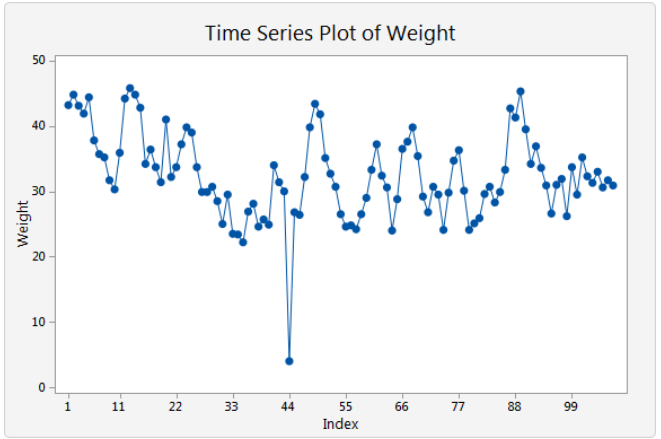
raimundo.sanchez@uai.cl

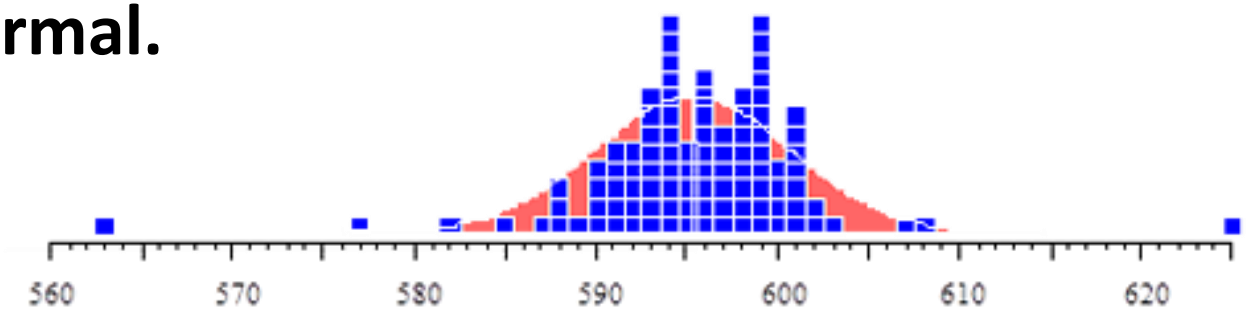
@raimun2

# Definición

- Valores atípicos (también llamados anomalías, o outliers) corresponden a datos que son considerablemente diferentes del resto de los datos.
- La noción de valor atípico es altamente subjetiva y depende del dominio.
- La mayoría de las definiciones se basan en la definición de una distribución normal.

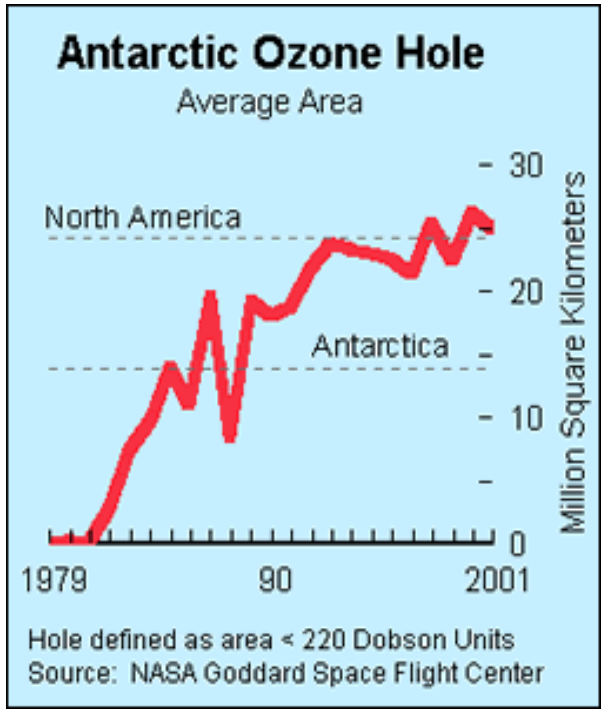


# Tipos de anomalías

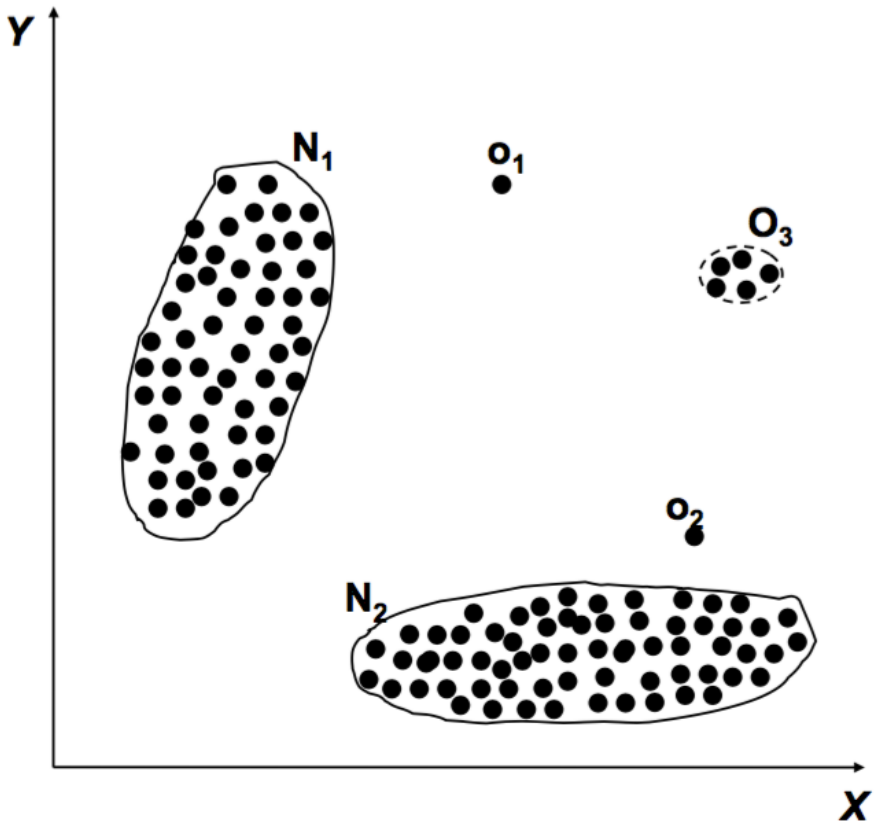
- Datos de diferentes clases
- Variación natural
- Errores de medición y recopilación de datos



# Anomalías puntuales

Una entidad de datos individual es anómala con respecto a los datos.

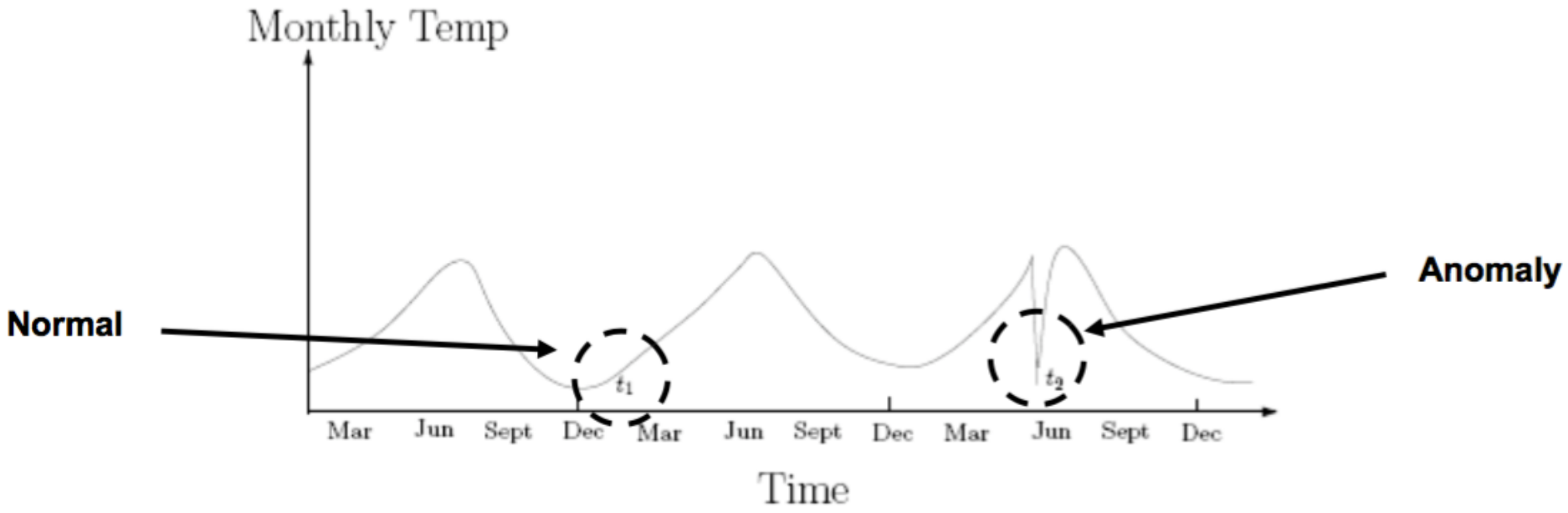
¿Cual es el mínimo numero de datos aislados que son considerados anomalías?



# Anomalías de contexto

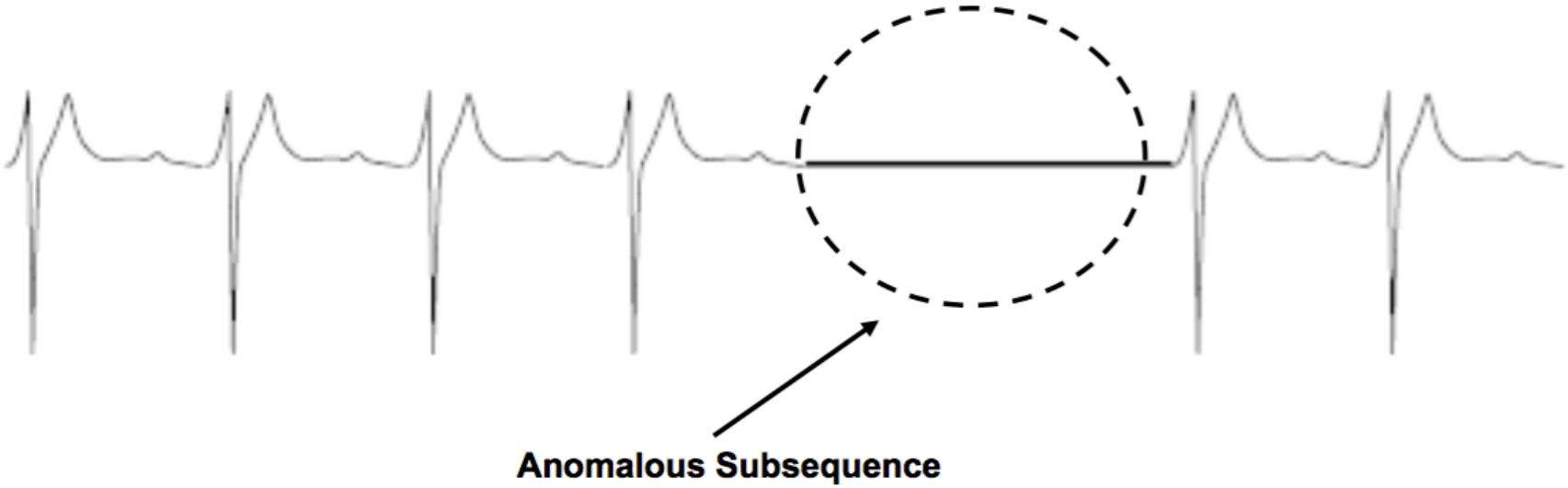
**Anomalías puntuales dentro de un contexto especifico.**

**Requiere una noción de contexto y también se conoce como anomalías condicionales.**



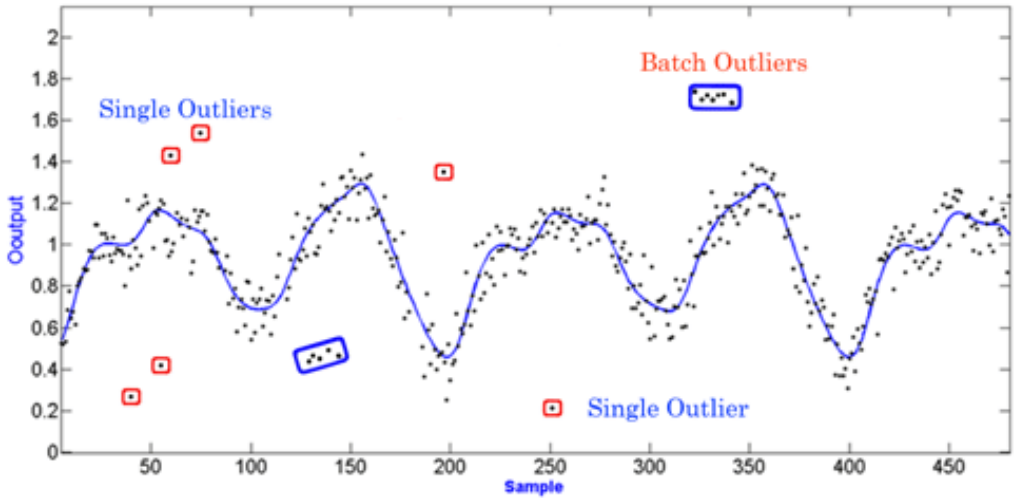
# Anomalías colectivas

- Es una colección de instancias de datos anómalas relacionadas.
- Requiere una relación entre entidades, como por ejemplo: datos secuenciales, espaciales, gráficos.
- Las instancias individuales dentro de una anomalía colectiva no son anómalas por sí solas.



# Desafíos

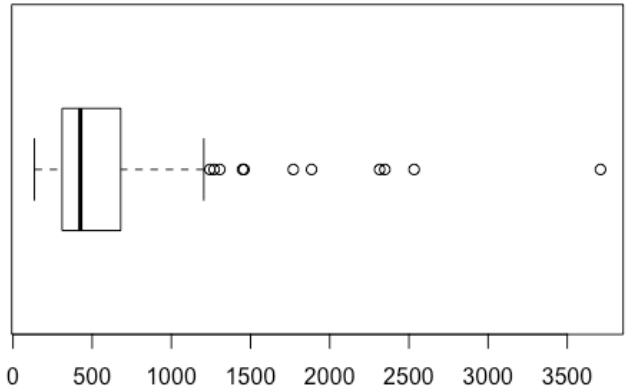
- ¿Cuántos atributos se utilizan para definir un valor atípico?
- ¿Cuántos valores atípicos hay en los datos?
- Las etiquetas de las clases son costosas
- Distribución de clases sesgada (encontrar agujas en un pajar)
- ¿cual es el conjunto mínimo de datos aislados considerados anómalos?



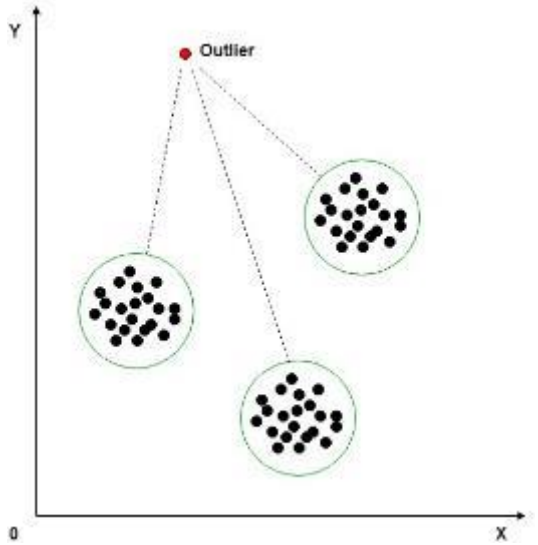
# Algunas aplicaciones

- Detección de fraude / intrusión
- Perturbaciones del ecosistema
- Monitoreo del sistema
- Bio vigilancia / salud pública
- Preprocesamiento de datos

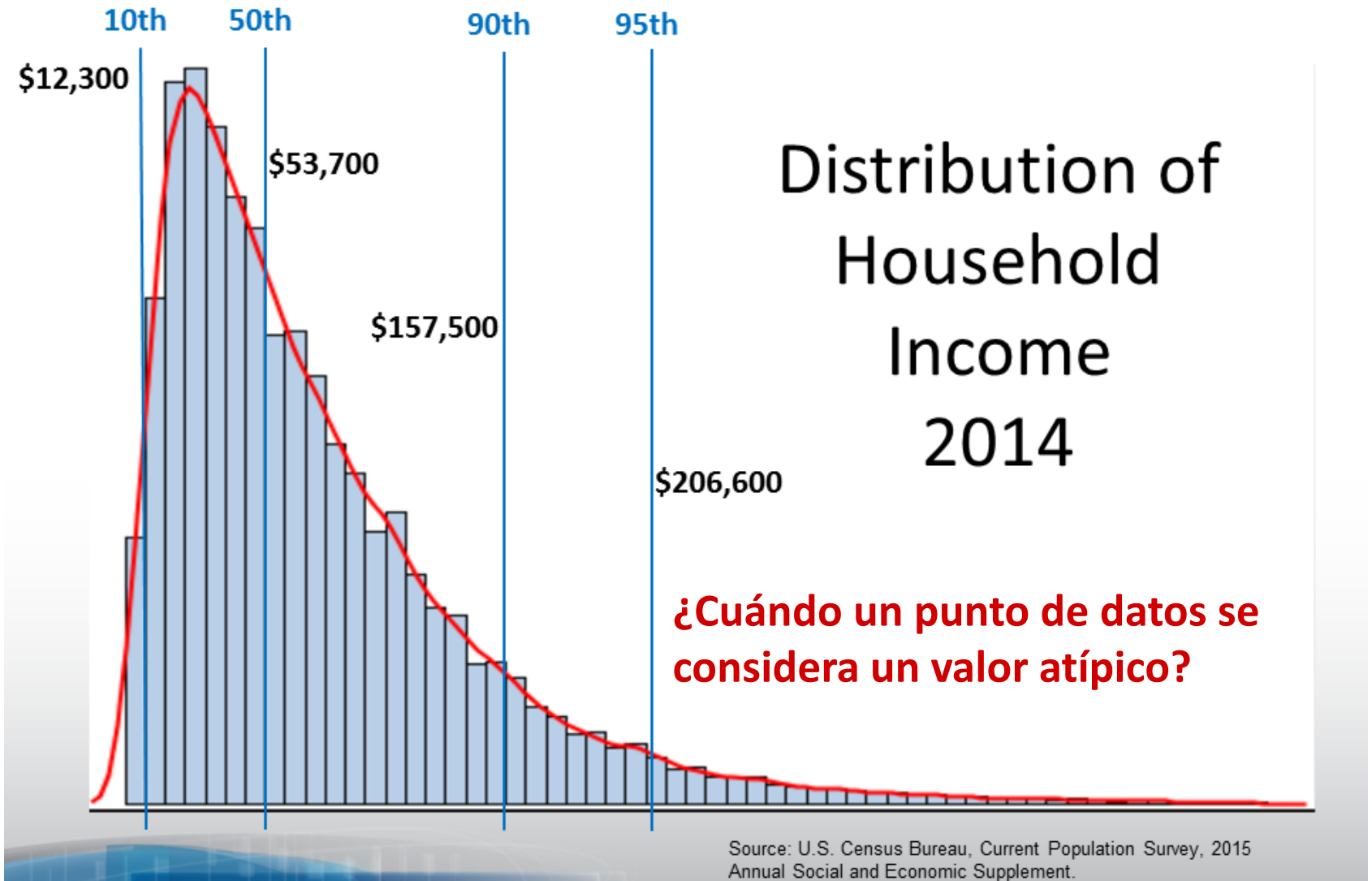
1D




2D



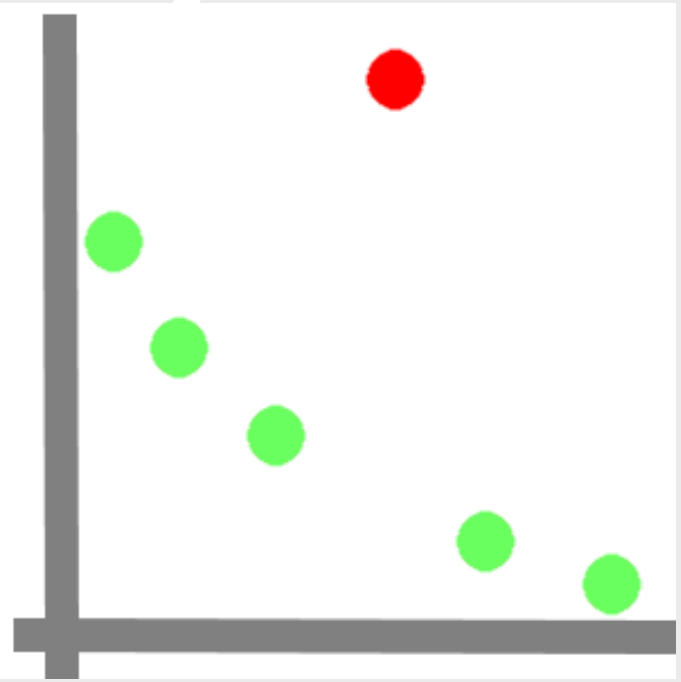
# Ejemplo: Distribución del ingreso



# Enfoques

Hay tres enfoques principales para la detección de datos atípicos

- **Sin supervisión:** No se asumen etiquetas. Se basa en la suposición de que las anomalías son muy raras en comparación con los datos normales.
- **Supervisado:** Etiquetas disponibles tanto para datos normales como para anomalías. Esto es similar a la clasificación con clases desequilibradas.
- **Semi supervisado:** Etiquetas disponibles solo para datos normales



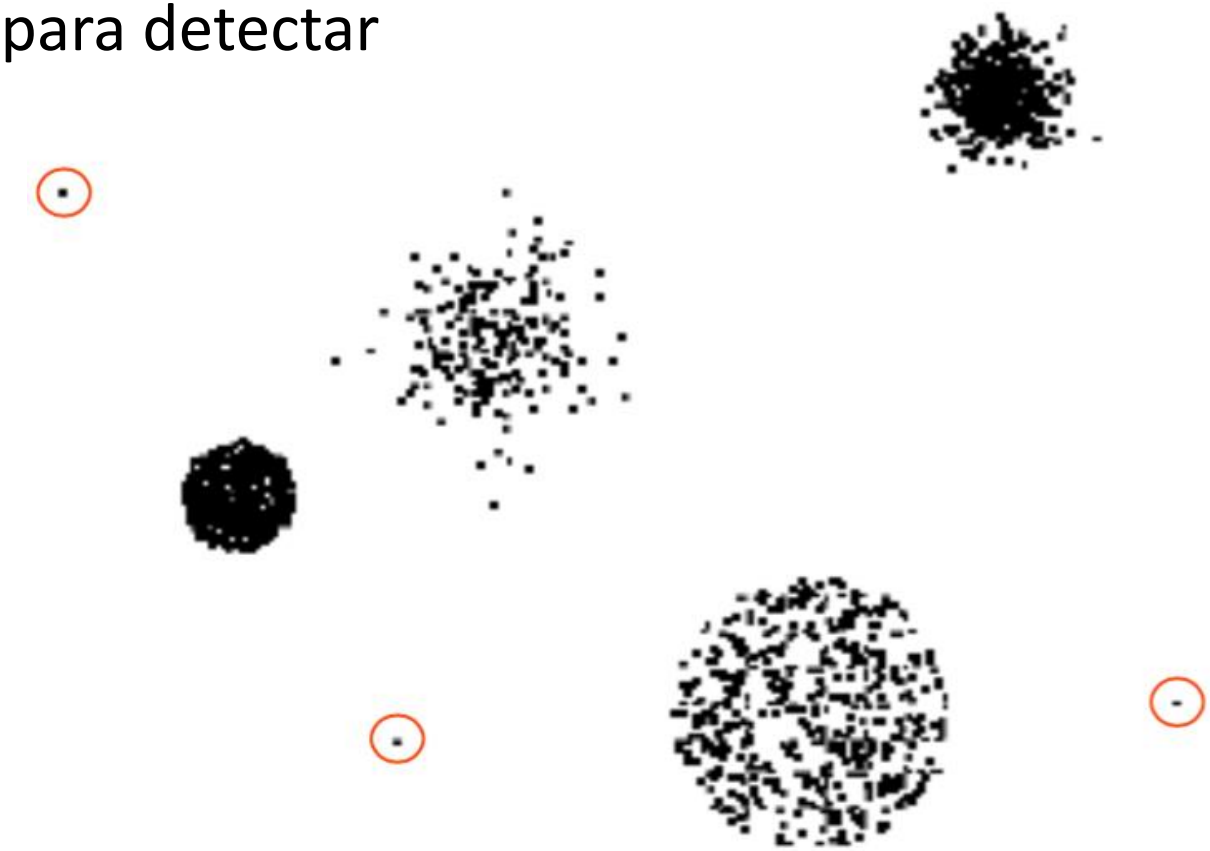
# Métodos sin supervisión

## Método general:

- Cree un perfil de comportamiento "normal" basado en estadísticas descriptivas de la población en general.
- Utilice desviaciones de lo "normal" para detectar anomalías.

## Tipos de métodos

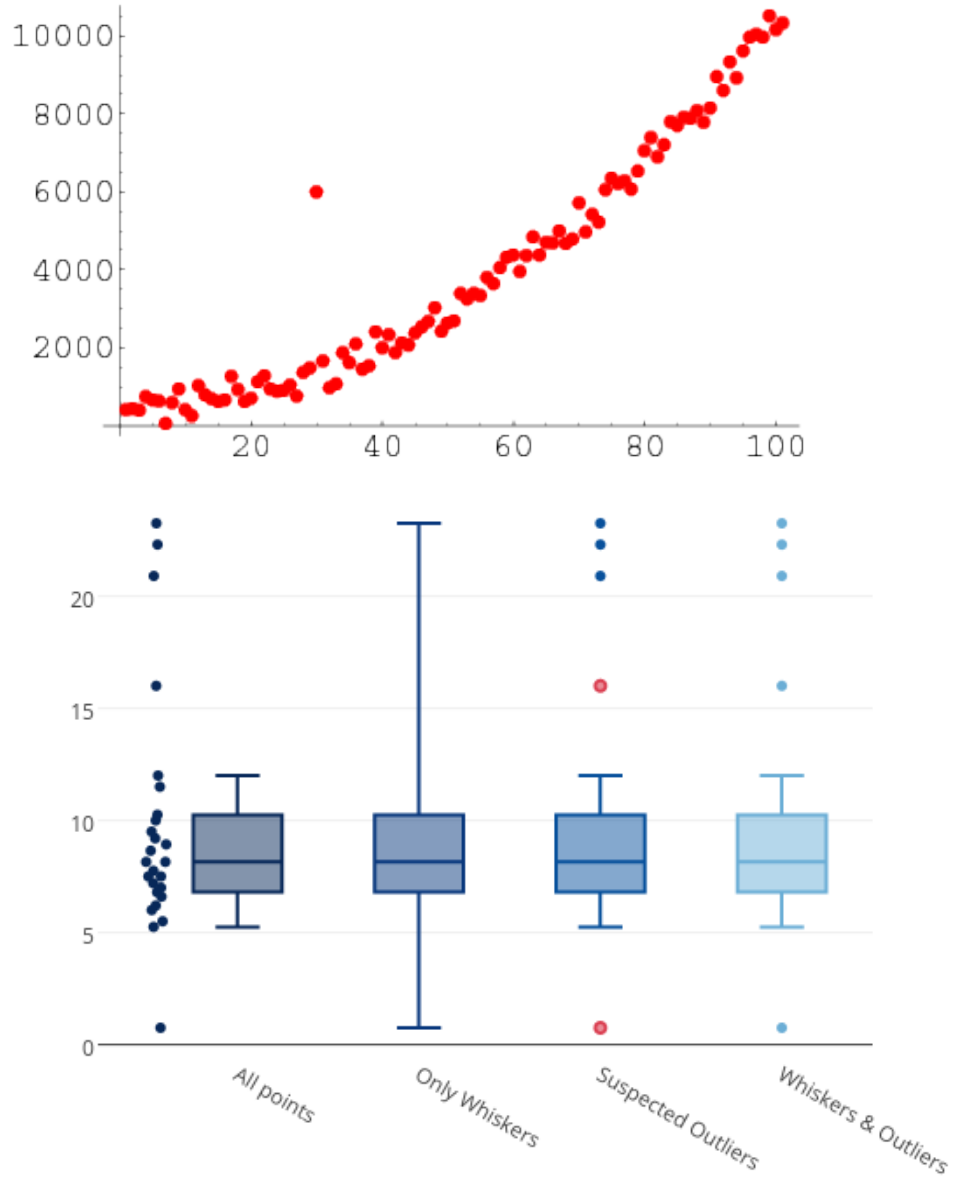
- Visual y estadístico
- Basado en distancia
- Basado en modelos



# Método visual

Según los gráficos de los datos se detectan los valores atípicos.

Estos métodos requieren mucho tiempo y son subjetivos para el usuario.



# Métodos estadísticos

- Se utiliza un modelo paramétrico que describa la distribución de los datos
- Se aplica una prueba estadística para determinar ubicación de los valores en la distribución, basándose en parámetros de distribución como la media y la varianza.
- La mayoría de las pruebas son para la distribución de un solo atributo y es posible que la distribución de datos no se conozca o sea difícil de estimar (especialmente en multivariantes).
- La Prueba de Grubb's es la mas utilizada para datos univariados

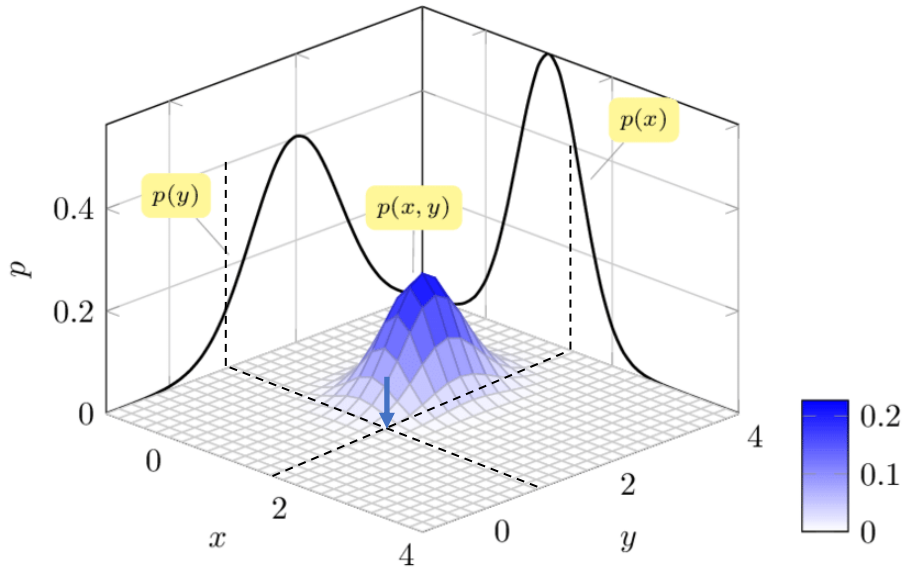
$$G = \frac{\max_{i=1,\ldots,n} |x_i - \overline{X}|}{S_X}$$

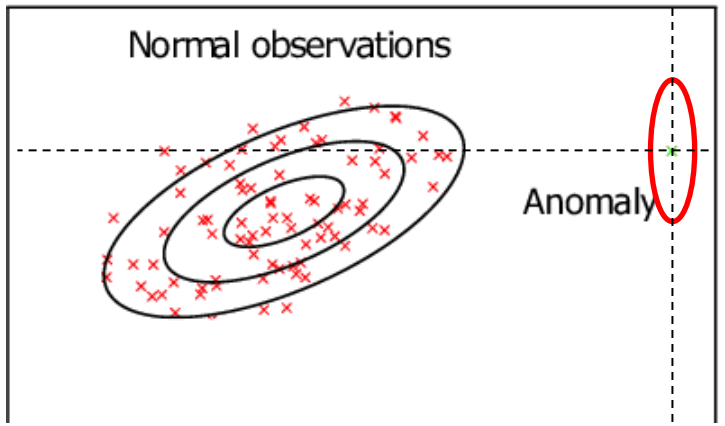
$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

Si se cumple, Xi se considera un valor atípico

# Métodos estadísticos: Múltiples atributos

Para las distribuciones gaussianas multivariadas se calcula la probabilidad de un punto con respecto a la distribución estimada, señalando los puntos con baja probabilidad como anómalos.



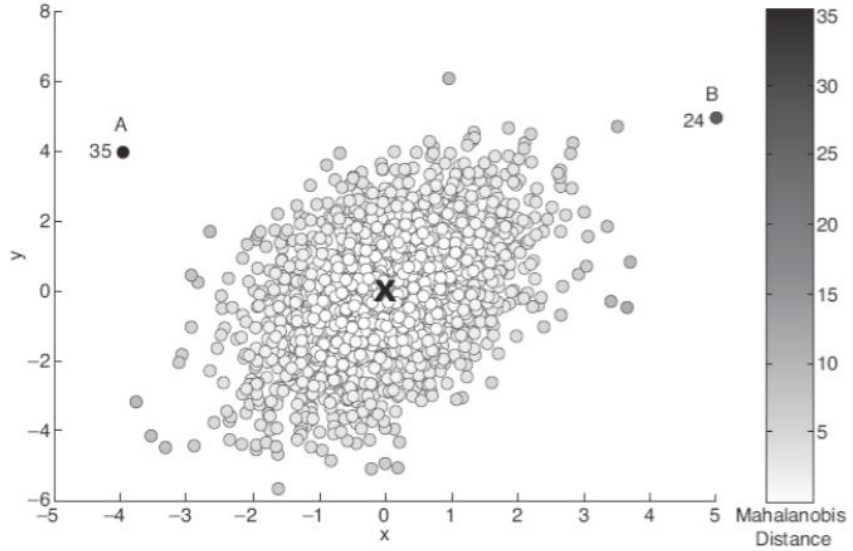



# Métodos de distancia

# Sin supervisión

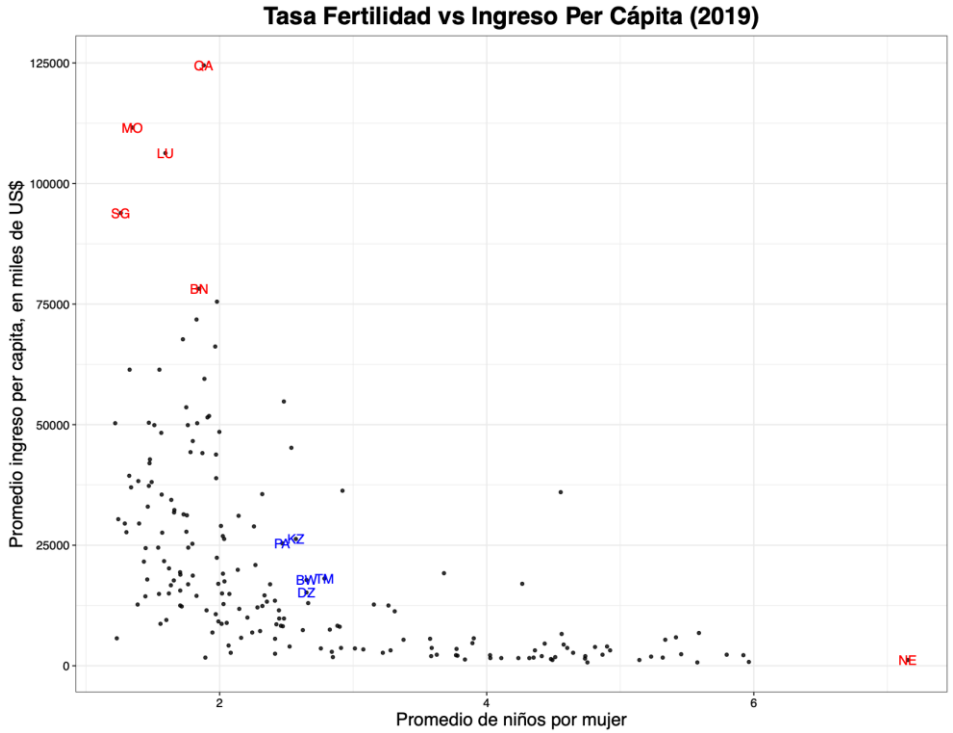
Utiliza una medida de distancia (comúnmente la distancia Mahalanobis) para considerar la covarianza entre los atributos.

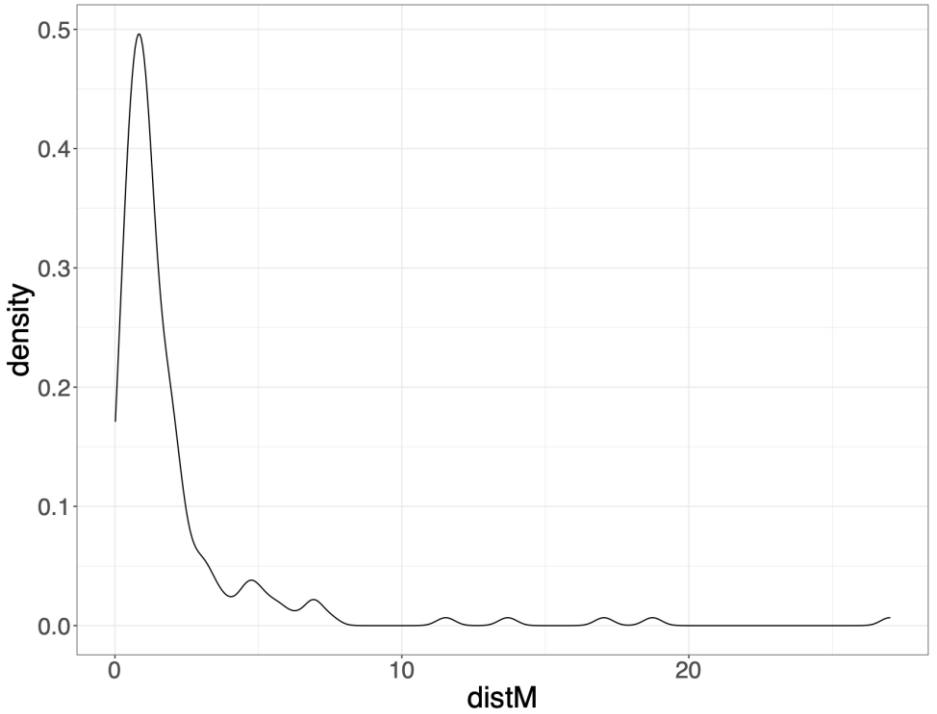
Calcula la distancia de cada punto al centroide, e identifica puntos con mayor distancia.



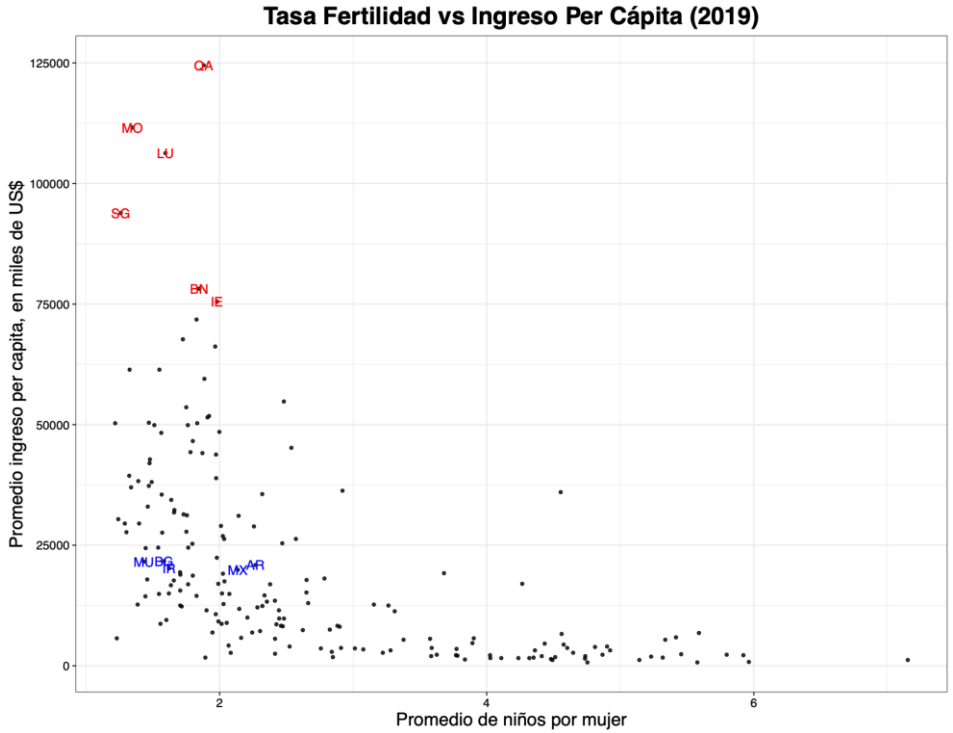

$$d_{MH}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})}$$

# Ejemplo distancia Mahalanobis



# Ejemplo distancia Euclidiana




Mismo proceso con distancia euclidiana

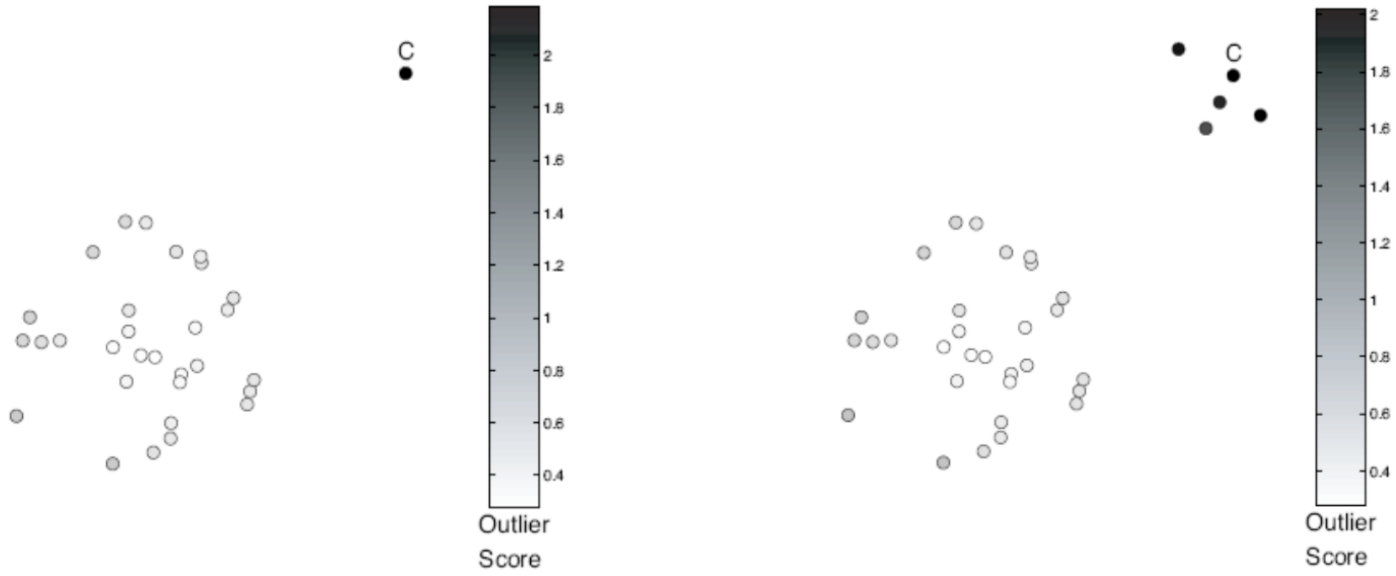
# Vecino más cercano

El método del vecino más cercano calcula la distancia entre cada par de puntos y permite definir valores atípicos basado en diferentes criterios:

- Los p puntos cuya distancia mínima a sus k vecinos más cercanos es mayor
- Los p puntos cuya distancia promedio a sus k vecinos más cercanos es mayor
- Puntos para los que hay menos de p puntos vecinos a una distancia D
- otros criterios ...

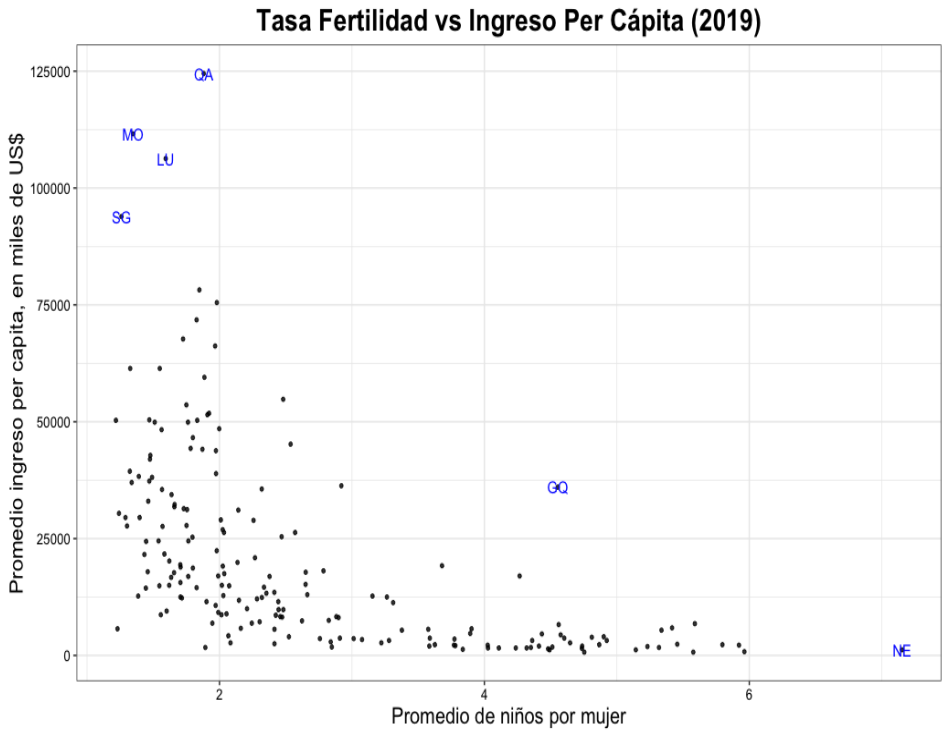
# Vecino más cercano

Los p puntos cuya distancia promedio a sus k vecinos más cercanos es mayor



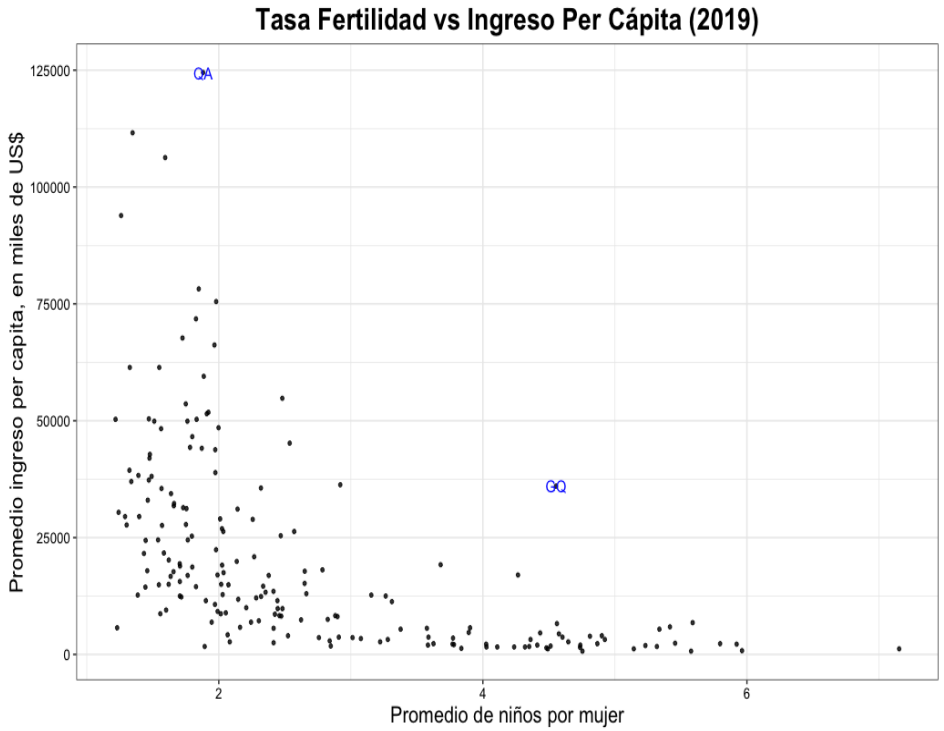

Puntajes atípicos basados en el 5$^{to}$ Vecino más cercano

# Ejemplo vecino más cercano



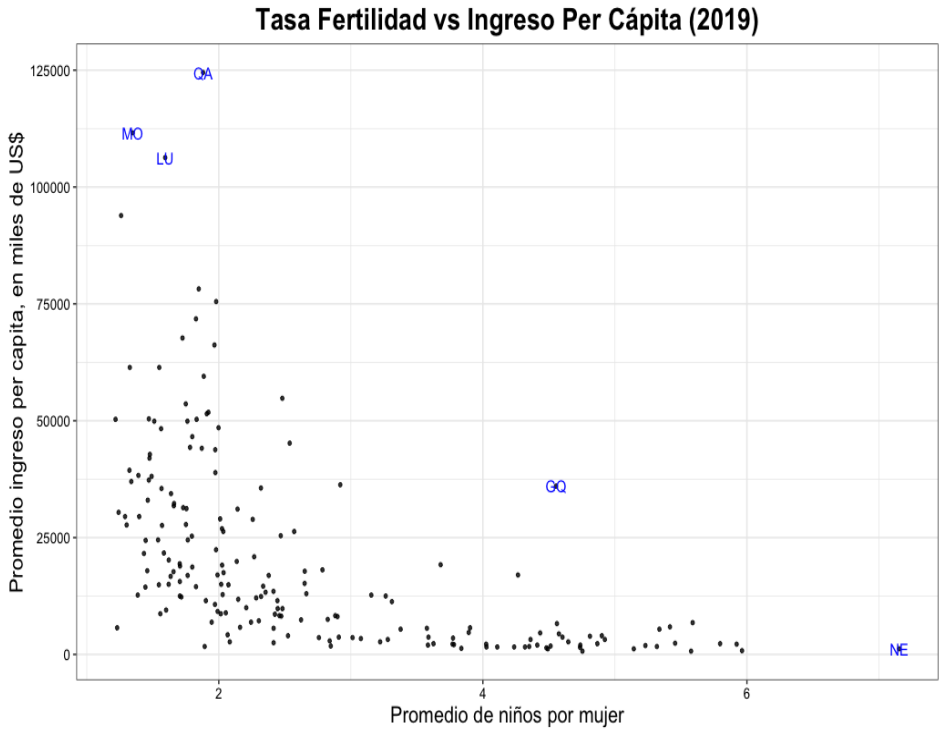

Aplicando el primer criterio con K = 4

# Ejemplo vecino más cercano




Aplicando el segundo criterio con K = 4

# Ejemplo vecino más cercano




Aplicando el tercer criterio con K = 4

# Agrupación

- Agrupa los datos en grupos de diferente densidad
- Elige puntos en un grupo pequeño como candidato
- Si un punto candidato está lejos (basado en la distancia) de todos los demás puntos no candidatos, es un valor atípico.

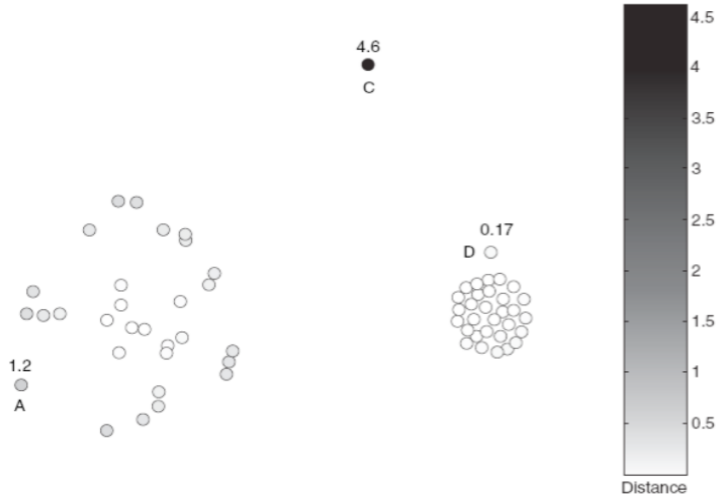



**Figure 10.9.** Distance of points from closest centroid.

# Método de densidad

- Encuentra los K vecinos más cercanos de cada punto
- Calcula la distancia máxima a los K más cercanos
- Calcula la distancia de alcance entre todos los puntos.

$$\text{reachability-distance}_k(A,B)=\max\{k\text{-distance}(B),\ d(A,B)\}$$

- Calcula la densidad de su vecindario local de k vecinos

$$\text{lrd}_k(A):=1/\left(\frac{\sum_{B\in N_k(A)}\text{reachability-distance}_k(A,B)}{|N_k(A)|}\right)$$

- Calcule el **factor de valor atípico local** como la relación entre la densidad de un punto y la densidad promedio de sus vecinos más cercanos

$$\text{LOF}_k(A):=\frac{\sum_{B\in N_k(A)}\frac{\text{lrd}_k(B)}{\text{lrd}_k(A)}}{|N_k(A)|}=\frac{\sum_{B\in N_k(A)}\text{lrd}_k(B)}{|N_k(A)|\cdot\text{lrd}_k(A)}$$

- LOF (x) >> 1 indica valores atípicos.

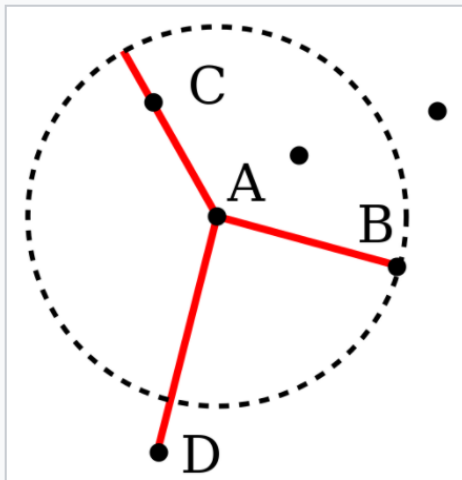
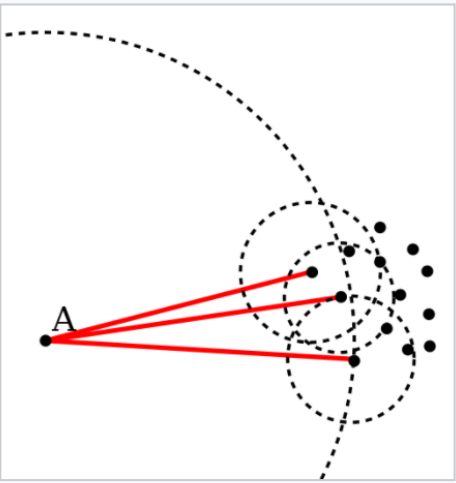
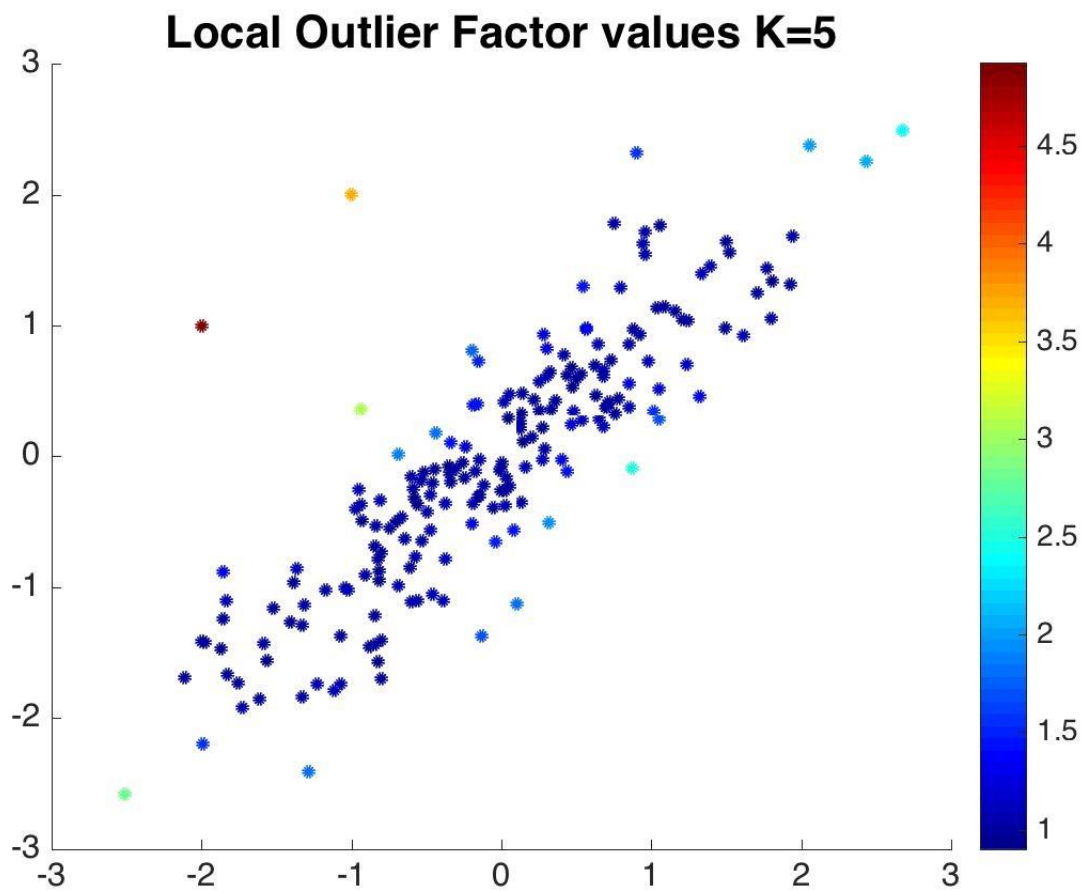



Illustration of the reachability distance. Objects *B* and *C* have the same reachability distance (k=3), while *D* is not a *k* nearest neighbor

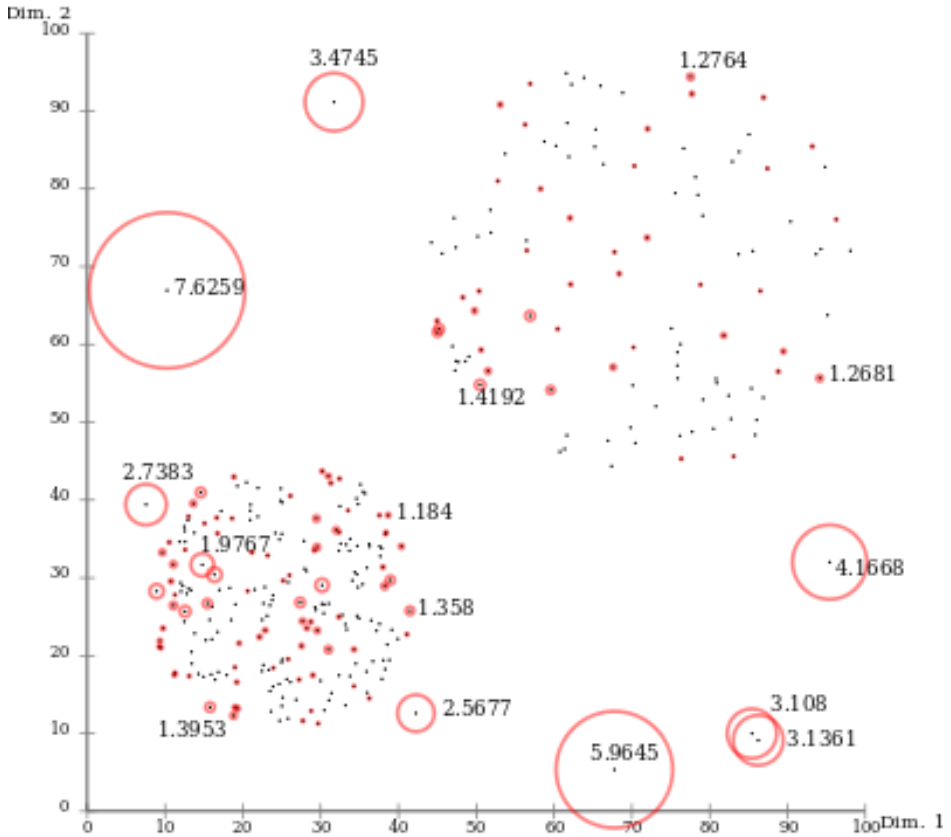



Basic idea of LOF: comparing the local density of a point with the densities of its neighbors. A has a much lower density than its neighbors.

# Método de densidad



# Método de densidad




Puntuaciones de LOF visualizadas por ELKI.

# Datos Atípicos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2