UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

"TÉCNICAS DE AGRUPAMIENTO PARA LA REGIONALIZACIÓN DE CAUDALES EN LA MIXTECA OAXAQUEÑA"

T E S I S

PARA OBTENER EL TÍTULO DE
INGENIERO EN COMPUTACIÓN

PRESENTA:
FELIX EMILIO LUIS PÉREZ

DIRECTOR DE TESIS:
DR. RAÚL CRUZ BARBOSA

ASESOR DE TESIS:
M.I. GABRIELA ÁLVAREZ OLGUÍN

Huajuapan de León, Oaxaca, Junio, 2012.

II

A mis padres.

Por hacerme un ser útil.

A Jean.

Por su apoyo, motivación y cariño.

Agradecimientos

Agradezco principalmente a mi director de tesis, el Dr. Raúl Cruz Barbosa, por el tiempo y la paciencia dedicada a las revisiones de este trabajo, así como su apoyo y consejos en el mejoramiento del mismo.

Agradezco también a mis sinodales, la M.C. Verónica Rodríguez López, el Dr. Antonio Orantes Molina y la Dra. Lluvia Carolina Morales Reynaga por su tiempo y disposición al revisar este trabajo, así como su valiosa contribución en el mejoramiento del mismo.

A los profesores, el Ing. Moisés Emmanuel Ramírez Guzmán y el Dr. Ricardo Pérez Águila por sus importantes observaciones durante el registro de este trabajo de tesis.

A los doctores Vitaliy Rybak y Felipe de Jesús Trujillo Romero, por permitirme utilizar sus horas de trabajo y clases, respectivamente, para culminar el presente proyecto de tesis.

A la Universidad Tecnológica de la Mixteca, por permitirme cursar y concluir mis estudios de licenciatura en los diferentes espacios de la misma.

A mis profesores, en especial aquellos que hacían su trabajo con verdadero entusiasmo.

A mis padres, porque sin su apoyo y cariño no hubiese terminado mis estudios, porque siempre han estado y me han impulsado en los momentos más difíciles de mi carrera, porque me enseñaron que el trabajo continuo genera mejores resultados, me enseñaron a ser una persona responsable, útil e insistente en mis objetivos. Gracias papá por esos llamados de atención que me hacían recordar el trabajo cuando la pereza iba ganando terreno. Gracias mamá porque con tu cariño y comprensión los regaños de papá eran menos duros. Gracias a ambos por ese cariño incondicional que me han brindado.

A mis hermanos Fanny y Fabian, porque ellos han sido la parte distractora en mis estudios, esa parte que cuando estoy preocupado por algo, un simple ¡ash manito! o ¡vamos a jugar!, son suficientes para desviar mi atención y olvidar por un momento los pendientes.

A Jean, porque siempre me ha impulsado a alcanzar mis objetivos, porque cuando necesito que alguien me escuche sé que puedo contar con ella, porque gracias a su cariño y comprensión puedo darme valor para enfrentar nuevos retos, porque cuando decía, ¿ya fuiste a ver a Barbosa?, y yo no sabía que contestar, ella sólo movía la cabeza para reprobar mis acciones y era suficiente para ponerme a trabajar de nuevo en la tesis.

A todas y cada una de las personas que se vieron involucradas, directa o indirectamente, con la realización de este trabajo, **gracias**.

Felix Emilio Luis Pérez.

Junio, 2012.

Índice general

Agradecimientos v

Publicación derivada xiii

1. Introducción 1
- 1.1. Planteamiento del problema . . . 3
- 1.2. Objetivos . . . 4
  - 1.2.1. Objetivo general . . . 4
  - 1.2.2. Objetivos específicos . . . 4
- 1.3. Organización de la tesis . . . 4

2. Técnicas de agrupamiento 5
- 2.1. Introducción . . . 5
- 2.2. Preprocesamiento: Pruebas de calidad de los datos . . . 8
  - 2.2.1. Prueba de independencia y aleatoriedad . . . 9
  - 2.2.2. Prueba de homogeneidad . . . 10
  - 2.2.3. Prueba de outliers . . . 11
- 2.3. Agrupamiento jerárquico . . . 12
  - 2.3.1. Características del agrupamiento jerárquico . . . 13
  - 2.3.2. Funciones de enlace entre grupos . . . 14
- 2.4. Agrupamiento difuso . . . 16
  - 2.4.1. Propiedades de los conjuntos difusos . . . 17
  - 2.4.2. Técnicas de agrupamiento difuso . . . 17
  - 2.4.3. Defusificación . . . 18
  - 2.4.4. Validación de *clusters* . . . 19
- 2.5. Postprocesamiento: Test de heterogeneidad de grupos . . . 20

3. Regresión lineal 23
- 3.1. Introducción . . . 23
- 3.2. Regresión lineal múltiple . . . 23
  - 3.2.1. Ajuste del modelo a los datos . . . 24
  - 3.2.2. Estimación de σ^2 . . . 27
- 3.3. Selección de variables . . . 28

3.3.1. Criterios para evaluar modelos de regresión lineal con subconjuntos de variables . . . 29
3.3.2. Técnicas computacionales de selección de variables . . . 31

4. Regionalización de caudales **35**
4.1. Introducción . . . 35
4.2. Caso de Estudio . . . 35
4.2.1. 5 estaciones hidrométricas . . . 39
4.2.2. 10 Estaciones hidrométricas . . . 40
4.3. Diseño experimental . . . 40
4.3.1. Esquema general . . . 42
4.3.2. Base de datos . . . 43
4.4. Implementación del sistema de regionalización . . . 44

5. Resultados **47**
5.1. Resultados experimentales . . . 47
5.2. Análisis de resultados de agrupamiento . . . 60
5.3. Modelos de estimación de caudales . . . 62

6. Conclusiones y Perspectivas **67**
6.1. Conclusiones . . . 67
6.2. Perspectivas . . . 68

Bibliografía **71**

A. Variables utilizadas en el proceso de agrupamiento **77**
A.1. Variables hidrológicas . . . 77
A.2. Variables climáticas . . . 78
A.3. Variables fisiográficas . . . 78

B. Pseudocódigo de algunos algoritmos de agrupamiento **83**
B.1. Agrupamiento jerárquico . . . 83
B.2. Agrupamiento Fuzzy C-Means . . . 84

C. Manual de usuario del software desarrollado **87**

Índice de figuras

2.1. Ejemplo de agrupamiento de un conjunto de datos (a) en 7 grupos distintos (b). 6
2.2. Clasificación de los algoritmos de agrupamiento. 7

4.1. Localización geográfica de la zona de estudio. 38
4.2. Esquema general de desarrollo para la regionalización de caudales en la Mixteca Oaxaqueña. 42

5.1. Agrupamiento del primer conjunto de prueba. 53
5.2. Agrupamiento del segundo conjunto de prueba. 53
5.3. Agrupamiento jerárquico con 5 estaciones y enlace promedio. 54
5.4. Agrupamiento jerárquico con 5 estaciones y enlace centroide. 55
5.5. Agrupamiento jerárquico con 5 estaciones y enlace *ward*. 55
5.6. Agrupamiento jerárquico con 10 estaciones y enlace promedio. 56
5.7. Agrupamiento jerárquico con 10 estaciones y enlace centroide. 56
5.8. Agrupamiento jerárquico con 10 estaciones y enlace *ward*. 57

C.1. Ejecución de la prueba de independencia para la estación Apoala. 89
C.2. Ejecución del algoritmo jerárquico para el caso de 5 estaciones. 91
C.3. Ejecución de la prueba de heterogeneidad con el primer grupo formado usando 5 estaciones hidrométricas. 92
C.4. Ejecución de mínimos cuadrados para obtener el modelo de regresión lineal del primer grupo usando 5 estaciones hidrométricas. 93

Índice de cuadros

2.1. Valores críticos para una distribución normal estándar. 9
2.2. Valores críticos de la prueba de Grubbs y Beck. 12

3.1. Datos para una regresión lineal múltiple. 25
3.2. Clasificación de los coeficientes de determinación múltiple. 30

4.1. Estaciones hidrométricas en la Mixteca. La columna Periodos indica los registros disponibles actualmente. 36
4.2. Registro de caudales en la Mixteca Oaxaqueña. 39
4.3. Caso de estudio con 5 estaciones hidrométricas. 40
4.4. Caso de estudio con 10 estaciones hidrométricas. 41
4.5. Variables climáticas y fisiográficas del estudio. 44

5.1. Resultados de la prueba de independencia. 48
5.2. Resultados de la prueba de homogeneidad. 49
5.3. Outliers encontrados para el caudal máximo. 50
5.4. Outliers encontrados para el caudal mínimo. 50
5.5. Outliers encontrados para el caudal medio. 51
5.6. Outliers encontrados para la lluvia media. 51
5.7. Agrupamiento Fuzzy C-Means con 2 grupos para 5 estaciones. 57
5.8. Agrupamiento Fuzzy C-Means con 3 grupos para 5 estaciones. 57
5.9. Agrupamiento Fuzzy C-Means con 4 grupos para 5 estaciones. 58
5.10. Agrupamiento Fuzzy C-Means con 5 grupos para 5 estaciones. 58
5.11. Validación de grupos para el caso de 5 Estaciones 58
5.12. Agrupamiento Fuzzy C-Means con 2 grupos para 10 estaciones. 59
5.13. Agrupamiento Fuzzy C-Means con 3 grupos para 10 estaciones. 59
5.14. Agrupamiento Fuzzy C-Means con 4 grupos para 10 estaciones. 59
5.15. Agrupamiento Fuzzy C-Means con 5 grupos para 10 estaciones. 60
5.16. Validación de grupos para el caso de 10 Estaciones 60
5.17. Prueba de heterogeneidad del primer grupo con 5 estaciones. 61
5.18. Prueba de heterogeneidad del segundo grupo con 5 estaciones. 61
5.19. Prueba de heterogeneidad del primer grupo con 10 estaciones. 62
5.20. Prueba de heterogeneidad del segundo grupo con 10 estaciones. 62

Publicación derivada

Como resultado parcial del presente trabajo, se publicó el siguiente artículo:

1. Luis-Pérez, F. E., Cruz-Barbosa, R. y Álvarez-Olguin, G. (2011). Regional Flood Frequency Estimation for the Mexican Mixteca Region by Clustering Techniques. *10th Mexican International Conference on Artificial Intelligence (MICAI 2011). Advances in Soft Computing, Lecture Notes in Artificial Intelligence*, Volumen 7095, páginas 249-260. ISBN: 978-3-642-25329-4. ISSN: 0302-9743.

En la publicación se muestran los resultados de un análisis de regionalización de caudales, en el que se involucraron 10 estaciones hidrométricas y 7 años de registro de caudales. La calidad de los datos se verificó por medio de análisis de aleatoriedad, independencia y outliers; en cada estación de medición. Así mismo, se obtuvieron regiones homogéneas con los algoritmos de agrupamiento jerárquico y Fuzzy C-Means. Se concluyó que ambos métodos coinciden en la formación de regiones, no obstante, se calcularon los índices de validación de *clusters* para identificar el número óptimo de grupos homogéneos.

Una vez delimitadas las regiones homogéneas, se obtuvieron modelos de estimación de caudales para cada región encontrada. En éste caso, el método que se ocupó para construir modelos de estimación de caudales, fue regresión lineal múltiple con selección de variables por agregación (forward selection).

Capítulo 1

Introducción

En términos de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad en México (CONABIO), regionalización se define como la división de un territorio en áreas de menor tamaño con características similares. Se utiliza como una herramienta metodológica para la planificación de actividades ambientales, pues permite el conocimiento de los recursos naturales en el área de estudio (CONABIO, 2008).

La regionalización de caudales, por su parte, permite la estimación de flujos de agua en una o más áreas hidrográficas (cuencas) con características similares. Generalmente, estas áreas se caracterizan por ser no aforadas, dicho de otra manera, no poseen mediciones directas a través de una estación hidrométrica (Nathan y McMahon, 1990).

En las estaciones hidrométricas se registran los caudales mínimos, medios y máximos que fluyen por un punto determinado de la cuenca. Esta información hidrológica, permite cuantificar la oferta hídrica de la cuenca, y estimar los caudales máximos para el diseño de obras hidráulicas en la región (Erazo, 2004).

De acuerdo con (Ouarda y otros, 2008), los métodos de regionalización de caudales involucran dos etapas principales. En la primer etapa, se identifican las cuencas hidrológicamente homogéneas y con ellas se forman las regiones de estudio. En la segunda etapa, se aplica un método de estimación regional de caudales para cada una de las regiones encontradas.

La búsqueda de regiones homogéneas, es el paso más importante en el proceso de regionalización (Smithers y Schulze, 2001), en él se determinan las regiones de características similares y consecuentemente las cuencas que pueden regirse por una misma ecuación de estimación de caudales. En 1990, (Nathan y McMahon, 1990) encontraron que las regiones homogéneas definidas por similitudes hidrológicas, físicas y climáticas de las cuencas, no necesariamente deben ser vecinas geográficamente.

En este sentido, existen diferentes técnicas de homogeneización de datos que pueden adaptarse al proceso de regionalización de caudales. Entre las más utilizadas se encuentran:

el análisis de correlación canónica (Ouarda y otros, 2001) y el análisis de grupos o *clustering* (Jain y otros, 1999). Dentro del análisis de *clusters*, el método más utilizado para este propósito es el agrupamiento jerárquico debido a su facilidad de uso.

Algunos ejemplos de trabajos relacionados de la homogeneización de datos se presentan a continuación.

Como ejemplos del análisis de correlación canónica para la delimitación de regiones homogéneas, se pueden mencionar los estudios realizados por (Shih-Min y otros, 2002) y (Leclerc y Ouarda, 2007). En el primer estudio se utilizaron estaciones hidrométricas de los estados de Alabama, Georgia y Mississippi en los Estados Unidos de Norteamérica. Por su parte, en el segundo caso se utilizaron 29 estaciones hidrométricas localizadas al sureste de Canadá y noreste de los Estados Unidos para delimitar las regiones hidrológicamente homogéneas.

En contraste, en (Jingyi y Hall, 2004) los autores utilizan el enlace tipo *ward* del agrupamiento jerárquico, el método Fuzzy C-Means y las redes neuronales de Kohonen para delimitar las regiones hidrológicamente homogéneas en el sureste de China. Un método alternativo es usado en (Chang y Donald, 2003), donde se utiliza el método de agrupamiento *k-means* con una selección de cuencas en Gran Bretaña, para realizar el correspondiente análisis de regionalización de caudales.

En México, una investigación importante en regionalización de caudales fue realizada por (Ouarda y otros, 2008). Los autores tomaron como casos de estudio los ríos Balsas, Lerma y Pánuco con sus correspondientes cuencas hidrográficas. Se utilizaron cuatro diferentes técnicas para delimitar las regiones hidrológicamente homogéneas: el análisis de agrupamiento jerárquico (Jain y otros, 1999), el análisis de correlación canónica (Muirhead, 1982), la versión modificada de correlación canónica (Girard, 2001) y el método de interpolación canónica o *kriging* (Chokmani y Ouarda, 2004).

De manera local, en el año 2008 fue realizada una primera regionalización hidrológica en la Mixteca Oaxaqueña (Hotait-Salas, 2008). La delimitación de regiones homogéneas fue determinada por los métodos de agrupamiento jerárquico y Andrews (Andrews, 1972). Como resultado se obtuvieron tres regiones hidrológicas homogéneas, para cada una se obtuvo el correspondiente modelo de estimación de caudales. Sin embargo, en el estudio se hicieron suposiciones de información inexistente en los registros mensuales de las variables hidrológicas utilizadas. Se utilizaron sólo las estaciones que presentaban un máximo de tres meses sin registro de caudales y la falta de dicha información fue solventada por valores medios obtenidos a partir de los demás años.

En el presente trabajo, se utilizan dos enfoques diferentes de *clustering* para encontrar las regiones homogéneas. Como primer método se utiliza el agrupamiento jerárquico aglomerativo, el cual asigna a cada una de las cuencas un único grupo de estudio. El segundo enfoque es el agrupamiento Fuzzy C-Means (Bezdek, 1981), con este último las cuencas mantienen diferentes grados de pertenencia a cada una de las regiones preestablecidas por el método.

1.1. Planteamiento del problema

En áreas donde el abasto de agua es insuficiente para cubrir las demandas de la sociedad, la evaluación de flujos de agua es un factor principal para administrar y optimizar el uso de la misma, tal es el caso de la región Mixteca en el estado de Oaxaca. La Mixteca Oaxaqueña, ubicada al sureste de la república Mexicana, es una región de relieve montañoso que se caracteriza por presentar problemas de escasez de agua, por tal razón en la zona, la correcta evaluación de la disponibilidad de recursos hídricos es un factor principal para administrar y optimizar el aprovechamiento de los mismos. En el caso de los flujos de agua superficial, conocer su disponibilidad implica determinar el volumen de escurrimiento de las cuencas, no obstante la falta de información hidrométrica para cuantificar esta variable es muy común. De acuerdo con los datos contenidos en el Sistema de Información de Aguas Superficiales (IMTA, 1997), a partir de 1940, se instalaron en la región 13 estaciones de aforo, de las cuales, sólo una se encuentra en operación actualmente; aunado a esto, diversas zonas y cuencas no cuentan con el registro histórico de sus caudales. Por tanto, al haber menos estaciones funcionando, la cantidad y calidad de caudales registrados en la región disminuye y en consecuencia, las evaluaciones hidrológicas que se realizan a partir de estos datos tienen menor fiabilidad.

Por su parte y en gran medida, las características físicas de las cuencas se deben a la actividad del agua sobre éstas, por lo que es lógico pensar en una fuerte relación entre las características físicas y geográficas de la cuenca y las variables que describen el comportamiento hidrológico de la misma.

La regionalización de caudales permite determinar la relación existente entre las características físicas, geográficas y climáticas de una cuenca con los caudales máximos, mínimos y medios asociados a ella en diferentes periodos estacionales. Además, por medio de la regionalización es posible estimar los volúmenes de escurrimiento generados para una región homogénea, sin necesidad de acudir a las mediciones de campo; no obstante, dichas estimaciones estarán limitadas por el grado de ajuste que presenten los modelos regionales a los datos históricos de dicha región.

El objetivo de este trabajo es identificar regiones hidrológicamente homogéneas en la Mixteca Oaxaqueña mediante la aplicación de métodos de agrupamiento, específicamente los métodos de agrupamiento jerárquico aglomerativo y Fuzzy C-Means. También, para cada una de las regiones encontradas se deberá obtener un modelo regional de estimación de caudales máximos, mínimos y medios, utilizando análisis de regresión lineal múltiple con selección de variables por agregación. Cabe mencionar que los resultados obtenidos en el presente estudio, servirán como punto de partida para evaluar la disponibilidad de agua superficial en la región.

1.2. Objetivos

1.2.1. Objetivo general

Obtener la regionalización de caudales para la Mixteca Oaxaqueña, mediante la utilización de técnicas de agrupamiento para la delimitación de regiones hidrológicamente homogéneas.

1.2.2. Objetivos específicos

- Realizar pruebas de calidad de los datos a utilizar.
- Implementar los algoritmos de agrupamiento jerárquico aglomerativo y Fuzzy C-Means, para aplicarlos a la homogeneización de regiones hidrológicas.
- Comparar la influencia de los diferentes algoritmos de agrupamiento en la homogeneización de las regiones.
- Obtener modelos de regresión lineal para la estimación de caudales máximos, mínimos y medios en cada una de las regiones homogéneas encontradas.

1.3. Organización de la tesis

El presente documento se encuentra dividido, para un mejor entendimiento, en seis capítulos principales.

En el capítulo 2, se describen las técnicas de agrupamiento empleadas para determinar regiones hidrológicamente homogéneas. De igual manera, se presenta la teoría correspondiente a las pruebas de calidad para el preprocesamiento de datos y las respectivas pruebas de postprocesamiento.

Los conceptos necesarios para determinar los modelos de estimación regional de caudales, haciendo énfasis en la regresión lineal múltiple y selección de variables con métodos paso a paso, son descritos en el capítulo 3.

Los detalles de diseño y desarrollo del proyecto, son presentados en el capítulo 4. En el capítulo 5 se plasman los resultados obtenidos en la delimitación de regiones homogéneas y sus correspondientes modelos de estimación de caudales.

Finalmente, en el capítulo 6 se proporcionan las conclusiones y el posible trabajo a futuro de esta tesis.

Capítulo 2

Técnicas de agrupamiento

2.1. Introducción

El problema de formar grupos en un conjunto de datos es muy importante para determinar el comportamiento de una población, de la cual sólo se tiene una cantidad n de sus elementos. En este sentido, los algoritmos de agrupamiento proveen un mecanismo para organizar grandes cantidades de información en pequeños grupos similares, en cuanto a su contenido se refiere (Jain y Dubes, 1988).

El análisis de agrupamiento, también llamado *clustering* (en inglés), tiene por objetivo la distribución de los elementos de un conjunto de datos en grupos homogéneos, en función de las similitudes entre ellos (Jain y otros, 1999). Por tanto e intuitivamente, los objetos dentro de un grupo, son más similares entre ellos que con objetos pertenecientes a otro grupo. Un ejemplo de agrupamiento es mostrado en la Figura 2.1, donde los datos de entrada son graficados en la Figura 2.1(a), y los *clusters* o grupos de salida se presentan en la Figura 2.1(b). Se puede observar que los puntos que pertenecen al mismo grupo son identificados con un mismo nombre o una misma etiqueta.

En nuestro caso, es importante entender la diferencia entre agrupamiento y clasificación supervisada. En la clasificación supervisada, se provee de una colección de observaciones etiquetadas o pre-clasificadas; el problema consiste entonces en asignar una de las etiquetas conocidas a cada nueva observación de entrada, haciendo algún tipo de comparación entre la observación de entrada y la colección de datos conocida. En el agrupamiento, el problema consiste en identificar dentro de una colección de observaciones no etiquetadas, uno o varios grupos (*clusters*) de observaciones con características similares (Jain y otros, 1999).

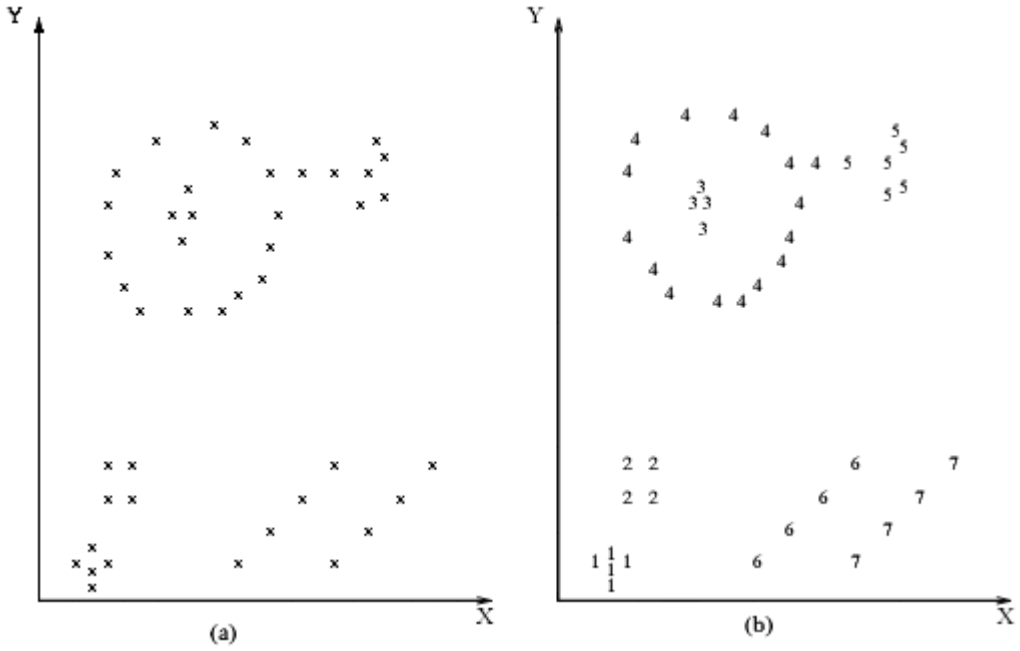

Figura 2.1: Ejemplo de agrupamiento de un conjunto de datos (a) en 7 grupos distintos (b).

Normalmente, se agrupan las observaciones de un experimento, sin embargo, el análisis de agrupamiento también puede aplicarse para separar las variables o atributos del conjunto de observaciones (Peña, 2002). En general, el análisis de *clusters* estudia tres tipos diferentes de problemas:

1. Partición de datos. Se presenta cuando se dispone de datos con sospecha a ser heterogéneos y se desea dividirlos en un número de grupos prefijado de manera que, cada elemento pertenezca a uno y sólo uno de los grupos. Así, todo elemento queda agrupado y cada grupo es internamente homogéneo.

2. Construcción de jerarquías. En estos casos se desean estructurar los elementos de un conjunto de forma jerárquica por su similitud. Una agrupación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores. Estrictamente estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos.

3. Agrupación de variables. En casos de estudio donde se tienen muchas variables, es interesante un análisis para dividir estas variables en grupos. Dicho estudio puede plantear modelos formales para reducir la dimensionalidad de los casos, de esta manera, las variables pueden agruparse o estructurarse en forma de jerarquías utilizando una matriz de relación entre variables.

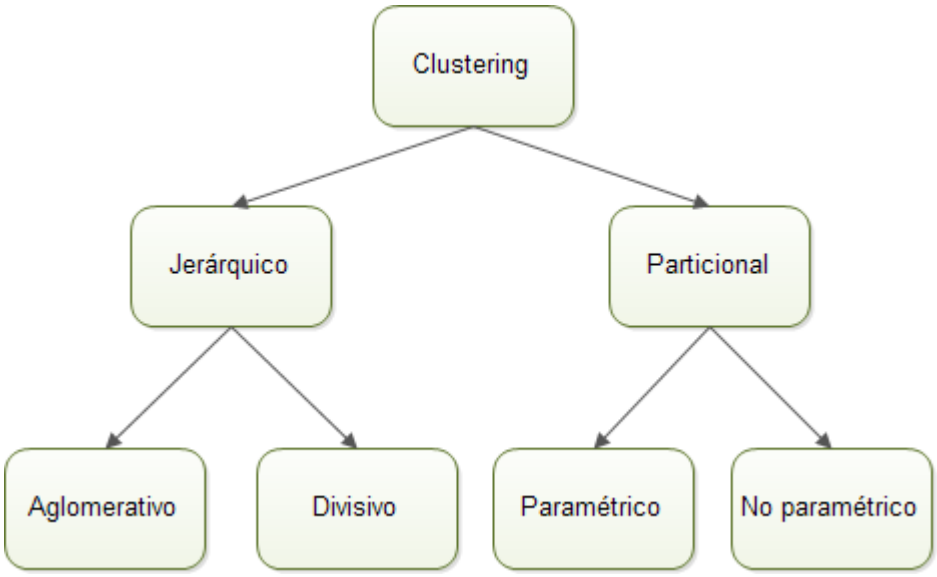

Figura 2.2: Clasificación de los algoritmos de agrupamiento.

La aplicación de estas técnicas de *clustering* son muy diversas y entre las más destacadas se pueden mencionar: toma de decisiones, aprendizaje no supervisado (agrupamiento y reducción de la dimensionalidad), minería de datos, segmentación de imágenes, entre otras.

Los métodos de *clustering* se pueden dividir en jerárquicos y particionales, como se muestra en la Figura 2.2 (Downs y Barnard, 2003). Los primeros construyen un árbol jerárquico de particiones anidadas de datos, donde cualquier seccionamiento del árbol a cierto nivel produce una partición específica de los datos. Éstos, a su vez pueden subdividirse en aglomerativos y divisivos dependiendo de la forma en que se genere el árbol de agrupamiento. Si todos los datos (items) comienzan en un solo grupo, y en cada paso se van separando en grupos de items hasta llegar a grupos que contienen un solo item, el agrupamiento es conocido como jerárquico divisivo. De otro modo, si los datos comienzan separados en grupos de un solo item y en cada paso se unen grupos similares hasta llegar a un solo grupo conteniendo todos los items, se dice que se realiza un proceso jerárquico aglomerativo.

Por su parte, los métodos de *clustering* particionales, a pesar de su variedad tienen como característica común, que requieren de entrada el número de grupos en que se particionarán los datos. Estas técnicas, se pueden dividir en métodos paramétricos y no paramétricos. Los primeros asumen que cada *cluster* de datos es descrito por una distribución de probabilidad específica, en cambio, los métodos no paramétricos tratan de encontrar particiones que optimicen un criterio de agrupamiento predefinido.

De manera general, en el presente trabajo retomaremos el problema de partición de datos, el cual agrupa un conjunto de observaciones en un número determinado de grupos (*clusters*).

Este agrupamiento toma como base la distancia o similitud entre cada una de las observaciones y de esta manera genera los grupos homogéneos de datos. Nos enfocaremos principalmente en los algoritmos de agrupamiento jerárquico y Fuzzy C-Means como técnicas para la delimitación de regiones hidrológicamente homogéneas. Estos métodos son ampliamente utilizados en problemas de regionalización de caudales, tal como se muestra en (Ouarda y otros, 2008; Guler y Thyne, 2004; Srinivas y otros, 2008).

2.2. Preprocesamiento: Pruebas de calidad de los datos

Con el objetivo de verificar la calidad de los datos hidrológicos, es necesario que éstos cumplan con algunos supuestos de calidad. Por ejemplo, la independencia referente a la aleatoriedad de los datos, la homogeneidad para determinar si los datos pertenecen a la misma distribución estadística y la detección de outliers o datos atípicos (Ouarda y otros, 2008). Para corroborar este tipo de aspectos existen diferentes pruebas hidrológicas que permiten comprobar si una muestra de datos hidrológicos cumple con cada uno de los supuestos antes mencionados (Torres-Gallardo y Peñaranda Gómez, 2006). Dichas pruebas se encuentran dentro de las denominadas técnicas estadísticas no paramétricas o contrastes de distribución libres, de las cuales sus mayores atractivos residen en que:

- No exigen ninguna condición suplementaria a la muestra sobre su procedencia de una población con cierto tipo de distribución, por tanto, son más fáciles de aplicar que las alternativas paramétricas.
- Son propias para usarse en muestras pequeñas de datos.
- Se pueden aplicar a datos cualitativos.

Las técnicas no paramétricas juegan un papel fundamental en la ordenación de datos, pues en la mayoría de casos se calculan las magnitudes de las observaciones, sólo para establecer una relación de menor a mayor entre ellas, denominadas "rangos" (Ríus-Díaz y otros, 1998).

Para cada una de las pruebas de calidad, es necesario establecer un nivel de significancia estadística, definida como la probabilidad de rechazar una hipótesis cuando esta es verdadera. De acuerdo a la experiencia de otros estudios (Chang y Donald, 2003; Chokmani y Ouarda, 2004; Ouarda y otros, 2008), el nivel de significancia más utilizado es de 5 %, es decir, el investigador tiene un 95 % de confianza para generalizar su hipótesis, y solo el 5 % de probabilidad de equivocarse.

Generalmente, los estadísticos obtenidos por las pruebas no paramétricas son valores normalizados, por tanto, para determinar si se acepta o no la hipótesis de la correspondiente prueba se hace una comparación entre el estadístico calculado y los valores críticos para una

Cuadro 2.1: Valores críticos para una distribución normal estándar.

Valores críticos		
α	α / 2	$Z_{\alpha/2}$
1 %	0.005	2.58
5 %	0.025	1.96
10 %	0.05	1.64

distribución normal estándar. Éstos últimos se encuentran tabulados o pueden calcularse en función de la distribución de la población.

El Cuadro 2.1 muestra los valores críticos para una distribución normal estándar. El valor crítico es representado por $Z_{\alpha/2}$ y se define como el valor de la abscisa en una determinada distribución que deja a su derecha un área igual a $\alpha/2$. De esta manera, α representa el nivel de significancia estadística y $1-\alpha$ el nivel de confianza correspondiente (Miller y otros, 1997).

2.2.1. Prueba de independencia y aleatoriedad

En general suele suponerse que los datos recolectados para un estudio constituyen una muestra aleatoria, de modo que cada observación o medida es tomada de la población de manera aleatoria e independiente. Tal suposición, puede ser probada mediante el empleo de un procedimiento no paramétrico conocido como prueba de rachas de Wald-Wolfowitz.

La hipótesis nula de aleatoriedad puede probarse mediante la observación del orden o de la secuencia en que se obtienen los elementos de la muestra. Si a cada elemento se le asigna uno de dos términos, como E y F (por Éxito y Fracaso), dependiendo de si el valor se encuentra arriba o abajo de la mediana estadística (referencia) respectivamente, la aleatoriedad de la secuencia puede ser investigada (Berenson y Levine, 1996).

Para estudiar si una secuencia observada es aleatoria o no, se considera como estadístico de prueba al número de rachas presentes en los datos. Una racha se define como una serie consecutiva de elementos similares que están limitados por elementos de un tipo diferente o por el inicio o fin de una secuencia.

Para probar la hipótesis nula de aleatoriedad, es necesario dividir el tamaño completo de la muestra n, en dos partes: $n1$ y $n2$. Donde $n1$ representa el número total de éxitos, o de valores superiores al valor de referencia y $n2$ el número de fracasos, o de valores inferiores a la referencia.

La estadística de la prueba, es representada por la letra Z, dicha estadística esta distribuida de manera aproximadamente normal y es calculada como:

$$Z = \frac{R - \mu_R}{\sigma_R^2} \tag{2.1}$$

Donde R es el número total de rachas observadas, su valor medio muestral es dado por:

$$\mu_R = \frac{2n_1n_2}{n} + 1 \tag{2.2}$$

Y la desviación estándar de R esta dada por:

$$\sigma_R = \sqrt{\frac{(\mu_R - 1)(\mu_R - 2)}{n - 1}} \tag{2.3}$$

El valor absoluto del estadístico Z obtenido por el procedimiento indicado, se contrasta con los valores de la tabla de distribución normal para un cierto nivel de significancia estadística α, si $|Z|$ se encuentra comprendido entre los límites de la tabla, se dice que los valores que integran la serie son aleatorias. De lo contrario, se rechaza tal afirmación y los datos no pasan la prueba de independencia o aleatoriedad.

2.2.2. Prueba de homogeneidad

La prueba de homogeneidad de Mann-Whitney es usada con frecuencia para determinar si un conjunto de datos es homogéneo o bien pertenece a la misma distribución estadística (Wilcoxon, 1945). El problema consiste en decidir si dos poblaciones son iguales o si es más probable que una produzca observaciones más grandes que la otra, para ello se comparan dos grupos de datos provenientes de la misma muestra.

Al igual que la prueba de independencia, es necesario dividir una muestra de tamaño n en dos submuestras de tamaño p y q respectivamente, con $p \leq q$. Para la muestra de tamaño n, se enumera cada dato, después se ordena de menor a mayor manteniendo la numeración inicial. Posteriormente, se calcula la suma de rangos pertenecientes a la primer submuestra (la de tamaño p), donde un rango se define como la posición que ocupa cierto elemento dentro del conjunto no ordenado de datos.

La prueba de Mann-Whitney considera los siguientes valores:

$$V = R - \frac{(p(p+1))}{2} \tag{2.4}$$

$$W = pq - V \tag{2.5}$$

Donde R es la suma de los rangos de la primer submuestra (de tamaño p) en las series combinadas de tamaño n, con $n > 20$.

El estadístico de la prueba de homogeneidad es representado por la letra U, y definido por el valor mínimo entre V y W. Para el estadístico U, se asume una distribución normal, con media y varianza dadas por las siguientes ecuaciones, respectivamente:

$$\mu_U = \frac{pq}{2} \tag{2.6}$$

$$var(U) = [\frac{pq}{n(n-1)}][\frac{n^3 - n}{12}] \tag{2.7}$$

Finalmente, el estadístico u, presenta una distribución normal con media cero y varianza unitaria y es usada para probar la hipótesis nula de homogeneidad con un nivel de significancia α.

$$u = \frac{U - \mu_U}{\sqrt{var(U)}} \tag{2.8}$$

Al igual que en la prueba de independencia, el valor de la estadística $|u|$ de cada serie de datos se compara con los valores críticos de la distribución normal estándar (Cuadro 2.1), y de esta manera se acepta o rechaza la hipótesis de homogeneidad en los datos (Torres-Gallardo y Peñaranda Gómez, 2006).

2.2.3. Prueba de outliers

En un conjunto de datos, se habla de la existencia de outliers cuando uno o más datos se desvían considerablemente de la distribución de los mismos. Esta situación puede darse por diferentes causas: una mala accesibilidad a los datos, olvido o errores en la captura de éstos, suposición de datos o simplemente porque el dato efectivamente se encuentra en un nivel alto o bajo con respecto a la media muestral de la serie.

La presencia de outliers causa dificultad al momento de ajustar los datos a una distribución específica. Por tanto, los outliers máximos y mínimos son posibles y tienen diferentes efectos en el análisis de datos, por ejemplo, los valores extremadamente altos, arrastran la media muestral hacia arriba y los valores extremadamente bajos, contrariamente arrastran la media muestral hacia abajo.

En la práctica, existen diferentes métodos que sirven para identificar outliers en un conjunto de datos. La prueba de Grubbs y Beck, desarrollada en 1972 y mostrada en (Torres-

Gallardo y Peñaranda Gómez, 2006) es una de las técnicas más abordadas para detección de outliers en estudios de hidrología.

En esta prueba se consideran los límites superior e inferior de la muestra, X_H y X_L, respectivamente, y son calculadas como:

$$X_H = exp(\bar{X} + K_n S) \tag{2.9}$$

$$X_L = exp(\bar{X} - K_n S) \tag{2.10}$$

Donde $\bar{X}$ y S son la media muestral y la desviación estándar, respectivamente, de los logaritmos naturales de la muestra, y K_n es la prueba estadística tabulada para diferentes muestras de tamaño n. En el Cuadro 2.2 se muestran los valores de K_n para diferentes tamaños de muestra con un nivel de significancia del 5 % y del 1 % (Garcia, 2007).

Los valores de la muestra mayores que X_H son considerados outliers altos, mientras que los valores menores que X_L son considerados outliers bajos.

Cuadro 2.2: Valores críticos de la prueba de Grubbs y Beck.

n	**5 %**	**1 %**	**n**	**5 %**	**1 %**	**n**	**5 %**	**1 %**
3	1.1531	1.1546	15	2.4090	2.7049	80	3.1319	3.5208
4	1.4625	1.4925	16	2.4433	2.7470	90	3.1733	3.5632
5	1.6714	1.7489	17	2.4748	2.7854	100	3.2095	3.6002
6	1.8221	1.9442	18	2.5040	2.8208	120	3.2706	3.6619
7	1.9381	2.0973	19	2.5312	2.8535	140	3.3208	3.7121
8	2.0317	2.2208	20	2.5566	2.8838	160	3.3633	3.7542
9	2.1096	2.3231	25	2.6629	3.0086	180	3.4001	3.7904
10	2.1761	2.4097	30	2.7451	3.1029	200	3.4324	3.8220
11	2.2339	2.4843	40	2.8675	3.2395	300	3.5525	3.9385
12	2.2850	2.5494	50	2.9570	3.3366	400	3.6339	4.0166
13	2.3305	2.6070	60	3.0269	3.4111	500	3.6952	4.0749
14	2.3717	2.6585	70	3.0839	3.4710	600	3.7442	4.1214

2.3. Agrupamiento jerárquico

Los algoritmos de agrupamiento jerárquico se caracterizan por tener una estructura en forma de árbol a lo que comúnmente se llama dendograma, en la que cada nivel es un agrupamiento posible de los objetos en la colección de datos (Jain y otros, 1999). De esta

manera, cada vértice o nodo del árbol representa un grupo de objetos, y la raíz del mismo puede contener a todos los elementos de la colección, formando un solo grupo. Por su parte, cada hoja en el último nivel del árbol formaría un grupo independiente, de modo que pueden existir tantas hojas como número de objetos en la colección de datos.

Los algoritmos de agrupamiento jerárquico se dividen en jerárquicos por aglomeración y jerárquicos por división de acuerdo al método que aplican para obtener el dendograma (Han y otros, 2001). El agrupamiento por aglomeración empieza con cada objeto en un grupo separado y en cada paso, se mezcla el par de grupos más similares o cercanos en contenido, este proceso puede continuar hasta que todos los objetos formen parte de un solo *cluster*. El agrupamiento por división, hacen el camino inverso, comienza en la raíz del árbol con todos los objetos en un solo grupo, y en cada paso, se divide el par de grupos más disimilares en contenido.

En el anexo B.1 se muestra el pseudocódigo correspondiente al algoritmo de agrupamiento jerárquico por aglomeración que fue utilizado para identificar las regiones hidrológicamente homogéneas en la Mixteca Oaxaqueña.

En este tipo de agrupamiento, la obtención de los *clusters* o grupos depende del criterio de disimilitud que se usa entre ellos. La medida o criterio de disimilitud que se utiliza con mayor frecuencia es la distancia euclidiana, la cual es basada en el teorema de Pitágoras y mostrada en la ecuación 2.11. No obstante, se pueden usar otras medidas de disimilitud entre grupos, tales como la distancia Manhattan o la distancia de Chebyshev.

$$d(X,Y) = ||X - Y|| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.11}$$

En general, el primer paso para el agrupamiento consiste en definir las variables que ayudan a formular el problema. Después, debe seleccionarse una medida de disimilitud apropiada, ésta determina que tan similares o diferentes son los objetos que se agrupan. Los grupos derivados deben interpretarse en términos de las variables utilizadas para formarlos, las cuales deben ser de preferencia, las más relevantes para el caso de estudio.

2.3.1. Características del agrupamiento jerárquico

Existen ciertas características que son propias de los algoritmos de agrupamiento jerárquico, dentro de las cuales se pueden mencionar las siguientes:

- Su forma de trabajo es simple e intuitiva: El enfoque utilizado por estos métodos es similar al que utiliza una persona para realizar un agrupamiento, especialmente los de aglomeración que buscan similitud entre los diferentes grupos para unirlos en uno solo.

- Su resultado es una serie de agrupamientos anidados: Esta característica permite revisar toda la estructura del árbol para observar a detalle los *clusters* formados en los diferentes niveles del agrupamiento.
- Son deterministas: Al aplicar dos veces el algoritmo jerárquico, en ambas ocasiones se seguirá el mismo camino para llegar a la solución y por tanto el resultado de agrupamiento será exactamente el mismo.
- No revisan las decisiones que toman los pasos anteriores: Una vez que dos elementos se han asignado a un grupo, ningún paso posterior los volverá a separar o juntar, según sea al caso (divisivo o aglomerativo), por lo que una asignación incorrecta en los primeros pasos no podrá corregirse en pasos posteriores.
- Requieren grandes tiempos de cómputo: La forma de buscar en cada paso, los grupos a unir o dividir, hacen que las implementaciones conocidas de estos algoritmos tengan tiempos de ejecución del orden de n^2 ó n^3, por lo que para conjuntos de datos grandes, el algoritmo se vuelve muy lento (Oren y Oren, 1998).

2.3.2. Funciones de enlace entre grupos

Las funciones de enlace entre grupos son aquellas que indican el grado de homogeneidad que puede existir entre dos grupos de observaciones en una colección de datos. Determinan el par de grupos que deben unirse o separarse para continuar la construcción del dendograma correspondiente. Cada enlace determina la distancia existente entre dos pares de grupos. Posteriormente el algoritmo de agrupamiento jerárquico toma la distancia más pequeña o más grande entre grupos, para juntar o separar los mismos.

Sean r y s, las representaciones del *cluster* R y S respectivamente, n_r el número de objetos en el grupo R, n_s el número de objetos en el grupo S y x_{ri} el *i-ésimo* objeto del grupo R, las funciones de enlace se pueden definir como:

- Enlace simple: Es aquel que se basa en la distancia mínima o la regla del vecino más cercano. Es decir, la distancia entre dos grupos, es determinada por la mínima distancia que existe entre un elemento del grupo R y un elemento del grupo S, su representación matemática esta dada por:

$$d(r,s) = min(dist(x_{ri}, x_{sj})) \quad \forall \quad i,j \tag{2.12}$$

- Enlace completo: Es el caso contrario al enlace simple, pues toma siempre los elementos más alejados entre el grupo R y S para determinar la distancia existente entre ambos grupos de datos. Matemáticamente se expresa como sigue:

$$d(r,s) = max(dist(x_{ri}, x_{sj})) \quad \forall \quad i,j \tag{2.13}$$

- Enlace promedio: Esta función toma como distancia entre dos grupos el promedio de todas las distancias entre elementos del *cluster* R y el *cluster* S. Matemáticamente se expresa como sigue:

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \tag{2.14}$$

- Enlace mediano: El cálculo de la distancias entre dos grupos es determinada por la distancia que existe entre los centros de masa de cada grupo, es decir, la distancia euclidiana entre medianas, expresada como:

$$d(r,s) = \| \tilde{x_r} - \tilde{x_s} \|_2 \tag{2.15}$$

  Donde, $\tilde{x_r}$ y $\tilde{x_s}$ son los centros de masa para los grupos R y S, respectivamente, y $\| \cdot \|_2$ representa la distancia euclidiana. En caso de que el grupo R fuese creado con la combinación de los grupos P y Q, $\tilde{x_r}$ es definida recursivamente como:

$$\tilde{x_r} = \frac{1}{2}(\tilde{x_p} + \tilde{x_q}) \tag{2.16}$$

- Enlace centroide: Se toma como distancia entre dos grupos, la distancia que existe entre los centroides de cada grupo, esto es, la distancia euclidiana entre medias, la cual definimos como sigue:

$$d(r,s) = \| \bar{x_r} - \bar{x_s} \|_2 \tag{2.17}$$

  Donde, $\bar{x_r}$ y $\bar{x_s}$ son los centroides para los grupos R y S respectivamente, dichos centroides están dados por:

$$\bar{x_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri} \tag{2.18}$$

- Enlace *ward*: Esta función de enlace especifica que la distancia entre dos *clusters* se calcula con el incremento del error en la suma de cuadrados (ESS, por sus siglas en ingles) después de unir dos grupos en uno solo. El método *ward* une o separa dos grupos minimizando este error en cada paso.

  El error en la suma de cuadrados del grupo R con n_r objetos, se define como la suma de los cuadrados de las distancias entre todos los objetos del grupo y el centroide del mismo, es decir:

$$ESS(r) = \sum_{i=1}^{n_r} (x_{ri} - \frac{1}{n_r} \sum_{j=1}^{n_r} x_{rj})^2 \tag{2.19}$$

Matemáticamente, la distancia entre los grupos R y S es descrita por la siguiente expresión:

$$d(r,s) = ESS(rs) - [ESS(r) + ESS(s)] \tag{2.20}$$

Donde, rs es el resultado de la combinación de los grupos R y S.

Sin embargo, según (Batagelj, 1988) una distancia equivalente puede ser determinada como:

$$d^2(r,s) = n_r n_s \frac{\| \bar{x_r} - \bar{x_s} \|_2^2}{(n_r + n_s)} \tag{2.21}$$

Donde, $\bar{x_r}$ y $\bar{x_s}$ son los centroides para los grupos R y S respectivamente. La segunda forma de calcular la distancia entre grupos puede resultar más sencilla, pues sólo depende de los centroides y número total de elementos en cada grupo.

En cada caso, las funciones de enlace proporcionan una distancia o una medida de disimilitud entre grupos, y queda a consideración del investigador el uso y aplicación de uno u otro método de enlace para la construcción de su respectivo dendograma.

2.4. Agrupamiento difuso

Zadeh en 1965 (Zadeh, 1965) definió el concepto de conjunto difuso, basándose en la idea de que existen conjuntos en los que no está claramente determinado si un elemento pertenece o no al conjunto. A veces, un elemento pertenece al conjunto con cierto grado, llamado grado de pertenencia.

En la práctica podemos encontrar muchos conceptos con incertidumbre, por ejemplo, cuando mencionamos que una persona es alta: ¿Qué es una persona alta?, ¿Cómo definimos a una persona alta?, ¿Cuánto mide una persona alta? Todos estos conceptos, habituales en el lenguaje natural del hombre pueden ser representados mediante un conjunto difuso.

En un conjunto difuso, cada objeto puede tomar valores entre [0,1] indicando el grado de pertenencia o membresía del objeto a un conjunto de datos. Un cero indica que no existe relación entre el objeto y el grupo. Un uno muestra una pertenencia completa del objeto al grupo en cuestión.

Sea X una colección de objetos, el conjunto difuso A perteneciente a X puede representarse como un conjunto de pares ordenados de valores, en el que cada elemento x esta acompañado de su grado de pertenencia $\mu(x)$ (Gath y Geva, 1989). Esto es:

$$A = \{\mu_A(x)/x\} \tag{2.22}$$

Donde $\mu_A(x)$ representa la función de membresía o grado de pertenencia del objeto x en el conjunto difuso A.

2.4.1. Propiedades de los conjuntos difusos

Al igual que el agrupamiento jerárquico, los conjuntos difusos cuentan con algunas propiedades que los caracterizan. Dichas propiedades son enlistadas como sigue:

- Prefijación de grupos. Al iniciar un algoritmo de agrupamiento difuso se debe prefijar el número de *clusters* a generar. De esta manera, el investigador tiene la libertad para determinar el total de grupos que desea formar para su estudio.
- Normalidad. Se dice que un conjunto difuso es normal si el valor más grande de la función de membresía esta representado por la unidad, es decir: $max(\mu(x)) = 1$, en otro caso, se dice que el conjunto difuso es subnormal.
- Punto de cruce. El punto de cruce de un conjunto difuso A, es el elemento cuyo grado de membresía en A es igual a 0.5.
- Punto difuso. Un punto difuso es un par de valores donde sólo un grado de membresía es asignado a una observación del conjunto de datos. Sea P un punto difuso en un universo n-dimensional X, y $x \in X$, el punto difuso P es descrito como:

  $$P = \mu/x \tag{2.23}$$

  Donde μ representa el grado de membresia de x. En este sentido, un conjunto difuso A puede ser considerado como la unión de dos o más puntos difusos P.

2.4.2. Técnicas de agrupamiento difuso

El agrupamiento es un herramienta matemática que intenta obtener las relaciones entre varios objetos de un conjunto de datos, organizando los patrones en grupos, así los patrones dentro de un mismo grupo son similares entre ellos y diferentes al resto de los grupos.

En el agrupamiento convencional, un objeto puede pertenecer o no a un grupo en particular, con grados de membresía iguales a uno o cero respectivamente; por su parte, el agrupamiento difuso asigna a cada observación un grado de pertenencia a todos los grupos existentes. Dichos valores de membresía pueden variar desde cero hasta uno y mientras más

grande sea el grado de membresia de un objeto hacia un grupo, más relación tendrá ese objeto con dicho grupo (Orozco y otros, 2005).

En este sentido, varias técnicas han sido diseñadas para el agrupamiento de datos difusos, dentro de las más utilizadas podemos encontrar el algoritmo Fuzzy C-Means; el algoritmo AVQ, por sus siglas en inglés: Adaptive Vector Quantization; y el algoritmo de mapas auto-organizados mejor conocido como self-organizing map (SOM).

El algoritmo de agrupamiento Fuzzy C-Means desarrollado por (Dunn, 1973) y mejorado por (Bezdek, 1981) es uno de los más utilizados en el área de reconocimiento de patrones. Esta basado en la minimización de la siguiente función objetivo.

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m||x_k - v_i||^2 \tag{2.24}$$

Donde m, es un valor constante mayor a uno, u_{ik} es el grado de membresía de la *k-ésima* observación en el cluster i, x_k representa la *k-ésima* observación dentro del conjunto de datos y v_i es el centroide del *i-ésimo* grupo.

En el anexo B.2, se muestra con más detalle el algoritmo Fuzzy C-Means, utilizado para la delimitación de regiones homogéneas en nuestro caso de estudio.

2.4.3. Defusificación

Una vez que se ha ejecutado el algoritmo de agrupamiento difuso, se tiene como resultado una matriz de pertenencias donde cada observación (columna) tiene un grado de membresía hacia cada uno de los grupos prefijados por el investigador (fila), esta matriz es llamada "matriz de partición difusa" (U). La función que se encarga de transformar el valor difuso de una observación, en un valor de pertenencia o ausencia hacia un grupo es llamado defusificador.

Para cada observación x_k, el método de máxima pertenencia mostrado en (Srinivas y otros, 2008), toma el elemento más grande en cada columna de la matriz U y le asigna un nuevo grado de pertenencia con valor de uno y al resto de los elementos en la columna les asigna un grado de pertenencia de cero. En otras palabras, el objeto es asignado al *cluster* donde presenta mayor relevancia. De manera formal, se puede describir como:

$$si \quad u_{jk} = max(u_{jk}) \quad \forall \quad 1 \leq j \leq c \quad u_{jk} = 1; \qquad u_{ik} = 0 \quad \forall\, i \neq j \tag{2.25}$$

Alternativamente, el método del centroide más cercano descrito también en (Srinivas y otros, 2008), toma cada elemento x_k y éste es asignado al *cluster* cuyo centroide se encuentre más cercano a él en términos de distancias euclidianas, es decir:

$$si \quad d_{jk} = min||v_j - x_k|| \quad \forall \quad 1 \leq j \leq c \quad u_{jk} = 1; \qquad u_{ik} = 0 \quad \forall \quad i \neq j \tag{2.26}$$

Donde v_j representa el centroide del *j-ésimo cluster*. Cualquiera de los procedimientos mencionados determinará los grupos finales del análisis de datos y dependerá del investigador tomar el método que mejor se adapte a sus necesidades.

2.4.4. Validación de *clusters*

Para obtener el número óptimo de *clusters* (c) que se forman en un conjunto de datos, se deben utilizar algunas medidas de validación para el agrupamiento (Pal y Bezdek, 1995). Dichas medidas de validación, indican cuál de las particiones difusas formadas por el algoritmo de agrupamiento Fuzzy C-Means es la que mejor describe al conjunto de datos.

Por tanto, cuatro medidas de validación de *clusters*, llamadas Coeficiente de Partición Difusa (V_{PC}), Entropía de Partición Difusa (V_{PE}), Índice de Realización Difusa (FPI) y Entropía de Clasificación Normalizada (NCE), pueden ser calculadas para diferente número de *clusters* (c) y matrices de pertenencia (U), y con éstas, determinar el número de grupos adecuado en cada caso. Además, estos índices los cuales no tienen relación directa con las propiedades de los datos, se han utilizado en diferentes estudios de hidrología (Guler y Thyne, 2004).

Los índices de validación V_{PC} y V_{PE} propuestos por (Bezdek, 1974), y los índices FPI y NCE introducidos por (Roubens, 1982) son definidos como:

$$V_{PC}(U) = \frac{1}{n} \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^2 \tag{2.27}$$

$$V_{PE}(U) = -\frac{1}{n} \left[\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} log(u_{ik}) \right] \tag{2.28}$$

$$FPI(U) = 1 - \frac{(c)\ (V_{PC}(U)) - 1}{c - 1} \tag{2.29}$$

$$NCE(U) = \frac{V_{PE}(U)}{log(c)} \tag{2.30}$$

En cada caso, la idea principal de las funciones de validación de grupos, es que la partición con menos promedio difuso será la que mejor desempeño tendrá para los datos agrupados.

En términos de los índices de validación de *clusters*, el mejor resultado de agrupamiento corresponde al máximo valor de V_{PC}, o bien, el mínimo de V_{PE}, FPI y NCE para $c = 2, ..., C$.

El rango de variación del índice V_{PC} es de $[1/c, 1]$, mientras que el de V_{PE} es de $[0, log(c)]$. Por su parte, el rango de variación de los índices FPI y NCE es de [0,1]. Por tanto, para el caso de particiones duras o fijas, el valor del índice de validación V_{PC} debe ser igual a 1, mientras los valores de V_{PE}, FPI y NCE deben ser todos iguales a 0.

2.5. Postprocesamiento: Test de heterogeneidad de grupos

Las regiones hidrológicas formadas por los algoritmos de agrupamiento, se consideran por si mismas homogéneas, no obstante, es necesario comprobar dicha homogeneidad mediante pruebas regionales. Hosking y Wallis en (Hosking y Wallis, 1993, 1997) proponen las pruebas de heterogeneidad H, para evaluar cuando una región propuesta puede ser considerada como homogénea. Con esta prueba se estima el grado de coherencia en un *cluster* de estaciones hidrométricas y se garantiza que éstas puedan ser tratadas como una región homogénea.

Las medidas de heterogeneidad H comparan la variabilidad de los L-momentos del grupo de estaciones que conforman una región homogénea con los L-momentos correspondientes a una región homogénea simulada. Dichas medidas de heterogeneidad están basadas en los L-coeficientes de variación, asimetría y curtosis (L-CV, L-skewness y L-kurtosis, respectivamente).

Los L-momentos (λ_r, $r = 1, 2, 3$, etc.) representan un sistema alternativo para describir la forma de diferentes distribuciones de probabilidad (Hosking y Wallis, 1993). Desde el punto de vista estadístico, son una combinación lineal de los momentos ponderados de probabilidad o momentos de probabilidad pesada que fueron desarrollados por (Greenwood y otros, 1979). Los estimadores muestrales de los L-momentos para un conjunto ordenado de n-muestras $x_1 \leq x_2 \leq ... \leq x_n$ son dados por:

$$\lambda_1 = \sum_{i=1}^{n} x_i/n \tag{2.31}$$

$$\lambda_2 = \sum_{\forall i>j} (x_i - x_j)/(n-1) \tag{2.32}$$

$$\lambda_3 = 2 \sum_{\forall i>j>k} (x_i - 2x_j + x_k)/n(n-1)(n-2) \tag{2.33}$$

$$\lambda_4 = 6 \sum_{\forall i>j>k>l} (x_i - 3x_j + 3x_k - x_l)/n(n-1)(n-2)(n-3) \tag{2.34}$$

Por su parte, el L-coeficiente de variación, es definido como $\tau_2 = \lambda_2/\lambda_1$ (Pearson, 1991). El resto de los coeficientes son definidos como $\tau_r = \lambda_r/\lambda_2$, para $r = 3, 4, 5$, etc. Hosking (Hosking y Wallis, 1993) muestra que τ_3 y τ_4 son las medidas correspondientes a los L-coeficientes de asimetría y curtosis, respectivamente. Con estos datos, las pruebas H de heterogeneidad pueden describirse como sigue:

Supóngase que la región a ser evaluada tiene N_R estaciones, donde la *i-ésima* estación tiene un total de n_i observaciones de caudales. $t^{(i)}$, $t_3^{(i)}$ y $t_4^{(i)}$ denotan los L-coeficiente de variación, asimetría y curtosis, respectivamente, para la estación i. Por tanto, el promedio regional de L-coeficientes de variación, asimetría y curtosis son calculados como:

$$t^R = \sum_{i=1}^{N_R} n_i t^{(i)} / \sum_{i=1}^{N_R} n_i \tag{2.35}$$

$$t_3^R = \sum_{i=1}^{N_R} n_i t_3^{(i)} / \sum_{i=1}^{N_R} n_i \tag{2.36}$$

$$t_4^R = \sum_{i=1}^{N_R} n_i t_4^{(i)} / \sum_{i=1}^{N_R} n_i \tag{2.37}$$

Cada una de las pruebas de heterogeneidad están basadas en diferentes medidas de dispersión. La primer prueba de heterogeneidad, se basa en el L-coeficiente de variación, la segunda prueba se basa en los L-coeficientes de variación y asimetría y la última prueba está basada en los L-coeficientes de asimetría y curtosis. Éstas medidas de dispersión se utilizan para calcular la variabilidad de los L-momentos en el conjunto de datos, representada como:

$$V = \left\{ \frac{\sum_{i=1}^{N_R} n_i (t^{(i)} - t^R)^2}{\sum_{i=1}^{N_R} n_i} \right\}^{1/2} \tag{2.38}$$

$$V_2 = \frac{\sum_{i=1}^{N_R} n_i \left\{ (t^{(i)} - t^R)^2 + (t_3^{(i)} - t_3^R)^2 \right\}^{1/2}}{\sum_{i=1}^{N_R} n_i} \tag{2.39}$$

$$V_3 = \frac{\sum_{i=1}^{N_R} n_i \left\{ (t_3^{(i)} - t_3^R)^2 + (t_4^{(i)} - t_4^R)^2 \right\}^{1/2}}{\sum_{i=1}^{N_R} n_i} \tag{2.40}$$

Para poder comparar la variabilidad de los L-momentos, se realiza un número muy grande de simulaciones de la región evaluada, generalmente alrededor de 500 simulaciones, cada una de ellas teniendo la distribución kappa como distribución de frecuencia. Dichas simulaciones se hacen estableciendo los promedios regionales de los L-coeficientes calculados: 1, t^R, t_3^R y t_4^R como parámetros de la distribución kappa (Hosking, 1994).

En este caso, cada simulación representa una región homogénea con N_R estaciones, y cada estación simulada tiene la misma cantidad de datos que su contraparte en el mundo real. Por lo que para cada región simulada se calculan nuevamente los valores de variabilidad V, V_2 y V_3.

Una vez que se tiene un número grande de simulaciones, se determinan la media muestral y la desviación estándar de las variabilidades simuladas V, V_2 y V_3. Estas estadísticas son usadas para estimar las medidas de heterogeneidad de los grupos formados por los algoritmos de agrupamiento, tal como se muestra en las siguientes ecuaciones:

$$H_1 = \frac{V - \mu_V}{\sigma_V} \tag{2.41}$$

$$H_2 = \frac{V_2 - \mu_{V_2}}{\sigma_{V_2}} \tag{2.42}$$

$$H_3 = \frac{V_3 - \mu_{V_3}}{\sigma_{V_3}} \tag{2.43}$$

La región evaluada se declara heterogénea si $H_i, i = 1, 2, 3$, es suficientemente grande. Hosking y Wallis en (Hosking y Wallis, 1997), sugieren que la región puede ser considerada "aceptablemente homogénea" si $H_i < 1$, "posiblemente heterogénea" si $1 \leq H_i < 2$ y "definitivamente heterogénea" si $H_i \geq 2$.

Capítulo 3

Regresión lineal

3.1. Introducción

El análisis de regresión lineal es una técnica estadística para la investigación y modelado de la relación existente entre una variable dependiente y una o más variables independientes de un caso de estudio (Montgomery y otros, 2001). Algunas de las aplicaciones de regresión lineal se pueden encontrar en campos de investigación como: ingeniería, ciencias físicas y químicas, economía, administración, ciencias biológicas, ciencias de la vida y ciencias sociales.

Una forma de expresar el modelo de regresión lineal es el siguiente:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.1}$$

Comúnmente, se dice que x es la variable independiente y y la variable dependiente. Sin embargo, estas definiciones causan confusión con el concepto de independencia estadística, por tanto, la variable x también es llamada "variable predictora o regresora" y la variable y "variable de respuesta". Como la ecuación 3.1 sólo tiene una variable regresora, se llama **modelo de regresión lineal simple**. Por su parte, el símbolo ε representa el error o la diferencia entre el valor observado de y y el valor estimado por la recta $(\beta_0 + \beta_1 x)$. Dicho error puede formarse por los efectos de otras variables incluidas en el modelo, por errores de medición, etc.

3.2. Regresión lineal múltiple

La regresión lineal múltiple es un método probabilístico utilizado para modelar la relación lineal entre una variable respuesta y dos o más variables predictoras. El método esta basado

en el concepto de mínimos cuadrados: El modelo se ajusta de tal manera que la suma de cuadrados de las diferencias entre los valores observados y pronosticados es reducido a un mínimo. Entonces, el modelo expresa el valor de la variable respuesta como una función lineal de dos o más variables predictoras y un término de error como sigue:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{3.2}$$

Donde, x_k es el valor del *k-ésimo* predictor, β_0 es la constante de regresión lineal, β_k es el *k-ésimo* coeficiente de predicción, la variable y representa el valor estimado para la variable respuesta, y la variable ε es el término de error entre los valores observado y estimado.

Este modelo describe un hiperplano en el espacio de k dimensiones y el parámetro β_k representa el cambio que se espera en la respuesta y por cambio unitario en x_k cuando las demás variables regresoras x_i $(i \neq k)$ se mantienen constantes.

Los modelos de regresión lineal se pueden usar con diversos fines, entre ellos se puede destacar:

- Descripción de los datos.
- Estimación de parámetros.
- Predicción.
- Control.

El objetivo más utilizado de la regresión lineal múltiple es el referente a la predicción y estimación de valores. Para ello, se hace uso de registros históricos de los datos y éstos se ajustan a una ecuación de regresión lineal que finalmente se utiliza para estimar los nuevos valores de la variable respuesta.

3.2.1. Ajuste del modelo a los datos

En la mayor parte de los problemas de regresión lineal se desconocen los valores de los coeficientes de predicción (coeficientes β_k) y la varianza del error en los datos (σ^2), por lo que dichos parámetros se deben estimar a partir de datos muestrales. El procedimiento más utilizado para estimar los coeficientes de regresión en la ecuación 3.2, es el método de mínimos cuadrados, descrito a continuación.

Sea n, el número de observaciones en los datos, con $n >> k$, y_i la *i-ésima* respuesta observada , y x_{ij} la *i-ésima* observación del regresor x_j, los datos muestrales del caso de estudio deberán observarse como en el Cuadro 3.1.

Cuadro 3.1: Datos para una regresión lineal múltiple.

Observación	Respuesta	Regresores			
i	y	x_1	x_2	...	x_k
1	y_1	x_{11}	x_{12}	...	x_{1k}
2	y_2	x_{21}	x_{22}	...	x_{2k}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
n	y_n	x_{n1}	x_{n2}	...	x_{nk}

Para el análisis de mínimos cuadrados se supone que las variables regresoras $x_1, x_2, ..., x_k$ son fijas, es decir, que son variables no aleatorias y que los datos se miden sin error. Sin embargo, el método sigue siendo válido para el caso en el que los regresores están dados por variables aleatorias. Esto es de suma importancia, ya que cuando los datos provienen de un estudio observacional, la mayor parte de los regresores son variables aleatorias. En cambio, cuando los datos son resultado de un experimento diseñado es más probable que los regresores sean variables fijas.

El modelo muestral de regresión lineal que corresponde a la ecuación 3.2 se puede reescribir de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \qquad i = 1, 2, ..., n \tag{3.3}$$

Por su parte, la función de mínimos cuadrados se describe formalmente como:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} e_i^2 \tag{3.4}$$

Despejando el término de error de la ecuación 3.3 y sustituyendo el mismo en la función de mínimos cuadrados, se obtiene:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2. \tag{3.5}$$

Si se expresa el modelo de regresión lineal múltiple en notación matricial, esto permite presentar en una forma más compacta el modelo, los datos y los resultados. Dicha notación matricial del modelo es presentada como:

$$Y = X\beta + \varepsilon \tag{3.6}$$

En donde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

En términos generales, Y es un vector de tamaño $n \times 1$ y representa las observaciones del experimento, X es una matriz de $n \times (k+1)$ representando los valores en cada una de las variables regresoras, β es un vector de $(k+1) \times 1$ correspondiente a los coeficientes de regresión y ε es un vector de $n \times 1$ para determinar los errores o diferencias con respecto a la muestra utilizada.

Tomando en cuenta los datos anteriores, se desea determinar el vector $\hat{\beta}$ de estimadores de mínimos cuadrados que minimice los errores en los datos por medio de la expresión:

$$S(\beta) = \sum_{i=1}^{n} e_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \tag{3.7}$$

Siguiendo el procedimiento mostrado en (Montgomery y otros, 2001) se obtiene que:

$$X'X\hat{\beta} = X'Y \tag{3.8}$$

La ecuación 3.8 es la ecuación normal de mínimos cuadrados en su forma matricial. Para resolver dicha ecuación, se multiplican ambos lados de 3.8 por la inversa de $X'X$. Así, el estimador de β por mínimos cuadrados queda como:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.9}$$

Siempre y cuando exista la matriz inversa $(X'X)^{-1}$. Dicha matriz existe si los regresores son linealmente independientes, es decir, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Finalmente, el vector de valores estimados de la variable respuesta $\hat{y}_i$, que corresponde a los valores observados y_i es calculado como:

$$\hat{Y} = X\hat{\beta} \tag{3.10}$$

La diferencia entre el valor observado y_i y el valor estimado $\hat{y}_i$ correspondiente es el residual $e_i = y_i - \hat{y}_i$. Por tanto los n residuales se pueden escribir sin ningún problema con la notación matricial como sigue:

$$\varepsilon = Y - \hat{Y} = Y - X\hat{\beta} \tag{3.11}$$

3.2.2. Estimación de σ^2

El estimador σ^2 es definido como la varianza de los errores ε, dicho de otra manera, el ruido que se produce en torno a la recta de regresión calculada (Montgomery y Runger, 1999). Tal estimador puede ser obtenido a partir de la suma de cuadrados residuales (SS_{res}) del modelo de regresión lineal como se muestra a continuación:

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \varepsilon'\varepsilon \tag{3.12}$$

Si se sustituye el valor de $\varepsilon = Y - X\hat{\beta}$ y se desarrolla la ecuación resultante, se obtiene la expresión:

$$SS_{res} = Y'Y - \hat{\beta}'X'Y \tag{3.13}$$

De acuerdo con (Montgomery y otros, 2001), la suma de cuadrados residuales tiene $n - (k + 1)$ grados de libertad asociados a ella, debido a que se estiman (k+1) parámetros en el modelo de regresión lineal correspondiente. De esta manera, el cuadrado medio de los residuales (MS_{res}) puede ser determinado como:

$$MS_{res} = \frac{SS_{res}}{n - (k + 1)} \tag{3.14}$$

También se demuestra que el valor esperado de MS_{res} es σ^2, por lo que el estimador insesgado de σ^2 es

$$\hat{\sigma}^2 = MS_{res} \tag{3.15}$$

3.3. Selección de variables

En la sección anterior se presupone que todas las variables regresoras incluidas en el modelo de regresión lineal son importantes. Sin embargo, en la mayoría de los problemas prácticos el analista tiene un conjunto muy extenso de variables candidatas, dentro de las cuales deberá determinar un subconjunto de variables regresoras ideales para el modelo de regresión lineal. La definición de un subconjunto adecuado de variables regresoras para el modelo de regresión lineal, es llamado problema de selección de variables (Montgomery y otros, 2001).

La construcción de un modelo de regresión lineal que sólo incluya el subconjunto de regresores ideales para el estudio implica dos objetivos contrapuestos:

- El primero de ellos desea que el modelo incluya tantos regresores como sea posible, para que el valor a predecir de y tenga la mayor influencia de las variables regresoras.
- El segundo desea que el modelo incluya la menor cantidad de regresores posibles, pues la varianza de la predicción y aumenta a medida que aumentan los regresores.

Una observación importante es que mientras más regresores haya en un modelo, los costos de recolección de datos y mantenimiento del modelo de regresión lineal serán mayores. Sin embargo, en la mayoría de casos se desea encontrar la mejor ecuación de regresión lineal o aquella que pueda describir de la mejor manera los datos en cuestión, y en el mejor de los casos, será aquella que se encuentra en un término medio entre los dos objetivos de la selección de variables.

Es importante mencionar que ninguno de los procedimientos de selección de variables, garantiza la producción de una ecuación de regresión lineal óptima para un determinado conjunto de datos. Queda completamente a consideración del analista usar los procedimientos de selección que mejor convengan al caso de estudio (Cox y Snell, 1974).

3.3.1. Criterios para evaluar modelos de regresión lineal con subconjuntos de variables

Uno de los aspectos clave en la selección de variables es determinar si un subconjunto de variables regresoras es mejor que otro, para ello se tienen diversos criterios para evaluar y comparar los diferentes modelos de regresión lineal con subconjuntos.

Estos criterios de evaluación tratan de determinar qué modelo de regresión lineal es el que mejor se adapta a los datos históricos de estudio, o bien, en qué medida se pueden describir los datos utilizando un modelo específico.

Coeficiente de determinación múltiple

Una medida de evaluación de un modelo de regresión lineal usado con frecuencia, es el coeficiente de determinación múltiple R^2, el cual se describe como sigue.

Sea R_p^2 el coeficiente de determinación múltiple para un modelo de regresión con subconjunto de p términos, es decir, $p-1$ regresores y un término β_0 de ordenada al origen. El coeficiente se calcula como:

$$R_p^2 = \frac{SS_R(p)}{SS_T} \tag{3.16}$$

Donde, $SS_R(p)$ y SS_T representan la suma de cuadrados de la regresión y la suma total de cuadrados, respectivamente. Dichos valores son obtenidos mediante:

$$SS_R(p) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.17}$$

$$SS_T = SS_R(p) + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.18}$$

Donde, y_i es el *i-ésimo* valor de la variable respuesta, $\hat{y}_i$ es el *i-ésimo* valor estimado de la variable respuesta, y $\bar{y}$ representa la media de la misma variable.

Generalmente, el valor del coeficiente de determinación múltiple R_p^2, se encuentra en el rango [0 1]. Por tanto, para valores de R_p^2 cercanos a uno, implica que la mayor parte de la variabilidad de la variable respuesta y está explicada en el modelo de regresión lineal (Aitkin, 1974).

Estrictamente, R^2 no mide la adecuación del modelo lineal porque con frecuencia R^2 es grande aún cuando las variables predictoras no tengan relación directa con la variable respuesta, es decir, aunque el valor R^2 sea grande, el modelo de regresión lineal no necesariamente será un predictor exacto (Montgomery y otros, 2001).

Según datos de (Rojo-Abuín, 2007), los valores del coeficiente R^2 pueden clasificarse como se muestra en el Cuadro 3.2.

Cuadro 3.2: Clasificación de los coeficientes de determinación múltiple.

Menor de 0.3	0.3 a 0.4	0.4 a 0.5	0.5 a 0.85	mayor a 0.85
Muy malo	Malo	Regular	Bueno	Sospechoso

Coeficiente de determinación múltiple ajustado

Para evitar las dificultades en la interpretación de R^2, algunos analistas prefieren utilizar la estadística R^2 ajustada, la cual se define para una ecuación de p términos y n observaciones de la siguiente manera:

$$R^2_{Aj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p) \tag{3.19}$$

De esta manera, el criterio para seleccionar un modelo con el subconjunto óptimo de variables es elegir aquel que tenga el valor máximo de $R^2_{Aj,p}$ (Haitovsky, 1969).

Cuadrado medio de residuales

El cuadrado medio de residuales para un modelo de regresión lineal con un subconjunto de p variables, se puede utilizar como criterio para la evaluación de éste. Dicho valor es definido por la expresión:

$$MS_{res}(p) = \frac{SS_{res}(p)}{n-p} \tag{3.20}$$

Donde, $MS_{res}(p)$ representa el cuadrado medio de los residuales para el subconjunto con p variables, y $SS_{res}(p)$ la suma de cuadrados residuales también para dicho subconjunto.

Los investigadores que optan por utilizar este criterio de evaluación de variables, grafican los valores de $MS_{res}(p)$ en función de las p variables y basan su elección en:

1. El valor de $MS_{res}(p)$ mínimo.

2. El valor de p tal que $MS_{res}(p)$ sea aproximadamente igual a MS_{res} para el modelo completo.

3. Un valor de p cercano al punto donde el $MS_{res}(p)$ mínimo tiende a crecer.

Cabe mencionar que el modelo de regresión lineal para el subconjunto con p variables, que minimiza $MS_{res}(p)$, también maximizará $R^2_{Aj,p}$.

Estadística C_p de Mallows

En (Mallows, 1964), el autor propuso un criterio que se relaciona directamente con el error cuadrático medio de un valor ajustado, es decir:

$$E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 - Var(\hat{y}_i) \tag{3.21}$$

Nótese que $E(y_i)$ es la respuesta esperada de la ecuación verdadera de regresión, y que $E(\hat{y}_i)$ es la respuesta esperada con el modelo de regresión ajustado al subconjunto de p términos. Entonces, $E(y_i) - E(\hat{y}_i)$ es el sesgo en el *i-ésimo* punto de los datos, en consecuencia, los términos del lado derecho de la ecuación 3.21 son los componentes de sesgo cuadrado y la varianza del error cuadrático medio, respectivamente. Por tanto, el criterio de Mallows puede ser calculado como:

$$C_p = \frac{SS_{res}(p)}{\hat{\sigma^2}} - n + 2p \tag{3.22}$$

En este caso, se prefieren los valores pequeños de C_p para aceptar el subconjunto de variables predictoras.

Con frecuencia se usa el cuadrado medio de los residuales como estimador insesgado de σ^2, sin embargo, esto hace que para la ecuación completa con $p = K + 1$ variables, el estimador σ^2 tenga sesgo despreciable y los valores de C_p sean pequeños, en tal caso, es recomendable utilizar otro estimador insesgado para el cálculo del criterio C_p de Mallows.

3.3.2. Técnicas computacionales de selección de variables

Para determinar el conjunto de variables que se van a usar en la ecuación de regresión lineal, es natural considerar el ajuste de modelos con diferentes combinaciones de regresores candidatos. Las técnicas de selección de variables más utilizadas en el área de regresión lineal son presentadas a continuación.

Todas las regresiones posibles

Este procedimiento requiere que el analista calcule todas las ecuaciones de regresión lineal posibles, es decir aquellas que tengan un regresor candidato, dos regresores candidato, etc. Todas esas ecuaciones se evalúan de acuerdo con algún criterio y se selecciona el modelo de regresión que mejor se ajuste a los datos según los diferentes criterios.

Como la evaluación de todos los regresores posibles resulta muy costoso computacionalmente, se han desarrollado varios métodos para evaluar solo una pequeña cantidad de modelos de regresión lineal utilizando un subconjunto de variables regresoras, agregando o eliminando estas. Este tipo de métodos son llamados "procedimientos paso a paso" y se clasifican principalmente en tres tipos: selección por agregación, selección por eliminación y regresión por segmentos (Montgomery y otros, 2001).

Selección de variables por agregación

Este procedimiento comienza con la hipótesis de que no existen variables regresoras en el modelo de regresión, aparte de la ordenada al origen. Se intenta determinar el subconjunto óptimo de variables, insertando una a una las variables candidatas a la ecuación de regresión lineal. El primer regresor que se selecciona para entrar en la ecuación es aquél que tenga la máxima correlación simple con la variable de respuesta y, esto es, el máximo índice de correlación de Pearson, definido como:

$$r = \frac{S_{xy}}{S_x S_y} \tag{3.23}$$

Siendo S_{xy} la covarianza de (x, y) y S_x, S_y las desviaciones estándar de las variables independiente y dependiente, respectivamente.

El segundo regresor que se escoge para entrar al modelo de regresión es el que presente la máxima correlación con la variable respuesta y, después de ajustar dicha variable y por el efecto del primer regresor introducido (Montgomery y otros, 2001). A este tipo de correlaciones se les llama "correlaciones parciales" y se definen como sigue:

$$F_j = \frac{SS_R(\beta_j|\beta_1)}{MS_{res}(x_j, x_1)} \tag{3.24}$$

Donde $SS_R(\beta_j|\beta_1)$ representa la suma de cuadrados de la regresión al introducir la variable j dado que el modelo de regresión lineal contiene a la primer variable regresora. $MS_{res}(x_j, x_1)$ indica el cuadrado medio de los residuales para el modelo de regresión que incluye ambas variables de regresión.

La variable candidata formará parte del modelo siempre y cuando el valor de la correlación parcial sea mayor a un valor predefinido F_{in}, de lo contrario el algoritmo termina y el modelo se queda con las variables contenidas hasta ese momento.

Selección de variables por eliminación

En la selección por eliminación se pretende determinar un buen modelo trabajando en dirección contraria a la selección por agregación, es decir, se comienza con un modelo que incluye el total de k regresores candidatos. A continuación, se calcula la correlación parcial de cada regresor, como si fuera la última variable que entró al modelo. La correlación mínima se compara con un valor preseleccionado F_{out}, y si es menor que dicho valor, se quita ese regresor del modelo. Se ajusta el modelo de regresión con $k-1$ regresores y se repite el proceso. El algoritmo termina cuando el valor mínimo de correlación parcial es mayor que F_{out}.

Regresión por segmentos

La regresión por segmentos es una modificación de la selección por agregación, en la que, cada paso reevalúa los regresores que han entrado al modelo utilizando las correlaciones parciales. Un regresor que fue agregado en una etapa anterior puede volverse redundante debido a la relación entre él y los nuevos regresores en la ecuación, por lo tanto, puede ser eliminado del modelo de regresión lineal.

En la regresión por segmentos se requiere de dos valores de corte, F_{in} y F_{out}. Algunos analistas prefieren definir $F_{in} = F_{out}$, aunque no es necesario, pues la mayoría opta por tomar $F_{in} > F_{out}$, con lo que se garantiza que es más difícil agregar un regresor que eliminar uno.

Capítulo 4

Regionalización de caudales

4.1. Introducción

Como se mencionó en el capítulo 1, la regionalización de caudales permite estimar los flujos de agua en cuencas no aforadas. Esta técnica hidrológica, se apoya a su vez en dos métodos de procesamiento de datos: La identificación de regiones hidrológicas con características similares o regiones homogéneas y la aplicación de un un método de estimación regional para la transferencia de información entre las regiones homogéneas encontradas.

El presente trabajo tiene como objetivo principal obtener las regiones hidrológicamente homogéneas para la Mixteca Oaxaqueña y obtener un modelo regional para la estimación de caudales máximos, mínimos y medios para cada una de las regiones homogéneas obtenidas. Dicha regionalización de caudales servirá como base para evaluar la disponibilidad de agua superficial en la región.

4.2. Caso de Estudio

La Región geográfica de la Mixteca se encuentra ubicada al sur del país y abarca los estados de: Puebla, Guerrero y Oaxaca, siendo este último el que ocupa la mayor parte del territorio mixteco. Se le llama así, principalmente por el establecimiento en esta área de la cultura Mixteca. La región se divide en tres partes debido a su margen de altitud: Mixteca baja, Mixteca alta y Mixteca de la costa. La primera compartida por los estados de Puebla y Oaxaca. La segunda y tercera división compartida por los estados de Guerrero y Oaxaca.

En términos de hidrología, la Mixteca cubre tres regiones hidrológicas a nivel nacional: la costa chica del río Verde, el río Balsas y el río Papaloapan (INEGI, 1988).

Para el análisis de regionalización, inicialmente se consideraron las 28 estaciones hidrométricas mostradas en el Cuadro 4.1, las cuales estuvieron funcionando en la región de la Mixteca Oaxaqueña y zonas aledañas. Sin embargo, se puede observar que muchas de ellas presentan grandes periodos de discontinuidad en el registro de sus caudales, por lo que se decidió prescindir de ellas en la regionalización.

Cuadro 4.1: Estaciones hidrométricas en la Mixteca. La columna Periodos indica los registros disponibles actualmente.

Clave	**Nombre**	**Periodos**	**Años**	**Total**
18337	Camotlan	1965	1	
		1967-1968	2	3
18338	Xatan	1967-1968	2	2
18342	Teponahuazo	1964-1978	15	
		1986-1992	7	22
18344	Mariscala	1966-1980	15	
		1988-1990	3	18
18348	Tonala	1964-1966	3	
		1971	1	
		1978-1981	4	
		1986-1989	4	12
18352	San Mateo	1971-1980	10	
		1987-1988	2	12
18354	Huajuapan de León	1966	1	
		1986-1987	2	
		1989-1994	6	9
18361	Tonahuixtla	1965-1967	3	
		1976-1977	2	
		1981-1982	2	
		1992	1	8
18374	La Huertilla	1965	1	
		1967	1	
		1969	1	
		1970-1972	3	
		1974	1	
		1976-1982	7	14

Cuadro 4.1 (continuación): Estaciones hidrométricas en la Mixteca.

Clave	Nombre	Periodos	Años	Total
18432	Ixcamilca	1953-1977	25	
		1981-1990	10	
		1992-1994	3	38
18433	Tamazulapan	1955-1967	13	
		1970-1977	8	21
18538	Tezoatlán	1973-1980	8	
		1987-1991	5	13
18573	Yundoo	1980-1982	3	
		1984-1987	4	7
20021	Ixtayutla	1961-1985	25	
		1987-1991	5	30
20025	Las Juntas	1973	1	
		1978-1981	4	
		1986-1987	2	
		1995	1	8
20029	Zacoalpan	1957	1	1
20033	Xochistlahuaca	1954-1968	19	19
20034	Nduave	1954	1	
		1956-1966	11	12
20041	Nusutia	1970-1983	14	
		1985-1989	5	19
20042	Yutacua	1970-1980	11	11
28064	Sto Domingo	1974-1975	2	
		1977-1987	11	13
28070	Calapilla	1955-1969	20	20
28072	Xiquila	1955-1980	25	25
28082	Apoala	1957-1979	23	23
28102	Axusco	1959	1	
		1963-1973	11	
		1975-1978	4	16
28104	Tomellín	1959-1960	2	
		1962-1964	3	
		1966-1968	3	
		1970-1982	13	
		1984	1	22
28131	Parian	1974-1980	7	7
28182	Stgo Apoala	1974-1978	5	5

Tomando en cuenta los caudales registrados dentro de periodos de tiempo similares, así como la posición geográfica de las estaciones hidrométricas, se limitó la zona de estudio a la Mixteca Oaxaqueña. La cual tiene una superficie total de $23,723\ km^2$ y está localizada entre las coordenadas geográficas 16° 30′ 24,3″, 18° 30′ 51,7″ de latitud norte y 96° 52′ 8,7″, 98° 54′ 21,8″ de longitud oeste, como se muestra en la Figura 4.1.

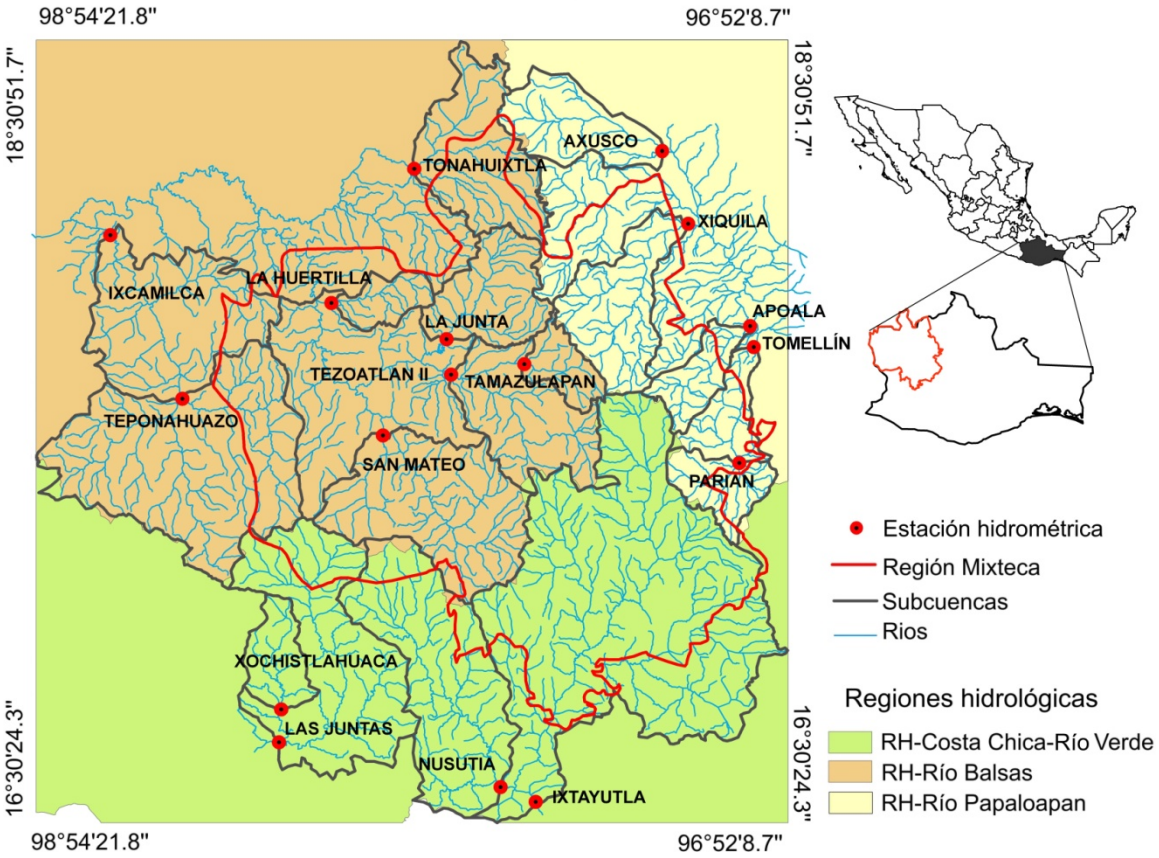

Figura 4.1: Localización geográfica de la zona de estudio.

Es posible observar que la zona de estudio comprende 8 subcuencas de la región hidrológica del río Balsas, 5 del río Papaloapan y 4 subcuencas del río Verde, haciendo un total de 17 subcuencas de estudio. Sin embargo, de las 17 subcuencas que forman parte de la Mixteca Oaxaqueña, no todas cuentan con el registro completo de sus caudales, pues en la mayoría de ellas, la estación hidrométrica asociada ha dejado de funcionar. Según datos del Instituto Mexicano de Tecnología del Agua (IMTA, 1997), el mayor registro de caudales para la región de estudio, se encuentra en la década de los 70's. Además, a la fecha la única subcuenca que registra sus caudales, corresponde a la estación hidrométrica de nombre "Tezoatlán II", perteneciente a la región hidrológica del río Balsas, al norte de la Mixteca Oaxaqueña.

En el Cuadro 4.2, se muestran las 17 estaciones hidrométricas preseleccionadas, con sus correspondientes registros de caudales en la década de los 70's. Los años marcados con "1" indican registros mensuales completos de los datos, los marcados con "2" son años incompletos y los que no tienen ninguna marca indican que no hay registro de caudales para ese año.

Cuadro 4.2: Registro de caudales en la Mixteca Oaxaqueña.

Estación	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Apoala	1	1	1	1	1	1	1	1	1	1	2				
Axusco	1	1	1	1	1	1	1	1	1						
Ixcamilca	1	1	1	1	1	1	1	1				1	1	1	1
Ixtayutla	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1
La Huertilla	2	2	2	2	2	2	2	2	2	1	1	1	2		
La Junta	2	2	2	2	2	2	2	2	1	1	1	1			
Las Juntas	1	1	1	1	1	1	1	1	1	1	1		1	1	
Nusutia	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Parian					1	1	1	1	1	1	1	2			
San Mateo	2	1	1	1	1	1	1	1	1	1	1				
Tamazulapan	2	1	1	1	1	1	1	1							
Teponahuazo	1	1	1	1	1	1	1	1	1						
Tezoatlán II			2	1	1	1	1	1	1	1	1				
Tomellin	1	1	1	1	1	2	1	1	1	1	1	1	2	2	2
Tonahuixtla	1	2	2	2	2	2	1	1	2	2	1	1	2	2	2
Xiquila	1	1	1	1	1	1	1	1	1	1	1				
Xochistlahuaca															

Se puede observar que la estación Xochistlahuaca no cuenta con ningún registro de sus caudales en el periodo mostrado, por lo cual quedó completamente descartada para el análisis de datos. Por su parte, estaciones como Tonahuixtla, La Huertilla y La Junta tienen muchos años incompletos, por lo que también fueron eliminadas para la regionalización de caudales.

De acuerdo con el registro de caudales mostrado en el Cuadro 4.2 y utilizando sólo años con datos completos, se pudieron extraer dos casos de estudio para la regionalización de caudales en la Mixteca Oaxaqueña: El primero, utilizando sólo 5 estaciones hidrométricas y el segundo utilizando 10 estaciones de aforo.

4.2.1. 5 estaciones hidrométricas

En el primer caso de estudio se ocupó el mayor número de años posibles con 5 estaciones de medición de caudales, siempre y cuando los datos estuviesen dentro del mismo periodo de tiempo.

Podemos observar que estaciones como Teponahuazo, San Mateo, Ixtayutla, Las Juntas, Nusutia, Xiquila, Apoala y Axusco, cuentan con el registro completo de sus caudales a partir de 1971, sin interrupciones y por largos periodos. Sin embargo, en las subcuencas de Teponahuazo y Axusco, los registros se terminan en 1978 y en Apoala, un año mas tarde. Basados en estas observaciones, fue posible eliminar las estaciones Teponahuazo, Axusco y Apoala para el primer caso de estudio.

Como resultado, obtuvimos las 5 estaciones hidrométricas con mayor registro común en el periodo que comprende del año 1971 a 1980. Dichas estaciones, contempladas para el primer caso de estudio se detallan en el Cuadro 4.3.

Cuadro 4.3: Caso de estudio con 5 estaciones hidrométricas.

Clave	Estación	Cuenca	Región hidrológica	Estado
20021	Ixtayutla	Río Verde	Balsas	Oaxaca
20025	Las Juntas	Río Ometepec	Costa Chica-Río Verde	Guerrero
20041	Nusutia	Río Yolotepec	Costa Chica-Río Verde	Oaxaca
18352	San Mateo	Río Mixteco	Balsas	Oaxaca
28072	Xiquila	Río Papaloapan	Papaloapan	Oaxaca

4.2.2. 10 Estaciones hidrométricas

En contraste con el primer caso de estudio, se optó por utilizar un menor número de años de registros, con el fin de tener más estaciones de aforo para el análisis y en consecuencia cubrir la mayor parte de la región de estudio.

En este caso, fue posible utilizar las mismas estaciones del primer caso de estudio más las estaciones Teponahuazo, Axusco y Apoala. Esta decisión nos permitiría utilizar los registros correspondientes al periodo 1971-1978.

Por otro lado, con un año menos de datos, es decir utilizando el periodo comprendido entre los años 1971 y 1977, se agregaron las estaciones Ixcamilca y Tamazulapan. Ambas estaciones con registros completos en el periodo señalado.

De este modo, para el segundo caso de estudio se seleccionaron las 10 estaciones hidrométricas mostradas en el Cuadro 4.4.

4.3. Diseño experimental

Como se describió en la sección anterior, en el análisis de datos se consideraron dos casos diferentes de estudio: En el primero de ellos, se utilizaron 5 estaciones hidrométricas y 10 años

Cuadro 4.4: Caso de estudio con 10 estaciones hidrométricas.

Clave	Estación	Cuenca	Región hidrológica	Estado
28082	Apoala	Río Papaloapan	Papaloapan	Oaxaca
28102	Axusco	Río Salado	Papaloapan	Oaxaca
18432	Ixcamilca	Río Mezcala	Balsas	Puebla
20021	Ixtayutla	Río Verde	Balsas	Oaxaca
20025	Las Juntas	Río Ometepec	Costa Chica-Río Verde	Guerrero
20041	Nusutia	Río Yolotepec	Costa Chica-Río Verde	Oaxaca
18352	San Mateo	Río Mixteco	Balsas	Oaxaca
18433	Tamazulapan	Río Salado	Balsas	Oaxaca
18342	Teponahuazo	Río Grande	Balsas	Guerrero
28072	Xiquila	Río Papaloapan	Papaloapan	Oaxaca

de registro de caudales. Por otro lado, en el segundo caso se consideraron 10 estaciones de aforo pero sólo 7 años con registro de caudales. En cada caso, se desean obtener las regiones o cuencas homogéneas y sus correspondientes ecuaciones de regresión lineal, para estimar los caudales máximos, mínimos y medios en diferentes periodos estacionales.

La identificación de regiones hidrológicas con características similares o cuencas homogéneas se llevó a cabo mediante métodos de *clustering* o agrupamiento, debido a su eficiencia en la clasificación de datos no etiquetados, es decir, datos que *a priori* no pertenecen a un grupo específico, tal es el caso, de las estaciones hidrométricas y sus respectivas cuencas hidrológicas.

Una vez identificadas las regiones homogéneas, la estimación de caudales fue modelada mediante análisis de regresión lineal múltiple con selección de variables paso a paso, tomando como predictores las variables climáticas y fisiográficas del área de estudio. Se optó por este método debido a que se desconoce la relación que presentan los caudales con respecto a las variables predictoras de la región.

En este contexto, el método de regresión lineal es comúnmente utilizado para desarrollar relaciones de cuencas con mediciones (Shih-Min y otros, 2002). Además, es la técnica más usada probablemente por su fácil aplicación e interpretación (Heuvelmans y otros, 2006). No obstante, el análisis de variables hidrológicas puede conseguirse con mejores técnicas, aunque éstas pudieran ser más complejas, que permitan estimar los caudales con menor incertidumbre, tal como se propone en (Heuvelmans y otros, 2006), donde los autores utilizan redes neuronales artificiales para la regionalización de parámetros en un modelo hidrológico.

4.3.1. Esquema general

A lo largo del presente documento, se ha planteado la regionalización de caudales en dos fases diferentes. En primer lugar, la delimitación de regiones homogéneas y posteriormente, la generación de un modelo de estimación de caudales por medio de una ecuación de regresión lineal. Sin embargo, cada una de estas fases se puede subdividir en módulos más pequeños de procesamiento de datos, con la finalidad de garantizar el correcto funcionamiento de las mismas. Por tanto, tomando como base nuestra área de estudio y detallando más cada una de las fases de regionalización, se generó el esquema de desarrollo mostrado en la Figura 4.2.

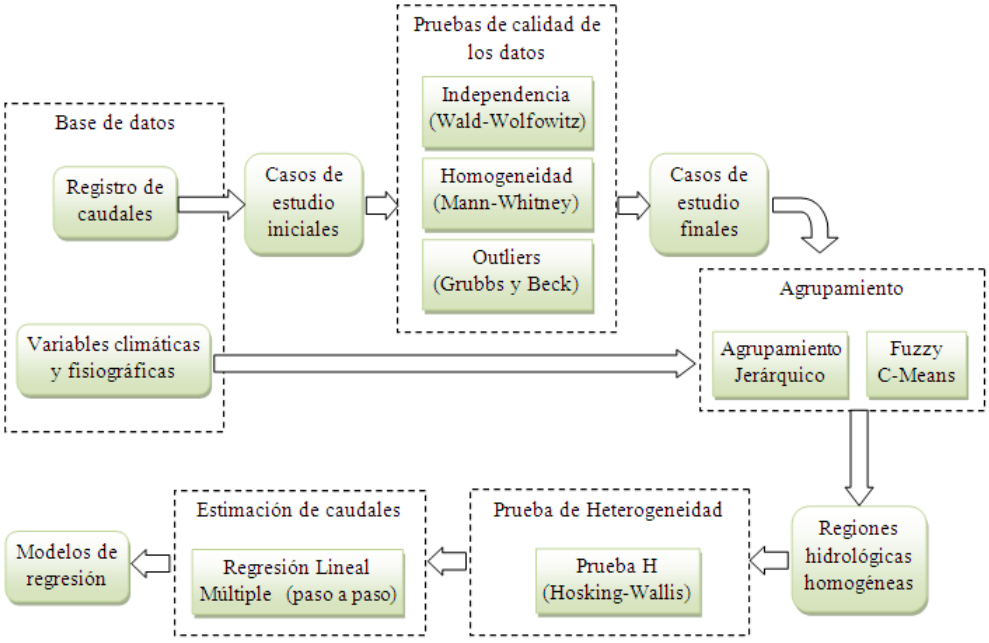

Figura 4.2: Esquema general de desarrollo para la regionalización de caudales en la Mixteca Oaxaqueña.

Se observa que, previamente a la identificación de regiones homogéneas, fue necesario aplicar pruebas de calidad a los datos hidrológicos (ver sección 2.2). Es decir, una vez que se tuvieron seleccionadas las estaciones para el análisis, la calidad de los datos hidrológicos fue revisada mediante la aplicación de tres pruebas: En primer lugar, se aplicó la prueba de independencia de Wald-Wolfowitz para verificar la aleatoriedad de los registros de caudales en cada estación de aforo y asegurar, en cierta forma, que los datos hallan sido tomados de una medición real y no de un caso artificial de caudales. La segunda prueba que se aplicó fue la de Mann-Whitney para corroborar la homogeneidad de los registros. Por último, se aplicó la prueba de Outliers de Grubbs y Beck para identificar los datos incongruentes en los caudales de cada estación. Todas estas pruebas fueron aplicadas con un nivel de significancia estadística del 5 %.

Ya que fue analizada la calidad de los datos hidrológicos, se pudieron aplicar los métodos de *clustering* utilizando las variables hidrológicas, climáticas y fisiográficas del estudio. Más puntualmente, se implementaron los métodos de Agrupamiento jerárquico aglomerativo y Fuzzy C-Means para encontrar las regiones hidrológicamente homogéneas. El agrupamiento jerárquico se implementó usando las funciones de enlace tipo promedio, centroide y *ward* para definir la distancia entre los grupos del análisis (ver sección 2.3). Por su parte, el agrupamiento Fuzzy C-Means fue aplicado con 2, 3, 4 y 5 grupos predefinidos; además se aplicaron cuatro índices de validación de *clusters* con la intención de determinar el mejor agrupamiento de los datos (ver sección 2.4).

De antemano, se asume que las regiones encontradas por los métodos de *clustering* eran por si mismas homogéneas. No obstante, fue necesario verificar dicha homogeneidad para cada una de ellas. En este sentido, las pruebas de heterogeneidad H (ver sección 2.5), propuestas por Hosking y Wallis, permitieron verificar la correspondiente homogeneidad de las regiones halladas por los algoritmos de agrupamiento.

Con la seguridad que las regiones encontradas eran internamente homogéneas, fue factible implementar el método de regresión lineal para obtener los modelos de predicción de caudales, específicamente se utilizó la técnica de regresión lineal paso a paso con selección de variables por agregación (ver capítulo 3). De esta manera, se obtuvieron modelos de regresión que no incluyeran todas las variables regresoras del estudio, pues mientras más regresores haya en un modelo, los costos de recolección de datos y mantenimiento del mismo, serán mayores.

4.3.2. Base de datos

Los datos utilizados, corresponden a las subcuencas descritas en los Cuadros 4.3 y 4.4. Para cada una de ellas, se extrajeron los registros mensuales de sus caudales y se cuantificaron los valores medios de 14 variables climáticas y fisiográficas, potencialmente útiles para la predicción de gastos (caudales) en la región. Dichas variables son mostradas en el Cuadro 4.5.

Los registros mensuales de los caudales máximos, mínimos y medios de cada subcuenca fueron extraídos del Sistema de Información de Aguas Superficiales, editado por el Instituto Mexicano de Tecnología del Agua (IMTA, 1997).

Por su parte, las tres variables climáticas (lluvia media mensual, precipitación media anual y temperatura) se obtuvieron de las series diarias de lluvia y temperatura del Extractor Rápido de Información Climatológica (ERIC III) del IMTA, que almacena los datos históricos del servicio meteorológico nacional de la CONAGUA, contenida en la base de datos CLICOM tal como se encontraba en enero de 2007 (IMTA, 2007).

Cuadro 4.5: Variables climáticas y fisiográficas del estudio.

Variable	Unidad	Tipo
lluvia media mensual	mm	Climática
longitud del cauce principal	km	Fisiográfica
covertura vegetal	%	Fisiográfica
temperatura		Climática
precipitación media anual	mm	Climática
área de la cuenca	km^2	Fisiográfica
densidad de drenaje	km^{-1}	Fisiográfica
elevación media de la cuenca	m	Fisiográfica
coeficiente de escurrimiento	mm	Fisiográfica
elevación máxima de la cuenca	m	Fisiográfica
elevación mínima de la cuenca	m	Fisiográfica
latitud	-	Fisiográfica
longitud	-	Fisiográfica
lluvia máxima anual en 24 horas con periodo de retorno de 2 años	mm	Fisiográfica

Finalmente, las variables fisiográficas se estimaron a partir de imágenes del satélite LANDSAT de 1979 (Barker, 1970) y mediante el procesamiento de información topográfica obtenida de (INEGI, 2000).

Es importante señalar que, con el fin de eliminar los problemas de escala y unidades, los datos originales fueron estandarizados mediante la expresión:

$$y_{i,j} = \frac{x_{i,j} - \bar{x}_i}{S_x} \tag{4.1}$$

donde $x_{i,j}$ representa el valor de la j-ésima estación en la i-ésima variable, $\bar{x}_i$ es la media de la variable i, S_x representa la desviación estándar y $y_{i,j}$ es la representación de la j-ésima estación en la i-ésima variable transformada.

4.4. Implementación del sistema de regionalización

Como parte del desarrollo de la tesis, se implementó una pequeña biblioteca de *clustering*. En ella, se desarrollan los métodos de agrupamiento utilizados para el análisis de caudales en la Mixteca Oaxaqueña, así como las pruebas de preprocesamiento y postprocesamiento de los datos. También se implementó el método de regresión lineal múltiple por mínimos cuadrados, en este caso se utilizaron las 14 variables climáticas y fisiográficas como regresores del modelo.

La biblioteca fue desarrollada en el lenguaje C++ y consta de 12 clases para su correcta portabilidad en cualquier programa desarrollado bajo este estándar de programación.

De las 12 clases desarrolladas, 3 de ellas hacen referencia a las pruebas de calidad de los datos. La clase "Wolfowitz" referente a la prueba de aleatoriedad e independencia en los caudales, la clase "Whitney" que implementa la prueba de homogeneidad en las estaciones hidrométricas y la clase "Grubbs" que detecta los outliers o puntos atípicos en los caudales de cada estación. En cada una de estas clases fué implementado un método "run()" que es el encargado de realizar todas las operaciones necesarias para la respectiva prueba de calidad.

Por otra parte, la clase "Hierarchical" y "Fuzzy" implementan los métodos de agrupamiento jerárquico aglomerativo y Fuzzy C-Means, respectivamente. La clase correspondiente al agrupamiento jerárquico puede realizar el análisis de los datos con las seis diferentes funciones de enlace mostradas en la sección 2.3.2. En lo referente a la clase de agrupamiento Fuzzy C-Means, ésta es capaz de obtener los cuatro índices de validación de *clusters* descritos en la sección 2.4.4 y realiza la defusificación de datos utilizando el método de máxima pertenencia.

Una clase más de nombre "Heterogeneity", es la encargada de verificar la homogeneidad de los grupos formados por los algoritmos de agrupamiento. Para cada grupo de estaciones, la clase es capaz de calcular los tres índices de heterogeneidad H_1, H_2 y H_3 descritos en la sección 2.5.

La clase "Regression" es la encargada de encontrar los modelos de estimación de caudales para cada grupo encontrado. Sin embargo, el modelo resultante fue difícil de implementar debido a su complejidad en la recolección de las 14 variables involucradas en el mismo. De forma alternativa, se decidió utilizar el paquete de software Minitab®, en su versión 15.0, para construir los modelos de estimación con técnicas de regresión paso a paso y utilizando los datos resultantes del proceso de agrupamiento.

El resto de las clases de la biblioteca, son objetos auxiliares para el almacenamiento de datos en memoria, asi como para la lectura y despliegue de los mismos en pantalla. Por ejemplo las clases "Cluster", "Instance" y "Matrix" son utilizadas para almacenar datos en memoria y las clases "Data" y "Screen" sirven para leer datos de un archivo de texto plano y escribir datos en pantalla, respectivamente.

La biblioteca completa fue programada sobre el sistema operativo ubuntu 10.10 de linux y compilado con la versión 4.4.5 de gcc (g++). Todo funcionando en una computadora portátil Dell XPS m1330 con dos procesadores a 2.0Ghz y 3Gb en memoria RAM.

Los requerimientos mínimos de hardware para utilizar la biblioteca quedan completamente en dependencia de la base de datos a procesar. En nuestro caso, con 10 estaciones hidrométricas y datos mensuales en el periodo 1971-1980 para cada una de ellas, el algoritmo de agrupamiento jerárquico, por ser de orden cuadrático, llegó a consumir un poco más de 10 minutos en el hardware especificado anteriormente.

Capítulo 5

Resultados

Como se describió en la sección 4.3.1, la metodología a seguir para la regionalización de caudales en la Mixteca Oaxaqueña fue la siguiente. Primero se aplicaron las pruebas de calidad a los datos hidrológicos. Posteriormente se identificaron las regiones homogéneas por medio de los algoritmos de agrupamiento. Luego, por cada región encontrada se verificó que dicha región fuese efectivamente homogénea y finalmente, se obtuvieron los modelos de regresión lineal para estimación de caudales en cada región homogénea encontrada.

En el presente capítulo se muestran los resultados obtenidos en cada una de estas etapas de desarrollo.

5.1. Resultados experimentales

Como primer paso, se aplicaron las pruebas de calidad a los datos de las 10 estaciones de estudio. Recordemos que si bien se formaron dos casos de análisis, las estaciones hidrométricas usadas en el primer caso de estudio (utilizando 5 estaciones), también fueron consideradas para el segundo análisis. Por tanto, para las estaciones que estaban incluidas en ambos casos de estudio (Ixtayutla, Las Juntas, Nusutia, San Mateo y Xiquila), se utilizaron sus registros de caudales en el periodo 1971-1980, y para el resto de las estaciones hidrométricas se utilizaron sólo sus caudales en el periodo 1971-1977. Estas pruebas de calidad, también se aplicaron a los datos mensuales correspondientes a la lluvia media de cada estación de aforo.

Los resultados de aplicar la prueba de independencia y aleatoriedad de Wald-Wolfowitz, con un nivel de significancia del 5 %, se muestran en el Cuadro 5.1.

Se puede observar que ningún valor de la tabla sobrepasa 1.96, que representa el valor crítico correspondiente a un nivel de significancia del 5 %, en una distribución normal estándar (ver sección 2.2). Esto quiere decir, que todas las estaciones hidrométricas cumplen con la

Cuadro 5.1: Resultados de la prueba de independencia.

	Valor del estadístico $\|Z\|$			
Estación	**Caudal máx.**	**Caudal mín.**	**Caudal med.**	**Lluvia med.**
Apoala	0.93	1.55	1.62	1.18
Axusco	0.60	1.83	0.97	0.83
Ixcamilca	0.87	1.18	1.18	1.06
Ixtayutla	1.32	1.30	1.31	0.84
Las Juntas	1.22	1.39	1.38	1.30
Nusutia	1.39	1.48	1.38	1.15
San Mateo	0.82	1.31	1.35	0.79
Tamazulapan	1.06	0.96	0.60	0.62
Teponahuazo	1.33	1.43	1.31	0.84
Xiquila	1.08	1.57	1.27	1.15

característica de aleatoriedad e independencia en sus registros de caudales y lluvia media. Sólo la estación Apoala se acerca al límite aceptado con un 5 % de significancia estadística, sobre todo en las series mensuales de caudal mínimo y caudal medio.

Con la verificación de que los datos cumplen con el supuesto de aleatoriedad, se puede suponer que los registros de caudales y lluvia media, objetos de nuestro estudio, fueron tomados en realidad en una estación hidrométrica y no fueron generados por una persona o software para simular dichos registros.

Una vez confirmado que los caudales y lluvia media son aleatorios e independientes, se aplicó la prueba de homogeneidad de Mann-Whitney. Esto, para asegurar que los datos fueron extraídos de una misma fuente y por tanto pertenecen estadísticamente a la misma población. La aplicación de esta prueba se realizó también con un nivel de significancia estadística del 5 %, y los resultados son mostrados en el Cuadro 5.2.

Se observa que tanto el caudal mínimo como el caudal medio de la estación Apoala no cumplen con la prueba de homogeneidad, pues sobrepasan el umbral de 1.96, sin embargo como el caudal máximo y la lluvia media pasan la prueba, se decide dejar la estación para el proceso de agrupamiento. Un caso similar se presenta con las estaciones Axusco y Las Juntas, donde las variables de caudal mínimo y lluvia media, respectivamente, son las que no aprueban el test.

El siguiente paso fue aplicar la prueba de outliers, para identificar los caudales y registros de lluvia media, que se desvían considerablemente de su distribución. Se utilizó para este propósito, la prueba de Grubbs y Beck con un nivel de significancia estadística del 5 %. La prueba fue implementada en cada uno de los caudales y la lluvia media de todas las estaciones del estudio.

Cuadro 5.2: Resultados de la prueba de homogeneidad.

	Valor del estadístico $\|u\|$			
Estación	**Caudal máx.**	**Caudal mín.**	**Caudal med.**	**Lluvia med.**
Apoala	1.51	3.58	3.18	0.90
Axusco	0.77	3.36	1.66	1.14
Ixcamilca	0.88	1.03	1.20	0.35
Ixtayutla	0.57	1.50	1.06	0.60
Las Juntas	0.15	0.88	0.66	2.30
Nusutia	0.04	1.19	0.52	0.39
San Mateo	0.25	1.68	0.55	0.19
Tamazulapan	0.06	0.83	0.30	0.54
Teponahuazo	1.04	1.37	1.47	0.06
Xiquila	0.76	0.50	0.59	0.66

Los resultados correspondientes a esta prueba son presentados a partir del Cuadro 5.3, donde se muestran los outliers detectados para el caudal máximo, hasta el Cuadro 5.6, que indica los outliers hallados para la serie de lluvia media de cada estación hidrométrica.

Como se puede observar, se encontraron en su mayoría outliers mínimos (valores menores al límite inferior XL), sólo en la estación Axusco se identificaron tres caudales mayores al limite superior (XH), esto en las series de caudales mínimos. Dichos outliers se eliminaron de la base de datos para evitar que éstos causaran ruido en la etapa de agrupamiento.

En el caso de los caudales máximos de cada estación hidrométrica (Cuadro 5.3), no se hallo ningún dato fuera de los límites superior e inferior (XH y XL, respectivamente). Esto refuerza la idea de que los datos fueron realmente tomados de la estación de aforo y no fueron generados por un agente externo al sistema de medición. También nos asegura que los datos no presentan errores en la medición, procesamiento y captura de los mismos.

En el caso de las series mensuales del caudal mínimo (Cuadro 5.4), se encontraron 12 datos fuera de los límites permitidos, de los cuales sólo tres estaban por encima del límite superior (XH). Éstos últimos se encontraron en la estación Axusco en el año de 1973, en los meses de octubre, noviembre y diciembre. Lo cual resulta un tanto extraño, ya que estos registros están fuera de la época de lluvia, cuando se presentan los eventos de mayor magnitud. El resto de los outliers fueron mínimos y en su mayoría con valores cero. Por lo que se decidió quitarlos para evitar inconsistencias en la fase de agrupamiento.

En lo que respecta a las series mensuales de caudal medio (Cuadro 5.5), sólo se encontró un outlier mínimo en la cuenca de la estación Apoala en el mes de mayo de 1973. Este registro indudablemente se debería quitar, pues en el mismo mes y año se encontró un outlier mínimo pero en el registro de caudales mínimos.

Cuadro 5.3: Outliers encontrados para el caudal máximo.

Estación	Límites		Outliers	
	XH	XL	Mínimos	Máximos
Apoala	1174.89	0.0051	—	—
Axusco	3764.71	0.0002	—	—
Ixcamilca	6838	0.4017	—	—
Ixtayutla	7270.27	3.4399	—	—
Las Juntas	6799.87	2.3548	—	—
Nusutia	8078.69	1.4782	—	—
San Mateo	9070.09	0.0993	—	—
Tamazulapan	230.86	0.0038	—	—
Teponahuazo	4570.98	0.3949	—	—
Xiquila	6885.69	0.0335	—	—

Cuadro 5.4: Outliers encontrados para el caudal mínimo.

Estación	Límites		Outliers	
	XH	XL	Mínimos	Máximos
Apoala	200.69	0.0001	0 (abril, 73) 0 (mayo,73) 0 (junio,73)	—
Axusco	0.28	0.0109	—	0.3 (octubre, 73) 0.3 (noviembre, 73) 0.3 (diciembre, 73)
Ixcamilca	140.50	0.7825	—	—
Ixtayutla	513.74	2.8984	1.46 (mayo, 78)	—
Las Juntas	586.04	1.7168	—	—
Nusutia	539.22	0.7449	—	—
San Mateo	88.91	0.0874	0.08 (junio, 80)	—
Tamazulapan	5.20	0.0004	0 (julio, 73) 0 (enero,76)	—
Teponahuazo	8816.04	0.0047	0 (mayo, 74) 0 (junio,74)	—
Xiquila	10.89	0.1741	—	—

Cuadro 5.5: Outliers encontrados para el caudal medio.

Estación	Límites		Outliers	
	XH	XL	Mínimos	Máximos
Apoala	114.83	0.0019	0 (mayo,73)	—
Axusco	5.39	0.0036	—	—
Ixcamilca	495.13	0.7824	—	—
Ixtayutla	1217.16	3.7881	—	—
Las Juntas	1278.84	2.0708	—	—
Nusutia	1111.91	1.4864	—	—
San Mateo	331.46	0.1431	—	—
Tamazulapan	4.09	0.0086	—	—
Teponahuazo	430.32	0.5731	—	—
Xiquila	49.87	0.1438	—	—

Cuadro 5.6: Outliers encontrados para la lluvia media.

Estación	Límites		Outliers	
	XH	XL	Mínimos	Máximos
Apoala	4608520	0	—	—
Axusco	679393000	0	—	—
Ixcamilca	247469000	0	—	—
Ixtayutla	35219.5	0.0152	0 (febrero, 71)	—
Las Juntas	5465780000	0	—	—
Nusutia	652615	0.0020	0 (enero, 71)	—
			0 (marzo,71)	
			0 (marzo,77)	
San Mateo	53012800	0	—	—
Tamazulapan	85762200	0	—	—
Teponahuazo	25728900	0	—	—
Xiquila	53759.7	0.0031	0 (febrero, 72)	—
			0 (diciembre,76)	

Finalmente, en los registros de lluvia media (Cuadro 5.6), se encontraron algunos outliers mínimos, todos con valor cero en los meses de diciembre, enero, febrero y marzo. Lo que es razonable, pues en esos meses no llueve en la región de estudio. Sin embargo, se decidió eliminarlos del conjunto de datos para evitar incongruencias en la fase de agrupamiento, pues los valores se encontraban fuera del límite inferior.

En conclusión, y como resultado de la prueba de outliers, se eliminaron tres registros para las estaciones de: Apoala, Axusco y Nusutia; dos para las estaciones de: Ixtayutla, Tamazulapan, Teponahuazo y Xiquila; y finalmente uno de San Mateo.

El siguiente paso, fue la identificación de grupos homogéneos mediante la aplicación de dos técnicas de agrupamiento. La primera, tomando en cuenta que cada observación de datos puede pertenecer a sólo un grupo homogéneo. La segunda, asignando cada observación, o registro de datos, a todos los grupos predefinidos, donde dicha asignación se realiza con un cierto grado de pertenencia a cada grupo.

Cabe señalar que para ejecutar los algoritmos de agrupamiento, el conjunto de datos debió estandarizarse con la ecuación 4.1. Esto se hace con la finalidad de eliminar los problemas de escala y unidades en las 17 variables utilizadas en este proceso: 3 caudales y 14 variables climáticas y fisiográficas por cada estación hidrométrica.

En el caso del agrupamiento jerárquico, se realizaron diferentes pruebas para determinar los tipos de enlace que producían los mejores resultados de agrupamiento, para ello se utilizaron dos conjuntos de prueba. El primer conjunto de datos fue generado de manera artificial con una distribución gaussiana de media 4 y desviación estándar 2, en un espacio bidimensional, formando tres grupos de datos. El segundo conjunto de prueba, fue la base de datos pública "Iris", extraída del repositorio virtual de la UCI (Frank y Asuncion, 2010). La base de datos "Iris"es una de las más utilizadas en la literatura de reconocimiento de patrones. El conjunto de datos contiene tres clases de 50 instancias cada una, donde cada clase hace referencia a un tipo de planta. De acuerdo con la descripción anexada a la base de datos, la primer clase es linealmente separable de las otras dos, pero la última clase difícilmente es separada de las dos primeras.

Tomando en cuenta la distribución del primer conjunto de datos y las características del segundo, se probaron los seis tipos de enlaces para agrupar los datos. En la Figura 5.1 se muestran los resultados para el primer conjunto de prueba y es posible observar que los enlaces promedio, centroide y *ward* son los que mejor construyen el dendograma de datos, mostrando claramente tres grupos en el conjunto. Por su parte, la Figura 5.2 muestra los resultados de agrupar la base de datos "Iris", indicando que los enlaces mediano, promedio, centroide y *ward* son capaces de separar en tres clases a dicho conjunto de datos.

Como resultado de estas pruebas, se determinó que los enlaces promedio, centroide y *ward* son los que mejor agrupan a los conjuntos de prueba. Por tanto, se decidió utilizar estos tres enlaces para agrupar los datos de las estaciones hidrométricas en cada uno de los casos de estudio establecidos.

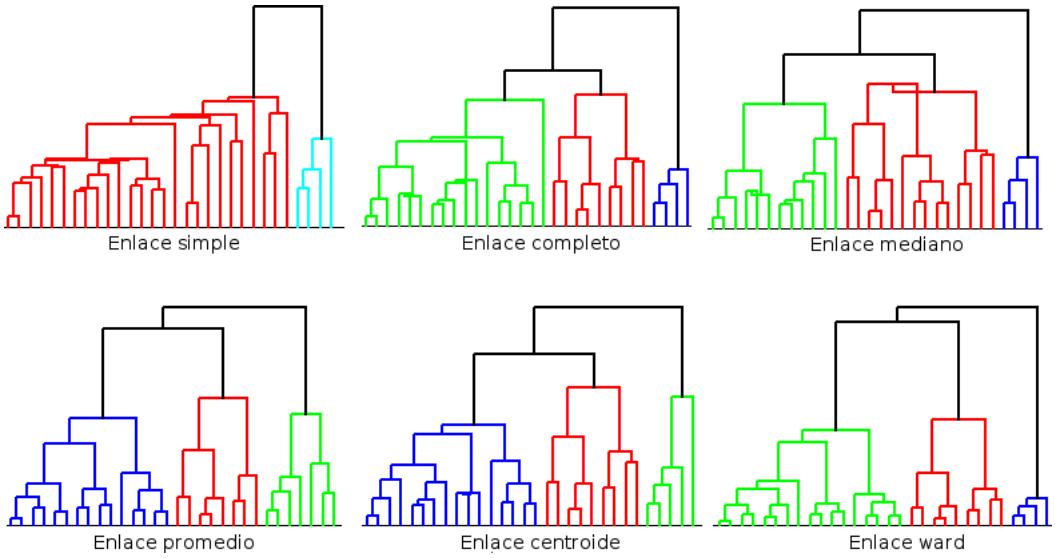

Figura 5.1: Agrupamiento del primer conjunto de prueba.

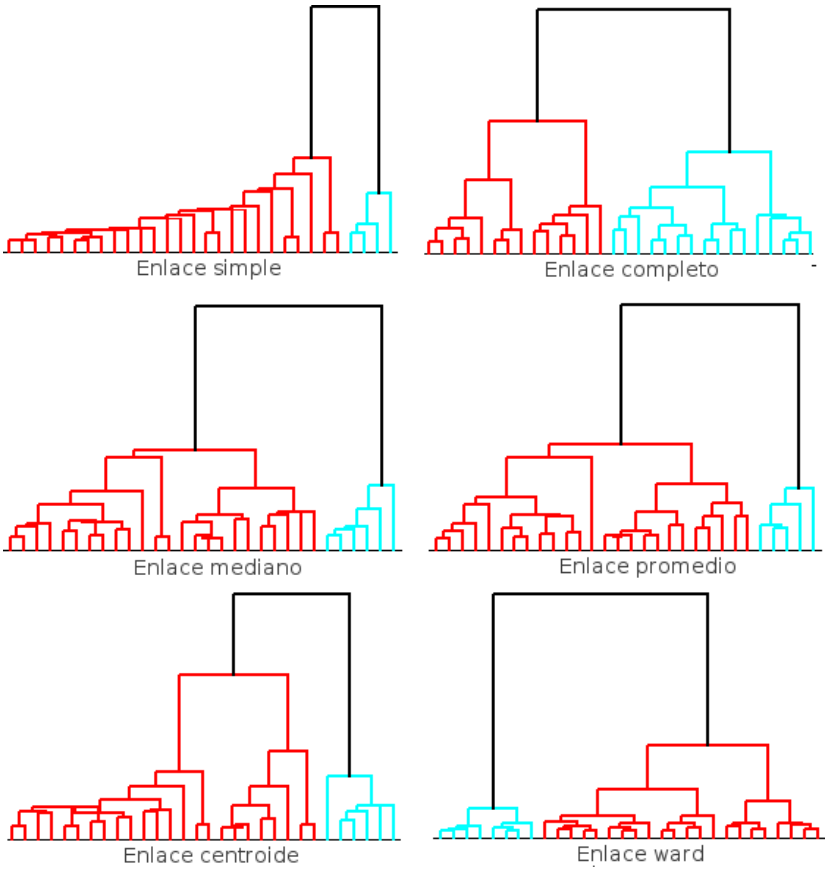

Figura 5.2: Agrupamiento del segundo conjunto de prueba.

De esta manera, la aplicación del algoritmo de agrupamiento jerárquico, utilizando sólo 5 estaciones hidrométricas y los tres tipos de enlace preseleccionados, arrojaron los resultados mostrados en las Figuras 5.3, 5.4 y 5.5.

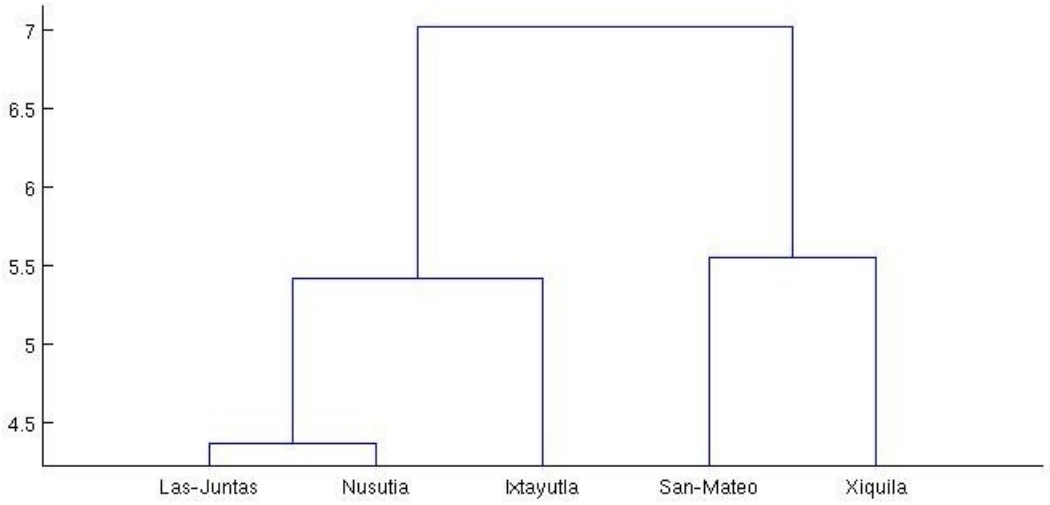

Figura 5.3: Agrupamiento jerárquico con 5 estaciones y enlace promedio.

Con estos resultados, se puede observar claramente que los enlaces promedio (Figura 5.3) y *ward* (Figura 5.5) producen resultados muy similares (con valores aproximados de distancia de corte de 5.7 y 46, respectivamente), en ambos casos se pueden identificar dos grupos homogéneos: El primero incluye las estaciones Las Juntas, Nusutia e Ixtayutla y el segundo grupo encuentra las estaciones San Mateo y Xiquila como homogéneas.

Al aplicar los mismos enlaces del algoritmo de agrupamiento jerárquico, pero con las 10 estaciones del segundo caso de estudio, se obtienen los resultados indicados en las Figuras 5.6, 5.7 y 5.8.

Nuevamente, en cada caso es posible identificar dos grupos, cada uno representa a una región homogénea. La primera región incluye las estaciones Apoala, Axusco, Tamazulapan y Xiquila; la segunda región está formada por las estaciones San Mateo, Teponahuazo, Ixcamilca, Ixtayutla, Las Juntas y Nusutia. Se puede observar que la distancia de corte para los enlaces promedio (Figura 5.6) y centroide (Figura 5.7) es menor a 5.5, mientras que para el enlace *ward* (Figura 5.8), este valor es menor a 50.

A diferencia del primer caso de estudio, se observa que los tres tipos de enlace producen exactamente los mismos grupos. Además, los enlaces promedio y centroide siguen la misma jerarquía en la construcción de las regiones homogéneas. El enlace *ward*, junta las estaciones Las Juntas y Nusutia antes de San Mateo y Teponahuazo, por eso es que la forma del dendograma varía un poco en relación con los otros dos enlaces de agrupamiento.

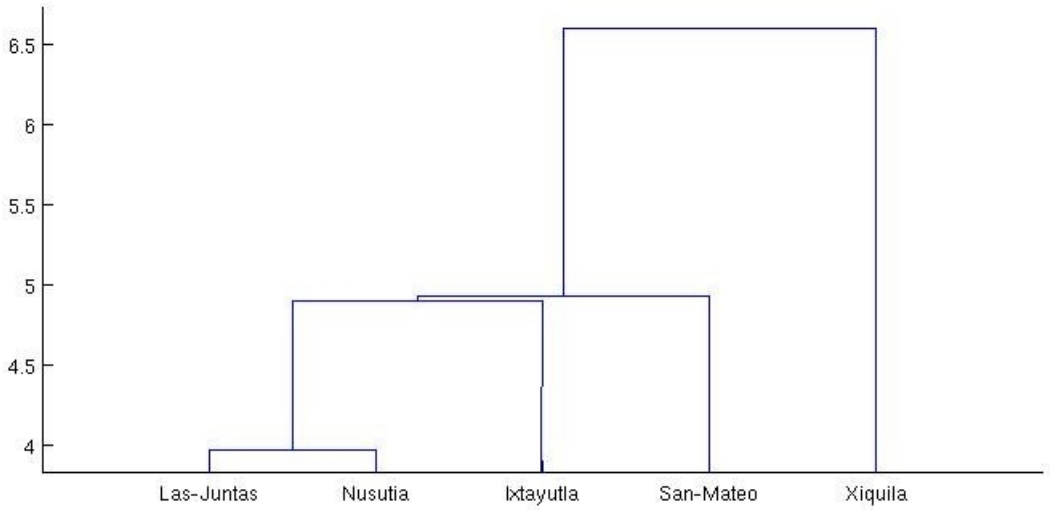

Figura 5.4: Agrupamiento jerárquico con 5 estaciones y enlace centroide.

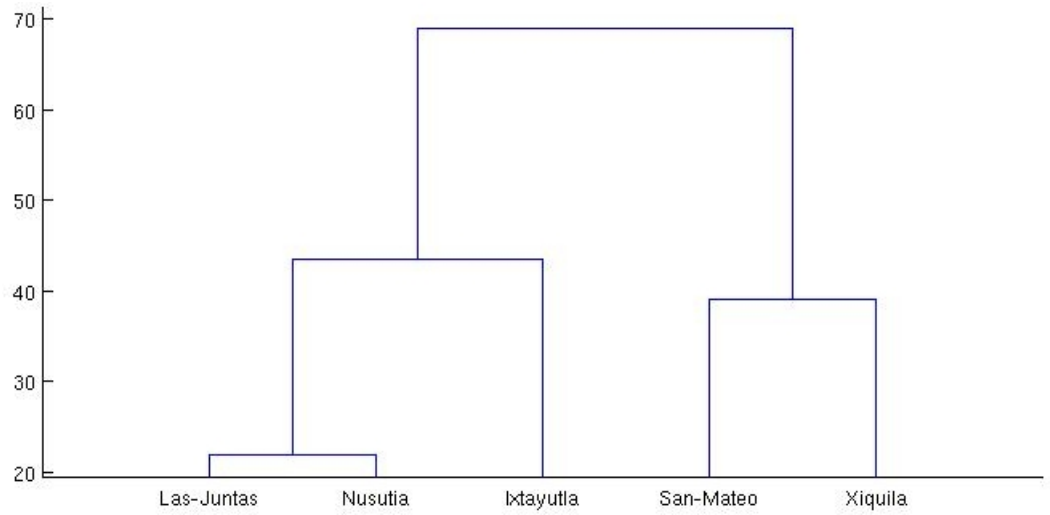

Figura 5.5: Agrupamiento jerárquico con 5 estaciones y enlace *ward*.

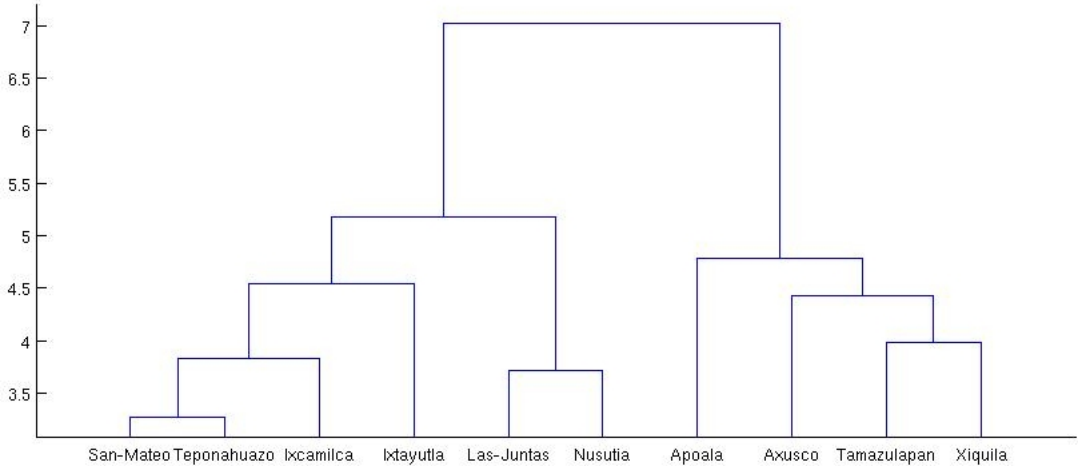

Figura 5.6: Agrupamiento jerárquico con 10 estaciones y enlace promedio.

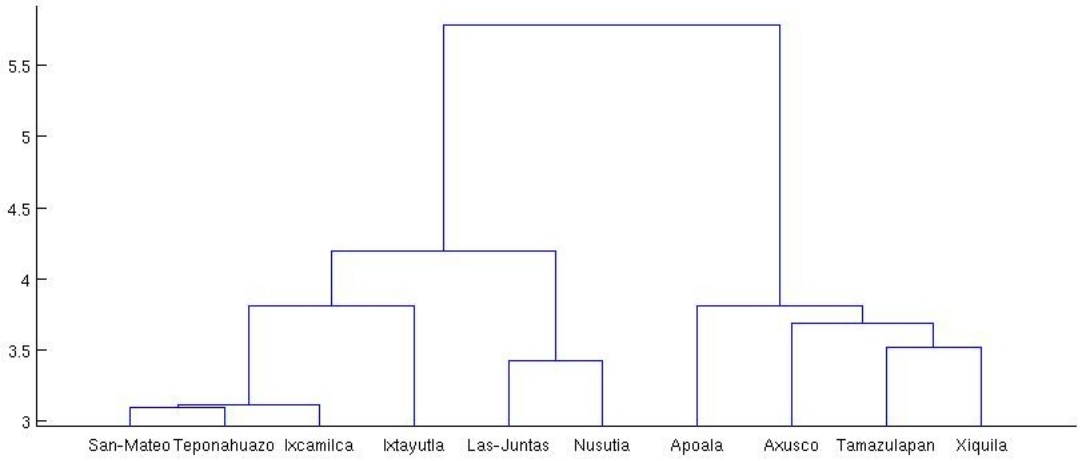

Figura 5.7: Agrupamiento jerárquico con 10 estaciones y enlace centroide.

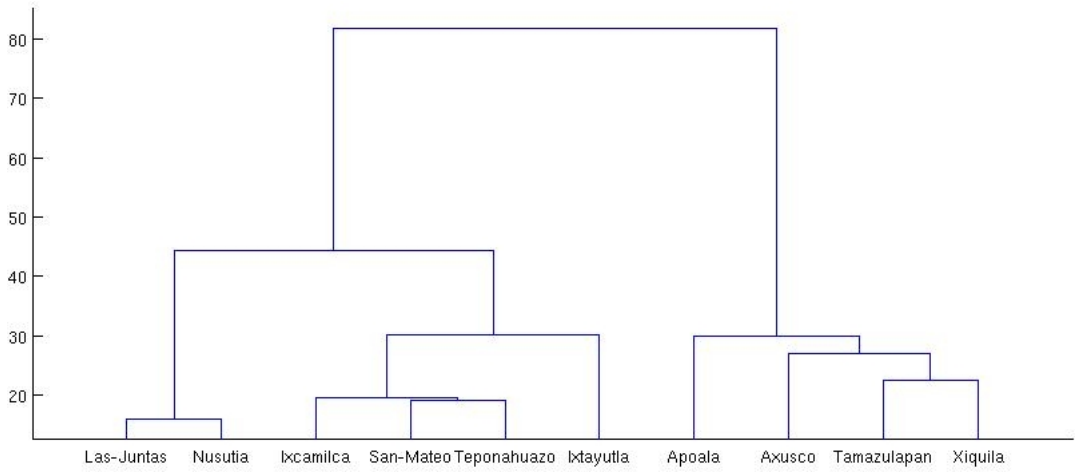

Figura 5.8: Agrupamiento jerárquico con 10 estaciones y enlace *ward*.

En lo que respecta al agrupamiento Fuzzy C-Means, los resultados se obtuvieron por medio de un defusificador para convertir los valores difusos en valores fijos de pertenencia. Uno de los defusificadores más utilizados es mostrado en (Srinivas y otros, 2008). Donde para cada instancia x_k se toma el valor más grande de la *k-ésima* columna en la matriz de pertenencia U y se le asigna el nuevo valor de pertenencia 1, y al resto de los elementos de la columna se les asigna el valor de pertenencia 0.

Siguiendo este razonamiento y tomando en cuenta que el algoritmo necesita un número de grupos predefinido, se obtuvo la delimitación de regiones para las 5 estaciones del primer caso de estudio. Se realizó el análisis formando 2, 3, 4 y 5 grupos homogéneos; los resultados obtenidos se presentan a partir del Cuadro 5.7, hasta el Cuadro 5.10.

Cuadro 5.7: Agrupamiento Fuzzy C-Means con 2 grupos para 5 estaciones.

Grupo 1	Grupo 2
Ixtayutla	Xiquila
Las Juntas	San Mateo
Nusutia	

Cuadro 5.8: Agrupamiento Fuzzy C-Means con 3 grupos para 5 estaciones.

Grupo 1	Grupo 2	Grupo 3
Ixtayutla	Xiquila	Las Juntas
San Mateo		Nusutia

En cada caso, se obtuvo una distribución diferente de las estaciones hidrométricas (pero prevaleciendo el hecho de que Las Juntas y Nusutia siempre permanecen juntas), por lo que es necesaria la aplicación de una medida de validación de grupos, para determinar el mejor

Cuadro 5.9: Agrupamiento Fuzzy C-Means con 4 grupos para 5 estaciones.

Grupo 1	Grupo 2	Grupo 3	Grupo 4
Ixtayutla	San Mateo	Xiquila	Nusutia Las Juntas

Cuadro 5.10: Agrupamiento Fuzzy C-Means con 5 grupos para 5 estaciones.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Ixtayutla	Las Juntas	Nusutia	San Mateo	Xiquila

agrupamiento de los datos. Por tanto, se aplican cuatro medidas de validación de *clusters*: El coeficiente de Partición Difusa V_{PC}, la Entropía de Partición Difusa V_{PE}, el Índice de Realización Difusa FPI y la Entropía de Clasificación Normalizada NCE. Cada una de ellas indicando la mejor distribución de las estaciones hidrométricas.

Los resultados correspondientes a la aplicación de estas medidas de validación son mostradas en el Cuadro 5.11.

Los índices V_{PC} y V_{PE} sugieren dos grupos como mejor delimitación de regiones homogéneas, por su parte los índices FPI y NCE sugieren cinco grupos como el óptimo de homogeneización.

Cuadro 5.11: Validación de grupos para el caso de 5 Estaciones

	Número de grupos			
Índice	2	3	4	5
V_{PC}	**0.713**	0.670	0.695	0.678
V_{PE}	**0.196**	0.260	0.271	0.300
FPI	0.574	0.494	0.406	**0.402**
NCE	0.652	0.547	0.451	**0.429**

En este caso, es conveniente tomar como mejor resultado, el análisis con dos grupos homogéneos, ya que el formar cinco grupos se estaría colocando una estación por grupo. Lo cual no tendría mucha relevancia para la regionalización de caudales, pues no habría transferencia de información entre cuencas.

Con estos resultados se puede observar que la delimitación de regiones homogéneas, usando 5 estaciones hidrométricas, es muy similar para ambos métodos de agrupamiento. Pues en ambos casos se encuentran dos regiones hidrológicamente homogéneas. La primera formada con las estaciones Las Junta, Nusutia e Ixtayutla. La segunda formada por las estaciones San Mateo y Xiquila.

Análogamente, se aplicó el agrupamiento Fuzzy C-Means a las 10 estaciones hidrométricas del segundo caso de estudio. Se definieron también 4 diferentes análisis, el primero con dos grupos homogéneos y el resto con tres, cuatro y cinco grupos. Los resultados de

dichos agrupamientos son presentados en los cuadros 5.12 a 5.15. Aquí, se observa como las estaciones de Apoala, Axusco, Tamazulapan y Xiquila permanecen juntas como grupo, independientemente del número de *clusters* preseleccionado para ejecutar el algoritmo.

Cuadro 5.12: Agrupamiento Fuzzy C-Means con 2 grupos para 10 estaciones.

Grupo 1	Grupo 2
Apoala	Ixcamilca
Axusco	Ixtayutla
Tamazulapan	Las Juntas
Xiquila	Nusutia
	San Mateo
	Teponahuazo

Cuadro 5.13: Agrupamiento Fuzzy C-Means con 3 grupos para 10 estaciones.

Grupo 1	Grupo 2	Grupo 3
Ixtayutla	Apoala	Ixcamilca
Las Juntas	Axusco	San Mateo
Nusutia	Tamazulapan	Teponahuazo
	Xiquila	

Cuadro 5.14: Agrupamiento Fuzzy C-Means con 4 grupos para 10 estaciones.

Grupo 1	Grupo 2	Grupo 3	Grupo 4
San Mateo	Apoala	Ixcamilca	Las Juntas
Teponahuazo	Axusco	Ixtayutla	Nusutia
	Tamazulapan		
	Xiquila		

Una vez más, fue necesario calcular los índices de validación para determinar cuál de los cuatro agrupamientos es el óptimo. En este caso, los valores resultantes para cada región se muestran en el Cuadro 5.16.

Se puede observar que los índices V_{PC}, V_{PE} y FPI, los cuales son altamente usados en la literatura hidrológica (Hall y Mins, 1999), sugieren dos grupos como la mejor partición de datos. En contraste la medida NCE, sugiere el uso de tres *clusters* como la mejor partición. Se puede observar que ésta última medida de validación selecciona tres grupos con una ventaja muy pequeña sobre la elección de dos grupos, pues el valor es casi igual en los dos casos.

Por lo anterior, se deciden tomar sólo dos grupos como la mejor partición de datos. Este resultado refuerza la delimitación de regiones con el agrupamiento jerárquico, pues también encuentra dos grupos de estaciones hidrométricas.

Cuadro 5.15: Agrupamiento Fuzzy C-Means con 5 grupos para 10 estaciones.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Axusco Tamazulapan Xiquila	Apoala	Ixcamilca Teponahuazo	Ixtayutla San Mateo	Las Juntas Nusutia

Cuadro 5.16: Validación de grupos para el caso de 10 Estaciones

	Número de grupos			
Índice	2	3	4	5
V_{PC}	**0.693**	0.566	0.419	0.435
V_{PE}	**0.208**	0.330	0.466	0.491
FPI	**0.612**	0.649	0.773	0.705
NCE	0.693	**0.692**	0.774	0.703

5.2. Análisis de resultados de agrupamiento

Haciendo una comparación de resultados óptimos entre el agrupamiento jerárquico y Fuzzy C-Means, es posible verificar que ambos algoritmos producen grupos muy similares, tanto para el caso de 5 estaciones hidrométricas como para el caso que ocupa las 10 estaciones de aforo.

En lo que respecta al primer caso de estudio, el algoritmo jerárquico determina que la mejor delimitación de regiones incluye dos zonas homogéneas. La primera formada con las estaciones Las Juntas, Nusutia e Ixtayutla y la segunda región formada por las estaciones San Mateo y Xiquila. Análogamente, el algoritmo Fuzzy C-Means encuentra las mismas regiones como mejor partición difusa de los datos.

Por su parte, el análisis de datos con las 10 estaciones de aforo tiene una convergencia mayor y unánime al identificar dos regiones independientes de datos. Es posible observar que los tres tipos de enlaces del algoritmo jerárquico indican la presencia de dos regiones homogéneas, así mismo el algoritmo Fuzzy C-Means identifica también dos regiones hídricas con características similares. En la primera región se incluyen las estaciones Apoala, Axusco, Tamazulapan y Xiquila. La segunda región es formada por las estaciones Ixcamilca, Ixtayutla, Las Juntas, Nusutia, San Mateo y Teponahuazo.

Como conclusión de esta etapa de procesamiento, se encontraron dos regiones homogéneas en cada uno de los casos de estudio. Ahora bien, para cada uno de los *clusters* identificados por los algoritmos de agrupamiento, fue necesario verificar la homogeneidad de los mismos. Esto con la finalidad de garantizar que los grupos formados fueran, en cierta manera, similares u homogéneos. Por tanto, se aplicaron las pruebas de Heterogeneidad H de Hosking y Wallis descritas en la sección 2.5.

De manera preliminar, se puede decir que cada grupo formado debiera ser internamente homogéneo, sin embargo, es necesario asegurar tal suposición por medio de un procedimiento establecido. Por tal motivo, los *test* de Heterogeneidad se aplicaron a cada delimitación de regiones encontrada. Para cada región, se realizó un total de 500 simulaciones de la misma, con el objetivo de hacer la comparativa de variabilidad en los L-momentos de la región original y la simulada (ver sección 2.5). La primera región evaluada fue el grupo uno del caso de estudio con 5 estaciones hidrométricas (el grupo que incluye las estaciones Ixtayutla, Las Juntas y Nusutia), los resultados obtenidos son mostrados en el Cuadro 5.17.

Cuadro 5.17: Prueba de heterogeneidad del primer grupo con 5 estaciones.

	Variable			
Prueba	Caudal máx.	Caudal mín	Caudal med.	Lluvia Med.
H1	-1.04	-3.84	-1.39	0.42
H2	-1.06	-4.01	-1.22	0.40
H3	-1.33	-3.85	-13.55	-5.01

En este caso, se observa que la mayoría de valores son negativos y aquellos que son positivos son menores de 1. Recordando las condiciones de homogeneidad de Hosking y Wallis, una región homogénea debería tener un estadístico H_i menor de uno. Por tanto, se puede asegurar que la primera región del caso de 5 estaciones es completamente homogénea.

La siguiente región evaluada fue el grupo que incluye sólo a las estaciones Xiquila y San Mateo. Los respectivos resultados de la prueba son los mostrados en el cuadro 5.18. Se puede observar, que nuevamente las pruebas no sobrepasan el valor de 1, y en consecuencia la región puede ser considerada como homogénea.

Cuadro 5.18: Prueba de heterogeneidad del segundo grupo con 5 estaciones.

	Variable			
Prueba	Caudal máx.	Caudal mín	Caudal med.	Lluvia Med.
H1	-7.77	-0.63	-2.46	-1.99
H2	-7.77	-0.63	-2.46	-1.99
H3	0.62	-1.44	-0.71	-1.78

Posteriormente, se aplicaron las pruebas de heterogeneidad a los dos grupos identificados para el caso de estudio con las 10 estaciones hidrométricas. Los resultados son mostrados en el cuadro 5.19 y 5.20.

En el primer grupo, las pruebas de heterogeneidad muestran que la mayoría de datos son aceptablemente homogéneos, sólo el caudal mínimo se presenta como posiblemente homogéneo por la prueba H_1. Aun así, las pruebas H_2 y H_3 permitieron considerar el caudal mínimo como homogéneo.

Finalmente, el segundo grupo de las 10 estaciones se acepto como homogéneo, ya que todas las pruebas (H_1, H_2 y H_3) indican que son menores a 1.

Cuadro 5.19: Prueba de heterogeneidad del primer grupo con 10 estaciones.

	Variable			
Prueba	Caudal máx.	Caudal mín	Caudal med.	Lluvia Med.
H1	-11.78	1.34	-1.74	-1.36
H2	-7.09	0.97	-1.14	-1.35
H3	0.74	-0.13	-1.83	-2.58

Cuadro 5.20: Prueba de heterogeneidad del segundo grupo con 10 estaciones.

	Variable			
Prueba	Caudal máx.	Caudal mín	Caudal med.	Lluvia Med.
H1	-0.72	-7.06	-1.89	-3.89
H2	-0.54	-4.23	-2.02	-4.84
H3	0.96	-2.36	-2.19	-8.30

En general, los cuatro grupos formados por los algoritmos de agrupamiento son validados como homogéneos. Por lo tanto, éstos pueden utilizarse para obtener los modelos de regresión lineal, que sean capaces de estimar los caudales para todas las estaciones que componen las regiones homogéneas.

5.3. Modelos de estimación de caudales

Con las regiones hidrológicas identificadas y la seguridad de que los datos en ellas son homogéneos, se aplicó la técnica de regresión lineal múltiple con selección paso a paso, para determinar los modelos de regresión que permitan la estimación de caudales en cada grupo identificado.

Se aplicó la variante de selección por agregación para obtener modelos con un mínimo de variables. Sólo aquellas que fuesen las más importantes en describir el comportamiento de los caudales. De esta manera, se facilita el trabajo de recolección de datos incluidos en el modelo de regresión lineal, pues los datos a recolectar son mínimos. Para obtener los modelos de regresión, se utilizaron las 14 variables climáticas y fisiográficas como regresores. Por ejemplo, para calcular el modelo de regresión del caudal máximo, se eliminaron las variables correspondientes a los caudales mínimo y medio, con el objetivo de evitar dependencias entre los mismos caudales.

Las variables independientes usadas para calcular el modelo de regresión fueron: lluvia media mensual, longitud del cauce principal, porcentaje de cobertura vegetal, temperatura, precipitación media anual, área de la cuenca, densidad de drenaje, elevación media de la cuenca, coeficiente de escurrimiento, elevación máxima de la cuenca, elevación mínima de la cuenca, latitud, longitud y lluvia máxima anual en 24 horas con periodo de retorno de

2 años. Consecuentemente, las variables dependientes fueron el caudal máximo (Q_{max}), el caudal mínimo (Q_{min}) y el caudal medio (Q_{med}).

Los modelos de regresión fueron obtenidos con la ayuda del paquete estadístico Minitab 15.0, instalado en un sistema operativo Windows 7 (x86, 32 bits). Dicho paquete fue configurado para utilizar una correlación parcial de entrada $F_{in} = 4$ (ver sección 3.3.2) y un nivel de significancia estadística del 5 %.

Con estos valores de configuración se aplicó el método al primer grupo para el caso de 5 estaciones, y se obtuvieron los modelos de regresión mostrados en las ecuaciones 5.1 a 5.3.

$$Q_{max} = -17805.5 + 1.497x_1 + 0.009x_2 \tag{5.1}$$

$$Q_{min} = 131.94 + 0.13x_1 - 1.88x_3 \tag{5.2}$$

$$Q_{med} = -6014.31 + 0.389x_1 + 0.003x_2 \tag{5.3}$$

Donde x_1 es la lluvia media mensual, x_2 es la latitud y x_3 el porcentaje de cobertura vegetal. Para el caudal máximo, el coeficiente de determinación múltiple (R^2) es de 0.47, para el caudal mínimo es de 0.21 y para el caudal medio es de 0.41. Con esas medidas los modelos propuestos tienen una buena tendencia a describir la variabilidad de los datos, excepto el caudal mínimo que tiene un coeficiente muy bajo.

Aplicando el mismo método al segundo grupo de estaciones del primer caso de estudio, se obtuvieron los modelos mostrados en las ecuaciones 5.4 a 5.6.

$$Q_{max} = 13.83 + 1.21x_1 \tag{5.4}$$

$$Q_{min} = -0.411 + 0.001x_2 + 0.017x_1 \tag{5.5}$$

$$Q_{med} = -4.843 + 0.116x_1 + 0.003x_2 \tag{5.6}$$

Donde x_1 es la lluvia media mensual y x_2 el área de la cuenca. En este caso, el caudal máximo tiene un coeficiente de determinación múltiple de 0.31, el caudal mínimo de 0.26 y el caudal medio de 0.48. Nuevamente el modelo del caudal mínimo no puede describir por completo la variabilidad de sus datos y en este caso, se suma al problema el modelo del caudal máximo.

Podemos observar que los diferentes modelos de regresión incluyen la variable de lluvia media mensual, por lo que es posible asegurar que dicha variable tiene una mayor relevancia que el resto de las variables regresoras.

Los mismos datos de configuración $F_{in} = 4$ y un nivel de significancia del 5 %, fueron aplicados para determinar los modelos de regresión lineal para el segundo caso de estudio. Después de aplicar el método al primer grupo del caso de estudio con 10 estaciones, se obtuvieron los modelos descritos por las ecuaciones 5.7 a 5.9.

$$Q_{max} = -133.642 + 0.57x_1 + 1.69x_2 + 0.109x_3 \tag{5.7}$$

$$Q_{min} = -5.73 + 0.053x_2 + 0.007x_3 + 0.002x_1 \tag{5.8}$$

$$Q_{med} = -133.688 + 0.027x_1 + 0.123x_2 + 0.015x_3 \tag{5.9}$$

Donde x_1 representa la lluvia media mensual, x_2 la longitud del cauce principal y x_3 la precipitación media anual. Los respectivos coeficientes de determinación múltiple son de 0.46 para el caudal máximo, 0.48 para el caudal mínimo y 0.46 para el caudal medio. Por tanto, estos modelos de regresión pueden describir mejor la variabilidad de los caudales de esta región.

Por último, se aplicó regresión lineal, con selección por agregación y los parámetros ya definidos anteriormente, a la segunda región hallada en el caso de estudio con 10 estaciones hidrométricas. Los respectivos modelos de estimación son los presentados de la ecuación 5.10 a la ecuación 5.12.

$$Q_{max} = -112.88 + 1.58x_1 + 1.87x_2 \tag{5.10}$$

$$Q_{min} = 158.62 - 0.067x_3 + 0.098x_1 + 0.020x_4 - 1.36x_2 - 0.033x_5 \tag{5.11}$$

$$Q_{med} = -106.25 + 0.334x_1 + 0.0351x_3 + 0.0180x_4 + 1.70x_6 - 0.0192x_7 \tag{5.12}$$

Aquí, x_1 es la lluvia media mensual, x_2 la longitud del cauce principal, x_3 la elevación mínima de la cuenca, x_4 el área de la cuenca, x_5 la precipitación media anual, x_6 representa el coeficiente de drenado y x_7 es la elevación media de la cuenca. También, se puede observar que todas las variables utilizadas en los modelos del primer grupo están incluidas en los modelos del segundo, lo cual es una gran ventaja al momento de recolectar los datos (ya que en este caso se necesitan a lo más 7 variables).

Para el segundo grupo identificado, el coeficiente de determinación múltiple para el caudal máximo es de 0.4, para el caudal mínimo es de 0.38 y para el caudal medio es de 0.5. Estos resultados muestran que el modelo más seguro es el correspondiente al caudal medio, el cual describe gran parte de la variabilidad de los datos.

En general, los modelos correspondientes al segundo caso de estudio (utilizando 10 estaciones) presentan una mejor fiabilidad para estimación de caudales, pues el coeficiente de determinación múltiple indica que en ellos se describe mejor la variabilidad de los datos.

Capítulo 6

Conclusiones y Perspectivas

6.1. Conclusiones

Los métodos de regionalización de caudales permiten la estimación de flujos de agua, principalmente en áreas o cuencas no aforadas. Tal es el caso de la Mixteca Oaxaqueña, donde actualmente sólo una estación hidrométrica se encuentra en operación, pues el resto de las estaciones han dejado de funcionar. En la actualidad sólo la estación Tezoatlán II es la que lleva el registro de sus caudales.

Por tal motivo, en el presente trabajo se realizó una regionalización de caudales para la Mixteca Oaxaqueña. Se formaron dos casos de estudio, uno con sólo 5 estaciones hidrométricas y otro utilizando 10 estaciones de aforo. El primer caso de estudio se formó con las estaciones Ixtayutla, Las Juntas, Nusutia, San Mateo y Xiquila, utilizando sus registros de caudales en el periodo 1971-1980. Por su parte, el segundo caso fue formado con las estaciones Apoala, Axusco, Ixcamilca, Ixtayutla, Las Juntas, Nusutia, San Mateo, Tamazulapan, Teponahuazo y Xiquila pero sólo con los caudales en el periodo 1971-1977.

Para cada estación, se hizó un análisis de calidad de datos con la finalidad de asegurar que éstos no fueran construidos de manera artificial o incluyan errores de medición y captura. Se aplicaron pruebas de independencia, homogeneidad y outliers; las dos primeras fueron verificadas sin mucho problema. Como resultado de la prueba de outliers se eliminaron tres registros de la estación Apoala, tres de la estación Axusco, dos de Ixtayutla, tres de Nusutia, un registro de San Mateo, dos de Tamazulapan, dos de Teponahuazo y dos de Xiquila.

En cada caso de estudio, se delimitaron las cuencas hidrológicamente homogéneas por medio de algoritmos de agrupamiento, específicamente los métodos de agrupamiento jerárquico aglomerativo y Fuzzy C-Means. En ambos casos, los algoritmos detectaron las mismas regiones homogéneas.

Para el primer caso de estudio, se encontraron dos regiones homogéneas: la primera incluye las estaciones Ixtayutla, Las Juntas y Nusutia, la segunda región incluye las estaciones Xiquila y San Mateo.

En el segundo caso, donde se ocuparon las 10 estaciones de aforo, también se encontraron dos regiones hidrológicas con características similares: por una parte la primer región toma en cuenta las estaciones Apoala, Axusco, Tamazulapan y Xiquila y la otra región homogénea considera las estaciones Ixcamilca, Ixtayutla, Las Juntas, Nusutia, San Mateo y Teponahuazo.

Posteriormente se verificó la homogeneidad de cada uno de los grupos encontrados por los métodos de *clustering*, y los resultados de la prueba confirmaron que los grupos homogéneos hallados para cada caso de estudio, cumplen con el supuesto de homogeneidad. Por lo tanto, en cada grupo fue posible determinar el modelo de regresión encargado de estimar los caudales para las estaciones involucradas en cada región homogénea.

La construcción de los modelos de regresión se hizo mediante la técnica de regresión paso a paso con selección por agregación y se determinó que el segundo caso de estudio tiene los mejores modelos de estimación de caudales, ya que estos describen mejor la variabilidad de los datos. Además, este caso de estudio involucra a la mayor parte de las estaciones en la Mixteca Oaxaqueña.

Para el segundo caso de estudio, que involucra la mayor parte de las estaciones, se observa que los modelos obtenidos sólo requieren de la lluvia media mensual, la longitud del cauce principal y la precipitación media anual; para estimar los caudales máximos, mínimos y medios en la primer región homogénea. Por su parte, los caudales de la segunda región pueden ser estimados a partir de la lluvia media mensual, la longitud del cauce principal, la elevación mínima de la cuenca, el área de la misma, la precipitación media anual, el coeficiente de escurrimiento y la elevación media de la cuenca. Se hace notar, que en cada caso es posible estimar los caudales máximos, mínimos y medios con pocas variables de regresión, lo cual es una característica deseable de los modelos al momento de recolectar datos.

Como producto de software, se desarrolló una biblioteca de *clustering* en C++, que incluye los métodos de agrupamiento jerárquico y Fuzzy C-Means, así como las pruebas de calidad en los datos y los *test* de heterogeneidad de grupos. También se implementó una clase dedicada a la generación de modelos de regresión lineal, pero utilizando todas las variables del estudio.

6.2. Perspectivas

Como parte de la continuación de este trabajo, se deberán tomar en cuenta diferentes técnicas para estimar los caudales de la región Mixteca. Por ejemplo, utilizar métodos no lineales para garantizar una mejor adecuación de los datos, pues con los modelos actuales se forzan los datos a una distribución lineal de los mismos. Como ejemplos de técnicas de estimación no lineal se pueden mencionar las redes neuronales, filtros de Kalman, entre otros.

En este sentido, resultaría interesante hacer una comparación de los diferentes métodos de regresión y estimación para determinar cual de ellos es el mejor para la predicción de caudales. Se deberá tomar en cuenta el número y la importancia de las variables utilizadas, así como la descripción de la variabilidad de los datos.

También, sería de importancia un análisis de correlación entre las variables utilizadas para el estudio. De esta manera se podrían eliminar variables que presenten una mayor relación con el resto de las mismas. Esto permitiría utilizar menos variables en la identificación de regiones homogéneas. En este mismo sentido, se podrían utilizar métodos de *clustering* y visualización con el fin de obtener un panorama visual de los agrupamientos y de las relaciones de las observaciones (registros) con los grupos obtenidos.

Finalmente, podrían utilizarse métodos de imputación de datos para completar los registros mensuales de las estaciones hidrométricas con años incompletos. De esta manera, se puede realizar una regionalización alternativa y comparar los modelos de estimación de caudales con los obtenidos en el presente trabajo.

Bibliografía

Aitkin, M. A: «Simultaneous inference and the choice of variables subsets». *Technometrics*, 1974, **16(2)**, pp. 221–227.

Andrews, D. F: «Plots of high dimensional data». *Biometrics*, 1972, **28**, pp. 125–136.

Barker, John: «The Landsat Program». National Aeronautics and Space Administration. [http://landsat.gsfc.nasa.gov], 1970. Último acceso, 13/Junio/2011.

Batagelj, Vladimir: «Generalized ward and related clustering problems». En: H. H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, volumen 19, pp. 67–74. North-Holland, Amsterdam, 1988.

Berenson, M. L. y Levine, D. M: *Estadística Básica en Administración, Conceptos y Aplicaciones*. Pearson, New York, 6ł edición, 1996.

Bezdek, J. C: «Cluster validity with fuzzy sets». *Journal of cybernetics*, 1974, **3(3)**, pp. 58–72.

——: *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

Chang, Shu. y Donald, H. B: «Spatial patterns of homogeneous pooling groups for flood frequency analysis». *Hydrological Sciences Journal*, 2003, **48(4)**, pp. 601–618.

Chokmani, K. y Ouarda, T. B. M. J: «Physiographical space-based kriging for regional flood frequency estimation at ungauged sites». *Water Resource Research*, 2004, **40**, p. 13.

CONABIO: «Comisión Nacional para el Conocimiento y Uso de la Biodiversidad en México». [http://www.conabio.gob.mx/conocimiento/regionalizacion/doctos/regionalizacion.html], 2008. Último acceso, 5/Marzo/2012.

Cox, D. R. y Snell, E. J: «The choice of variables in observational studies». *Royal Statistical Society*, 1974, **23(1)**, pp. 51–59.

Downs, G. M. y Barnard, J. M: «Clustering methods and their uses in computational chemistry». En: K. B. Lipkowitz y D. B. Boyd (Eds.), *Reviews in Computational Chemistry*, volumen 18. Hoboken, New Jersey, USA, 2003.

Dunn, J. C: «A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters». *Journal of Cybernetics*, 1973, **3**, pp. 32–57.

Erazo, Adriana Maria: «Regionalización de caudales máximos y medios en el Salvador». Servicio Nacional de Estudios Territoriales, 2004. Servicio Hidrológico Nacional, El Salvador.

Escalante-Sandoval, Carlos. y Reyes-Chávez, Lilia: *Técnicas estadísticas en hidrología*. Universidad Nacional Automoma de México, México, 2002.

Frank, A. y Asuncion, A.: «UCI Machine Learning Repository». [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences, 2010.

Garcia, David Fernando.: «Diseño e implementación de un proceso de recuperación de plata metálica en la Empresa Incineradores Industriales S.A. E.S.P.» Universidad Tecnológica de Pereira, Facultad de Tecnologías. Colombia, 2007. Tesis de Licenciatura.

Gath, I. y Geva, A. B: «Unsupervised optimal fuzzy clustering». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989, **11(7)**, pp. 773–780.

Girard, C: «Estimation régionale des crues basée sur l'analyse canonique des correlations». En: INRS-Eau (Ed.), *Mémoire de M.Sc*, Universitié du Québec, Canada, 2001.

Greenwood, J.; Landwehr, J.; Mátalas, N. y Wallis, J: «Probability weighted moments: Definition and relation to parameters of several distributions expressed in inverse form». *Water Resources Research*, 1979, **15(6)**, pp. 1049–1054.

Guler, C. y Thyne, G. D: «Delineation of hydrochemical facies distribution in a regional groundwater system by means of Fuzzy C-Means clustering». *Water Resources Research*, 2004, **40**.

Haitovsky, Yoel: «A note on the maximization of R^2». *American Statistician*, 1969, **23**, p. 1.

Hall, M. J. y Mins, A. W: «The classification of hydrologically homogeneous region». *Hydrological Science Journal*, 1999, **44**, pp. 693–704.

Han, J.; Kamber, M. y Tung, A. K: «Spatial clustering methods in data mining: A Survey». En: H. Miller y J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery*, pp. 1–29, 2001.

Heuvelmans, Griet.; Muys, Bart. y Feyen, Jan: «Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets». *Journal of Hidrology*, 2006, **319**, pp. 245–265.

Hosking, J. R. M: «The four-parameter kappa distribution». *IBM Journal of Research and Development*, 1994, **38**, pp. 251–258.

Hosking, J. R. M. y Wallis, J. R: «Some statistics useful in regional frequency analysis». *Water Resources Research*, 1993, **29(2)**, pp. 271–281.

——: *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, New York, 1997.

Hotait-Salas, Natalia.: «Propuesta de regionalización hidrológica de la Mixteca Oaxaqueña, México, a través de análisis multivariante». Universidad Politécnica de Madrid, 2008. Tesis de Licenciatura.

IMTA: «Sistema de Información de Aguas Superficiales». Instituto Mexicano de Tecnología del Agua, 1997. Versión 1.0. México..

——: «Extractor rápido de información climatológica». Instituto Mexicano de Tecnología del Agua, 2007. Versión ERIC III. México..

INEGI: «Carta Hidrológica de Aguas Superficiales, clave E14-9, escala 1:250 000. Oaxaca, México». Instituto Nacional de Estadística Geografía e Informática, 1988.

——: «Conjunto de datos vectoriales, escala 1:250 000. México». Instituto Nacional de Estadística Geografía e Informática, 2000.

Jain, A. K. y Dubes, Richard. C: *Algorithm for Clustering Data*. Prentice-Hall, Advanced Reference Series, New Jersey, 1988.

Jain, A. K.; Murty, M. N. y Flinn, P. J: «Data clustering: A review». *ACM Computing Surveys*, 1999, **31(3)**.

Jingyi, Zhang. y Hall, M. J: «Regional flood frequency analysis for the Gan-Ming River basin in China». *Journal of Hydrology*, 2004, **296**, pp. 98–117.

Leclerc, Martin. y Ouarda, T. B. M. J: «Non-stationary regional flood frequency analysis at ungauged sites». *Journal of Hydrology*, 2007, **343**, pp. 254–265.

Mallows, C. L: «Choosing variables in a linear regression: A graphical aid». Central Regional Meeting of the Institute of Mathematical Statistics, 1964. Manhattan, Kansas.

Miller, I.; Freud, J. E. y Johnson, R. A: *Probabilidad y Estadística para Ingenieros*. Prentice Hall Hispanoamericana, México, 5ł edición, 1997.

Montgomery, Douglas. C.; Peck, Elizabeth. A. y Vining, G. Geoffrey: *Introduction to Linear Regression Analysis.* Wiley-Interscience, New York, thirdł edición, 2001.

Montgomery, Douglas. C. y Runger, George. C: *Applied Statistics and Probability for Engineers.* John Wiley and Sons, New York, secondł edición, 1999.

Muirhead, Robb. J: *Aspects of Multivariate Statistical Theory.* Wiley-Interscience, 1982.

Nathan, R. J. y McMahon, T. A: «Identification of homogeneous regions for the purposes of regionalization». *Journal of Hydrology*, 1990, **121**, pp. 217–238.

Oren, Zamir. y Oren, Etzioni: «Web document clustering: A feasibility demonstration». *Proceedings of ACM/SIGIR*, 1998, pp. 46 – 54.

Orozco, Álvaro; Guarnizo, Cristian y Echeverry, Julian: «Organización de espigas usando agrupamiento Fuzzy C-Means». *Scientia et Technica*, 2005, **11(28)**, pp. 37–40.

Ouarda, T. B. M. J.; Ba, K. M.; Diaz-Delgado, C.; Carsteanu, A.; Chokmani, K.; Gingras, H.; Quentin, E.; Trujillo, E. y Bobée, B: «Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study». *Journal of Hydrology*, 2008, **348**, pp. 40–58.

Ouarda, T. B. M. J.; Girard, C.; Cavadias, G. S. y Bobée, B.: «Regional flood frequency estimation with canonical correlation analysis». *Journal of Hydrology*, 2001, **254**, pp. 157–173.

Pal, N. R. y Bezdek, J. C: «On cluster validity for the fuzzy c-means model». *IEEE Transactions on Fuzzy Systems*, 1995, **3**, pp. 370–379.

Peña, Daniel: *Análisis de datos multivariantes.* volumen 1. Mc Graw Hill, Madrid, España, 2002.

Pearson, C. P: «New Zealand regional flood frequency analysis using L-Moments». *Journal of hydrology*, 1991, **30(2)**, pp. 53–64.

Ríus-Díaz, Francisca.; Barón-Lopez, Javier.; Sánchez-Font, Elisa. y Parras-Guijosa, Luis: *Bioestadística: Métodos y aplicaciones.* Universidad de Málaga. Publicaciones, Málaga, España, 1998. ISBN 847496-653-1.

Rojo-Abuín, J. M: «Regresión lineal multiple». Instituto de Economía y Geografía. Consejo Superior de Investigaciones Científicas. Madrid, España, 2007. Reporte Técnico.

Roubens, M: «Fuzzy clustering algorithms and their cluster validity». *European Journal of Operations Research*, 1982, **(10)**, pp. 294–301.

Shih-Min, Chiang; Ting-Kuei, Tsay y Stephan, J. N: «Hydrologic regionalization of watersheds. II: Applications». *Journal of Water Resources Planning and Management*, 2002, **128(1)**.

Smithers, J. C. y Schulze, R. E: «A methodology for the estimation of short duration design storms in South Africa using a regional approach based on L-moments». *Journal of Hydrology*, 2001, **24**, pp. 42–52.

Srinivas, V. V.; Tripathi, Shivam.; Rao, Ramachandra. y Govindaraju, Rao: «Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering». *Journal of Hidrology*, 2008, **348**, pp. 146–166.

Torres-Gallardo, Andrea. y Peñaranda Gómez, Gabriel Alfredo: «Regionalización de caudales mínimos por métodos estadísticos de la cuenca Magdalena Cauca». Universidad de la Salle, Facultad de Ingeniería Ambiental y Sanitaria. Bogotá, Colombia, 2006. Tesis de Licenciatura.

Wilcoxon, F: «Individual comparisons by ranking methods». *Biometrics Bulletin*, 1945, **1**, pp. 80–83.

Zadeh, L. A: «Fuzzy Sets». *Information and Control*, 1965, **8(3)**, pp. 338–353.

Anexo A

Variables utilizadas en el proceso de agrupamiento

Las variables utilizadas en el proceso de agrupamiento fueron en total 17, de ellas, tres se refieren a los caudales de la Mixteca Oaxaqueña, tres más de tipo climáticas y once variables fisiográficas. Cada una de ellas se detalla y se define en el presente anexo. Algunas de las variables se definen de acuerdo al glosario hidrológico internacional de la UNESCO. El cual se encuentra disponible en la dirección electrónica:
http://webworld.unesco.org/water/ihp/db/glossary/glu/aglo.htm.

A.1. Variables hidrológicas

Estas variables se cuantificaron de manera mensual.

Gasto máximo

El gasto o caudal de un río es definido como el volumen de agua que pasa por una sección dada, en un tiempo determinado. (Escalante-Sandoval y Reyes-Chávez, 2002). Generalmente se expresa en metros cúbicos por segundo. Por su parte, el gasto máximo se refiere a el máximo volumen de agua que recorre una sección del río en un tiempo determinado. En nuestro caso, el máximo volumen de agua que pasa durante un mes.

Gasto mínimo

Se refiere al volumen mínimo de agua que recorre una sección del río en un periodo igual a un mes.

Gasto medio

La variable del gasto medio es calculada como la media aritmética de todos los caudales registrados por la estación hidrométrica en un día, así mismo puede calcularse el gasto medio mensual y de ser necesario, el gasto medio anual.

A.2. Variables climáticas

Estas variables, a excepción de la lluvia media mensual se cuantificaron en periodos anuales.

Lluvia media mensual

La lluvia media es una característica climatológica de la cuenca hidrográfica y es determinada por la cantidad de lluvia o agua que incide en la cuenca en un tiempo determinado. La lluvia media mensual es calculada como el promedio de la lluvia máxima y mínima registrada en un mes.

Temperatura

La temperatura es una propiedad climática de la cuenca que se refiere a las nociones comunes de frío o calor correspondientes a la misma.

Precipitación media anual

En meteorología, la precipitación es cualquier forma meteorológica de partículas acuosas que caen del cielo y llegan a la superficie terrestre. Esto incluye lluvia, llovizna, nieve y granizo.

A.3. Variables fisiográficas

Todas las variables fisiográficas fueron cuantificadas en periodos de tiempo de un año.

Longitud del cauce principal

La longitud del cauce principal es medida en km y se estima para la corriente de mayor orden en la cuenca sin incluir los ramales, pues éstos se consideran como caudales tributarios al cauce principal.

Cobertura vegetal

Ésta variable se refiere en particular a los bosques y cultivos que añaden su influencia en la naturaleza geológica de la cuenca pues condicionan la retención, evaporación y escurrimiento del agua. En general, la vegetación es capaz de controlar la acción y movimiento del agua.

Área de la cuenca

El área drenada de una cuenca es la superficie medida en km^2 y tienen como punto de salida una estación de aforo o un sitio de interés. Hoy en día esta característica se obtiene fácilmente mediante herramientas computacionales como lo son los Sistemas de Información Geográfica (SIG).

Densidad de drenaje

La densidad de drenaje se define como la longitud total de los cauces dentro de la cuenca, dividida entre el área total de drenaje. Matemáticamente se define como sigue:

$$Dd = \frac{\sum_{j=1}^{n} l_j}{A}$$

donde n representa el número total de cauces de la cuenca y l_j la longitud del *j-ésimo* cauce.

Comúnmente, se encuentran bajas densidades de drenaje en regiones de rocas resistentes o de suelos muy permeables con vegetación densa y donde el relieve es débil. En cambio, se obtienen altas densidades de drenaje en áreas de rocas débiles o de suelos impermeables, vegetación escasa y relieve montañoso.

Elevación media de la cuenca

La elevación media se obtiene fácilmente mediante una malla generada sobre el plano topográfico del sitio de estudio. Para cada una de las intersecciones dentro de la cuenca se

obtiene el valor de la elevación(E_i), por lo que sólo se requiere obtener el promedio entre al menos 100 puntos de la malla para determinar la pendiente media.

$$Em = \frac{\sum_{i=1}^{n} E_i}{n}$$

Coeficiente de escurrimiento

El escurrimiento se define como la parte de la precipitación que fluye por gravedad por la superficie del terreno, o en el interior del mismo. Por su parte el coeficiente de escurrimiento se refiere a un método de estimación de volumen escurrido y puede ser evaluado con las siguientes fórmulas:

$$cuando \quad k <= 0{,}15 \qquad C = k(\frac{P - 250}{2,000})$$

$$cuando \quad k > 0{,}15 \qquad C = k(\frac{P - 250}{2,000}) + \frac{k - 0{,}15}{1{,}5}$$

Donde:

C = coeficiente de escurrimiento anual.

P = precipitación anual.

k = parámetro que depende del tipo de suelo.

Elevación máxima

La Elevación máxima de la cuenca está determinada por la cota superior que aparece dentro de la cuenca, es indicada por las curvas de nivel o en su caso un valor de referencia altitudinal.

Elevación mínima

La elevación mínima constituye la cota menor que aparece en en área de la cuenca, no coincidiendo necesariamente con el punto final del curso o cauce principal.

Latitud y Longitud

Las variables latitud norte y longitud oeste corresponden a las coordenadas del centro de gravedad de la cuenca. El centro de gravedad es el lugar geométrico donde se concentra toda la superficie drenada por la cuenca.

Lluvia máxima anual en 24 horas con periodo de retorno de 2 años

La lluvia es definida por tres variables: magnitud o lámina, duración y frecuencia. La magnitud de la lluvia es la lámina total ocurrida (en milímetros) en la duración de la tormenta. La frecuencia de la lluvia, es expresada por su periodo de retorno o intervalo de recurrencia, que es el tiempo promedio en años en el cual, el evento puede ser igualado o excedido cuando menos una vez.

Anexo B

Pseudocódigo de algunos algoritmos de agrupamiento

B.1. Agrupamiento jerárquico

Dado un conjunto de datos con n objetos a agrupar y una matriz de distancias de tamaño $n \times n$, el procedimiento para obtener el agrupamiento jerárquico aglomerativo es el mostrado por el siguiente algoritmo:

1. Inicializar cada objeto en un solo grupo, es decir, si se tienen n objetos, se deberán tener también n grupos. Cada grupo deberá contener sólo un objeto.
2. Obtener las distancias entre grupos, esto es, calcular las distancias entre cada par de observaciones del conjunto de datos.
3. Encontrar el par de grupos más cercanos o similares y mezclarlos en un solo grupo, esto reducirá el número de grupos.
4. Calcular las distancias entre el nuevo grupo y los grupos existentes para actualizar la matriz de distancias. Dicha actualización se deberá realizar utilizando alguno de los criterios de similitud mostrados en la sección 2.3.2.
5. Repetir los pasos 3 y 4 hasta que todos los objetos se encuentren en un solo grupo de tamaño n, o bien, hasta que se cumpla una restricción previamente establecida.

Recordemos que la distancia entre grupos se puede determinar mediante diferentes caminos, entre ellos, utilizando los enlaces simple, completo o algún otro criterio de similitud entre grupos.

B.2. Agrupamiento Fuzzy C-Means

El algoritmo Fuzzy C-Means(FCM), descrito en (Bezdek, 1981) es una de las técnicas más utilizadas en el agrupamiento difuso y puede ser desarrollado como sigue:

Sea $X = \{x_1, ..., x_n\}$ el conjunto de datos a agrupar, donde cada punto $x_k (k = 1, ..., n)$ es un vector en R^p, $U_{c,n}$ es una matriz de números reales de tamaño $c\,x\,n$ y c es un entero, $2 \leq c < n$. Entonces, la partición del espacio difuso en X esta dado por:

$$M_{fcn} = \{U \in U_{cn} : u_{ik} \in [0,1], \quad \sum_{i=1}^{c} u_{ik} = 1, \quad 0 < \sum_{k=1}^{n} u_{ik} < n\} \tag{B.1}$$

donde u_{ik} es el grado de pertenencia de x_k en el *i-ésimo cluster*.

El objetivo del algoritmo Fuzzy C-Means es buscar la partición difusa óptima y la minimización de la correspondiente función objetivo

$$J(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m ||x_k - v_i||^2 \tag{B.2}$$

donde, $V = (v_1, v_2, ..., v_c)$ es la matriz de centros de *cluster* o centroides, $v_i \in R^p$ es el centroide del *i-ésimo cluster*, $||.||$ es la norma euclidiana entre el *k-ésimo* elemento y el *i-ésimo cluster*, y el exponente difuso $m \in [1, \infty]$ es una constante que influye directamente en los valores o grados de membresía.

El pseudocódigo del algoritmo FCM es el siguiente:

1. Seleccionar un valor para c, m y definir un error máximo ε entre cero y uno.

2. Generar aleatoriamente la partición difusa U^0 de acuerdo a las restricciones en la ecuación B.1.

3. Calcular los nuevos centroides por medio de la ecuación:

$$\hat{c}_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \tag{B.3}$$

4. Teniendo los nuevos centroides, actualizar la matriz de pertenencias utilizando la siguiente expresión:

$$\hat{u}_{ik} = \left[\sum_{j=1}^{c} \left(\frac{||x_k - v_i||^2}{||x_k - v_j||^2} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{B.4}$$

5. Calcular el error mediante la ecuación:

$$E = max[|u_{ik} - \hat{u}_{ik}|] \tag{B.5}$$

6. si $E < \varepsilon$, o bien, se ha llegado a un límite de iteraciones, el algoritmo se detiene, de lo contrario, regresar al paso 3.

Anexo C

Manual de usuario del software desarrollado

El software desarrollado consiste en 4 programas. El primero, realiza las pruebas de Independencia, Homogeneidad y outliers a un conjunto de datos de entrada; el segundo, es capaz de aplicar los métodos de agrupamiento jerárquico aglomerativo y Fuzzy C-Means; el tercer programa, calcula el grado de heterogeneidad en un conjunto de datos de entrada; l cuarto, obtiene un modelo de regresión lineal a partir de un conjunto de datos muestrales.

Cada uno de los programas se encuentra disponible en su versión ejecutable (.exe). Además, necesitan de un conjunto de datos de entrada, el cual es proporcionado por medio de un archivo de texto plano, que puede tener cualquier formato, siempre y cuando tenga la siguiente estructura:

$n\ m$
$val_{11}\ \ val_{12}\ \ \ldots\ val_{1m}\ \ grupoItem_1$
$val_{21}\ \ val_{22}\ \ \ldots\ val_{2m}\ \ grupoItem_2$
$\ldots$
$val_{n1}\ \ val_{n2}\ \ \ldots\ val_{nm}\ \ grupoItem_n$

Donde n y m indican el número de items o registros y el número de variables observadas, respectivamente. Esto es, el tamaño de la matriz de datos que procesará el correspondiente algoritmo (incluyendo el grupo de cada item). El grupo que se coloca en la columna final del archivo, no es tomada en cuenta para el procesamiento de los datos, no obstante es necesario colocarla para que la lectura de los datos sea de manera correcta. En caso de que los datos no tengan asociado un grupo, se deberá colocar un texto cualquiera en dicha columna.

Después de indicar el tamaño de la matriz de datos, es necesario introducir cada item u observación en una fila. Los valores correspondientes a cada una de las variables deberán separarse por un espacio y como último valor de la fila, colocar el grupo del item.

Pruebas de calidad

Este programa es el encargado de aplicar las pruebas de calidad de independencia, homogeneidad y outliers a un determinado conjunto de datos. Dicho programa proporciona una serie de valores que deberán ser tratados de acuerdo a lo mostrado en la sección 2.2. Para ejecutar correctamente este programa se deben realizar los siguientes pasos:

1. Ejecute el archivo de nombre "tests.exe". Desde una consola linux, introduzca el comando: **./tests.exe**.
2. El programa solicitará el nombre del archivo que contiene el conjunto de datos a analizar. Introduzca el nombre de dicho archivo y presione "Enter".
3. Aparecerá un menú de opciones donde se enlistan las pruebas Wald-Wolfowitz (independencia), Mann-Whitney (homogeneidad) y Grubbs (outliers).
4. Introduzca el número correspondiente a la prueba que desea realizar y presione "Enter".
5. El software mostrará los resultados de la prueba seleccionada en el paso anterior.

En la Figura C.1 se muestra la secuencia de pasos descritos previamente para obtener la prueba de independencia de la estación Apoala.

Como se puede observar, el programa muestra el valor numérico de la prueba en cada una de las variables de la base de datos. En el ejemplo de la Figura C.1, se utilizó un conjunto de datos de 4 variables (Gasto máximo, gasto mínimo, gasto medio y lluvia media), por lo que el programa muestra los valores de la prueba de independencia en cada una de estas variables. El mismo formato de salida se obtiene para la prueba de homogeneidad de Mann-Whitney. Sin embargo, en la prueba de outliers, el programa muestra en pantalla los límites superior e inferior de la muestra de datos y enlista los items que se encuentran fuera de dicho rango (indicando que son outliers del conjunto).

Algoritmos de agrupamiento

Este programa implementa los algoritmos de agrupamiento jerárquico y Fuzzy C-Means. En el caso del agrupamiento jerárquico, se pueden aplicar los seis tipos de enlaces para la construcción de grupos (véase sección 2.3). Por su parte, el algoritmo de agrupamiento Fuzzy C-Means calcula los cuatro índices de validación de *clusters* mostrados en la sección 2.4.

El algoritmo de agrupamiento jerárquico implementa como criterio de finalización el número de *clusters*, de esta manera el usuario puede determinar el número de grupos que desea al finalizar el agrupamiento. Así mismo, el algoritmo de agrupamiento Fuzzy C-Means toma como criterio de finalización un error mínimo, es decir, cuando la diferencia entre un

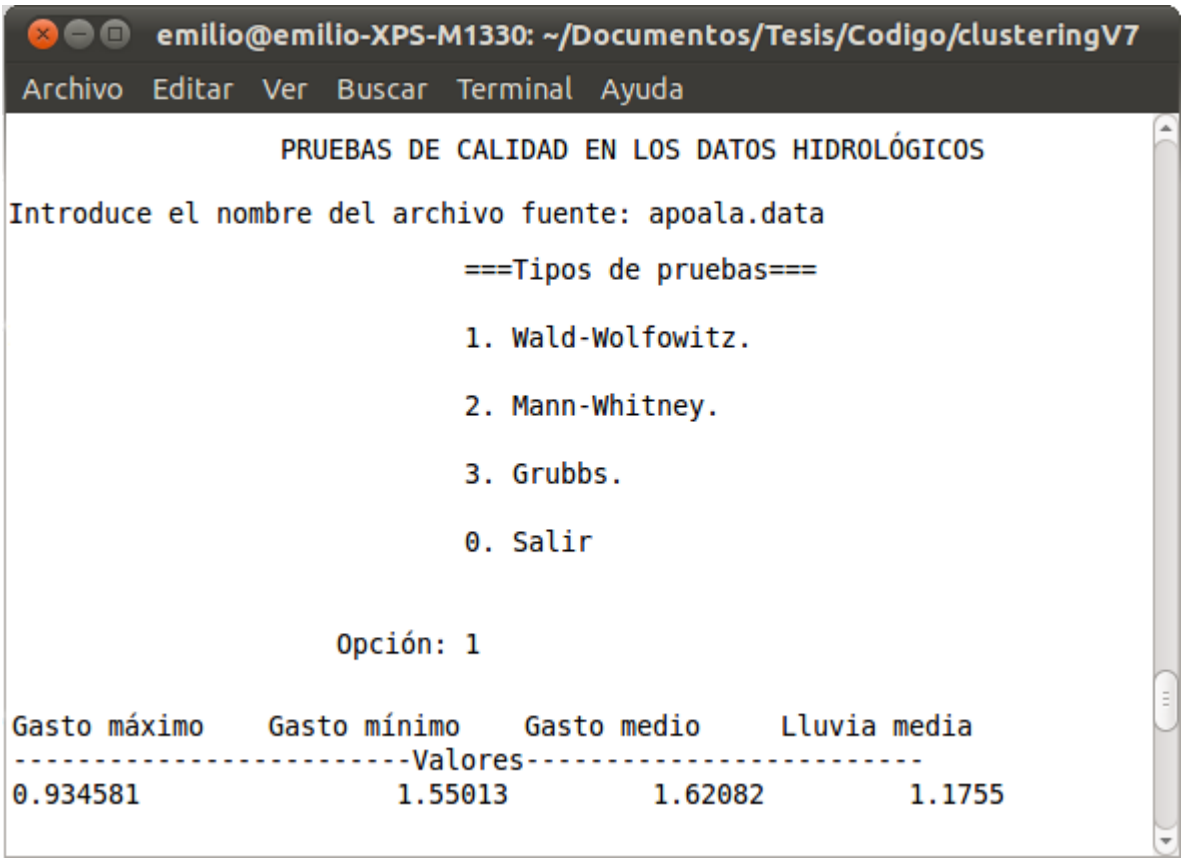

Figura C.1: Ejecución de la prueba de independencia para la estación Apoala.

agrupamiento y el previo es menor a un valor definido, el algoritmo se detiene y proporciona la salida del agrupamiento. La ejecución de este programa se lleva a cabo mediante el siguiente procedimiento:

1. Ejecute el programa de nombre "main.exe". Desde una consola linux, introduzca el comando: **./main.exe**.

2. El programa solicitará el tipo de agrupamiento a realizar (jerárquico o Fuzzy C-Means). Seleccione una opción y presione "Enter".

3. El programa solicitará el nombre del archivo que contiene el conjunto de datos a analizar. Introduzca el nombre de dicho archivo.

4. Introduzca el número de *clusters* que desea obtener.

5. En caso de haber seleccionado un agrupamiento jerárquico, deberá seleccionar el tipo de enlace para construir el agrupamiento. El software realizará el agrupamiento y generará

un archivo de salida. Finalmente se mostrará nuevamente el menú con los tipos de enlace para agrupar el mismo conjunto de datos pero con otro tipo de enlace.

6. En caso de haber seleccionado el agrupamiento Fuzzy C- Means, el software realizará el agrupamiento correspondiente y mostrará en pantalla el valor de los índices de validación de *clusters*. Posteriormente, generará el archivo de salida y volverá a mostrar el menú principal para seleccionar un tipo de agrupamiento.

En la Figura C.2 se muestra la secuencia de pasos a seguir para encontrar los grupos homogéneos usando el algoritmo de agrupamiento jerárquico, y el conjunto de datos que incluye las 5 estaciones de aforo de nuestro primer caso de estudio.

Una vez que se le proporciona al software el tipo de enlace, se comenzará con el agrupamiento de los datos y se genera un archivo con los grupos formados por el algoritmo. Ambos algoritmos de agrupamiento (jerárquico y Fuzzy C-Means), producen un archivo de salida de nombre "salida.txt", donde se enlistan los items del conjunto de datos. Debido a que los items fueron agrupados según su similitud, no se encuentran enlistados con el mismo orden del archivo de entrada, por tanto, las dos últimas columnas del archivo hacen referencia al grupo y número de lista original del item, respectivamente. De esta manera, se puede saber a que grupo queda asociado cada item del conjunto de datos original.

Pruebas de heterogeneidad de grupos

El tercer programa desarrollado, obtiene los índices de heterogeneidad de grupos descritos en la sección 2.5. Es decir, para un conjunto de datos de entrada define qué tan homogéneo es dicho conjunto, calculando los índices de heterogeneidad H_1, H_2 y H_3. Para ejecutar este programa se deberá realizar lo siguiente:

1. Ejecute el programa de nombre "postprocessing.exe". Desde una consola linux, introduzca el comando: **./postprocessing.exe** y presione "Enter".

2. El programa solicitará el nombre del archivo donde se encuentran los datos a evaluar.

3. Introduzca el nombre del archivo y presione la tecla "Enter".

4. El programa solicitará un número de variable. Este número corresponde a la columna donde se encuentran los datos de la variable que deseamos evaluar. Introduzca el número y presione "Enter".

5. El programa mostrará en pantalla el resultado de cada una de las pruebas de Heterogeneidad.

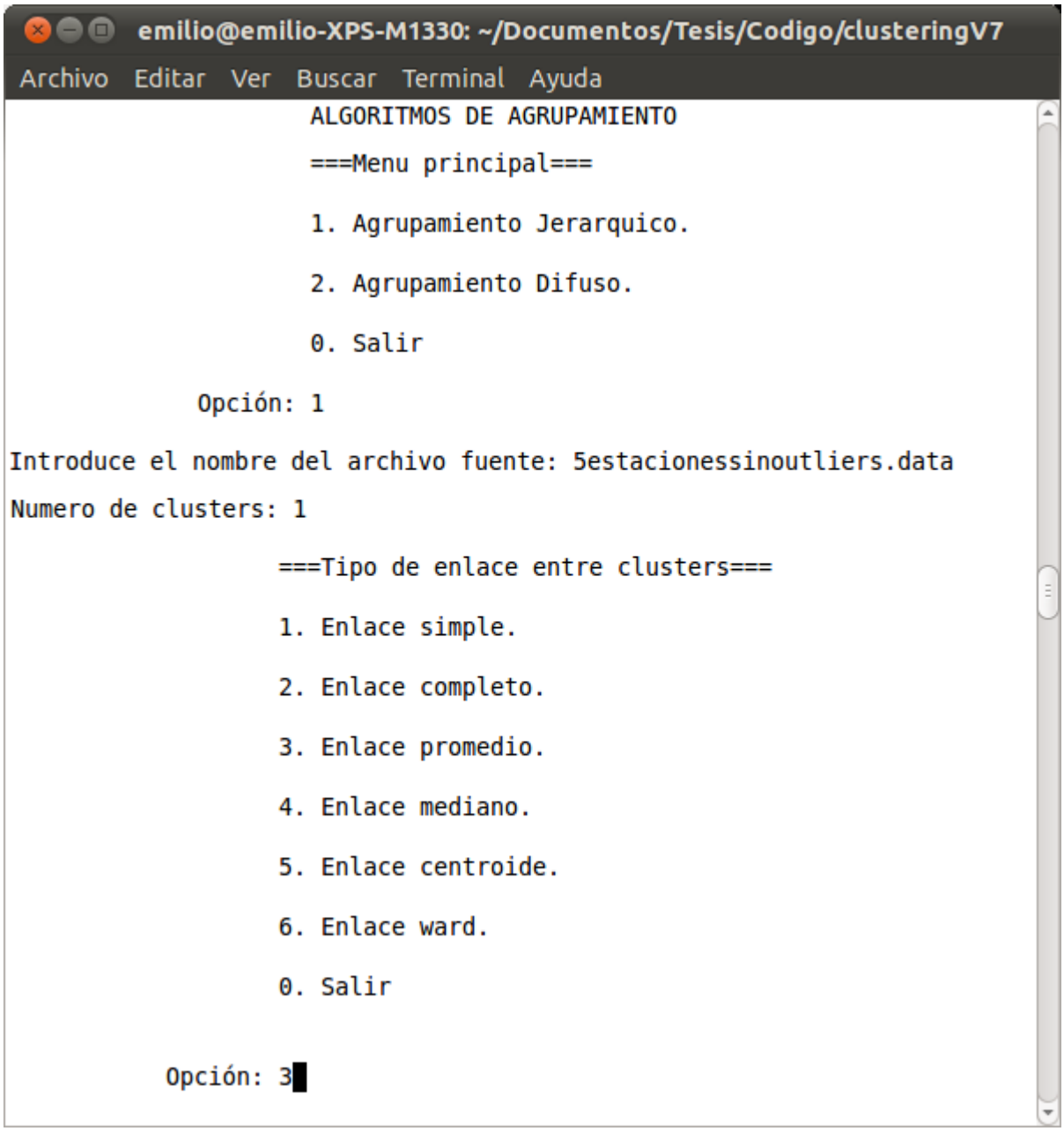

Figura C.2: Ejecución del algoritmo jerárquico para el caso de 5 estaciones.

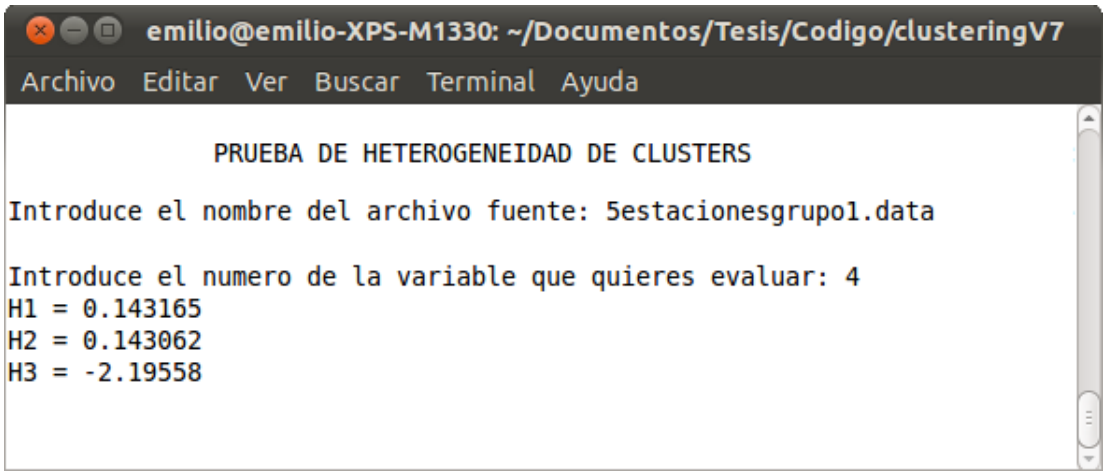

Figura C.3: Ejecución de la prueba de heterogeneidad con el primer grupo formado usando 5 estaciones hidrométricas.

En la Figura C.3 se muestra un ejemplo de este programa. Se utiliza el conjunto de datos formado por el primer grupo homogéneo encontrado en el caso de 5 estaciones hidrométricas. En este caso se hace la evaluación de la cuarta variable, es decir la que corresponde a la lluvia media mensual.

Recordemos que el programa necesita realizar un cierto número de simulaciones, por lo que el valor de las pruebas difícilmente será el mismo en dos ejecuciones diferentes. Sin embargo, deberán ser valores similares para concluir en un mismo resultado de los datos (homogéneos, posiblemente heterogéneos o definitivamente heterogéneos). Para nuestro ejemplo, los datos resultan ser homogéneos, pues los valores de los tres índices se encuentran por debajo de 1 (véase sección 2.5).

Regresión lineal

El último programa desarrollado es capaz de construir un modelo de regresión lineal utilizando todas las variables del estudio. En este caso, se utiliza el método de mínimos cuadrados para estimar los coeficientes de regresión y la varianza del error en los datos. La ejecución del programa se realiza de la siguiente manera:

1. Ejecute el programa de nombre "linearRegression.exe". Desde una consola linux, introduzca el comando: **./linearRegression.exe** y presione "Enter".
2. Del menú que se desprende seleccione la opcion "Modelar función".
3. El programa solicitará el nombre del archivo donde se encuentran los datos a modelar. Introduzca el nombre del archivo y presione la tecla "Enter".

4. El programa solicitará un número de variable. Este número corresponde a la variable que representará la respuesta del sistema. Introduzca el número y presione "Enter".

5. El programa mostrará en pantalla los coeficientes de regresión para cada variable y la varianza del error en los datos.

En la Figura C.4 se muestra la ejecución del programa utilizando el primer grupo del caso de estudio con 5 estaciones. Además, se obtiene el modelo para la primer variable que representa el caudal máximo. De la lista de valores mostrados, el primero corresponde a la constante de regresión lineal β_0 y el resto de los valores corresponden a cada uno de los coeficientes de regresión del modelo.

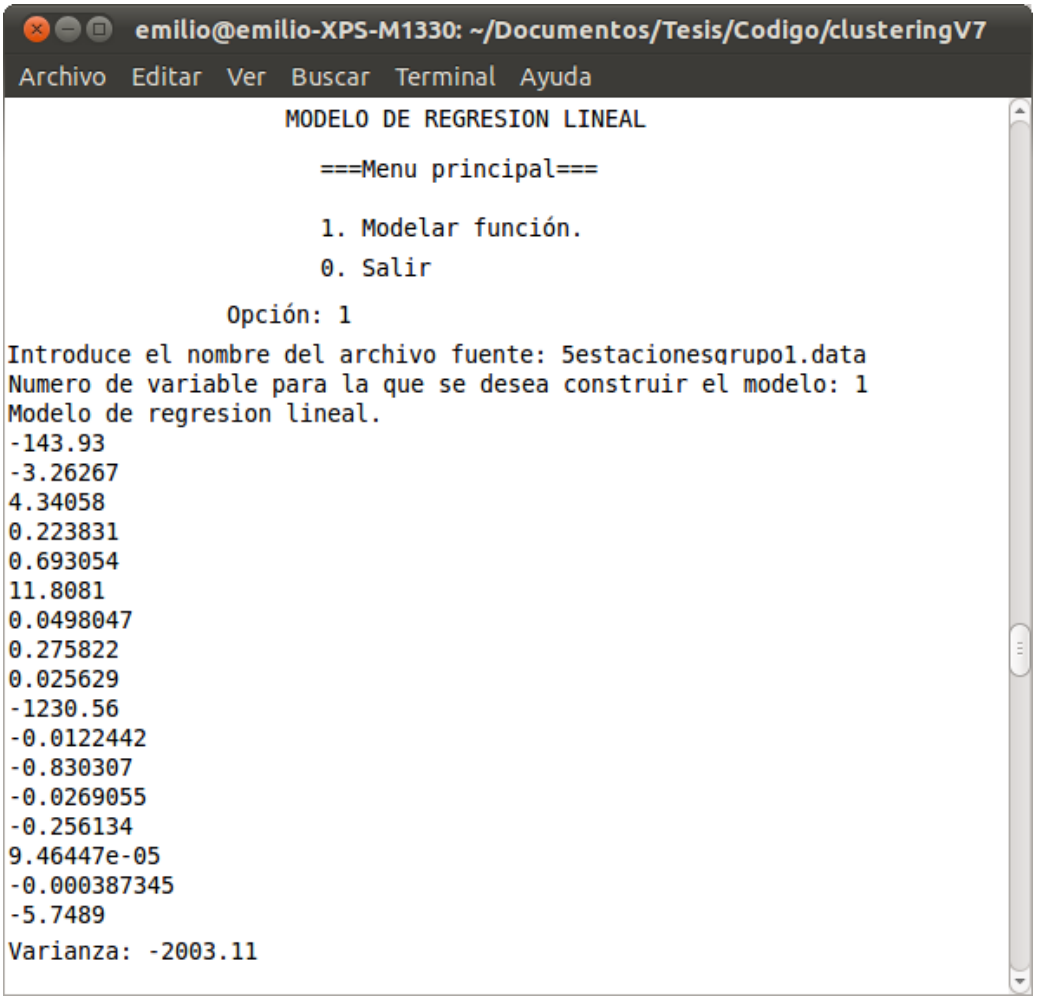

Figura C.4: Ejecución de mínimos cuadrados para obtener el modelo de regresión lineal del primer grupo usando 5 estaciones hidrométricas.