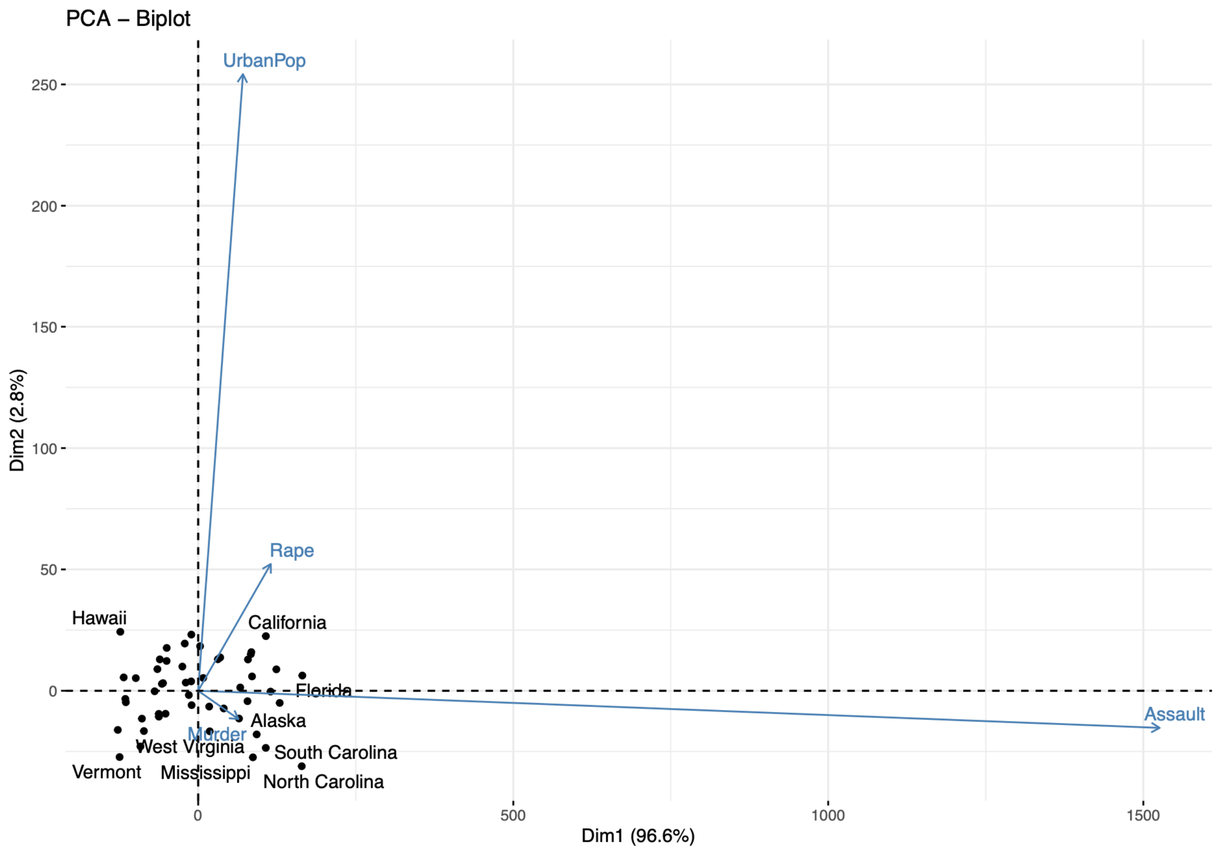
Más ejemplos de reducción de dimensionalidad

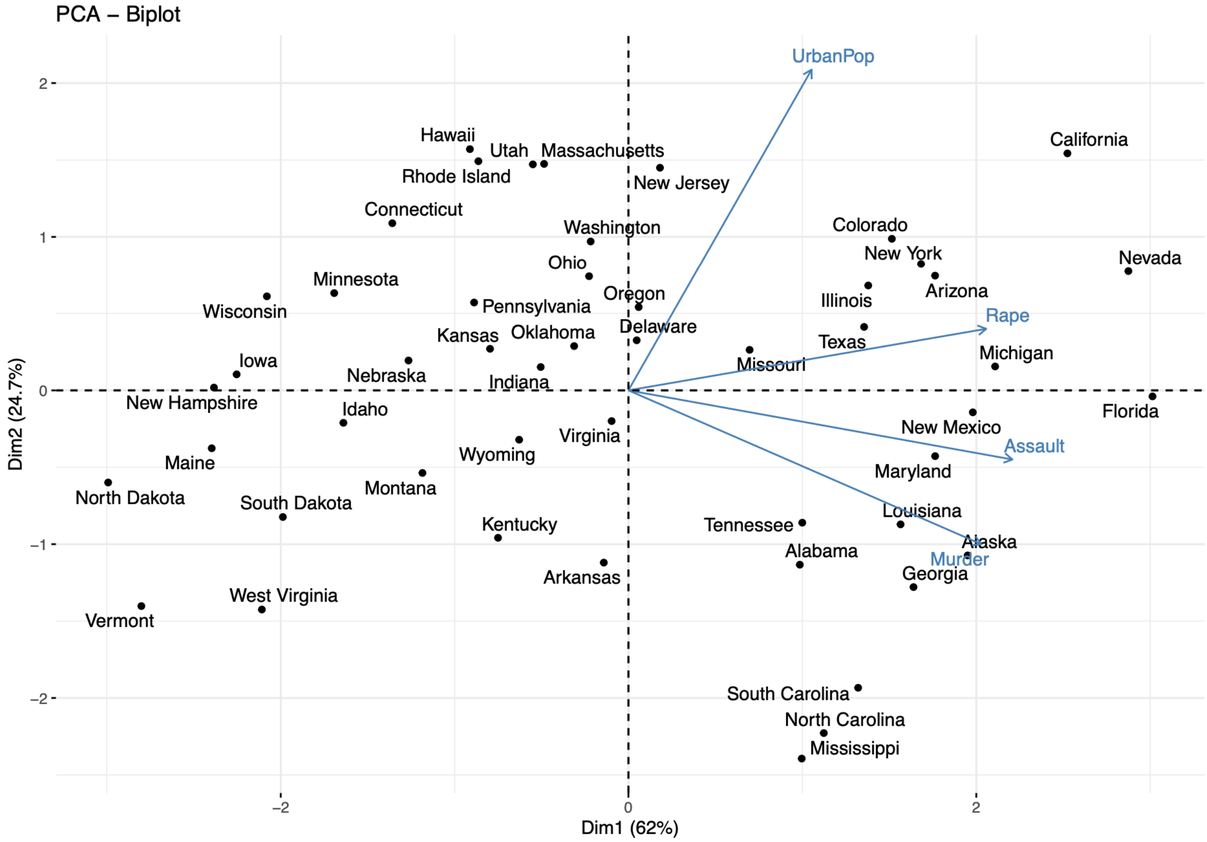
Maine, 6 de mayo de 2025

PCA: estandarizar vs no estandarizar

```
data("USArrests")
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```





Reducción de dimensionalidad no lineal

tSNE: t-distributed stochastic neighbor embedding (Hinton y van der Maaten)

Paso 1:

- construir una distribución de probabilidad sobre pares datos
- objetos más parecidos, tienen mayor probablidad

x_j

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Reducción de dimensionalidad no lineal

tSNE: t-distributed stochastic neighbor embedding (Hinton y van der Maaten)

Paso 1:

- construir una distribución de probabilidad sobre pares datos
- objetos más parecidos, tienen mayor probablidad

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Sean $\mathbf{x}_1, \ldots, \mathbf{x}_N$ datos

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

↓

Es la probabilidad que x_i elija a $x_j$omo su vecino.

Los vecinos se eligen en proporción a su densidad de probabilidad bajo una gaussiana centrada en x_i .

Reducción de dimensionalidad no lineal

tSNE: t-distributed stochastic neighbor embedding (Hinton y van der Maaten)

Paso 1:

- construir una distribución de probabilidad sobre pares datos
- objetos más parecidos, tienen mayor probablidad

Sean $\mathbf{x}_1, \ldots, \mathbf{x}_N$ datos

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

↓

Se adapta a la densidad de los datos

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

tSNE: t-distributed stochastic neighbor embedding (Hinton y van der Maaten)

Paso 2:

- definir una distribución de probabilidad similar sobre los puntos del mapa de baja dimensión y minimizar la divergencia de Kullback-Leibler

$$\mathbf{y}_1, \cdots, \mathbf{y}_N \in \mathbb{R}^d$$

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

$$\mathrm{KL}\left(P \parallel Q\right) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

tSNE: t-distributed stochastic neighbor embedding (Hinton y van der Maaten)

Paso 2:

- definir una distribución de probabilidad similar sobre los puntos del mapa de baja dimensión y minimizar la divergencia de Kullback-Leibler

$$\mathbf{y}_1, \cdots, \mathbf{y}_N \in \mathbb{R}^d$$

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

$$\mathrm{KL}\left(P \parallel Q\right) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

UMAP: versión no euclideana de tSNE (aproxima una variedad)

Un mal y conocido ejemplo UMAP vs PCA

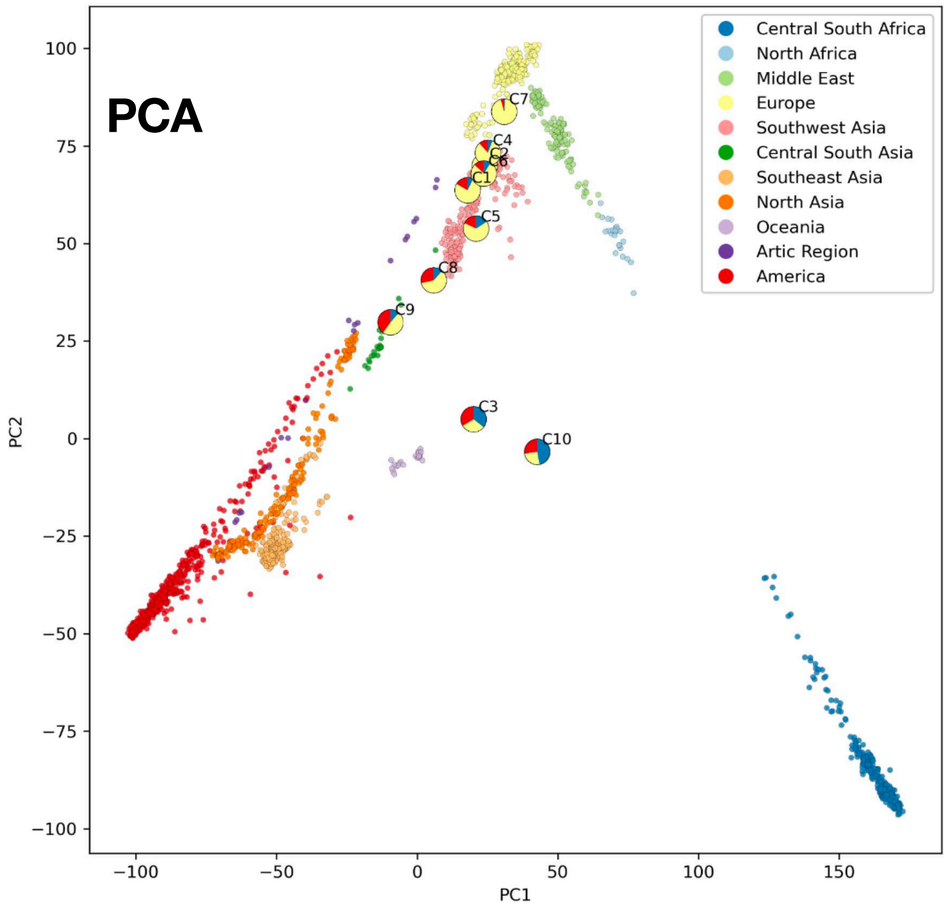


Figure 1

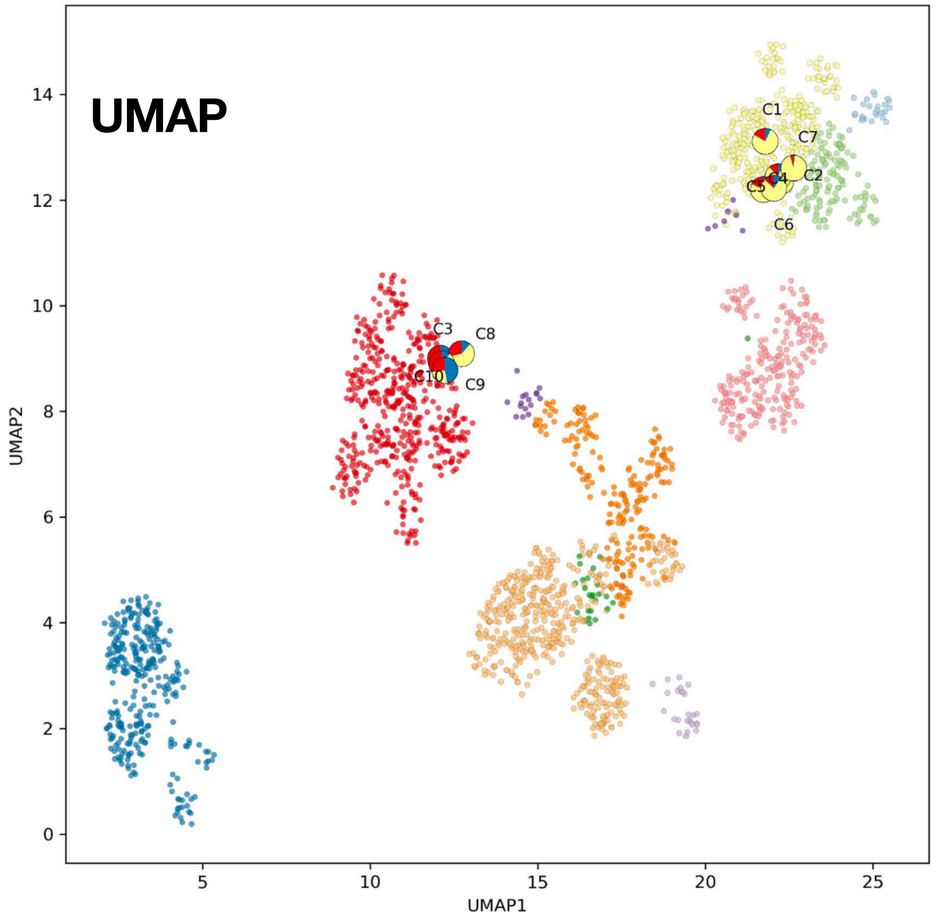
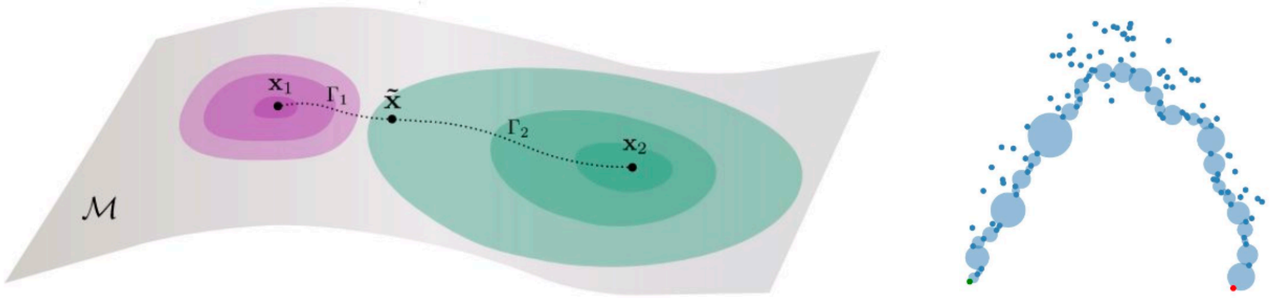


Figure 2

UMAP prioriza las distancias locales, perdiendo "noción" de las distancias globales.

Útil cuando los datos están organizados en clusters y quieron encontrarlos.



Fermat distance representation. Images taken from https://www.aristas.com.ar/fermat/index

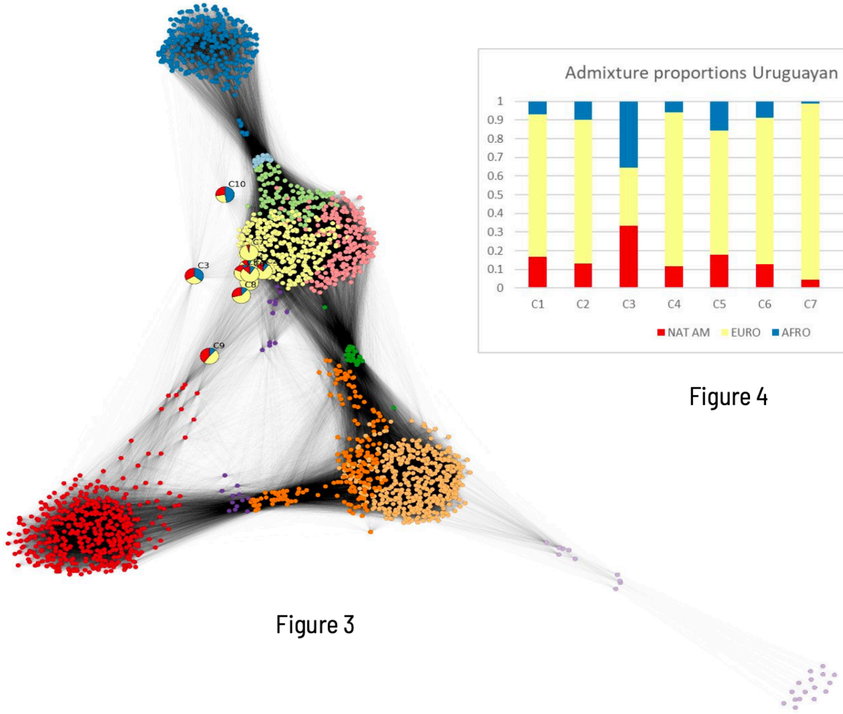


Figure 3

Figure 4

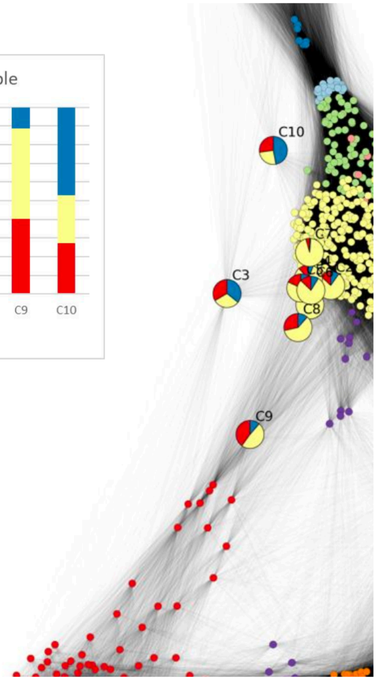


Figure 5

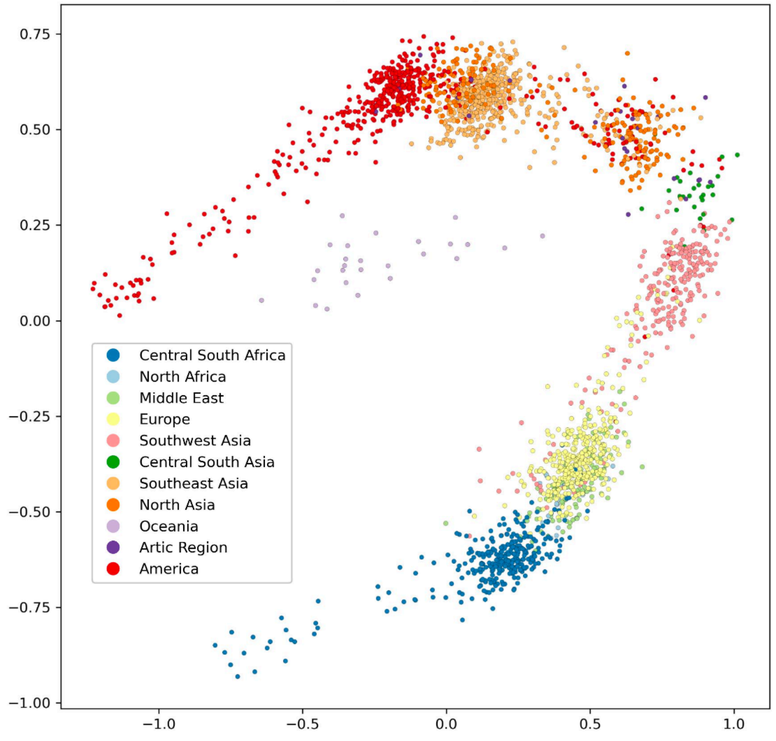


Figure 6: Node2vec embedding of Fermat distance graph.