Unsupervised Learning

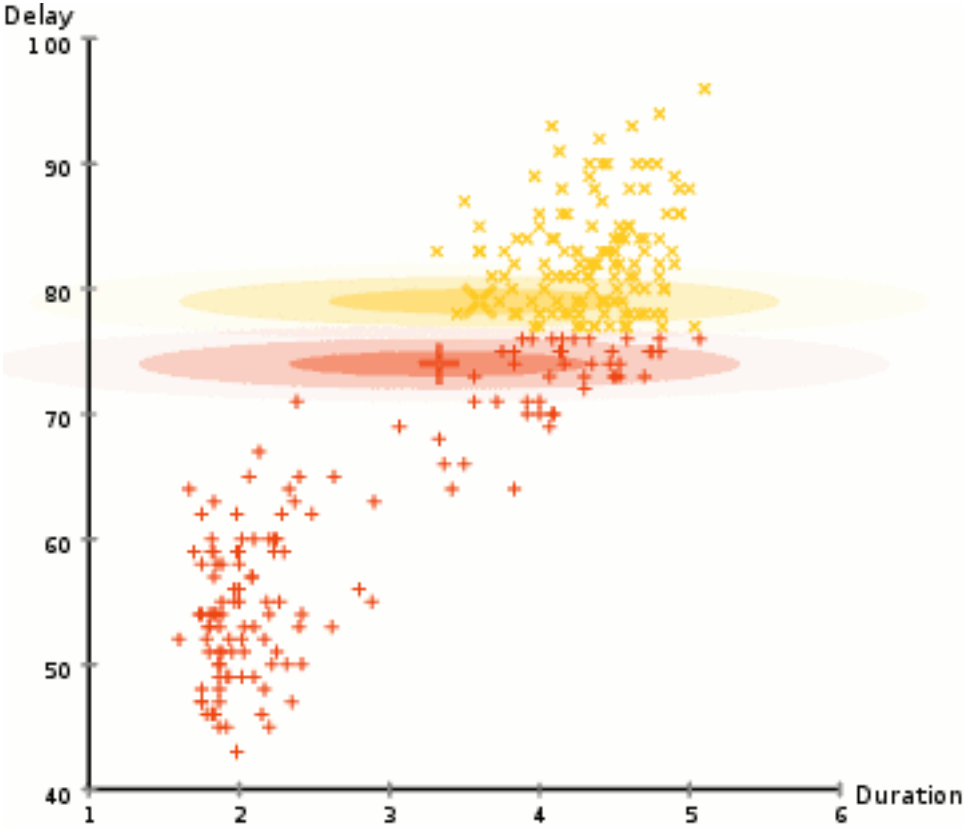
Germán Chaparro Molano

UNIVERSIDAD ECCI

Clasificación no supervisada

- Nadie ha visto antes los datos
- No es trivial hacer una clasificación
  - Complejidad
  - Volumen de datos
- ¿Qué podemos hacer?
- Buscamos estructura intrínseca en los datos
- Problema: No hay con qué comparar, *accuracy* = ????
- Evaluación relativa/cualitativa
- Ventaja: podemos evaluar clasificación supervisada (y viceversa)
- **Aprendizaje de máquina + aprendizaje humano**

Modelos de Mixtura Gaussiana



IAML: Mixture models and EM

Victor Lavrenko and Charles Sutton

Mixture Models

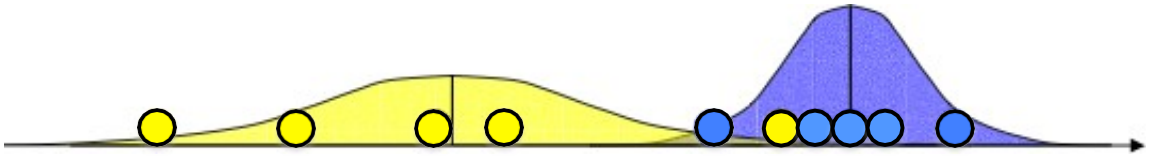
- Classification methods
  - hard classification: classes do not overlap
    - element either belongs to class or it does not
  - soft classification: classes may overlap
    - strength of association between classes and instances
- Mixture models
  - probabilistically---grounded way of doing soft classification
  - each source: a generative model (Gaussian or multinomial)
  - parameters (e.g. mean/covariance are unknown)
- Expectation Maximization (EM) algorithm
  - automatically discover all parameters for the K "sources"

Mixture Models in 1D

- Observations $x_1 \dots x_n$
  - K=2 Gaussians with unknown μ, σ^2
  - Estimation trivial if we know the source of each observation

$$\mu_b = \frac{x_1 + x_2 + \dots + x_{n_b}}{n_b}$$

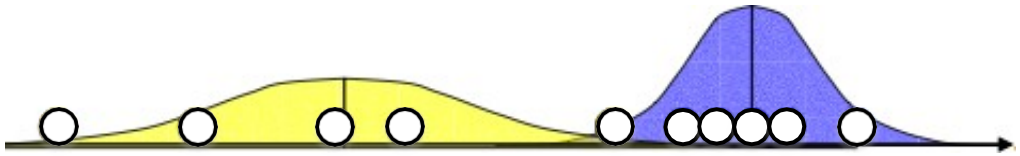
$$\sigma_b^2 = \frac{(x_1 - \mu_1)^2 + \dots + (x_n - \mu_n)^2}{n_b}$$



- What if we don't know the source?
- If we knew parameters of the Gaussians (μ, σ^2)
  - can guess whether point is more likely to be a or b

$$P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

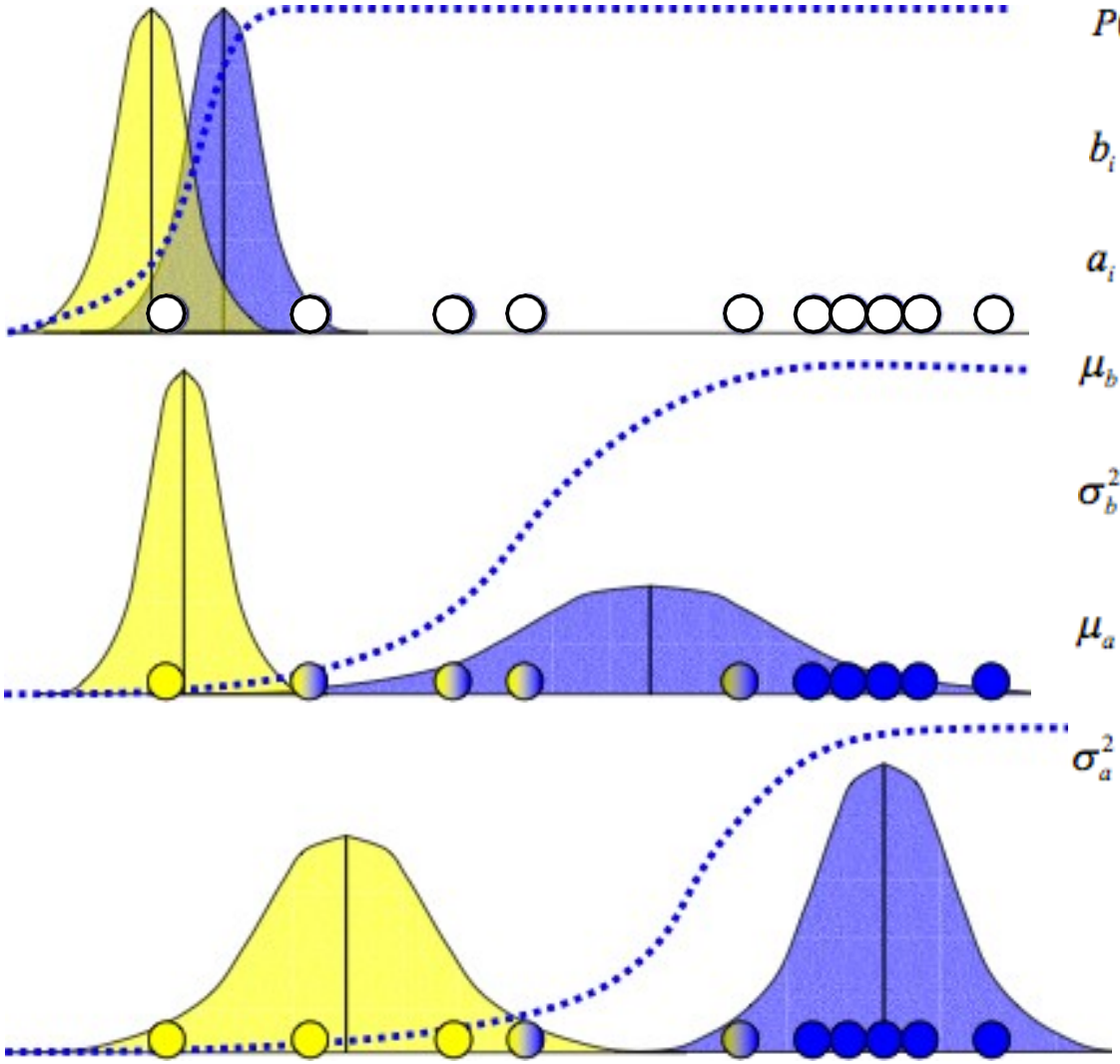
$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$



Expectation Maximization (EM)

- Chicken and egg problem
  - need (μ_a, σ_a^2) and (μ_b, σ_b^2) to guess source of points
  - need to know source to estimate (μ_a, σ_a^2) and (μ_b, σ_b^2)
- EM algorithm
  - start with two randomly placed Gaussians (μ_a, σ_a^2), (μ_b, σ_b^2)
  - E-step: for each point: $P(b|x_i)$ = does it look like it came from b?
  - M-step: adjust (μ_a, σ_a^2) and (μ_b, σ_b^2) to fit points assigned to them
  - iterate until convergence

EM:1D example



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \ldots + b_n x_{n_b}}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1 (x_1 - \mu_1)^2 + \ldots + b_n (x_n - \mu_n)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_{n_b}}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1 (x_1 - \mu_1)^2 + \ldots + a_n (x_n - \mu_n)^2}{a_1 + a_2 + \ldots + a_n}$$

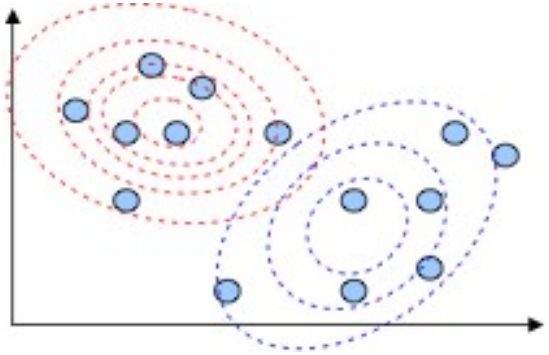
could also estimate priors:

$P(b) = (b_1 + b_2 + \ldots b_n) / n$

$P(a) = 1 - P(b)$

Gaussian mixture models: d>1

- Data with d attributes, from k sources
- Each source c is a Gaussian
- Iteratively estimate parameters:
  - prior: what % of instances came from source c?
  - mean: expected value of attribute j from source c
  - covariance: how correlated are atiributes j and k in source c?
  - based on: our guess of the source for each instance



Gaussian mixture models: d>1

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \doteq \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$p(\mathbf{x}) = \sum_{m=1}^{M} \frac{c_m}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_m|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right]$$

$$= \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m), \qquad (c_m > 0).$$

$$\boldsymbol{\Theta} = \{c_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\} \qquad \mathcal{L} = \prod_i p_i(\mathbf{x};\boldsymbol{\Theta})$$

How to pick K?

- Probabilistic model $L = \log P(x_1 \ldots x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i \mid k) P(k)$
  - tries to "fit" the data (maximize likelihood)
- Pick K that makes *L* as large as possible?
  - *K* = *n*: each data point has its own "source"
  - may not work well for new data points
- Split data into training (T) and validation (V) sets
  - for each *K*: fit parameters of T, measure likelihood of V
  - sometimes still best when *K* = *n*
- Occam's razor: pick "simplest" of all models that fit
  - Bayes Inf. Criterion (BIC): $\min_p(\ln(n)k - 2\ln(\hat{L}))$
  - Akaike Inf. Criterion (AIC): $\min_p \{ 2p - L \}$

L … **likelihood**, how well our model fits the data

p … **number of parameters** how "simple" is the model

Estudio de Sitio

- Estudios de viabilidad para observaciones en ondas milimétricas para Colombia - arXiv:1705.06121

Publications of the Astronomical Society of the Pacific, 129:105002 (20pp), 2017 October https://doi.org/10.1088/1538-3873/aa83fe

CrossMark

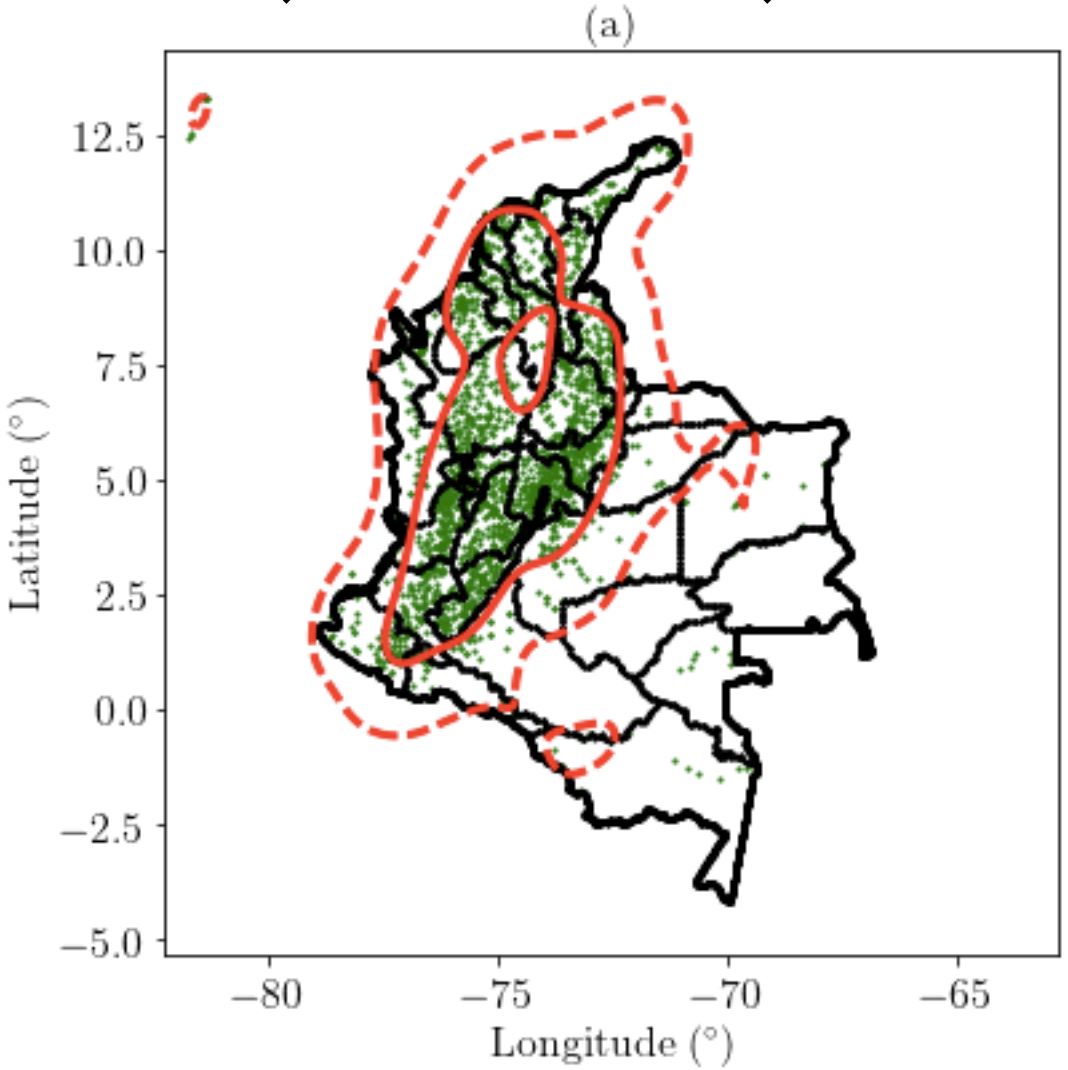
Low Dimensional Embedding of Climate Data for Radio Astronomical Site Testing in the Colombian Andes

Germán Chaparro Molano[1], Oscar Leonardo Ramírez Suárez[1], Oscar Alberto Restrepo Gaitán[1,2], and Alexander Marcial Martínez Mercado[1,3,4,5]

[1] Grupo de Simulación, Análisis y Modelado, Vicerrectoría de Investigación, Universidad ECCI, Bogotá, Colombia
gchaparrom@ecci.edu.co, oramirezs@ecci.edu.co
[2] Radio Astronomy Instrumentation Group, Universidad de Chile, Santiago de Chile, Chile; orestrepog@ecci.edu.co
[3] Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá, Colombia
[4] Grupo de Simulación del Sistema Climático Terrestre, Universidad Nacional de Colombia, Bogotá, Colombia
[5] Departamento de Ciencias Básicas, Universidad ECCI, Bogotá, Colombia
Received 2017 May 25; accepted 2017 July 31; published 2017 September 1

Datos del IDEAM (1980-2010)

- 2046 estaciones meteorológicas en todo el país
- Precipitación, Días con Lluvia, Humedad Relativa, Brillo Solar
- Promedios mensuales multianuales (Enero, Febrero,…)
- Datos 12-dimensionales
- Criterio de selección – Machine Learning



Climate data for National Meteorological Observatory, Bogotá (1971–2000)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °C (°F)	26.4 (79.5)	25.2 (77.4)	26.6 (79.9)	24.4 (75.9)	25.0 (77)	28.6 (83.5)	25.0 (77)	23.3 (73.9)	26.0 (78.8)	25.1 (77.2)	25.6 (78.1)	24.4 (75.9)	28.6 (83.5)
Average high °C (°F)	20.2 (68.4)	20.3 (68.5)	19.4 (66.9)	20.1 (68.2)	19.0 (66.2)	19.2 (66.6)	18.6 (65.5)	18.8 (65.8)	19.2 (66.6)	19.5 (67.1)	19.6 (67.3)	19.9 (67.8)	19.6 (67.3)
Daily mean °C (°F)	14.3 (57.7)	14.5 (58.1)	14.9 (58.8)	14.9 (58.8)	15.0 (59)	14.5 (58.1)	14.6 (58.3)	14.1 (57.4)	14.3 (57.7)	14.3 (57.7)	14.4 (57.9)	14.6 (58.3)	14.4 (57.9)
Average low °C (°F)	7.6 (45.7)	8.4 (47.1)	9.5 (49.1)	9.7 (49.5)	9.7 (49.5)	9.5 (49.1)	9.2 (48.6)	8.9 (48)	8.7 (47.7)	9.0 (48.2)	9.2 (48.6)	8.0 (46.4)	9.0 (48.2)
Record low °C (°F)	−1.5 (29.3)	−5.2 (22.6)	−0.4 (31.3)	0.2 (32.4)	0.2 (32.4)	1.1 (34)	0.4 (32.7)	0.4 (32.7)	0.3 (32.5)	1.8 (35.2)	0.5 (32.9)	−1.1 (30)	−5.2 (22.6)
Average precipitation mm (inches)	50 (1.97)	68 (2.68)	91 (3.58)	135 (5.31)	120 (4.72)	54 (2.13)	35 (1.38)	45 (1.77)	70 (2.76)	137 (5.39)	127 (5)	81 (3.19)	1,012 (39.84)
Average rainy days (≥ 1 mm)	9	12	14	18	19	17	15	14	16	21	16	11	181
Average relative humidity (%)	75	76	75	77	77	75	74	74	75	76	77	76	76
Mean monthly sunshine hours	156	128	107	88	83	94	114	117	109	96	103	138	1,328
Source: Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM)[33]													

Reducción de Dimensionalidad

- Análisis de Componentes Principales (Eigenvalores y Eigenvectores de la matriz de covarianza)
- Preservamos 2-sigma de la varianza de los datos
- Reducimos la dimensionalidad de los vectores (datos) de 12 a 3(2)

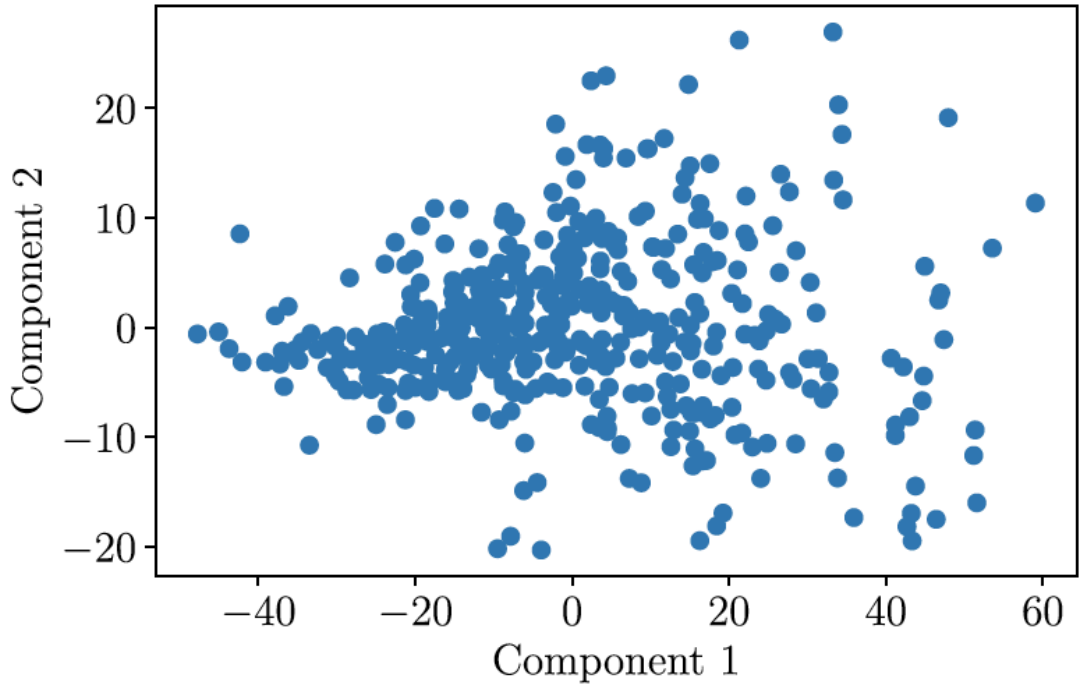
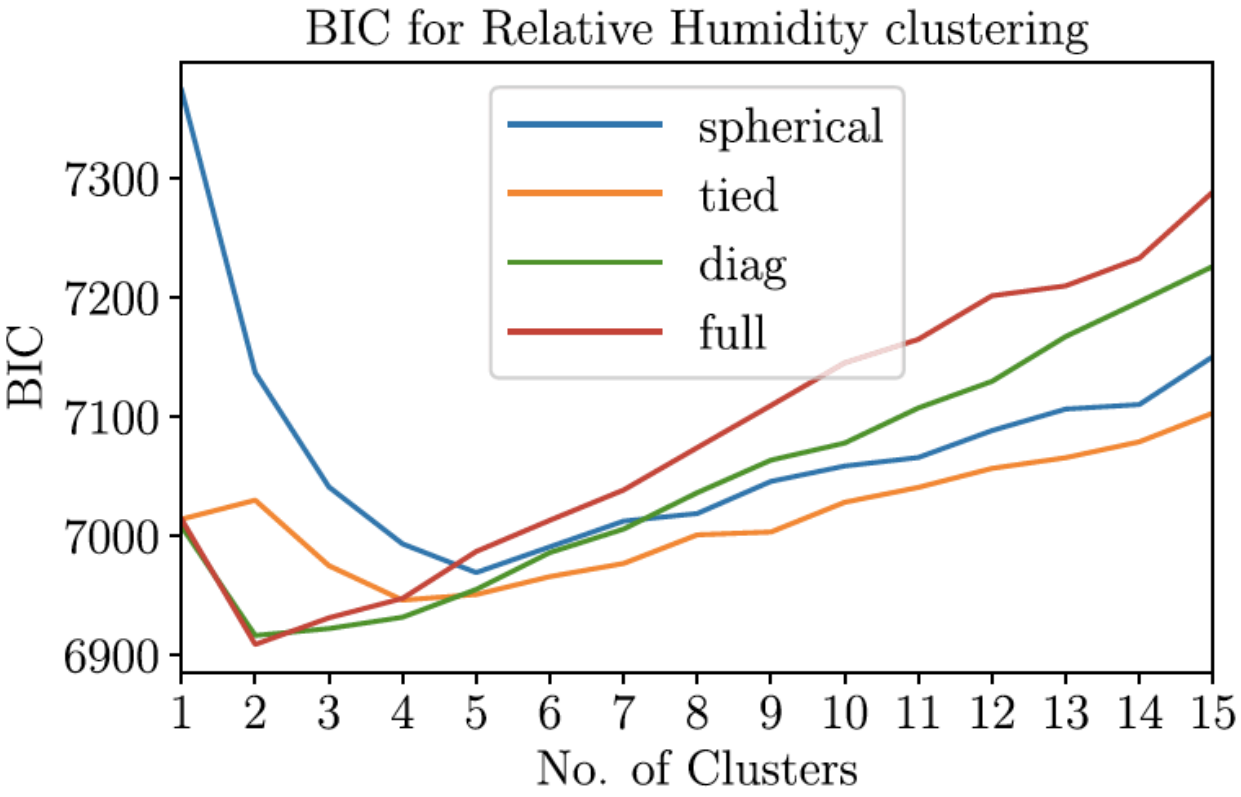


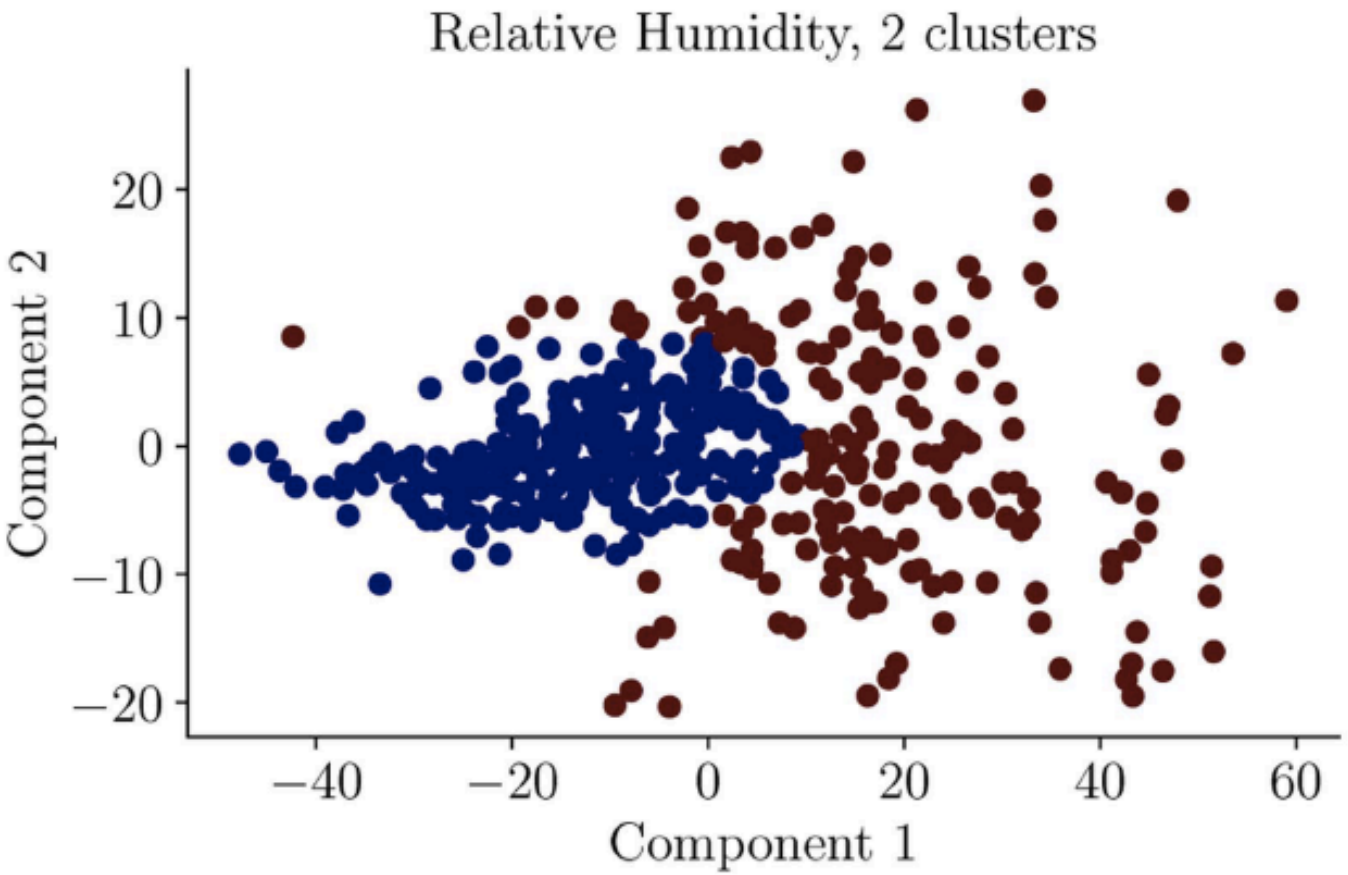
Figure 3. Relative humidity data projected across two principal components. The dimensionality of the data has been reduced from 12 to two while covering 95% of the variance of the data.

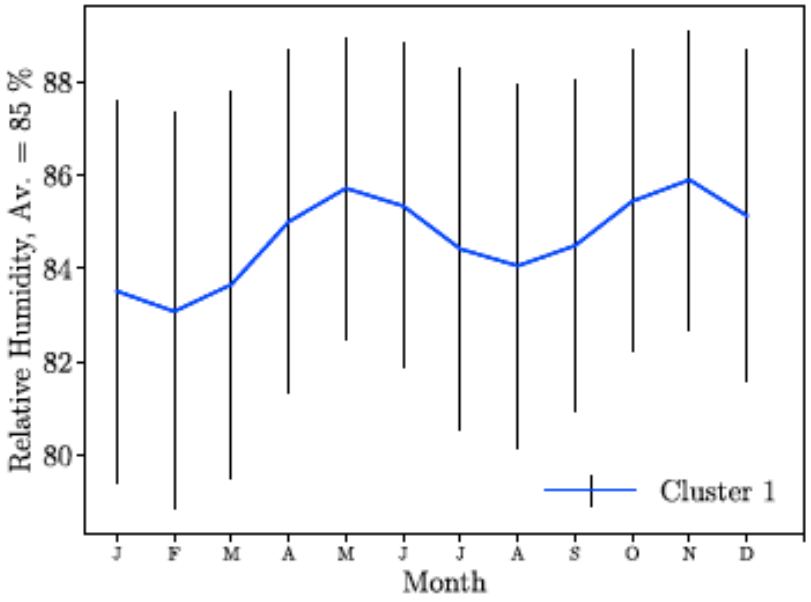
Criterio de Información Bayesiano

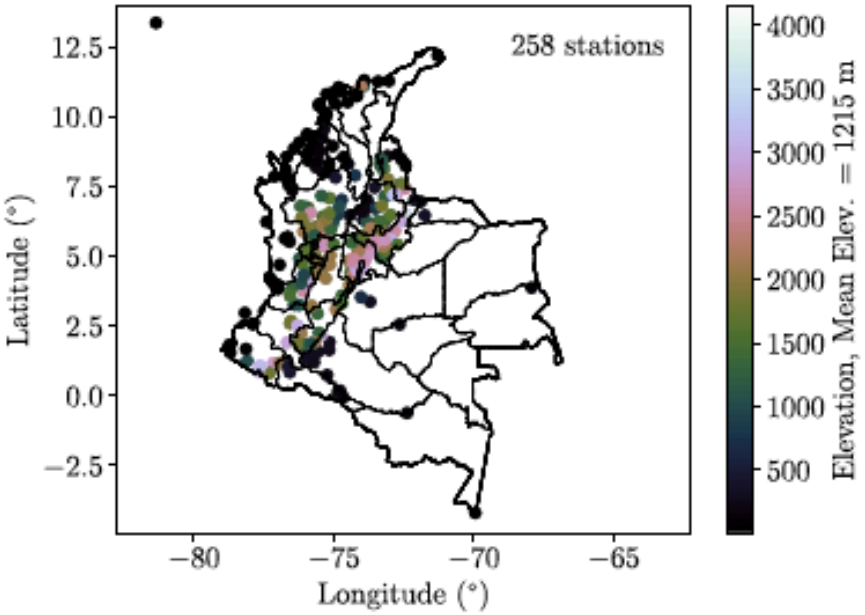
$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L})$$

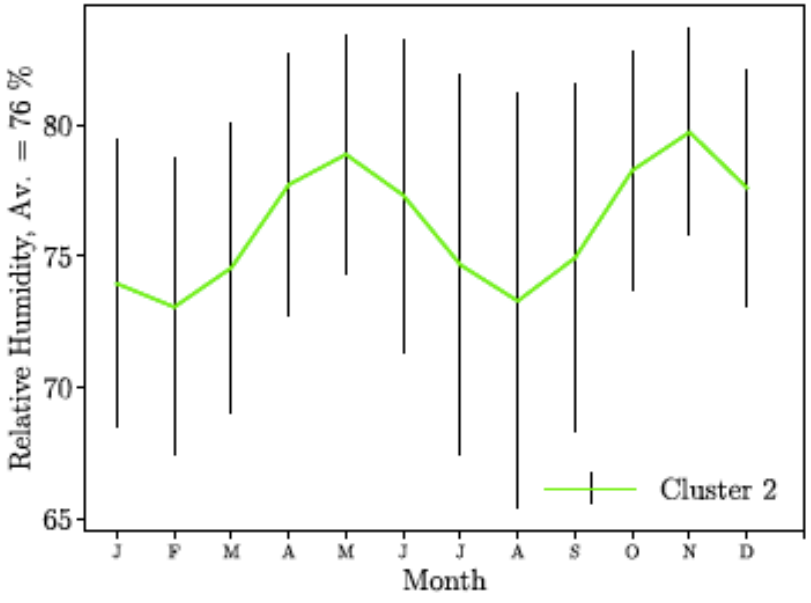


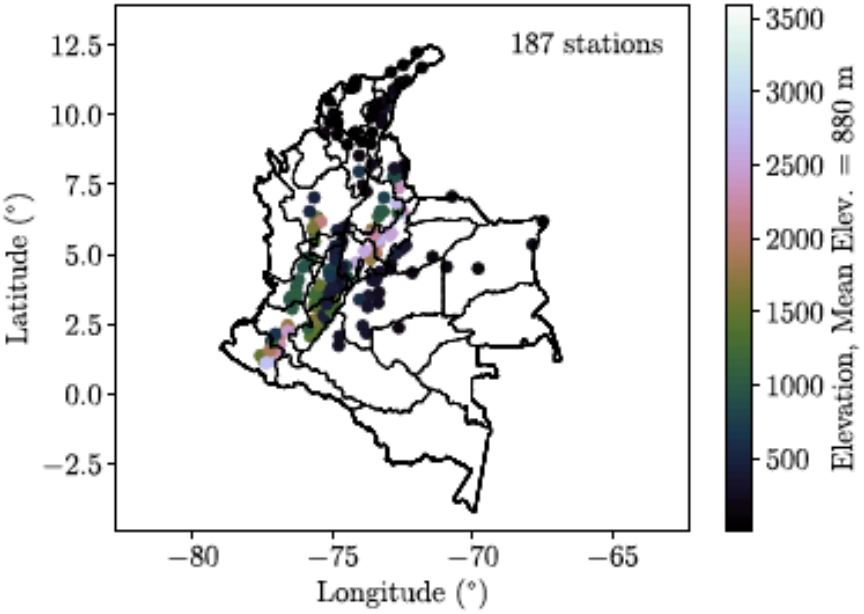
Low dimensional embedding



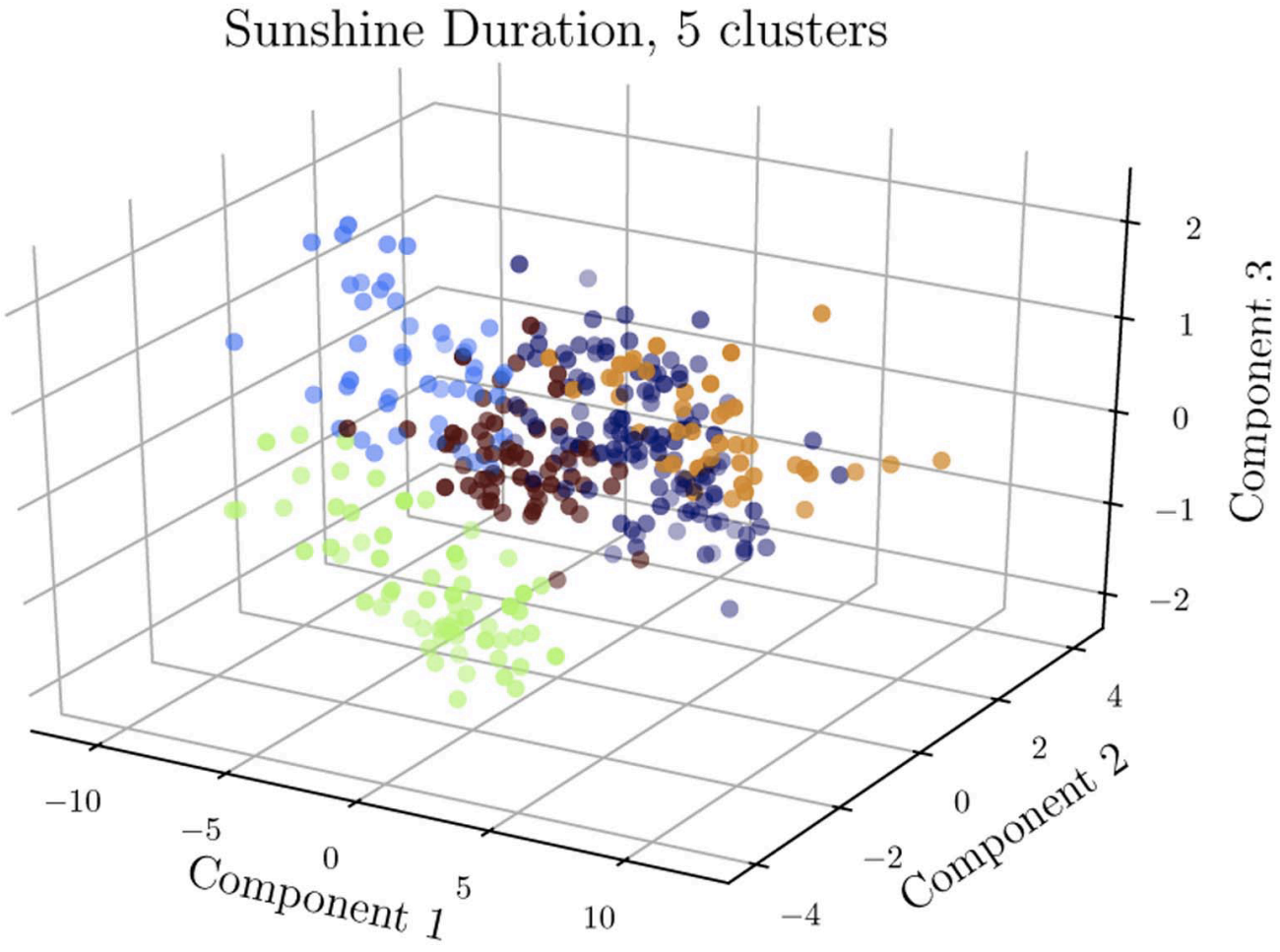

Relative Humidity, Av. = 85 %
88
86
84
82
80
Cluster 1
J F M A M J J A S O N D
Month

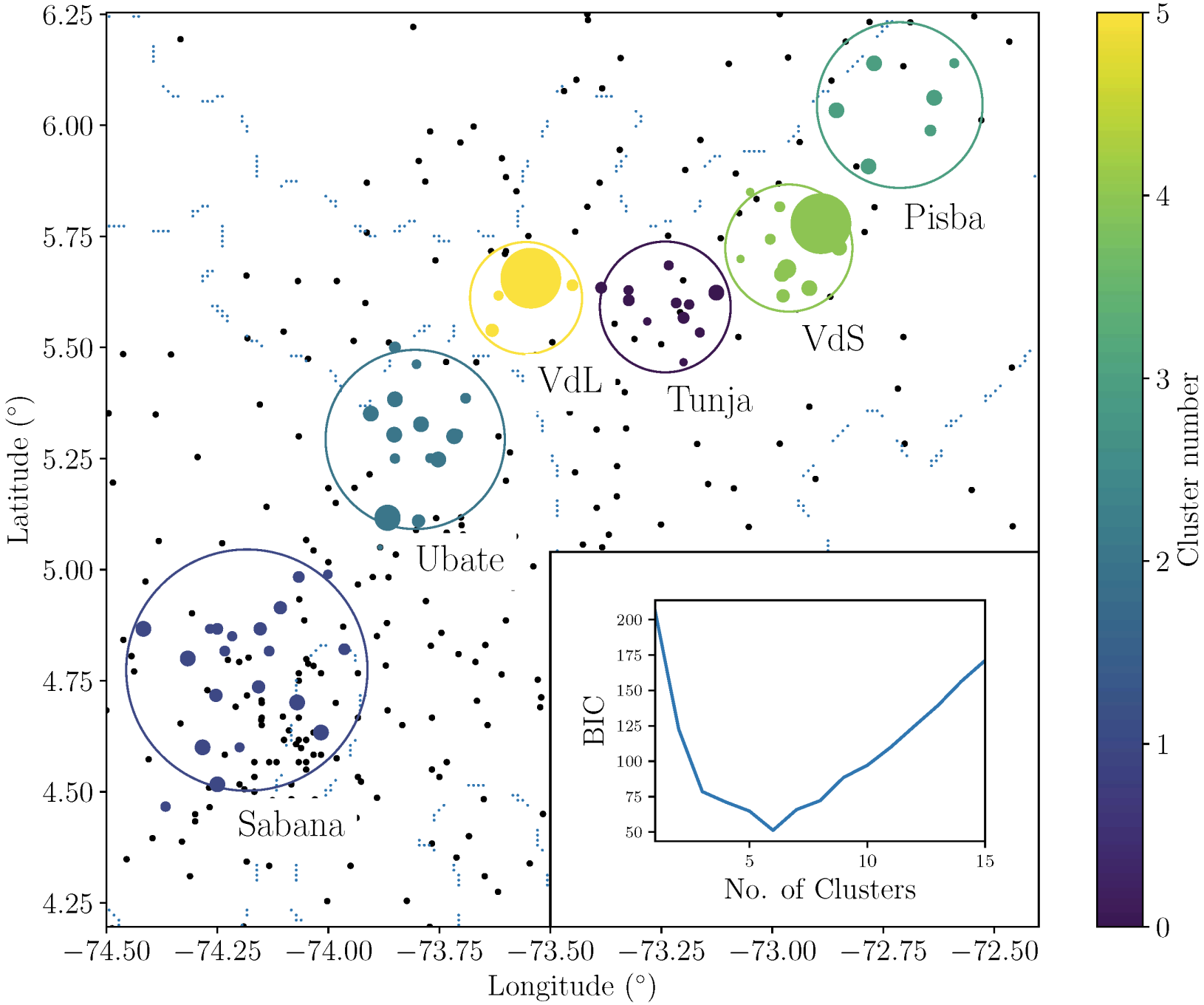

258 stations
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
Latitude (°)
−80 −75 −70 −65
Longitude (°)
4000
3500
3000
2500
2000
1500
1000
500
Elevation, Mean Elev. = 1215 m


Relative Humidity, Av. = 76 %
80
75
70
65
Cluster 2
J F M A M J J A S O N D
Month


187 stations
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
Latitude (°)
−80 −75 −70 −65
Longitude (°)
3500
3000
2500
2000
1500
1000
500
Elevation, Mean Elev. = 880 m

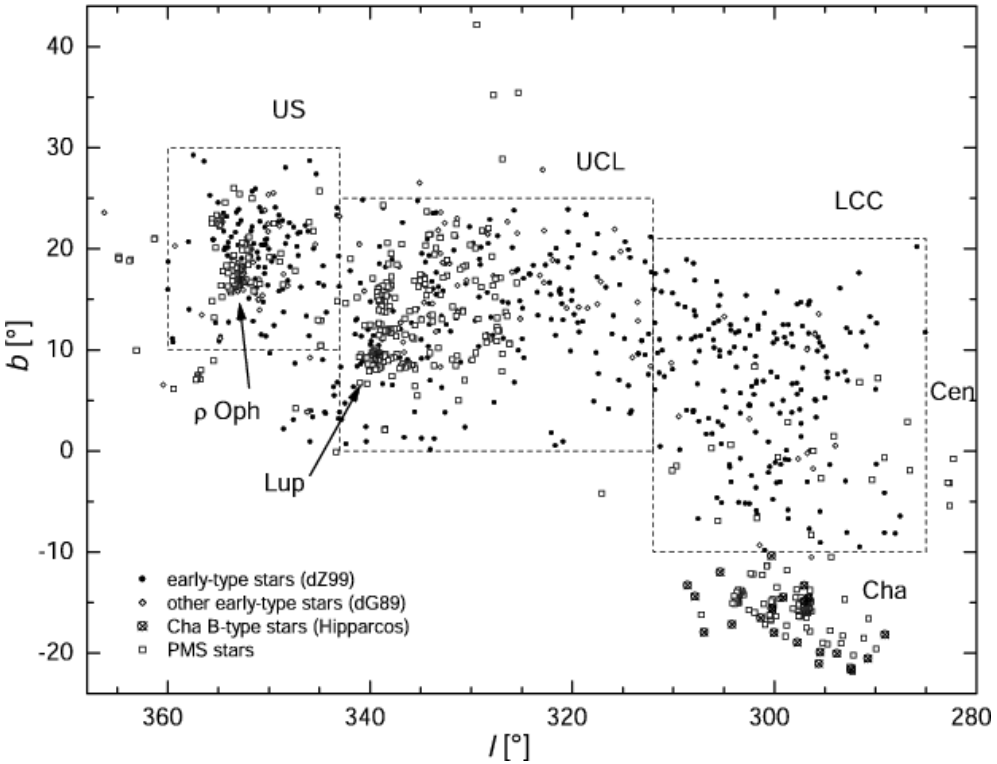
Low dimensional embedding




Pisba
VdS
VdL
Tunja
Ubate
Sabana
Latitude (°)
Longitude (°)
6.25
6.00
5.75
5.50
5.25
5.00
4.75
4.50
4.25
−74.50
−74.25
−74.00
−73.75
−73.50
−73.25
−73.00
−72.75
−72.50
Cluster number
0
1
2
3
4
5
BIC
200
175
150
125
100
75
50
5
10
15
No. of Clusters

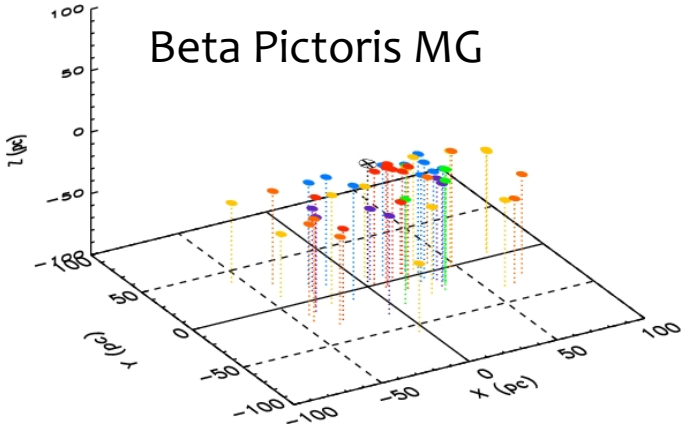
Moving Groups – Young Stellar Objects

Sco-Cen association

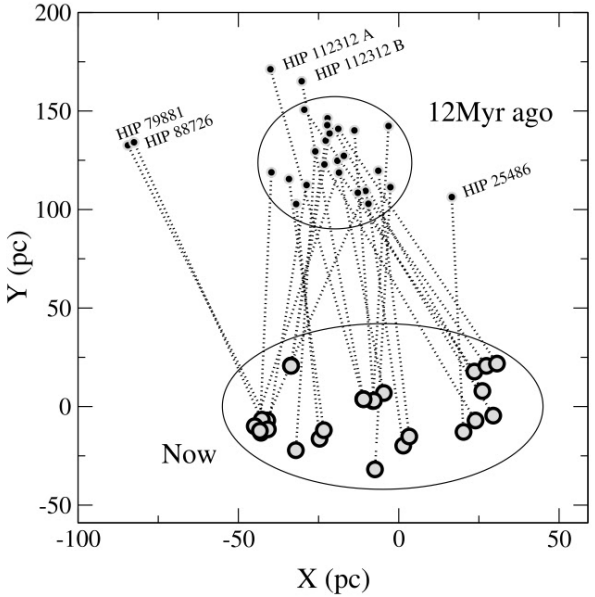


Sartori, Lépine, & Dias (2003)

Beta Pictoris MG

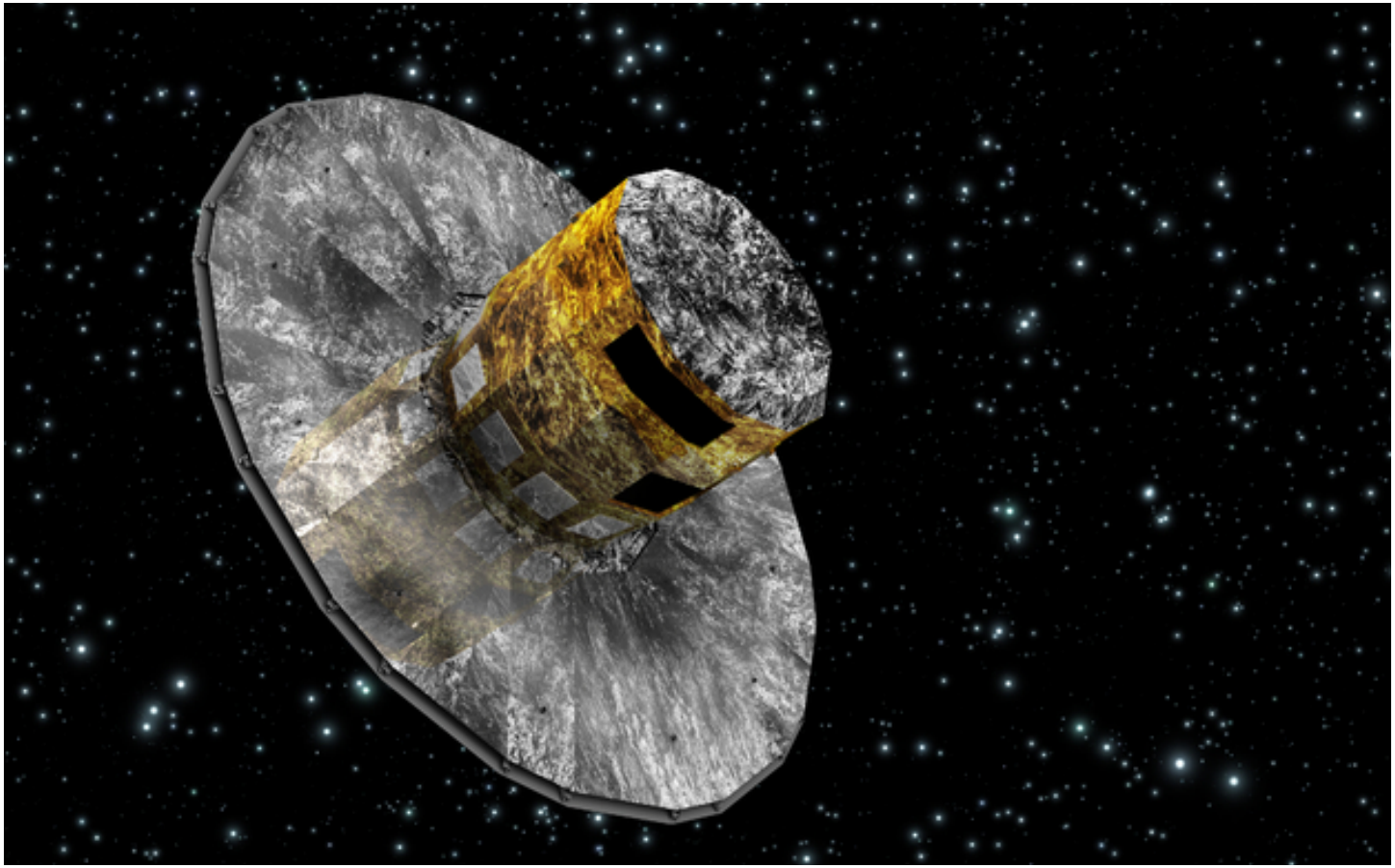


Schlieder, Lépine, & Simon (2012)

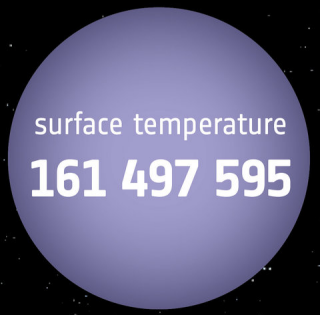


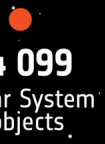
Song et al. (2003)

Gaia: Astrometría de alta precisión (2013 – 2022?)



→ HOW MANY STARS WILL THERE BE IN THE SECOND GAIA DATA RELEASE?





The second data release of ESA's Gaia mission is scheduled for publication on 25 April 2018.

Taller
moving_group_clustering.ipynb