Identificación y caracterización de patrones climáticos en la ciudad de Manizales, usando técnicas de series de tiempo y de conglomerados

Elizabeth Solórzano Tovar

Universidad Nacional de Colombia
Facultad de Ciencias Exactas y Naturales
Departamento de Matemáticas y Estadística
Manizales, Colombia
2016

Identificación y caracterización de patrones climáticos en la ciudad de Manizales, usando técnicas de series de tiempo y de conglomerados

Elizabeth Solórzano Tovar

Tesis presentada como requisito parcial para optar al título de:
Magíster en Ciencias - Matemática Aplicada

Director:
Doctor Mauricio Orozco Alzate

Línea de Investigación:
Reconocimiento de Patrones
Grupo de Investigación:
Control y Procesamiento Digital de Señales

Universidad Nacional de Colombia
Facultad de Ciencias Exactas y Naturales
Departamento de Matemáticas y Estadística
Manizales, Colombia
2016

Dedicatoria

A mis hijas

Agradecimientos

Quiero expresar mis más sinceros agradecimientos al director del trabajo por el apoyo y acompañamiento permanente para el desarrollo del trabajo.

Resumen

En este trabajo se avanza en la identificación de patrones climáticos en la ciudad de Manizales a partir de datos hidrometeorológicos obtenidos de la red de estaciones coordinadas por el Instituto de Estudios Ambientales IDEA de la Universidad Nacional de Colombia - Sede Manizales. Se estructura un marco matemático conceptual completo, formalizado y de nomenclatura apropiada para los métodos de conglomerados usados en el trabajo y para los algoritmos involucrados. Se describe la base de datos usada, se muestran los resultados de agrupamientos por simplificación de variables y por análisis de conglomerados, obteniéndose concordancia entre estos resultados. Datos históricos de distintos años se usan para corroborar las conclusiones obtenidas.

Palabras clave: métricas, distancias, disimilitud, análisis de conglomerados, métodos jerárquicos, métodos de partición, patrones climáticos.

Abstract

Title: **Identification and characterization of weather patterns in the city of Manizales, using time-series techniques and clustering**

This work deals with the identification of weather regimes in Manizales city from a hydrometeorological database managed by the IDEA (Environmental Studies Institute of *Universidad Nacional de Colombia*). A suitable mathematical framework is structured for the clustering techniques and algorithms involved in the study. The database is introduced together with the definition of the methodological design for the variables simplification and clustering analysis. The results obtained for the different techniques keep concordance. Historical data are used for validation purpose.

Keywords: metrics, distances, dissimility, clustering, hierarchical clustering, agglomerative clustering, partition methods, weather regimes.

Contenido

1 Introducción

El propósito principal de este trabajo radica en obtener una comprensión conceptual que, junto con herramientas y estrategias de análisis de series de tiempo y conglomerados, permita obtener una identificación y caracterización de patrones climáticos en la ciudad de Manizales - Colombia; esto, a partir del análisis estadístico de los datos obtenidos en las redes meteorológicas operadas por el Instituto de Estudios Ambientales de la Universidad Nacional de Colombia - Sede Manizales (IDEA). Dichos patrones identificados y caracterizados habrán de servir como insumo para la gestión integral del riesgo en la ciudad.

Para dar alcance a tal propósito se plantean objetivos específicos como: Obtener procesos de reducción de la complejidad de las variables medidas, a nuevas variables que faciliten su estudio y análisis; establecer modos naturales, comunes y no triviales de agrupamiento del clima de Manizales; e identificar variables de mayor relevancia en el comportamiento global del clima en la ciudad.

El IDEA en cumplimiento de su función de "formular, orientar y desarrollar proyectos de investigación interdisciplinaria y programas en el campo de los estudios ambientales" ([15]) ha avanzado, en convenio con otras entidades de intereses afines, en la estructuración de una red de monitoreo climático para el departamento de Caldas. Claramente, el alto potencial de uso de la información recopilada incluye aspectos como la gestión del riesgo, estudios calidad del ambiente y la identificación de patrones climáticos. Es esta última posibilidad la que motiva este trabajo, máxime si se tiene presente el fuerte impacto que tiene el clima de Manizales en la ocurrencia de siniestros de alto costo económico y social.

El estudio se focaliza en las mediciones registradas por las estaciones meteorológicas denominada "Posgrados", "Emas", "Enea" e "Ingeominas", ubicadas en distintos sectores geográficos de la ciudad de Manizales. Particularmente en los registros obtenidos para las variables temperatura, radiación solar, humedad relativa y precipitación durante los años $2009, 2010$ y 2011.

La estructura del trabajo es la siguiente: En el capítulo 2 se construye el marco teórico-conceptual referente a los métodos de análisis de conglomerados; haciéndose énfasis en los métodos jerárquicos aglomerativos y en los métodos de partición, por ser de alta conveniencia en el estudio de la base de datos objeto de este trabajo. Se incluye además una revisión

del estado del arte de los métodos de conglomerados, junto con su uso en hidrometeorología y con los avances en la gestión de la información climática de Manizales.

En el capítulo 3 se hace una descripción de la composición, estructura y delimitación de la base de datos usada en el trabajo. Luego se detalla el análisis de tendencia central efectuado para la reducción de la complejidad de las variables en la base de datos. Finalizando con la presentación de la metodología empleada en el estudio de conglomerados adelantado para la identificación de los patrones climáticos.

Los resultados de los estudios de tendencia central y dispersión de los datos son presentados y discutidos en el capítulo 4. En tanto que, lo obtenido en los estudios de análisis de conglomerados se muestra en el capítulo 5. Por último, el capítulo 6 presenta las conclusiones finales y las recomendaciones para trabajos futuros.

2 Marco teórico

En este capítulo se establece la fundamentación conceptual necesaria para el estudio objeto de este trabajo. Empezamos con la presentación de los conceptos propios del análisis de conglomerados, luego presentamos la formalización de estas ideas con las definiciones y algoritmos concretos a emplearse, en esta parte resultan fundamentales las referencias [16], [8], [11], [12]. Finalmente, se incluye también una revisión del estado del arte de los aspectos relevantes de este trabajo como lo son los avances pertinentes en análisis de conglomerados, el uso de conglomerados en estudio de datos metereológicos y el estado de los estudios de datos del clima en la región de Manizales.

El análisis de agrupamientos o análisis de conglomerados (del inglés *cluster analysis*) es la herramienta más conocida para el análisis de datos multivariados no estructurados. Esto gracias a que permite el agrupamiento de un conjunto de objetos en categorías o clases (que llamaremos *agrupamientos o conglomerados*, del inglés *cluster*) según el grado de similitud entre ellos.

En esta tarea se asume que los objetos pueden ser divididos razonablemente en conglomerados que contienen objetos similares y que, a su vez, se puede diferenciar suficientemente un conglomerado de otro. Esto, dado que se pretende clasificar los diferentes elementos en clases disyuntas que guarden una similitud importante al interior de cada una, pero que tengan una disimilitud marcada entre ellas. En general, un conglomerado es considerado como un conjunto de elementos (objetos, puntos) en los que cada elemento está "cerca" (en un sentido apropiado) a los otros elementos de su agrupamiento, en tanto que los miembros de diferentes conglomerados están "muy lejos" uno de otro. De esta manera, los conglomerados pueden ser entonces considerados como "regiones de alta densidad" dentro de un espacio multidimensional, separados por "regiones de baja densidad", [16], [8], [11], [12].

Según [8], los propósitos más frecuentes para la construcción y análisis de conglomerados son la identificación de una estructura natural en los objetos, la búsqueda de esquemas conceptuales útiles que expliquen el agrupamiento de algunos objetos, la formulación de hipótesis mediante la descripción y exploración de los conglomerados conformados y la verificación o confirmación de si las estructuras definidas mediante otros procedimientos están realmente presentes en los datos.

Para alcanzar estos propósitos se deben tener en cuenta tres consideraciones. La primera ¿Cómo se puede medir la similitud entre los objetos?, la segunda ¿Cómo se forman los conglomerados? y la tercera ¿Cómo escoger el número de agrupamientos que se deben tener en cuenta para el análisis por conglomerados de un conjunto de datos dado? Para la primera consideración se debe tener en cuenta la escogencia de un método que nos permita determinar o reconocer objetos como similares o disímiles. Es esta necesidad la que da origen a las denominadas medidas de similitud y medidas de disimilitud. Claramente, las unas son complementarias a las otras y usualmente a partir de una medida de similitud se define una de disimilitud y viceversa. Por lo que se habla de medidas de similitud sin mayor distinción.

Las medidas de similitud se pueden clasificar en: medidas de distancia, medidas angulares y los denominados coeficientes de asociación. Las primeras incluyen, pero no se limitan (en este trabajo) a, las que reúnen las propiedades de métrica sobre un espacio métrico en sentido matemático. Las de mayor uso son las distancias: *euclidiana, Mahalanobis, Manhattan* y *Minkowski*. Las medidas angulares más utilizadas son el coseno del ángulo entre los dos puntos a comparar y el coeficiente de correlación de *Pearson*. Son de gran utilidad dos tipos de medidas cuando los objetos a clasificar son caracterizados por aspectos cuantitativos. Los coeficientes de asociación, a su vez, se emplean para datos en escala nominal, es decir que cada variable toma valores binarios de 0 (cero) para ausencia y 1 (uno) para presencia. En estos últimos se destacan el coeficiente de asociación simple, el coeficiente de *Jaccard, Rogers* y *Tanimoto*, entre otros (ver Sección 2.1 y [8] para detalles).

Para la segunda consideración se debe apuntar a la construcción de metodologías, métodos o procedimientos que permitan agrupar o determinar qué objetos son más similares y ubicarlos dentro de un determinado conglomerado. Cada método representa una perspectiva diferente para la conformación de los conglomerados, por lo que se recomienda escoger un método que se acople a la tipología esperada en los datos (categóricos, escalares, nominales, vectoriales), su distribución, la presencia de perturbaciones, las variables a considerar y la medida de similitud usada; o bien la comparación de varios métodos. Existen varios algoritmos para la conformación de conglomerados, como lo son los métodos jerárquicos, los métodos de partición, las nubes dinámicas, los métodos gráficos y la clasificación difusa.

Los métodos jerárquicos empiezan calculando una matriz de distancias entre los objetos, luego se forman grupos ya sea por un proceso aglomerativo o por un proceso de división. Para el caso del proceso aglomerativo se empieza asumiendo que cada objeto es un conglomerado, y luego los grupos cercanos se mezclan sucesivamente hasta que todos los objetos quedan dentro de un mismo conglomerado. Para el caso del proceso de divisivo, se inicia con todos los objetos dentro de un mismo conglomerado, éste es dividido en dos conglomerados, éstos en otros dos hasta que cada conglomerado llega a constar de uno sólo elemento.

Los métodos jerárquicos aglomerativos son los más utilizados por la simplicidad aritmética de sus cálculos. Para su implementación se cuenta con diversos métodos que se definen para la aglomeración de los objetos después de conformada la matriz de similitud. Entre estos criterios podemos destacar: *enlace simple, enlace completo y enlace promedio.*

El enlace simple o del "vecino más cercano" se caracteriza porque fusiona los conglomerados que están a la menor distancia o dentro de un límite de similitud dispuesto y porque la distancia entre dos conglomerados se mide como la menor distancia observada desde los puntos de un conglomerado a los puntos del otro conglomerado. En el enlace completo, o del "vecino más lejano" se establece que cualquier candidato a incluirse en el conglomerado existente, debe estar dentro de un nivel de similitud determinado con todos los miembros de ese conglomerado. Es decir, dos conglomerados son unidos sólo si los miembros más distantes de los dos conglomerados están lo suficientemente cerca de manera conjunta, "el suficientemente cerca" es dado por el nivel de similitud impuesto en cada etapa del algoritmo. Finalmente, en el enlace mediante el promedio, la distancia que se establece entre dos conglomerados es el promedio entre todos los pares de objetos de los dos conglomerados. Es decir, se une el caso u objeto al conglomerado si se logra un determinado nivel de similitud con el valor promedio obtenido. El promedio más comúnmente usado es la media aritmética de las similitudes entre los objetos.

Por su parte, los métodos de partición son métodos no jerárquicos que para su utilización tienen en cuenta los siguientes pasos:

1. Se particiona al conjunto en un determinado número de agrupamientos y a cada uno de ellos se le calcula el centroide.
2. Cada objeto es asignado al conglomerado cuyo centroide esté más cercano a él.
3. Se calcula el nuevo centroide de los conglomerados, éstos no son actualizados hasta tanto no se comparen sus centroides con todos los casos.
4. Se repiten los pasos 2. y 3. hasta que los casos resulten invariantes o cumplan un criterio de optimalidad.

Entre los métodos de partición más comunes está el método de las k-medias, éste hace una selección o partición inicial de los objetos, para luego modificar su configuración hasta obtener la mejor partición en términos de una función objetivo (ver Sección 2.3 para detalles). También están los métodos basados en la traza, estos minimizan la varianza dentro de los grupos, buscando detectar las diferencias entre ellos (ver [8] para detalles).

Otra opción en la literatura es la de nubes dinámicas, estos métodos utilizan una partición del conjunto de individuos, con el propósito de mejorarla u optimizarla respecto a una regla.

La optimización recurre a la utilización de procesos iterativos de cálculos, generalmente basados en métodos numéricos. Estos procesos requieren un criterio que permita comparar las calidades de dos particiones o clasificaciones que tienen el mismo número de agrupaciones. Dicho proceso termina cuando no se pueda mejorar la calidad de tal partición.

Los métodos gráficos son una alternativa más, cuyas técnicas más utilizadas incluyen los glifos, las estrellas, los rostros de Chernoff y los gráficos de Fourier.

- Un glifo (del inglés *glyph*) consta de un círculo de radio r con p rayos que salen de él. La posición y la longitud de cada rayo refleja el valor de la coordenada asociada con cada una de las p variables, las cuales pueden ser cualitativas o cuantitativas.
- En la técnica de estrellas, las variables se ubican sobre los radios de una estrella regular. La magnitud de cada variable si es cualitativa, se ubica sobre cada radio, así un valor máximo se representa en los extremos y un valor nulo en el centro de la circunferencia, el polígono que une los puntos ubicados sobre cada radio determina un individuo.
- Los rostros de Chernoff se basan en la representación de un vector de observaciones mediante características faciales como, por ejemplo, la cabeza, la boca, la nariz, los ojos, las cejas y las orejas. Para un problema particular se asigna a cada una de las variables un rasgo facial determinado.
- Por último, en los gráficos de Fourier, la proyección de la transformación de los vectores de respuesta p dimensionales por series de Fourier revelan los grupos o conglomerados, a manera de bandas que contienen ondas "paralelas".

Adicionalmente, existe la posibilidad de definir conglomerados difusos (del ingles *fuzzy*). Un conjunto difuso (borroso) puede pensarse como una clase de objetos con algún grado de pertenencia a éste. Esta relación de pertenencia algunas veces no está claramente definida. Una manera más formal de definir cómo funciona este método es la siguiente. Sea X una colección de objetos. Un conjunto difuso A de X es caracterizado por una función de pertenencia (característica) $f_A(x)$ la cual asocia a cada punto x de X un número real en el intervalo $[0, 1]$; donde el valor de $f_A(x)$ representa el "grado de pertenencia" de x a A. Un valor de $f_A(x)$ cercano a 1 corresponde a un alto grado de pertenencia de x a A. Cuando A es un conjunto en el sentido clásico, su función de pertenencia toma únicamente los valores 1 ó 0, de acuerdo con la pertenencia o la no pertenencia de x a A.

Retomemos ahora la última consideración que enunciamos antes para alcanzar los propósitos del agrupamiento: la decisión acerca del número apropiado de conglomerados a seleccionar. Se debe tener una estrategia para decidir sobre la cantidad de conglomerados a construir, cuya escogencia debe ir acompañada del conocimiento del especialista asociado al estudio en consideración y de la homogeneidad del número de conglomeredos, ya que si disminuye el

número de conglomerados escogidos, necesariamente disminuye la homogeneidad dentro de los mismos. Se destacan tres opciones para este requerimiento.

Una estrategia para tomar esta decisión se basa en el dendrograma que sugiere un número de conglomerados en cada paso (ver sección 2.1.1); sin embargo, un problema de esta estrategia es que se debe tener un alto vínculo con el especialista y el campo de aplicación, para poder decidir el lugar de corte del árbol, de modo que obtega un número óptimo de grupos.

Otro método se basa en la gráfica del número de conglomerados de un árbol jerárquico versus los coeficientes de fusión, que corresponde al valor numérico bajo el cual varios casos se mezclan para formar un grupo. El método consiste en representar en una gráfica los valores del coeficiente de fusión sobre el eje X y el número de conglomerados en el eje Y, se traza una línea que une los puntos de coordenadas del coeficiente de fusión y el número de grupos, el punto donde la línea trazada se hace horizontal sugiere el número apropiado de grupos.

La tercera estrategia para la selección del número de conglomerados adecuados es la del coeficiente silueta. Este mide la calidad de los conglomerados mediante la comparación de la disimilitud del objeto con respecto a su conglomerado y a los otros conglomerados. Esto genera una función cuya representación es una gráfica que tiene en el eje X los valores del coeficiente silueta y en el eje Y se disponen los conglomerados en consideración con una línea horizontal por cada objeto indicando su valor silueta. La estrategia consiste en efectuar el análisis para distintos números de conglomerados. Luego, se determina el valor medio de la silueta para los números de conglomerados considerados y se representa gráficamente. A continuación, en el eje X, se colocan los números de conglomerados considerados y en el eje Y los valores promedios de la silueta sobre todos los objetos en cada caso. Finalmente, se escoge como valor óptimo el número de conglomerados para el cual el coeficiente silueta alcanza el valor máximo; ver más detalles en la sección 2.4.

2.1. Agrupación jerárquica

Empezamos considerando un conjunto $\Im$ finito, digamos $\Im = \{X_1, X_2, \ldots, X_M\}$, la tarea de clasificar los elementos de $\Im$ en $K \leq M$ conglomerados consiste en obtener K subconjuntos no vacíos $C_1, C_2, ..., C_K$ de $\Im$ tales que

$$\begin{cases} C_i \cap C_j = \emptyset, \ si \ i \neq j, \\ \\ \bigcup_{i=1}^{K} C_i = \Im. \end{cases} \tag{2-1}$$

Claramente este problema tiene múltiples soluciones, pero nuestro interés es que la partición obtenida de los elementos de $\Im$ sea tal que los elementos en cada C_i guarden similitud im-

portante entre ellos y marcada disimilitud con los elementos de los demás C_j, para $j \neq i$.

Nos enfocamos ahora en la agrupación jerárquica como método para la obtención de análisis de conglomerados. La idea básica de la agrupación jerárquica es recoger objetos en agrupaciones mediante la combinación de los objetos más cercanos y a su vez los conglomerados formados se fusionan para formar agrupaciones más grandes, hasta que todos los objetos estén en un sólo conglomerado. Esto da más información sobre la estructura del conjunto de datos y muestra qué conglomerados son similares o distintos. Esto representa una diferencia importante frente a métodos de partición, ya que en ese caso los objetos se colocan en k conglomerados distintos sin contar con indicadores de conexión entre los elementos o entre las agrupaciones resultantes. Los métodos de agrupamiento jerárquico están destinados a reducir la variabilidad total, representar de manera sintética el resultado de las comparaciones entre los objetos y producir una representación gráfica de la información contenida en la tabla de datos.

Se dijo más arriba que hay dos tipos de métodos de agrupamiento jerárquico, los jerárquicos aglomerativos y los jerárquicos divisivos. Un método jerárquico aglomerativo parte con una situación en la que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas están en un único conglomerado, a menudo llamados métodos "abajo hacia arriba". Los algoritmos de agrupamiento jerárquico divisivos, llamados métodos de "arriba hacia abajo", siguen el sentido inverso, parten de conglomerados con un gran número de objetos y en pasos sucesivos se van dividiendo hasta que cada objeto queda en un conglomerado distinto.

2.1.1. Dendrograma

El resultado final de todos los métodos de agrupamiento jerárquico es un dendrograma (llamado también diagrama de árbol jerárquico), donde la solución con k conglomerados se obtiene mediante la fusión de dos de las agrupaciones de la solución con $(k+1)$ conglomerados. El dendrograma puede dibujarse horizontal o verticalmente, dependiendo de la elección del usuario o de la decisión del *software*; ambos tipos dan la misma información.
El dendrograma permite leer la "altura" del criterio de vinculación a la cual los conglomerados se combinan para formar un nuevo conglomerado más grande. Los elementos con mayor similitud entre sí se combinan en alturas bajas, mientras que los elementos que son más diferentes se combinan más arriba en el dendrograma. Por lo tanto, la diferencia en las alturas define cuán cercanos son los elementos.
Una partición de los datos en un número específico de conglomerados se puede obtener por un "corte" del dendrograma a una altura adecuada. Si trazamos una línea horizontal en el dendrograma a una altura determinada, entonces el número k de líneas verticales cortadas por la línea horizontal se identifica como una solución con k agrupaciones, y la intersec-

ción de la línea horizontal y una de esas líneas verticales k representa un conglomerado, y los elementos que se encuentran al final de todas las ramas por debajo de la intersección constituyen los miembros de la agrupación.
En resumen, un algoritmo de agrupamiento sigue algunos pasos básicos como son:

1. Inicializar los conglomerados, esto no es más que asumir que cada objeto constituye un conglomerado: $\{C_1, C_2, ..., C_M\}$.

2. Calcular las distancias entre todos los pares de conglomerados.

3. Buscar los dos conglomerados más cercanos, digamos C_i, C_j, estos se unen y se genera uno sólo, C_{ij}.

4. Se repiten los pasos 2. y 3. del proceso mientras queden pares para comparación.

Finalmente todo el proceso se representa como árboles binarios. Este proceso es detallado en la sección 2.2.

2.1.2. Similitudes, disimilitudes y distancias

Como vimos arriba, la agrupación jerárquica requiere poder establecer indicadores de proximidad o de lejanía de los objetos en estudio. Para ello se hace uso de las denominadas similitudes, disimilitudes y distancias. Las cuales a su vez requieren que los elementos de estudio puedan ser representados en conjuntos o espacios apropiados.
Las representaciones de los individuos de un conjunto finito $\Im$ son, usualmente, de dos tipos:

1. Una representación a lo largo de unos ejes reales que describen las analogías y diferencias entre los individuos (elementos de $\Im$). Los ejes se interpretan como factores o causas de variabilidad y la información obtenida es de tipo espacial.

2. Una representación como un grafo con estructura de árbol (dendrograma), como forma de representar clasificaciones jerárquicas entre los individuos. La información es de tipo agrupativo.

En ambos casos el punto de partida es una matriz de distancias $D = (\delta_{ij})$ de dimensión $M \times M$, siendo M el número de individuos del conjunto $\Im$. El concepto de distancia entre dos objetos o individuos observados permite interpretar geométricamente muchas técnicas clásicas de análisis multivariante, equivalentes a representar estos objetos como puntos de un espacio métrico adecuado.
La herramienta básica para el agrupamiento jerárquico es una medida de la disimilitud o de la proximidad (es decir, distancia) de un elemento con respecto a otro elemento. Luego la definición que se utiliza de distancia en cualquier aplicación dada es a menudo una cuestión de elección cuidadosa en la que resulta de importancia no sólo el tipo de dato sino el conocimiento previo del experto sobre el comportamiento del fenómeno en estudio.

Definición 1. *Una disimilitud o semi-métrica sobre un conjunto $\Im$, es una función* $\delta : \Im \times \Im \to \mathbb{R}$ *tal que:*

· $\delta(x, y) \geq 0$, *para todo* $x, y \in \Im$

· $\delta(x, x) = 0$, *para todo* $x \in \Im$

· $\delta(x, y) = \delta(y, x)$, *para todo* $x, y \in \Im$

Una disimilitud es métrica si satisface

· $\delta(x, y) > 0$, para todo $x, y \in \Im$, distintos

· $\delta(x, y) \leq \delta(x, z) + \delta(z, y)$, para todo $x, y, z \in \Im$

Una ultramétrica es una disimilitud que cumple:

· $\delta(x, y) \leq \max\{\delta(x, z), \delta(z, y)\}$, para todo $x, y, z \in \Im$

Cuando el conjunto $\Im$ es finito, digamos $\Im = \{x_1, x_2, \ldots, x_M\}$, se suele denotar por δ_{ij} a $\delta(x_i, x_j)$.

En general, la palabra distancia se utiliza como sinónimo de métrica. Además, usaremos la letra d para denotar disimilitudes que sean métricas y usaremos δ para disimilitudes que no necesariamente lo sean.
Presentamos en seguida varias medidas de disimilitud, siendo la más conocida la distancia euclidiana. Para ello, sean $\mathbf{x}_i = (x_{i1},\ ..., x_{ip})^\top$ y $\mathbf{x}_j = (x_{j1},\ ..., x_{jp})^\top$ dos puntos de $\Im \subseteq \mathbb{R}^p$, entonces algunas medidas de disimilitud se definen como sigue:

1. **Distancia euclidiana:** Es la más conocida y utilizada. Mide la distancia geométrica entre los puntos, en sentido cartesiano.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left[(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)\right]^{1/2} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right]^{1/2}.$$

2. **Distancia *cityblock***: También conocida como distancia *Manhattan*, es la sumatoria de las distancias entre las componentes de los puntos sobre cada eje.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|.$$

3. **Distancia de Minkowski:** Generaliza las dos anteriores, las cuales se obtienen, respectivamente, haciendo $m = 2$ y $m = 1$. Se puede utilizar para todo valor real $m \geq 1$.

$$d_m(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^m\right]^{1/m}.$$

4. **Disimilitud euclidiana cuadrada:** Aunque no es una métrica tiene dos ventajas importantes. Conduce a los mismos minimizadores que la distancia euclidiana y es suave, en el sentido de poseer derivadas en todo $\mathbb{R}^p$

$$\delta\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left[\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \left(\mathbf{x}_i - \mathbf{x}_j\right)\right] = \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2.$$

5. **Disimilitud coseno:** Esta disimilitud se obtiene como 1 menos el coseno del ángulo formado por los dos vectores en consideración.

$$\delta\left(\mathbf{x}_i, \mathbf{x}_j\right) = 1 - \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} \left(x_{ik}\right)^2 \sum_{k=1}^{p} \left(x_{jk}\right)^2}}.$$

6. **Disimilitud de correlación:** El coeficiente de correlación es una medida de proximidad o similitud entre dos series de datos. Por lo tanto, a partir de él se puede definir una medida de disimilitud, como 1 menos dicho coeficiente:

$$\delta\left(\mathbf{x}_i, \mathbf{x}_j\right) = 1 - \frac{n \sum_{k=1}^{p} x_{ik} x_{jk} - \sum_{k=1}^{p} x_{ik} \sum_{k=1}^{p} x_{jk}}{\sqrt{n \sum_{k=1}^{p} x_{ik}^2 \ - \left(\sum_{k=1}^{p} x_{ik}\right)^2} \sqrt{n \sum_{k=1}^{p} x_{jk}^2 \ - \left(\sum_{k=1}^{p} x_{jk}\right)^2}}.$$

Definición 2. *Una similitud es una función* $s : \Im \times \Im \to \mathbb{R}$ *tal que:*

· $0 \leq s(x, y) \leq s(x, x) = 1$, para todo $x, y \in \Im$.

· $s(x, y) = s(y, x)$, para todo $x, y \in \Im$.

Cuando el conjunto $\Im$ *es finito, digamos* $\Im = \{x_1, x_2, \ldots, x_M\}$, *se suele denotar por* s_{ij} *a* $s(x_i, x_j)$. *Y las condiciones de función de similitud se leen:*

· $0 \leq s_{ij} \leq s_{ii} = 1$, para todo i, j.

· $s_{ij} = s_{ji}$, para todo i, j.

Finalmente, se debe observar que mediante la transformación

$\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$,

se puede definir una disimilitud a partir de una función de similitud dada s_{ij}. Claramente, se trata de una extensión natural de la relación entre norma y producto punto euclídeo. A saber, para $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$\|x - y\|^2 = < x, x > + < y, y > -2 < x, y > .$

2.2. Agrupamientos jerárquicos aglomerativos

Las disimilitudes definidas recién nos permiten obtener matrices de disimilitud entre objetos individuales. No obstante, una vez empecemos a formar conglomerados estaremos en la necesidad de contar con criterios de disimilitud definidos sobre conjuntos de observaciones. Este papel lo juegan los métodos de agrupación. Cada método de agrupación se define por la forma en la que se establece la disimilitud entre dos conglomerados (que pueden ser elementos individuales), para luego unir o "combinar" los más cercanos y formar un nuevo conglomerado, más grande.
Los métodos más utilizados de agrupación jerárquica aglomerativa son: vinculación única (o del vecino más cercano), vinculación completa (o del vecino más lejano), y el de vinculación promedio. El de vinculación única utiliza una métrica de distancia mínima entre conglomerados, la vinculación completa utiliza una métrica de máxima distancia y el de vinculación promedio calcula la distancia promedio entre todos los pares de elementos dentro de los dos grupos diferentes, como un elemento de cada grupo. Hay también una versión ponderada de la vinculación media, donde los pesos reflejan los tamaños dispares (posiblemente) de los grupos en cuestión.
El algoritmo que utilizan los agrupamientos jerárquicos aglomerativos es el siguiente:

1. Entrada: Conjunto de observaciones $\Im = \{x_i, i = 1, 2, 3, ..., M\}$. Se genera la colección inicial de conglomerados $\left\{C_i^{(1)}, i = 1, 2, 3, ..., M\right\}$, donde $C_i^{(1)} = \{x_i\}$. Es decir, cada observación genera un agrupamiento inicial donde cada conglomerado contiene un sólo elemento.

2. Calcular $D^{(1)} = (\delta_{ij})$, matriz de disimilitud de tamaño $M \times M$, entre los M conglomerados, y donde

   $$\delta_{ij} = \delta\left(C_i^{(1)}, C_j^{(1)}\right),\ i, j = 1, 2, 3, ..., M.$$

3. Encontrar la disimilitud más pequeña en $D^{(1)}$, digamos δ_{IJ}. Combinar entonces las agrupaciones I e J para formar un nuevo conglomerado IJ.

4. Calcular las disimilitudes, $\delta_{IJ,K}$, entre el nuevo conglomerado IJ y todos las demás agrupaciones $C_K^{(1)}$ para $K \neq IJ$. Estas disimilitudes dependen del método de vinculación que se utiliza. Para las opciones consideradas arriba, las expresiones para este cálculo son:

   - Vinculación-única (vecino más cercano): $\delta_{IJ,K} = \text{mín}\{\delta_{I,K}, \delta_{J,K}\}$
   - Vinculación-completa (vecino más lejano): $\delta_{IJ,K} = \text{máx}\{\delta_{I,K}, \delta_{J,K}\}$

- Vinculación-promedio (promedio de agrupación): Si N_{IJ} y N_K son el número de elementos en los conglomerados IJ y K, respectivamente, definimos

$$\delta_{IJ,K} = \sum_{\{i:X_i \in C_{IJ}\}} \sum_{\{k:X_k \in C_K\}} \frac{1}{N_{IJ} N_K} \delta_{ik},$$

- Vinculación-*ward*. Aplica sólo para la distancia euclidiana y se define como:

$$\delta_{IJ,K} = \frac{N_I + N_K}{N_I + N_J + N_K} \delta_{I,K} + \frac{N_J + N_K}{N_I + N_J + N_K} \delta_{J,K} - \frac{N_K}{N_I + N_J + N_K} \delta_{I,J},$$

donde N_K es el número de elementos en el conglomerado K.

5. Formar una nueva matriz de tamaño $(M-1)\times(M-1)$, $D^{(2)}$, mediante la eliminación de filas y columnas de I y J, y formando una nueva fila y columna IJ con la disimilitudes calculadas en el paso 4.

6. Repitir los pasos 3, 4 y 5 un total de $M-1$ veces. En el paso i-ésimo, $D^{(i)}$ es una matriz simétrica de tamaño $((M-i+1)\times(M-i+1))$, $i = 1, 2, 3, ..., M$. En el último paso $(i = M)$, $D^{(M)} = 0$ y todos los elementos se combinan para formar un sólo grupo.

7. Salida: Listar los conglomerados que se fusionan en cada paso, el valor (o la altura) de la disimilitud de cada combinación, y en un dendograma, el resumen del procedimiento de la agrupación.

2.2.1. Agrupamiento de Hausdorff

Se introduce en esta sección la distancia de Hausdorff, tal y como es presentada en [4]. Esta distancia se define bien sobre los subconjuntos compactos y no vacíos de un espacio métrico $(\Im, d)$. La estrategia consiste en utilizarla como medida de disimilitud entre conglomerados en el algoritmo de agrupamiento jerárquico aglomerativo presentado arriba, como una alternativa a la vinculación única, la completa y la promedio.

Distancia de Hausdorff

Dado un espacio métrico $(\Im, d)$ con una métrica d, la distancia entre un punto $a \in \Im$ y un subconjunto compacto y no vacío B de $\Im$ está dada por:

$$\delta(a; B) = \inf_{b \in B} d(a, b). \tag{2-2}$$

Este concepto es ilustrado en la Figura **2-1**
Ahora, dados compactos no vacíos $A, B \subseteq \Im$, definimos:

$$\widetilde{\delta}(A, B) = \sup_{a \in A} \delta(a, B) = \sup_{a \in A} \inf_{b \in B} d(a, b). \tag{2-3}$$

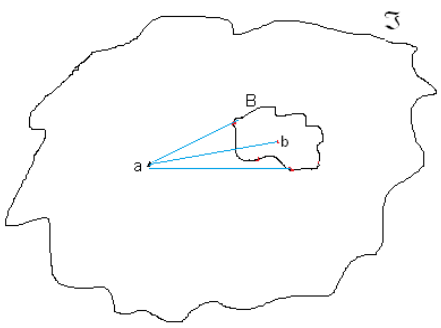


Figura 2-1: Distancia de un punto a un conjunto

Al ser conjuntos compactos, esa expresión mide la más grande entre todas las distancias $\delta(a, B)$, con $a \in A$. Nótese que esta función no es simétrica ya que en general no se tiene $\widetilde{\delta}(A, B) = \widetilde{\delta}(B; A)$, y por tanto no podemos considerarla distancia. No obstante, a partir de ella podemos definir una distancia apropiada, así:

Definición 2. *La distancia de Hausdorff entre dos conjuntos compactos no vacíos* $A, B \subseteq \Im$ *se define como:*

$$d_H(A; B) = \text{máx}\left\{\widetilde{\delta}(A; B), \widetilde{\delta}(B; A)\right\}, \tag{2-4}$$

o equivalentemente,

$$d_H(A; B) = \text{máx}\left\{\sup_{a\in A} \text{ínf}_{b\in B}\, d(a, b), \sup_{b\in B} \text{ínf}_{a\in A}\, d(a, b)\right\}. \tag{2-5}$$

Esta distancia resulta ser claramente simétrica. Nótese además que la distancia de Hausdorff entre A y B es el menor número positivo r tal que cada punto de A está a lo sumo a una distancia r de algún punto de B, y cada punto de B está a lo sumo a una distancia r de algún punto de A. Esta interpretación es ilustrada en las Figuras **2-2** y **2-3**.

En el caso en que $\Im$ sea finito, todos sus subconjuntos no vacíos son compactos y por ello se puede definir bien la distancia de Hausdorff entre cada par de ellos. Ahora, dados dos conjuntos finitos $A = \{x_1, x_2, \ldots, x_N\}$ y $B = \{y_1, y_2, \ldots, y_M\}$, la ecuación (2-5) se lee:

$$d_H(A; B) = \text{máx}\left\{\max_{1\le i\le N} \min_{1\le j\le M} d(x_i, y_j), \max_{1\le j\le M} \min_{1\le i\le N} d(x_i, y_j)\right\}. \tag{2-6}$$

El cómputo se ilustra en la siguiente forma: empezamos con disponer las distancias entre elementos de A y B en una matriz (d_{ij}) de orden $(N \times M)$, en donde $d_{ij} = d(x_i, y_j)$ es la distancia del elemento $i-$ésimo de A al $j-$ésimo de B. En la forma,

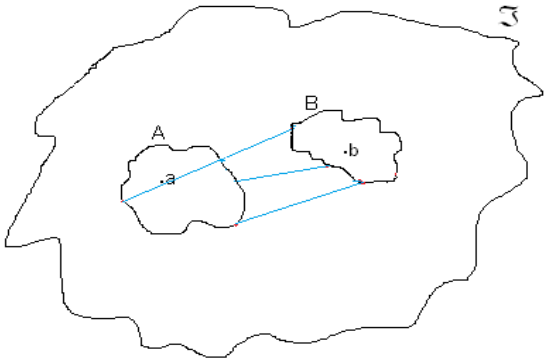


Figura 2-2: Distancia de Hausdorff entre los conjuntos A y B.

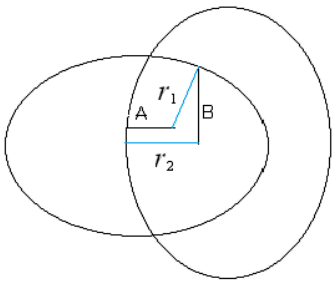


Figura 2-3: Distancia de Hausdorff entre dos segmentos. Aquí $\widetilde{\delta}(A;B) = r_2$ y $\widetilde{\delta}(B;A) = r_1$, así $d_H(A;B) = r_2$

$$D_H = \begin{bmatrix} d_{11} & d_{12} & & ... & d_{1M} \\ d_{21} & d_{22} & & ... & d_{2M} \\ d_{31} & d_{32} & d_{33} & ... & d_{3M} \\ \vdots & & & \ddots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & ... & d_{NM} \end{bmatrix} \begin{matrix} \text{mín} \\ \text{mín} \\ \text{mín} \\ \\ \text{mín} \end{matrix} \tag{2-7}$$

$$\begin{matrix} \text{mín} & \text{mín} & \text{mín} & \text{mín} & \text{máx} \end{matrix} \tag{2-8}$$

Construida la matriz de distancias D_H, se encuentra la distancia mínima sobre cada fila, y, posteriormente, el máximo entre todos estos mínimos, el resultado es $\widetilde{\delta}(A;B)$. Se procede de modo similar sobre las columnas, obteniendo $\widetilde{\delta}(B;A)$, y los dos números finales son comparados. El mayor de ellos es la distancia de Hausdorff entre A y B, $d_H(A;B)$.

A modo de ilustración considere $A = \{-4, -2, 0\}$ y $B = \{2, 4, 6, 8, 10\}$. Entonces,

$$D_H = \begin{bmatrix} 6 & 8 & 10 & 12 & 14 \\ 4 & 6 & 8 & 10 & 12 \\ 2 & 4 & 6 & 8 & 10 \end{bmatrix} \begin{matrix} 6 \\ 4 \\ 2 \end{matrix} \tag{2-9}$$

$$\begin{matrix} 2 & 4 & 6 & 8 & 10 \end{matrix} \tag{2-10}$$

En este caso tenemos $\widetilde{\delta}(A;B) = max\{6, 4, 2\} = 6$, en tanto que $\widetilde{\delta}(B;A) = max\{2, 4, 6, 8, 10\} = 10$ y con ello $d_H(A;B) = 10$. Este ejemplo muestra que, incluso en el caso finito, $\widetilde{\delta}$ no es simétrica.

Ahora, la utilización de la distancia de Hausdorff como método de vinculación en el algoritmo de agrupamiento jerárquico aglomerativo se sigue de modo natural. Simplemente, para calcular las disimilitudes requeridas en los pasos 2. y 4. del algoritmo se utiliza la ecuación 2-5 (o 2-6, en casos finitos) para cada par de conglomerados A y B a comparar. En el primer nivel cada elemento es un conglomerado y la distancia de Hausdorff entre cualquier par de ellos coincide con la distancia entre los puntos respectivos, es decir:

$$d_H(\{x_i\}, \{x_j\}) = d(x_i, x_j) = d_{ij}. \tag{2-11}$$

Cabe anotar que las particiones obtenidas por este método de vinculación de Hausdorff, son intermedias a las obtenidas por los algoritmos más comúnmente usados de vinculación única y vinculación completa.

En efecto, si A y B son dos subconjuntos no vacíos y compactos de $\Im$, los algoritmos de vinculación "única" y "completa" pueden escribirse, respectivamente, en la forma

$$\delta_S(A;B) = \inf_{a\in A, b\in B} \delta(a, b), \qquad \delta_C(A;B) = \sup_{a\in A, b\in B} \delta(a, b). \tag{2-12}$$

Los cuales a su vez, para el trabajo con conjuntos finitos, cuentan con expresiones simplificadas, así:

$$\delta_S(A;B) = \min_{1\leq i\leq N, 1\leq j\leq M} \delta_{ij}, \qquad \delta_C(A;B) = \max_{1\leq i\leq N, 1\leq j\leq M} \delta_{ij}. \tag{2-13}$$

En todo caso, se sigue que:

$\delta_S(A;B) \leq d_H(A,B) \leq \delta_C(A;B)$.

Terminamos esta sección haciendo notar que la vinculación por Hausdorff queda definida por una disimilitud que es métrica sobre los subconjuntos compactos y no vacíos de un espacio métrico $(\Im, d)$. Por lo que el análisis de disimilitud y conglomerados con ella cuentan con una fundamentación más fuerte. Las bondades al implementarse serán discutidas posteriormente.

2.3. Técnica de agrupamiento K-medias

El algoritmo de K-medias es un método muy rápido en la agrupación ya que su idea básica es la reasignación continua de objetos en grupos diferentes de forma que la distancia dentro del grupo se minimiza. K-medias utiliza un algoritmo iterativo para reducir al mínimo la suma de las distancias de un punto a un centroide sobre todos los k grupos deseados [20].

Siguiendo a [36], el algoritmo se resume así:

1. K centroides iniciales (o conglomerados iniciales) son escogidos, aquí K es ingresado por el usuario e indica el número deseado de agrupaciones.

2. Cada punto de los datos es asignado al centroide más cercano y la colección de puntos asignados a un centroide forma un conglomerado.

3. El centroide de cada conglomerado es actualizado, basándose en los puntos asignados a él.

4. Se repiten los pasos 2. y 3. hasta que no haya cambio en los puntos de los conglomerados.

Como en [36], podemos presentar K-medias como sigue: dado un conjunto de datos $\Im = \{x_1, x_2, \ldots, x_M\}$ a ser asignados a K conglomerados, K-medias puede expresarse como la minimización de una función objetivo que depende de la proximidad de los puntos del conjunto $\Im$ a los centroides de los conglomerados, así:

$$\min_{\{m_k\}, 1\leq k\leq K} \sum_{k=1}^{K} \sum_{x\in C_k} \pi_x \delta(x, m_k), \tag{2-14}$$

donde π_x es el peso asignado al punto $x \in \Im$, y un $x \in C_k$ cuando $\delta(x, m_k) = \min_{1\leq j\leq K} \delta(x, m_j)$. Entonces, el algoritmo de arriba propone un proceso iterativo para minimizar (2-14), usando

como forma de actualización del m_k el definirlo como el centroide de los C_k obtenidos en cada paso, así:

$$m_k = \sum_{x \in C_k} \pi_x x / N_k,$$

donde N_k denota el número de elementos en C_k. Esta estrategia se corresponde con una optimización vía gradiente descendiente [36].
Las bondades de K-medias le ameritaron obtener el segundo lugar en un *top 10* construido para algoritmos de minería de datos [37] y lo han posicionado como referente obligatorio para la validación del desempeño de otros métodos de análisis de conglomerados [36].

2.4. Selección del número de conglomerados

Una de las tareas más cruciales en el proceso de análisis de conglomerados es determinar o identificar el número apropiado de conglomerados para una muestra particular de datos. Una de las herramientas que resulta de gran utilidad en el abordaje de este aspecto es el conocido como coeficiente silueta [29]. Su importancia radica en que apunta a la medición de la calidad de los conglomerados, esto mediante la comparación de las disimilitudes de un objeto con el conglomerado al que pertenece contra la disimilitud del objeto a los otros conglomerados. Es decir, contrasta las disimilitudes intra-conglomerado contra las entre-conglomerados. Adicionalmente, al hacerse el análisis sobre cada objeto, permite contar con un indicador de la calidad de la asignación de un objeto específico al conglomerado correspondiente. A continuación se presenta la definición de dicho coeficiente.

2.4.1. Coeficiente silueta

Dada una partición en K conglomerados $\{C_1, C_2, \ldots, C_K\}$, de un conjunto de observaciones $\{x_1, x_2, \ldots, x_M\}$, en la cual asumiremos que $K \geq 2$ y que cada conglomerado tiene más de un elemento, definiremos para cada x_i su coeficiente silueta. Para ello, sea $C_{k(i)}$ el conglomerado al cual x_i fue asignado. Denotemos por $a(i)$ la disimilitud promedio de x_i a los demás elementos de $C_{k(i)}$, es decir:

$$a(i) = \frac{1}{N_{k(i)} - 1} \sum_{x_j \in C_{k(i)}} \delta(x_j, x_i),$$

donde N_k denota nuevamente el número de elementos en C_k.
Definimos ahora $g(i, j)$ como la disimilitud promedio del objeto x_i al conglomerado C_j, esto para $i = 1, 2, \ldots, M$ y los $1 \leq j \leq K$, con $j \neq k(i)$. Es decir,

$$g(i, j) = \frac{1}{N_j} \sum_{x_m \in C_j} \delta(x_m, x_i).$$

Procedemos ahora a definir $b(i)$ como el más pequeño de estos $g(i,j)$. Entonces $b(i)$ puede interpretarse como la distancia media de x_i al conglomerado más cercano a él, usualmente denominado conglomerado vecino. Se lee,

$$b(i) = \min_{j \neq k(i)} g(i,j).$$

Con estos insumos definimos el coeficiente silueta $Sh(i)$ del objeto x_i, como

$$Sh(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}. \tag{2-15}$$

Ahora, si durante el proceso de construcción de los conglomerados el elemento x_i resulta bien asignado al conglomerado respectivo, es de esperarse que el valor de la disimilitud al interior del conglomerado, $a(i)$, sea mucho menor en valor que la disimilitud media con el conglomerado vecino $b(i)$. Esto llevaría a que el denominador en (2-15) sea $b(i)$ y a que el numerador sea positivo y cercano a $b(i)$. Por lo que el valor de $Sh(i)$ ha de ser próximo a $+1$. Recíprocamente, si el valor de $Sh(i)$ es cercano a -1, entonces el numerador es negativo. Con ello, $a(i)$ superaría a $b(i)$ por lo que el denominador debe ser $a(i)$ en este caso. Pero al ser, $Sh(i)$ cercano a -1 llevaría entonces a que $a(i)$ es mucho mayor que $b(i)$. Lo cual a su vez significa que el elemento x_i es más disímil de los otros elementos de su conglomerado que de los elementos del conglomerado vecino, siendo indicador claro de una asignación no apropiada.

Aunque obtener los valores $+1$ o -1 son situaciones límite, puede utilizarse el valor del coeficiente silueta como medida de calidad de la clasificación de los elementos en los distintos conglomerados. Siendo valores positivos indicadores de buena clasificación y valores negativos lo contrario. Esto es particularmente útil gracias a que el valor de $Sh(i)$ siempre está entre -1 y $+1$. Por último, un valor de $Sh(i)$ de cero (0) equivaldría a afirmar que el objeto en cuestión se identifica tanto con el conglomerado al cual está asignado actualmente como a su vecino más próximo o que x_i está a "mitad de camino" entre los dos conglomerados.

Una re-escritura de la ecuación (2-15) que resulta particularmente conveniente es:

$$Sh(i) = \begin{cases} 1 - a(i)/b(i), & \text{si } a(i) < b(i) \\ 0, & \text{si } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{si } a(i) > b(i) \end{cases}. \tag{2-16}$$

Cuando un conglomerado está compuesto de un único elemento x_i se define su coeficiente silueta como $Sh(i) = 0$. Aprovechando la condición de neutralidad descrita arriba para tal valor.

2.4.2. Gráfica del coeficiente silueta

Se introduce en esta sección cómo construir una representación gráfica de los coeficientes silueta. Se empieza por ordenar los valores de los coeficientes de modo descendiente dentro de cada conglomerado. Luego se puede generar un gráfico de barras horizontales, donde la cordenada en el eje X será el valor de cada coeficiente y sobre el eje Y se disponen los distintos conglomerados a diferentes alturas (Ver Figura **2-4**). El orden en este caso está marcado por la etiqueta asignada a los conglomerados por el algoritmo y por el valor del coeficiente silueta, al graficarse de modo descendente.

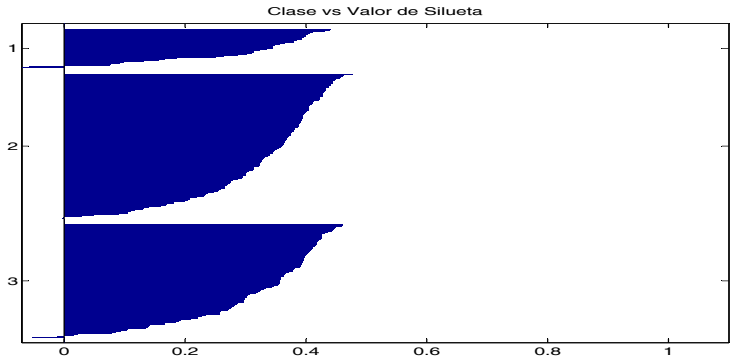


Figura 2-4: Ilustración de la gráfica del coeficiente silueta, para tres conglomerados en este caso a partir de una muestra de 365 observaciones extraida de la base de datos de este trabajo.

En principio, la importancia de la silueta radica en ser una herramienta de interpretación y validación de los resultados de un proceso de análisis por conglomerados [29]. No obstante, puede utilizarse para generar un procedimiento de escogencia del número de conglomerados conveniente para un conjunto particular de datos. La estrategia consiste en llevar a cabo el análisis para distintos números de conglomerados, en un rango establecido por el usuario y utilizando para la obtención de los conglomerados un algoritmo de preferencia. Luego para cada número de conglomerados en consideración, digamos k, se calcula el valor promedio de $Ş h(i)$ sobre todos los elementos x_i, generando así un valor medio $\overline{Ş h(k)}$. Por último, como valor óptimo para el número de elementos se selecciona aquel k para el cual $\overline{Ş h(k)}$ alcanza su máximo (Figura **2-5**).

2.5. Estado del arte

La utilización de técnicas de análisis de conglomerados para la gestión de información meteorológica es ampliamente referenciada en la literatura, manteniendo vigencia como área de

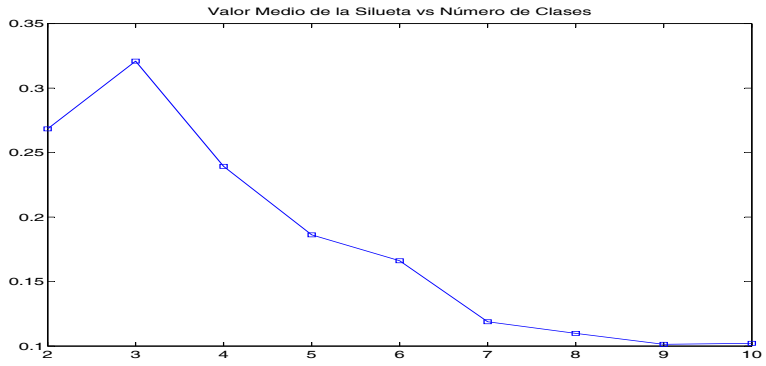


Figura 2-5: Ilustración de la selección del número de conglomerados a partir de promedios de coeficientes silueta, tres (3) resulta ser el valor escogido.

estudio. La mayor utilización se da en la identificación de tipos de climas que se presentan en zonas específicas del planeta, usando análisis por conglomerados sobre información hidrometeorológica. Algunos trabajos de este tipo son: [6] donde se utiliza un algoritmo jerárquico que usa Ward como método de vinculación de los conglomerados, para el estudio de datos del invierno boreal entre 1946 y 1985. Encontrando 3 agrupaciones que pudieron asociarse con tres regiones específicas del hemisferio norte.
En [22] se utiliza un análisis de agrupamiento particional sobre datos de 44 inviernos en latitud media en sectores del Pacífico y del Atlántico para indentificar patrones recurrentes del clima. Los resultados son comparados con identificaciones por otro método, que busca en cambio patrones cuasi-estacionarios, y con trabajos previos, pudiendo encontrar similitudes e información complementaria.
A su vez, [19] y [32] son trabajos denominados por sus autores de "re-análisis" y constituyen un esfuerzo de muchas instituciones climatológicas para estudiar datos del clima a escala global, entre 1957 y 1996 en el primero y entre 1957 y 2002 en el segundo. Entre los problemas considerados en este esfuerzo están problemas metodológicos para la asimilación de datos de diversas fuentes y la predicción numérica del clima a gran escala.
En [30] se usa un análisis de conglomerados por K-medias para estudiar regímenes climáticos que relacionen la variabilidad interanual de la lluvia invernal en Portugal con datos de presión a nivel del mar y flujos atmosféricos de gran escala. En tanto que en [24] se utiliza K-medias para resumir la variabilidad atmosférica sobre regiones de Senegal, Sahel occidental y Atlántico norte tropical, en los veranos boreales entre 1961 y 1998. Se hallan 5 tipos de clima que logran relacionarse con la evolución de los monzones.
Otros trabajos que siguen la previamente mencionada tarea de "re-análisis"son [5], [17] y [26]. El primero usa los regímenes climáticos para estudiar variables en la superficie del océano contra circulaciones atmosféricas de gran escala. El segundo usa componentes principales y K-medias sobre la base de datos en [19] para obtener 12 tipos de clima para Nueva Zelanda.

El tercero usa K-medias sobre información de circulación atmosférica entre 1979 y 2009 para identificar características fundamentales del clima en modo anular sureño, también conocido como oscilación antártica.
Una revisión de técnicas para la clasificación de patrones de circulación en climatología sinóptica es hecha en [14]. Se considera el papel del análisis de componentes principales y del análisis por conglomerados en la tarea, y se plantea como objetivos una revisión histórica del clima, la identificación de cambios recientes y el estudio de modelos del clima.
Un par de trabajos adicionales que cabe mencionar son [34] y [18]. El primero modela y analiza datos de precipitación sobre 20 estaciones en Grecia. Para ello, deben agruparlas en 7 conglomerados. Luego deben incorporar un índice de la oscilación del Atlántico Norte para poder contar con un modelo más realista y flexible. En el segundo se propone un algoritmo de agrupación jerárquica y orientada a objetos, se basa en medir la disimilitud de la predicción de precipitación con un índice de amenaza sobre los objetos. Obteniendo conglomerados mejor ajustados a la evaluación subjetiva de expertos y resulta además mejorar la sensibilidad a desplazamientos espaciales y cantidad de precipitación. Para ello, utilizan datos del Centro de Análisis y Predicción de Tormentas de la NOOA.
El alto costo social y económico de las emergencias y daños ocasionados por fenómenos del clima en Caldas motivó a la Universidad Nacional de Colombia Sede Manizales, a través de su Instituto de Estudios Ambientales -IDEA-, y a CORPOCALDAS a adelantar una serie de convenios y proyectos orientados a implementar un sistema de gestión integral del riesgo ambiental para la región. Dentro de sus principales objetivos se busca aunar esfuerzos para "desarrollar proyectos de procesamiento de la información, investigación, monitoreo, capacitación y transferencia tecnológica en temas relacionados con la gestión del riesgo, el medio ambiente y los recursos naturales renovables del Departamento de Caldas". Esfuerzos similares han sido emprendidos por el IDEA en conjunto con la Unidad de Gestión del Riesgo de la Alcaldía de Manizales (anteriormente llamada Oficina Municipal de Prevención y Atención de Desastres, OMPAD), enfocándose en la instrumentación de estaciones meteorológicas estratégicamente ubicadas en la ciudad de Manizales, [33], [21].
El montaje de estas estaciones de monitoreo ambiental ha sido adelantado también por otros actores clave de la región como la UDEGER (Unidad Departamental de Gestión del Riesgo), CHEC (Central Hidroeléctrica de Caldas) y EMAS (Empresa Metropolitana de Aseo de Manizales). Es así como se ha configurado un conjunto de redes de monitoreo ambiental para el Departamento de Caldas que hoy cuenta con más de 100 estaciones de diverso tipo. Esta información ha sido registrada a lo largo de varios años, lo que la hace idónea para realizar estudios del clima local. Un grupo importante de estas estaciones permiten ver en línea el estado de todas las variables climáticas, y parte de este cúmulo de información es catalogado y almacenado por el IDEA. En la Tabla **2-1** se muestra un resumen de las estaciones actualmente dispuestas en el Departamento, por tipo e institución responsable. Los tipos catalogados son:

- M: Meteorológica

Institución	M	HM	ALA	ALE	R	C	TOTAL
UGR	10					1	11
UNAL	1	1					2
EMAS	1						1
CORPOCALDAS[a]	8	17			3	2	30
CORPOCALDAS[b]		13					13
CORPOCALDAS[c]	6	6	6	1		1	20
CHEC[d]	1						1
CHEC[e]		8				1	9
UDEGER		4	7	1	1	1	14
TOTAL ACTUAL	27	49	13	2	4	6	101

Tabla 2-1: Relación de estaciones de monitoreo meterológico activas en el Departamento Caldas indicando el tipo de estación y la entidad propietaria de la misma. Fuente: IDEA.

- HM: Hidro-meteorológica
- ALA: Alarma sonora
- ALE: Alerta
- R: Repetidora
- C: Central

Fruto de este esfuerzo, adicional a esta infraestructura instalada, El IDEA a través de sus Grupos de Trabajo Académico en Ingeniería Hidráulica y Ambiental y en Gestión del Riesgo y más recientemente con el Grupo GAIA de Informática y Computación ha desarrollado numerosos proyectos para consolidar y lograr un adecuado manejo de las redes mencionadas y el aprovechamiento de la información allí generada. Habiéndose logrado avances importantes en tres ámbitos:

1. Implementación de espacios físicos para la ubicación de la central de recepción de datos meteorológicos, que cuenta con los requerimientos de ventilación y amplitud adecuada para equipos de funcionamiento 24/7.

2. Desarrollo de software propio para la adquisión y control de los datos de las estaciones meteorológicas, lo mismo que el desarrollo de una bodega de datos para almacenamiento, consulta, análisis y visualización base de la información capturada de las estaciones [9].

3. Disposición de información ambiental y su análisis a la comunidad, a través de páginas web como:
    - http://idea.manizales.unal.edu.co/, que además de ser la web principal del IDEA permite visualizar la información de las estaciones en tiempo real.
    - http://www.gestiondelriesgomanizales.com/, donde se publican informes y documentación sobre el tema.
    - http://cdiac.manizales.unal.edu.co/, que permite acceso a la bodega de datos construida para almacenamiento y generación de indicadores de línea base ambiental para el Departamento de Caldas.

Mención especial requiere una serie de trabajos realizados por investigadores del IDEA y sus colaboradores para la gestión de la información obtenida de la red meteorológica. Particularmente interesantes resultan trabajos como [7] donde se busca identificar zonas de la ciudad con variabilidad homogénea en la intensidad de la lluvia, usando datos de precipitación en 2008 en 11 estaciones distribuidas en la ciudad. A su vez en [33] se identifica una fuerte correlación entre un indicador de lluvia acumulada en regiones de la ciudad con la ocurrencia de deslizamientos, posibilitando la habilitación de un sistema de alertas. Finalmente, mencionamos [27] donde se investiga una distribución espacio-temporal de la lluvia en Manizales, usando información de nueve (9) de las estaciones de la red entre 2006 y 2014.

Se ha revisado en este capítulo los conceptos base del análisis de conglomerados, el estado del arte en su uso para datos meteorológicos y el estado del estudio de datos climáticos de Manizales. Cabe resaltar que la estructuración de un marco matemático apropiado para expresar las ideas y métodos estudiados en las secciones 2.1, 2.2 y 2.3 es un aporte de este trabajo, pues requiere la conciliación de varias teorías generadas en diversos campos de la ciencia en que se han generado los métodos de conglomerados aquí estudiados y las formulaciones que para ello presentan diversos autores.

3 Marco experimental

El propósito de este capítulo es hacer una descripción de la base de datos metereológicos utilizada en el estudio. Se indica su composición, estructura y delimitación. Luego se presenta la estructura propuesta para el análisis de tendencia central de las variables de interés, para finalizar con la descripción del estudio de agrupamiento realizado de los datos.

3.1. Base de datos

Como ya se indicó en la sección 2.5, el Instituto de Estudios Ambientales (IDEA) de la Universidad Nacional de Colombia - Sede Manizales ha desarrollado una serie de proyectos con distintos actores del departamento de Caldas para configurar una red de monitoreo climático estratégicamente distribuida en la región. Del total de las 101 estaciones indicadas arriba, hay 37 de ellas que cuentan con capacidad de telemetría al IDEA; posibilitando el almacenamiento y manejo de dichos datos en la estación central del instituto. En todo caso debe tenerse en cuenta que la cantidad de registros disponibles para cada estación es distinto, dado que su entrada en funcionamiento no fue simultánea y la ocurrencia de daños e interrupciones es distinta en cada caso [33], [21].
Cada estación establece mediciones de las variables climáticas: *temperatura, precipitación, humedad relativa, presión barométrica, radiación solar, evapotranspiración, velocidad del viento* y *dirección del viento*. Las estaciones envían una medición para cada variable cada 5 minutos, las 24 horas del día durante los 365 días del año. Esto proporciona 288 observaciones diarias, y consecuentemente un total de mediciones por año de $288 * 365 = 105120$ por variable por estación. Toda esta información es almacenada y catalogada por el IDEA; sin embargo su gestión, análisis y estudio requiere aún mucho trabajo para optimizar su aprovechamiento.
Para los propósitos de este trabajo, cada variable meteorológica medida por las estaciones es considerada de modo independiente. Adicionalmente, como unidad de análisis se utilizan los días. Es decir, se considera cada día una observación en la que se evalúan 288 características. Contando entonces con 365 observaciones de 288 características, esto para cada variable meteorológica en consideración por cada estación de medición y por cada año.
Para el presente estudio nos concentramos en las estaciones de medición denominadas **Posgrados, Emas, Enea e Ingeominas** y más concretamente en los datos por ellas registrados en los años $2009, 2010$ y 2011. Las principales razones para la escogencia de tales estaciones son su ubicación geográfica en distintas zonas de la ciudad y el que cuentan con buenas

condiciones de mantenimiento, lo que redunda en menores interrupciones en su operación y medidas más fiables que las otras estaciones.
Adicionalmente, limitamos el estudio a las variables temperatura, radiación solar, humedad relativa y precipitación. Así, la información base de este trabajo, por cada una de las cuatro (4) estaciones consideradas, consta de cuatro (4) matrices de datos, una por cada variable, que almacenan hasta 1095 observaciones, una por cada día, constituidas por 288 características cada una.
Posteriormente se adelantó una verificación de la integridad de los datos, obteniendo que los datos de los días 63 y 64 del año 2010 en la estación Posgrados eran inconsistentes debido muy probablemente a una falla en la estación. Para corregir esta situación, los datos del día 63 fueron remplazados por el valor promedio aritmético de los días 61 y 62, en tanto que los valores del día 64 por el promedio aritmético de los días 65 y 66.
En la estación Emas se encontró que faltaba todo el mes de septiembre del año 2009. En tanto en la estación Enea faltaron los días entre el 07 de enero y el 16 de febrero de 2009. Para la estación Ingeominas la información estaba completa en cuanto a cantidad, no obstante los algoritmos detectaron fallas para la variable humedad relativa en los días $134, 135, 148, 149, 150$ y 151 del año 2009, los días $63, 64, 129, 130$ y 131 de 2010, lo mismo que los días $24, 25$ y 26 de 2011. Por ello fueron removidos del análisis para cada variable en que se detectaba la falla.
Mención especial requieren las variables radiación y precipitación. Para la radiación se trabaja con la información de la radiación entre las 6:00am y las 5:55pm, esto teniendo en cuenta que durante la noche la radiación marca 0, así cada día tendrá 144 mediciones. Para la precipitación, a partir de los datos iniciales, se generó una nueva variable que se denomina *precipitación acumulada*. La cual, en cada instante, se obtiene como la lluvia acumulada en los últimos 15 días. Es decir, cada valor para la nueva variable es la sumatoria de la cantidad precipitación registrada en las últimas $15 * 288 = 4320$ mediciones, esto recordando que se tiene una medición cada 5 minutos. Se tuvo presente aquí los datos de los días faltantes. Adicionalmente, como para los primeros 15 días de 2009 se requeriría información de 2008 para encontrar la precipitación acumulada, se asume en este caso que no hubo precipitación en los últimos quince días de 2008. Finalmente, cabe anotar que el trabajo con precipitación acumulada ha de tener efectos de filtrado y homogenización, pero hay documentación importante sobre la relación entre la cantidad de lluvia antecedente y la ocurrencia de deslizamientos [23], [25], [2]. Lo que hace mucho más interesante trabajar con dicha variable.

3.1.1. Instrumentos de medición

A continuación se presentan definiciones de las variables metereológicas consideradas en este trabajo, los instrumentos de medición que se utilizan para determinar sus valores y su respectivos rangos. Para ello se toma como base la guía para métodos e instrumentos de observación meteorológica de la organización meteorológica mundial [35], complementada

con referencias como [21], [28] y [1].

- **Temperatura:** Según la Organización Mundial de Metereología (WMO), en 1992, se definió la temperatura como la magnitud física que caracteriza el movimiento aleatorio medio de las moléculas en un cuerpo físico. La temperatura se caracteriza por el comportamiento donde dos cuerpos en contacto térmico tienden a una temperatura igual. Por lo tanto, la temperatura representa el estado termodinámico de un cuerpo, y su valor se determina por la dirección del flujo neto de calor entre esos dos cuerpos.

  Sin embargo, la definición de la temperatura como una magnitud física en relación con el "estado de un cuerpo", es difícil. Una solución se encuentra definiendo una escala de temperatura aprobada internacionalmente sobre la base de puntos de congelación universal y puntos triples. La escala actual es la Escala Internacional de Temperatura de 1990 (ITS-90). Para el rango meteorológico (-80 a $+60°C$) esta escala se basa en una relación lineal con la resistencia eléctrica de platino y el punto triple del agua, definido como $273,16$ kelvin [35].

  Para fines meteorológicos, las temperaturas se miden para una variedad de medios. La variable más comunmente medida es la temperatura del aire (a varias alturas). Otras variables son la temperatura del suelo y del agua de mar. En [35] se define la temperatura del aire como "la temperatura indicada por un termómetro expuesto al aire en un lugar protegido de la radiación solar directa". Aunque esta definición no se puede utilizar como la definición de la propia cantidad termodinámica, es adecuado para la mayoría de aplicaciones.

  La temperatura termodinámica (T), en unidades de grados Kelvin (K), (también definidas como "temperatura Kelvin"), se considera la temperatura básica. El kelvin es la fracción $1/273,16$ de la temperatura termodinámica del punto triple del agua. La temperatura (t), en grados Celsius (o "Celsius de temperatura") se define por la ecuación $t/°C = T/K - 273{,}16$.

- **Precipitación:** La precipitación se define como los productos líquidos o sólidos de la condensación del vapor de agua que cae de las nubes o el deposito de aire en el suelo. Incluye las precipitaciones de lluvia, granizo, nieve, rocío, la escarcha y la niebla. La cantidad total de precipitación que llega a la tierra en un período determinado se expresa en términos de la profundidad vertical de agua (o agua equivalente en el caso de formas sólidas) a la cual se cubriría una proyección horizontal de la superficie de la Tierra. Las nevadas se expresan también por la profundidad de la nieve fresca y recién caída sobre una superficie plana horizontal. La unidad de precipitación es profundidad lineal, generalmente se da en milímetros (volumen / área), o kg m^2 (masa / área) de precipitación líquida. Las cantidades diarias de precipitación deben ser leídas al (0,2) mm más cercano y, si es posible, a $(0,1)$ mm; cantidades semanales o mensuales deben

ser leídas al (1) mm más próximo (al menos). Las mediciones diarias de precipitación deben ser tomadas a horas fijas comunes a toda la red o redes de interés [35].

Desde el punto de vista de la ingeniería hidrológica, la precipitación es la fuente primaria del agua de la superficie terrestre y sus mediciones forman el punto de partida de la mayor parte de los estudios concernientes al uso y control del agua. Para su medición se utiliza un aparato llamado pluviómetro de cazoletas basculantes y se hace relacionando la capacidad de la cazoletas con el número de oscilaciones por la lluvia caída [21].

- **Radiación solar:** Esta variable se define como el proceso físico por medio del cual se transmite energía en formas de ondas electromagnéticas [21].

  La radiación puede clasificarse en dos grupos de acuerdo con su origen, solar o terrestre. La "radiación"puede implicar un proceso o se puede aplicar a varias cantidades. Por ejemplo, "radiación solar"podría significar la energía solar, la exposición solar o la irradiación solar.

  La energía solar es la energía electromagnética emitida por el sol. La radiación solar incidente sobre la parte superior de la atmósfera terrestre se llama radiación solar extraterrestre; 97 % de ella se limita al rango espectral entre 290 y 3,000 nanómetros, y se conoce como radiación solar (o a veces, radiación de onda corta). Una parte de la radiación solar extraterrestre penetra a través de la atmósfera a la superficie de la Tierra, mientras que parte de ella se dispersa y/o es absorbida por las moléculas de gas, partículas de aerosol, gotas de las nubes y los cristales de las nubes en la atmósfera [35]. Su medición se hace a través de sensores de radiación solar basados en un fotodiodo de silicio que se adapta adecuadamente al espectro solar. Este instrumento detecta longitudes de onda de 300 a 1100 nanómetros y su rango de medición es de cero (0) a (1800) vatios por metros cuadrados (W/m^2) [21].

- **Humedad relativa:** La medición de la humedad atmosférica es un requisito importante en la mayoría de ámbitos de la actividad meteorológica. La medición de la humedad debe estar cerca de la superficie de la Tierra. Hay muchos métodos diferentes en uso, y existe una amplia literatura sobre el tema. Definiciones simples de las cantidades utilizadas con mayor frecuencia en las mediciones de humedad son las siguientes, [35]:

  - Relación de mezcla r: La relación que se da entre la masa de vapor de agua y la masa de aire seco.
  - Humedad específica q: La relación entre la masa de vapor de agua y la masa de aire húmedo.
  - Temperatura de punto de rocío T_d: La temperatura a la que el aire húmedo saturado con respecto al agua a una presión dada tiene una relación de mezcla de saturación igual a la relación de mezcla dada.

- Humedad relativa U: La relación en porcentaje de la presión de vapor observada a la presión de vapor de saturación con respecto al agua, a la misma temperatura y presión.
- Presión de vapor $e´$: La presión parcial de vapor de agua en aire.

Una definición para humedad relativa es la relación que existe entre la presión parcial que ejerce el vapor contenido en el aire a la temperatura ambiente y la que ejercería si el aire estuviera saturado en esa misma temperatura. Esta variable es medida por un sensor que registra fluctuaciones. Su rango es de 0 a 100 % [21].

- **Presión barométrica:** Para definir la presión barométrica, se debe presentar algunas definiciones [3].

  Presión atmosférica es la presión que ejerce la atmósfera que rodea la Tierra sobre todos lo objetos que se hallan en contacto con ella. La presión atmosférica cambia con la altitud, a mayor altitud menor presión atmosférica, un aumento en la altitud de $1000m$ representa una disminución de presión atmosférica de aproximadamente $100hPa$.

  Presión atmosférica normalizada: es la presión ejercida por la atmósfera bajo condiciones normalizadas, igual a 760 mmHg (1013, 25 hPa). La cual idealmente se presenta a una altitud de (0) m.s.n.m. (metros sobre el nivel medio del mar), temperatura ambiente de 20^oC, humedad relativa del 65 % y densidad del aire de $1, 2kg/m^3$.

  Altitud: Es la distancia vertical entre un punto situado sobre la superficie terrestre o la atmósfera y el nivel medio del mar.

  La presion barométrica, entonces se define como la presión atmosférica local más una corrección por la altitud geopotencial local. Ésta oscila alrededor de la presión atmosférica normalizada ($760mmHg$).

  Existen varios instrumentos para medir la presión barométrica, dentro de los cuales se destacan: barómetro de Fortin basado en el principio de Torricelli y los barómetros metálicos como el holostérico y el aneroide.

- **Evapotranspiracón (EVP):**

  Para definir esta variable, cabe mencionar o definir por separado lo que es la evaporación y la transpiración [35].

  La evaporación se define como la cantidad de agua evaporada de una superficie abierta de agua o del suelo.
  La transpiración es el proceso por el cual agua de la vegetación se transfiere a la atmósfera en forma de vapor.
  Evapotranspiración real (o efectiva) es la cantidad de vapor de agua evaporada desde el suelo y las plantas cuando el suelo está en su contenido de humedad natural.

La evapotranspiración potencial es la cantidad máxima de agua capaz de ser evaporada en un clima determinado a partir de una extensión continua de la vegetación que cubra todo el suelo y bien cubierto de agua. Incluye la evaporación del suelo y la transpiración de la vegetación de una región específica en un intervalo de tiempo específico, expresada como profundidad del agua, usualmente en milímetros.

- **Velocidad y Dirección del viento:** El viento se conoce como el movimiento del aire que está presente en la atmósfera, especialmente en la troposfera, producido por causas naturales. El viento es producto de: El movimiento de rotación y de traslación terrestres que dan origen a diferencias considerables en la radiación solar o el desigual calentamiento del aire, que produce las diferencias de presión. Los vientos pueden clasificarse en cuatro clases principales: dominantes, estacionales, locales y, por último, ciclónicos y anticiclónicos [10].

  La velocidad del viento se define como una cantidad vectorial tridimensional con fluctuaciones aleatorias a pequeña escala en el espacio y el tiempo superpuestos de un flujo organizado a gran escala. No obstante, puede considerarse principalmente como una cantidad vectorial bidimensional especificada por dos números representando dirección y velocidad. Para medirla se utiliza un anenómetro de copas, el cual registra velocidades del viento entre cero (0) y (76) metros por segundo (m/seg).

  La dirección del viento depende de la distribución de las presiones, ya que tiende a soplar desde la región de altas presiones hacia las de presiones más bajas. Se llama dirección del viento al punto del horizonte de donde viene o sopla. Para distinguir uno de otro se les aplica el nombre de los principales rumbos de la brújula, según la conocida rosa de los vientos. El instrumento de medición para la variable dirección del viento se conoce con el nombre de veleta, ésta registra la dirección del viento en un rango de cero (0) a 360 grados (°) o también en la conocida *rosa de los vientos.*

3.2. Estudio de tendencia central, condicionamiento y correlación

Para fines de una mejor visualización del comportamiento, tendencias y posibles correlaciones de las variables de interés, se adelanta un estudio de tendencia central y dispersión. A tal fin, dada una variable (digamos X), a cada una de las mediciones diarias (X_i) de 288 componentes se les calcula los siguientes indicadores: *media, media±desviación estándar, mínimo, máximo, cuartil 01, mediana, cuartil 03*. Estos indicadores han de generar 8 series de tiempo de hasta 1095 mediciones (un valor diario), que son estudiadas para cada variable por cada estación. Luego, para cada estación, se trabaja con los datos de media de cada variable y se le aplica la prueba *Kolmogorov-Smirnov* para establecer, con un 95 % de confianza, si dicha serie

proviene de una distribución *Normal, Logística o Weibull*.
Se incluye también una exploración de correlaciones, se toma el grupo de variables y se revisa el comportamiento de la serie de tiempo de la media de una contra la otra, incluyendo histogramas de las medias de cada variable. Esto ha de permitir identificar posibles relaciones directas o inversas de las variables, que aunque no son objeto de este trabajo pueden ser indicativos de generación de patrones similares.

3.3. Estudios de agrupamiento

Para el estudio de agrupamiento de cada una de las variables se realiza una comparación de distintos resultados obtenidos con las técnicas consideradas en la sección 2.1. Concretamente, se implementa el agrupamiento jerárquico aglomerativo para distintas escogencias de disimilitud y diversos métodos de vinculación o enlace de los conglomerados, con ello se establecen los conglomerados obtenidos. A modo de validación, los resultados son contrastados con los obtenidos por el método de partición K-medias y se explora la relación entre las agrupaciones obtenidas y la magnitud de los datos.

En esta tarea resulta de gran utilidad el *Statistics toolbox* de MATLAB®, el cual cuenta ya con implementaciones de las disimilitudes definidas en 2.1.2, y otras más, usando opciones apropiadas del comando `clusterdata`. También cuenta no sólo con implementaciones de los métodos de vinculación simple, completa y promedio (con las opciones `single, complete, average, ward` del comando `linkage`), sino que también permite usar los métodos `weighted, median` y `centroid` [31]. La implementación del agrupamiento de Hausdorff tal y como es propuesto en [4], fue hecha por los autores para este estudio.

La escogencia del número de grupos para el caso jerárquico se hizo usando como insumos el dendrograma obtenido y el valor sugerido siguiendo la estrategia basada en el coeficiente silueta descrito en la sección 2.4, ([20], [29]). Se aprovecha para ello la implementación MATLAB® del comando `silhouette`. Para el caso de K-medias se usa el criterio silueta y la información obtenida en el estudio jerárquico.

Se ha revisado en este capítulo la estructura de la base de datos utilizada y los algoritmos implementados para adelantar los estudios requeridos tanto para el estudio de tendencia central como para los estudios de agrupamiento.

4 Tendencia central: Resultados

En este capítulo se presentan los resultados de los estudios de tendencia central adelantados sobre las variables meteorológicas de interés para este trabajo. Como se indicó en la sección 3.2, nos concentramos en revisar el comportamiento de indicadores de tendencia central y dispersión generados para cada variable. Se genera un valor de los indicadores por cada observación (día), por cada estación, por cada variable.

En primer lugar se considera todo el grupo de variables **Temperatura - Radiación - Humedad - Precipitación** en la estación posgrados y se grafica el histograma del valor medio diario de cada una y las medias de unas contra otras, para el período considerado $2009 - 2011$. Esto como herramienta de visualización para identificar normalidad o no en la distribución de las medias y eventuales proporcionalidades de una variable contra otra. Los resultados obtenidos pueden apreciarse en la Figura **4-1**. Los coeficientes de correlación entre estas medias se muestran en la Tabla **4-1**.

En seguida, se procede a establecer los valores paramétricos que permiten ajustar la media diaria de cada variable a distribuciones normal, logística y Weibull. Luego se utiliza la prueba *Kolmogorov-Smirnoff* para determinar la bondad de dicho ajuste. Este ejercicio es hecho sobre todo el período $2009 - 2011$ e independientemente para cada una de las estaciones, Posgrados (Posg), Emas, Enea e Ingeominas (Ing). Los resultados son mostrados en las tablas **4-2**, **4-3**, **4-4** y **4-5**.

En el caso de la distribución normal la pareja de parámetros (μ, σ) reportados en las tablas denotan media y varianza, para logística denotan media y escala y para Weibull son escala y forma. Se reporta también en las tablas el $p - valor$ obtenido por la prueba Kolmogorov-Smirnoff, valores inferiores a 0,05 llevan a que se rechace la hipótesis nula de que los datos estudiados (media diaria de la variable) provienen de la distribución considerada, a un nivel de significancia de 5 %. Los mejores ajustes obtenidos son ilustrados en las Figuras **4-2**, **4-4**, **4-6** y **4-8**.

Finalmente, se generan dos histogramas adicionales por variable, que muestran el comportamiento de otros estadísticos de interés. El primero basado en indicadores diarios de media y varianza, el segundo en mediana y cuartiles. Los resultados para las variables **temperatura, radiación, humedad** y **precipitación** son mostrados en las Figuras **4-3**, **4-5**, **4-7** y **4-9**, respectivamente.

R	Temperatura	Radiación	Humedad	Precipitación
Temperatura	1,0000	0,7260	−0,8586	−0,3194
Radiación	0,7260	1,0000	−0,7738	−0,1136
Humedad	−0,8586	−0,7738	1,0000	0,3210
Precipitación	0,3194	−0,1136	0,3210	1,0000

Tabla 4-1: Coeficientes de correlación entre las medias diarias de las variables de estudio, en la estación Posgrados en el período 2009 − 2011.

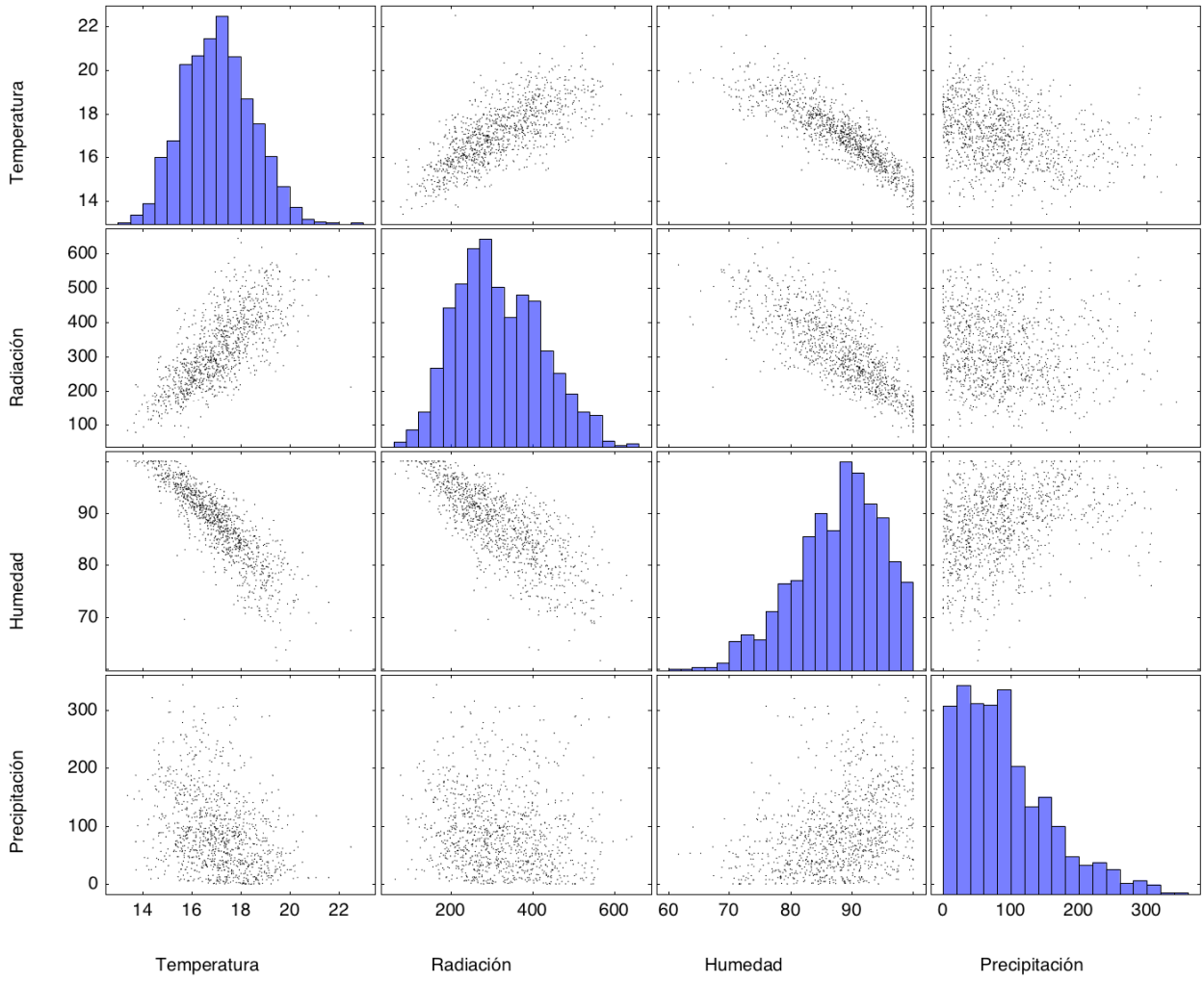


Figura 4-1: Correlación de Variables: Temperatura vs Radiación vs Humedad vs Precipitación

		Normal		Logistica		Weibull	
Posg	p	$0,607$		$0,2541$		$2,81E-4$	
	(μ,σ)	$17,07$	$1,39$	$17,05$	$0,80$	$17,72$	$12,57$
Emas	p	$0,2421$		$0,1661$		$1,41E-5$	
	(μ,σ)	$17,02$	$1,13$	$17,00$	$0,65$	$17,56$	$15,51$
Enea	p	$0,2005$		$0,2224$		$1,01E-4$	
	(μ,σ)	$16,10$	$1,04$	$16,07$	$0,59$	$16,60$	$15,44$
Ing	p	$0,2105$		$0,1617$		$4,91E-4$	
	(μ,σ)	$17,20$	$1,61$	$17,16$	$0,93$	$17,94$	$11,12$

Tabla 4-2: Prueba de bondad de ajuste de la variable temperatura media diaria entre $2009-2011$ a distintas distribuciones en las 4 estaciones de estudio.

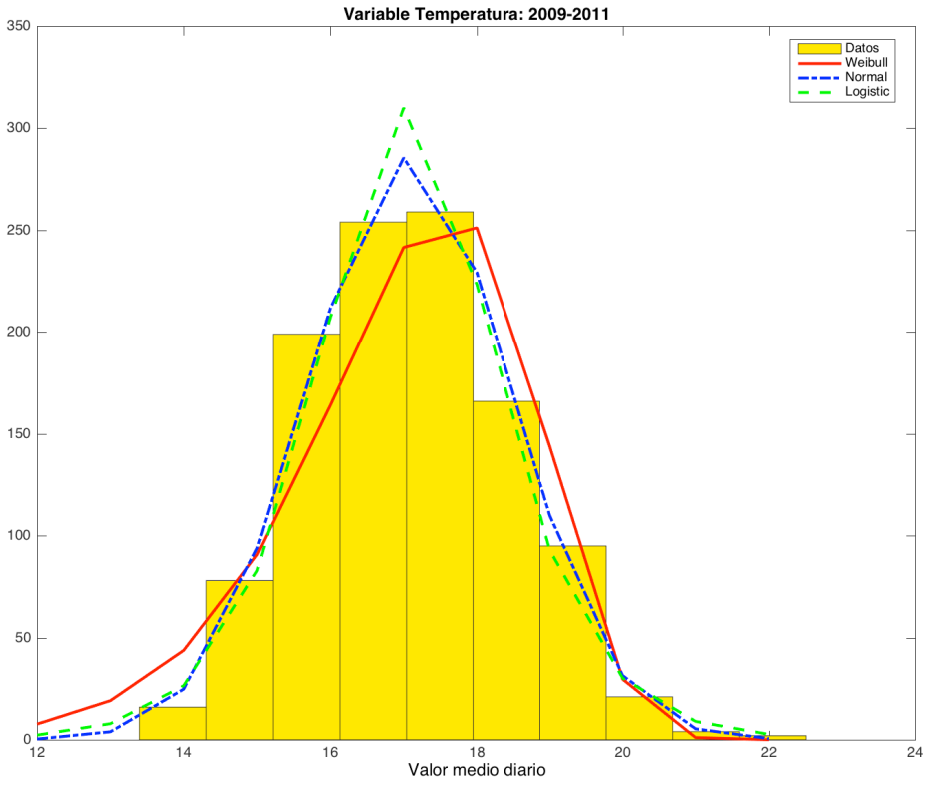
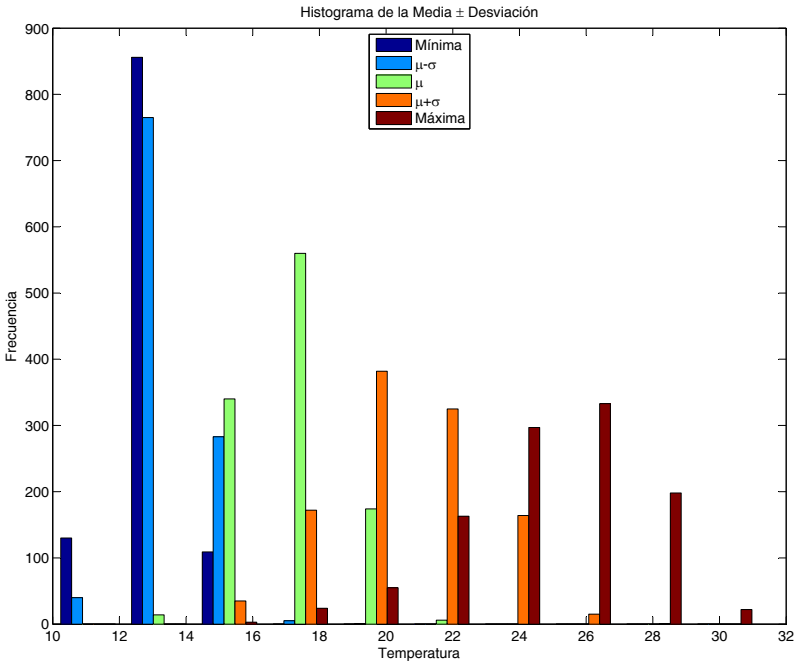


Figura 4-2: Prueba de ajuste a distribución para temperatura media diaria en la estación Posgrados.

(a)



(b)

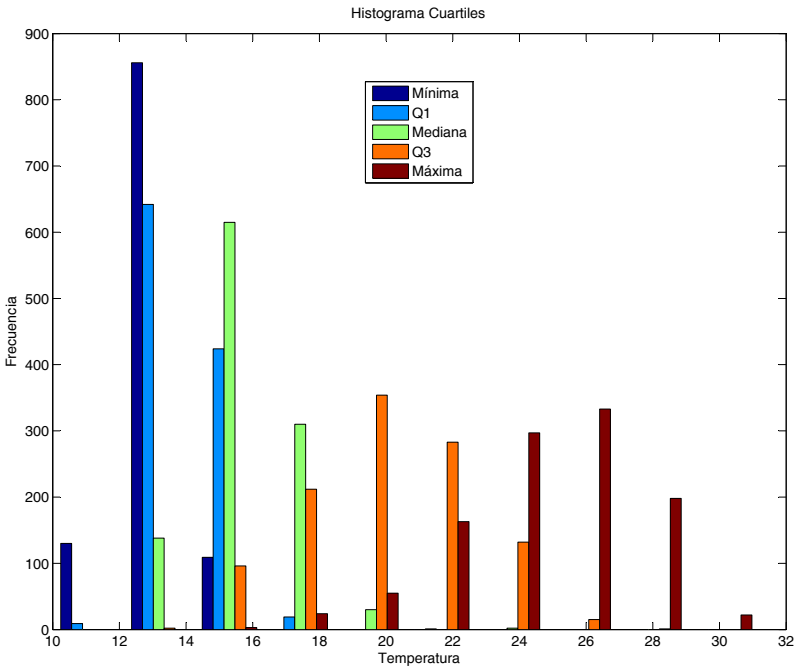


Figura 4-3: Tendencia central de temperatura media diaria, estación Posgrados, años 2009−2011: (a) Histograma para $\mu \pm \sigma$, (b) Histograma de cuartiles.

		Normal		Logistica		Weibull	
Posg	p	$0,0022$		$0,0168$		$0,0091$	
	(μ, σ)	$317,68$	$105,03$	$313,69$	$61,26$	$354,47$	$3,29$
Emas	p	$0,0011$		$0,008$		$0,02$	
	(μ, σ)	$298,76$	$102,77$	$293,36$	$59,35$	$334,15$	$3,13$
Enea	p	$0,1599$		$0,2629$		$7,67E-5$	
	(μ, σ)	$318,46$	$96,26$	$316,00$	$54,54$	$348,28$	$3,22$
Ing	p	$0,0012$		$0,0148$		$0,0413$	
	(μ, σ)	$268,93$	$101,21$	$264,80$	$58,58$	$299,13$	$2,71$

Tabla 4-3: Prueba de bondad de ajuste de la variable radiación media diaria entre 2009 – 2011 a distintas distribuciones en las 4 estaciones de estudio.

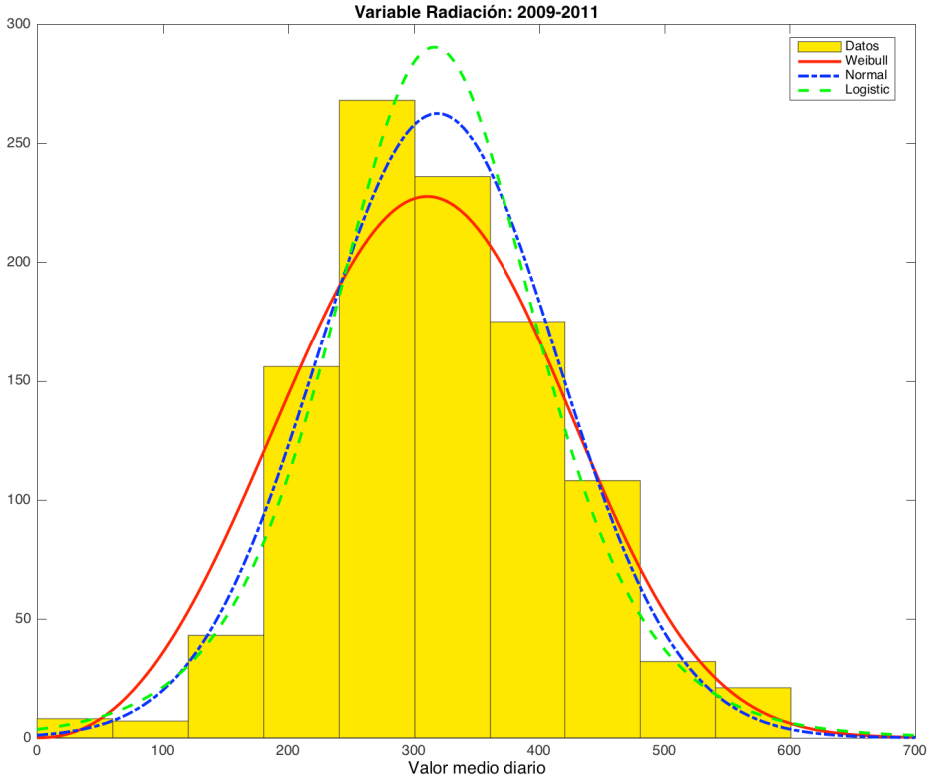
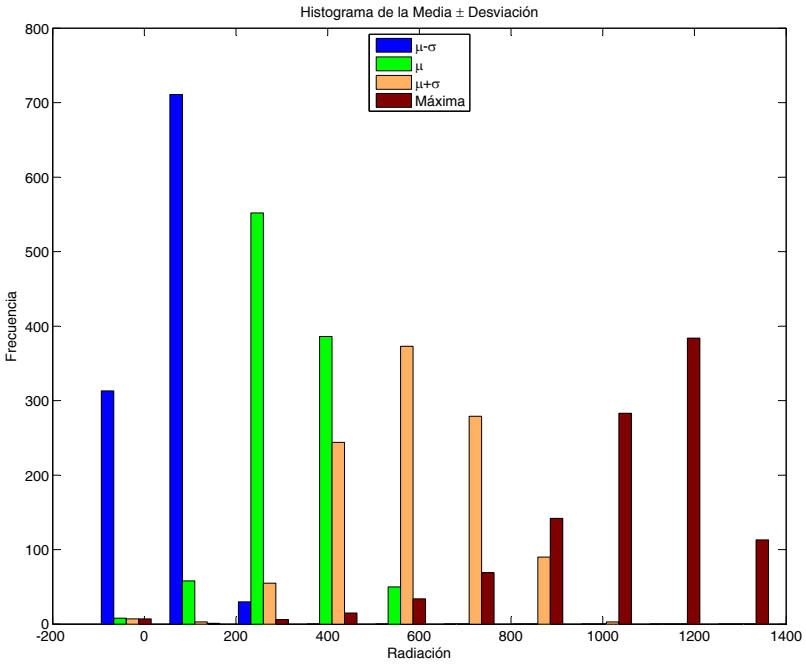


Figura 4-4: Prueba de ajuste a distribución para radiación media diaria en la estación Enea.

(a)



(b)

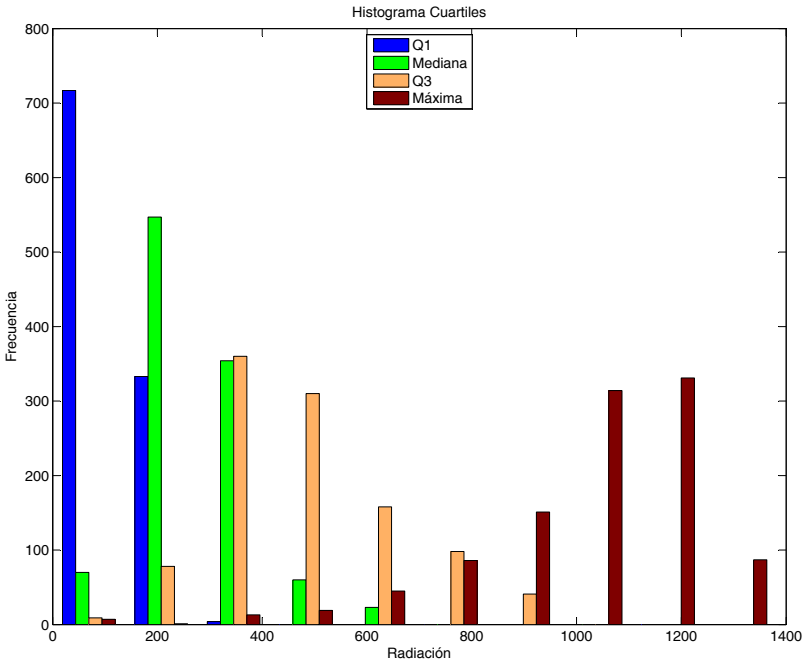


Figura 4-5: Tendencia central de radiación media diaria, estación Enea, años 2009 – 2011: (a) Histograma para $\mu \pm \sigma$, (b) Histograma de cuartiles.

		Normal		Logistica		Weibull	
Posg	p	$7,801E-4$		$0,0041$		$0,4152$	
	(μ,σ)	$87,75$	$7,15$	$88,15$	$4,11$	$90,92$	$14,81$
Emas	p	$0,0175$		$0,1746$		$0,6326$	
	(μ,σ)	$88,47$	$5,04$	$88,74$	$2,86$	$90,76$	$20,66$
Enea	p	$0,0527$		$0,0590$		$0,0596$	
	(μ,σ)	$86,74$	$7,09$	$86,93$	$4,14$	$89,95$	$14,00$
Ing	p	$0,3300$		$0,2775$		$0,1482$	
	(μ,σ)	$75,40$	$7,95$	$75,56$	$4,53$	$78,94$	$10,32$

Tabla 4-4: Prueba de bondad de ajuste de la variable humedad media diaria entre 2009 – 2011 a distintas distribuciones en las 4 estaciones de estudio

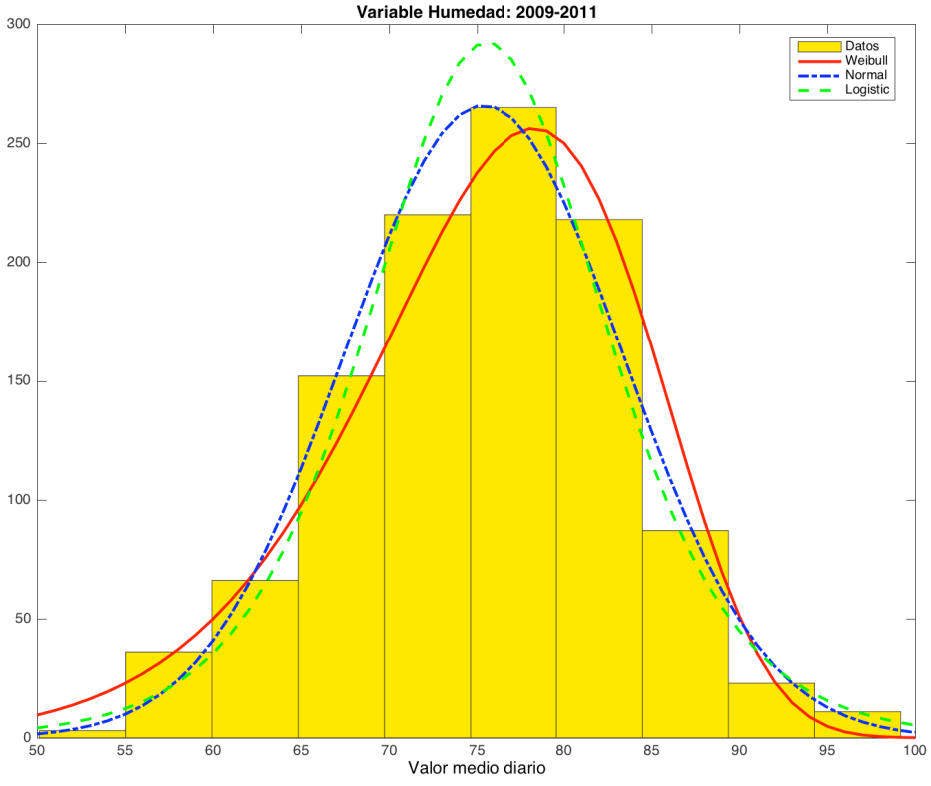
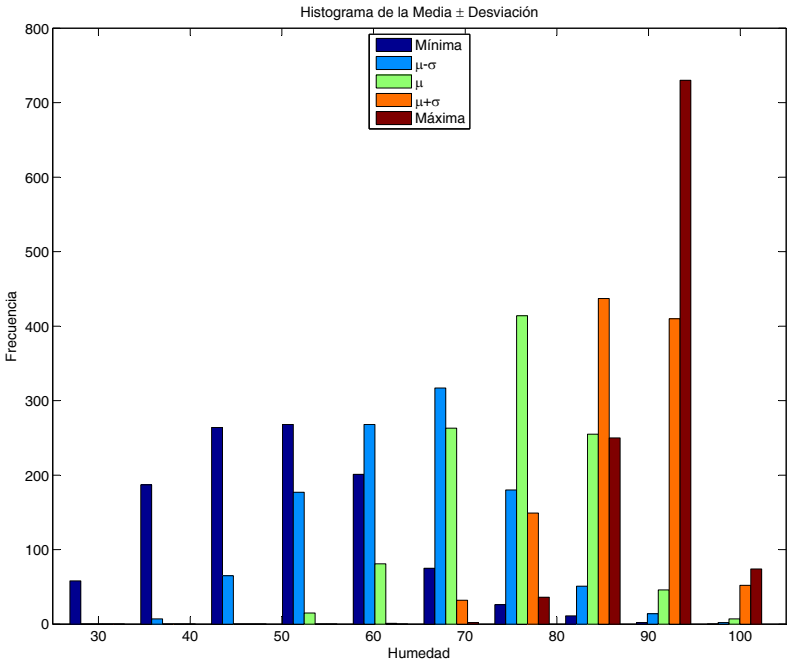


Figura 4-6: Prueba de ajuste a distribución para humedad media diaria en la estación Ingeominas.

(a)



(b)

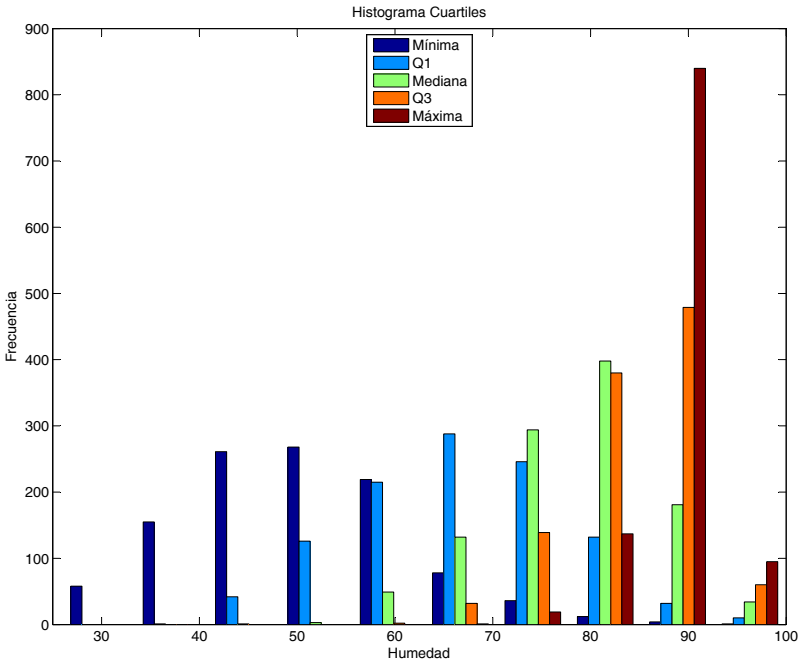


Figura 4-7: Tendencia central de humedad media diaria, estación Ingeominas, años 2009 – 2011: (a) Histograma para $\mu \pm \sigma$, (b) Histograma de cuartiles.

		Normal		Logistica		Weibull	
Posg	p	$7,29E-9$		$4,75E-9$		$0,054$	
	(μ,σ)	$91,17$	$67,33$	$84,11$	$37,25$	$98,12$	$1,29$
Emas	p	$3,19E-5$		$4,40E-6$		$0,0057$	
	(μ,σ)	$93,74$	$62,83$	$88,67$	$35,89$	$101,35$	$1{,}36$
Enea	p	$1,08E-6$		$5,12E-7$		$6,29E-7$	
	(μ,σ)	$65,57$	$46,12$	$61,42$	$25,82$	$68,74$	$1,19$
Ing	p	$1,61E-7$		$2,30E-6$		$0,0018$	
	(μ,σ)	$120,81$	$81,69$	$113,4$	$46,12$	$131,67$	$1,40$

Tabla 4-5: Prueba de bondad de ajuste de la variable media diaria de la precipitación acumulada entre $2009-2011$ a distintas distribuciones en las 4 estaciones de estudio

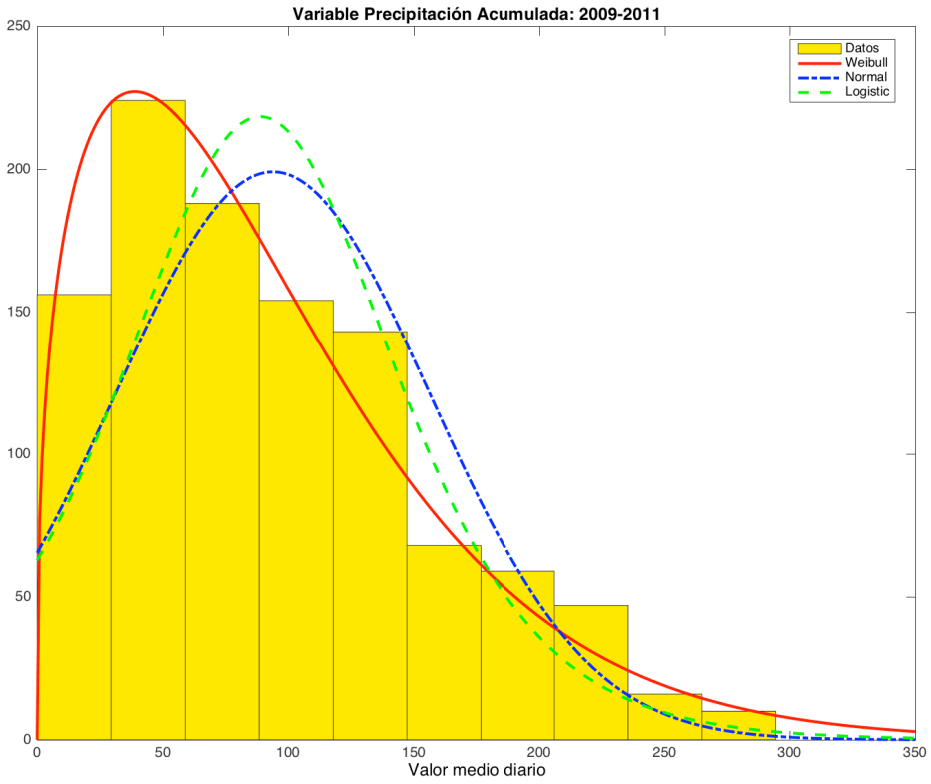
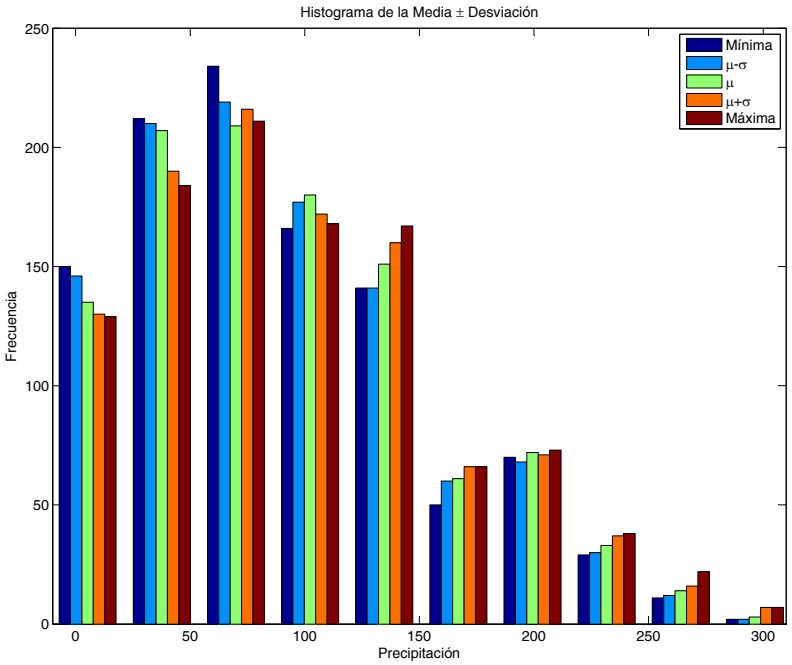


Figura 4-8: Prueba de ajuste a distribución para la media diaria de la precipitación acumulada en la estación Emas.

(a)



(b)

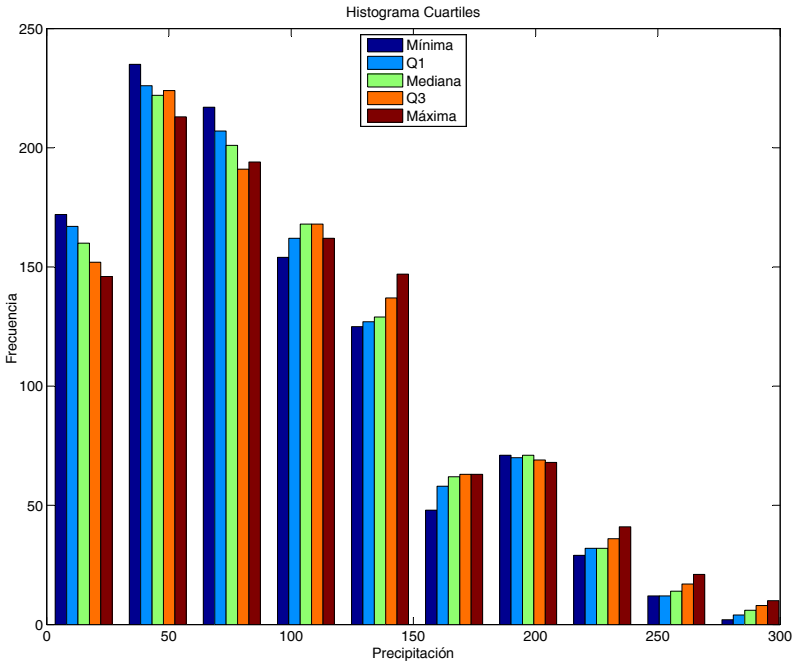


Figura 4-9: Tendencia central de la media diaria de la precipitación acumulada, estación Emas, años 2009−2011: (a) Histograma para $\mu \pm \sigma$, (b) Histograma de cuartiles.

4.1. Discusión

Se presenta en esta sección una lectura de los resultados encontrados arriba para el estudio de tendencia central de las variables meteorológicas de interés. Empezamos comentando los resultados de la graficación de medias para el grupo de variables **Temperatura-Radiación-Humedad-Precipitación** (Figura **4-1**). Se observa que los histogramas de mayor tendencia central corresponden a Temperatura y Radiación, en tanto Humedad se acumula a la derecha y precipitación a la izquierda. Se aprecian también rasgos de proporcionalidad directa entre Temperatura y Radiación y proporcionalidad inversa entre estas dos y Humedad, salvando una alta dispersión. La variable Precipitación no refleja relaciones similares con las otras variables.
Estas proporcionalides apreciadas resultan validadas por los coeficientes de correlación reportados en la Tabla **4-1**. Allí se observa que las correlaciones más fuertes son las relaciones inversas entre la humedad y las variables temperatura y radiación, seguida por la relación directa entre temperatura y radiación. Se reitera también la débil correlación de la variable precipitación con las demás. Cabe reportar que a un nivel de significancia de 5 % resultan rechazadas todas las hipótesis nulas asociadas a asumir independencia entre estas variables.

Pasamos a comentar los resultados para la variable **temperatura**. Como puede observarse en la Tabla **4-2** la prueba de Kolmogorov-Smirnoff resulta favorable para la temperatura media diaria, respecto de las distribuciones Normal y Logística en todas las estaciones consideradas en el período de estudio $2009 - 2011$. Para la distribución Weibull en ninguno de los casos se superó el umbral. Por ello se puede descartar que la temperatura media diaria provenga de una distribución Weibull en ese período para esas estaciones, con una confianza del 95 %. En la Figura **4-2** se aprecia un ajuste de calidad para el caso normal y logístico, en la estación Posgrados.
Continuando con la temperatura en la estación Posgrados, se puede observar que la temperatura mínima está casi siempre cerca de los $13°C$. En tanto, la temperatura máxima presenta un comportamiento muy disperso con valores que oscilan desde los 16 a los $31°C$, aunque la mayor concentración se da entre los 24 y los $27°C$. Ahora, si usamos la temperatura media o la mediana de cada día como criterio de agrupamiento para formar conglomerados, la Figura **4-3** sugiere que obtendríamos tres (3).

Las pruebas de bondad de ajuste para la variable **radiación** media diaria indican que sólo en la estación Enea para las distribuciones normal y logística no se rechaza la hipótesis nula a un nivel de significancia de 5 %. En ninguna de las estaciones se alcanzó umbral para la distribución Weibull. Por esto, la Figura **4-4** muestra el ajuste obtenido para la estación Enea. Ahora, en la Figura **4-3** se aprecia cómo durante 75 % del día la radiación está esencialmente entre 0 y $800W/m^2$. A su vez la radiación máxima asume valores en un rango muy amplio, no obstante se concentra fuertemente en los valores más grandes, concretamentamente casi

siempre supera los $900W/m^2$ y llegando hasta los $1350W/m^2$. Indicando esto que en la mayoría de los días se alcanzan niveles de radiación muy altos, con respecto a los registros de la ciudad.

Para el caso de la media diaria de la **humedad relativa** los resultados de ajuste a distribución tienen fuertes cambios dependiendo la estación considerada. Para la estación Posgrados la única hipótesis no rechazada es la Weibull. En Emas no se rechazan Logística ni Weibull. En Enea e Ingeominas se aceptan las tres distribuciones. En la Figura **4-6** se muestra el caso Ingeominas.
En las gráficas de la Figura **4-7**, se tiene que la humedad mínima puede llegar a registrar valores en un rango bastante amplio, entre 25 y 100 %. La mediana es casi siempre superior a 70 % y la humedad relativa máxima siempre superior a 80 %. Queriendo decir esto que aunque puede haber días en los que la humedad llegue a registrar valores tan bajos como 30 % ese mismo día superará el 80 %.

Cerramos esta sección comentando los resultados de tendencia central para la variable **precipitación acumulada**. Puede decirse que las pruebas de ajuste con mejores resultados fueron para la distribución Weibull, aunque sólo se supera el umbral de aceptación para la estación Posgrados. En la Figura **4-8** se muestra el resultado para la estación Emas.
Debe tenerse en cuenta que como en cada instante de tiempo la variable precipitación acumulada marca la cantidad total de precipitación de los últimos 15 días desde ese momento, los estadísticos en la Figura **4-9** presentan una variación menor. No obstante, podemos hablar de dos regímenes uno para valores inferiores a $150mm$ y otro para los superiores.

5 Patrones de acumulación: Resultados

En este capítulo presentamos los resultados de la aplicación de métodos de agrupación jerárquicos y de partición sobre los datos disponibles en la base de datos delimitada para este estudio. Se trabaja sobre las mismas variables meteorológicas de interés: temperatura, radiación solar, humedad y precipitación acumulada, en el período $2009 - 2011$ y en las estaciones Posgrados, Enea, Emas e Ingeominas.
Se adelantaron cuatro estudios de conglomerados por cada variable, uno por cada estación. El estudio se basa en métodos jerárquicos aglomerativos tal y como son descritos en la sección 2.1. Se trabaja con las disimilitudes consideradas en la sección 2.1.2, acompañadas de los métodos de vinculación presentados en la sección 2.2 (única, completa, promedio, *ward*) y con la vinculación por Hausdorff (sección 2.2.1). Teniendo precaución de usar el método *ward* sólo con distancia euclidiana.
De las combinaciones disimilitud-método que resultaron convergentes, se escogió para fines de ilustración y referenciación la de mejor desempeño para ser mostrada en este capítulo. Para la escogencia del número de conglomerados a encontrar se utiliza el criterio de coeficiente silueta (sección 2.4) como valor inicial, luego se compara ese resultado con la estructura del dedrograma, pudiendo modificarse para conservar consistencia entre dicha estructura y valores altos de silueta promedio.
Los dendrogramas para las cuatro estaciones en consideración, para las variables temperatura, radiación solar, humedad y precipitación pueden encontrarse en las Figuras **5-1**, **5-8**, **5-15** y **5-22**, respectivamente. La aplicación del criterio valor medio de silueta puede apreciarse en las Figuras **5-2**, **5-9**, **5-16** y **5-23**.
Presentamos también aquí los valores de la silueta de los elementos agrupados en cada conglomerado para la escogencia final, esto como un importante indicador de la calidad de la asociación de los elementos al conglomerado asignado (Figuras **5-3**, **5-10**, **5-17** y **5-24**). Adicionalmente, para facilitar la relación de los conglomerados encontrados con las distintas épocas del año, se muestra la clase a que es asignado cada día de los años en estudio para cada una de las variables y estaciones. Temperatura: Figuras **5-4**, **5-5**, radiación: Figuras **5-11**, **5-12**, humedad: Figuras **5-18**, **5-19** y precipitación: Figuras **5-25**, **5-26**.

A modo de validación de los resultados obtenidos se adelantaron dos pruebas. La primera consistió en la implementación del método de partición de K-medias bajo las distintas disimilitudes consideradas arriba, de modo que se aplicara un método radicalmente distinto a los datos y comparar los resultados. Para la escogencia del número de conglomerados se

tiene en cuenta el criterio silueta para K-medias y el número de conglomerados usado en el caso a validar. Los resultados pueden revisarse en las Figuras **5-6**, **5-13**, **5-20** y **5-27**.
La segunda prueba es más visual y se basa en el comando `clustergram` de MATLAB® que junto con el dendrograma muestra un grafico de la magnitud de los datos agrupados. Esto en procura de hacer identificaciones entre ellos. Ver Figuras **5-7**, **5-14**, **5-21** y **5-28**.

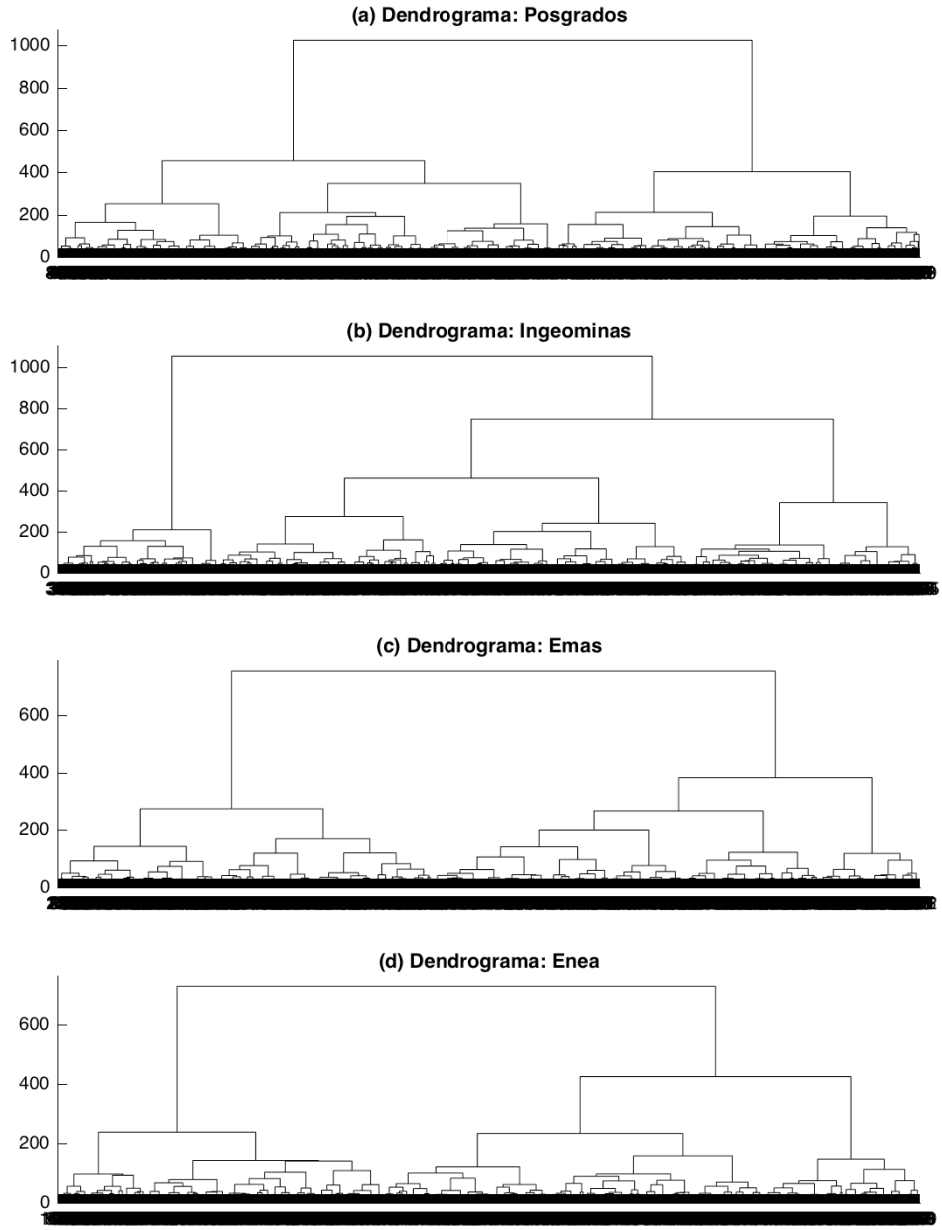


Figura 5-1: Dendrogramas de la variable temperatura usando *Linkage*, con método *ward* y distancia euclidiana, período 2009 – 2011.

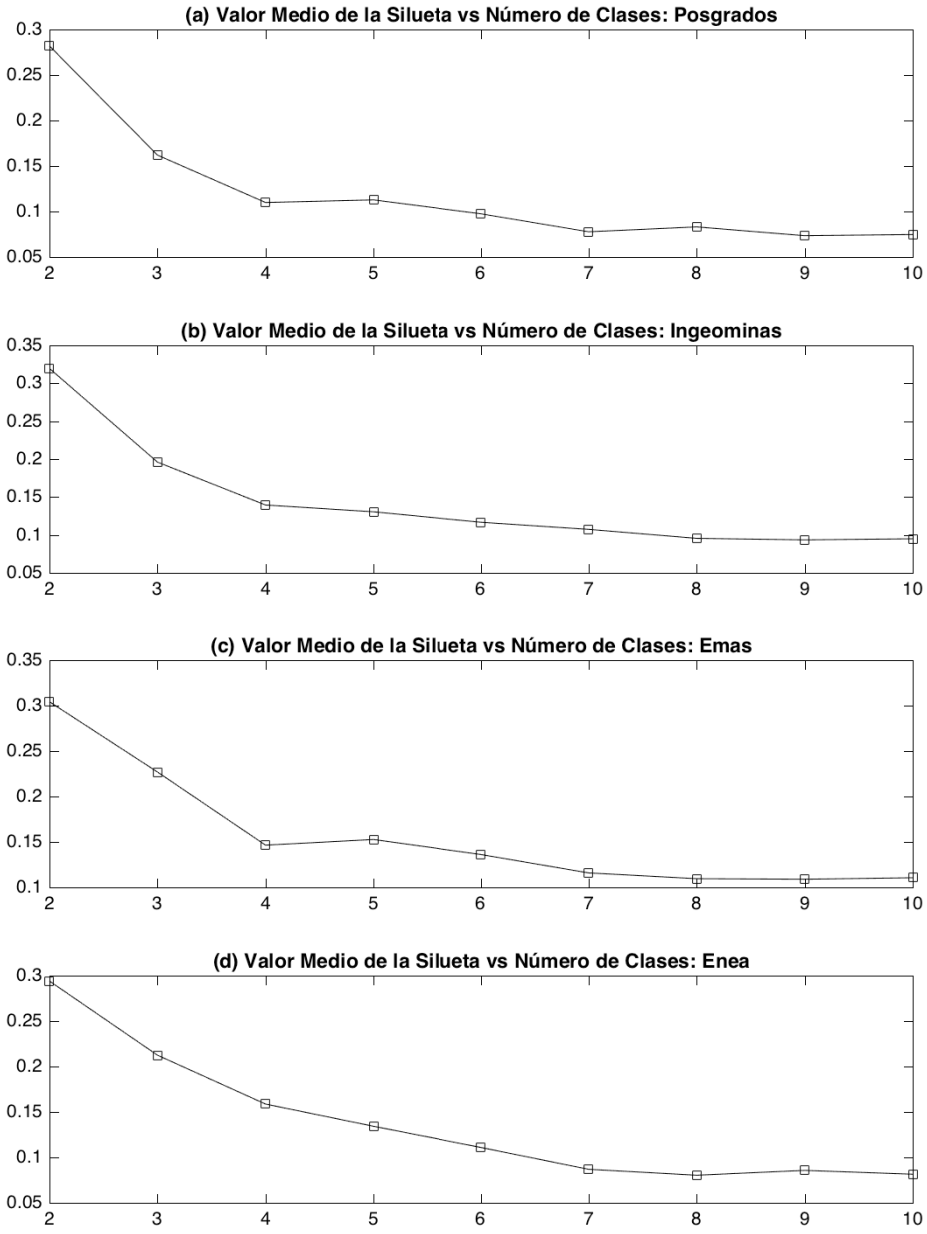


Figura 5-2: Valor medio de silueta en la variable temperatura, período 2009 – 2011.

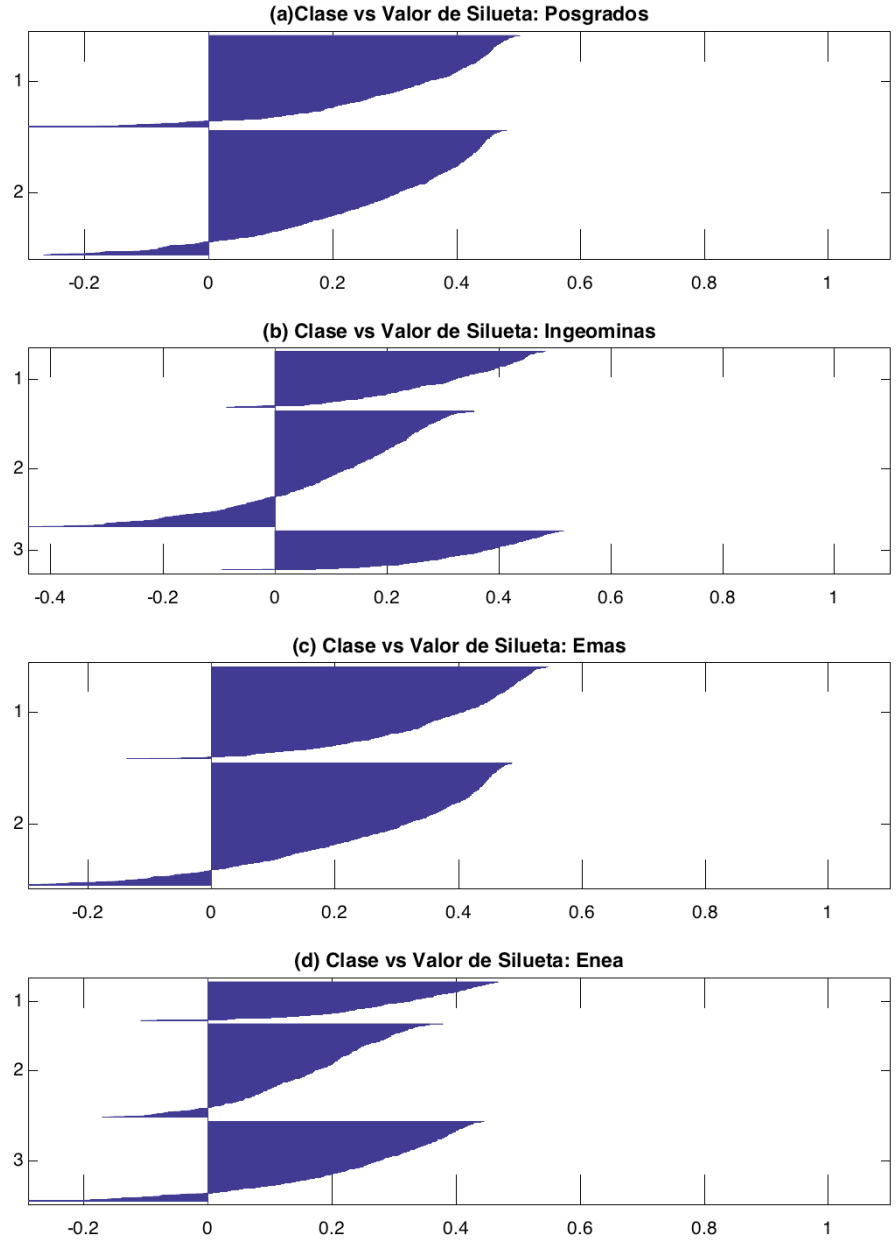


Figura 5-3: Valor de silueta en la variable temperatura para el número de conglomerados escogido, período 2009 – 2011.

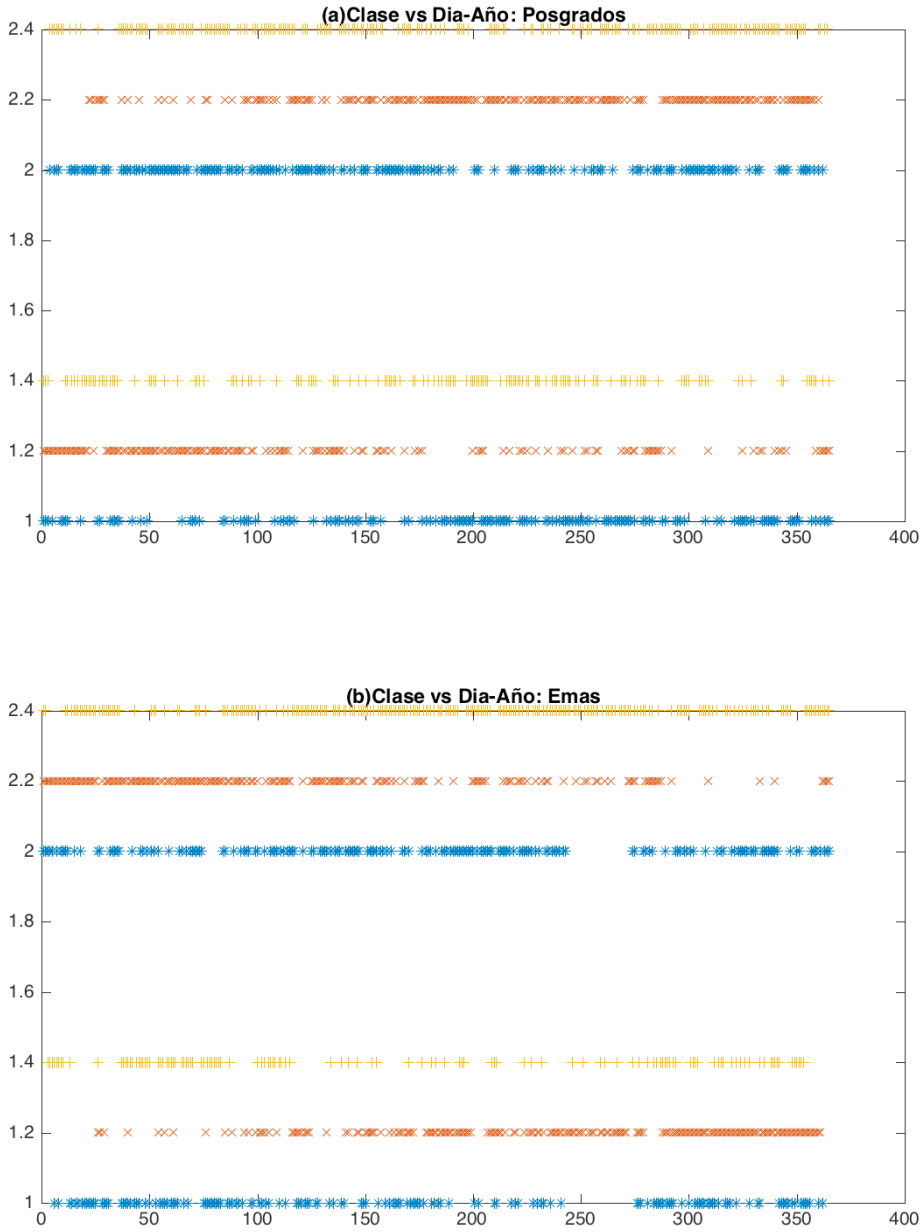


Figura 5-4: Asignación de conglomerado a cada día, para la variable temperatura en las estaciones Posgrados y Emas. Azul: 2009, Rojo: 2010, Amarillo: 2011.

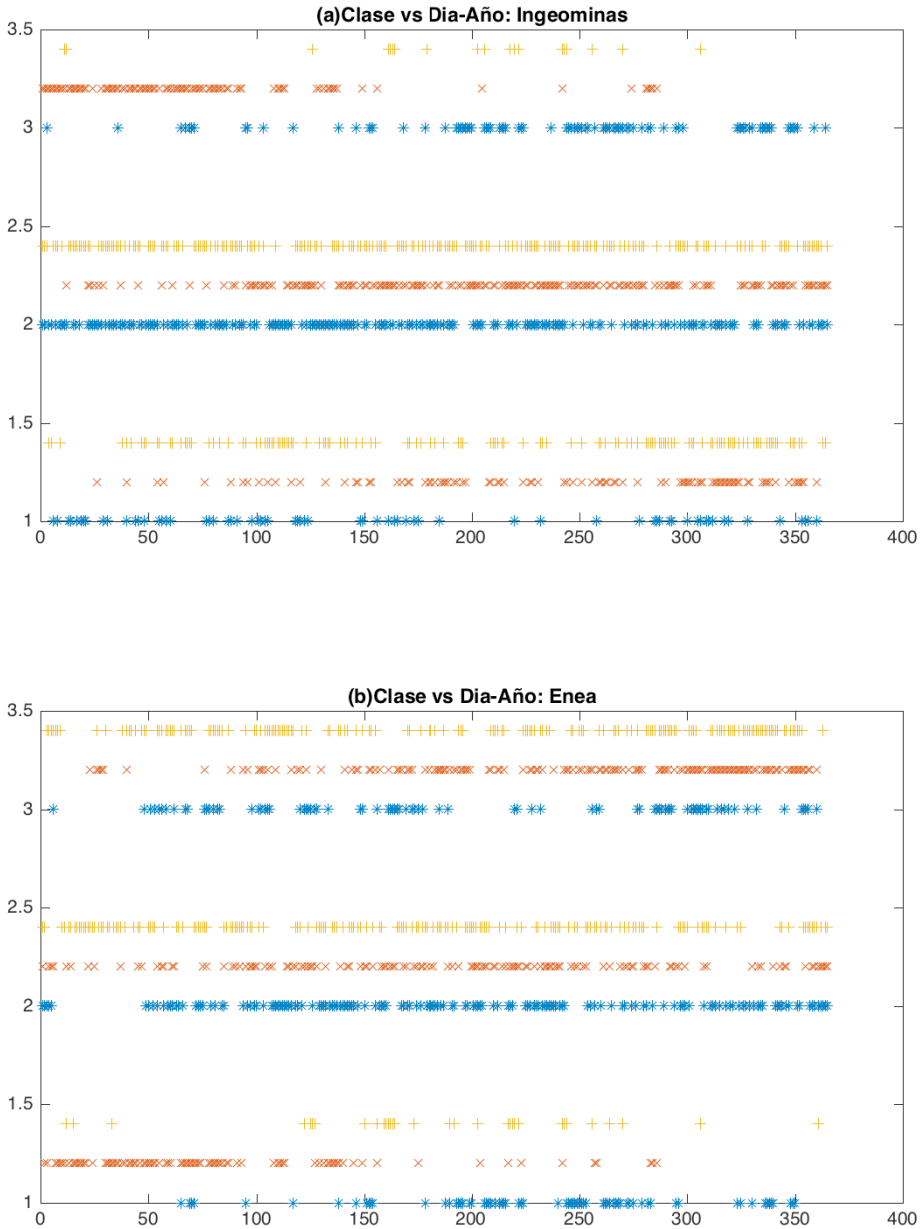


Figura 5-5: Asignación de conglomerado a cada día, para la variable temperatura en las estaciones Ingeominas y Enea. Azul: 2009, Rojo: 2010, Amarillo: 2011.

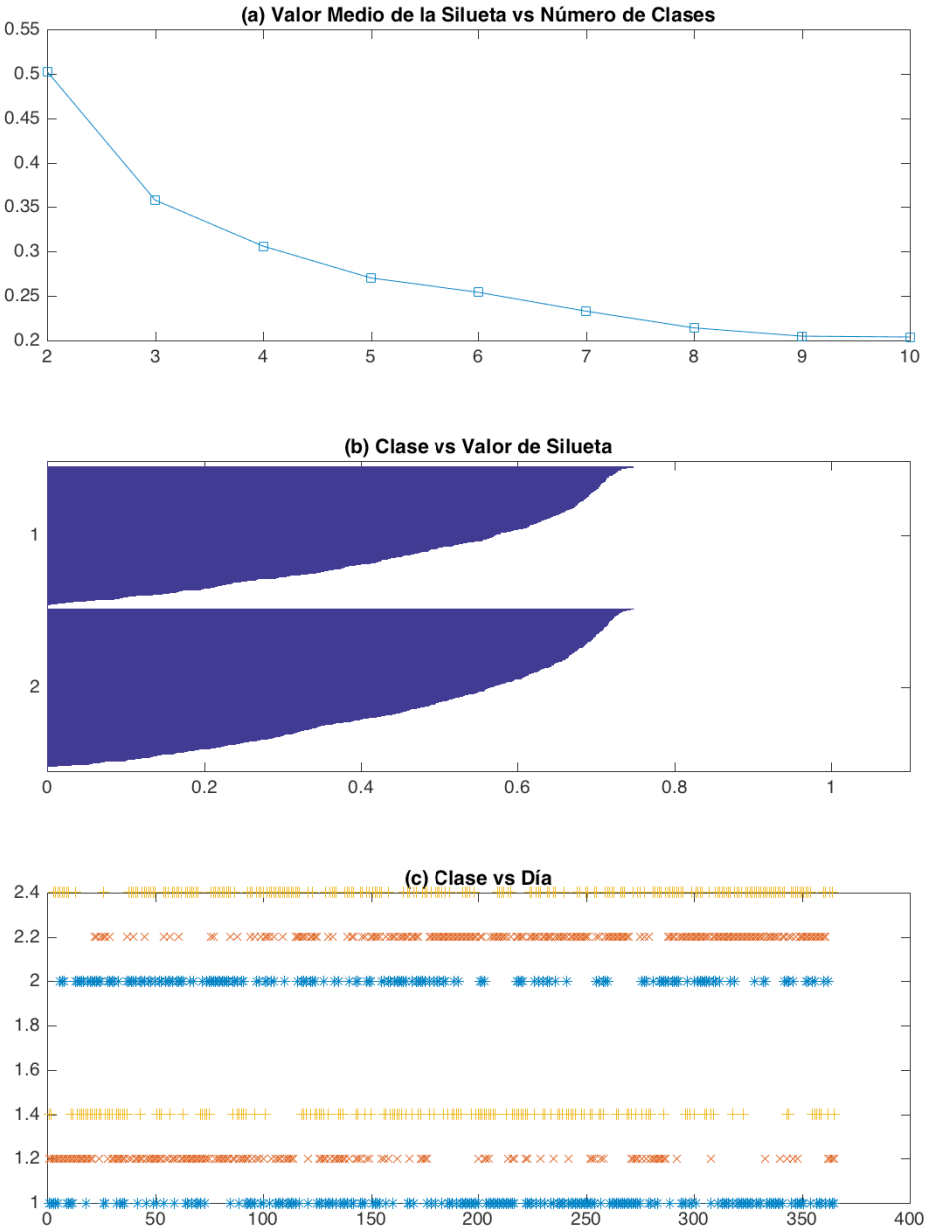


Figura 5-6: Patrones de acumulación de la variable temperatura usando K-medias, con disimilitud euclidiana cuadrada, en la estación Posgrados, período $2009-2011$.

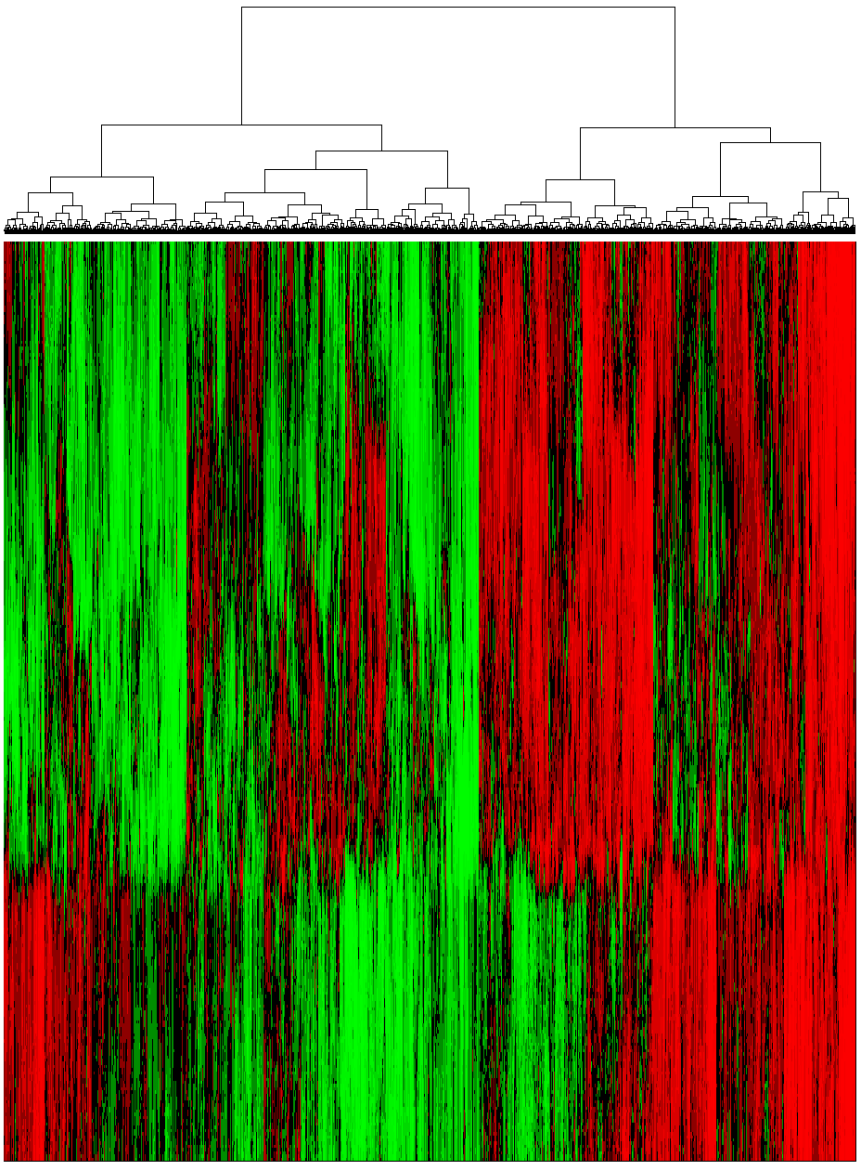


Figura 5-7: Gráfico de *clustergram* para la variable temperatura en la estación Enea, período 2009 – 2011.

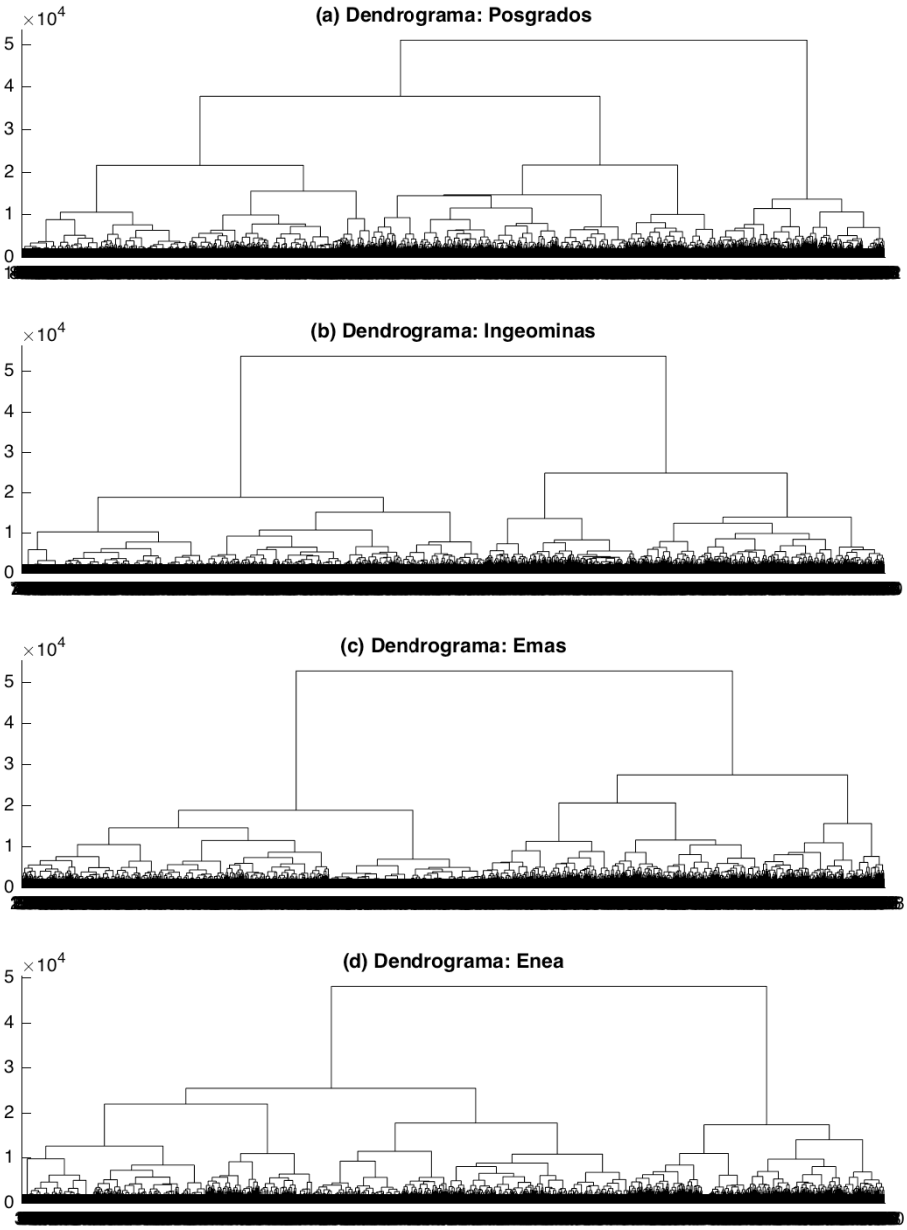


Figura 5-8: Dendrogramas de la variable radiación usando *Linkage*, con método *ward* y distancia euclidiana, período 2009 – 2011.

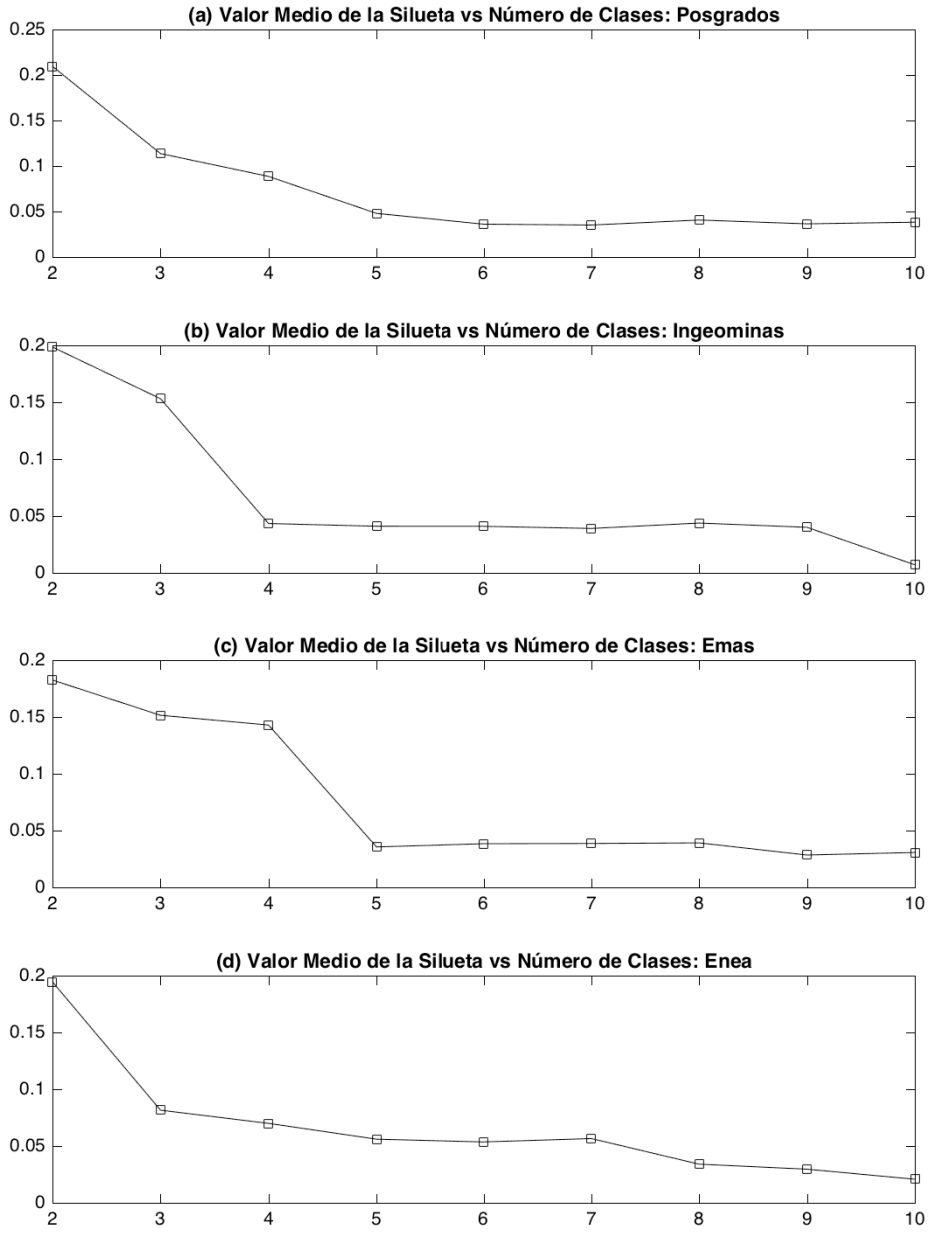


Figura 5-9: Valor medio de silueta en la variable radiación, período 2009 – 2011.

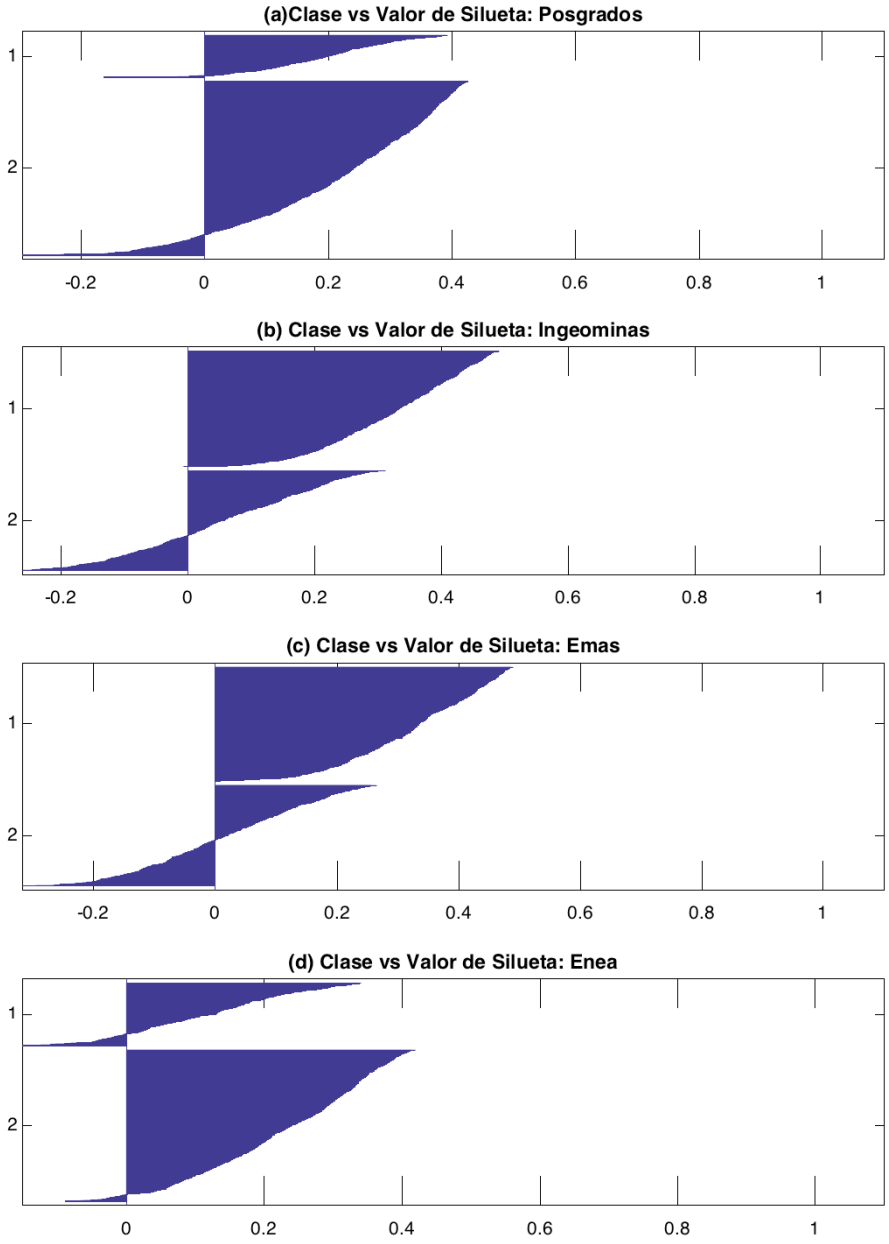


Figura 5-10: Valor de silueta en la variable radiación para el número de conglomerados escogido, período 2009 – 2011.

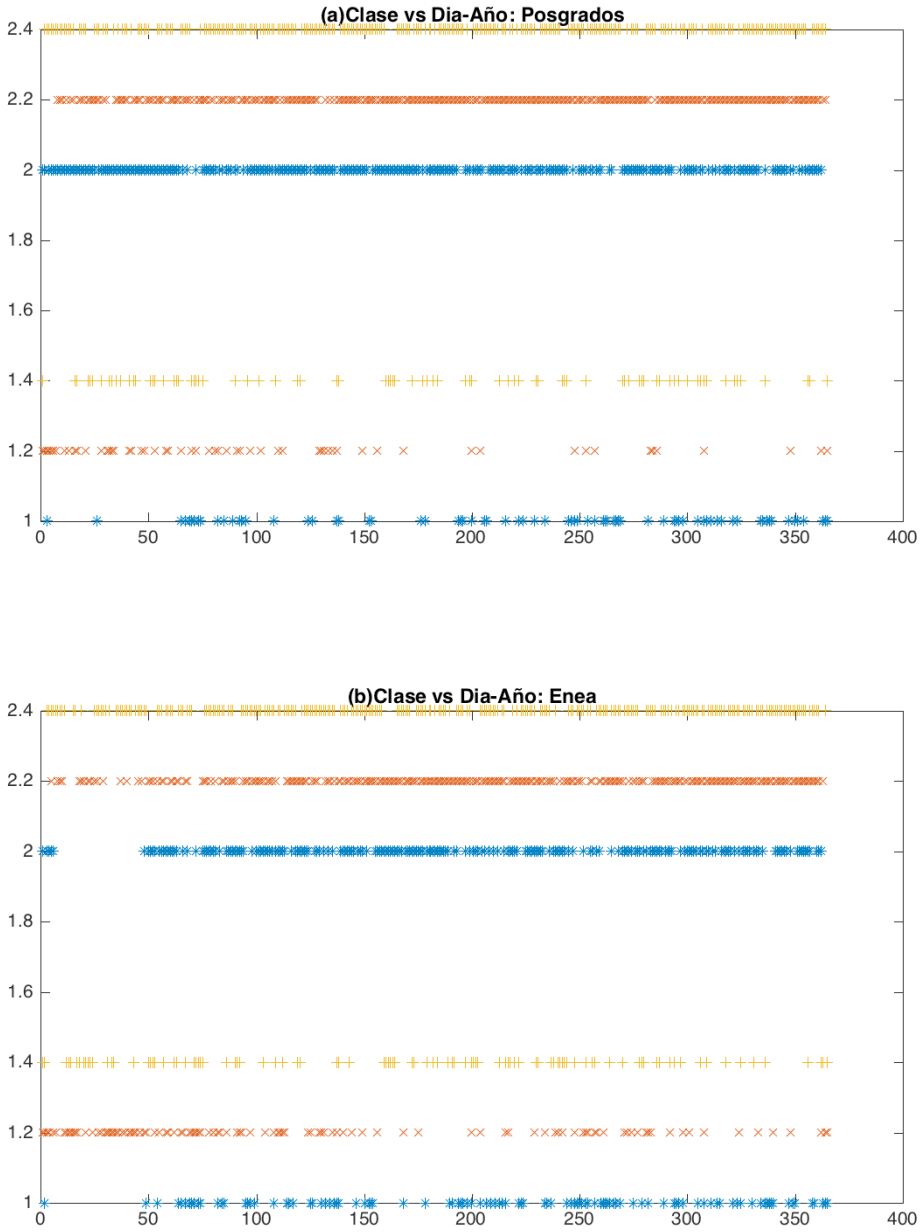


Figura 5-11: Asignación de conglomerado a cada día, para la variable radiación en las estaciones Posgrados y Enea. Azul: 2009, Rojo: 2010, Amarillo: 2011.

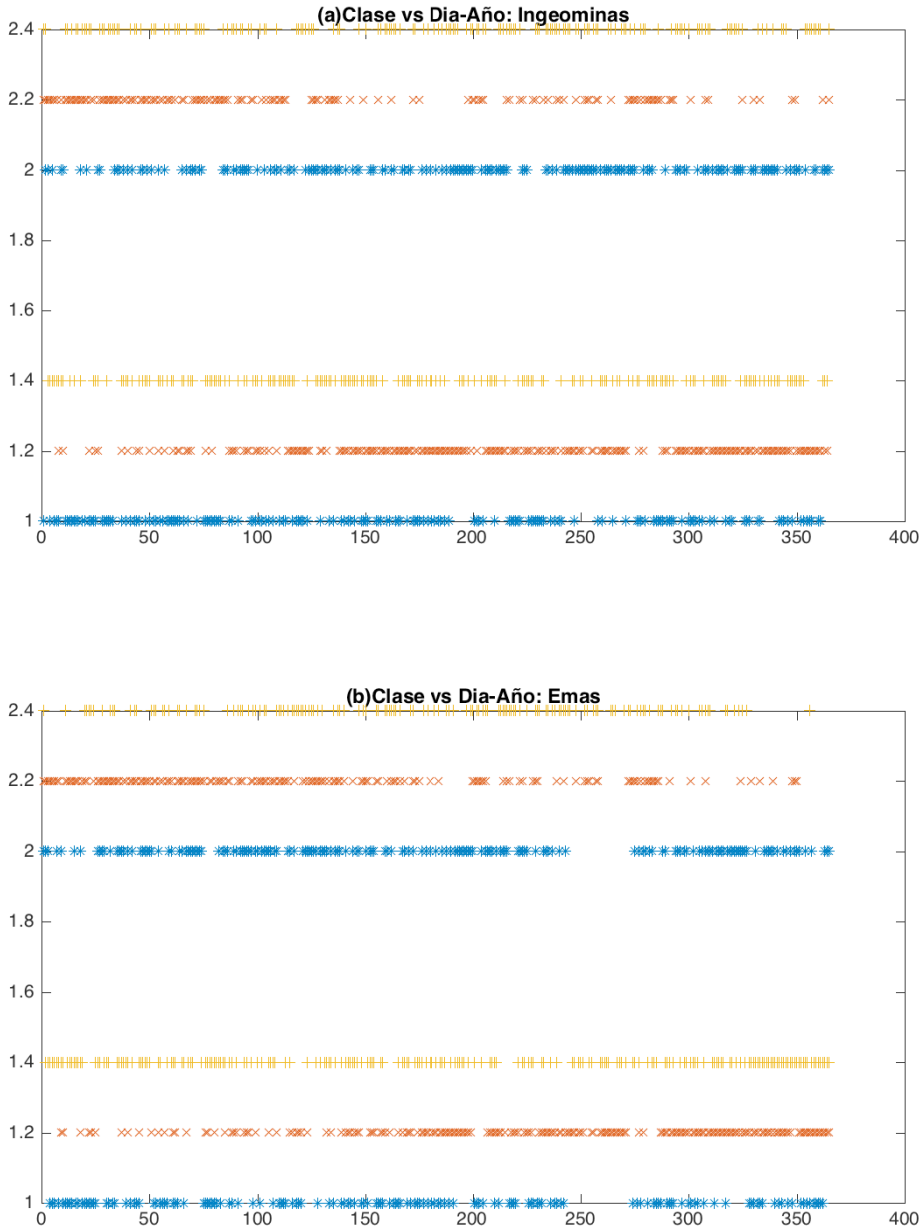


Figura 5-12: Asignación de conglomerado a cada día, para la variable radiación en las estaciones Ingeominas y Emas. Azul: 2009, Rojo: 2010, Amarillo: 2011.

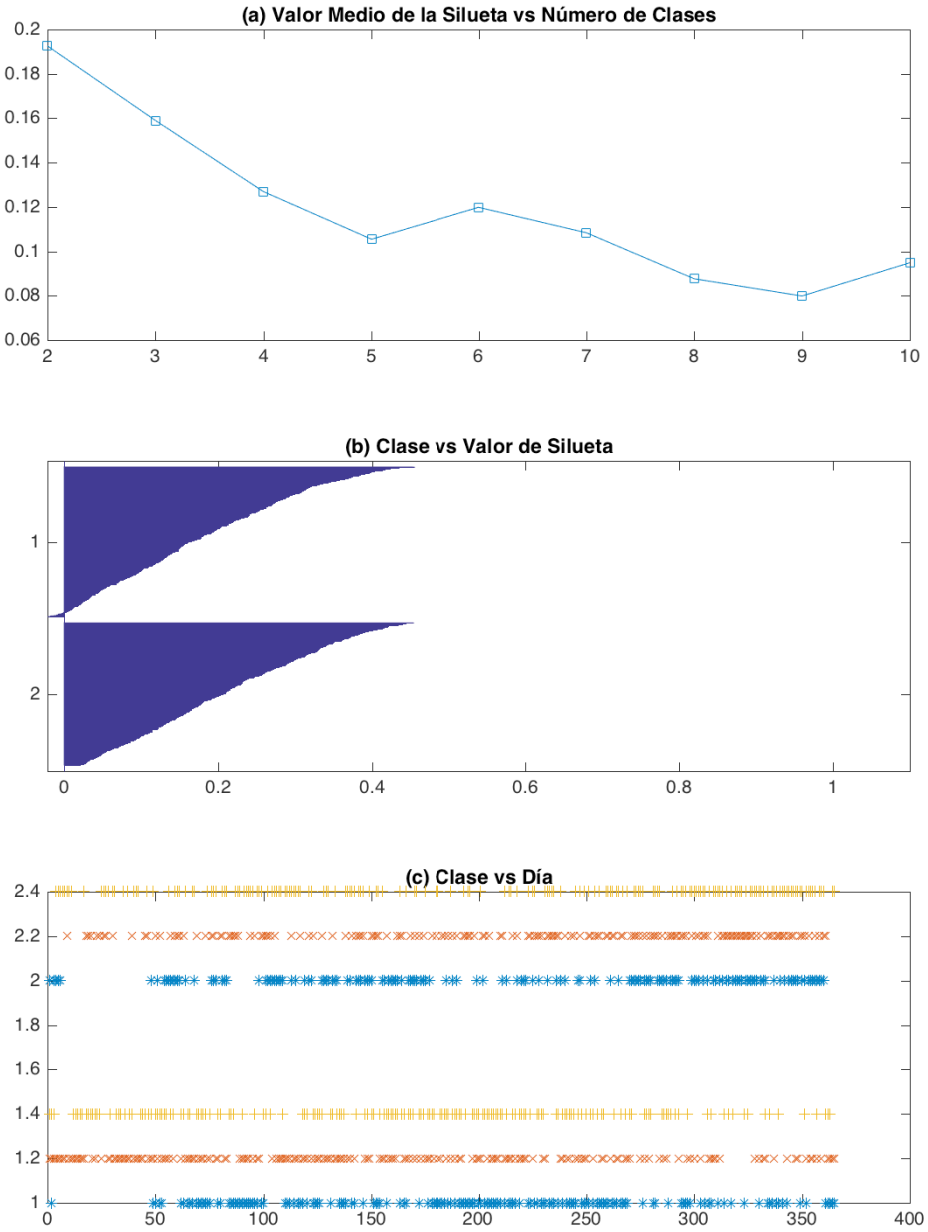


Figura 5-13: Patrones de acumulación de la variable radiación usando K-medias, con disimilitud coseno, en la estación Enea, período 2009 – 2011.

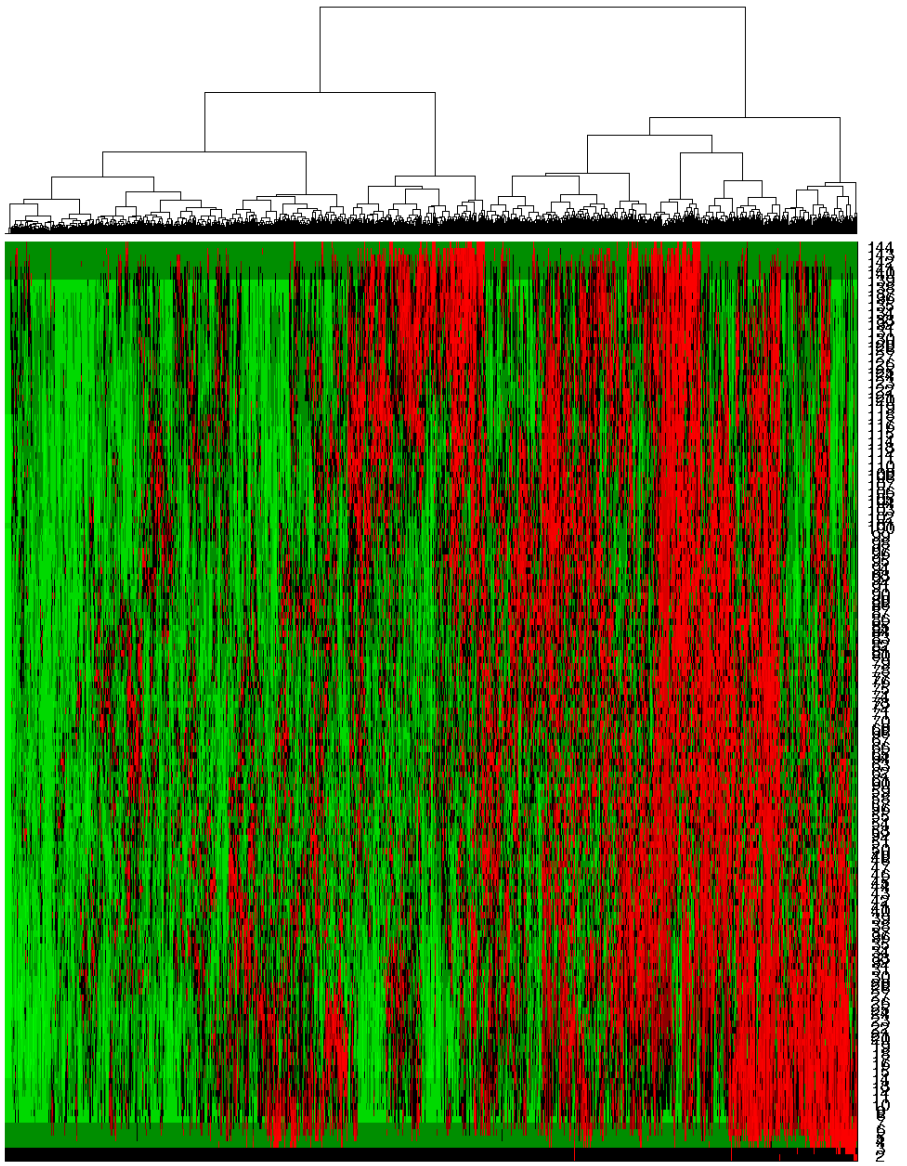


Figura 5-14: Gráfico de *clustergram* para la variable radiación en la estación Ingeominas, período 2009 − 2011.

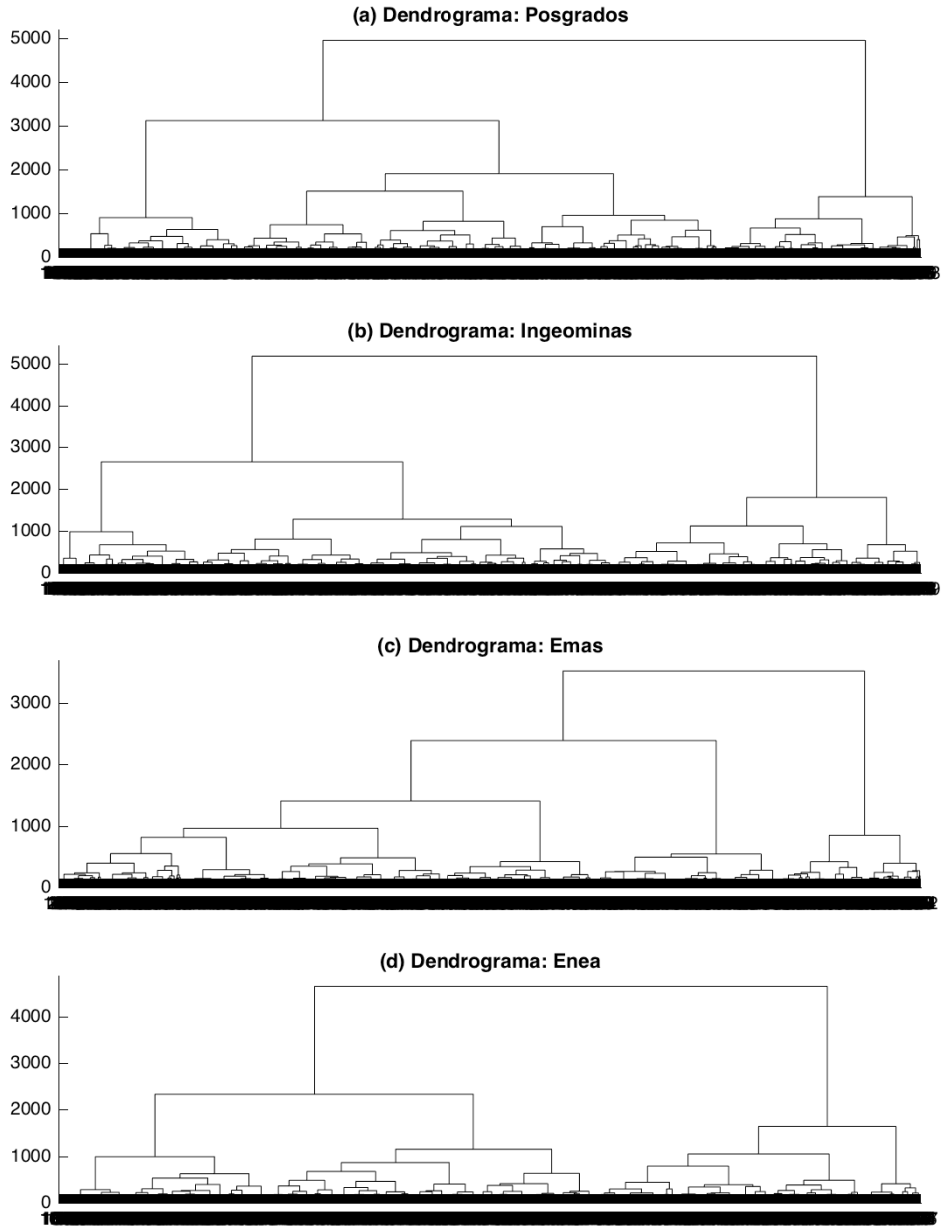


Figura 5-15: Dendrogramas de la variable humedad usando *Linkage*, con método *ward* y distancia euclidiana, período $2009 - 2011$.

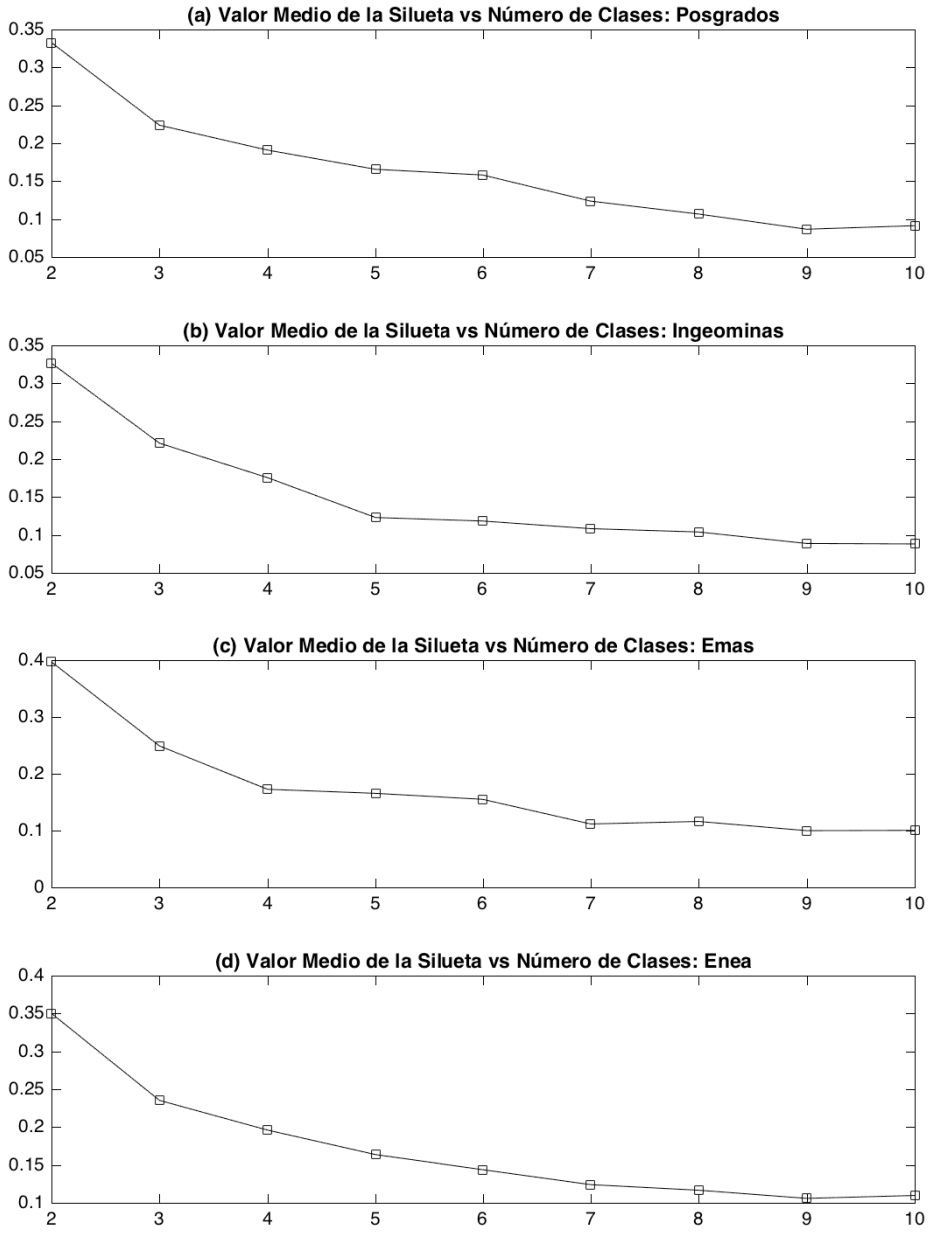


Figura 5-16: Valor medio de silueta en la variable humedad, período $2009 - 2011$.

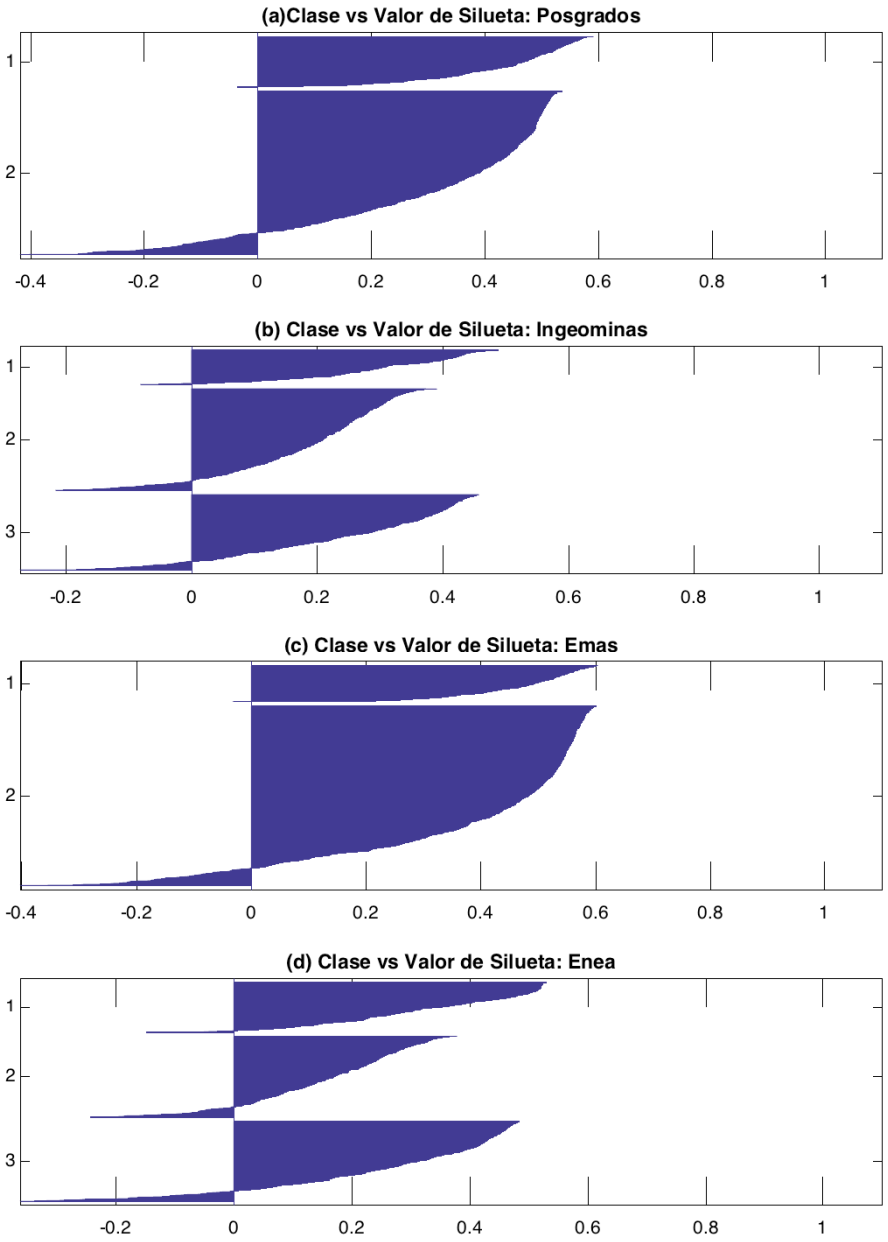


Figura 5-17: Valor de silueta en la variable humedad para el número de conglomerados escogido, período $2009 - 2011$.

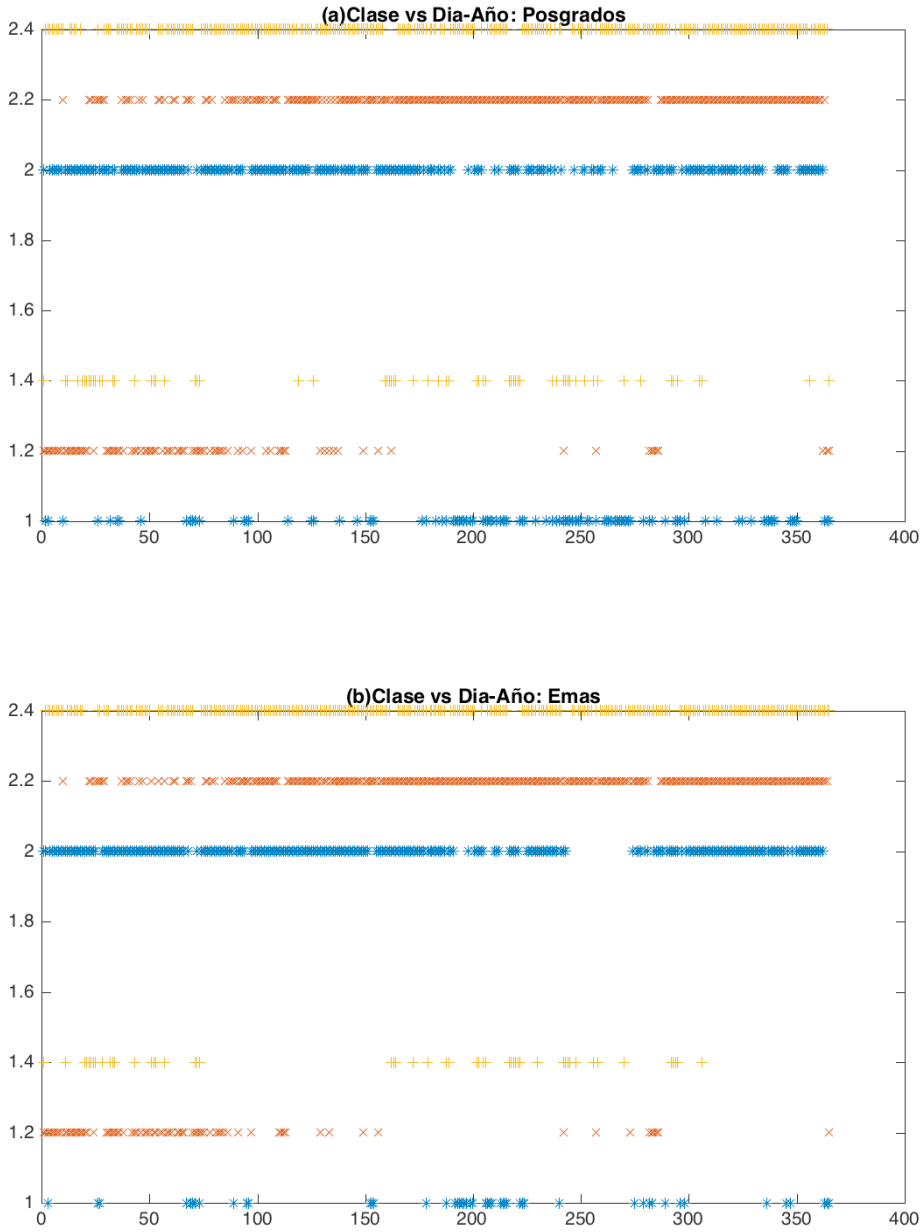


Figura 5-18: Asignación de conglomerado a cada día, para la variable humedad en las estaciones Posgrados y Emas. Azul: 2009, Rojo: 2010, Amarillo: 2011.

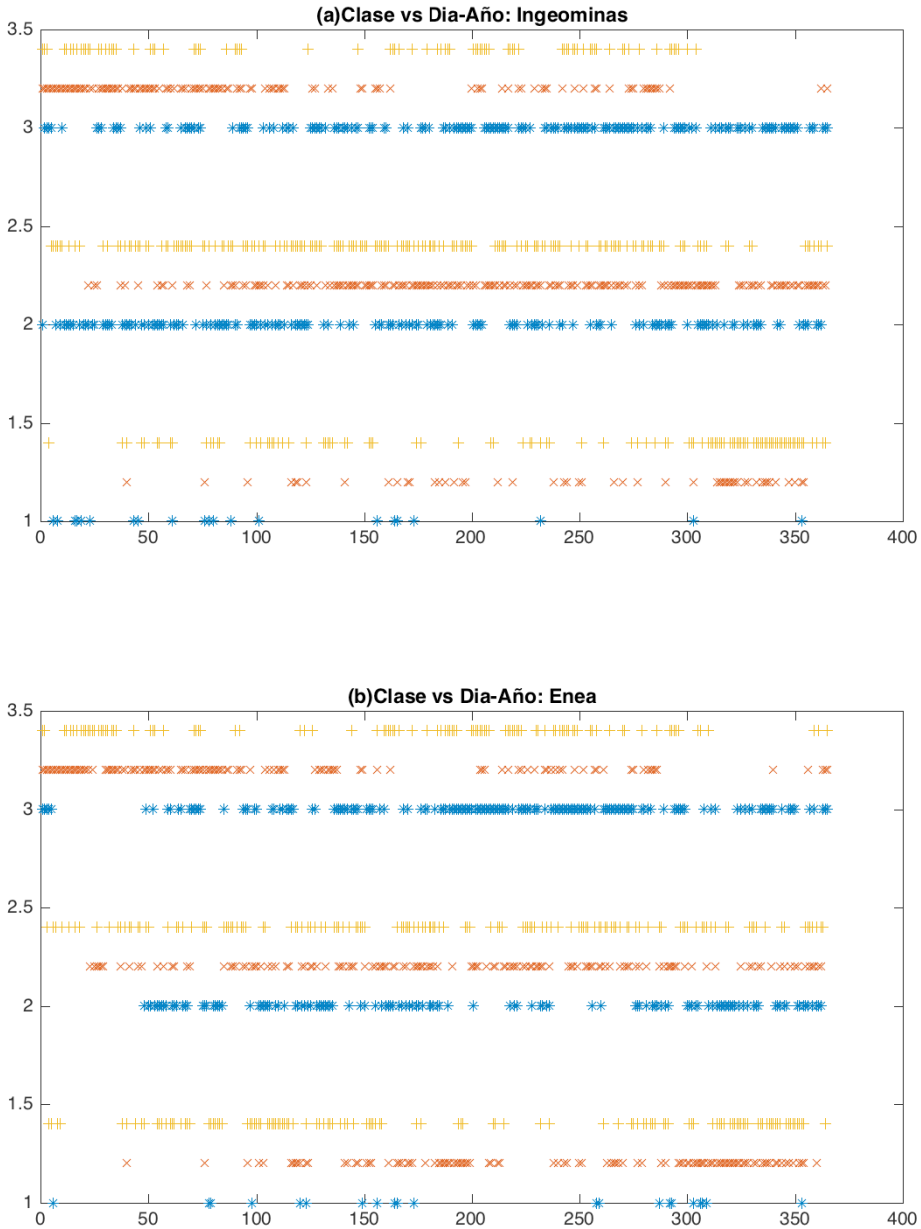


Figura 5-19: Asignación de conglomerado a cada día, para la variable humedad en las estaciones Ingeominas y Enea. Azul: 2009, Rojo: 2010, Amarillo: 2011.

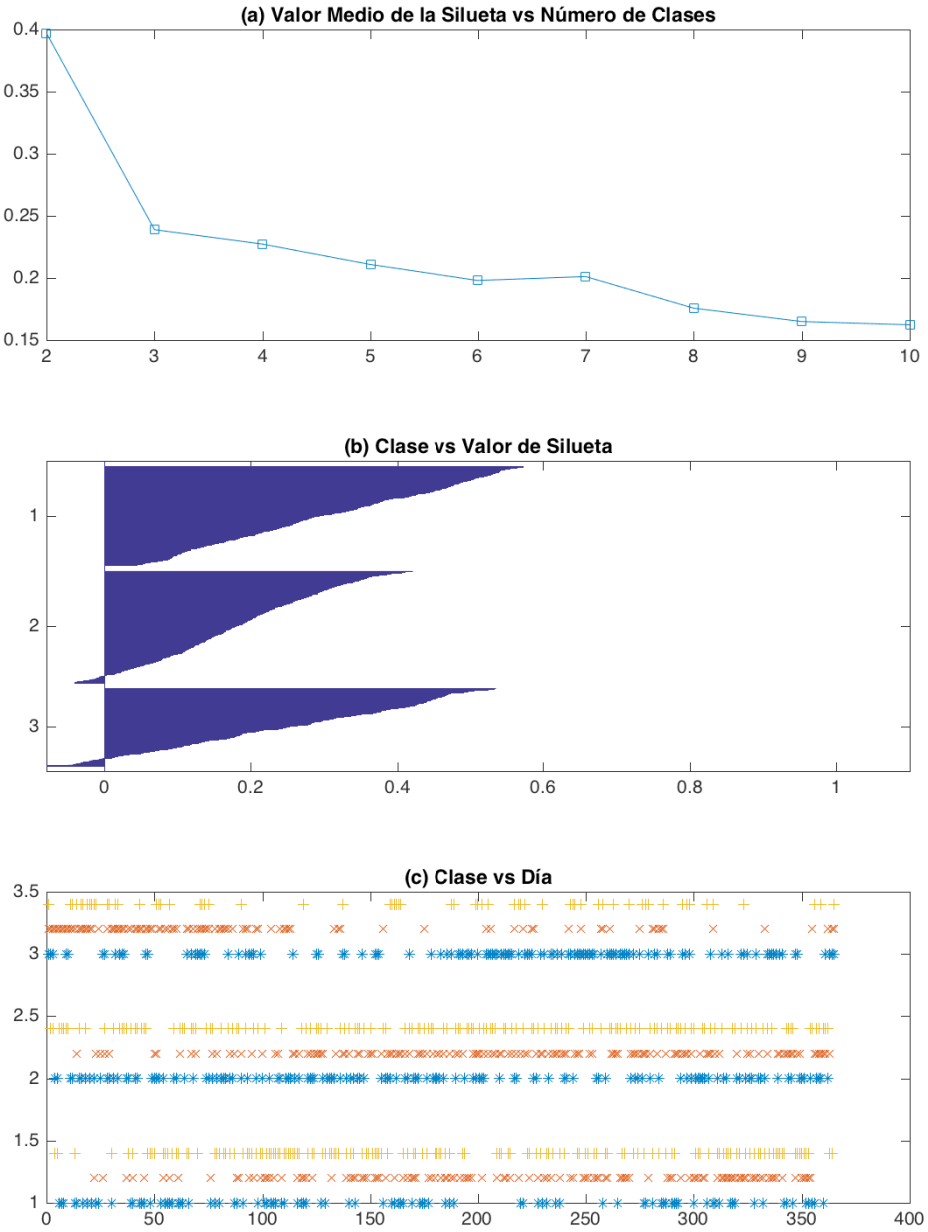


Figura 5-20: Patrones de acumulación de la variable humedad usando K-medias, con disimilitud coseno, en la estación Ingeominas, período 2009 – 2011.

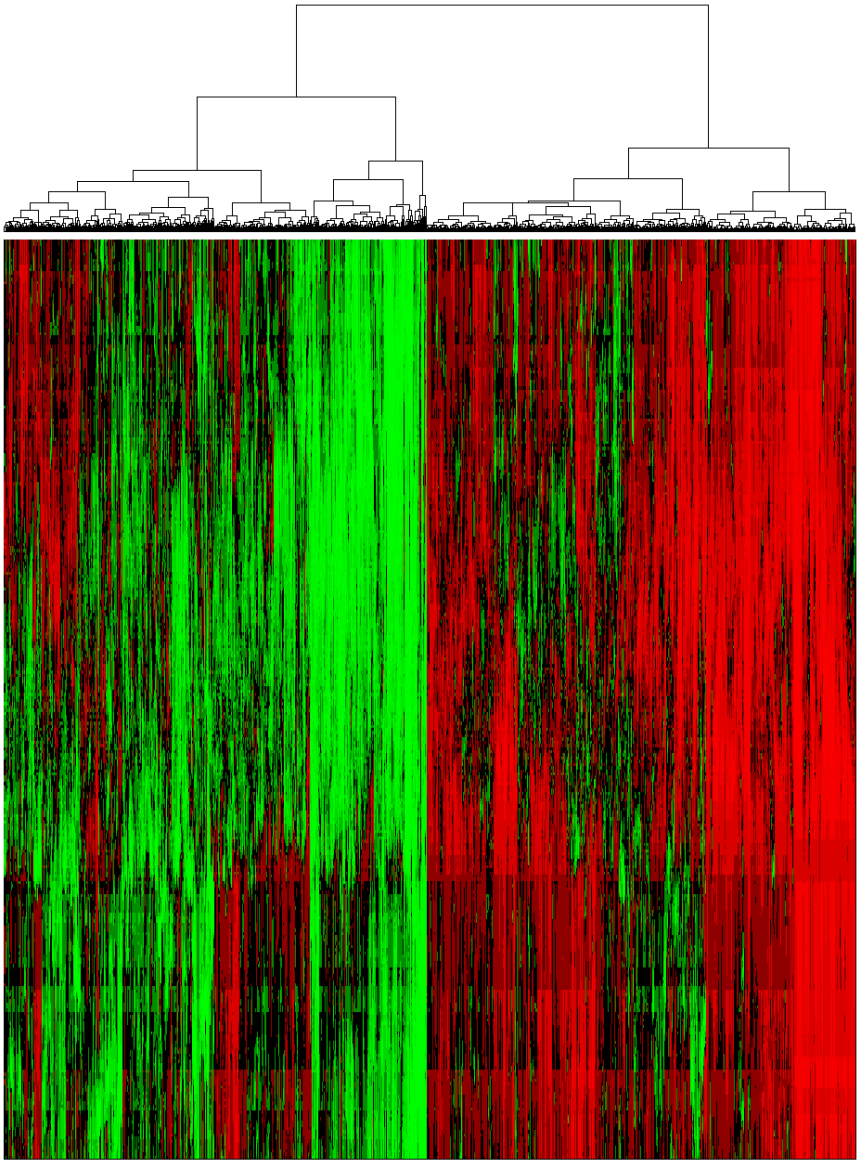


Figura 5-21: Gráfico de *clustergram* para la variable humedad en la estación Emas, período 2009 – 2011.

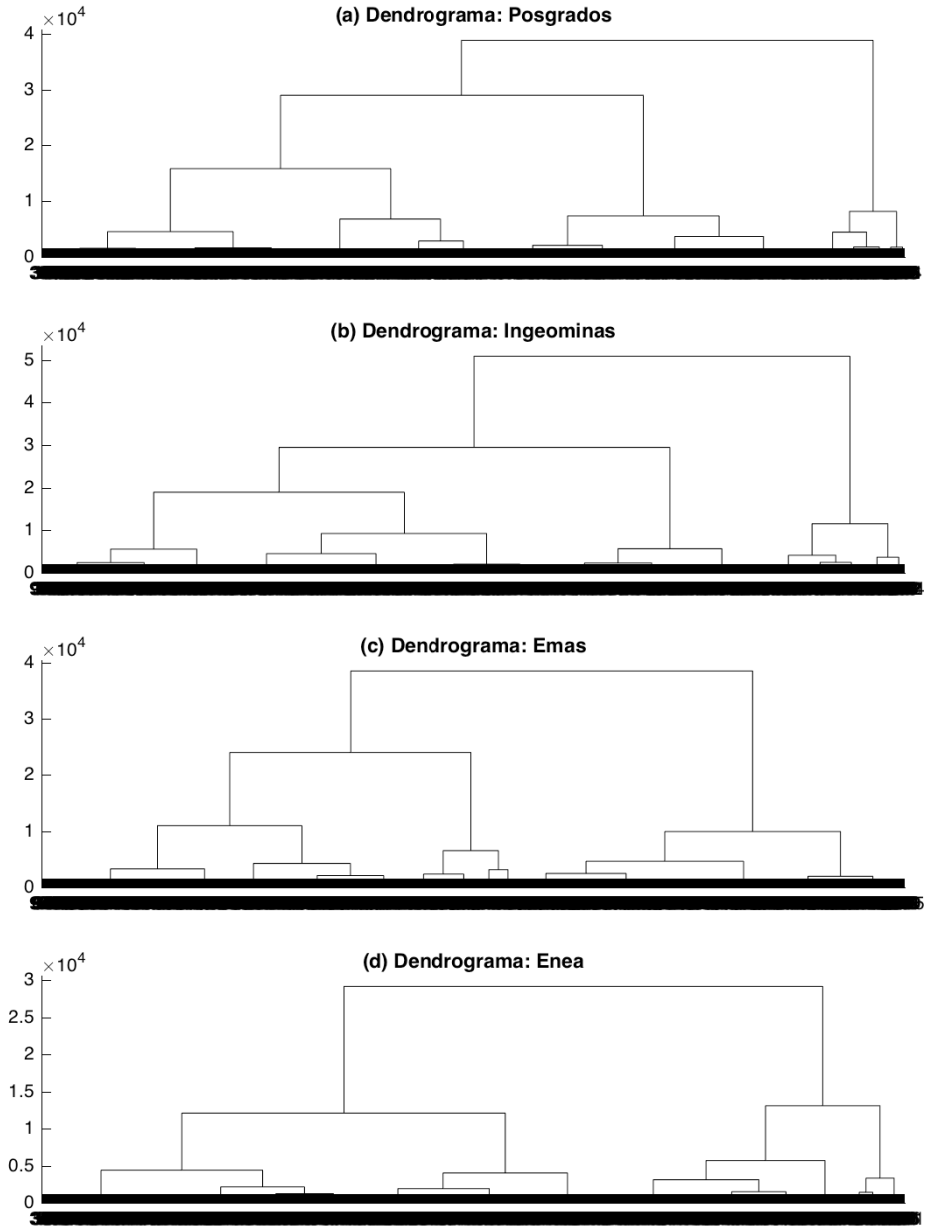


Figura 5-22: Dendrogramas de la variable precipitación usando *Linkage*, con método *ward* y distancia euclidiana, período 2009 – 2011.

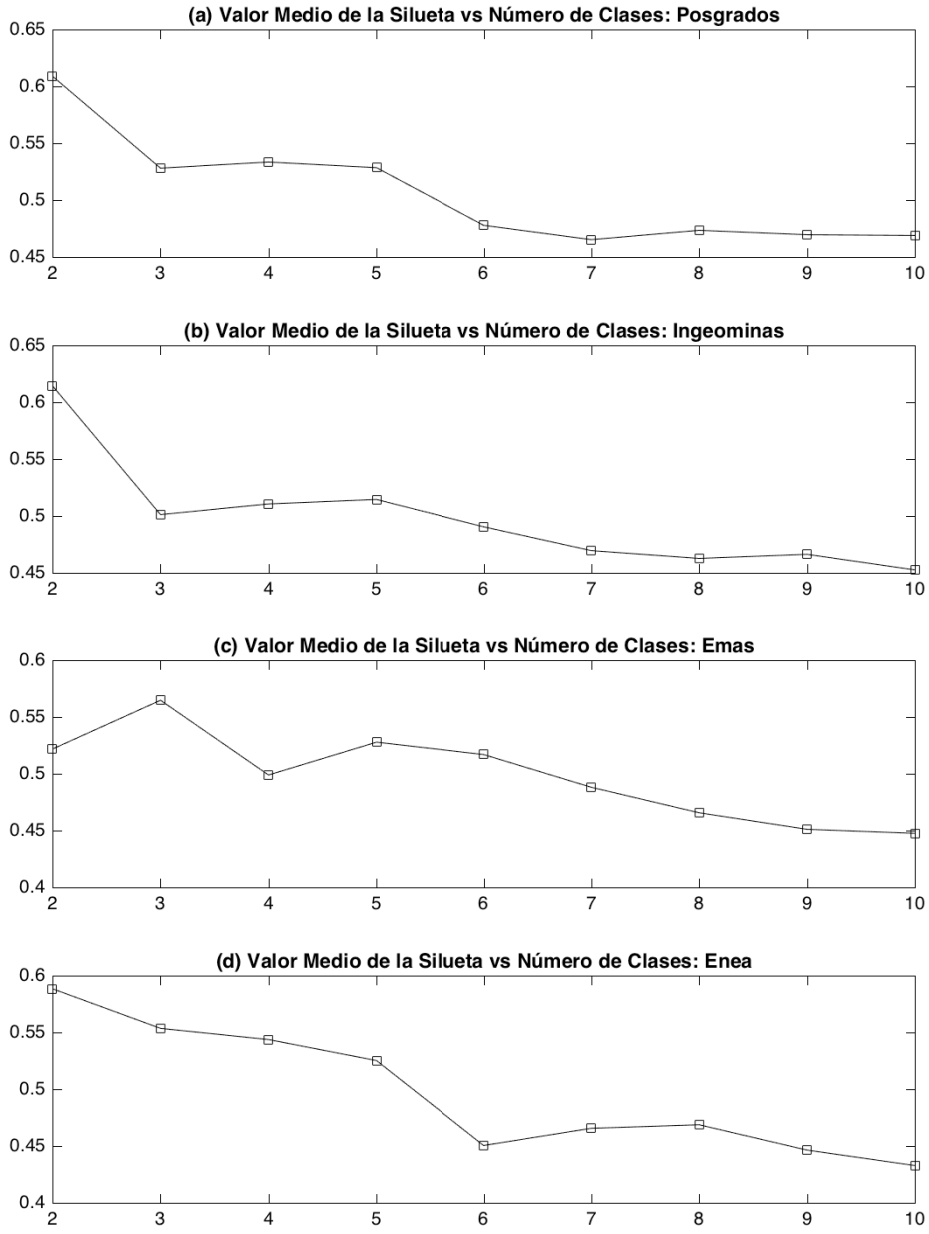


Figura 5-23: Valor medio de silueta en la variable precipitación, período 2009 – 2011.

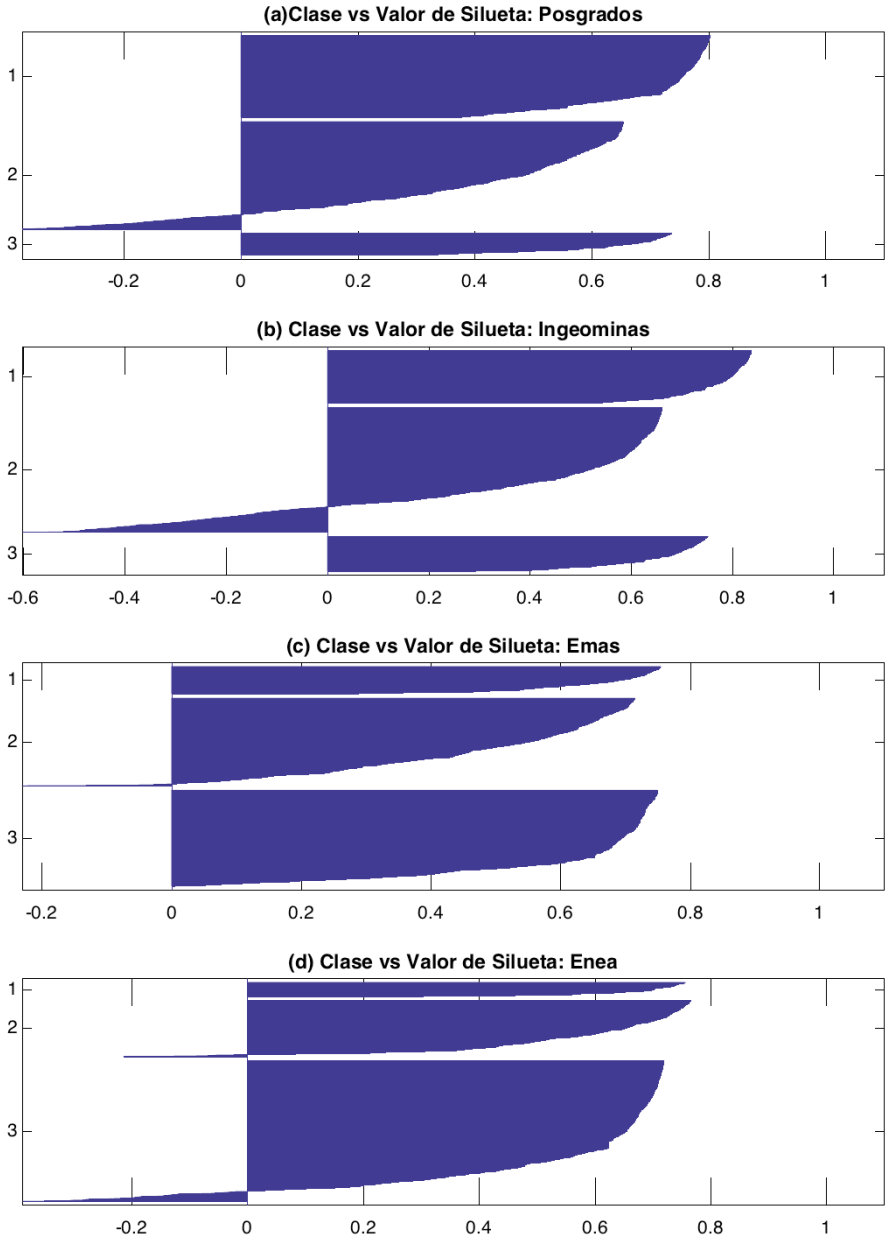


Figura 5-24: Valor de silueta en la variable precipitación para el número de conglomerados escogido, período 2009 – 2011.

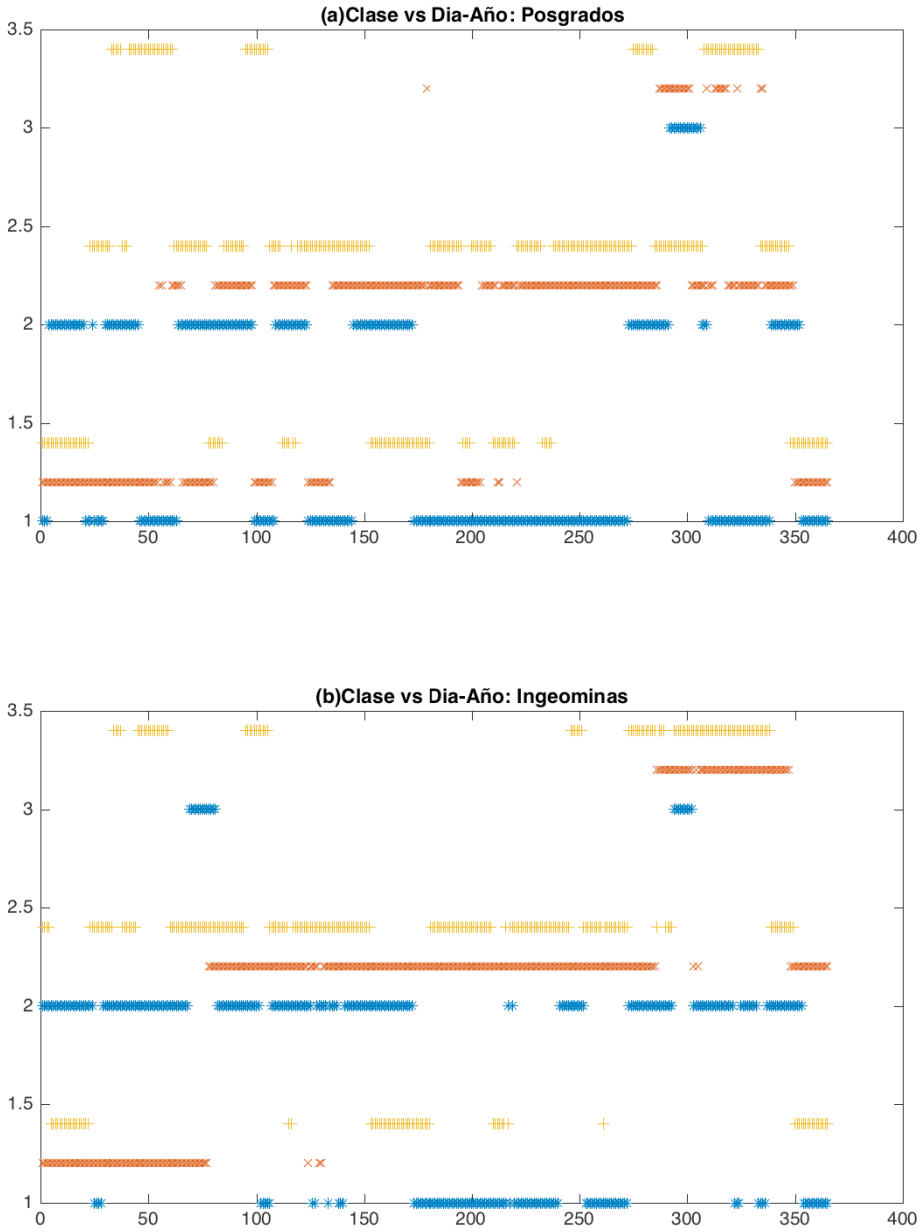


Figura 5-25: Asignación de conglomerado a cada día, para la variable precipitación en las estaciones Posgrados e Ingeominas. Azul: 2009, Rojo: 2010, Amarillo: 2011.

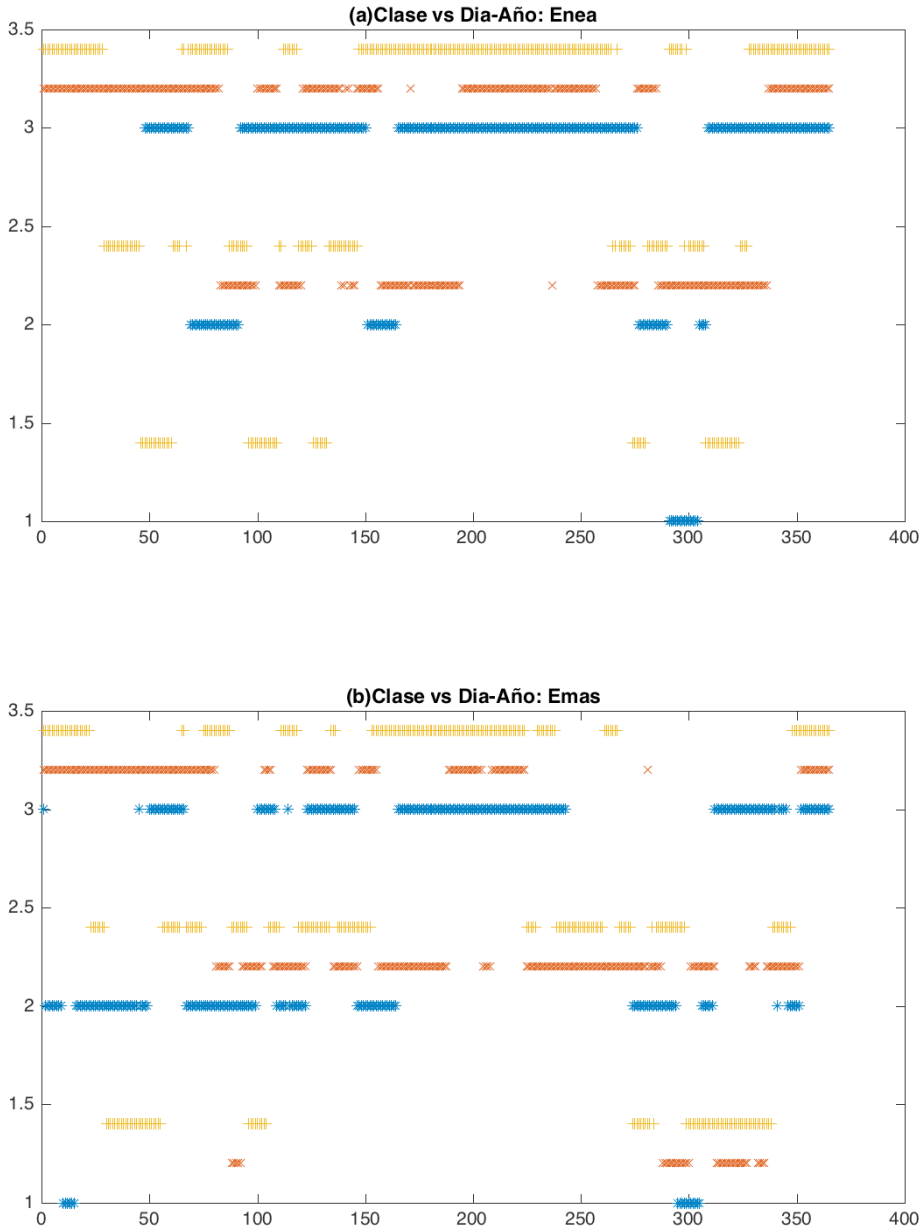


Figura 5-26: Asignación de conglomerado a cada día, para la variable precipitación en las estaciones Enea y Emas. Azul: 2009, Rojo: 2010, Amarillo: 2011.

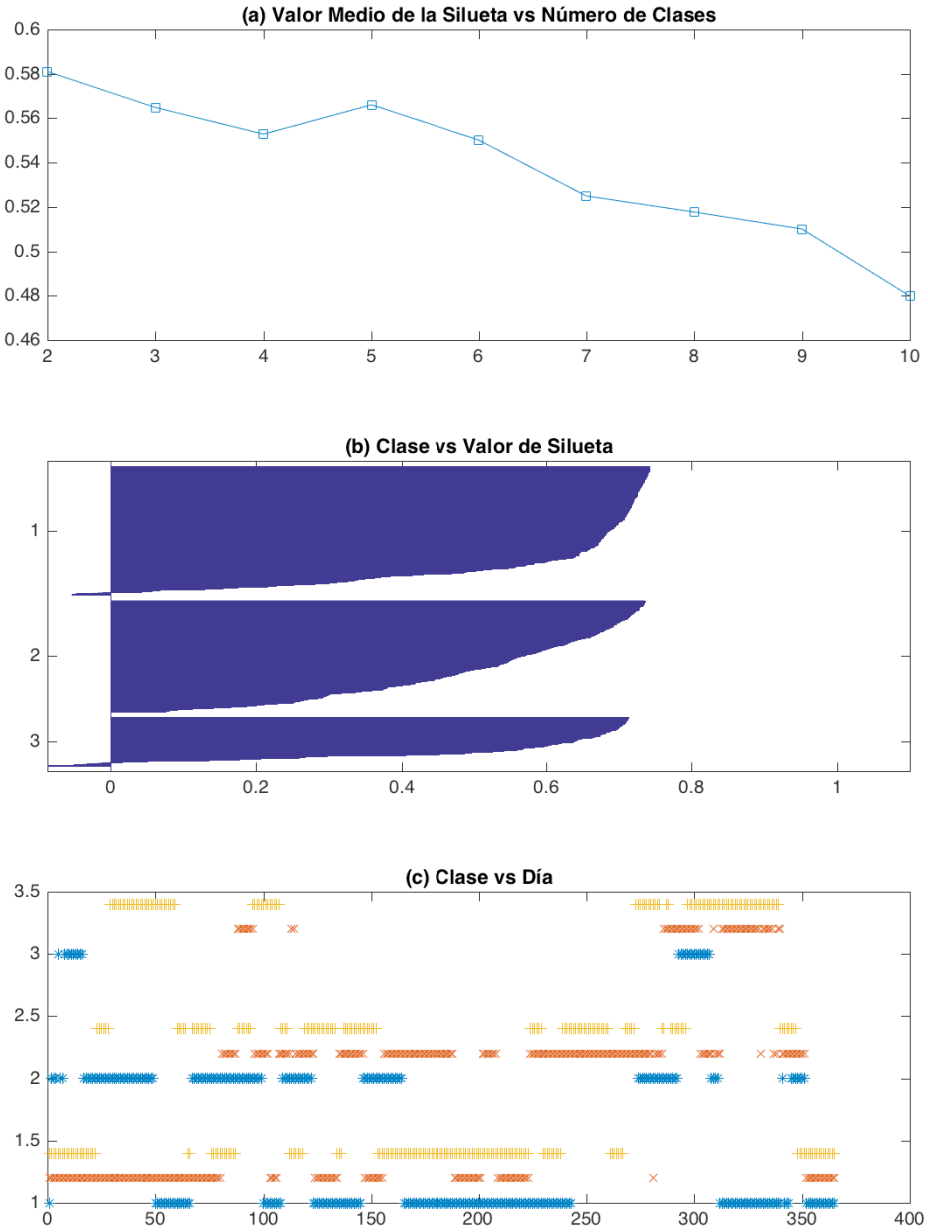


Figura 5-27: Patrones de acumulación de la variable precipitación usando K-medias, con disimilitud *cityblock*, en la estación Emas, período 2009 – 2011.

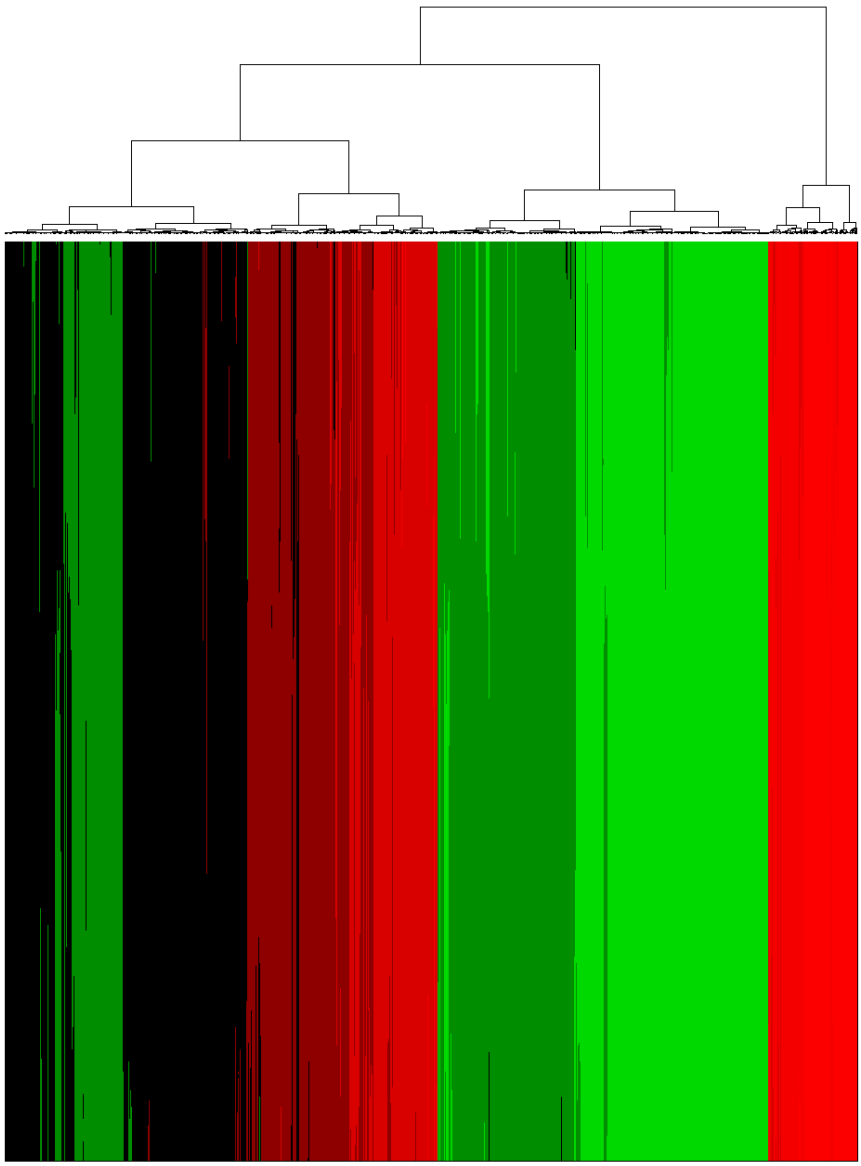


Figura 5-28: Gráfico de *clustergram* para la variable precipitación en la estación Posgrados, período 2009 − 2011.

5.1. Discusión

En esta sección se discuten los resultados de los estudios de agrupamiento para las variables de interés: temperatura, radiación solar, humedad y precipitación. Incluyendo un análisis variable a variable de la relación de los conglomerados encontrados con patrones del clima.

Empezamos con la variable **temperatura**. Para las estaciones Posgrados y Emas el criterio de la silueta plantea 2 conglomerados como la cantidad apropiada, esto se corresponde muy bien con la estructura obtenida en el dendrograma respectivo. Por ello se trabaja con dicha cantidad en tales casos. Para las estaciones Ingeominas y Enea, el criterio silueta también sugiere trabajar con 2 conglomerados, no obstante, a juicio de la autora, la escogencia de 3 conglomerados para trabajar es más acorde al dendrograma disponible (Figuras **5-1**, **5-2**).
De otro lado, al comparar los resultados de clase contra día-año (Figura **5-4**) para las estaciones Posgrados y Emas, se aprecia una correspondencia muy fuerte entre los conglomerados encontrados. Salvo que el agrupamiento etiquetado como 1 en Posgrados está asociado con el 2 de Emas y viceversa. Debe notarse también como el conglomerado 2 de Posgrados (1 en Emas) tiene una alta presencia a principios de 2009 y va perdiendo fuerza a lo largo del año, continúa con poca frecuencia a inicios de 2010 y se afianza a final del año. Pero en 2011, simplemente está disperso.
Adicionalmente, aunque para las estaciones Ingeominas y Enea se encontraron tres conglomerados (Figura **5-5**), se observa una relación importante entre los conglomerados 1 de Ingeominas y 3 de Enea, con el 2 de Posgrados. Lo mismo que entre el 3 de Ingeominas y 1 de Enea, con el conglomerado 1 de Posgrados. El conglomerado 2 de Ingeominas y Enea se correspondería a un patrón propio de esas zonas que persistió a las tendencias marcadas en Posgrados y Emas.
Finalmente, al observar los resultados obtenidos por K-medias para 2 conglomerados en la estación Posgrados los resultados anotados arriba se reafirman. La distribución de los días presentan variaciones pequeñas, pero beneficiosas desde el punto de vista del valor del coeficiente silueta. A su vez, para el `clustergram` se reafirma la conveniencia de las tres clases para la estación Enea. La primera para días con magnitud moderada al principio y cierran con magnitud alta, otra con presencia de magnitudes altas al principio y moderadas al final y la tercera con predominancia de magnitudes altas todo el día.

Pasamos a discutir los resultados de la variable **radiación**. En este caso se trabajó con 2 conglomerados como la cantidad escogida para los análisis, esto guardando correspondencia con la estructura de los dendrogramas y lo sugerido por el criterio silueta.
En seguida pasamos a notar que en la Figura **5-11** se puede observar un relacionamiento entre los conglomerados 1 de las estaciones Posgrados y Enea. En ambos casos este conglomerado tiene presencia débil a principios de 2009, se fortalece a final del año, continúa con presencia importante a principios de 2010 y se debilita a lo largo del año. Para 2011 este conglomerado

se presenta disperso en 2011. Nótese la complementariedad de este comportamiento con lo reportado para la variable temperatura.
Respecto de las estaciones Ingeominas y Emas (Figura **5-12**), podemos decir que el conglomerado 2 se comporta como los conglomerados 1 de Posgrados y Enea, pero la densidad de días asignados a los conglomerados es bastante diferente. Por lo que la intensidad del patrón sería distinta.
La prueba de validación para K-medias devuelve para la estación Enea los mismos dos patrones en el mismo orden (Figura **5-13**), no obstante se presentan correcciones sobre la densidad de puntos pertenecientes a los conglomerados bien soportado en la mejora del gráfico silueta (Figura **5-10**). A su vez, la prueba de `clustergram` sobre la estación Ingeominas muestra un conglomerado asociado a niveles de radiación altos a lo largo de todo el día y el otro con mayor presencia de valores moderados de radiación.

Para la variable **humedad** los resultados de dendrograma y criterio silueta nos conducen a una situación similar al caso temperatura. De hecho la estructura de los conglomerados obtenidos es también muy similar. Nuevamente, la validación con K-medias y `clustergram` es importante y coherente con lo obtenido previamente.

Finalmente, revisados los resultados de dendrograma y criterio silueta para la variable **precipitación acumulada** se optó por trabajar con 3 conglomerados para cada una de las 4 estaciones (Figuras **5-22**, **5-23**). Los conglomerados obtenidos para las estaciones Posgrados e Ingeominas guardan similitudes importantes con respecto a las partes del año en que se presentan a lo largo del período de estudio (Figura **5-25**). Lo mismo para lo encontrado en las estaciones Enea y Emas (Figura **5-26**).
La validación con K-medias sobre la estación Emas resulta bastante exitosa en los conglomerados obtenidos, salvo numeración (Figura **5-27**). Es muy importante observar las similitudes entre este resultado y las agrupaciones obtenidas para temperatura y humedad en las estaciones Ingeominas y Enea.

6 Conclusiones y recomendaciones

6.1. Conclusiones

Se presentan en esta sección las conclusiones obtenidas de los distintos aspectos estudiados en el trabajo, claramente el alcance de estas conclusiones está limitado al período de tiempo observado $(2009-2011)$ y al área de influencia o cobertura dentro de la ciudad de Manizales de las estaciones metereológicas consideradas (Posgrados, Emas, Enea e Ingeominas):

- Se logró estructurar un marco matemático conceptual de soporte a las técnicas y algoritmos de conglomerados jerárquicos aglomerativos, incluyendo el método de vinculación de Haussdorff. Esto involucra la formalización conceptual con definiciones apropiadas y la unificación de nomenclatura. Esta tarea es importante debido a que las distintas técnicas provienen de diversos campos del saber y el contar con un esquema matemático unificado de definiciones, propiedades y nomenclatura facilita y formaliza su estudio. Aunque el marco esté limitado esencialmente a datos de tipo escalar es lo suficientemente amplio para el propósito del trabajo.

- En lo referente a desarrollo tecnológico es importante señalar que si bien el uso del *Statistics Toolbox* de MATLAB® resultó de gran importancia, debe reportarse el desarrollo de rutinas de preproceso que permiten preparar los datos provenientes de la bodega de datos del IDEA para su utilización en el *Statistics Toolbox*. También cabe mencionar que fue preciso desarrollar rutinas macro que implementan el criterio de selección de número de conglomerados y otras para el método de vinculación de Haussdorff, que interactúan de modo conveniente con el *Toolbox*. Estos desarrollos permiten un acople importante de los esfuerzos previos, como bodega de datos, con las herramientas de análisis de conglomerados aquí estudiados y con otras herramientas de MATLAB®.

- El estudio adelantado de análisis de tendencia central y dispersión proporcionó una reducción bastante importante de la complejidad del problema. Esencialmente, la estrategia permitió reducir cada variable a 8 series de tiempo de un único valor diario: mínimo, media - varianza, media, media + varianza, cuartil 1, mediana, cuartil 3 y máximo. Permitiendo concluir varias cosas:

    1. Existe una correlación fuerte (mayor a 0.75) entre las medias diarias de las variables temperatura, radiación y humedad. A su vez el relacionamiento con la variable precipitación resulta bajo.

2. De la variable temperatura podemos decir que: su media diaria satisface criterios de normalidad, la mínima diaria está casi siempre en $13°C$ y la máxima sobre los 26.
3. La variable radiación presenta comportamiento normalizado en la estación Enea. Importante resaltar allí que aunque el 75 % del día la radiación es inferior a $850W/m^2$ el máximo ha de estar por encima de 900, salvo pocos casos. Esto es importante para adoptar medidas de protección UV y potenciales usos en energías renovables [13].
4. En la variable humedad cabe destacar el buen acople de la distribución Weibull a sus datos. Se destaca también que aunque se pueden llegar a registrar valores tan bajos como 30 % ese mismo día puede superar el 80 %.
5. De la precipitación acumulada, se puede comentar que sólo la distribución Weibull logró superar el umbral de aceptación en la prueba de bondad de ajuste.

- Para el análisis de conglomerados se optó por trabajar directamente con la base de datos y no con la simplificación hecha en el estudio de tendencia central. Esto a fin de mantener independencia entre ellos. Hecho el comparativo de desempeño de los distintos métodos jerárquicos, el más exitoso fue el método *ward* (con la distancia euclidiana). Podemos plantear las siguientes conclusiones:
  1. En la variable temperatura se identificaron dos patrones para las estaciones Posgrados y Emas, que guardan una correspondencia alta entre ellos. Para las estaciones Ingeominas y Enea, adicional a los dos patrones ya mencionados se identificó un tercero que afecta la intensidad de los dos marcados en las otras estaciones.
  2. Para la variable radiación se identificaron dos patrones en todos los casos. No obstante, esta vez se observan correspondencia entre las estaciones Posgrados y Enea por un lado y entre Emas e Ingeominas por otro. La diferencia detectada entre un grupo y otro radica escencialmente en la densidad de los días del conglomerado, que podría explicar por una diferencia en la intesidad del patrón en las áreas de influencia de las estaciones.
  3. Para la variable humedad se ratifican los resultados obtenidos para la variable temperatura. Esta cercanía ya había sido evidenciada en el estudio de tendencia central.
  4. En el caso de la precipitación acumulada se identificaron 3 conglomerados para cada estación. Aquí se observó mayor correspondencia entre los pares de estaciones Posgrados e Ingeominas por un lado y Emas y Enea por otro. Las diferencias más fuertes se dan en las asignaciones resultantes para los días de los distintos años.
  5. En todo caso, se detecta la presencia de un patrón climático común en todas las estaciones y todas las variables, en lo referente al período de ocurrencia. Se trata

de un patrón que se va consolidando a lo largo de 2009, arranca con fuerza en 2010 y va perdiendo fuerza, pero en 2011 se encuentra disperso a lo largo del año. Esto podría estar relacionado con el fenómeno de La Niña activo por esos días.

- Las herramientas de validación por K-medias y `clustergram`, brindaron no sólo tranquilidad sobre la pertinencia y consistencia de los resultados obtenidos, sino que proporcionaron redistribuciones de los conglomerados que mejoran la medida de silueta (para K-medias) y facilitan la interpretabilidad de los resultados (para `clustergram`).

- Finalmente, se puede decir que si bien los estudios de tendencia central resultan obligatorios para simplificar el manejo y mejorar el conocimiento de las variables de interés, se logró obtener de ellos información caracterizadora del clima de la ciudad. También se observa que el esquema metodológico propuesto para el estudio de conglomerados, el cual considera: distintas disimilitudes, varios métodos de vinculación, una selección de número de conglomerados que combina un método automático (silueta media) con la percepción del experto de la forma como se estructuran los conglomerados (lectura de dendrograma), acompañados de un esquema de validación apropiado; mostró su efectividad en la identificación de patrones climáticos. Llegando incluso a poderse establecer relaciones entre los conglomerados obtenidos en diferentes estaciones y entre las distintas variables de estudio.

6.2. Recomendaciones

En esta sección se presentan algunas proyecciones de trabajos futuros:

- Debe adelantarse trabajo adicional de modo que se aproveche el conocimiento obtenido sobre las distribuciones que se ajustan a las variables metereológicas del clima en Manizales, las correlaciones identificadas entre ellas y los patrones climáticos obtenidos. Todo, a fin de avanzar en la estructuración de un modelo climático propio para Manizales, permitiendo entre otras cosas hacer regresiones que mejoren la predicción del clima y la gestión de riesgos asociados al mismo.

- Si bien este estudio se adelantó sobre todo el período $2009 - 2011$ y con cuatro estaciones lo que representa un volumen de información muy importante, haciendo ajustes menores a la metodología aquí desarrollada, puede extenderse a períodos mayores y con más estaciones involucradas. A futuro podrían trabajarse también metodologías multicanal y que aprovechen información de georeferenciación. Lo que permitiría poner a prueba los conglomerados encontrados sobre períodos más largos de tiempo e ir en procura de identificar la presencia de fenómenos periódicos o cuasi-periódicos de mayor escala.

- Se debe mejorar la base de datos con la incorporación de datos de radar que vinculen información de circulación atmosférica para una mejor comprensión de la lluvia sobre la ciudad. Esto a su vez presentará retos para incorporar este tipo de datos a los estudios por conglomerados.

Bibliografía

[1] Aparicio-Mijares, F. J.: *Fundamentos de Hidrología de Superficie*. Limusa, 1992

[2] Aristizábal, E. ; Martínez, H. ; Vélez, J. I.: Una revisión sobre el estudio de movimientos en masa detonados por lluvia. En: *Rev. Acad. Colomb. Cienc.* 34 (2010), p. 209–227

[3] Asociados, MetAs & M.: Presión atmosférica, presión barométrica y altitud: Conceptos y aplicaciones. En: *La Guía MetAs* 5 (2005), Nr. 2

[4] Basalto, N. ; Bellotti, R. ; De-Carlo, F. ; Facchi, P. ; Pantaleo, E. ; Pascazio, S.: Hausdorff clustering of financial time series. En: *Physica A* 379 (2007), p. 635–644

[5] Cassou, C. ; Minvielle, M. ; Terray, L. ; Périgaud, C.: A statistical-dynamical scheme for reconstructing ocean forcing in the Atlantic. Part I: weather regimes as predictors for ocean surface variables. En: *Climate Dynamics* 36 (2010), p. 19–39

[6] Cheng, X. ; Wallace, J. M.: Cluster Analysis of the Nothern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns. En: *Department of Atmospheric Sciences* 50 (1991), Nr. 16, p. 2674–2696

[7] Cortés, A. C.: *Análisis de la variabilidad espacial y temporal de la precipitación en una ciudad de media montaña andina*, Universidad Nacional de Colombia - Sede Manizales, Tesis de Grado, 2010

[8] Díaz-Monroy, L. G. ; Morales-Rivera, M. A.: *Análisis estadístico de datos multivariados*. Universidad Nacional de Colombia - Sede Bogotá, Facultad de Ciencias, 2012

[9] Duque-Méndez, N. D. ; Orozco-Alzate, M. ; Vélez-Upegui, J. J.: Hydrometeorological data analysis using OLAP techniques. En: *Dyna* 81 (2013), Nr. 185, p. 160–167

[10] ETESA: *¿Qué es el viento?* http://www.hidromet.com.pa/viento, 11-06-2016

[11] Everitt, B. S. ; Landau, S. ; Leese, M. ; Stahl, D.: *Cluster Analysis*. 5a Ed. Wiley, 2011 (Wiley Series in Probability and Statistics 848)

[12] Hartigan, J. A.: *Clustering Algorithms*. John Wiley & Sons., 1975.

[13] Hernández, J. ; Sáenza, E. ; Vallejo, W. A.: Estudio del Recurso Solar en la Ciudad de Bogotá para el Diseño de Sistemas Fotovoltaicos Interconectados Residenciales. En: *Revista Colombiana de Física* 42 (2010), p. 161–165

[14] Huth, R. ; Beck, C. ; Philipp, A. ; Demuzere, M. ; Ustrnul, Z. ; Cahynová, M. ; Kysely, J. ; Einar-Tveito, O.: Classifications of Atmospheric Circulation Patterns, Recent Advances and Applications. En: *Trends and Directions in Climate Research* 1146 (2008), p. 105–152

[15] IDEA, Unal: *Página Web IDEA*. http://www.idea.manizales.unal.edu.co, Accessed: 02-11-2015.

[16] Izenman, A. J.: *Modern Multivariate Statistical Techniques*. Springer, 2013

[17] Jiang, N.: A new objective procedure for classifying New Zealand synoptic weather types during 1958–2008. En: *Royal Meteorological Society* 31 (2010), p. 863–879

[18] Johnson, A. ; Wang, X. ; Koong, F. ; Xue, M.: Hierarchical Cluster Analysis of a Convection-Allowing Ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the Object-Oriented Cluster Analysis Method for Precipitation Fields. En: *American Meteorological Society* 139 (2011), p. 3673–3693

[19] Kalnay, E. ; Kanamitsu, M ; Kistler, R. ; Collins, W. ; Deaven, D. ; Gandin, L. ; Iredell, M. ; Saha, S. ; White, G. ; Woollen, J. ; Zhu, Y. ; Chelliah, M. ; Ebisuzaki, W. ; Higgins, W. ; Janowiak, J. ; Ropelewski, K. C. ; Wang, J. ; Leetmaa, A. ; Reynolds, R. ; Jenne, R. ; Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project. En: *American Meteorological Society* 77 (1996), Nr. 3, p. 437–471

[20] Martínez, F. ; Troncoso, A. ; Riquelme, J. C. ; Riquelme, J. M.: Partitioning-Clustering Techniques Applied to the Electricity Price Time Series. En: *Intelligent Data Engineering and Automated Learning-IDEAL* (2007), p. 990–999

[21] Mejía, F. ; Londoño, J. P. ; Pachón, J. A.: Red de estaciones metereológicas para prevención de desastres en Manizales - Colombia. En: *Proceedings: Taller internacional sobre gestión del riesgo a nivel local*. Manizales, Colombia, 2006, p. 25pp

[22] Michelangeli, P. A. ; Vautard, R. ; Legras, B.: Weather Regimes: Recurrence and Quasi Stationarity. En: *American Meteorological Society* 52 (1995), Nr. 8, p. 1237–1256

[23] Moreno, H. A. ; Vélez, M. V. ; Montoya, J. D. ; Rhenals, R. L.: La lluvia y deslizamientos de tierra en Antioquia: Análisis de su ocurrencia interanual, intraanual y diaria. En: *Revista EIA* 5 (2006), p. 59–69

[24] Moron, V. ; Robertson, A. W. ; Ward, M. N. ; Ndiaye, O.: Weather Types and Rainfall over Senegal. Part I: Observational Analysis. En: *American Meteorological Society* 21 (2007), p. 266–287

[25] Narvaéz, D. F.: *Análisis de lluvia como elemento detonante en la ocurrencia de movimientos de masa en las comunas Atardeceres y Macarena, sector occidental de la ciudad de Manizales*, Universidad Nacional de Colombia - Sede Manizales, Tesis de Grado, 2007

[26] Pohl, B. ; Fauchereau, N.: The Southern Annular Mode Seen through Weather Regimes. En: *American Meteorological Society* 25 (2011), p. 3336–3354

[27] Rincón, D. F. ; Vélez, J. J. ; Chang, P.: Spatio-temporal description of the rainfall in the andean city of Manizales (Colombia) for storm design. En: *E-proceedings of the 36th IAHR World Congress*. Netherlands, 2015, p. 3pp

[28] Rodríguez, R. M. ; Capa, A. B. ; Portela, A.: *Metereología y Climatología*. FECYT (Fundación Española para la Ciencia y la Tecnología), 2004

[29] Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. En: *Computational and Applied Mathematics* 20 (1986), p. 53–65

[30] Santos, J. A. ; Corte-Real, J. ; Leite, S. M.: Weather Regimes and their connection to the Winter Rainfall in Portugal. En: *Royal Meteorological Society* 25 (2004), p. 33–50

[31] The Mathworks, Inc: *Matlab R2015b Documentation*. http://www.mathworks.com/help/matlab/, Accessed: 02-11-2015.

[32] Uppala, S. M. ; Kallberg, P. W. ; Simmons, A. J. ; Andrae, U. ; Da-Costa, V. ; Fiorino, M. ; Gibson, J. K. ; Haseler, J. ; Hernandez, A. ; Kelly, G. A. ; Li, X. ; Onogi, K. ; Saarinen, S. ; Sokka, N. ; Allan, R. P. ; Andersson, E. ; Arpe, K. ; Balmaseda, M. A. ; Beljaars, A. C. M. ; Van-De-Berg, L. ; Bidlot, J. ; Bormann, N. ; Caires, S. ; Chevallier, F. ; Dethof, A. ; Dragosavac, M. ; Fisher, M. ; Fuentes, M. ; Hagemann, S. ; Hólm, E. ; Hoskins, B. J. ; Isaksen, L. ; Janssen, P. A. E. M. ; Jenne, R. ; Mcnally, A. P. ; Mahfouf, J. F. ; Morcrette, J.J. ; Rayner, N. A. ; Saunders, R. W. ; Simon, P. ; Sterl, A. ; Trenberth, K. E. ; Untch, A. ; Vasiljevi, D. ; Viterbo, P. ; Woollen, J.: The ERA-40 re-analysis. En: *Royal Meteorological Society* 131 (2005), Nr. 612, p. 2961–3012

[33] Vélez-Upegui, J. J. ; Duque-Méndez, N. D. ; Mejía-Fernández, F. ; Orozco-Alzate, M.: Red de monitoreo climático para dar apoyo a la prevención y atención de desastres en Manizales, Colombia. En: *Proceedings: III Congreso de Meteorología Tropical*. La Habana, Cuba, 2012, p. 12pp

[34] Vovoras, D. ; Tsokos, C. P.: Statistical analysis and modeling of precipitation data. En: *Nonlinear Analysis* 71 (2009), p. 1169–1177

[35] WMO: *Guide to Meteorological Instruments and Methods of Observation*. 7a. World Metereological Organization, 2008

[36] Wu, J.: *Advances in K-Means Clustering*. Berlin : Springer, 2012

[37] Wu, X. ; Kumar, V. ; Quinlan, J. R. ; Ghosh, J. ; Yang, Q. ; Motoda, H. ; McLachlan, G. J. ; Ng, A. ; Liu, B. ; Yu, P. S. ; Zhou, Z-H. ; Steinbach, M. ; Hand, D. J. ; Steinberg, D.: Top 10 algorithms in data mining. En: *Knowl Inf Syst* 14 (2008), p. 1–37