Agrupamiento de Conjunto de información de Precipitación usando R

Oscar García-Cabrejo

Deparment of Civil and Environmental Engineering
University of Illinois at Urbana-Champaign

Spatio-temporal clustering of precipitation data using ensemble relational self-organizing maps with alternative metrics

Oscar Garcia-Cabrejo[a,1,*], Giovanni Quiroga[b]

[a] *2515 Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign*
[b] *Universidad Santo Tomas, Facultad de Ingenieria Civil, Tunja, Boyaca, Colombia*

Abstract

Definition of zones with similar characteristics of precipitation is critical for building accurate rainfall-runoff models that are an important tool in water management problems. In recent years, a variant of the Self-Organizing Map (SOM) called Two Level SOM has been used to define these clusters using the statistical and multiscale properties of precipitation. There are three problems when Two Level SOM are used in hydrologic applications:

Los resultados aquí presentados hacen parte de este artículo enviado a la revista *Advances in Water Resources*

Contenido

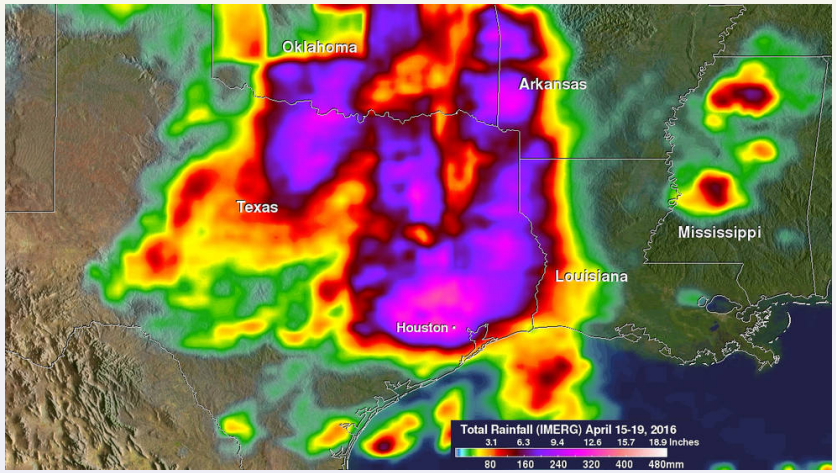
Introducción

Fuente: www.http://conserve-energy-future.com

Introducción

- El incremento en las emisiones de CO_2 ha causado incremento de temperatura (Calentamiento Global)
- Incremento de Temperatura afecta patrones de precipitación y sequía (Colombia 2010-2016)

Manejo Recurso Hídrico



Fuente: www.nasa.gov

Manejo Recurso Hídrico

- Estudiar variación precipitación en espacio y tiempo

Manejo Recurso Hídrico

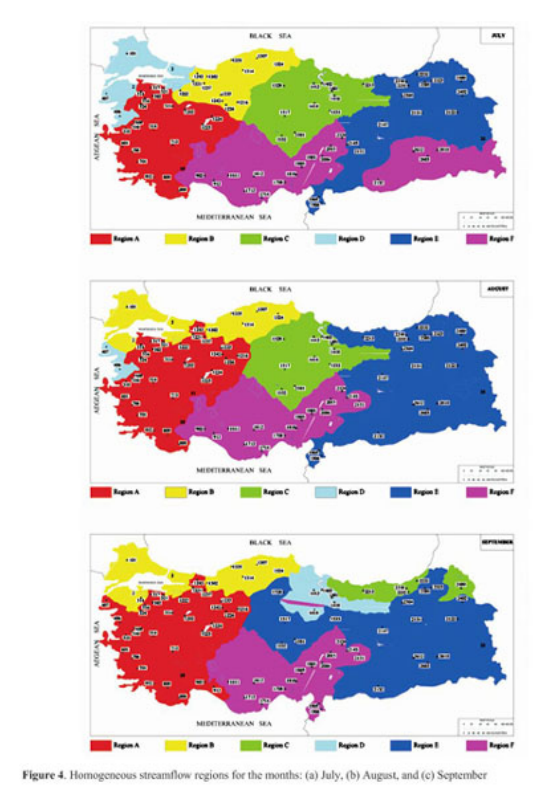


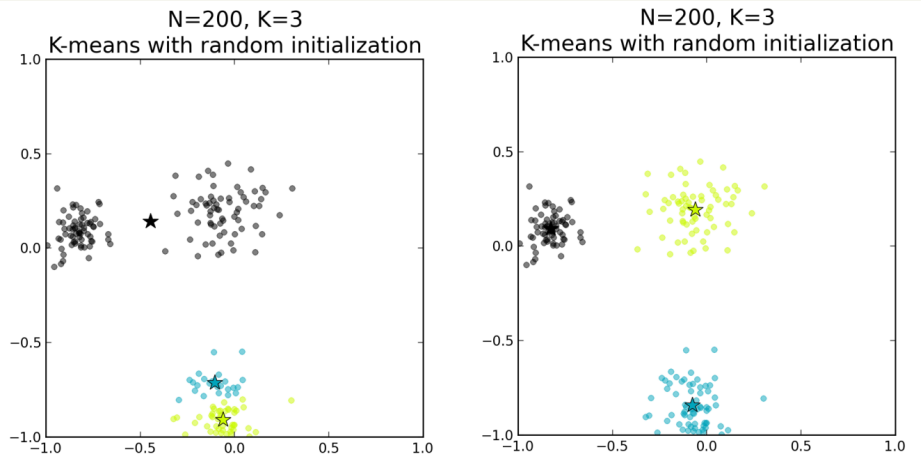
Figure 4. Homogeneous streamflow regions for the months: (a) July, (b) August, and (c) September

Manejo Recurso Hídrico

- Identificar áreas con caracteristicas hidrológicas similares (espacio y tiempo)
- Regionalización hidrológica
- Regionalización hidrológica = Agrupamiento (espacio + tiempo)

Fuente: http://callisto.ggsrv.com

Problema



Problemas Agrupamiento

1. Inicialización aleatoria

Fuente: https://datasciencelab.files.wordpress.com/

Problema

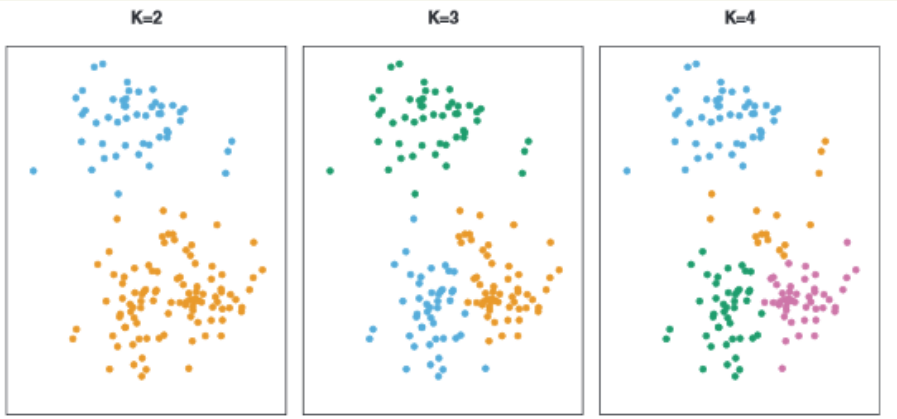


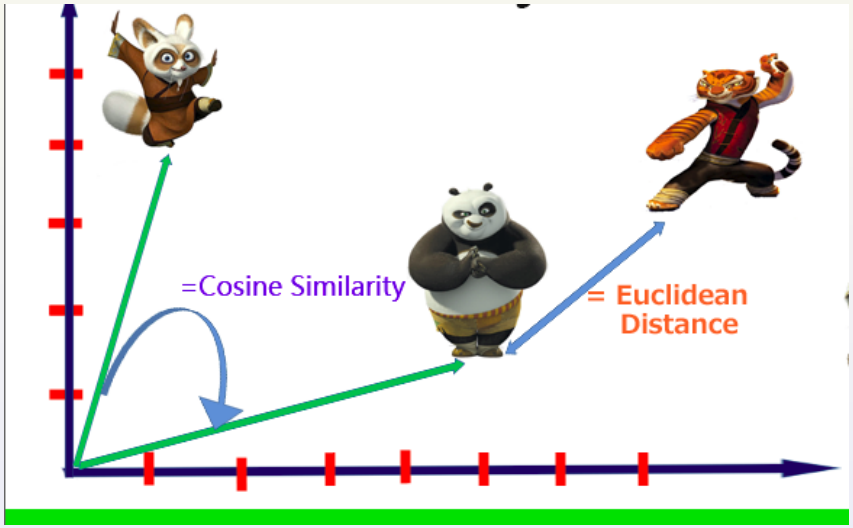
FIGURE 10.5. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.*

Problemas Agrupamiento

1. Inicialización aleatoria
2. Definición del número de grupos

Fuente: www.andrewjahn.com

Problema



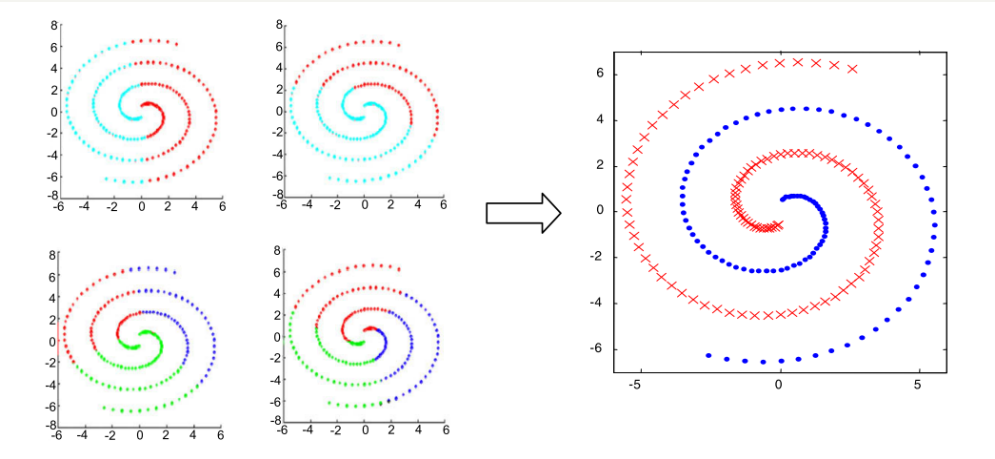
Fuente:
www.bigdata-madesimple.com

Problemas Agrupamiento

1. Inicialización aleatoria
2. Definición del número de grupos
3. Medida de disimilaridad

Solución Propuesta

Agrupamiento



- Problemas 1-2

  Agrupamiento de Conjunto (Ensemble Clustering). Metodología de acumulación de evidencia.

Fuente: A

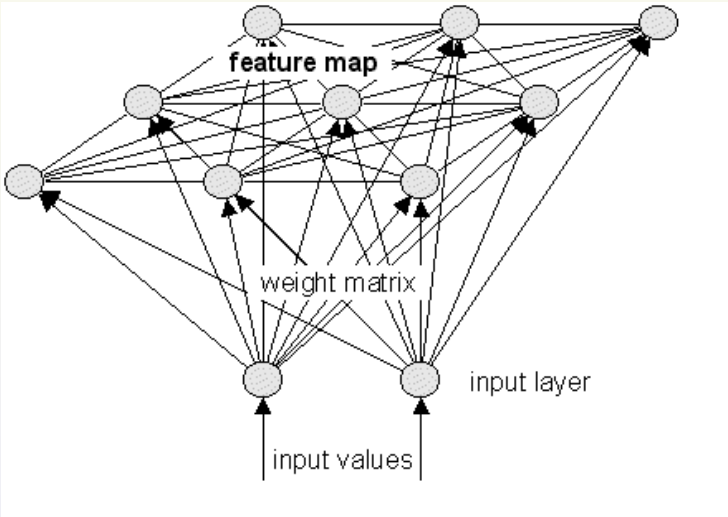
Solución Propuesta

Fuente: http://simaaron.github.io

Agrupamiento

- Problemas 1-2
  Agrupamiento de Conjunto (Ensemble Clustering). Metodología de acumulación de evidencia.
- Problema 3 Medida disimilaridad dependiendo caracterśticas hidrológicas

Solución Propuesta

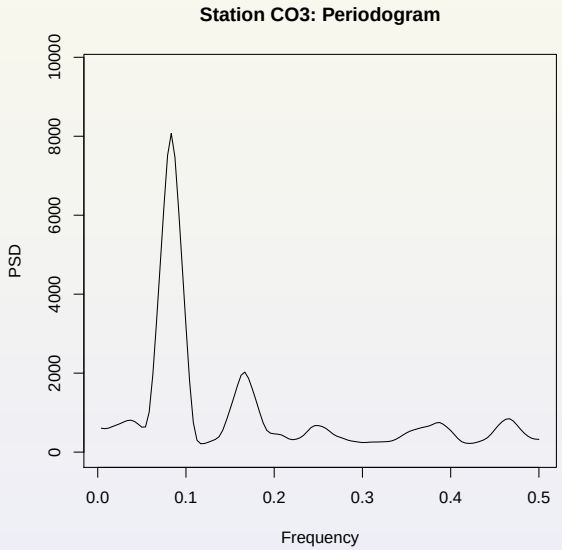


Fuente: http://cse-wiki.unl.edu

Agrupamiento

- Técnica de agrupamiento: Redes Neuronales Auto-organizadas relacionales de Kohonen de 2 niveles
- Variable: Función Densidad Espectral Precipitación
- Medida disimilaridad: (Itakura-Saito, Predicción)
- Implementación en R

Solución Propuesta



Fuente: Autor

Agrupamiento

- Técnica de agrupamiento: Redes Neuronales Auto-organizadas relacionales de Kohonen de 2 niveles
- Variable: Función Densidad Espectral Precipitación
- Medida disimilaridad: (Itakura-Saito, Predicción)
- Implementación en R

Solución Propuesta

$$D_{\mathrm{IS}}(\hat{f}_1(\omega), \hat{f}_2(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{\hat{f}_1(\omega)}{\hat{f}_2(\omega)} - \log \frac{\hat{f}_2(\omega)}{\hat{f}_1(\omega)} - 1 \right] d\omega$$

Agrupamiento

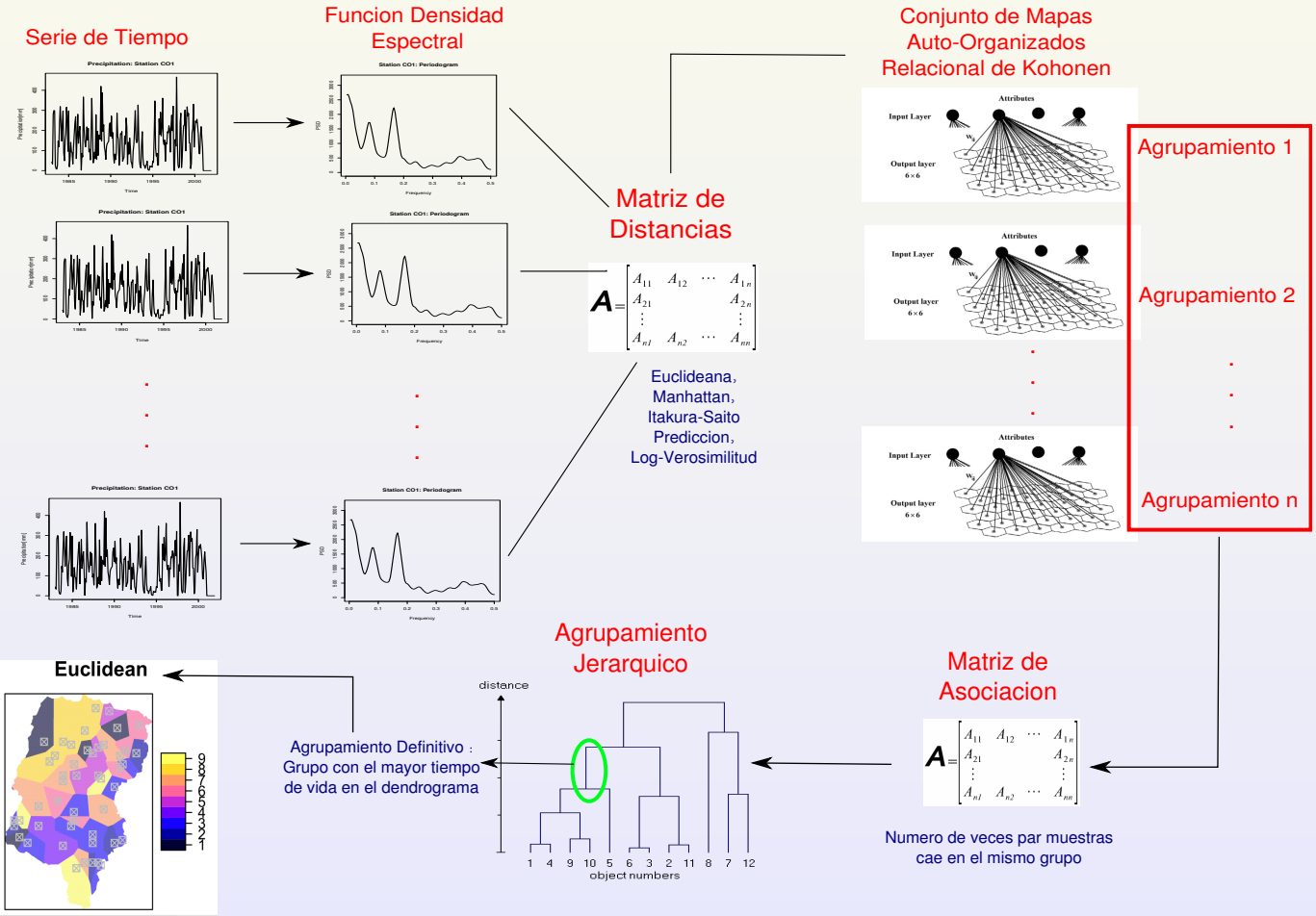
- Técnica de agrupamiento: Redes Neuronales Auto-organizadas relacionales de Kohonen de 2 niveles
- Variable: Función Densidad Espectral Precipitación
- Medida disimilaridad: (Itakura-Saito, Predicción)
- Implementación en R

Solución Propuesta

Agrupamiento

- Técnica de agrupamiento: Redes Neuronales Auto-organizadas relacionales de Kohonen de 2 niveles
- Variable: Función Densidad Espectral Precipitación
- Medida disimilaridad: (Itakura-Saito, Predicción)
- Implementación en R

Metodología



Medidas de Disimilaridad

Distancia Euclidiana

$$D_{\text{Eucl}}(\hat{f}_1(\omega), \hat{f}_2(\omega)) = \sqrt{\sum_{k=1}^{K} |\hat{f}_1(\omega_k) - \hat{f}_2(\omega_k)|^2} \tag{1}$$

Distancia de Manhattan

$$D_{\text{Eucl}}(\hat{f}_1(\omega), \hat{f}_2(\omega)) = \sum_{k=1}^{K} \left| \hat{f}_1(\omega_k) - \hat{f}_2(\omega_k) \right| \tag{2}$$

Medidas de Disimilaridad

Distancias Espectrales 1

Divergencia Itakura-Saito:

$$DI_{\mathrm{IS}}(\hat{f}_1(\omega), \hat{f}_2(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{\hat{f}_1(\omega)}{\hat{f}_2(\omega)} - \log \frac{\hat{f}_2(\omega)}{\hat{f}_1(\omega)} - 1 \right] d\omega \qquad (3)$$

Divergencia de Predicción

$$DI_{\mathrm{PRED}}(\hat{f}_1, \hat{f}_2) = \sqrt{\int_{-\pi}^{\pi} \left(\log \frac{\hat{f}_1(\omega)}{\hat{f}_2(\omega)} \right)^2 \frac{d\theta}{2\pi} - \left(\log \frac{\hat{f}_1(\omega)}{f_2(\hat{\omega}) \frac{d\theta}{2\pi}} \right)^2} \qquad (4)$$

Medidas de Disimilaridad

Distancias Espectrales 2

Distancia de Log-verosimilitud:

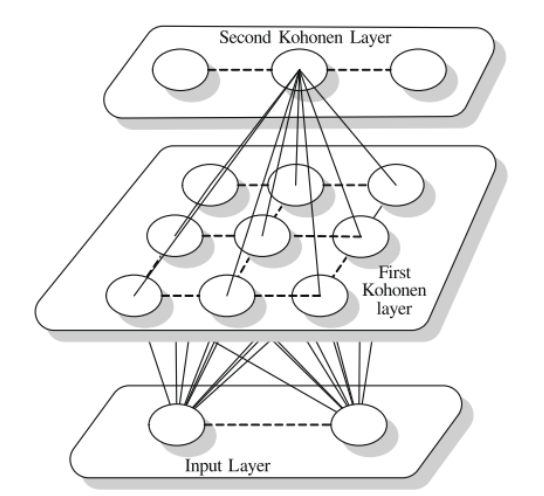
$$DI_{\text{LogLik}}(\hat{f}_1(\omega), \hat{f}_2(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{f}_1(\omega) \log \left[\frac{\hat{f}_1(\omega)}{\hat{f}_2(\omega)} \right] d\omega \qquad (5)$$

Distancia usando divergencias:

$$D = \frac{1}{2} \left[DI(\hat{f}_1(\omega), \hat{f}_2(\omega)) + DI(\hat{f}_2(\omega), \hat{f}_1(\omega)) \right] \qquad (6)$$

Implementación: Fortran2003 $\rightarrow$ C $\rightarrow$ R (usando función `.Call`)

Mapa Auto-Organizado de Kohonen

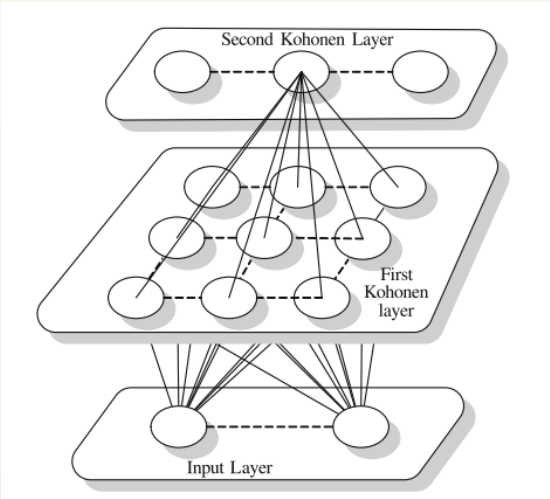


Fuente: Hsu, K.-c. & Li, S. (2010)

Definición

- Tipo de red neuronal artificial
- Proyecta información multivariada en espacio de 2 dimensiones
- Preserva topología de conjunto de datos (muestras cercanas en el espacio de alta dimensionalidad son muestras cercanas en espacio de 2 dimensiones)

Mapa Auto-Organizado de Kohonen

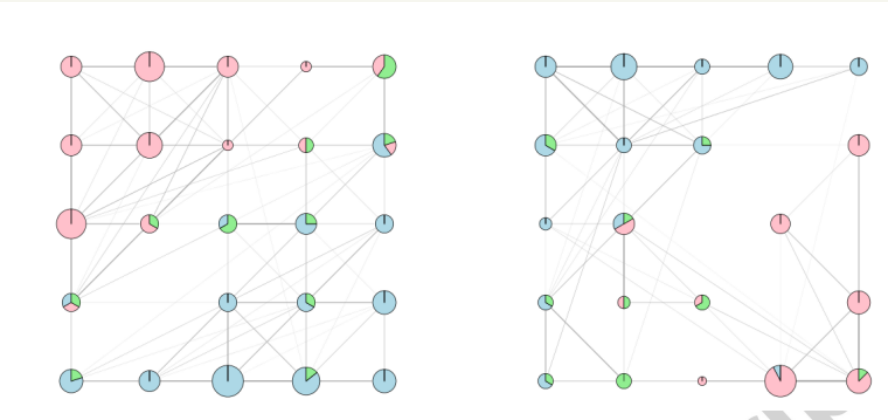


Fuente: Hsu, K.-c. & Li, S. (2010)

Agrupamiento

- Primera capa: Datos de alta dimensionalidad proyectados a espacio de 2 dimensiones
- Segunda capa: Datos proyectados en 2 dimensiones son proyectados a 1 dimensión
- Número de grupos = Número nodos segunda capa

Mapas Auto-Organizados Relacionales



Ejemplos de Mapas Auto-organizados relacionales obtenidos para un conjunto de libros de política usando diferentes métricas. Fuente: Olteanu & Villa, 2015

Definición

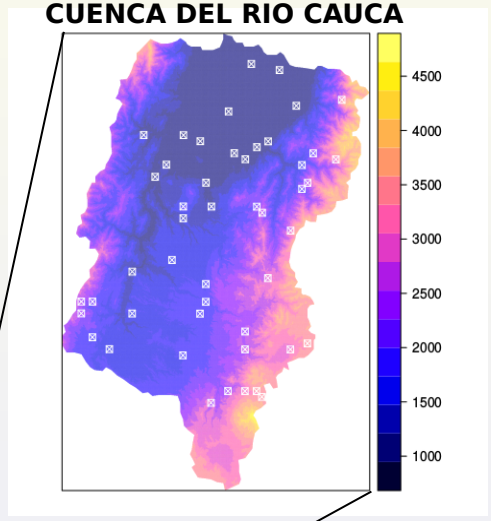
- Mapas Auto-organizados que operan con matrices de distancia o disimilaridad
- Mapas Auto-organizados Relacionales de dos niveles fueron implementados en R
- Más detalles en este link

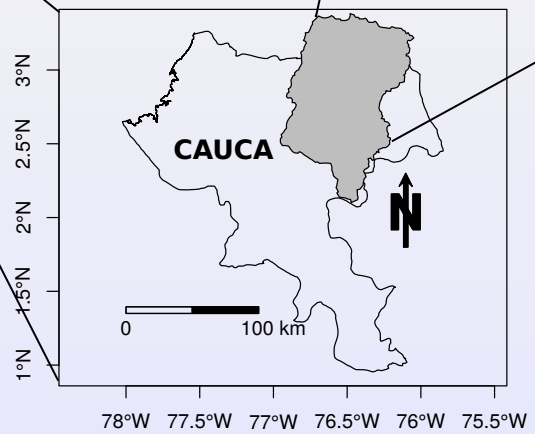
Investigación Reproducible

Metodología

- Basada en *Reproducible Research with R and Rstudio* 2nd Ed por Christopher Gandrud
- Implementada en R y Rstudio con paquete `knitr`
- Versionamiento usando SVN
- Análisis Estadístico automatizado con `make` en Linux
- R scripts documentados usando `Rmarkdown`

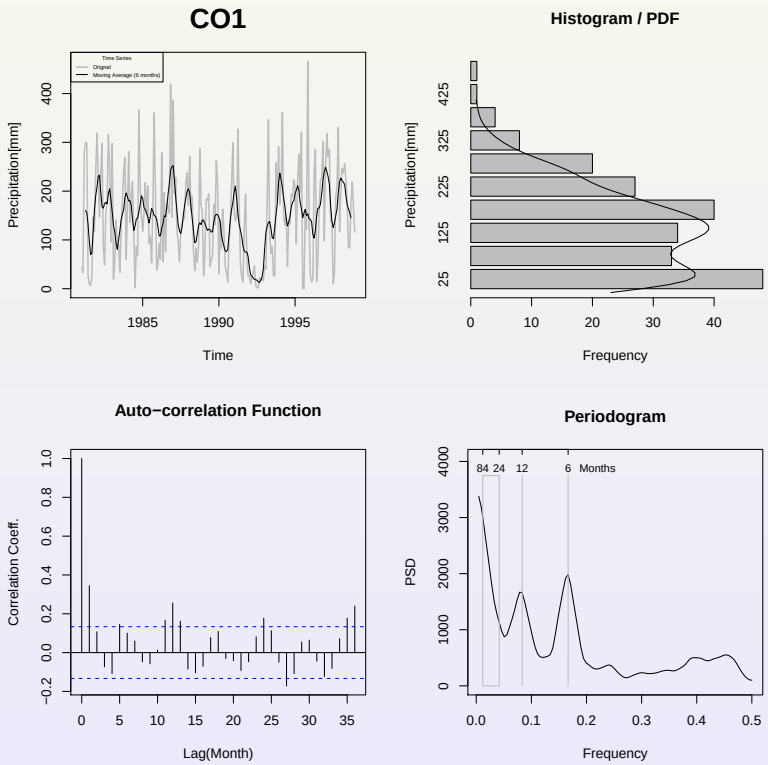
Cuenca Alta del Rio Cauca, Colombia





Area 8440 km2
49 Estaciones
Registro 1981-2000
topografía compleja

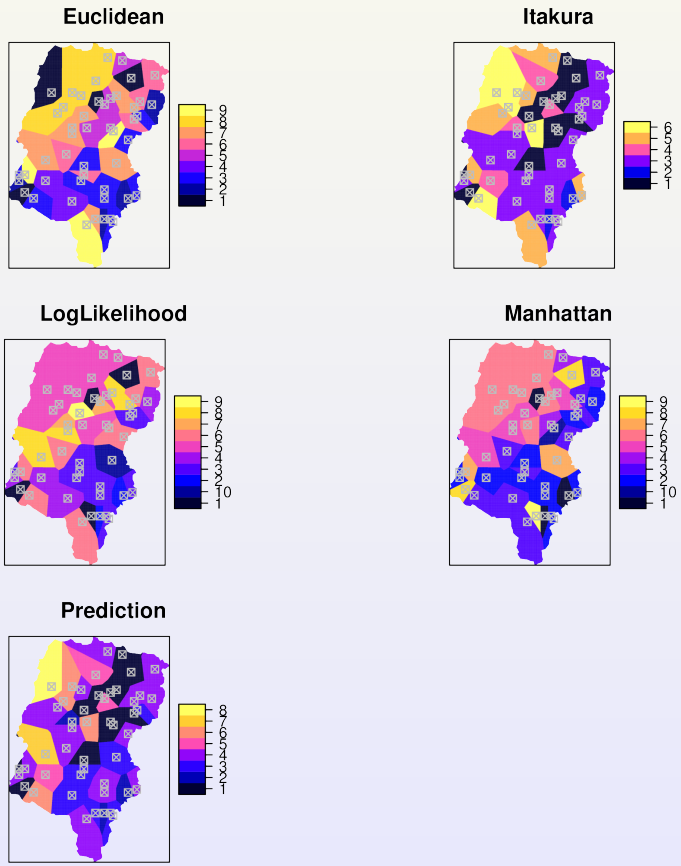
Información de Precipitación



Características

- Rango temporal 1981 – 2000
- Distribución asimétrica
- Componentes periódicas: 6 y 12 meses
- R: Paquete base (`acf`,`spectrum`), graphics

Resultados



Resultados

- Claras diferencias entre las zonas definidas usando cada distancia
- Visualización: Paquetes `sp,raster`

Conclusiones

- Análisis Exploratorio de Datos y Espectral de la Precipitación facilitada paquetes R.
- Redes Relacionales de Kohonen de 2 niveles es metodología apropiada para realizar regionalización de precipitación usando Agrupamiento de Conjunto usando la Función de Densidad Espectral con distancias apropiadas.
- Zonificación por características de precipitación se convierte en herramienta importante en gestión de recurso hídrico.