

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS



**ANÁLISIS DE CONGLOMERADOS EN ESTACIONES PLUVIOMÉTRICAS
MEDIANTE EL EXPONENTE DE HURST Y VARIOGRAMAS EN LA
CUENCA RÍO SAN JUAN**

POR

FRANCISCO GERARDO BENAVIDES BRAVO

**COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS CON ORIENTACIÓN EN MATEMÁTICAS**

JUNIO, 2017

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
CENTRO DE INVESTIGACIÓN EN CIENCIAS FÍSICO MATEMÁTICAS



**ANÁLISIS DE CONGLOMERADOS EN ESTACIONES PLUVIOMÉTRICAS
MEDIANTE EL EXPONENTE DE HURST Y VARIOGRAMAS EN LA
CUENCA RÍO SAN JUAN**

POR

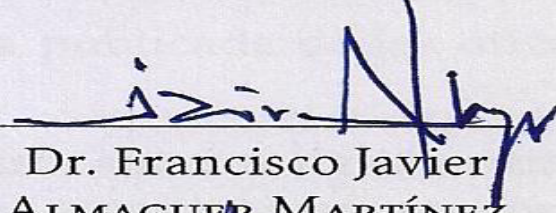
FRANCISCO GERARDO BENAVIDES BRAVO

**COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS CON ORIENTACIÓN EN MATEMÁTICAS**

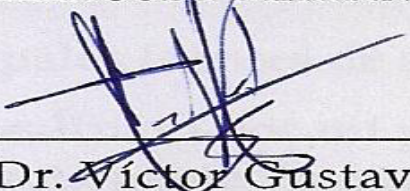
JUNIO, 2017

Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas
Centro de Investigación en Ciencias Físico Matemáticas

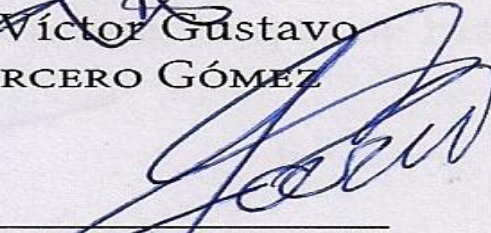
Los miembros del Comité de Tesis recomendamos que la Tesis "ANÁLISIS DE CONGLOMERADOS EN ESTACIONES PLUVIOMÉTRICAS MEDIANTE EL EXPONENTE DE HURST Y VARIOGRAMAS EN LA CUENCA RÍO SAN JUAN", realizada por el alumno Francisco Gerardo Benavides Bravo, matrícula 0095431, sea aceptada para su defensa como opción al grado de Doctor en Ciencias con Orientación en Matemáticas (Modelación).



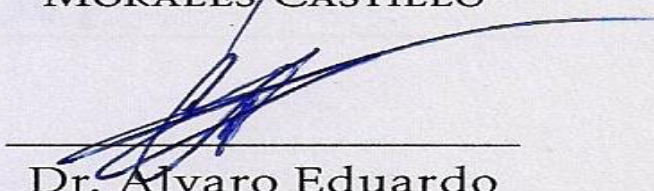
Dr. Francisco Javier
ALMAGUER MARTÍNEZ




Dr. Víctor Gustavo
TERCERO GÓMEZ



Dr. Javier
MORALES CASTILLO



Dr. Alvaro Eduardo
CORDERO FRANCO



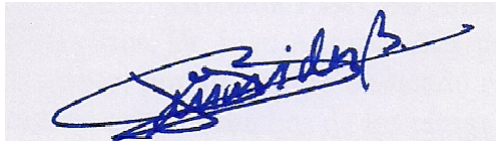
Dr. Gerardo Jesús
ESCALERA SANTOS

Declaración del Autor

Yo, Francisco Gerardo BENAVIDES BRAVO, declaro que la presente tesis titulada «ANÁLISIS DE CONGLOMERADOS EN ESTACIONES PLUVIOMÉTRICAS MEDIANTE EL EXPONENTE DE HURST Y VARIOGRAMAS EN LA CUENCA RÍO SAN JUAN» y los trabajos presentados en ella son de mi propio proceso de investigación y autoría. Confirmando que:

- Este trabajo fue realizado en su totalidad o principalmente, para obtener la candidatura al grado de Doctor en esta Universidad.
- Cualquier parte de esta tesis no ha sido previamente sometida a obtención de un grado de titulación en esta universidad o cualquier otra institución.
- Donde he consultado la obra publicada de los otros investigadores, lo he atribuido con claridad.
- He citado los trabajos de los otros autores, dando siempre la fuente. Con quizá algunas excepciones de citas, esta tesis es enteramente mi propio trabajo.
- He reconocido todas las principales fuentes de ayuda.
- La tesis se basa en el trabajo realizado por mí persona o en forma conjunta con mis asesores, y he dejado claro, exactamente, lo que se hizo por ellos y lo que he contribuido yo.

Firma:



Fecha: 15 de junio de 2017

Agradecimientos

Deseo expresar mi agradecimiento al Dr. Francisco Javier ALMAGUER MARTÍNEZ, director de esta tesis, mi guía durante estos últimos 5 años de mi vida. Él, junto al Dr. Víctor Gustavo TERCERO GÓMEZ docente e investigador, ahora del Instituto Tecnológico de Monterrey, quienes con su apoyo influyeron en mi formación académica. También agradezco la participación del Dr. Javier MORALES CASTILLO, al Dr. Alvaro Eduardo CORDERO FRANCO y al Dr. Gerardo Jesús ESCALERA SANTOS por haber aceptado participar en este comité y tomarse el tiempo de revisar esta tesis proporcionando, además, invaluable recomendaciones, a todos ellos, gracias, ya que sin sus observaciones, consejos, ideas, críticas, no hubiese sido posible el desarrollo y culminación de esta tesis.

Deseo agradecer a todos mis compañeros del Doctorado en Ciencias con Orientación en Matemáticas, Dra. María Esther GRIMALDO REYNA y en especial a mi amigo Lic. Roberto SOTO VILLALOBOS y al M.C. Mario Alberto AGUIRRE LÓPEZ, M.C. Jesús ARRIAGA GARZA, quienes constantemente me apoyaron y alentaron en esta proyecto que emprendí con mucho esfuerzo .

A mis compañeros y amigos de trabajo Ing. Marta Gabriela RÍOS NAVA, Ing. Ángela Gabriela BENAVIDES RÍOS, Lic. Miguel Ángel SALAZAR SALINAS, Dr. Juan Ramón GARCÍA JIMÉNEZ, MIA. María Gricelda PÁMANES AGUILAR, a la Ing. Gricelda RÍOS PÁMANES, al Dr. Ricardo PULIDO RÍOS, Dr. Juan Antonio ALANIS RODRÍGUEZ, al M.C. Eduardo BENITEZ TÁMEZ, mi gran reconocimiento por sus palabras de apoyo, por su tiempo dedicado en la discusión de gran parte de esta tesis, lo que me permitió aclarar muchos de los temas que contiene la misma.

Al igual que a la Facultad de Matemáticas de la U.A.N.L. y al Instituto Tecnológico de NUEVO LEÓN, por el apoyo brindado para la realización de este programa de Doctorado.

Agradezco la ayuda prestada por la Comisión Nacional del Agua (CONAGUA) ya que proporcionó la información para la realización de esta tesis, sin este apoyo, nunca hubiese sido posible el desarrollo de la misma.

A los generadores de las ideas que fortalecieron este trabajo, gracias por compartir su ciencia.

Finalmente, les doy una gran felicitación

$$\sum_{k=1}^{10,095,431} \{gracias_{[atodoslosqueporconfianzaolesdijegraciasporquepensequesobradecirlo]}\}^k \quad (1)$$

En un carácter muy, pero muy especial agradezco y dedico este trabajo a: MI FAMILIA, MARTA GABRIELA (GABY) la compañera de la mayor parte de mi vida, y a mis hijos: ÁNGELA GABRIELA, FRANCISCO GERARDO, por el sacrificio, espera, aguante, apoyo, comprensión, ayuda y por el compartir un papá que en varias ocasiones se sintió perdido y derrotado con lo que día tras día crecía más y más: lo muy intenso o insoportable de mí, por el tiempo que les reste, por las cosas que deje de hacer con ellos y para ellos .

También la dedico a, mis padres: HÉCTOR† Y VIRTUDEST†, mis hermanos: HÉCTOR Y VIRTUDEST†.

Mensaje del autor

«Una serie de temas y sus variantes, permiten que se pueda escribir sobre estos, aún cuando prácticamente no sean claros, la formación académica que se adquiere, permite profundizar en esta información, acumular y generar conocimientos, dejando un trabajo de investigación y, además, nos otorga la posibilidad de trabajos futuros.»

BENAVIDES

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Resumen

Facultad de Ciencias Físico Matemáticas

Centro de Investigación en Ciencias Físico Matemáticas

Doctor en Ciencias con Orientación en Matemáticas

ANÁLISIS DE CONGLOMERADOS EN ESTACIONES PLUVIOMÉTRICAS MEDIANTE EL EXPONENTE DE HURST Y VARIOGRAMAS EN LA CUENCA RÍO SAN JUAN

por Francisco Gerardo BENAVIDES BRAVO

Un tema importante en el estudio del comportamiento de las series temporales y, en particular, series de tiempo meteorológicas, es la dependencia a largo plazo. En esta tesis se analiza el comportamiento de las variaciones de precipitación en diferentes períodos, utilizando el análisis de las correlaciones de largo alcance. Variogramas y exponente de Hurst se aplicaron a los datos históricos de diferentes estaciones pluviométricas de la cuenca del río San Juan, en la región hidrográfica RH-24 México. La base de datos fue proporcionada por la Comisión Nacional del Agua (CONAGUA). A los variogramas, se les obtuvo el exponente de Hurst y se utilizó como una entrada para llevar a cabo un análisis de agrupamiento de estaciones de lluvia. Grupos de muestras homogéneas que pueden ser útiles en un análisis de frecuencia regional se obtuvieron a través del proceso.

Palabras clave: Series de Tiempo, correlaciones de Largo-alcance, lluvia, semivariogramas, exponente de Hurst, análisis de clústers.

Software: Matlab, GNU-R

An important topic in the study of the time series behavior and, in particular, meteorological time series, is the long-range dependence. This thesis explores the behavior of rainfall variations in different periods, using long-range correlations analysis. Variograms and Hurst exponent were applied to historical data in different pluviometric stations of the río San Juan watershed, at the hydrographic RH-24 Mexico region. The database was provided by the Water National Commission (CONAGUA). Using the semivariograms, the Hurst exponent was obtained and used as an input to perform a cluster analysis of rainfall stations. Groups of homogeneous samples that might be useful in a regional frequency analysis were obtained through the process.

Keywords: Time series, Long-range correlations, rainfall, semivariogram, Hurst exponent, cluster analysis

Software: Matlab, GNU-R

Índice

Índice	2
1. Introducción	4
1.1. Antecedentes y descripción del problema.	9
1.2. Justificación del problema.	14
1.3. Planteamiento del problema.	14
1.4. Hipótesis.	15
1.5. Objetivo General.	15
1.6. Objetivo Específico.	15
2. Metodología	16
2.1. Marco metodológico	17
2.2. Depuración de datos	18
2.3. Descripción de series de tiempo	18
2.4. Estimación de la dimensión fractal en la práctica	20
2.5. Medición del exponente de Hurst (H).	21
2.5.1. Mediciones del exponente de Hurst	23
2.6. Variograma	25
2.7. Identificación de regiones homogéneas	26
2.8. Validación	27
3. Resultados	32
3.1. Discusión	33
3.2. Descripción del problema	35
3.3. Resultados	37
4. Conclusiones y trabajo futuro	55
5. Bibliografía	59
6. Anexos	63
6.1. Descripción de la cuenca del río San Juan	64
6.2. Técnicas de regionalización	66
6.3. Recolección y depurado de información	68
6.4. Gráficos de las Series de Tiempo	73
6.5. Fractales	78
6.5.1. Relación entre dimensiones Fractales	80
6.6. Gráficos del exponente de Hurst	83

6.7. Variogramas y sus Gráficos	87
6.8. Gráficos de variogramas enlazados	94
6.9. L-Momentos	98
6.9.1. Estimación de parámetros	101
6.10. Histogramas y Distribuciones de probabilidad	108

Capítulo 1

Introducción

Introducción

El comportamiento de las diversas variables climatológicas, en el tiempo y en el espacio, han llamado la atención del ser humano desde la antigüedad, ya que estas condicionan parte del medio ambiente en el cual este desarrolla sus actividades.

El presente trabajo tiene la finalidad de analizar el comportamiento de 33 estaciones pluviométricas instaladas en la región hidrológica río Bravo-San Juan (RH-24) o cuenca del río San Juan, y efectuar una clasificación de acuerdo a sus características de comportamiento de estas.

La partición espacial de un conjunto de estaciones en grupos sustancialmente homogéneos con respecto a parámetros similares, facilita estudios posteriores de variabilidades en el tiempo o correlaciones con otras variables externas. Particularmente, la clasificación de estas, según su ciclo anual, conduce a la obtención de una regionalización según el régimen de lluvias. Esta regionalización puede ser útil en la búsqueda de relaciones con agentes físicos externos a la hora de ajustar pronósticos, y quizá en múltiples aplicaciones hidrológicas.

La mayor parte de los métodos de regionalización desarrollados se efectúa mediante el análisis estadístico de frecuencias de máximos de una variable hidrológica. La estimación de parámetros, a partir de una muestra pequeña, presenta dificultades debido a la incertidumbre existente respecto a su representatividad. Es incierto efectuar análisis de dispersiones o de coeficientes de variación debido a que las bases de datos son irregulares, y establecer relaciones entre estaciones es casi fortuito. Lo anterior conduce a métodos que asumen una región homogénea respecto a ciertas características estadísticas, lo que permite aprovechar el conjunto de información disponible en la región. La fase más importante en la utilización de información regional, es la de definir las estaciones de precipitación que se consideran similares entre sí, y que puedan ser agrupadas según el grado de heterogeneidad que se quiera asumir para tener un beneficio en el tratamiento conjunto de la información. Aunque no existe un procedimiento que asegure correctamente la definición de una región para el análisis de precipitación, Lettenmaier y Potter (1985), reportan las ventajas de agrupar los datos de precipitación máxima de distintas estaciones con coeficientes de variación ($C_v = \sigma/\bar{x}$, σ = desviación típica, \bar{x} = media) bajos y homogéneos haciendo referencia a la relación entre el tamaño de la media y la variabilidad de la variable. En los análisis de precipitación máxima, la mayoría de los métodos toma como base la regionalización.

Una parte de la teoría de eventos extremos se basa en el supuesto de que dichos valores se extraen de poblaciones idénticas, independientes e igualmente distribuidas (iid) ya que esto simplifica las operaciones de muchos métodos estadísticos. Estas condiciones no se cumplen en las series de tiempo meteorológicas porque comúnmente muestran estacionalidades y dependencias.

En el caso de una serie estacionaria, se remedia el problema de la dependencia analizando la serie en bloques. Esto se justifica con el hecho que los valores extremos de cada

bloque tienden a volverse independientes entre sí conforme aumenta el tamaño de los bloques.

Por ejemplo, una serie de temperaturas horarias muestra la correlación fuerte del ciclo diurno. Si se toman bloques de un día, los valores extremos diarios muestran una correlación menor pero apreciable, durante típicamente varios días. Si los bloques son de un mes, la correlación entre los eventos extremos se pierde y se puede suponer que se cumple la condición de independencia, Javier Soley, Noviembre, 2010[1].

En el análisis de excedencias, las dependencias entre las poblaciones pueden producir que las excedencias de eventos extremos ocurran en grupos o cúmulos y no aleatoriamente. En este caso, se debe buscar una manera de identificar los grupos o cúmulos dejando, para el análisis, únicamente el valor máximo de cada grupo o cúmulo.

Una serie temporal, que no es otra cosa más que una sucesión de valores de una variable observada en intervalos de tiempo igualmente espaciados y que generalmente es aleatoria, puede verse influenciada por como fueron almacenados sus registros históricos; mensual, bimestral, trimestral, semestral, días, horas, etc. A este tipo de series le llamaremos “*estacionales*” ya que esa manera de registrarlos las afecta.

Se dice que estas poseen tendencia, cuando observamos cierto comportamiento en ellas o cuando sus valores oscilan alrededor de curvas que podemos modelar mediante una ecuación y la cual nos permite establecer criterios como, el de estacionalidad, creciente ó decreciente, tendencia, ciclicidad o aleatoriedad.

En nuestro trabajo, las series de tiempo son registros históricos que fueron analizadas con el interés de particionar la región en conglomerados, esto es, regiones que poseen un comportamiento similar, ajustando una función de probabilidad que permita caracterizarlos.

El caso de las series estacionales no es tan simple, ya que no existe una metodología general. Un método para analizar una serie estacional es asignarle una dependencia en tiempo a los parámetros de la distribución de probabilidad con la que se modelan los datos. Por ejemplo, una serie de varianzas constante y con un valor medio que varía en tiempo, podría modelarse con una distribución normal $N(a_0 + a_1 t + a_2 t^2; \sigma^2)$ y calcular los coeficientes a_i , usando técnicas de máxima verosimilitud.

En otros casos es posible analizar la serie, dividiéndola en segmentos o estaciones más pequeñas, ya que la función de máxima verosimilitud puede subdividirse para tomarlas en cuenta. Por ejemplo, una serie que muestra variaciones fuertes durante el año podría subdividirse en los 12 meses del año y plantear la función de verosimilitud como un producto de doce funciones (una para cada mes).

La complejidad de los modelos medioambientales es que virtualmente cualquier proceso físico lleva consigo una variabilidad en el espacio y tiempo, ya que la interacción supone que los casos cercanos están en el espacio. Se disponen de datos que se revisaron buscando cierto patrón espacial mediante observaciones gráficas.

Esto puede permitir apreciar cierto comportamiento y analizar más a fondo períodos particulares en donde se pueda hacer un mejor análisis, o bien, una clasificación de zonas homogéneas, que formen conglomerados o clústeres espaciales, además, si observamos los acumulados de algunos meses, nos permite decir que si, en algún período particular, después de ciertos acumulados mensuales, se le agrega la presencia de un evento extremo, puede ocurrir una avenida de agua como las ocurridas con los huracanes, Beulah, Gilberto y Alex.

En particular, nuestro trabajo se centrará en efectuar conglomerados de las estaciones pluviométricas con similares comportamientos, utilizando para ello el exponente de Hurst, aplicándolo a los datos de manera directa así como a su variograma, caracterizaremos distribuciones de probabilidad a cada una de las estaciones pluviométricas llevando a cabo conglomerados, efectuaremos simulaciones que analizaremos con métodos como el modelo autorregresivo de promedio móvil o **ARMA** (*acrónimo del inglés autoregressive moving average*), y compararemos estos nuevos conglomerados de la simulación con los conglomerados originales mediante el coeficiente de Jaccard para analizar semejanza entre los conglomerados simulados y originales, tanto de manera directa a los datos así como a los variogramas mediante el exponente de Hurst, se efectuaran particiones que llamaremos naturales, esto es, ordenamientos de menor a mayor y también utilizaremos el método de k-medias para investigar cual es el más eficiente.

El variograma, utilizado en la Geoestadística impulsada por George Matheron[2] (1930-2000), es una técnica que se aplicó a nuestros datos la cual permitió observar un suavizamiento en las series de tiempo, se optó por medir esa variabilidad aplicando un análisis de fractalidad ó análisis de Reescalado, encontrando así el denominado exponente de Hurst[3], que permite medir dicha variabilidad y capturar su magnitud, con la cual generaremos intervalos de valores numéricos y por análisis de jerarquización de k-medias, conjeturando que, aquellos que estén en ciertos intervalos, poseen características semejantes, y con ello poder clasificar en conglomerados las estaciones que fueron analizadas.

La Comisión Nacional del Agua (CONAGUA), organismo administrativo desconcentrado de la Secretaría de Medio Ambiente y Recursos Naturales, creado en 1989 con la responsabilidad de administrar, regular, controlar y proteger las aguas nacionales en el país, proporcionó las bases de datos de 41 estaciones pluviométricas distribuidas en la cuenca del río San Juan que fue la zona de estudio elegida para el presente trabajo. Las estaciones constan de datos desde 1980 hasta el 2012, distribuidos en meses.

Aquí se propone establecer una metodología para identificar comportamiento similar en estaciones pluviométricas, analizando los datos de los registros históricos de estas, aplicando índices de fractalidad, el exponente de Hurst, tanto a los datos directos como a los variogramas de estos y que nos permitirá llevar a cabo una conglomeración de estaciones. Una vez hecho esto, se estimará los parámetros de comportamiento como: promedio, desviación, curtosis y asimetría, mediante L-Momentos, para proponer una función de probabilidad que los caracterice.

Organización de la tesis

A lo largo de los próximos capítulos se emplea una terminología que es importante aclarar para una correcta comprensión de los mismos. El elemento fundamental de análisis considerado en esta investigación son las series temporales representadas por un conjunto de datos con estructura compleja y cuando se habla de estos se hace referencia al conjunto de todos aquellos atributos propios de cada una de las 33 estaciones que se analizó. La organización de esta tesis comprenderá lo siguiente:

Capítulo 1 Introducción, antecedentes, descripción del problema, justificación.

Capítulo 2 Metodología aplicada en el desarrollo de la Tesis.

Capítulo 3 Presentación de Resultados. Aquí se discuten las contribuciones del presente trabajo.

Capítulo 4 Conclusiones.

Anexos

- * Se presenta una explicación breve de la revisión de los datos y su filtración para llevar a cabo el análisis. * Introducción a las series de tiempo.
- * Explicación de Fractales.
- * Explicación de variogramas, técnica aplicada a la presente investigación.
- * Rango reescalado, técnica aplicada posteriormente a los variogramas para la obtención del exponente de fractalidad.
- * Se esboza el porque establecer una clasificación de regiones homogéneas utilizando el exponente de Hurst[3].
- * Aplicación de L-momentos, utilizados para calcular, ubicación, asimetría y curtosis, parámetros necesarios para el ajuste de una función de probabilidad.

1.1. Antecedentes y descripción del problema.

En México existen 722 cuencas de aguas superficiales y 653 acuíferos agrupadas en 37 regiones hidrológicas y 13 gerencias regionales administrativas de la CONAGUA con numerosas estaciones pluviométricas como se muestra en la Fig.1.1.

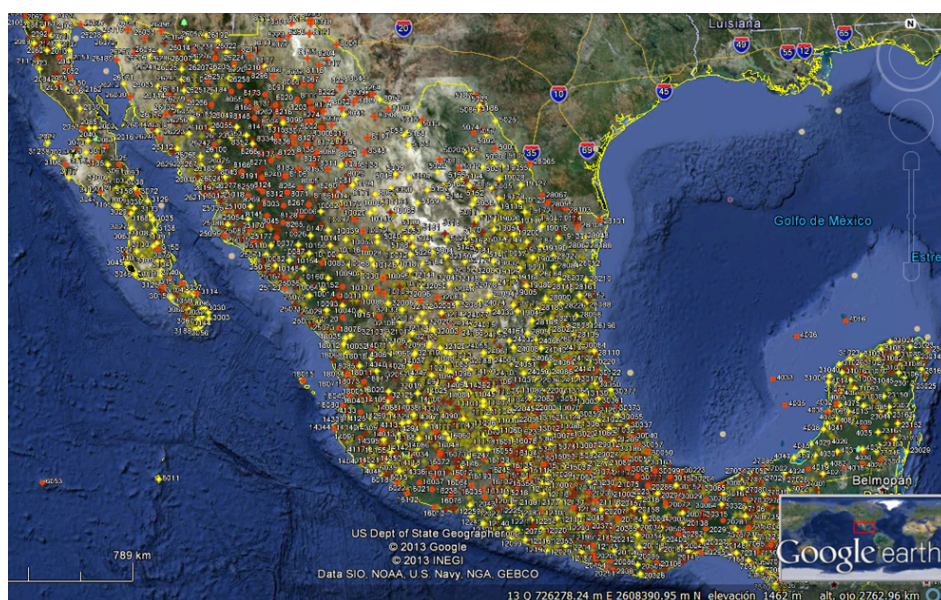


FIGURA 1.1: Estaciones pluviométricas de México. Imagen obtenida de: "Google Earth"

La cuenca del río San Juan integra diversos centros urbanos, entre los cuales se encuentra la Ciudad de Monterrey, con su área metropolitana, y Saltillo (capital de Coahuila), que en conjunto a las zonas citricolas y agrícolas conforman un sistema económico que es crítico tanto para la región como para el país. En el Anexo 6.1 damos una descripción de la ubicación de la cuenca del río San Juan.

Eventos atípicos han mostrado lo susceptible de la región ante fenómenos extremos como el caso de los huracanes, Beulah en 1967, Gilberto en 1988, Emily en 2005 y el ocurrido más recientemente, Alex, en el 2010.

Este tipo de riesgos hidrológicos ocasionales de grandes avenidas, junto a otro riesgo de grandes sequías, conforman parte de los comportamientos extremos de la zona de análisis.

Si se desea lograr una mejor planificación urbana, elaboración de planes de contingencia y paliación de riesgos, se requiere contar con controles regionales del perfil de riesgo de estos comportamientos extremos.

Entre las estrategias de análisis de riesgo se encuentra la modelación espacial de niveles de precipitación para diferentes períodos en las zonas bajo estudio.

En esta investigación, se realizó una revisión de la información histórica proveniente de 41 bases de datos que dan seguimiento a las estaciones meteorológicas de la cuenca del río San Juan. Cada estación está caracterizada por su ciclo anual en un período de 1984-2011. De estas 41 estaciones se seleccionaron 33 debido a que algunas poseen menos de esta información y otras que dentro de ese período carecen de algunos años de registro. Los datos fueron recopilados por la CONAGUA. Se observó el comportamiento de las diferentes estaciones y se efectuó un análisis de fractalidad y se agrupó en conglomerados mediante el exponente de Hurst, también se calculó los variogramas de estas y se agruparon mediante el exponente de Hurst. El ajuste de modelos probabilísticos nos permite efectuar simulaciones y validar si es recomendable esta metodología para pronósticos de precipitación de lluvias para distintos períodos de interés.

Los fenómenos extremos causados por precipitaciones en la región generalmente ocurren en verano y otoño.

Estas lluvias están asociadas a la influencia de sistemas atmosféricos de gran escala, como la zona de convergencia intertropical y los vientos alisios, así como a sistemas transitorios y ciclones tropicales. La zona intertropical de convergencia es una región de bajas presiones alrededor del globo y cerca al Ecuador, donde tiene lugar el encuentro de los vientos del este (alisios) provenientes de cada hemisferio, eso da lugar a movimiento de ascenso de la masa de aire y a una abundante precipitación.

El decir fenómeno extremo, es como sinónimo de acontecimiento poco común bajo la perspectiva humana. Sin embargo, la formación de una gota de lluvia es un fenómeno natural de la misma manera que un huracán, esta expresión también se refiere, en general, a los peligrosos fenómenos naturales también llamados "*desastres naturales*", la lluvia, por ejemplo, no es en sí un "*desastre*", pero, "*puede serlo*", dependiendo de la perspectiva humana, si ciertas condiciones se reúnen. La mala planificación urbana, con la construcción de estructuras en lugares vulnerables e inundaciones, puede causar efectos desastrosos para los seres humanos. Los huracanes, como un factor importante de las lluvias de verano en México, se les asocia con este tipo de situaciones de desastre y pérdida de vidas humanas, constituyen los principales sistemas productores de lluvia en verano para regiones como el noreste del país. Un caso de este tipo ocurre cuando el flujo que acompaña a un ciclón tropical choca con una cadena de montañas, lo que produce un ascenso de la masa de aire cargada de humedad, causando una magnificación del proceso convectivo. Ejemplos de esta situación se presentaron en Nuevo León y Tamaulipas durante el huracán Gilberto en septiembre de 1988, el huracán Paulina (afectando a Guerrero y Oaxaca) en 1997 y Alex en 2010. Las inundaciones extraordinarias, así como la persistencia de la sequía en grandes extensiones de la república, es lo que motiva a efectuar algunos planteamientos que pueden servir para entender mejor estos fenómenos y, con esa base, disminuir los daños que se derivan de los escurrimientos extremos.

En el estudio de los escurrimientos máximos por precipitaciones extremas, las inundaciones constituyen la principal preocupación. Dependiendo de su duración, la magnitud de las áreas afectadas y el tipo de afectación, las inundaciones provocadas por lluvias intensas en cuencas con respuesta rápida provocan los denominados “*flash floods*” (inundaciones repentinas), los cuales se acompañan casi siempre de una gran cantidad de lodo. Algunas de estas se han presentado en el Valle de México, en el arroyo Topo Chico y Santa Catarina en Monterrey y en las serranías de Puebla y Veracruz.

Inundaciones de larga duración, por la lentitud con que se producen, causan lamentablemente pérdidas humanas e importantes pérdidas económicas tanto en zonas urbanas como en zonas rurales. Encharcamientos se presentan casi siempre en sectores urbanos, cuya principal consecuencia es el retraso en el desarrollo de las actividades productivas de la población.

Dada la diversidad de características de aquellas avenidas que pueden producir inundaciones de distinto tipo, el problema de la estimación del riesgo es complejo y es necesario caracterizarlo ya que esto permitiría diseñar una política de contingencia o mitigación de los daños con medidas preventivas o resolutivas.

Debido a los riesgos presentados anteriormente es necesario realizar estudios regionales de lluvias extremas. Al llevar a cabo el estudio regional, es necesario identificar los sitios similares y agruparlos en conglomerados y, de acuerdo con el tipo de afectación, analizar con probabilidades los registros históricos con objeto de definir, en cada caso, la probabilidad de que en un año cualquiera se presente una inundación asociada a una lluvia extrema o cuál será el nivel de retorno correspondiente.

Para la modelación de estos fenómenos se puede utilizar la metodología de análisis regional de frecuencias, modelos jerárquicos basados en estadística bayesiana y procesos **max-estables**. En el Anexo 6.2 damos una breve explicación del porque es requerido diferentes análisis al respecto. El primero de ellos ha sido aplicado en estudios en EE.UU, el Noroeste de México, Chile, Turquía y algunas regiones de Europa. Este análisis permite realizar predicciones decrecientes, es decir, estimaciones asociadas a una determinada probabilidad de excedencia, con base en todos los datos observados en varias estaciones hidrométricas de una región.

La condición implícita en este planteamiento es que las estaciones utilizadas sean homogéneas en algún sentido estadístico-hidrológico. Las pruebas de homogeneidad regional han evolucionado bastante desde el enfoque de Dalrymple[4], hasta las pruebas que emplean características climáticas y/o fisiográficas de las cuenca. Esta metodología se fundamenta en el supuesto de que los datos provienen de sitios en una región homogénea, en la cual se asume una distribución de frecuencias idéntica, excepto por un factor de escala específico del sitio, en la cual, además, se agregan para la mejora de la estimación de la relación en todos los sitios, Hoskin and Wallis[5].

Para la aplicación de esta metodología se siguen cinco etapas: (1) revisión y preparación de los datos, (2) identificación de regiones homogéneas, (3) selección de la distribución de frecuencia, (4) cálculo de parámetros y estimación de la función de cuantiles, y (5) mapeo del periodo de retorno.

Algunas de las distribuciones que más se utilizan para el ajuste de dicha información hidrológica son: distribuciones de: Gumbel, Pearson, Log Pearson, Gamma, Normal, Log Normal, General de Valores Extremos, entre otras.

La base de esta investigación fueron datos provenientes de 41 estaciones pluviométricas, a las que se les realizó un análisis estadístico de reescalado o índice fractal de Hurst[3] y variogramas, clasificando en conglomerados, considerando un comportamiento similar en un intervalo, se tomó la mejor distribución de probabilidad, estimando los parámetros: media, desviación, asimetría y curtosis mediante L-momentos.



FIGURA 1.2: Ubicación de las estaciones pluviométricas de estudio: "Google Earth"

Cuando se dispone de observaciones limitadas de eventos hidrológicos, se ve comprometida la capacidad de proveer una adecuada caracterización, análisis y predicciones de un fenómeno. Sin embargo, el análisis puede mejorarse mediante la identificación de muestras homogéneas que pueden usarse en combinación para hacer mejores estimaciones de un modelo de probabilidad.

Esta es una de las principales preocupaciones dentro de la práctica del análisis regional de frecuencia (ARF), donde el resultado final es obtener una estimación de eventos extremos dentro de una zona geográfica, la cual se puede utilizar como entrada en un análisis de riesgos, administración del agua, zonificación y aplicación del agua a usos de la tierra, Hosking y Wallis[5].

Sin embargo, la estimación de eventos extremos se considera un problema complejo, sobre todo porque la información es generalmente limitada, existe correlación serial, muchos puntos de cambio que pudieran estar presentes y observaciones a seguir sobre tendencias y patrones estacionales.

Para enfrentar estos problemas, muchos estudios de series de tiempo hidrológicas han sido aplicado con éxito en el pasado Machiwal y JAI[24], sin embargo otros esfuerzos de

investigación habían centrado su atención en pruebas de detección de tendencia, dejando a un lado otras propiedades importantes tales como de estacionalidad, la homogeneidad, la periodicidad y la persistencia. Al abordar estas propiedades, sería posible una mejor selección de muestras homogéneas, y como consecuencia, los profesionales podrían lograr mejores predicciones.

Trabajos previos sobre el análisis de series de tiempo en climatología con aplicaciones en la precipitación se remontan a Bhuiya[25], con el desarrollo de una prueba de estacionariedad después que componentes periódicos y de tendencia se restaron de la serie hidrológica.

Otras pruebas de tendencia de Buishand[26] las utilizó para evaluar la diferencia de precipitación entre zonas rurales y urbanas de Amsterdam y Rotterdam. Buishand[27,28] construyó varias pruebas de homogeneidad y la media de la serie usando las sumas acumulativas, pruebas probabilísticas e inferencia bayesiana. Kothyari[29], et al. evaluó tres estaciones en la India, Agra, Dehradun y Delhi para probar cambios en la precipitación y temperatura, proporcionando la evidencia de un cambio en el número de días de lluvia durante la temporada de monzones y un incremento en la temperatura.

Giakoumakis y Baloutsos[30] realizaron un análisis de tendencia en la serie histórica de precipitaciones anuales de la cuenca del Eveno Riven en Grecia. Aplicando diferentes pruebas de aleatoriedad, disminuyendo las tendencias encontradas en los registros de precipitación. Otros autores relacionados con análisis de tendencias, puntos de cambio encontradas en la literatura y homogeneidad son Ángel y Huff [31], Mirza[32], et al., Tarhule y Woo[33], De Lués[34] et al., Kripalani, Adamowski y Kulkarni[35], Bougadis [36], Yu[37] et al., Kumar[38], et al.

Una revisión exhaustiva de estos trabajos puede encontrarse en Machiwal y JAI[24], con descripciones de acontecimientos relacionados con en análisis de series de tiempo hidrológicas.

Desarrollos recientes en el análisis hidrológico se incluyen en las obras de Golian[39], et al., con una clasificación y clusterización, llevando a cabo un enfoque de los datos de la precipitación, utilizando este como un método de clasificación natural y el algoritmo de (FCM) *fuzzy c-means*. Shi[40], et al., analizaron las variaciones en las tendencias de datos de precipitación utilizando un método de regresión lineal, la prueba de Mann-Kendall y el exponente de Hurst.

1.2. Justificación del problema.

Las mediciones hidrológicas se hacen con el fin de obtener información de los procesos hidrológicos. Esta información se utiliza para poder entender mejor estos procesos y para el diseño, análisis y toma de decisiones. Sin embargo en nuestro país gran parte del territorio no cuenta con equipo de medición o en ocasiones éste no opera de manera adecuada, lo que repercute directamente en la calidad o confiabilidad de la información.

La regionalización hidrológica, transfiere información a puntos sin medición y para eso requiere que tengan un comportamiento hidrológico semejante.

La importancia de establecer una metodología que permita tener una información más confiable, es necesaria para resolver esta problemática.

El objetivo de este trabajo es identificar regiones hidrológicamente homogéneas y establecer una metodología que permita resolver dicha tarea.

1.3. Planteamiento del problema.

El estudio de una regionalización parte del principio que se requiere subdividir o fraccionar cierta zona en un conjunto y sistema de regiones menores dentro de los límites de manera que facilite el ejercicio del control, ya sea administrativo, asignación de los recursos, políticas de dirección. Regionalización es un procedimiento para modificar el orden territorial en unidades más pequeñas con características comunes y representa una herramienta metodológica básica en la planeación ambiental, pues permite un manejo adecuado donde se puede analizar la dependencia entre objetos de análisis. Con el objetivo de la identificación de regiones homogéneas, es necesario encontrar correlaciones entre las diferentes bases de datos y es necesario recortar, en algunos casos, información para encontrarlas.

El presente trabajo pretende responder y aportar información en relación a las siguientes preguntas: ¿ Existe la posibilidad de capturar la magnitud en la variación de los datos mediante el análisis fractal ?, ¿ Es el exponente de Hurst otra herramienta estadística para hacer esto ?, ¿ Podrá servirnos como relación ?, y de ser así, ¿ Puede esta medida, permitirnos efectuar cúmulos que tengan características similares ?.

1.4. Hipótesis.

Procedimientos como el análisis regional de frecuencias son utilizados para ajustar una muestra de datos a un tipo de distribución, asociando la forma de la distribución a un número finito de parámetros. La asimetría y la curtosis se utilizan comúnmente para establecer la proximidad de los valores observados (muestra) a diferentes tipos de distribuciones. Sin embargo, el cálculo de estos estadígrafos es sensible al tamaño de la muestra. ¿Puede el exponente de Hurst, formar parte de una estrategia para identificar regiones con comportamiento similar?

1.5. Objetivo General.

Establecer una modificación a la metodología para la identificación de Regiones Homogéneas dentro del análisis regional de frecuencias utilizando ahora el exponente de fractalidad de Hurst.

1.6. Objetivo Específico.

Proponer una regionalización de las estaciones de la región RH-24, mediante el análisis de comportamiento fractal, y así identificar comportamientos similares en las diferentes bases de datos de nuestras estaciones pluviométricas para nuestro caso de estudio. Posteriormente aplicar el método de L-Momentos y estimar parámetros para el ajuste de las funciones de probabilidad que se utilizan tradicionalmente, adecuándose a la información hidrológica (Generalizada de valor extremo, lognormal, Pearson tipo III, Gamma o Gumbel).

Capítulo 2

Metodología

2.1. Marco metodológico

El presente trabajo integra un procedimiento estadístico en el marco de la teoría de los procesos estocásticos, donde las unidades de análisis son series de registros de precipitaciones históricas de lluvia en la cuenca del río San Juan. La variable de estudio es la cantidad acumulada mensual de lluvia registrada en milímetros.

Aunque en la cuenca del río San Juan hay muchas estaciones meteorológicas, gestionadas por diferentes instituciones públicas y privadas, un buen número de estas se encuentran inactivas, otras han empezado a operar en las últimas décadas o tienen series de menos de 10 años continuos.

Después de un análisis en función de su altitud, longitud y latitud se complementaron las series de datos. De las 41 bases de datos de las estaciones gestionadas se seleccionaron 33 y estas se detallan en el capítulo 3. Se analizaron las series de tiempo y se les aplicó un método orientado a corregir los valores anómalos, correcciones de datos, relleno de celdas, análisis de posibles(outliers), para tener las bases homogéneas agrupadas por meses, con datos comprendidos desde 1984-2011.

El procedimiento estadístico que se utilizará es el de análisis de rangos reescalados ó R/S que será aplicado a los datos directos así como también a los variogramas, el cual se utiliza para cuantificar las correlaciones de largo alcance de datos de las diferentes estaciones pluviométricas con registros mensuales.

Teniendo en cuenta el análisis de la muestra de la serie histórica, se realizó un análisis de rango reescalado R/S para obtener una medida particular del exponente de Hurst[3]. El proceso se repite para cada estación pluviométrica en la región bajo análisis.

Este exponente de Hurst se utilizó como referencia para identificar si los datos presentan un comportamiento aleatorio puro, o si tiene tendencias subyacentes. Una vez obtenido este coeficiente, se aplicó un análisis de conglomerados (clusters), analizando frecuencias para clasificar las estaciones homogéneas.

Una ventaja del exponente de Hurst es la simplicidad de su algoritmo que puede ser utilizado para medir la condición de persistencia o antipersistencia de un proceso y proporciona una métrica que puede utilizarse para clasificar series de tiempo diferentes.

2.2. Depuración de datos

La depuración fue una actividad inevitable dentro del análisis de datos estadísticos. Esta primera actividad, nos permitió detectar algunos errores en la codificación de las variables cuantitativas como: digitación, datos inconsistentes, valores ausentes, fuera del rango, duplicados de las variables cuantitativas que nos interesa analizar, códigos numéricos que sirven para almacenar información con fines de procesamiento y no numéricos (datos en formato de texto). Los diferentes tipos de error que se presentaron en los datos, se analizaron con técnicas desarrolladas por diferentes investigadores para detectarlos y corregirlos.

Dado que existe una multiplicidad de técnicas para la depuración de los datos, resulta que no es trivial decidir cuál o cuáles deben ser utilizadas en nuestro caso particular. En el Anexo 6.3, se presentan algunos pasos como guía metodológica que puede apoyar para el análisis de datos de acuerdo con la naturaleza y la distribución de los mismos.

2.3. Descripción de series de tiempo

Una serie temporal cronológica es un conjunto de observaciones ordenadas en el tiempo, que pueden representar la evolución de una variable a lo largo de él. También podemos considerarla como la realización de un proceso estocástico en tiempo discreto, donde los elementos están ordenados y corresponden a instantes equidistantes del tiempo. El conjunto de observaciones se simboliza: $\{X(t_i), i = 1, 2, \dots, n\}$ donde t_i es independiente e indica sucesivos instantes o tiempos determinados (quinquenos, años, trimestres, meses, ..., etc).

“ X ” es la variable cuyo comportamiento a través del tiempo se desea estudiar o sea que la serie de tiempo es una serie estadística (información cuantitativa) cuyos valores han sido observados en el tiempo.

La serie puede simbolizarse como $\{X_n, n = 1, 2, \dots, n\}$, la cual representara el paso del tiempo, configurando un proceso estocástico que tendrá su propia función de distribución con sus respectivos momentos. Normalmente, para reconocer y caracterizar las distribuciones resulta complejo, basta con especificar la media y la varianza para cada $X(t)$, y la covarianza para variables referidas a distintos valores de t :

$$\mu_t = E[X(t)] \quad (2.1)$$

$$\sigma_t^2 = Var(X(t)) = E[X(t) - \mu_t]^2 \quad (2.2)$$

$$\gamma_{t_1, t_2} = Cov(X(t_1), X(t_2)) = E[(X(t_1) - \mu_{t_1})(X(t_2) - \mu_{t_2})] \quad (2.3)$$

donde $E[X(t)]$ es el valor esperado de $X(t)$, μ_t es la media, σ_t^2 es la varianza y γ_{t_1, t_2} la covarianza. $X(t)$ puede ser:

- Fuertemente estacionaria si todas las funciones de distribución conjuntas son constantes, o dicho con más propiedad, son “*invariantes con respecto a un desplazamiento en el tiempo*” (variación de t). Es decir, considerando que $t, t+1, t+2, \dots, t+k$ reflejan períodos sucesivos: $F(X_i, X_{i+1}, \dots, X_{i+k}) = F(X_{i+m}, X_{i+1+m}, \dots, X_{i+k+m})$ para todo t, k y m .
- Débilmente estacionario si:
 - Las esperanzas matemáticas de las variables aleatorias no dependen del tiempo, son constantes: $E[X_i] = E[X_{i+m}]$ para todo m .
 - Las varianzas no dependen del tiempo (y son finitas): $Var[X_i] = Var[X_{i+m}] \neq \infty$
 - Las covarianzas entre dos variables aleatorias del proceso correspondientes a períodos distintos de tiempo (distintos valores de t) sólo dependen del lapso de tiempo transcurrido entre ellas: $Cov(X_{t_1}, X_{t_2}) = Cov(X_{t_1+m}, X_{t_2+m})$

De esta última condición se desprende que, si un fenómeno es estacionario, sus variables pueden estar relacionadas linealmente entre sí, pero de forma que la relación entre dos variables sólo depende de la distancia temporal k transcurrida entre ellas.

Sea ρ la Función de Autocorrelación (FAC) de $X(t)$. Si la serie de tiempo es débilmente estacionaria, FAC estaría dada por:

$$\rho_X(t_1, t_2) = \frac{E[\{X(t_1) - \mu_X(t_1)\}\{X(t_2) - \mu_X(t_2)\}]}{\sqrt{Var(X(t_1))Var(X(t_2))}} \quad (2.4)$$

La serie de tiempo, $X(t)$, tiene la propiedad de Dependencia de Largo Alcance (DLA) si $\sum_{k=-\infty}^{k=\infty} \rho(k)$ diverge, esta puede determinarse de dos maneras:

- En el dominio del tiempo, donde se manifiesta con un alto grado de correlación entre sitios separados a cierta distancia.
- En el dominio de la frecuencia, donde se manifiesta como un nivel significativo de las frecuencias próximas a cero.

Existen diferentes técnicas para su estimación. Por ejemplo: el rango reescalado R/S , el método de momentos absolutos, el método modificado R/S o método de $Lo[12]$, periodograma, Wavelets, entre otros. En esta investigación empleamos el método R/S , el cual es uno de los más aplicados, debido a la efectividad y sencillez de su algoritmo. Éste fue propuesto por Hurst(1951), afinado, posteriormente por Mandelbrot y Wallis[13](1969) y Mandelbrot[14] (1972). En el Anexo 6.4 se muestran las series de tiempo de las 33 estaciones pluviométricas.

2.4. Estimación de la dimensión fractal en la práctica

Algunos fenómenos u objetos de la vida real pueden mostrar propiedades fractales, como lo mencionamos en el Anexo 6.5, y podemos calcular su dimensión fractal. Es aquí donde puede ser útil obtener la dimensión fractal de un conjunto de datos de una muestra. Este cálculo no se puede obtener de forma exacta sino que debe estimarse.

Esto se usa en una variedad de áreas de investigación tales como la física, análisis de imagen, acústica, ceros de la función zeta de Riemann e incluso procesos electroquímicos.

Tales relaciones de escala familiares pueden definirse matemáticamente por la regla de escala general, donde la variable N es el número de particiones, ϵ es el factor de escala, y D es la dimensión fractal:

$$N \propto \epsilon^{-D} \quad (2.5)$$

Para nuestro caso de estudio, al revisar las series de tiempo de las precipitaciones fenómeno climático, con amplia variabilidad y comportamiento aleatorio, de diferentes estaciones pluviométricas, mostraron ser irregulares, y nos llevó a enfrentarnos a una pregunta, ¿Qué tan irregulares son?. Lo que podíamos decir, depende de su resolución. Si utilizamos técnicas para el análisis de series de tiempo, las conclusiones pueden ser diferentes dependiendo del tipo de herramienta. De hecho, cada vez que utilizemos una nueva técnica obtenemos un resultado con más y más detalle. Pero entonces, ¿cuál será la mejor manera de efectuar un análisis fractal de series espacio temporales?.

La pregunta así formulada es incorrecta, pues como se ha dicho, depende de la resolución de medida.

Al aplicar el Análisis Reescalado (R/S) se pudo capturar su dimensión fractal, mediante el exponente de Hurst, mostrando persistencia significativa a largo plazo, el cual es uno de los factores más importantes que caracterizan a las precipitaciones, debido a los errores aleatorios sistemáticos en ellas (Mandelbrot y Wallis, 1969; McGregor y Nieuwolt, 1998) donde se puede considerar la propiedad de memoria a largo plazo donde la dependencia temporal persiste, aún entre observaciones, mostrando que pueden ser caracterizadas a través de la dimensión fractal. De esto, los valores obtenidos pueden ser utilizadas como instrumento de clasificación o agrupación.

En un determinado rango, se obtiene una línea recta como función al representar en una gráfica *log-log*. La pendiente de esa recta es el exponente de Hurst, que puede caracterizar las diferentes cambios variacionales, y la distingue de otras.

Chang[42] extendió la aplicación del exponente de Hurst mediante el desarrollo de un enfoque de cálculo estimándolo sobre series de tiempo comparadas con un movimiento

browniano fraccional de tiempo discreto o ruido fraccional Gaussiano. Yu[43], et al., también estudió correlaciones a largo plazo usando al exponente de Hurst y realizó un análisis fractal múltiple o multifractal de las series de precipitación (véase Kantelhardt[44]) basado en un modelo de cascada multiplicativa y un análisis multifractal de fluctuación de poblamientos.

Otros trabajos recientes sobre análisis de series temporales pueden encontrarse en Carbone[45], et al., con la construcción de un modelo de simulación de tormentas con una distribución exponencial doble. Chou[46] investigó la complejidad a diferentes escalas temporales, precipitación y escorrentía de series de tiempo utilizando el método de muestreo-entropía, y finalmente, García Marén[47], et al., realiza un análisis de frecuencia regionales sobre datos de precipitación en Málaga, España, donde la agrupación de las estaciones en las regiones homogéneas se ha hecho siguiendo un análisis de clúster, con múltiples valores fractales de las diferentes series.

2.5. Medición del exponente de Hurst (H).

El famoso hidrólogo británico Harold Edwin Hurst (1880-1978), trabajó las fluctuaciones de los niveles del Río Nilo, durante largos períodos de tiempo, su interés era proyectar las capacidades de las reservas y tomar medidas de precaución en épocas de sequía. Para esto, ideó una nueva metodología estadística, la cual consiste en saber si las tendencias de la serie de tiempo tienen persistencia o no, luego de medir la duración de ciclos de las series de tiempo y posteriormente determinar si la serie de tiempo es fractal, ó comprobar si esta posee memoria. Esto, con el fin poder proyectar los resultados a futuro.

El método contiene una serie de pasos básicos, que son necesarios para calcular un valor H denominado exponente de Hurst, indispensable para determinación de la persistencia o antipersistencia de una serie de tiempo, además proporciona información sobre la dimensión fractal, dato importante para el desarrollo de nuestra tesis, ya que con el configuraremos regiones con un mismo comportamiento.

El procedimiento para estimar el exponente de Hurst, a partir de una serie temporal de longitud N son, particionar la serie en un conjunto de d -subseries de tiempo más cortas, cada una de longitud m , efectuando lo siguiente pasos para cada subserie desde $n=1, \dots, d$:

- A cada partición de tamaño m se le calcula la media E_n , y la desviación estándar S_n ;
- Normalizar los datos (X_{in}), sustrayendo a cada uno, la media de la subserie;

$$Z_{in} = X_{in} - E, \quad i = 1, 2, \dots, m; \quad (2.6)$$

- Obtener las sumas parciales para cada serie de tiempo.

$$Y_{in} = \sum_{j=1}^i Z_{in}, i = 1, 2, \dots, m; \quad (2.7)$$

- Calcular el rango de cada subserie;

$$R_n = \max_{i=1:m}(Y_{in}) - \min_{i=1:m}(Y_{in}) \quad (2.8)$$

- Se reescala o normaliza el rango calculando;

$$\frac{R_n}{S_n} \quad (2.9)$$

- Una vez calculados los reescalados para cada subserie de longitud m se promedian:

$$\left\langle \frac{R}{S} \right\rangle_m = \frac{1}{d} \sum_{n=1}^d \frac{R_n}{S_n} \quad (2.10)$$

Hurst[12] encuentra que la relación del estadístico (R/S) está dado por la siguiente ley de potencia:

$$\left\langle \frac{R}{S} \right\rangle_m \approx c * m^H \quad (2.11)$$

donde H , es el exponente de Hurst y c es una constante positiva.

Dos factores que intervienen en la determinación del coeficiente de Hurst son: la forma en que la serie temporal es dividida en un conjunto de subseries, donde el rango de los valores de t sobre el cual la pendiente de $\log(\langle R/S \rangle_t)$ y $\log(t)$ es calculado y el segundo factor que interviene en la determinación de H , es el resultado del comportamiento asintótico del rango reescalado, es decir, cuando el valor de t tiende a infinito.

El análisis reescalado $\langle R/S \rangle_t$ sobre varios valores de t , es estimado aplicando \log/\log a la expresión dada, esto es:

$$\log\left(\left\langle \frac{R}{S} \right\rangle_m\right) = \log(c) + H * \log(m) \quad (2.12)$$

Para obtener el coeficiente H , se lleva a cabo un ajuste lineal de los puntos de la relación $\langle R/S \rangle_t$ vs. $\log(t)$ por el método de mínimos cuadrados.

La pendiente de dicha línea es entonces el **coeficiente de Hurst, H** .

Este exponente es un índice fractal que proporciona información sobre una medida de memoria a las correlaciones a largo plazo que se presentan en las series de tiempo.

En la práctica los valores del exponente de Hurst se encuentran entre 0 y 1. Apoyado en dicho valor, ver Mendelbort y Wallis[13], se pueden catalogar las series de tiempo y se puede establecer que las series son:

- Anti-persistentes, lo que significa que para un incremento es más probable que sea seguido por un decremento, y viceversa.
- Aleatoria, corresponde a falta de correlación en la serie (denominado, ruido blanco Gaussiano).
- Persistente, es decir, a un incremento es muy probable que le siga un incremento, y a un decremento es muy probable un decremento

Aunque el parámetro de Hurst esta bien definido matemáticamente, medirlo es problemático. Los datos se medirán a frecuencias de intervalos, grandes/pequeños, donde están contenidas las lecturas menores o mayores (no lluvia ó huracanes, tormentas), los cuales afectan el cálculo de H .

Otros estimadores son parciales y convergen lentamente de acuerdo a la cantidad de datos disponibles, además son vulnerables a las tendencias y periodicidad en los datos, y quizá, a otras fuentes de ruido. Muchos estimadores asumen formas funcionales específicas con modelos subyacentes y se aplican erróneamente al querer simplificar un trabajo.

Este método, en esta investigación, se eligió porque el análisis R/S es una técnica bien conocida, que se ha utilizado durante algún tiempo al efectuar mediciones del parámetro de Hurst.

Otras obras incluyen el exponente de Hurst como los desarrollos de Golder[41], et al., donde utiliza el exponente de Hurst para explorar las correlaciones a largo plazo, y observaciones de precipitación acumulada las modelan usando la ley de probabilidad $alpha - estable$ para hacer frente a distribuciones de colas pesadas.

Estimar el valor de H , puede ser un paso importante en el análisis de series de tiempo real. Nos permite clasificar (por lo menos aproximadamente) la serie según su correlación a largo plazo dependiendo de el número de particiones en la serie,

Mendelbrot[14] da una justificación formal para el uso de esta prueba.

La estimación del exponente de Hurst, puede obtenerse como se muestra en la siguiente subsección.

2.5.1. Mediciones del exponente de Hurst

Hurst simple (H_s): Una serie de tiempo de longitud completa N no se divide en un número de series más cortas, aquí $n = N$. Solo calculamos el rango redimensionado medio. Esto es, tomamos la serie completa y calculamos el coeficiente como una sola partición.

Hurst simple corregido (Hsc): Este se calcula efectuando el análisis anterior restandole el valor esperado del mismo, $(R/S)_{Hs} - E[R/S]$.

Hurst empírico(He): Este es el algoritmo de rango reescalado que se explicó al inicio de esta sección 2.5.

Hurst ANNIS-LLOYD(Hal): Este cálculo se denomina así ya que es adjudicado a A.A. ANNIS y E.H. LLOYD. Dado que el análisis anterior produce estimaciones sesgadas del exponente que rige esa ley de potencia y que para pequeñas particiones existe una desviación significativa, encuentran una mejor estimación para calcular esa ley de potencias y es efectuando el cálculo siguiente: $ERSal = \sqrt{0,5 * \pi * n}$

$$E[R/S] = \begin{cases} \frac{\Gamma(\frac{n-1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}}, & \text{para } n \leq 340 \\ \frac{1}{\sqrt{n\frac{\pi}{2}}} \sum_{i=1}^{n-1} \sqrt{\frac{n-i}{i}}, & \text{para } n > 340 \end{cases} \quad (2.13)$$

dónde Γ es la función gamma de Euler. El coeficiente se obtiene ajustando estos cálculos a una recta de mínimos cuadrados de $(\log n, \log(R/S - E[R/S] - ERSal))$

Hurst teórico(Ht): Este valor es la pendiente de la recta de mínimos cuadrados obtenida del los cálculos \log/\log de los rangos reescalados corregidos y las particiones realizadas.

Calculados estos cinco coeficientes se utilizó el promedio de ellos para efectuar las clasificaciones de los grupos.

2.6. Variograma

El variograma, $\gamma(h)$, es un ajuste o modelado espacial considerado como un estimador de la varianza poblacional y de análisis estructural, donde la población debe tener una tendencia de estacionalidad, se utiliza para describir la relación de observaciones pareadas separadas por una “lag” (retardo o distancia) h y en otros casos con una dirección.

Es una técnica geoestadística, la cual permite una medida cuantitativa de la persistencia a largo plazo en series de tiempo no estacionarias Witt[16], Haslett[17], Dmowska[18] et al. Establece correlaciones a través del tiempo a un fenómeno regionalizado en el espacio, generando patrones que pueden ser utilizados para describir el comportamiento de un conjunto de observaciones.

Matemáticamente, el variograma estima la diferencia cuadrada prevista entre variables aleatorias vecinas, dando un soporte fundamental y permitiendo representar cuantitativamente esta relación. Este proceso continúa para cada punto de medición.

Teniendo en cuenta una serie de tiempo o procesos estocásticos $\{X_t, t \geq 0\}$, la función de autocovarianza en el punto $(t, t+h)$ se define como $C_X(t, t+h) = E[X_t X_{t+h}] - E[X_t]E[X_{t+h}]$ con $E[X_t]$ la media del proceso en tiempo t .

Sin embargo, para una sola serie de tiempo, $\{X_n, n = 1, 2, \dots, n\}$, se espera que el valor puede ser estimado suponiendo una hipótesis de ergodicidad, i.e., es decir, un principio estadístico de equivalencia según la cual “*el promedio a través del tiempo y el promedio a través del ensamble son los mismos*” Lefevbre[19].

Así las diferencias $X_{t+h} - X_t$, que se obtendría con un proceso infinitamente reproducible, son “simulados” ó “clonados” de la “serie madre”.

Por lo tanto, el valor medio de las diferencias $X_{t+h} - X_t$ es estimado por:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{t=1}^{n(h)} (x_{t+h} - x_t)^2 \quad (2.14)$$

Una explicación mas a detalle se da en el Anexo 6.7. Este estimador de momentos, es un promedio de diferencias al cuadrado, que puede ser influenciado por un número pequeño de valores que ocasionan discrepancias al final de los cálculos debido a las particiones realizadas. Pero, se considera un estimador robusto, ya que disminuye la importancia de las diferencias grandes al cuadrado.

También se considera robusto en el sentido de que es resistente a distribuciones normales contaminadas y afloramientos posiblemente generados por distribuciones de colas pesadas. Esto puede verse por el uso de la raíz cuadrada de las diferencias, en lugar de diferencias de cuadrados, en el estimador insesgado.

La aplicación de variogramas a nuestros datos mostró periodos largos de comportamiento similar, con diferente duración, un comportamiento o patrón cíclico, pero no periódico.

2.7. Identificación de regiones homogéneas

La identificación de regiones homogéneas es a menudo la etapa más complicada, al requerir de la toma de decisiones subjetivas.

En zonas montañosas, las características fisiográficas influyen en la distribución espacial irregular de las variables climáticas. Por lo que en estos sitios la división en regiones se debe realizar teniendo en cuenta tanto aspectos hidroclimáticos como fisiográficos, sin tomar en cuenta la continuidad geográfica de la cuenca. Nathan et al. (1990), mencionaron que regiones homogéneas definidas por la similitud hidrológica de las cuencas o las características de éstas, pueden no tener significancia geográfica.

La Cuenca del río San Juan, es una región de relieve montañoso donde actualmente existen mas de 80 estaciones hidrométricas en operación. El objetivo de este trabajo fue identificar zonas homogéneas dentro de la Cuenca para obtener una clasificación de estaciones con comportamiento similar.

Esta clasificación se puede llevar a cabo utilizando distintos tipos de datos y empleando métodos univariantes o multivariantes, lo cual varía en función del tipo de zonificación que se quiere encontrar en relación a una o más variables de interés.

La evaluación comparativa basada en la similitud estructural permite identificar regiones a través de comparaciones sistemáticas con otros y así marcar una diferencia. Esta metodología de identificación de regiones homogéneas, permite la evaluación comparativa y puede ser de gran ayuda en la toma de decisiones estratégicas.

En este trabajo se presenta un método para la detección de zonas homogéneas a través de un análisis fractal sobre los datos que caracterizan las series de tiempo de las 31 estaciones pluviométricas de la Cuenca del río San Juan.

El objetivo en esta investigación es formar conglomerados de estaciones que satisfagan aproximadamente la condición de homogeneidad, esto es, que podamos decir de alguna manera que poseen la misma distribución de frecuencias, excepto por un factor de escala.

2.8. Validación

El objetivo del análisis estadístico es asegurar que el problema se aborda en forma adecuada, que el número de condiciones y casos de la metodología que se examina sea suficiente para obtener inferencias estadísticas válidas a partir de los resultados.

En la validación se utilizan diversas pruebas y procedimientos estadísticos. Si un modelo determinado no simula en forma adecuada la respuesta del sistema real, entonces resulta necesario volver a examinar las dos primeras etapas (identificación del problema y planteamiento del modelo) con el objeto de identificar los factores o relaciones que no se hayan considerado.

En este caso, para revisar si la metodología que se propone es adecuada, se llevan a cabo simulaciones de datos aleatorios con diferentes distribuciones de probabilidad: Distribución Normal, $N \sim (0, \sigma^2)$, distribución General de valores extremos $Gev \sim (\mu, \sigma, \kappa)$, ($\mu = \text{ubicacion}, \sigma = \text{escala}, \kappa = \text{forma}$) distribución Gamma $G \sim (\alpha, \beta)$ ($\alpha = \text{forma}, \beta = \text{escala}$).

La información que aportan las funciones de autocorrelación y autocorrelación parcial, resulta de fundamental importancia para decidir el tipo de proceso generador del conjunto de datos que conforman la serie de tiempo. Para esto utilizamos los modelos Autorregresivos y de Promedios Móviles (ARMA por sus siglas en inglés Autorregresive and Movable Average).

La idea de estos modelos es que los valores actuales de la serie X_t dependen de los p valores previos: X_{t-1}, \dots, X_{t-p} .

Definición: Un modelo autoregresivo de orden p , denotado por AR(p), es de la forma

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \xi \tag{2.15}$$

donde X_t es estacionario, $\phi_1, \phi_2, \dots, \phi_p$ son constantes ($\phi_p \neq 0$) y ξ es un ruido.

En las simulaciones que se efectuaron, el ruido se consideró como: blanco con media 0 y varianza σ_ξ^2 , Gamma y General de valores extremos.

Se generó 25 bases con 336 datos cada una, los cuales se dispusieron como muestra el siguiente cuadro 2.1:

CUADRO 2.1: Tabla de datos mensuales correspondiente a n años que serán simuladas

Año/mes	1	2	...	12
1	y_1	y_2	...	y_{12}
2	y_{13}	y_{14}	...	y_{24}
3	y_{25}	y_{26}	...	y_{36}
\vdots	\vdots	\vdots	\vdots	\vdots
n	$y_{1+12(n-1)}$	$y_{2+12(n-1)}$...	$y_{12+12(n-1)}$

Cada fila es una realización de una serie temporal. Supóngase que cada una de estas n series es generada por un modelo ARMA(P), o más específicamente, la serie correspondiente al j -ésimo mes, y_{j+12t} , con $t = 0, 1, \dots, (n-1)$, satisface la ecuación en diferencias:

$$y_t = \Phi_1 y_{t-12} + \xi \quad (2.16)$$

denominado modelo $MA(1)_{12}$ donde ξ será un ruido: Gaussiano, Gamma ó General de valores extremos en nuestras simulaciones.

Se simuló 5 series de datos para cada valor de $\Phi = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Esto es 5 grupos con $\Phi = 0.1$, 5 grupos con $\Phi = 0.3, \dots$, 5 grupos con $\Phi = 0.9$ con distribución, Normal, Gamma y Gev. a los cuales se les llamará grupos T .

Es difícil definir cuando el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado dependiendo del tipo de grupos que se busquen y de las características de datos inherentes. La técnica que utilizaremos es la de validación externa Pairwise Measures (Medidas en parejas) descrita en el libro DATA MINING AND ANALYSIS de Zaki and Meira (capítulo 17), También utilizaremos el coeficiente de Jaccard (Paul Jaccard 1868-1944) que mide el grado de similitud entre dos conjuntos, sea cual sea el tipo de elementos.

Dados los agrupamientos C y las estimadas por la aproximación de Hurst T , las medidas pareadas utilizan la información de etiqueta de partición y cluster sobre todos los pares de puntos.

Si $D = \{x_i\}_{i=1}^n$ es nuestro conjunto de datos, repartidos en 5 grupos. $y_i \in 1, 2, \dots, k$ denota la información básica del agrupamiento verdadero o etiqueta para cada punto. $\hat{y}_i \in 1, 2, \dots, r$ denota la etiqueta del agrupamiento para cada x_i estimada por algún algoritmo, que en nuestro caso es la clasificación de el exponente de Hurst, tanto para los datos obtenidos así como los variogramas de los mismos.

Generalmente existe una relación inversa entre estos dos objetivos, que puede ser capturado por una medida de forma explícita o está implícito en su calculo. Todas las medidas externas dependen de una tabla de contingencia N de $r \times k$ en la cual se introduce los agrupamientos base C y los obtenidos por algún algoritmo T , definida como sigue:

$$N(i, j) = n_{ij} = |C_i \cap T_j| \quad (2.17)$$

En otras palabras, el n_{ij} denota el número de puntos que son comunes a C_i y la partición T_j , se examinó las tablas de contingencia, obteniendo la pureza de los agrupamientos y capturando el coeficiente de Jaccard para medir el grado de similitud entre los dos conjuntos como se describe en el libro de Zaki and Meira.

La pureza cuantifica el grado en que un agrupamiento C_i contiene entidades de solamente una partición. En otras palabras, mide que tan puro es cada grupo. La pureza del cluster C_i se define como:

$$pureza_i = \frac{1}{n_i} \max_{j=1}^k \{n_{ij}\} \quad (2.18)$$

La pureza del agrupamiento C se define como la suma ponderada de los valores de pureza

$$pureza = \sum_{i=1}^r \frac{n_i}{n} pureza_i = \frac{1}{n} \sum_{i=1}^r \max_{j=1}^k \{n_{ij}\} \quad (2.19)$$

donde el cociente n_i/n denota la razón de puntos en el agrupamiento C_i .

Dado agrupamiento C y T , las medidas por pares acertadas utilizan la información de las etiquetas de partición y del cluster sobre todos los pares de puntos.

Sean $x_i, x_j \in D$ dos puntos cualquiera $i \neq j$. y_i describe la etiqueta de la verdadera partición \hat{y}_i denota la etiqueta del agrupamiento de. Si x_i y x_j pertenecen al mismo grupo, es decir, $\hat{y}_i = \hat{y}_j$, lo llamamos un evento positivo, y si no pertenecen al mismo grupo, es decir, $\hat{y}_i \neq \hat{y}_j$, lo llamamos un acontecimiento negativo. Dependiendo de si concuerdan entre el grupo de etiquetas C_i y el grupo de etiquetas de partición T_i , hay cuatro posibilidades a considerar:

Verdaderos Positivos (VP): x_i y x_j pertenecen a la misma partición T y están en el mismo agrupamiento C . Esto es un par verdadero positivo porque el evento positivo $\hat{y}_i = \hat{y}_j$, corresponde a la partición verdadera, $y_i = y_j$. El número de pares positivos verdaderos esta dado por:

$$VP = |(x_i, x_j) : y_i = y_j \wedge \hat{y}_i = \hat{y}_j| \quad (2.20)$$

Falsos Negativos (FN): x_i y x_j pertenecen a la misma partición T Pero no pertenecen a el mismo grupo C . Esto es, el evento es negativo, $\hat{y}_i \neq \hat{y}_j$ no corresponde a la partición verdadera, $y_i = y_j$. Este par es, por tanto, un falso negativo, y el número de todos los pares falsos negativos esta dado por:

$$FN = |(x_i, x_j) : y_i = y_j \wedge \hat{y}_i \neq \hat{y}_j| \quad (2.21)$$

Falso Positivos (FP): x_i y x_j no pertenecen a la misma partición T y pero si pertenecen a el mismo grupo C . Este par es un falso positivo porque el evento es positivo, $\hat{y}_i = \hat{y}_j$. Es en realidad falsa, es decir, no concuerda con la partición verdadera-real, lo cual indica que, $y_i \neq y_j$. El número de pares falsos positivos se da como:

$$FP = |(x_i, x_j) : y_i \neq y_j \wedge \hat{y}_i = \hat{y}_j| \quad (2.22)$$

Verdaderos Negativos (VN): x_i y x_j no pertenecen a la misma partición T ni pertenecen al mismo grupo C . Este par es así un verdadero negativo, esto es, $\hat{y}_i \neq \hat{y}_j$, corresponde a la partición verdadera, $y_i = y_j$. El número de pares positivos verdaderos esta dado por:

$$VN = |(x_i, x_j) : y_i \neq y_j \wedge \hat{y}_i \neq \hat{y}_j| \quad (2.23)$$

La cantidad de pares que existen son, $\binom{n}{2} = \frac{n(n-1)}{2}$ con la siguiente identidad:

$$N = VP + FN + FP + VN \quad (2.24)$$

Haciendo uso de la ecuación 2.24 podemos obtener el coeficiente de Jaccard, el cual mide el grado de similitud entre dos conjuntos, sea cual sea el tipo de elementos. Este coeficiente esta definido como sigue:

$$Jaccard = \frac{VP}{VP + FN + FP} \quad (2.25)$$

El coeficiente de Jaccard mide la fracción de pares de puntos positivos verdaderos, pero después de ignorar verdaderos negativos(VN). Siempre toma valores entre 0 y 1, correspondiente este último a la igualdad total entre ambos conjuntos, esto es, un perfecto agrupamiento de C (es decir, totalmente de acuerdo con el T particionado). El coeficiente de Jaccard es asimétrico en cuanto a los verdaderos positivos y negativos porque ignora los verdaderos negativos. Enfatiza la similitud en términos de los pares de puntos que pertenecen juntos tanto en el agrupamiento particionado como en el agrupamiento verdadero, pero descuentan las parejas de puntos que no pertenecen juntas.

Para examinar las posibilidades de agrupamiento existen métodos que permiten dar solución a esto. Una solución se encuentra en los llamados métodos jerárquicos y no jerárquicos.

Métodos jerárquicos aglomerativos: se comienza con los objetos o individuos de modo individual; de este modo, se tienen tantos clusters iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único cluster. **Métodos jerárquicos divididos:** se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén.

Métodos no jerárquicos Se usan para agrupar objetos, pero no variables, en un conjunto de k clusters ya predeterminado. No se tiene que especificar una matriz de distancias ni se tienen que almacenar las iteraciones. Todo esto permite trabajar con un número de datos mayor que en el caso de los métodos jerárquicos. Se parte de un conjunto inicial de clusters elegidos al azar, que son los representantes de todos ellos; luego se van cambiando de modo iterativo. Se usa habitualmente el método de las $k - medias$.

De esta manera, se obtendrá el exponente de Hurst, tanto a los simulados como a sus variogramas, se les clasificará según el orden y el método no jerárquicos de $k - medias$, con el que evaluaremos la semejanza de los conglomerados simulados y los conglomerados generados con el coeficiente de Hurst a estas particiones se les denominará como C .

$k - medias$ es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es uno de los métodos más utilizados en minería de datos.

Para revisar los resultados de las simulaciones, se realizará una prueba de diferencia de medias y un análisis de varianza utilizando un nivel de significancia del 0.05 mediante los cálculos: $\bar{X}_2 = \bar{X}_H =$ Promedios de Hurst, $\bar{X}_1 = \bar{X}_{HK} =$ Promedios de Hurst k -medias, $\bar{X}_2 = \bar{X}_{HV} =$ Promedios de Hurst variograma, $\bar{X}_1 = \bar{X}_{HVK} =$ Promedios de Hurst variograma k -medias. $s_2 = s_H =$ desviación de Hurst, $s_1 = s_{HV} =$ desviación de Hurst variograma. $s_{HK} =$ desviación de Hurst k -medias, $s_{HVK} =$ desviación de Hurst variograma k -medias, mediante las siguientes ecuaciones:

$$t_c = \frac{\bar{X}_2 - \bar{X}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad s_p = \sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{(n_1 + n_2 - 2)}} \quad (2.26)$$

$$F = \frac{s_2^2}{s_1^2} \quad (2.27)$$

Capítulo 3

Resultados

3.1. Discusión

El exponente de Hurst, como parte de un análisis fractal, se utilizó para evaluar la dependencia de largo alcance y la posibilidad de tendencias en los datos.

En este trabajo, se llevó a cabo un enfoque de agrupamiento que se utilizó para concentrar las estaciones en muestras homogéneas, utilizando el exponente de Hurst tanto a los datos directamente como a los variogramas.

Como caso de estudio, se tomó una muestra de 33 estaciones pluviométricas, de la cuenca del río San Juan, en la región RH-24 de México (Fig.3.1).

Un mapa de la cuenca del río San Juan se muestra en la figura 3.2. Esta región se encuentra en México entre los Estados de Nuevo León Coahuila y Tamaulipas con una superficie aproximada de $32,972 \text{ km}^2$. Algunas de las estaciones de lluvias se muestran en la figura 3.3. Los datos utilizados han sido proporcionados por la CONAGUA, la institución local responsable de la gestión del agua en el país.



FIGURA 3.1: La cuenca del Río San Juan obtenida de Google Earth. Se muestra en negro, la cuenca del río San Juan.

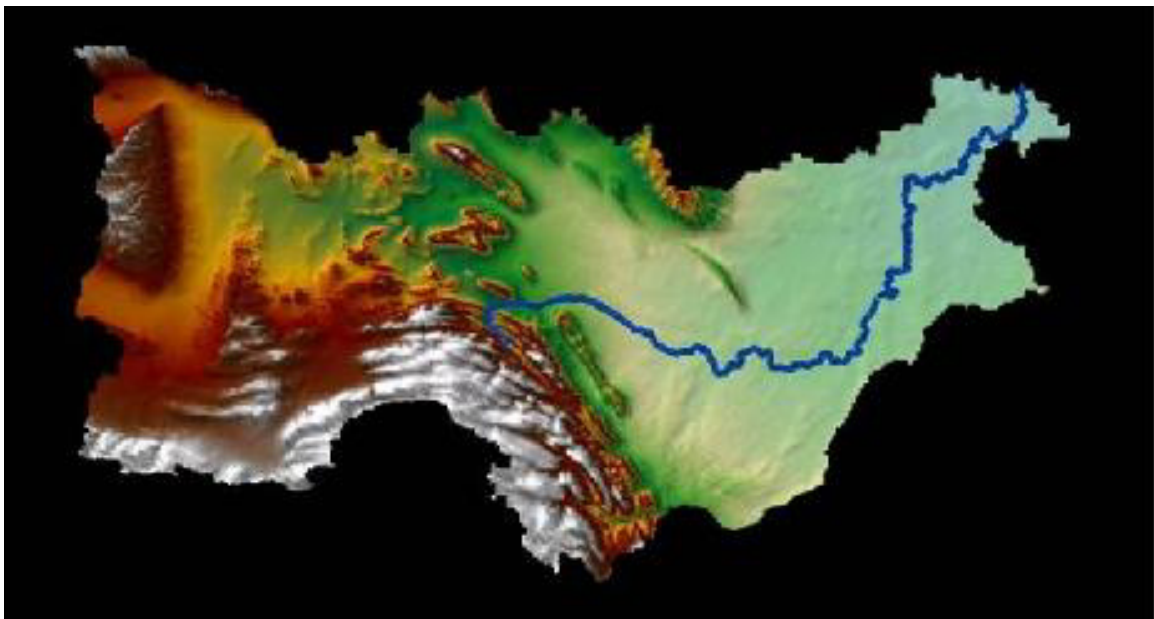


FIGURA 3.2: Cuenca del río San Juan (río San Juan en azul). Imagen obtenida de: “Administración del agua en la cuenca del río San Juan, en el sur del río Bravo Región Hidrológica de México” de <http://earthzine.org/2012/08/13/>

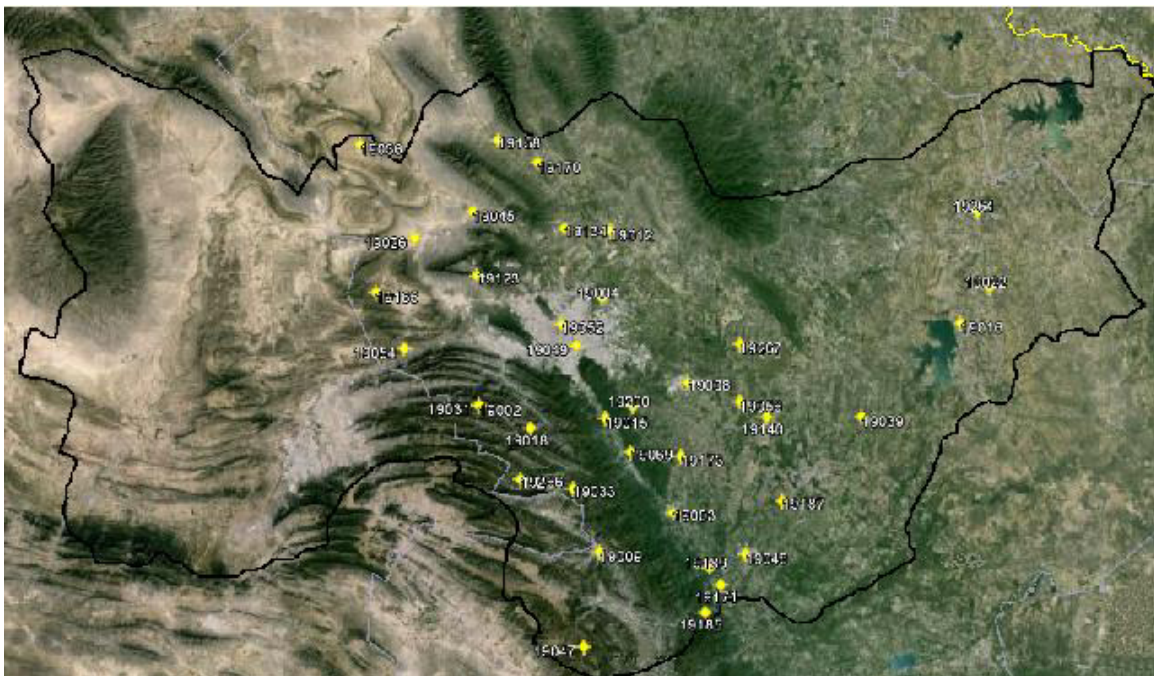


FIGURA 3.3: Localización Geográfica (de Google Earth) de las estaciones pluviométricas de la cuenca del río San Juan. Obtenida de la base de datos: www.conagua.gob.mx

3.2. Descripción del problema

En la práctica, para realizar un análisis regional hidrológico, se requiere identificar regiones homogéneas donde los datos siguen patrones similares que puede analizarse conjuntamente para mejorar la identificación de modelos de probabilidad que a su vez puede utilizarse para estimar eventos extremos y su frecuencia en términos de periodos de retorno. Este análisis generalmente se ejecuta cuando se trata con sequías, contaminación, movimiento del viento, temperatura, presión atmosférica y observaciones de precipitación, para nombrar algunos.

Esta investigación se ocupa del problema de encontrar grupos de estaciones pluviométricas creando conglomerados homogéneas, considerando el exponente de Hurst aplicado directamente a los datos y también a los variogramas. Los datos de la precipitación de una muestra de 33 estaciones de la región hidrográfica México RH-24, cuenca del río San Juan, fueron utilizados como estudio de caso para evaluar la propuesta.

El variograma, se utilizó como cuantificador de las correlaciones de largo alcance para los datos de las diferentes estaciones pluviométricas con registros mensuales. Teniendo en cuenta el análisis de la muestra de la serie histórica, se realizó un análisis de rango reescalado o análisis R/S , para obtener una medida del exponente de comportamiento fractal en nuestro caso el exponente de Hurst[48].

Este exponente se utilizó como una medida para cada una de las estaciones pluviométricas en particular. El proceso se repitió para cada estación pluviométrica en la región bajo análisis. El exponentes de Hurst se utilizó como referencia para identificar las estaciones que presentan patrones similares.

Una ventaja del exponente de Hurst es la simplicidad de su algoritmo que puede ser utilizado para medir la condición de persistencia o antipersistencia de un proceso, y proporciona una métrica que puede utilizarse para clasificar series de tiempo diferentes.

El variograma $\gamma(h)$ se utilizó para describir la relación de las observaciones pareadas y separadas por una distancia h . Es una técnica geoestadística que permite una medida cuantitativa de la persistencia a largo plazo en series de tiempo no estacionarias Witt[16], Haslett[17], Dmowska[18], et al.

En las correlaciones en el tiempo y el espacio, se crean patrones que pueden utilizarse para describir el comportamiento de un conjunto de observaciones, y el variograma estima la diferencia al cuadrado prevista entre variables aleatorias vecinas. Este cálculo se realiza sobre los valores de h diferentes.

Teniendo en cuenta una serie de tiempo o proceso estocástico $\{X_t, t \geq 0\}$, la función de autocovarianza en el punto $(t, t+h)$ se define como $C_X(t, t+h) = E[X_t X_{t+h}] - E[X_t]E[X_{t+h}]$ con $E[X_t]$ la media del proceso en tiempo t . El variograma $\gamma(h)$ está dada por la mitad de la varianza de la diferencia entre pares de observaciones en diferentes "localizaciones" en el tiempo.

El coeficiente de Hurst calculado con el método del rango reescalado ó también llamado crecimiento del rango, mide el crecimiento de las fluctuaciones de las estaciones, al aumentar el intervalo de tiempo Δt . Esto significa que cuanto mayor es el exponente de Hurst de una población más rápido aumenta el rango de las fluctuaciones. Si no se tiene en cuenta el valor de la constante c de proporcionalidad en la fórmula fractal para un mismo tamaño poblacional, los valores más grandes del coeficiente de Hurst se podrían asociar a un mayor aumento en los valores o disminución de los mismos (Sugihara y May).

Para estimar el exponente o coeficiente de Hurst de una serie temporal $\{X_k\}$, con $k \in 1, 2, \dots, N$ la serie es dividida en un grupo de d -subseries de longitud m . Realmente, la medida de m es un valor promedio.

Una forma estándar, aunque no la única, de obtener el tamaño m de la subserie, es particionar la serie original en potencias de base 2. De esta manera, en cada una de las particiones sucesivas, el valor aproximado de m es: $N, N/2, N/2^2, N/2^3, \dots$, y así sucesivamente. Para cada subserie $n = 1, 2, \dots, d$, se efectúan los pasos de la sección 2.6.

Dos factores que intervienen en la determinación del coeficiente de Hurst son: la manera en que es dividida la serie de tiempo en grupos de subseries y el comportamiento asintótico del análisis de rango reescalado.

El primero, el rango de valores m se utilizan para calcular la pendiente de $\log(\langle R/S \rangle_m)$, dada la relación

$$\log\left(\left\langle \frac{R}{S} \right\rangle_m\right) = \log(c) + H \log(m) \quad (3.1)$$

El segundo, la determinación de H es el resultado del comportamiento asintótico del rango reescalado, i.e., cuando el valor de m tiende a infinito.

El análisis de reescaldado $\langle R/S \rangle_m$ sobre algunos valores de m es estimado usando log/log expresión dada en (3.1). Para obtener el coeficiente H , se utiliza el método de mínimos cuadrados. La pendiente es el coeficiente de Hurst, H .

Este exponente es considerado un índice fractal, Mandelbrot and Wallis[13], y proporciona información acerca de correlaciones a largo plazo de una serie de observaciones; para una revisión teórica del exponente de Hurst véase Mandelbrot[14].

En la práctica, el exponente de Hurst puede tomar valores entre 0 y 1, donde:

- $0 < H < 0,5$ indica no persistencia en una serie, i.e., a un incremento es más probable que sea seguida por un decremento y viceversa.
- $H = 0.5$ indica ausencia de correlación serial (ruido blanco Gaussiano).
- $0,5 < H < 1$ indica persistencia, es decir, un incremento corre el riesgo de ser seguido por un incremento y un decremento por otro decremento.

3.3. Resultados

En esta sección se muestran los resultados obtenidos en el desarrollo de esta investigación.

Se revisaron 33 estaciones, el cuadro 3.1 muestra la información general de las estaciones que fueron analizadas.

Para ilustrar el procedimiento, se muestra el análisis de tres de las 33 estaciones de lluvia, el resto de ellas se encuentran en el Anexo 6.4.

Los valores medidos de precipitación mensual en milímetros de las estaciones de Apodaca, El Cuchillo y La Boca, se muestran en la Fig. 3.4, y el total se presentan en el Anexo 6.4.

Como puede verse, los comportamientos y/o relaciones entre diferentes estaciones son difíciles de evaluar utilizando el análisis “visual” de las series de tiempo..

Inicialmente se aplicó el método de rango reescalado para encontrar el exponente de Hurst a las bases de datos originales, obteniendo el denominado exponente de Hurst empírico como se muestra en el cuadro 3.2

La Fig.3.5, muestra un gráfico *log/log* del tamaño de particiones vs. rangos reescalados. Las pendientes de las ecuaciones lineales, obtenidas por el método de mínimos cuadrados, representa el exponente de Hurst. Estos valores son mayores a 0.5, lo cual indica que en estas series de datos, existe una persistencia en el comportamiento de las lluvias según lo mencionado anteriormente en la sección 3.2, el resto de los valores para las 33 estaciones se encuentran en el cuadro 3.2 donde también puede observarse que la mayoría de los valores de Hurst son mayores a 0.5, lo que indica una persistencia en las series de tiempo.

Para un mejor análisis se calcularon diferentes coeficiente de Hurst: Coeficientes de Hurst Simple (**Hs**), Hurst simple corregido (**Hsc**), Hurst empírico (**He**), Hurstal (**Hal**), Hurst Teórico (**Ht**) y Hurst Promedio (**HP**) como muestra el cuadro 3.3 en el que el tiempo de dependencia en todos ellos se hace evidente. Las gráficas donde se muestra el exponente de Hurst obtenido por un ajuste de mínimos cuadrados lineal se muestra en el Anexo 6.6.

CUADRO 3.1: Información general de las estaciones pluviométricas de la cuenca río San Juan.

Estación	Nombre	Latitud	Longitud	Datos
19015	El Cerrito	25 30 36	100 11 36	1984-2012
19039	Las Enramadas	25 30 05	099 31 17	1984-2012
19018	El Pajonal	25 29 23	100 23 20	1984-2012
19012	Ciénega de Flores	25 57 08	100 10 20	1984-2012
19003	Allende	25 17 01	100 01 13	1984-2012
19069	La Boca	25 25 46	100 07 44	1984-2012
19009	Casillas	25 11 47	100 12 51	1984-2012
19187	California	25 18 23	099 44 02	1984-2012
19173	Palmitos	25 25 02	099 59 50	1984-2012
19002	Agua Blanca	25 32 39	100 31 23	1984-2012
19134	Salinas Victoria	25 57 33	100 17 34	1984-2012
19031	La Cruz	25 32 47	100 31 23	1984-2012
19036	La Popa	26 09 50	100 49 40	1984-2012
19185	El Canadá	25 02 48	099 56 29	1984-2012
19056	San Juan	25 32 36	099 50 25	1944-2012
19016	El Cuchillo	25 43 05	099 15 21	1960-2012
19052	Monterrey(Obs)	25 44 01	100 16 01	1984-2012
19026	Icamole	25 56 28	100 41 13	1984-2012
19200	La Cienega	25 32 10	100 07 15	1984-2012
19004	Apodaca	25 47 37	100 11 50	1984-2012
19047	Mimbres	24 58 26	100 15 31	1984-2012
19140	Tepehuaje	25 30 19	099 46 15	1984-2012
19048	Montemorelos	25 10 55	099 49 56	1984-2012
19022	General Bravo	25 48 05	099 10 32	1984-2012
19158	Rancho de Gomas	26 10 11	100 27 52	1984-2012
19045	Mina	26 00 08	100 32 00	1984-2012
19054	Rinconada	25 40 52	100 43 03	1984-2012
19165	Chupaderos del Indio	25 48 49	100 47 24	1984-2012
19171	Lampacitos	25 06 38	099 53 57	1984-2012
19264	Dr. Coss	25 51 16	099 56 36	1984-2012
19170	El Hojase	26 06 55	100 21 38	1984-2012
19033	Laguna de Sanchez	26 06 55	100 21 38	1984-2012
19123	Grutas de García	26 06 55	100 21 38	1984-2012

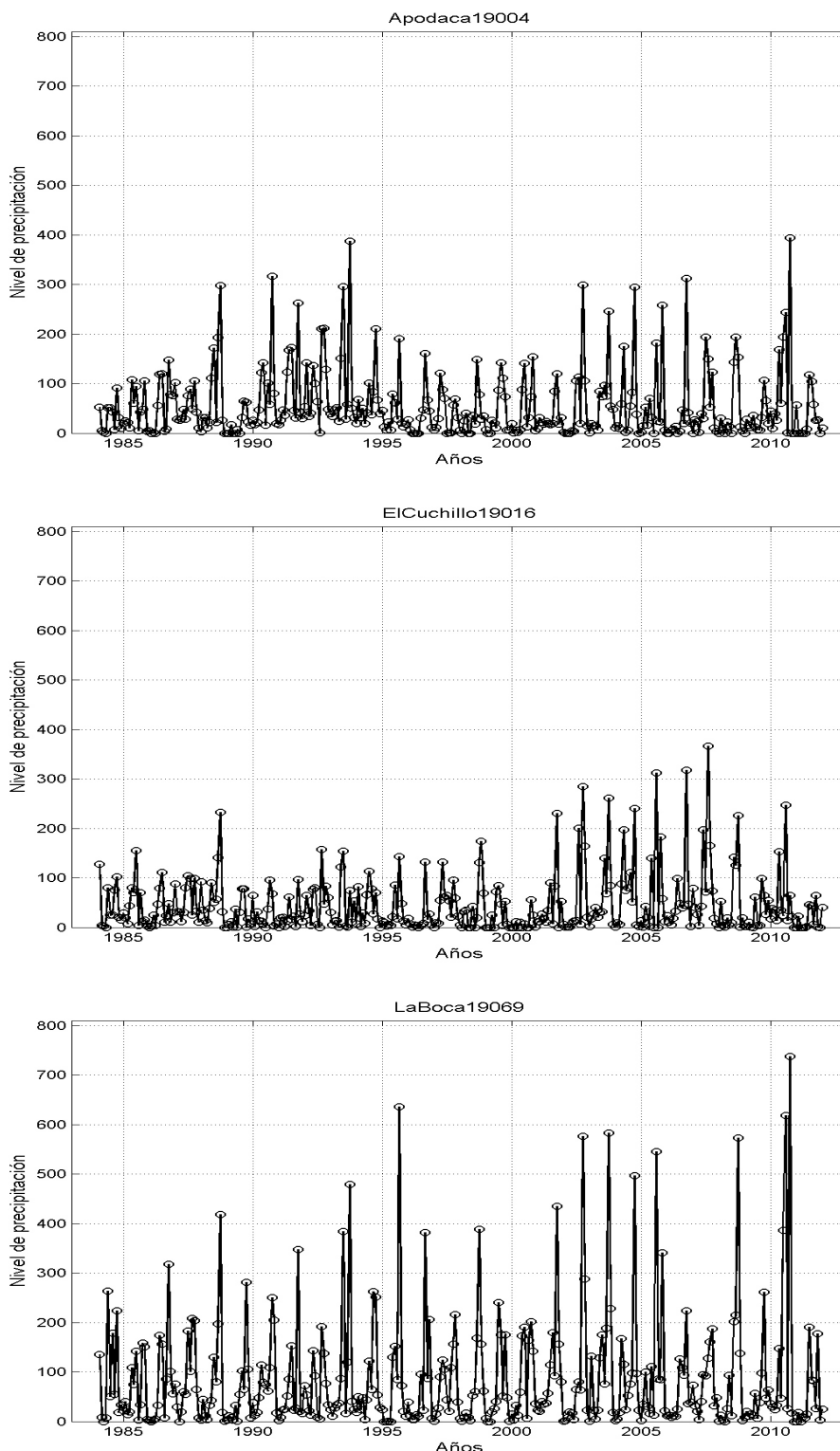


FIGURA 3.4: Mediciones de precipitación de tres estaciones de La cuenca del río San Juan, México. Desde la parte superior a la parte inferior, respectivamente: Apodaca, Estación número 19004, Enero 1940-Diciembre 2012; El Cuchillo, Estación número 19016, Enero 1939-Diciembre 2012 ; La Boca, Estación número 19069, Enero 1923-Diciembre 2012.

CUADRO 3.2: Coeficientes de Hurst empíricos (He) de cada estaciones pluviométricas de la cuenca río San Juan. Coeficiente de Hurst empico aplicado a variograma.

Estación	He	Hurst Variograma
19015	0.4368	0.5629
19009	0.4908	0.6642
19069	0.5232	0.5980
19054	0.5516	0.8831
19018	0.5718	0.7548
19002	0.5732	0.7621
19264	0.5760	0.7256
19047	0.5827	0.6010
19045	0.5849	0.8248
19012	0.5899	0.6692
19171	0.6048	0.8039
19187	0.6059	0.7454
19031	0.6409	0.6492
19134	0.6147	0.6252
19165	0.6147	0.8660
19048	0.6163	0.7882
19185	0.6167	0.7777
19036	0.6202	0.7869
19158	0.6218	0.7484
19003	0.6225	0.7047
19200	0.6229	0.7945
19056	0.6255	0.7746
19039	0.6258	0.7892
19052	0.6373	0.7514
19004	0.6501	0.7742
19140	0.6565	0.8221
19033	0.6505	0.8521
19022	0.6559	0.7937
19170	0.6712	0.9479
19173	0.6741	0.8468
19016	0.7124	0.7730
19123	0.7567	0.7871
19026	0.8413	0.9806

Para abordar el problema de encontrar muestras homogéneas, un análisis de conglomerados se llevó a cabo, utilizando las estimaciones del exponente de Hurst formando cinco conglomerados, como muestran las divisiones del cuadro 3.4 y el histograma 3.6.

Después de identificar un conjunto de distribuciones posibles con p -valores más grandes que un nivel significativo de 0.05, en todos los casos, se seleccionó la distribución con el más alto promedio p -valor para cada estación. Gamma y General de Valores Extremos eran las distribuciones que dieron el mejor ajuste en los datos analizados.

Los parámetros de ubicación, dispersión y curtosis, fueron calculados con los L-momentos. El cuadro 3.4 muestra en la cuarta columna el tipo de distribución ajustada y la quinta columna muestra los parámetros de dicha distribución. En el Anexo 6.9 se da una breve explicación del cálculo de L-Momentos. En el Anexo 6.10 se muestran los gráficos de las distribuciones con mejor ajuste obtenidas mediante librerías de Matlab.

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-\frac{x}{\beta}}, \quad (3.2)$$

y

$$f(x; k, \sigma, \mu) = \frac{1}{\sigma} e^{-(1+kz)^{-1/k}} (1+kz)^{-1-1/k}, \quad z = \frac{x-\mu}{\sigma} \quad (3.3)$$

respectivamente.

Sin embargo, cuando se obtiene el variograma de cada una de estas estaciones, como se muestra en la Fig. 3.7, una comportamiento se hace evidente, cada círculo muestra la relación a meses cercanos, esto es, el primer círculo muestra las variaciones con espacios de un mes: enero-febrero, febrero-marzo,..., noviembre-diciembre. El segundo círculo muestra las variaciones cada dos meses: enero-marzo, febrero-abril,..., octubre-diciembre, el tercer círculo muestra las variaciones de enero-abril, febrero-mayo,...,septiembre-diciembre y así sucesivamente, en esta inspección más cercana se muestra un comportamiento estacional que se repite cada 12 observaciones en el variograma como puede verse en la Fig. 3.8. Esto puede explicarse debido al hecho de que se utilizaron las observaciones mensuales en el análisis. En el Anexo 6.7 se muestran los variogramas de las 33 estaciones.

La figura 3.9, muestra las ecuaciones lineales, calculadas por mínimos cuadrados, en donde la pendiente de ellas es el exponente de Hurst aplicado a variogramas. El exponente de Hurst a variogramas, de todas las estaciones se encuentra en el cuadro 3.5. Se puede observar que los coeficientes de Hurst son mayores a 0.5 y algunos están cercanos a 1, lo que indica una dependencia larga positiva de los datos en los variogramas.

Aprovechando esta información también se efectuó particiones para clasificar en función de los variogramas como muestra las divisiones en el cuadro 3.5 y el histograma de la figura 3.10.

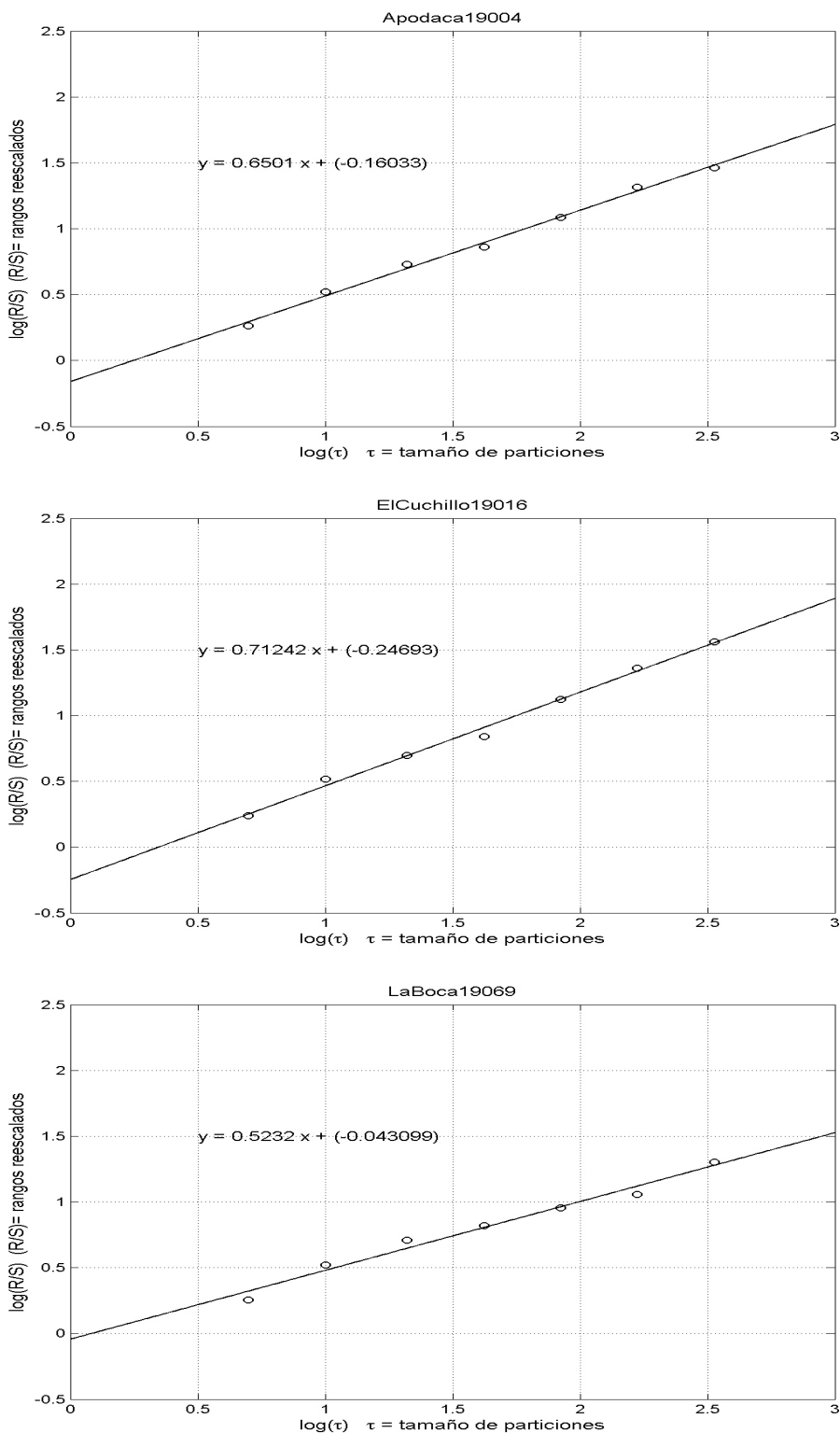


FIGURA 3.5: Exponente de Hurst promedio (H_p) a datos originales por mínimos cuadrados de las tres estaciones de la cuenca del río San Juan, México. Desde la parte superior a la parte inferior, respectivamente: Apodaca, Estación número 19004, $m = 0,6501$; El Cuchillo, Estación número 19016, $m = 0,71242$; La Boca, Estación número 19069, $m = 0,5232$.

CUADRO 3.3: Coeficientes de Hurst Simple(He), Hurst simple corregido (Hrs), Hurst empírico (He), Hurstal(Hal), Hurst Teórico (Ht) y Hurst Promedio (HP). Mediante (6.9) de estaciones pluviométricas de la cuenca río San Juan.

Estación	Hs	Hrs	He	Hal	Ht	HP
19015	0.4405	0.3273	0.3978	0.3535	0.5503	0.4139
19009	0.4705	0.4090	0.3893	0.3520	0.5503	0.4342
19069	0.5165	0.4686	0.3771	0.3356	0.5503	0.4496
19047	0.5427	0.5052	0.3763	0.3448	0.5503	0.4639
19002	0.5450	0.5186	0.4586	0.4177	0.5503	0.4980
19264	0.5404	0.5458	0.5064	0.4637	0.5503	0.5213
19036	0.5928	0.6021	0.4643	0.4250	0.5503	0.5269
19018	0.5370	0.5358	0.5383	0.4893	0.5503	0.5301
19056	0.5772	0.6174	0.4928	0.4509	0.5503	0.5377
19012	0.5652	0.5722	0.5432	0.4905	0.5503	0.5443
19054	0.5025	0.5011	0.6141	0.5536	0.5503	0.5443
19045	0.5257	0.5295	0.5997	0.5497	0.5503	0.5510
19187	0.5658	0.6169	0.5583	0.5054	0.5503	0.5594
19033	0.6106	0.6651	0.5203	0.4781	0.5503	0.5649
19200	0.5645	0.6194	0.5811	0.5307	0.5503	0.5692
19052	0.5926	0.6379	0.5804	0.5241	0.5503	0.5770
19134	0.5747	0.5942	0.6159	0.5632	0.5503	0.5797
19158	0.5879	0.6179	0.5993	0.5437	0.5503	0.5798
19031	0.5584	0.5553	0.6499	0.5974	0.5503	0.5822
19039	0.5746	0.6263	0.6133	0.5606	0.5503	0.5850
19048	0.5542	0.6294	0.6558	0.5966	0.5494	0.5971
19171	0.5406	0.5688	0.7148	0.6477	0.5503	0.6044
19140	0.5957	0.6421	0.6441	0.5918	0.5503	0.6048
19022	0.5949	0.6508	0.6435	0.5926	0.5503	0.6064
19185	0.5727	0.6116	0.6817	0.6171	0.5503	0.6067
19003	0.5619	0.6392	0.7005	0.6383	0.5503	0.6180
19170	0.6021	0.6443	0.6881	0.6337	0.5503	0.6237
19173	0.6125	0.6988	0.6839	0.6263	0.5503	0.6343
19165	0.5524	0.5947	0.7810	0.7080	0.5503	0.6373
19004	0.5790	0.6382	0.7770	0.7166	0.5503	0.6522
19016	0.6179	0.7319	0.8076	0.7432	0.5503	0.6902
19026	0.7011	0.8551	0.7680	0.7107	0.5503	0.7170
19123	0.6737	0.8377	1.0236	0.9501	0.5503	0.8071

CUADRO 3.4: Conglomerados de las estaciones pluviométricas mediante la aplicación del exponente de Hurst promedio(HP).

Clúster	Estación	HP	Distribución	Parámetros
4	19015	0.4139	Gamma(α, β)	(0.5663, 136.1047)
	19009	0.4342	GEV(k, σ, μ)	(4.5647, 2.1172, 0.4636)
	19069	0.4496	Gamma(α, β)	(0.5302, 161.6313)
	19047	0.4639	GEV(k, σ, μ)	(0.6367, 23.7109, 18.5646)
11	19002	0.4980	GEV(k, σ, μ)	(0.8549, 22.9085, 16.5826)
	19264	0.5213	GEV(k, σ, μ)	(5.2249, 15.2275, 2.9142)
	19036	0.5269	GEV(k, σ, μ)	(3.4686, 0.0035, 0.0010)
	19018	0.5301	GEV(k, σ, μ)	(5.0506, 12.0178, 2.3793)
	19056	0.5377	Gamma(α, β)	(0.6680, 89.4467)
	19012	0.5442	Gamma(α, β)	(0.6943, 65.1631)
	19054	0.5443	GEV(k, σ, μ)	(4.5781, 2.2694, 0.4954)
	19045	0.5510	GEV(k, σ, μ)	(1.0453, 8.5772, 5.7794)
	19187	0.5594	Gamma(α, β)	(0.6392, 95.3088)
	19033	0.5649	GEV(k, σ, μ)	(1.3751, 18.3613, 10.4899)
	19200	0.5692	Gamma(α, β)	(0.5432, 119.4911)
14	19052	0.5770	GEV(k, σ, μ)	(0.8071, 19.1296, 15.1422)
	19134	0.5797	Gamma(α, β)	(0.7065, 55.4153)
	19158	0.5798	GEV(k, σ, μ)	(1.2889, 9.9186, 5.8144)
	19031	0.5822	GEV(k, σ, μ)	(0.9576, 24.7552, 16.3712)
	19039	0.5850	GEV(k, σ, μ)	(4.9383, 21.8802, 4.4299)
	19048	0.5971	Gamma(α, β)	(0.5441, 134.6999)
	19171	0.6044	GEV(k, σ, μ)	(0.9713, 22.2791, 14.9927)
	19140	0.6048	Gamma(α, β)	(0.5804, 98.0680)
	19022	0.6064	Gamma(α, β)	(0.6157, 72.0019)
	19185	0.6067	Gamma(α, β)	(0.5962, 70.8507)
	19003	0.6180	GEV(k, σ, μ)	(0.6442, 32.9121, 28.7095)
	19170	0.6230	GEV(k, σ, μ)	(5.1656, 7.1451, 1.3831)
	19173	0.6343	Gamma(α, β)	(0.5384, 118.6863)
	19165	0.6373	GEV(k, σ, μ)	(3.5250, 0.1478, 0.0415)
3	19004	0.6522	GEV(k, σ, μ)	(0.8467, 20.9170, 15.2324)
	19016	0.6902	Gamma(α, β)	(0.5520, 79.1747)
	19026	0.7170	GEV(k, σ, μ)	(5.1724, 5.5123, 1.0656)
1	19123	0.8071	GEV(k, σ, μ)	(5.1039, 4.0464, 0.7927)

CUADRO 3.5: Conglomerados de las estaciones pluviométricas mediante la aplicación de exponente de Hurst a variogramas.

Clúster	Estación	Hurst Variograma	Distribución	Parámetros
4	19015	0.5629	Gamma(α, β)	(0.5663, 136.1047)
	19069	0.5980	Gamma(α, β)	(0.5302, 161.6313)
	19047	0.6010	GEV(k, σ, μ)	(0.6367, 23.7109, 18.5646)
	19134	0.6252	Gamma(α, β)	(0.7065, 55.4153)
6	19031	0.6492	GEV(k, σ, μ)	(0.9576, 24.7552, 16.3712)
	19009	0.6642	GEV(k, σ, μ)	(4.5647, 2.1172, 0.4636)
	19012	0.6692	Gamma(α, β)	(0.6943, 65.1631)
	19003	0.7047	GEV(k, σ, μ)	(0.6442, 32.9121, 28.7095)
	19033	0.7148	GEV(k, σ, μ)	(1.3751, 18.3613, 10.4899)
	19264	0.7256	GEV(k, σ, μ)	(5.2249, 15.2275, 2.9142)
16	19187	0.7454	Gamma(α, β)	(0.6392, 95.3088)
	19158	0.7484	GEV(k, σ, μ)	(1.2889, 9.9186, 5.8144)
	19052	0.7514	GEV(k, σ, μ)	(0.8071, 19.1296, 15.1422)
	19018	0.7548	GEV(k, σ, μ)	(5.0506, 12.0178, 2.3793)
	19002	0.7621	GEV(k, σ, μ)	(0.8549, 22.9085, 16.5826)
	19016	0.7730	Gamma(α, β)	(0.5520, 79.1747)
	19004	0.7742	GEV(k, σ, μ)	(0.8467, 20.9170, 15.2324)
	19056	0.7746	Gamma(α, β)	(0.6680, 89.4467)
	19185	0.7777	Gamma(α, β)	(0.5962, 70.8507)
	19036	0.7869	GEV(k, σ, μ)	(3.4686, 0.0035, 0.0010)
	19123	0.7871	GEV(k, σ, μ)	(5.1039, 4.0464, 0.7927)
	19048	0.7882	Gamma(α, β)	(0.5441, 134.6999)
	19039	0.7892	GEV(k, σ, μ)	(4.9383, 21.8802, 4.4299)
	19022	0.7937	Gamma(α, β)	(0.6157, 72.0019)
	19200	0.7945	Gamma(α, β)	(0.5432, 119.4911)
	19171	0.8039	GEV(k, σ, μ)	(0.9713, 22.2791, 14.9927)
5	19045	0.8248	GEV(k, σ, μ)	(1.0453, 8.5772, 5.7794)
	19173	0.8468	Gamma(α, β)	(0.5384, 118.6863)
	19140	0.8521	Gamma(α, β)	(0.5804, 98.0680)
	19165	0.8660	GEV(k, σ, μ)	(3.5250, 0.1478, 0.0415)
	19054	0.8831	GEV(k, σ, μ)	(4.5781, 2.2694, 0.4954)
2	19170	0.9479	GEV(k, σ, μ)	(5.1656, 7.1451, 1.3831)
	19026	0.9806	GEV(k, σ, μ)	(5.1724, 5.5123, 1.0656)

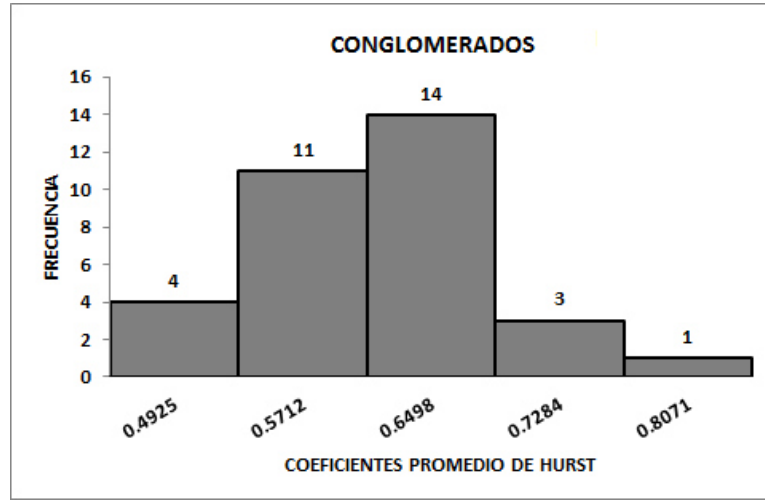


FIGURA 3.6: Histograma de las estaciones pluviométricas utilizando el exponente de Hurst aplicado a los datos de entrada.

Una vez realizado esto se sugirió la validación de la metodología, simulando series de datos con Distribución Normal $N \sim (0, \sigma^2)$, distribución General de valores extremos $Gev \sim (\mu, \sigma, \kappa)$, ($\mu = \text{ubicacion}, \sigma = \text{escala}, \kappa = \text{forma}$) y distribución Gamma $G \sim (\alpha, \beta)$ ($\alpha = \text{forma}, \beta = \text{escala}$).

El modelo de análisis regresivo que se utilizó para la simulación fue el mencionado en la sección 2.8:

$$y_t = \Phi_1 y_{t-12} + \xi \tag{3.4}$$

Se simuló 5 series de datos para cada valor de $\Phi = \{ 0.1, 0.3, 0.5, 0.7, 0.9 \}$ hasta formar 25 series con las distribuciones antes mencionadas asignando parámetros de $\kappa, \mu, \sigma, \alpha$ y β , según el tipo de distribución:

$$\begin{aligned} y_t &= \Phi_i y_{t-12} + G(\alpha, \beta) \\ y_t &= \Phi_i y_{t-12} + Gev(\kappa, \mu, \sigma) \\ y_t &= \Phi_i y_{t-12} + N(\mu, \sigma^2) \end{aligned} \tag{3.5}$$

Para cada una de estas ecuaciones en diferencias se ejecutaron 1000 simulaciones, calculando los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP), falsos negativos (FN), además se encontró el coeficiente de pureza de la relación entre grupos y el coeficiente de **Jaccard**. Esto se realizó tanto para analizar los valores iniciales, como para variogramas.

Enseguida se describe la información obtenida de las 1000 simulaciones para cada una de las distribuciones de prueba.

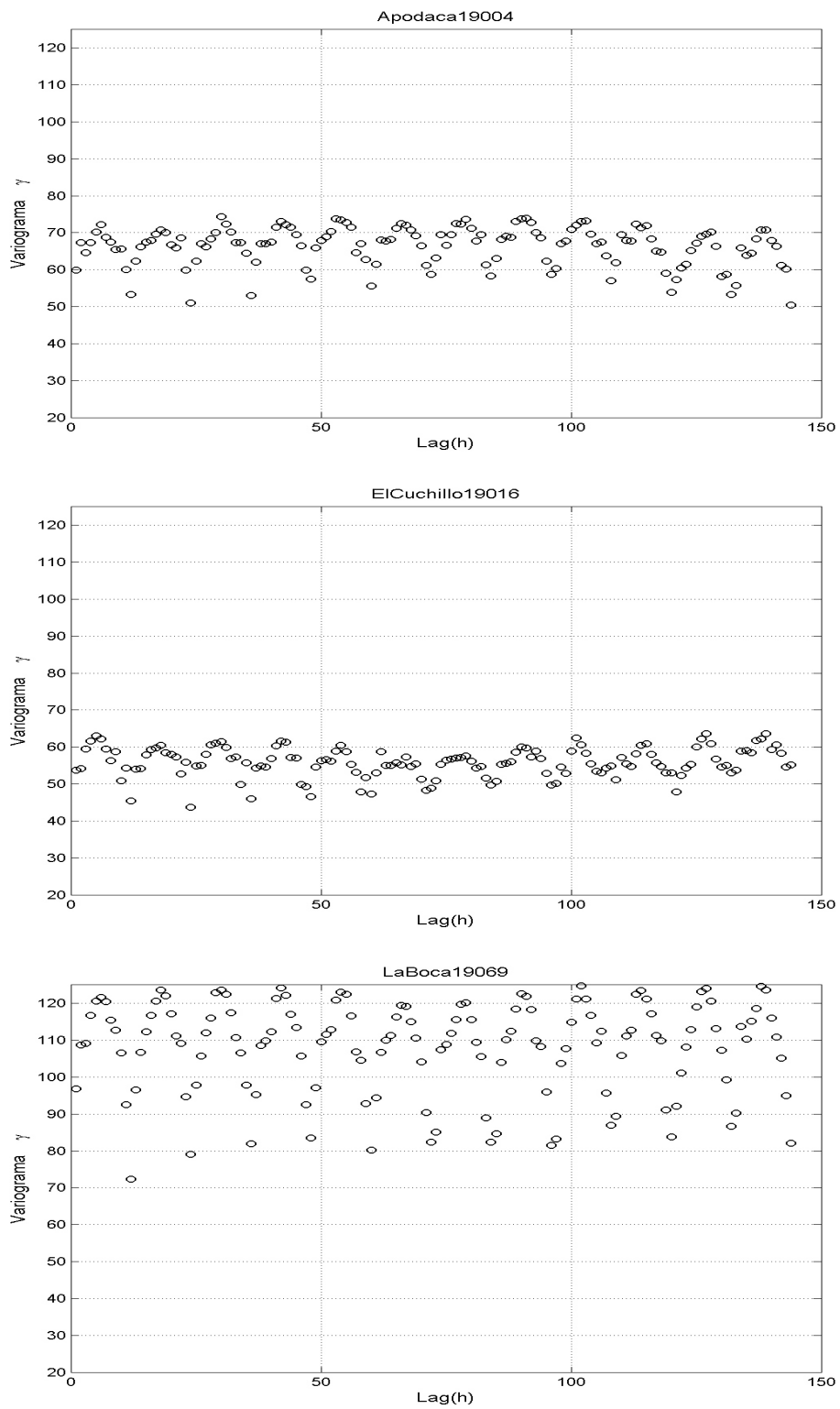


FIGURA 3.7: Variogramas
Gráfico de variogramas correspondientes a las series temporales de lluvia de las tres estaciones anteriores.

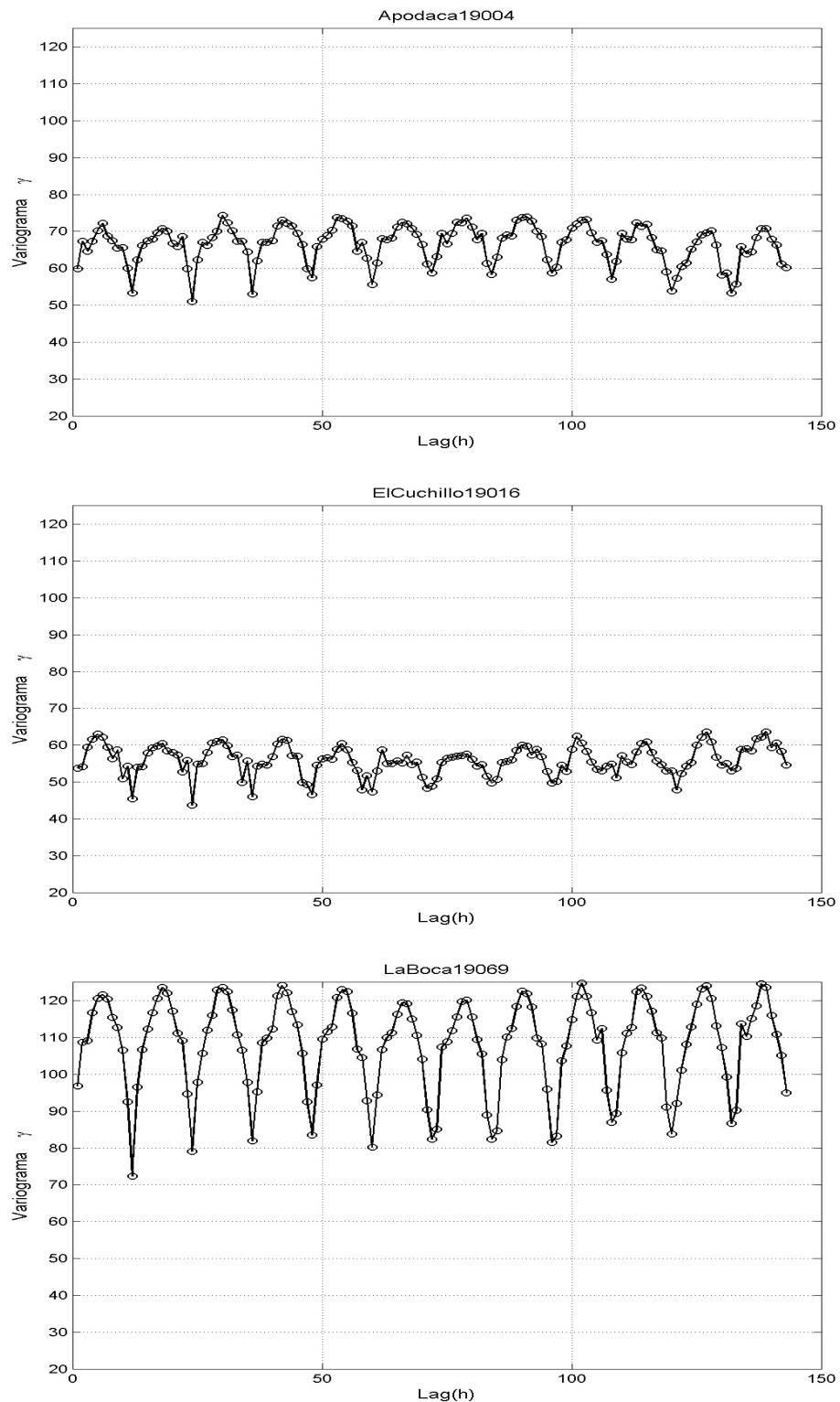


FIGURA 3.8: Variogramas

Gráfico del cálculo de variogramas. Desde la parte superior a la parte inferior, respectivamente: Apodaca, Estación número 19004, El Cuchillo, Estación número 19016, La Boca, Estación número 19069, variogramas correspondientes a las series temporales de lluvia de las tres estaciones anteriores.

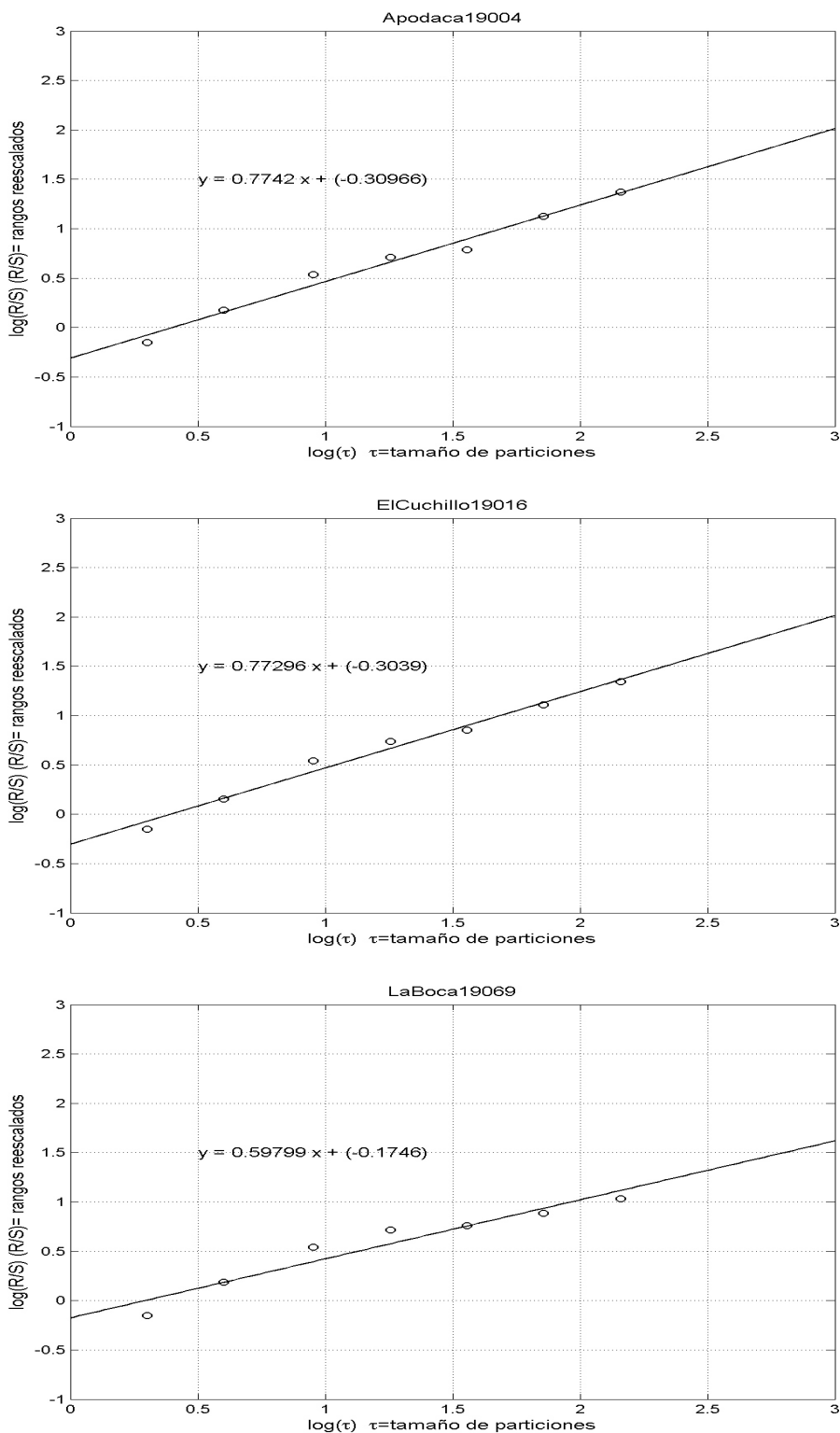


FIGURA 3.9: Mediciones del exponente de Hurst a variogramas de tres estaciones de la cuenca del río San Juan, México. El valor de la pendiente es el exponente de Hurst. Calculado por mínimos cuadrados. Desde la parte superior a la parte inferior, respectivamente: Apodaca, Estación número 19004, Enero 1940-Diciembre 2012, $m = 0,7742$; El Cuchillo, Estación número 19016, Enero 1939-Diciembre 2012, $m = 0,77296$; La Boca, Estación número 19069, Enero 1923-Diciembre 2012, $m = 0,59799$.

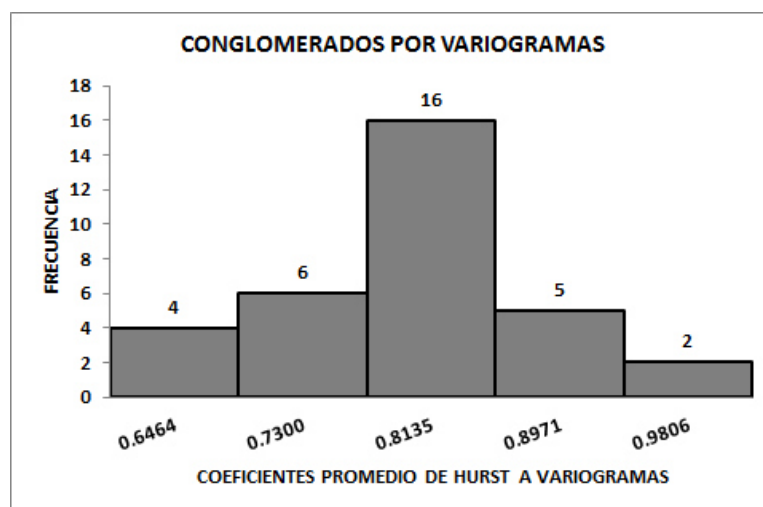


FIGURA 3.10: Histograma de las estaciones pluviométricas utilizando el exponente de Hurst aplicando variograma.

Los cuadros 3.6-3.9 muestran una selección de la 1ª, 10ª, 100ª y 1000ª simulación con distribución **Gamma**, $G \sim (\alpha, \beta)$ ($\alpha = forma, \beta = escala$) y los valores **VP**, **FN**, **FP** y **VN**, la pureza del agrupamiento y el índice de **Jaccard** el cual mide el grado de similitud entre agrupamiento clasificados con base a las particiones generadas por los valores de Φ , mencionados en 2.16 y las ecuaciones 3.5.

El cuadro 3.6 muestra el resumen de valores obtenidos de las 1000 simulaciones. A cada simulación se le calculó el exponente de Hurst, estos se agruparon por frecuencias y se compararon con los conglomerados iniciales.

CUADRO 3.6: Resumen comparativo de conglomerados, iniciales v.s. simulados mediante la aplicación del exponente de Hurst por frecuencias (**H**).

Simulación	VP	FN	FP	VN	PUREZA	JACCARD
1	12	38	46	204	0.48	0.125
10	18	32	65	185	0.6	0.1565
100	9	41	58	192	0.44	0.0833
1000	12	38	48	202	0.52	0.1224

El cuadro 3.7 muestra el resumen entre la partición inicial y los coeficientes de Hurst clasificados por el método de **k-medias(HK)**.

El cuadro 3.8 muestra el resumen entre la partición inicial y los coeficientes de Hurst aplicados a variogramas (**HV**) clasificados por frecuencias.

El cuadro 3.9 muestra el resumen entre la partición inicial y la partición generada por el coeficientes de Hurst aplicados a variogramas y clasificados por el método de k-medias (**HVK**).

El cuadro 3.10 muestra el resumen final de las 1000 simulaciones Gamma.

CUADRO 3.7: Resumen comparativo de conglomerados, iniciales v.s. los simulados mediante la aplicación de exponente de Hurst clasificados por k-medias (Hurst k-medias (HK)).

Simulación	VP	FN	FP	VN	PUREZAKM	JACCARDKM
1	12	38	44	206	0.52	0.1277
10	13	37	54	196	0.52	0.125
100	7	43	53	197	0.4	0.068
1000	14	36	47	203	0.56	0.1443

CUADRO 3.8: Resumen comparativo de conglomerados, iniciales v.s. exponentes de Hurst a variogramas clasificados por frecuencias(HV).

Simulación	VP	FN	FP	VN	PUREZAVAR	JACCARDVAR
1	13	37	43	207	0.52	0.1398
10	8	42	60	190	0.44	0.0727
100	15	35	54	196	0.52	0.1442
1000	14	36	45	205	0.52	0.1474

CUADRO 3.9: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst a variogramas con k-medias (HVK).

Simulación	VP	FN	FP	VN	PUREZAVK	JACCARDVK
1	9	41	51	199	0.44	0.0891
10	8	42	60	190	0.44	0.0727
100	10	40	51	199	0.44	0.099
1000	10	40	44	206	0.48	0.1064

CUADRO 3.10: Resumen de 1000 simulaciones H, HV, HK, HVK.

	H	HV	HK	HVK
PUREZA				
PROMEDIO	0.5074	0.4884	0.5053	0.4839
DESVIACIÓN	0.0637	0.06	0.0625	0.058
JACCARD				
PROMEDIO	0.1274	0.1148	0.1282	0.1137
DESVIACIÓN	0.0346	0.0325	0.036	0.0325

Los cuadros 3.11-3.14 muestran una selección de la 1^a, 10^a, 100^a y 1000^a simulación con distribución, **Gev**, y los valores VP, FN, FP, VN, la pureza del agrupamiento y el coeficiente de Jaccard, de las comparaciones entre agrupamientos clasificados en base a las particiones generadas por los valores de Φ , mencionados en 2.16 y las ecuaciones 3.5 y las clasificaciones mediante el exponente de Hurst, Hurst k-medias, Hurst variograma y Hurst variograma k-medias.

El cuadro 3.11 muestra el resumen entre la partición inicial y la generada por los coeficientes de Hurst clasificado por frecuencias.

CUADRO 3.11: Resumen comparativo de conglomerados, iniciales v.s. los simulados mediante el exponente de Hurst por frecuencias(H).

Simulación	VP	FN	FP	VN	PUREZA	JACCARD
1	13	37	43	207	0.52	0.1398
10	21	29	59	191	0.64	0.1927
100	28	22	58	192	0.72	0.2593
1000	23	27	47	203	0.68	0.2371

El cuadro 3.12 muestra el resumen entre la partición inicial y los coeficientes de Hurst clasificados por el método de k-medias (**HK**).

CUADRO 3.12: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst por k-medias(HK).

Simulación	VP	FN	FP	VN	PUREZAKM	JACCARDKM
1	11	39	46	204	0.48	0.1146
10	18	32	54	196	0.56	0.1731
100	29	21	36	214	0.76	0.3372
1000	28	22	44	206	0.72	0.2979

El cuadro 3.13 muestra el resumen entre la partición inicial y los coeficientes de Hurst con variogramas clasificados por frecuencias.

CUADRO 3.13: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst variograms por frecuencias (**HV**).

Simulación	VP	FN	FP	VN	PUREZAVAR	JACCARDVAR
1	13	37	57	193	0.52	0.1215
10	12	38	52	198	0.48	0.1176
100	13	37	68	182	0.48	0.1102
1000	14	36	59	191	0.56	0.1284

El cuadro 3.14 muestra el resumen entre la partición inicial y los coeficientes de Hurst a variogramas aplicando k-medias (**HVK**).

El cuadro 3.15 muestra el resumen final de las 1000 simulaciones.

CUADRO 3.14: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst a variogramas por k-medias (HVK).

Simulación	VP	FN	FP	VN	PUREZAVK	JACCARDVK
1	17	33	46	204	0.6	0.1771
10	12	38	52	198	0.48	0.1176
100	18	32	55	195	0.6	0.1714
1000	14	36	47	203	0.52	0.1443

CUADRO 3.15: Resumen de 1000 simulaciones HF, HV, HK, HVK.

	H	HV	HK	HVK
PUREZA				
PROMEDIO	0.6180	0.4992	0.6126	0.4812
DESVIACIÓN	0.0719	0.0671	0.0699	0.0621
JACCARD				
PROMEDIO	0.2154	0.1134	0.2163	0.1089
DESVIACIÓN	0.0578	0.0309	0.0579	0.0307

De los cuadros 3.16-3.19 corresponden al resumen de simulaciones, **Normal**, muestran los valores VP, VN, FP, FN, la pureza del agrupamiento y el coeficiente de Jaccard, para las comparaciones entre agrupamientos clasificados en base a las particiones generadas por los valores de Φ , antes establecidos.

El cuadro 3.16 muestra el resumen entre la partición inicial y la generada por los coeficientes de Hurst clasificadas por frecuencias.

CUADRO 3.16: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst mediante frecuencias (H).

Simulación	VP	FN	FP	VN	PUREZA	JACCARD
1	13	37	49	201	0.52	0.1313
10	15	35	70	180	0.52	0.125
100	10	40	50	200	0.44	0.1
1000	14	36	47	203	0.56	0.1443

El cuadro 3.17 muestra el resumen entre la partición inicial y los coeficientes de Hurst clasificados por el método de k-medias (HK).

El cuadro 3.18 muestra el resumen entre la partición inicial y los coeficientes de Hurst a variogramas clasificado por frecuencias.

El cuadro 3.19 muestra el resumen entre la partición inicial y los coeficientes de Hurst a variogramas clasificados por el método de k-medias (HVK).

El cuadro 3.20 muestra el resumen final de las 1000 simulaciones con distribución Normal $N \sim (\mu, \sigma^2)$.

CUADRO 3.17: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst clasificado por k-medias (HK).

Simulación	VP	FN	FP	VN	PUREZAKM	JACCARDKM
1	13	37	49	201	0.52	0.1313
10	15	35	71	179	0.52	0.124
100	8	42	58	192	0.4	0.0741
1000	14	36	47	203	0.56	0.1443

CUADRO 3.18: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst a variogramas HV.

Simulación	VP	FN	FP	VN	PUREZAVAR	JACCARDVAR
1	12	38	66	184	0.44	0.1034
10	14	36	43	207	0.48	0.1505
100	14	36	72	178	0.48	0.1148
1000	10	40	62	188	0.48	0.0893

CUADRO 3.19: Resumen comparativo de conglomerados, iniciales v.s. exponente de Hurst a variogramas por k-medias (HVK).

Simulación	VP	FN	FP	VN	PUREZAVK	JACCARDVK
1	9	41	53	197	0.44	0.0874
10	13	37	40	210	0.48	0.1444
100	10	40	48	202	0.44	0.102
1000	8	42	52	198	0.44	0.0784

CUADRO 3.20: Resumen de 1000 simulaciones HF, HFV, HKM, HVK.

	H	HV	HK	HVK
PUREZA				
PROMEDIO	0.5067	0.4826	0.5007	0.4772
DESVIACIÓN	0.0613	0.0586	0.0601	0.0588
JACCARD				
PROMEDIO	0.1262	0.1098	0.1256	0.1083
DESVIACIÓN	0.0345	0.0301	0.0348	0.0309

Capítulo 4

Conclusiones y trabajo futuro

Conclusiones

Como caso de estudio para desarrollar esta investigación, utilizamos una muestra de estaciones de lluvia, las cuales están contenidas en la cuenca del río San Juan de la región hidrográfica RH-24 México. Fueron utilizadas 33 de las 41 estaciones en el análisis. Se revisaron las series de tiempo y se le analizó su comportamiento mediante el exponente de Hurst o rango reescalado (R/S) para efectuar una clasificación de cada una de ellas. La técnica del variograma también fue utilizada posteriormente. La primera metodología fue con el fin de efectuar conglomerados de acuerdo a la persistencias o no persistencia de los datos, y la segunda con el objetivo de efectuar conglomerados mediante a variabilidad.

El exponente de Hurst provee una medida para determinar si una serie de tiempo es ruido blanco gaussiano o tiene una tendencia subyacente y puede ser utilizada para agrupar las estaciones pluviométricas de acuerdo con a los valores obtenidos al aplicarse el análisis de rango reescalado.

El mapa de variogramas brinda una forma muy eficiente y automática de determinar si un conjunto de datos presenta un comportamiento que puede ser claramente observado y además seguido del análisis de R/S o estimación del exponente de Hurst, revela mayor información utilizable como herramienta de un análisis de conglomerados.

La dependencia de largo alcance se encuentra en cada uno de los variogramas evaluados con el exponente de Hurst, sin embargo, se encontró todavía más útil como una herramienta para un análisis de conglomerados.

Una bondad del proceso de ajuste se ejecuta con cada serie, y los resultados mostraron que existen distribuciones dominantes dentro de un conjunto factible (que se encuentra de forma independiente en cada estación).

Se encontró que las distribuciones de probabilidad anidan dentro de cada grupo. Esto es indicativo de que los patrones homogéneos fueron identificados dentro de los grupos, y los grupos fueron heterogéneos entre sí.

Del análisis de validación se puede decir lo siguiente:

De resumen final de la tabla 3.10 de las simulaciones desarrollando mediante la distribución Gamma, $G \sim (\alpha, \beta)$. Se efectuó una prueba de diferencia de medias y un análisis de varianza del cual podemos concluir:

En lo que respecta a la pureza:

- No hay evidencia estadística suficiente para decir que las desviaciones de los indicadores de Hurst sean distintas a los de Hurst K-medias.
- Se encontró diferencia estadísticamente significativa (nivel menor a 0.02) entre los promedios de los indicadores de Hurst y los de Hurst K-medias.
- Se encontró diferencia significativa (nivel menor a 0.01) entre promedios de los indicadores Hurst variograma v.s. Hurst variograma K-medias.

- Hay diferencia entre las varianzas (nivel menor a 0.01) de los indicadores de Hurst variograma y Hurst variograma K-medias.

En lo que respecta al coeficiente de Jaccard para las simulaciones, con la distribución Gamma, para cada indicador:

- Utilizando el índice de Jaccard no se encontró diferencia significativa para los promedios, ni para las varianzas de los índices de Hurst v.s. Hurst K-medias y para Hurst variograma v.s. Hurst variograma K-medias.

De resumen final de la tabla 3.15 de las simulaciones desarrollando mediante la distribución General de valores extremos, $Gev \sim (\mu, \sigma, \kappa)$, se aplicó una prueba de diferencia de medias y un análisis de varianza que revelan lo siguiente:

En lo que respecta a la pureza:

- En promedio el índice de Hurst es mayor que Hurst K-medias con nivel de significancia menor a 0.02 y, que no hay diferencia significativa en la variabilidad.
- En cuanto a los indicadores de Hurst variograma v.s. Hurst variograma K-medias, se encontraron diferencias significativas en promedios y en varianzas con niveles de significancia menores a 0.01.

En lo que respecta al coeficiente de Jaccard para las simulaciones, con la distribución Gev, para cada indicador:

- No se detecta diferencia significativa ni para promedios ni para varianzas en la comparación Hurst v.s. hurst K-media a nivel de significancia 0.05.

De resumen final de la tabla 3.20 de las simulaciones desarrollando mediante la distribución Normal, también se aplicó una prueba de diferencia de medias y un análisis de varianza que muestran los siguiente resultados:

En lo que respecta a la pureza:

- No hay evidencia estadística suficiente para decir que las desviaciones de los indicadores de Hurst sean distintas a los de Hurst K-medias.
- No hay diferencia estadísticamente significativa entre los promedios de los indicadores de Hurst y los de Hurst K-medias.
- Se encontró diferencia significativa (nivel menor a 0.01) entre promedios de los indicadores Hurst variograma v.s. Hurst variograma K-medias.
- Hay diferencia entre las varianzas (nivel menor a 0.01) de los indicadores de Hurst Variograma y Hurst Variograma K-medias.

En lo que respecta al coeficiente de Jaccard para las simulaciones, con la distribución Normal $N \sim (0, \sigma^2)$, para cada indicador:

- Utilizando el índice de Jaccard no se encontró diferencia significativa para los promedios ni para las varianzas de los índices de Hurst v.s. Hurst- kmedias.

- Se encontró diferencia en las varianzas de los indicadores Hurst variograma v.s. Hurst variograma K-medias con nivel de significancia 0.01. Mientras que no hay evidencia de que exista una diferencia entre los promedios en el indicador Hurst variograma y Hurst variograma K-medias a nivel de significancia de 0.05.

A la luz de estos resultados podemos concluir que en promedio el índice de Hurst y el de Hurst variograma determinan en mayor medida que los de Hurst k-medias y Hurst variograma k-medias respectivamente el grado de coincidencia (pureza) en los conglomerados, con la ventaja adicional de la simplicidad en su algoritmo e interpretación.

Para la determinación de la pureza el índice de Hurst resulta ser un mejor indicador, en promedio que el de k-medias, debido a que en promedio detecta más coincidencias y con la misma variabilidad o por lo menos, no distinta a la de Hurst K-medias.

Para estas aseveraciones se utilizaron pruebas estadísticas para diferencia de medias y cocientes de varianzas.

El estudio de las estaciones de lluvia con variogramas y el análisis *R/S* proporciona una herramienta alternativa que permite a los profesionales encontrar correlaciones a largo plazo y además analizar la agrupación de las series hidrológicas.

El análisis de agrupamiento se utiliza, intentando identificar las estaciones dentro de la cuenca con similar régimen hidrológico.

Los parámetros de ubicación, dispersión, asimetría y curtosis de las distribuciones ajustadas a cada conglomerado, se obtuvieron utilizando L-momentos, los cuales dan una mejor caracterización de la distribución de probabilidad de nuestras muestras.

Se identificaron cinco regiones de la cuenca, que son hidrológicamente más parecidas entre sí que a captaciones en otros lugares pero, no del mismo modo distintivas entre sí ya que sus distribuciones son, generalmente, distribuciones; Gamma y Gev.

En trabajos futuros, esta metodología podrá ser utilizada en el estudio de las complejas series de tiempo, para mejorar la comprensión de estas y, si se analizan más variables, se podría efectuar una clasificación de los conglomerados de manera más efectiva que solo considerando una variable como, la pluviométrica.

Si se considerasen más factores de igual importancia relacionados a la hidrología del área, igualmente importantes, sería como un complejo mosaico de zonas hidrológicamente homogéneas, y más difícil relacionarlas a una gran escala realista, pero mediante este tipo de análisis, se podría hacer una mejor clasificación de conglomerados.

Capítulo 5

Bibliografía

Bibliografía

- [1] Soley, F. Javier. Descripción de dos métodos de rellenado de datos ausentes en series de tiempo meteorológicas. *Revista de Matemática: Teoría y Aplicaciones*, vol. 16, núm. 1, enero-junio, 2009, pp. 60-75.
- [2] Georges Matheron. *Curso de Geoestadística*.
- [3] Hurst, H. E. (1951). Long-Term Storage Capacity of Reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1), 770-799.
- [4] Tate Dalrymple. Flood-frequency analyses, *Manual of Hydrology: Part 3 Water Supply Paper 1543-A*.
- [5] Hosking, J. R. M., & Wallis, J. R. (2005). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press.
- [6] Flood frequency analysis. A Ramachandra Rao, Khaled H. Hamed.
- [7] Lena M. Tallaksen, Henny A.J. Van Lanen. *Hydrological Drought, processes and estimation methods for streamflow and groundwater*.
- [8] Durrans and Tomic. *Advances in Data-based Approaches for Hydrologic Modeling*.
- [9] Murugesu Sivapalan. *Effects of spatial variability and scale with implications to hydrologic modeling*.
- [10] Dominic Mazvimavi. *A review of aspects of hydrological sciences research in Africa over the last decade*.
- [11] T. M. L. WIGLEY. *Global temperature variations between 1861 and 1984*.
- [12] Andrew W. Lo. *Behaviors as the product of evolution in stochastic environments*.
- [13] Mandelbrot, B. B., & Wallis, J. R. (1969). Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. *Water Resources Research*, 5(5), 967-988.
- [14] Mandelbrot, B. (1972). Statistical methodology for nonperiodic cycles: from the covariance to R/S analysis. In *Annals of Economic and Social Measurement*, 1(3), 259-290.
- [15] Aristid Lindenmayer. *Modelled the behaviour of cells of plants. L-systems nowadays are also used to model whole plants*.
- [16] Witt, A., Malamud, B. D. (2013). Quantification of Long-Range Persistence in Geophysical Time Series: Conventional and Benchmark-Based Improvement Techniques. *Surv Geophys*, 34, 541–651.

- [17] Haslett, J. (2002). On the sample variogram and the sample autocovariance for non-stationary time series. *Journal of the Royal Statistical Society: Series D*, 46(4), 475-484.
- [18] Dmowska, R., Saltzman, B. (Editors) (1999). *Advances in geophysics, Vol 40. Long-Range Persistence in Geophysical Time Serie*. Academic Press, USA.
- [19] Lefevbre, M. (2007). *Applied stochastic processes*. Springer, USA.
- [20] A. I. McKerchar. Application of seasonal parametric linear stochastic models to monthly flow data.
- [21] M. G. Schaefer. Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington State.
- [22] M.C. Acreman, C.D. Sinclair. Classification of drainage basins according to their physical characteristics - An application for flood frequency analysis in Scotland. *Journal of Hydrology*, 84, 365- 380.
- [23] Nathaniel B. Guttman. 1992-1993 Winter precipitation in southwest Arizona.
- [24] Machiwal, D., & Jha, M. K. (2012). Current Status of Time Series Analysis in Hydrological Sciences. In *Hydrologic Time Series Analysis: Theory and Practice* (pp. 96-136). Springer Netherlands.
- [25] Bhuiya, R. K. (1971). Stochastic analysis of periodic hydrologic process. *Journal of the Hydraulics Division*, 97(7), 949-962.
- [26] Buishand, T. A. (1979). Urbanization and changes in precipitation, a statistical approach. *Journal of Hydrology*, 40(3), 365-375.
- [27] Buishand, T. A. (1982). Some methods for testing the homogeneity of rainfall records. *Journal of hydrology*, 58(1), 11-27.
- [28] Buishand, T. A. (1984). Tests for detecting a shift in the mean of hydrological time series. *Journal of Hydrology*, 73(1), 51-69.
- [29] Kothiyari, U. C., Singh, V. P., & Aravamuthan, V. (1997). An investigation of changes in rainfall and temperature regimes of the Ganga Basin in India. *Water resources management*, 11(1), 17-34.
- [30] Giakoumakis, S. G., & Baloutsos, G. (1997). Investigation of trend in hydrological time series of the Evinos River basin. *Hydrological sciences journal*, 42(1), 81-88.
- [31] Ángel, J. R., & Huff, F. A. (1997). Changes in heavy rainfall in Midwestern United States. *Journal of Water Resources Planning and Management*, 123(4), 246-249.
- [32] Mirza, M. Q., Warrick, R. A., Ericksen, N. J., & Kenny, G. J. (1998). Trends and persistence in precipitation in the Ganges, Brahmaputra and Meghna river basins. *Hydrological Sciences Journal*, 43(6), 845-858.
- [33] Tarhule, A., & Woo, M. K. (1998). Changes in rainfall characteristics in northern Nigeria. *International Journal of Climatology*, 18(11), 1261-1271.

- [34] Luis, M. D., Raventós, J., González-Hidalgo, J. C., Sánchez, J. R., & Cortina, J. (2000). Spatial analysis of rainfall trends in the region of Valencia (East Spain). *Int. J. Climatol*, 20(12), 1451-1469.
- [35] Kripalani, R. H., & Kulkarni, A. (2001). Monsoon rainfall variations and teleconnections over South and East Asia. *International Journal of Climatology*, 21(5), 603-616.
- [36] Adamowski, K., & Bougadis, J. (2003). Detection of trends in annual extreme rainfall. *Hydrological Processes*, 17(18), 3547-3560.
- [37] Yu, P. S., Yang, T. C., & Kuo, C. C. (2006). Evaluating long-term trends in annual and seasonal precipitation in Taiwan. *Water Resources Management*, 20(6), 1007-1023.
- [38] Kumar, V., Jain, S. K., & Singh, Y. (2010). Analysis of long-term rainfall trends in India. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 55(4), 484-496.
- [39] Golian, S., Saghafian, B., Sheshangosht, S., & Ghalkhani, H. (2010). Comparison of classification and clustering methods in spatial rainfall pattern recognition at Northern Iran. *Theoretical and applied climatology*, 102(3-4), 319-329.
- [40] Shi, P., Ma, X., Chen, X., Qu, S., & Zhang, Z. (2013). Analysis of variation trends in precipitation in an upstream catchment of Huai River. *Mathematical Problems in Engineering*, 2013.
- [41] Golder, J., Joelson, M., Neel, M. C., & Di Pietro, L. (2014). A time fractional model to represent rainfall process. *Water Science and Engineering*, 7(1), 32-40.
- [42] Chang, Y. C. (2014). Efficiently Implementing the Maximum Likelihood Estimator for Hurst Exponent. *Mathematical Problems in Engineering*, 2014.
- [43] Yu, Z. G., Leung, Y., Chen, Y. D., Zhang, Q., Anh, V., & Zhou, Y. (2014). Multifractal analyses of daily rainfall time series in Pearl River basin of China. *Physica A: Statistical Mechanics and its Applications*, 405, 193-202.
- [44] Kantelhardt, J. W. (2009). Fractal and multifractal time series. In *Encyclopedia of Complexity and Systems Science* (pp. 3754-3779). Springer, New York.
- [45] Carbone, M., Turco, M., Brunetti, G., & Piro, P. (2015). A Cumulative Rainfall Function for Subhourly Design Storm in Mediterranean Urban Areas. *Advances in Meteorology*, 2015.
- [46] Chou, C. M. (2014). Complexity analysis of rainfall and runoff time series based on sample entropy in different temporal scales. *Stochastic Environmental Research and Risk Assessment*, 28(6), 1401-1408.
- [47] García-Marín, A. P., Estévez, J., Medina-Cobo, M. T., & Ayuso-Muñoz, J. L. (2015). Delimiting homogeneous regions using the multifractal properties of validated rainfall data series. *Journal of Hydrology*, 529, 106-119.

Capítulo 6

Anexos

6.1. Descripción de la cuenca del río San Juan

La cuenca del río San Juan pertenece a la región hidrológica río Bravo-San Juan o RH-24 se muestra en la Fig. 6.1, esta comprende parte de los estados de Coahuila, Nuevo León y Tamaulipas; así como los diferentes aprovechamientos hidráulicos que en esta existen, tales como el Distrito de Riego 026 Bajo río San Juan, el Distrito de Riego 031 Las Lajas, el acueducto China-Monterrey, el acueducto regional China-Aldamas y las unidades de riego que se ubican en el río Pesquería, arroyo Ayancual y río San Juan.

Esta cuenca tiene una superficie de 32,972 km² y es el segundo afluente en importancia de los aportadores mexicanos al río Bravo cuya confluencia ocurre a 58 km aguas abajo de la cortina de la presa Internacional Falcón y 383 km aguas arriba de la desembocadura en el Golfo de México.

La cuenca se localiza entre los paralelos 25°15' y 26°45' de latitud Norte y los meridianos 99°15' y 101°45' de longitud Oeste.



FIGURA 6.1: Ubicación de la cuenca: "Google Earth"

Este río es uno de los más importantes de la región Noreste del país, abarcando territorio de tres estados que son Coahuila con 13,123 km², con 18,860 km y Tamaulipas con 989 km² (Secretaría de Agricultura y Recursos Hidráulicos (SARH), 1973).

La presa Rodrigo Gómez “La Boca”, se construyó en la cabecera del río San Juan en el municipio de Santiago Nuevo León, es ubicada entre las coordenadas en UTM Nuevo León, X Mínima 383871.168, X Máxima 386780.921, con Y Mínima 2810047.796 y Y Máxima 2815299.763. Cuenta con una superficie aproximada de 455Ha. y un volumen de almacenamiento de 40 millones de metros cúbicos (Mm³). Es alimentada principalmente por el arroyo La Chueca con dirección sureste y capta las aguas de arroyos perennes que descargan de la Sierra Madre Oriental a altitudes entre 2,000 a 2,300 msnm, estos arroyos son: Cavazos, Cristalinas, Dolores, Escamilla, Puerco y San Antonio; continua como efluente después de la compuerta con el nombre de río San Juan.

6.2. Técnicas de regionalización

En la década de los años 80, da inició formalmente el denominado análisis regional de los datos para predicción en sitios donde no se tiene tal información y obteniendo estimaciones más confiables para aplicarlos en cuencas con datos insuficientes.

Aunque este método da el índice de avenida, data de 20 años antes (Dalrymple, 1960[4]), es hasta finales de los 80 que se hace una exploración de este enfoque, en cuanto a métodos o procesos y sus ventajas (A. Ramachandra, Khaled H. Hamed 1988[6]).

El análisis regional de frecuencias de crecientes (ARFC), permite realizar predicciones, es decir, estimaciones asociadas a una determinada probabilidad de excedencia, con base en todos los datos observados en varias estaciones hidrométricas de una región, y que incluso se utilizan en las pruebas que emplean características climáticas y/o fisiográficas de cuencas (Lena M. and Henny[7]).

El análisis regional de frecuencia es la estimación de cuan a menudo un evento específico puede ocurrir. La estimación de la frecuencia de eventos extremos es de particular importancia, ya que existen numerosas fuentes de incertidumbre sobre los procesos físicos que amplifican a los eventos observados, es así como una aproximación estadística para el análisis de datos es necesaria.

Los métodos estadísticos son nobles con la existencia de incertidumbre, permiten cuantificar sus efectos y los procedimientos para el análisis de frecuencias estadísticas son un agregado singular, que también nos es de utilidad.

Por ejemplo, las observaciones meteorológicas del medio ambiente, de una misma variable, en diferentes lugares de medición con una frecuencia de eventos, tiene similares o diferentes cantidades observadas, entonces la conclusión exacta podría alcanzarse analizando todos los datos de las muestras, usando únicamente una simple muestra.

En aplicaciones ambientales este método es conocido como **ANÁLISIS REGIONAL DE FRECUENCIAS**, donde los datos de las muestras analizadas son típicamente observaciones de la misma variable en un número de lugares de medición dentro de lo que se define como **“REGIÓN”**. Los principios de análisis de frecuencia regional, sin embargo, son aplicables, siempre que las muestras múltiples de los datos similares están disponibles.

Diversas técnicas para llevar a cabo una regionalización hidrológica se han desarrollado para facilitar y hacer más rápido los análisis. Durrans & Tomic[8](1996) indicaron que ciertas técnicas se pueden clasificar en dos tipos.

La primera, dedicada a la predicción de caídas de agua en cuencas (Sivapalan[9], 2003), en la que la relación de ciertas características hidrológicas (por ejemplo, el pico de

descarga o de bajo flujo) con características climáticas y fisiográficas se establecen para medir las cuencas hidrográficas. Esa relación se puede aplicar sin datos en cuencas y predecir características hidrológicas mediante observaciones fisiográficas y características climáticas.

La segunda, la de análisis regional, conocida como análisis regional de frecuencia o análisis de frecuencia regional. Su objetivo es mejorar la estimación en algunos sitios medidos por medio del uso de la información de otros sitios calibrados con datos de períodos más largos en una región homogénea. Hosking y Wallis[5], consideran este método como una forma de “Negociación en espacio-temporal”.

Métodos de regresión múltiple también han sido utilizado para este propósito durante muchos años (p. ej. Mazvimavi[10] et al., 2004).

Con el desarrollo de geotecnología de la información, como los sistemas de información geográfica (SIG) y teledetección, más y más información está disponible actualmente.

6.3. Recolección y depurado de información

Los datos fueron proporcionados por la CONAGUA. Las bases son archivos de excel los cuales se tienen que importar para su analisis, importándolos desde su origen.

Se complementaron los datos en el mismo formato, tabla con filas y columnas:

- Mismo tipo de datos en cada columna
- Todas las columnas y filas visibles
- Se buscó que no hubiese filas en blanco

Se ejecutaron, primero, las tareas que no necesiten la manipulación de columnas, posteriormente, las que requiera de manipulación.

Al revisar la información de las bases de datos se encontraron:

- Logotipos
- Comas en lugar de puntos
- Espacios entre números
- Letras en la captura
- Falta de información en ellas como S/D, s.d. (sin dato), NaN (Not a Number), INAP (no aplica), etc.
- Información que no corresponde a los datos como notas del capturista o registros del analista.

Esto fue un problema común en todas de las bases de datos. La tabla 1. muestra lo anteriormente expuesto.

Se utilizó librerías de Matlab como: xlsread (nombre del archivo, número de hoja). Instrucciones que permiten leer archivos desde Excel, representando valores que no son números con un valor especial llamado NaN, que significa “no es un número”. De esta forma, se logró detectar celdas con texto, comas, espacios, etc.

El cuadro 6.1 y 6.2 muestra lo anterior.

Ya depurada la base de datos, mediante interpolación se completó las faltantes y/o se tomaron en cuenta las observaciones de estaciones cercanas para el caso de celdas datos al inicio de la base o al final de esta.

El cuadro 6.3 muestra los datos ya efectuada la operación.

CUADRO 6.1: COMISIÓN NACIONAL DEL AGUA
ORGANISMO DE CUENCA RÍO BRAVO
DIRECCIÓN TÉCNICA
DATOS DE PRECIPITACIÓN MENSUAL EN mm.

AÑOS	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1987	50.0	63.0	35.5	68.0	79.5	356.1	146.1	241.5	356.0	69.0	13.0	9.0
1988	36.3	11.8	12.5	41.0	69.5	209.0	130.5	236.0	388.0	41.0	20,8	9.0
1989	30.0	41.5	8.5	54.0	15.5	51.5	116,0	44.5	251.0	192.0	28.5	74.0
1990	17.5	21.5	47.5	122.0	141.5	16.0	101.5	58.0	316.5	80.0	19.5	17.0
1991	26.5	47.5	35.5	123.5	167.5	172.5	46.0	30.5	262.0	42.0	29.5	54.0
1992	142.5	33.5	40.0	136.5	100.5	63.5	1.0	210.5	212.0	129.0	48.0	39.0
1993	33.5	46.5	51.5	23.8	151.0	298.5	28.0	58.5	367.5	50.5	31.0	19.5
1994	67.5	31.0	50.5	19.5	41.5	101.0	36.5	83.5	210.3	67.2	39.0	46.0
1995	28.5	27.0	50.0	S/D	S/D	S/D	s/d	S/D	S/D	S/D	S/D	S/D
1996	S/D	S/D	S/D	S/D	S/D	S/D	S/D	S/D	30.1	179.1	7.2	10.2
1997	18.7	61.8	72.2	121.2	128.8	68.1	21.4	69.3	42.8	226.4	36.7	14.8
1998	7.7	17.9	13.7	9.0	0.0	12.5	28.3	262.2	232.0	38.7	63.3	5.2
1999	0,0	0.0	25.8	25.7	34.1	103.9	145.3	0	72.4	25.7	0.0	14,3
2000	0.0	33.9	12.6	13.7	100.1	104.7	20.5	93.3	72.1	279.1	44.5	2
2001	47.1	5.3	68.7	78.0	15.4	99.6	25.9	25.8	564.9	42.0	131.2	24.4
2002	4.3	S/D	INAP	INap	INAP	S/D	118.6	93.2	407.2	308.2	8.6	5.3
2003	179.2	20.2	40.4	0.0	45.2	80.7	33.3	249.3	529.2	70.6	54.6	0.0
2004	0.0	38.7	221.3	167.0	19.4	23.3	28.8	253.8	415.3	70.1	10.3	6.0
2005	40.6	135.1	55.0	29.6	127.7	17.8	397.4	33.3	65.7	361.2	12.5	14.3
2006	0.0	28.7	19.6	10.4	49.7	56.5	74.0	87.0	244.6	19.3	36.0	INAP
2007	61.8	43.5	4.5	65.0	97.0	124.0	89.3	163.0	99.9	42.0	46.0	2.5
2008	25.5	5.0	34.5	91.5	58.0	1.0	169.0	169.0	789.5	130.0	2.0	14.5
2009	32.5	16.0	22.5	9.0	62.5	58.0	17.0	31.0	106.6	65.6	19.5	44.5
2010	S.D.	S.D.	20.0	299.2	S.D.	93.5	444.8	17.5	169.5	35.3	0.0	S.D.
2011	21.7	0.0	23.0	0.0	152.2	103.1	82.8	83.8	87.5	89.0	2.5	27.7

Cuadro 6.1 con asignaciones S/D, s/d, S.D., INAP, comas, espacios etc..

CUADRO 6.2: DATOS CON ASIGNACIÓN NAN

AÑOS	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1987	50.0	63.0	35.5	68.0	79.5	356.1	146.1	241.5	356.0	69.0	13.0	9.0
1988	36.3	11.8	12.5	41.0	69.5	209.0	130.5	236.0	388.0	41.0	NaN	9.0
1989	30.0	41.5	8.5	54.0	15.5	51.5	NaN	44.5	251.0	192.0	28.5	74.0
1990	17.5	21.5	47.5	122.0	141.5	16.0	101.5	58.0	316.5	80.0	19.5	17.0
1991	26.5	47.5	35.5	123.5	167.5	172.5	46.0	30.5	262.0	42.0	29.5	54.0
1992	142.5	33.5	40.0	136.5	100.5	63.5	1.0	210.5	212.0	129.0	48.0	39.0
1993	33.5	46.5	51.5	23.8	151.0	298.5	28.0	58.5	367.5	50.5	31.0	19.5
1994	67.5	31.0	50.5	19.5	41.5	101.0	36.5	83.5	210.3	67.2	39.0	46.0
1995	28.5	27.0	50.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1996	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	30.1	179.1	7.2	10.2
1997	18.7	61.8	72.2	121.2	128.8	68.1	21.4	69.3	42.8	226.4	36.7	14.8
1998	7.7	17.9	13.7	9.0	0.0	12.5	28.3	262.2	232.0	38.7	63.3	5.2
1999	NaN	0.0	25.8	25.7	34.1	103.9	145.3	0	72.4	25.7	0.0	NaN
2000	0.0	33.9	12.6	13.7	100.1	104.7	20.5	93.3	72.1	279.1	44.5	2
2001	47.1	5.3	68.7	78.0	15.4	99.6	25.9	25.8	564.9	42.0	131.2	24.4
2002	4.3	NaN	NaN	NaN	NaN	NaN	118.6	93.2	407.2	308.2	8.6	5.3
2003	179.2	20.2	40.4	0.0	45.2	80.7	33.3	249.3	529.2	70.6	54.6	0.0
2004	0.0	38.7	221.3	167.0	19.4	23.3	28.8	253.8	415.3	70.1	10.3	6.0
2005	40.6	135.1	55.0	29.6	127.7	17.8	397.4	33.3	65.7	361.2	12.5	14.3
2006	0.0	28.7	19.6	10.4	49.7	56.5	74.0	87.0	244.6	19.3	36.0	NaN
2007	61.8	43.5	4.5	65.0	97.0	124.0	89.3	163.0	99.9	42.0	46.0	2.5
2008	25.5	5.0	34.5	91.5	58.0	1.0	169.0	169.0	789.5	130.0	2.0	14.5
2009	32.5	16.0	22.5	9.0	62.5	58.0	17.0	31.0	106.6	65.6	19.5	44.5
2010	NaN	NaN	20.0	299.2	NaN	93.5	444.8	17.5	169.5	35.3	0.0	NaN
2011	21.7	0.0	23.0	0.0	152.2	103.1	82.8	83.8	87.5	89.0	2.5	27.7

Cuadro 6.2 con asignaciones NaN utilizando matlab.

CUADRO 6.3: DATOS SIN ASIGNACIÓN NAN

AÑOS	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1987	50.0	63.0	35.5	68.0	79.5	356.1	146.1	241.5	356.0	69.0	13.0	9.0
1988	36.3	11.8	12.5	41.0	69.5	209.0	130.5	236.0	388.0	41.0	20.8	9.0
1989	30.0	41.5	8.5	54.0	15.5	51.5	116.0	44.5	251.0	192.0	28.5	74.0
1990	17.5	21.5	47.5	122.0	141.5	16.0	101.5	58.0	316.5	80.0	19.5	17.0
1991	26.5	47.5	35.5	123.5	167.5	172.5	46.0	30.5	262.0	42.0	29.5	54.0
1992	142.5	33.5	40.0	136.5	100.5	63.5	1.0	210.5	212.0	129.0	48.0	39.0
1993	33.5	46.5	51.5	23.8	151.0	298.5	28.0	58.5	367.5	50.5	31.0	19.5
1994	67.5	31.0	50.5	19.5	41.5	101.0	36.5	83.5	210.3	67.2	39.0	46.0
1995	28.5	27.0	50.0	53.4	70.6	90.03	31.5	78.8	120.2	123.2	23.1	28.1
1996	23.6	44.4	61.1	87.3	99.7	79.1	26.4	74.1	30.1	179.1	7.2	10.2
1997	18.7	61.8	72.2	121.2	128.8	68.1	21.4	69.3	42.8	226.4	36.7	14.8
1998	7.7	17.9	13.7	9.0	0.0	12.5	28.3	262.2	232.0	38.7	63.3	5.2
1999	3.9	0.0	25.8	25.7	34.1	103.9	145.3	0	72.4	25.7	0.0	14.3
2000	0.0	33.9	12.6	13.7	100.1	104.7	20.5	93.3	72.1	279.1	44.5	2
2001	47.1	5.3	68.7	78.0	15.4	99.6	25.9	25.8	564.9	42.0	131.2	24.4
2002	4.3	12.8	554.5	39.0	30.3	91.2	118.6	93.2	407.2	308.2	8.6	5.3
2003	179.2	20.2	40.4	0.0	45.2	80.7	33.3	249.3	529.2	70.6	54.6	0.0
2004	0.0	38.7	221.3	167.0	19.4	23.3	28.8	253.8	415.3	70.1	10.3	6.0
2005	40.6	135.1	55.0	29.6	127.7	17.8	397.4	33.3	65.7	361.2	12.5	14.3
2006	0.0	28.7	19.6	10.4	49.7	56.5	74.0	87.0	244.6	19.3	36.0	8.4
2007	61.8	43.5	4.5	65.0	97.0	124.0	89.3	163.0	99.9	42.0	46.0	2.5
2008	25.5	5.0	34.5	91.5	58.0	1.0	169.0	169.0	789.5	130.0	2.0	14.5
2009	32.5	16.0	22.5	9.0	62.5	58.0	17.0	31.0	106.6	65.6	19.5	44.5
2010	27.1	8.0	20.0	299.2	107.4	93.5	444.8	17.5	169.5	35.3	0.0	36.1
2011	21.7	0.0	23.0	0.0	152.2	103.1	82.8	83.8	87.5	89.0	2.5	27.7

Cuadro 6.3 con asignaciones numéricas mediante interpolación.

La forma recomendada para importar datos de Excel, es guardar estos datos en un fichero csv (comma separated values), y posteriormente importarlos con el comando read.csv o read.csv2. Este procedimiento es válido para otro tipo de formato en el que tengamos los datos (SAS, SPSS, MINITAB,...).

Otras observaciones a los datos son los valores extremos ó atípicos, conocidos como Outliers. Aunque no necesariamente son errores, se generan por un comportamiento diferente a los datos normales ó como problemas en los sensores, distorsiones en el proceso, mala calibración de instrumentos así como errores humanos ó como en nuestro caso por la misma naturaleza de los fenómenos meteorológicos cuyo registro histórico se captura por los pluviómetros. También sobre este tema se han realizado múltiples investigaciones, entre las cuales se encuentran trabajos tipo resumen, trabajos comparativos y trabajos sobre técnicas específicas, entre muchos otros.

Hasta aquí se han descrito algunos elementos importantes que deben tomarse en consideración, los cuales pueden ser de utilidad y quizá, servir como guía, que oriente a otros investigadores a su aplicación al análisis de datos, acorde a las peculiaridades de estos.

Dado que los posibles problemas presentados por los datos son muchos y la cantidad de técnicas existentes también es alta, elaborar una guía metodológica completa, la cual contemple los diferentes problemas y que detecte la mayor cantidad posible para corregirlos, es una tarea ardua que vale la pena para garantizar la buena toma de decisiones.

6.4. Gráficos de las Series de Tiempo

El marco del trabajo propuesto ha sido probado sobre datos procedentes de la CO-NAGUA. El objetivo del análisis de una serie temporal es el conocimiento de su patrón de comportamiento, para así prever su evolución futura, suponiendo que las condiciones no variarán.

Dado que no se trata de fenómenos deterministas, sino sujetos a una aleatoriedad, el estudio del comportamiento pasado ayuda a inferir la estructura que permita predecir su comportamiento futuro. La particular forma de la información disponible de una serie cronológica (se dispone de datos en periodos regulares de tiempo) hace que las técnicas habituales de inferencia estadística no sean válidas para estos casos, ya que nos encontramos ante elementos procedentes de poblaciones de características y distribución desconocidas.

Actualmente existen muchas razones por las cuales es importante poder pronosticar utilizando datos de cualquier tipo. Los diferentes métodos para realizar un pronóstico dependen básicamente de la experiencia, la cantidad de información disponible, del nivel de dificultad que presenta la situación y del grado de exactitud o confianza necesaria en el pronóstico.

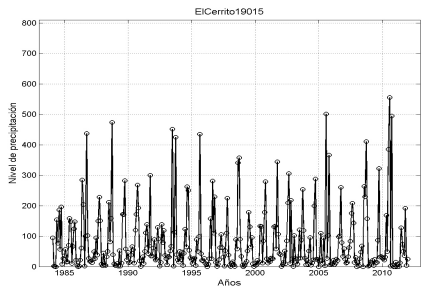
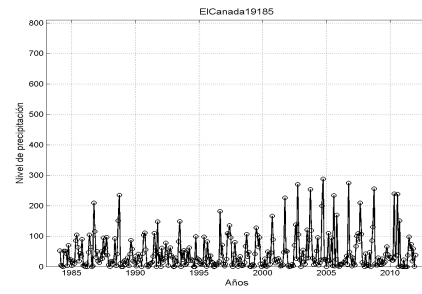
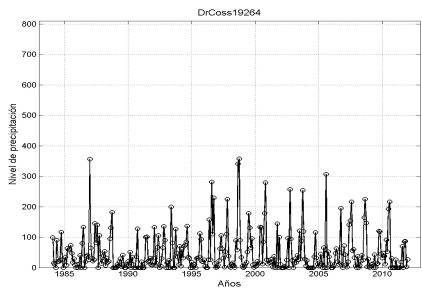
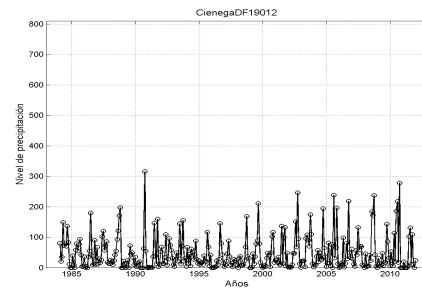
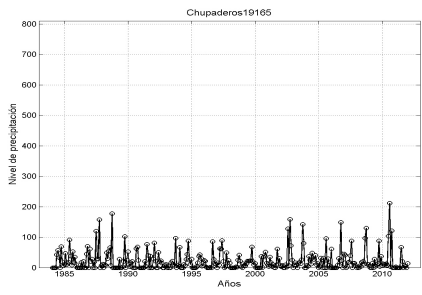
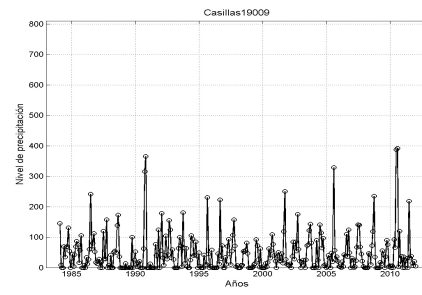
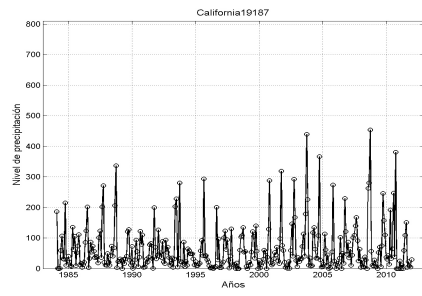
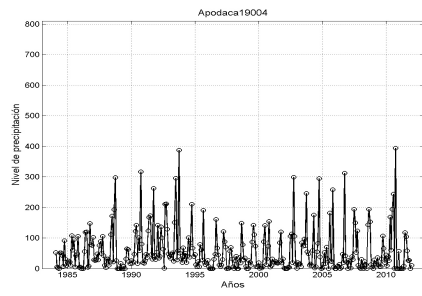
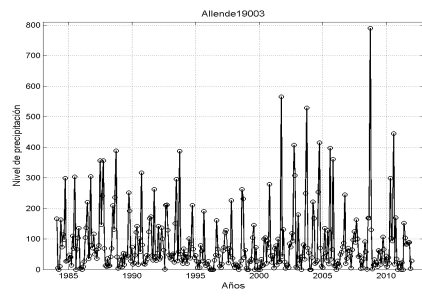
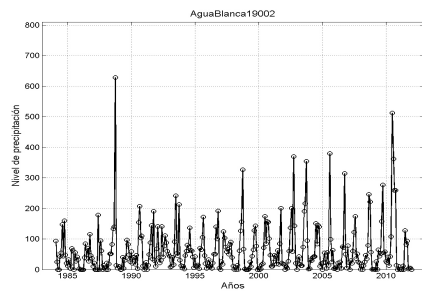
El Método de la persistencia, es la manera más simple de hacer un pronóstico, en él se asume que las condiciones atmosféricas no cambiarán en el tiempo. Si los patrones atmosféricos varían poco o se mueven lentamente en el tiempo, funcionan adecuadamente, pero, si las condiciones varían significativamente, de manera repentina este método falla y no es eficiente para pronosticar.

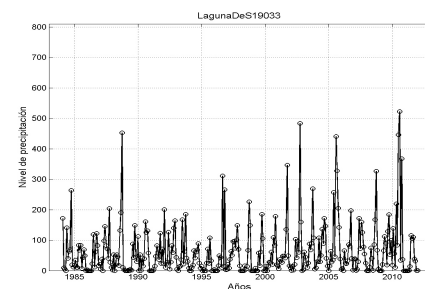
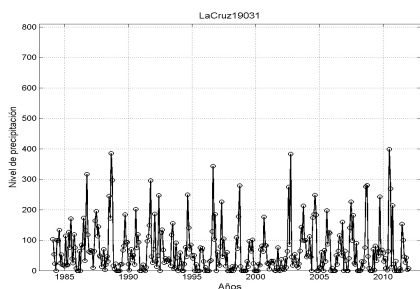
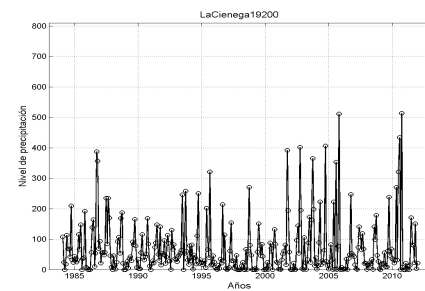
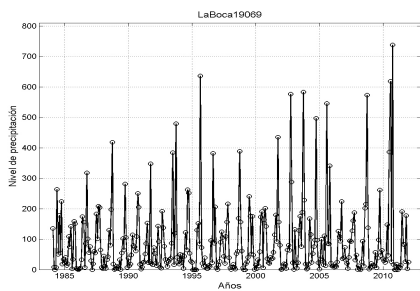
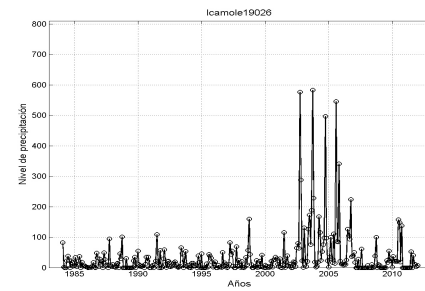
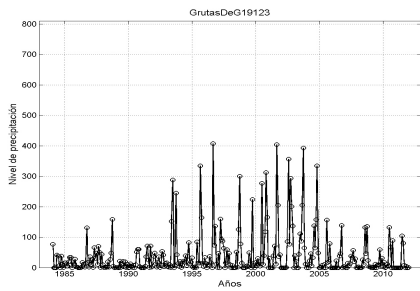
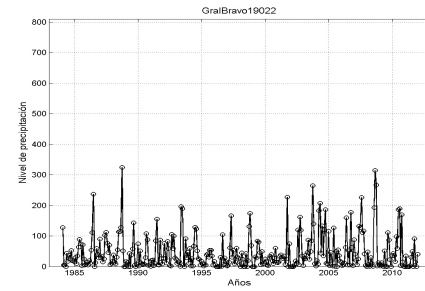
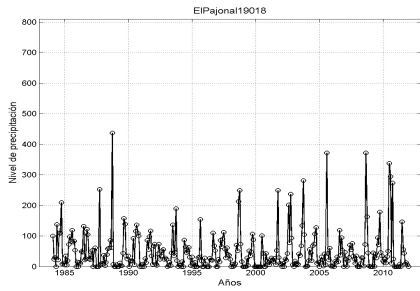
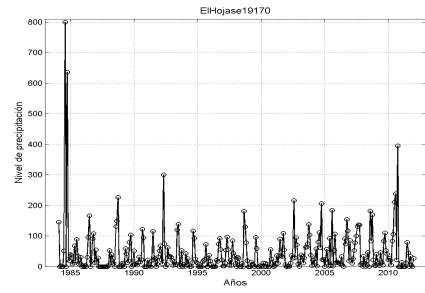
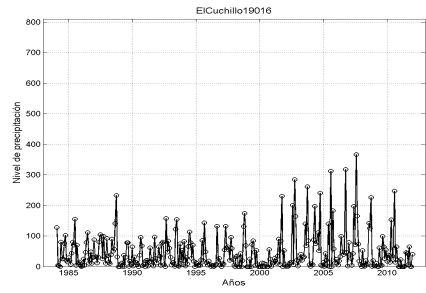
El Método de la tendencia, involucra diferentes cálculos como, medición de velocidad del viento, presiones atmosféricas, frentes, cúmulos de nubes y precipitación. Con esta información se puede predecir donde se esperan características semejantes en un tiempo posterior.

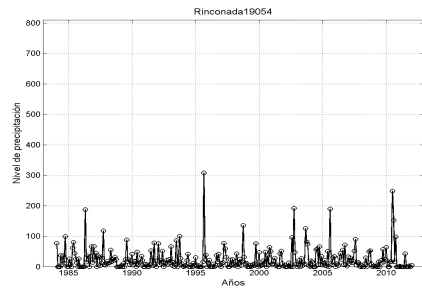
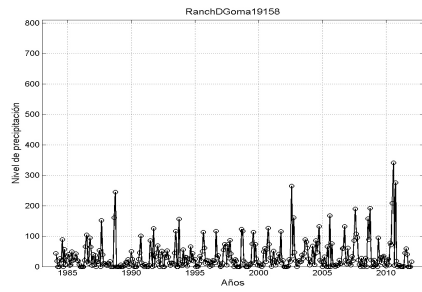
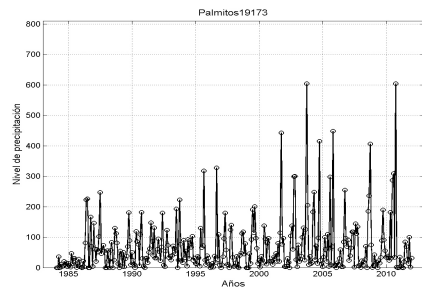
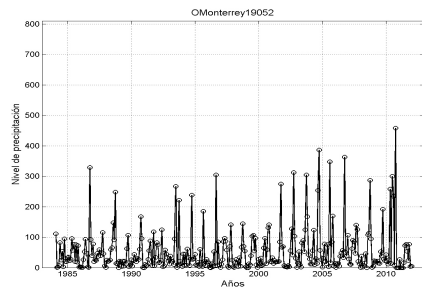
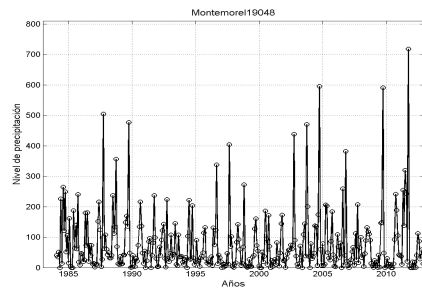
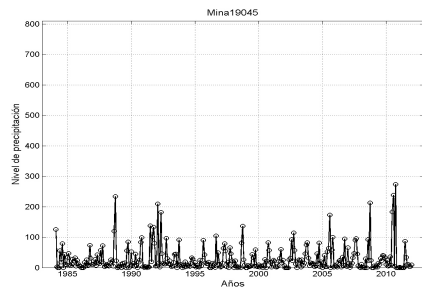
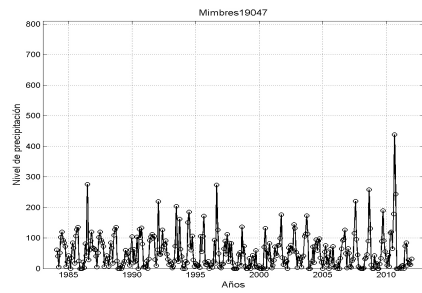
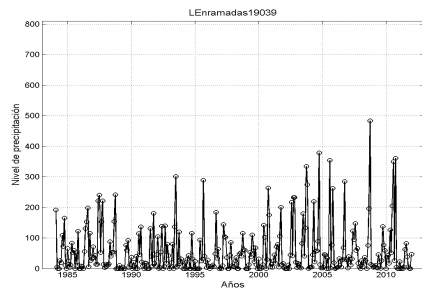
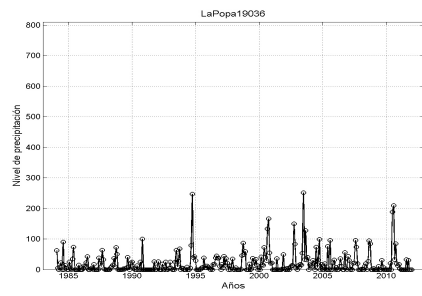
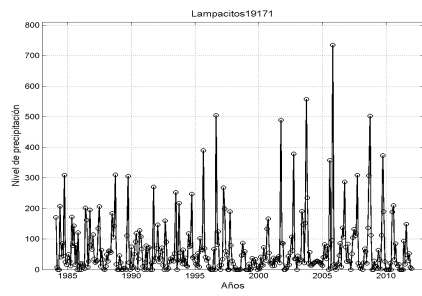
El Método climatológico, involucra el uso de promedios estadísticos de las diferentes variables atmosféricas, con registros históricos de años y algunos lo utilizan ya que parece simple realizar un pronóstico con él.

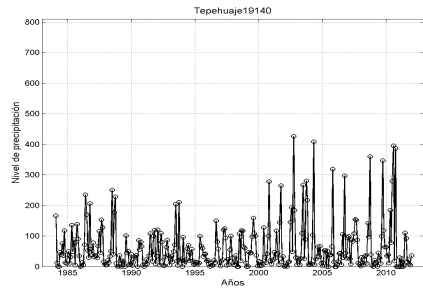
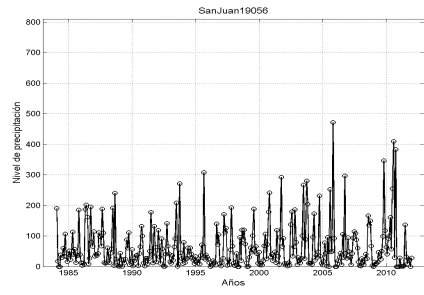
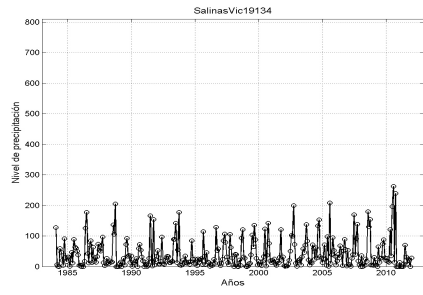
El Método análogo, es un poco más complicado, ya que se tienen que examinar condiciones actuales y recordar escenarios semejantes del pasado y su pronóstico futuro es equivalente al del pasado.

En este anexo se muestran las 33 series de tiempo de las estaciones pluviométricas analizadas.









6.5. Fractales

A temprana edad estudiamos medidas de longitudes, Dimensión 1, para después adentrarnos en formas geométricas elementales, figuras que pueden ser dibujadas o representadas en un plano, triángulos, cuadrados, rectángulos, polígonos, círculos y su contorno: la circunferencia, cosas bidimensionales, Dimensión 2.

Posterior a esto, nos adentramos a la geometría de los cuerpos que ocupan un volumen: prismas, cilindros, conos, esferas, pirámides y sus diferentes combinaciones figuras en Dimensión 3.

Observamos que estos representan una geometría muy regular, pero del cálculo aparente a la realidad esto no es del todo correcto.

Para representar la naturaleza es necesario utilizar un tipo de geometría que se aproxime más a esas observaciones que ejecutamos y que las incorpore lo más cerca de su propia naturaleza.

De esta manera es, como se presenta una geometría, que quizá no sea la última, para analizar este tipo de situaciones: la geometría fractal.

Esta palabra puede adjudicarse a Benoit Mandelbrot (1924-2010) quien la acuñó por primera vez en su libro "La geometría Fractal de la naturaleza", recopilando un buen número de figuras geométricas y procesos naturales bajo un marco global que denominó fractales. Del adjetivo en latín fractus, derivada del verbo frangere: romper, quebrar en pedazos, desacoplar.

Este término no pudo ser más apropiado ya que uno de las primeras figuras fue la curva creada en 1904 por el matemático sueco N. F. H. Von Koch, la cual exhibe claramente un concepto clave en los fractales: la autosimilitud. La información geométrica que contiene la curva está contenida en cualquier trozo por minúsculo que sea, su longitud diverge. Su distancia no puede ser calculada.

La curva es continua y no derivable en numerosos puntos. No podemos determinar la tangente en incontables de sus puntos.

Lo que Benoit interpreta con esto es que todos los objetos cumplen con una serie de propiedades, las cuales pueden ser descritas con una nueva geometría, la de las formas irregulares, debido a que todo tiene demasiada irregularidad como para ser descrito en términos geométricos tradicionales.

Un fractal, es por definición un conjunto cuya dimensión de Hausdorff-Besicovitch excede la dimensión topológica.

Desde la perspectiva de la física teórica y aplicada, para discutir una variedad de series espacio dimensional estamos sujetos a una comprensión del espacio multidimensional y que sólo puede quedarse en conceptos matemáticos.

Los fractales son un pensamiento pionero desde donde pueden ser abordados una gran diversidad de problemas, donde podemos incluir los de la naturaleza, y desde un punto crítico, analizar espacios no solamente unidimensionales, bidimensionales, tridimensional o espacio de 4 dimensiones. La primera, segunda o tercera dimensión o el mundo 1D, 2D o 3D, el de las mediciones, el de objetos planos, el de objetos sólidos y el espacio infinito, no ocupa pasar a un pensamiento mayor o menor sino solo aceptar lo que en él se presente o pueda ser analizado.

El tiempo, otra dimensión, un poco difícil de manejar, pero cuando lo empezamos a explorar, las cosas pueden ser más específicas, aunque en múltiples ocasiones la manipulación de esta dimensión, realmente es más difícil.

La ampliación de fractal, es generada por computadoras, mediante ecuaciones iteradas, no lineales, en diferentes dimensiones, incluso en el plano complejo. Estos pueden ser utilizados para desarrollar diversos procesos, como los pluricelulares. A. Lindenmayer, desarrolló los L-Systems, en los que se analizan diferentes comportamientos encontrando relaciones fractales o dimensión fractal. Su potencia reside en la recursividad.

En los 60's la física descubrió el caos determinista los cuales, estos, pueden generar dinámicas impredecibles, como son los conocidos atractores de Lorenz, Hénon y la curva logística de Feigenbaum, que no son otra cosa que fractales. El estudio de ellos, ha generado un avance en los sistemas dinámicos.

En los ochenta Barnsley, desarrolló fractales a partir de la aplicación iterada de contracciones afines. Aplicó rotaciones, contracciones, inversiones y traslaciones lo cual le permitió generar una gran cantidad de fractales, como el famoso helecho. Estos han sido utilizados para encriptación de imágenes, generación de estas y como una buena herramienta para ciberartistas.

Muchas de las figuras presentadas por Mandelbrot, ya habían sido propuestas como ejemplos paradójicos en la discusión de conceptos como curva o dimensión. Los atractores, que surgen de dinámicas que modelan sistemas reales, aparecen en el abstracto espacio de fases.

Estos conceptos pueden ser abordados con más detalle, cuando se utiliza el concepto de fractalidad. Podemos decir que la naturaleza está plagada de estos.

La geometría fractales es una rama de las matemáticas, estudia los objetos que poseen una dimensión D no entera ó dimension fractal y que presentan propiedades de escala muy particulares. D , es un número real que garantiza el concepto de dimensión ordinaria para objetos que no permiten espacio tangente.

La fractalidad ha encontrado aplicación en diversos campos de la ciencia y la tecnología, proporcionando modelos matemáticos alternativos para obtener distribuciones y propiedades de diversos fenómenos, como propiedades de rocas, yacimientos, generando distribuciones con cierta heterogeneidad y cierto orden a la vez.

Esta dimensión, D , con exponente no entero, parece llenar espacios a escalas más y más finas.

Existen diferentes definiciones de esta, como la dimensión de Hausdorff-Besicovitch, la dimensión de Minkowski-Bouligand, la dimensión de empaquetado, la dimensión de homotecia y la dimensión de Rényi. Estas no pueden ser tratadas como universales, ya que las diferencias en la estructura interna del fractal, ocasionan discrepancia entre ellas. Aunque para un buen número de fractales clásicos los valores de las diferentes definiciones de dimensión fractal coinciden en dimension, en general no son equivalentes.

6.5.1. Relación entre dimensiones Fractales

Para algunas de las anteriores dimensiones fractales ha podido probarse la siguiente serie de desigualdades:

$$D_T \leq \dots D_\alpha \leq D_2 \leq D_1 \leq D_0 = D_{MB} \leq D_E \leq D_C \quad (\alpha > 2) \quad (6.1)$$

$$D_T \leq D_{HB} \leq D_{MB} \leq D_E \leq D_C \quad (6.2)$$

Donde:

- D_T : es la dimensión topológica, la cual siempre es entera.
- D_α : es la dimensión de Rényi de parámetro α .
- D_2 : es la dimensión de correlación.
- D_1 : es la dimensión de entropía o dimensión de Kolmogórov.
- D_E : es la dimensión de empaquetado.
- D_{MB} : es la dimensión de Minkowski-Bouligand o de conteo de cajas, a veces llamada dimensión de Hausdorff.
- D_{HB} : es la dimensión de Hausdorff-Besicovitch, que para los fractales clásicos suele ser un número irracional.
- D_C : es la dimensión del espacio euclídeo, que contiene al fractal, este también es un número entero.

El primero en emprender un análisis riguroso fue Cantor en su carta a Dedekind, fechada el 20 de junio de 1877. Le siguió Peano en 1890, y los pasos finales datan de la década de 1920.

Según Mandelbrot, un tratado matemático sobre la teoría de la dimensión asume no unicidad, pero lo más importante es su concepto, ya que al ampliarlo a las dimensiones, presenta diversas facetas matemáticas que, aparte de ser conceptualmente diferentes, dan distintos resultados numéricos.

Algunas aclaraciones: Edward Szpilrajn(1930)(más tarde llamado Edward Marczewski), establece un teorema en el que cada orden parcial estricto está contenido en un orden total, donde:un orden parcial estricto es una irreflexiva y transitiva relación y a su vez, un orden total es un orden parcial estricto que también es total. De acuerdo a esto establece la siguiente desigualdad: $D_T \leq D_{HB}$, la cual es uno de los principales resultados de la geometría fractal.

Para fractales de autosimilitud entre dimensiones de Rényi se cumple $D_2 \leq D_1 \leq D_0$, a todas las escalas, difieren en el caso de objetos multifractales.

Las dimensiones de Minkowski-Bouligand y Hausdorff-Besicovitch, para conjuntos cerrados se cumple. Si un conjunto es no cerrado la dimensión de Hausdorff-Besicovitch puede diferir de las otras dos, por ejemplo el conjunto de números racionales del intervalo $[0,1]$ tiene $D_{HB} = 0$, pero en cambio tiene $D_0 = D_{MB} = 1$.

Si los fractales están definidos formalmente, como los mencionados anteriormente, puede calcularse su dimensión fractal. No obstante, algunos fenómenos u objetos de la vida real pueden mostrar propiedades fractales, y es aquí donde puede ser útil obtener la dimensión fractal de un conjunto de datos de una muestra. Este cálculo no se puede obtener de forma exacta sino que debe estimarse.

Esto se usa en una variedad de áreas de investigación tales como la física, análisis de imagen, acústica, ceros de la función zeta de Riemann é incluso procesos electroquímicos.

En la práctica estas dimensiones fractales son muy sensibles al ruido numérico o experimental, y también a las limitaciones en la cantidad de datos. Las tesis basadas en estimaciones de dimensiones fractales deben tomarse con cuidado ya que hay un límite superior inevitable, a menos que se presenten cantidades muy grandes de datos.

Los más sencillos de implementar son, el conteo por cajas (box counting) y la dimensión de correlación (basada en generar un número de puntos aleatorios en un entorno del fractal y medir cuántos de ellos caen sobre el conjunto fractal). Una técnica que se ha hecho popular es utilizar la transformada de Fourier, para mediciones como el espectro de potencia o señales de comunicación.

El concepto de una dimensión fractal se basa en puntos de vista no convencionales de escala y dimensión. Las nociones tradicionales de geometría dan forma a la escala previsible de acuerdo a las ideas intuitivas y familiares sobre el espacio en el que están contenidas, de manera que, por ejemplo, la medición de una línea, usándola como una medida unitaria y luego tomar $\frac{1}{3}$ de su tamaño, dará para la primer medida un total de longitud 3 veces la cantidad de medidas que el particionado. Esto se mantiene en 2 dimensiones, también. Si se mide el área de un cuadrado a continuación, las medidas de uno nuevo, con una caja de longitud de lado $\frac{1}{3}$ del tamaño del original, se encontrará que es 9 veces el número de plazas que con la primera medida. Tales relaciones de escala familiares pueden definirse matemáticamente por la regla de escala general, donde la variable N , es el número de particiones, ϵ es el factor de escala, y D es la dimensión fractal:

$$N \propto \epsilon^{-D} \quad (6.3)$$

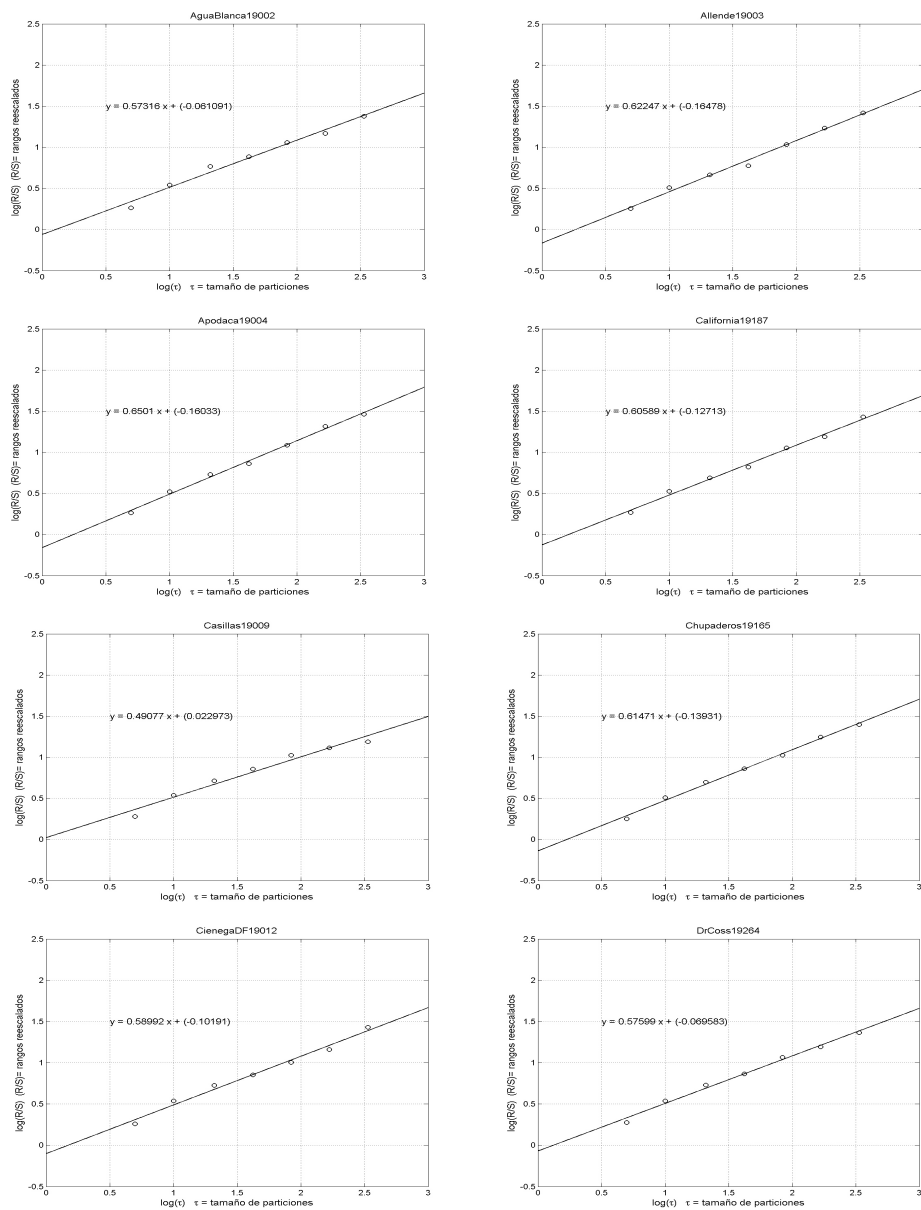
Esta regla de escala tipifica normas convencionales sobre la geometría y la forma para encontrar la dimensión de un fractal.

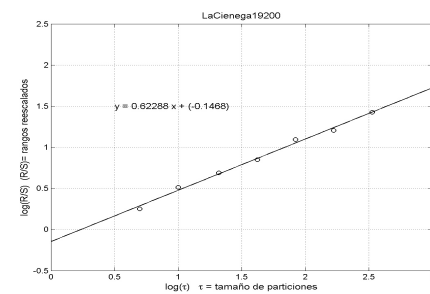
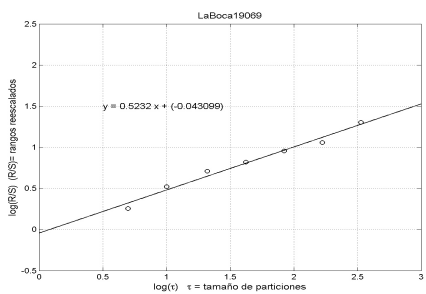
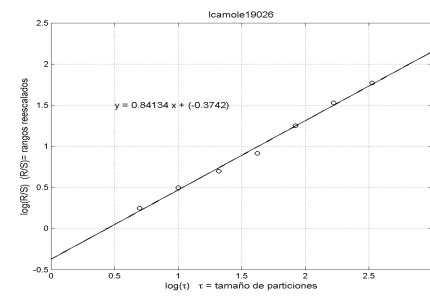
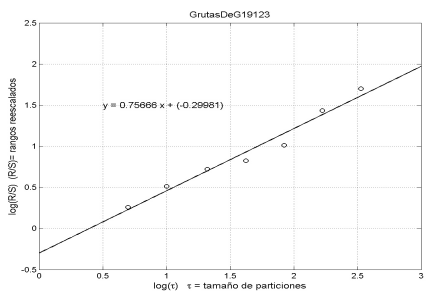
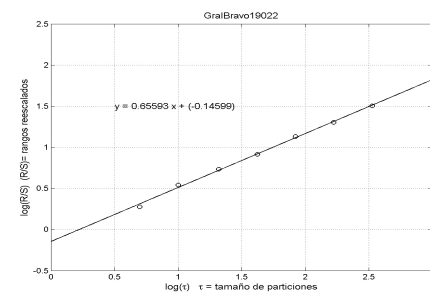
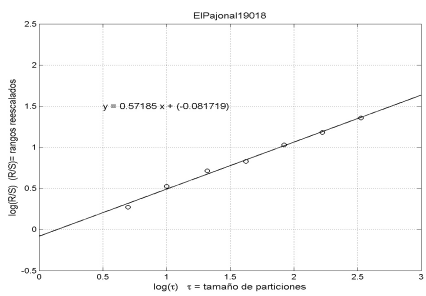
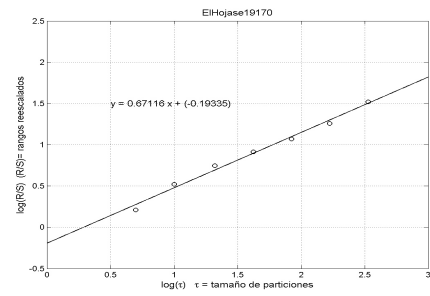
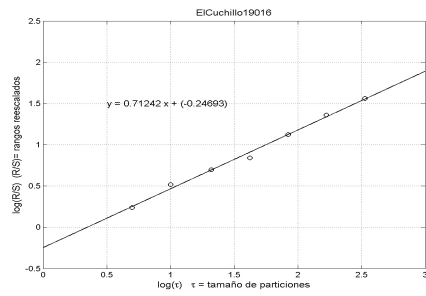
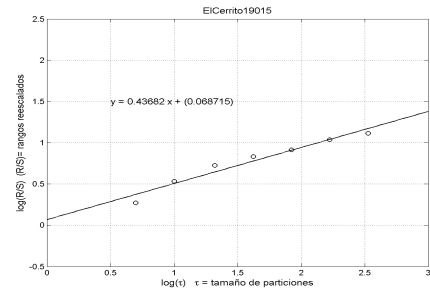
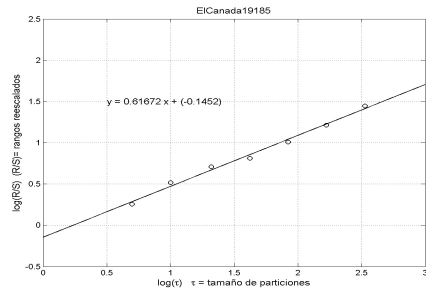
El concepto de dimensión fractal descrito es una vista básica de una construcción complicada.

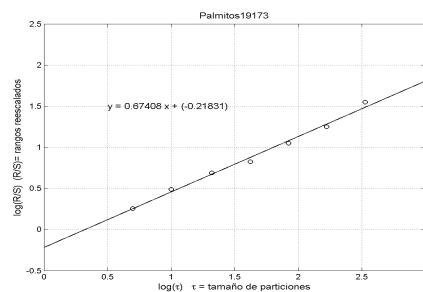
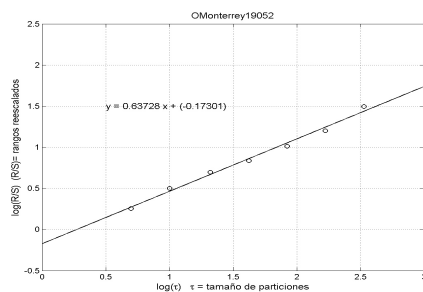
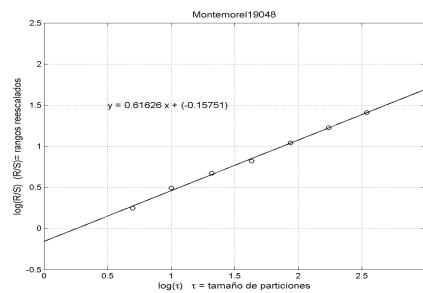
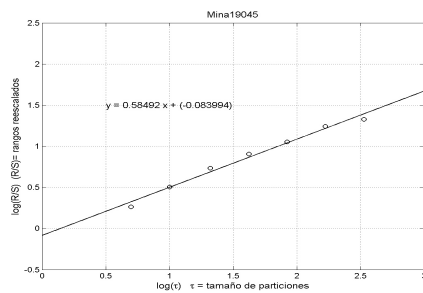
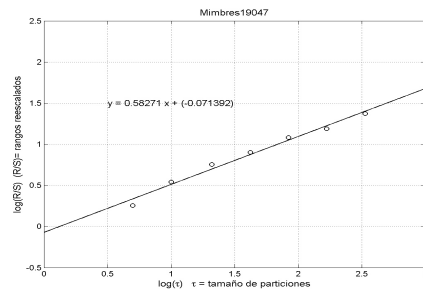
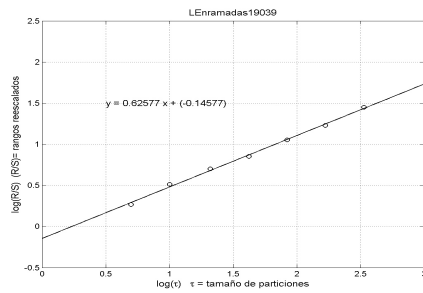
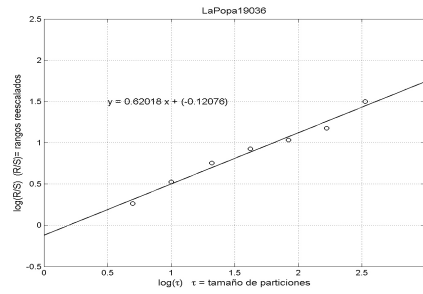
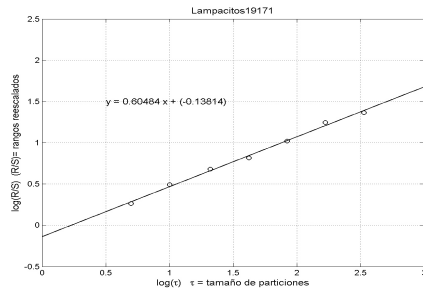
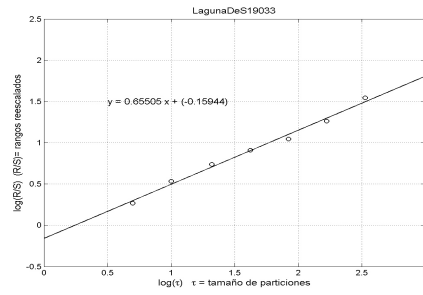
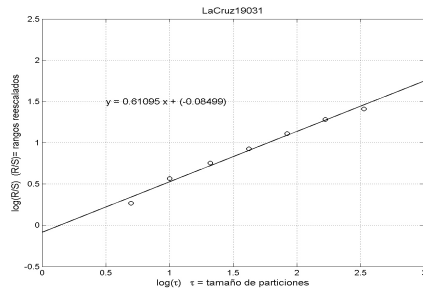
Una validación también se puede hacer mediante la comparación de distintas propiedades fractales, la cuales pueden estar implícitas en el modelo que utilicemos, y también acorde a los datos medidos. Incluso el comportamiento de nuestros datos, quizá este siendo impulsado por una compleja interacción entre la agregación y coalescencia donde dos o más elementos pueden acompañarse uniendo sus dimensionalidades en un solo conjunto.

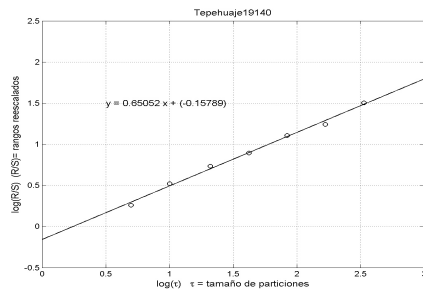
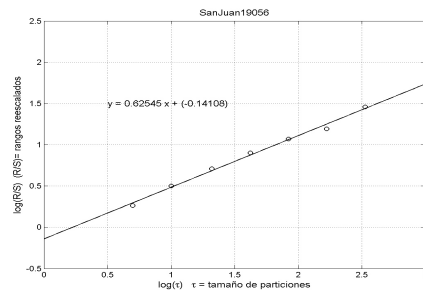
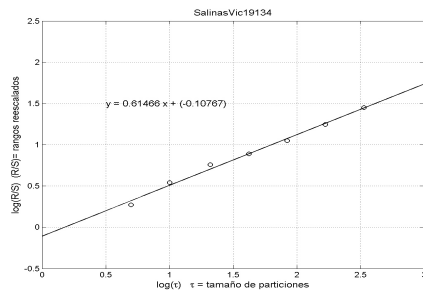
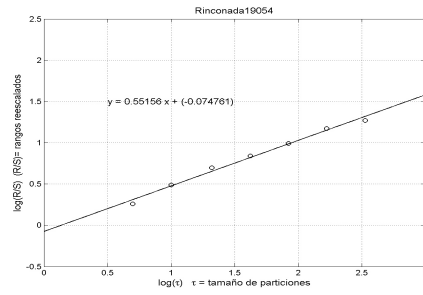
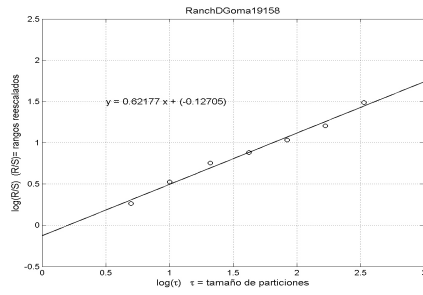
6.6. Gráficos del exponente de Hurst

En este anexo se muestran las gráficas de las 33 rectas de mínimos cuadrados de las estaciones pluviométricas analizadas, la pendiente en cada ecuación es el denominado exponente fractal de Hurst.









6.7. Variogramas y sus Gráficos

En este anexo se detalla la utilización de los variogramas y se muestran los gráficos de los 33 variogramas de las estaciones pluviométricas analizadas.

El variograma, $\gamma(h)$, es un ajuste o modelado espacial considerado como un estimador de la varianza poblacional y de análisis estructural, donde la población debe tener una tendencia de estacionalidad, se utiliza para describir la relación de observaciones pareadas separadas por una distancia h y en otros casos con una dirección.

Es una técnica geoestadística, la cual permite una medida cuantitativa de la persistencia a largo plazo en series de tiempo no estacionarias Witt[16], Haslett[17], Dmowska[18] et al. Establece correlaciones a través del tiempo y un fenómeno regionalizado en el espacio, generando patrones que pueden ser utilizados para describir el comportamiento de un conjunto de observaciones.

Matemáticamente, el variograma estima la diferencia cuadrada prevista entre variables aleatorias vecinas, dando un soporte fundamental y permitiendo representar cuantitativamente esta relación. Este proceso continúa para cada punto de medición.

Teniendo en cuenta una serie de tiempo o procesos estocásticos $\{X_t, t \geq 0\}$, la función de autocovarianza en el punto $(t, t+h)$ se define como $C_X(t, t+h) = E[X_t X_{t+h}] - E[X_t]E[X_{t+h}]$ con $E[X_t]$ la media del proceso en tiempo t .

El variograma $\gamma(h)$ está dada por la mitad de la varianza de la diferencia entre pares de observaciones en diferentes “puntos” en el tiempo,

$$\begin{aligned} \gamma(h) &= \frac{1}{2} \text{Var}(X_{t+h} - X_t) = \frac{1}{2} E[(X_{t+h} - X_t - E[X_{t+h} - X_t])^2] \\ &= \frac{\text{Var}(X_t) + \text{Var}(X_{t+h})}{2} - \text{Cov}(X_t, X_{t+h}) \\ &= \frac{\text{Var}(X_t) + \text{Var}(X_{t+h})}{2} - \sqrt{\text{Var}(X_t)\text{Var}(X_{t+h})}\rho_X(t, t+h) \end{aligned} \quad (6.4)$$

donde $-1 \leq \rho_X(t, t+h) \leq 1$ es la función de autocorrelación (la función de autocovarianza normalizada).

En el caso especial cuando $\rho(x_{t+h}, x_t) = 0, \forall(t, h)$, se dice que los procesos estocásticos $\{X_t, t \geq 0\}$ no están correlacionados y el semivariogram se reduce a la media aritmética de la varianza entre los procesos t and $t+h$,

$$\gamma(h) = \frac{Var(X_t) + Var(X_{t+h})}{2} \quad (6.5)$$

Si el campo aleatorio $\{X_t, t \geq 0\}$ tiene media constante, $E[X_t] = E[X_{t+h}] = \mu \quad \forall t$, el variograma (6.4), adopta la forma simple,

$$\gamma(h) = \frac{1}{2}E[(X_{t+h} - X_t)^2] \quad (6.6)$$

Si $X(t)$ y $X(t+h)$ son variables aleatorias independientes $\forall t$, de nuevo, aunque por una razón diferente, el variograma se reduce al caso especial (6.5).

En principio, dado un proceso estocástico $\{X_t, t \geq 0\}$, el valor esperado de las diferencias $E[X_{t+h} - X_t]$, en el momento t y con lag h , se estima empíricamente por el promedio sobre un “suficientemente grande” ensamble de realizaciones o trayectorias en el tiempo.

Sin embargo, para una sola serie de tiempo, $\{X_n, n = 1, 2, \dots, n\}$, se espera que el valor puede ser estimado suponiendo una hipótesis de ergodicidad, i.e., es decir, un principio estadístico de equivalencia según la cual “*el promedio a través del tiempo y el promedio a través del ensamble son los mismos*” Lefevbre[19].

Así las diferencias $X_{t+h} - X_t$, que se obtendría con un proceso infinitamente reproducible, son “simulados” ó “clonados” de la “serie madre”.

Por lo tanto, el valor medio de las diferencias $X_{t+h} - X_t$ se estima por,

$$E[X_{t+h} - X_t] = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (x_{i+h} - x_i) \quad (6.7)$$

donde $n(h)$ es el número de diferencias con un lag h . Para $h = 1, 2, 3, \dots$ los promedios (6.7) son, respectivamente,

$$\begin{aligned} E[X_{t+1} - X_t] &= \frac{x_n - x_1}{n - 1} \\ E[X_{t+2} - X_t] &= \frac{x_{n-1} + x_n - (x_1 + x_2)}{n - 2} \\ E[X_{t+3} - X_t] &= \frac{x_{n-2} + x_{n-1} + x_n - (x_1 + x_2 + x_3)}{n - 3} \\ &\vdots \\ E[X_{t+h} - X_t] &= \frac{1}{n(1 - \frac{h}{n})} \left(\sum_{j=0}^{k-1} x_{n-j} - \sum_{l=1}^k x_l \right) \end{aligned} \quad (6.8)$$

De acuerdo a (6.8) para un valor máximo de h “relativamente moderado” ó $h/n < 1$, excepto en la presencia de valores extremos aislados, las dos sumas en (6.8) son aproximadamente del mismo orden, de manera que el valor medio empírico (6.7) puede ser aproximado por, $E[X_{t+h}] \approx E[X_t] = m = \text{constante}$. Esta es una característica observada en las series de tiempo de las estaciones pluviométricas.

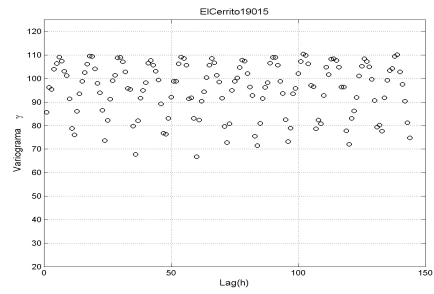
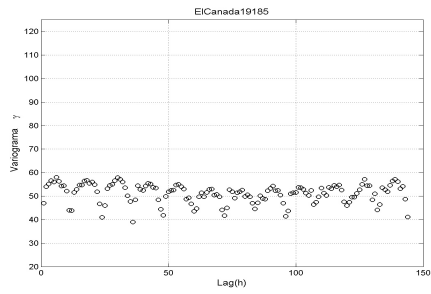
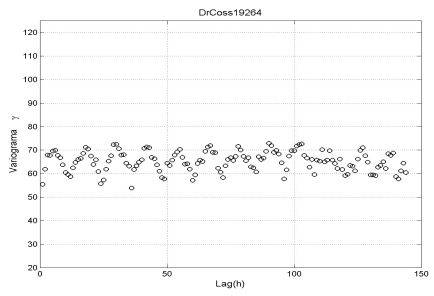
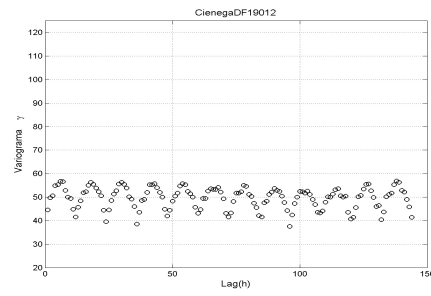
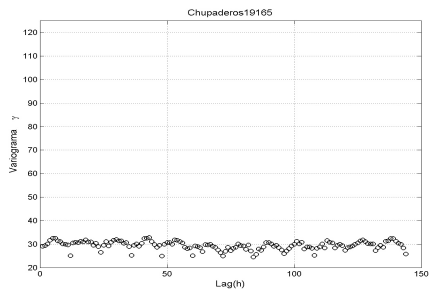
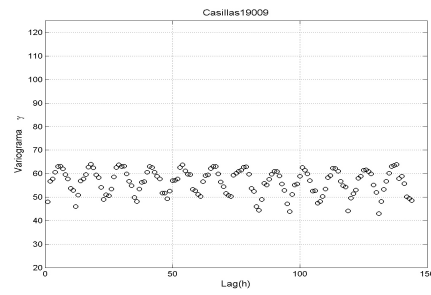
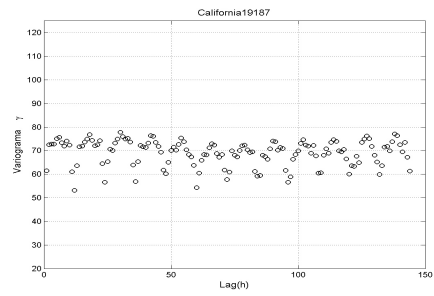
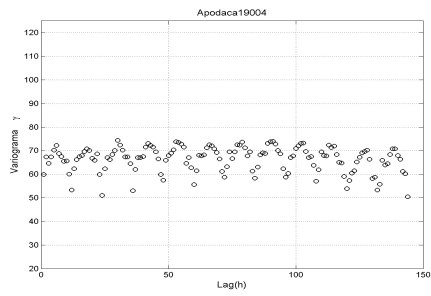
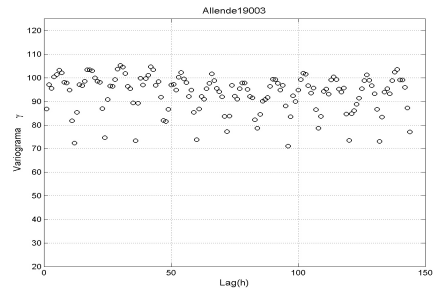
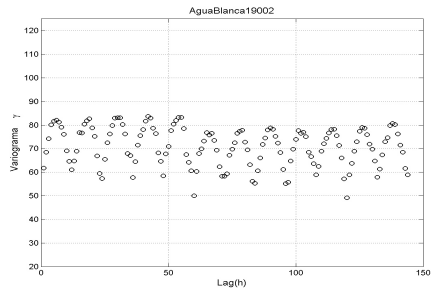
Por lo tanto, el correspondiente estimador de (6.6) es simplemente:

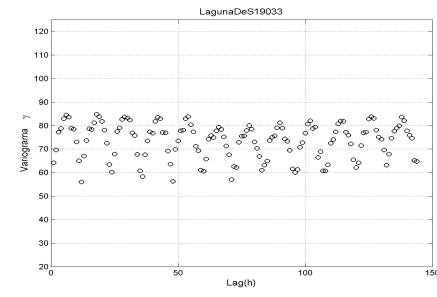
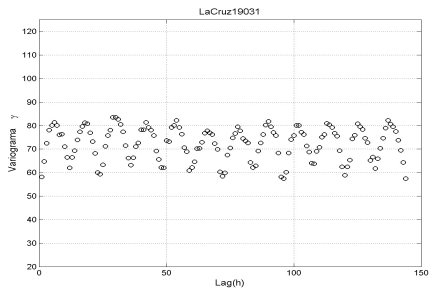
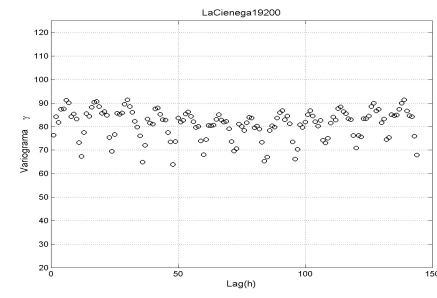
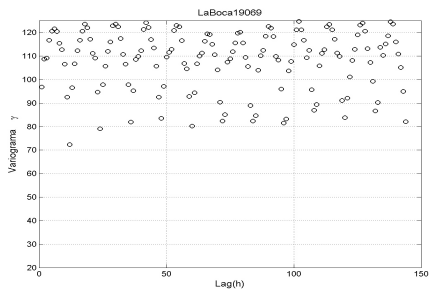
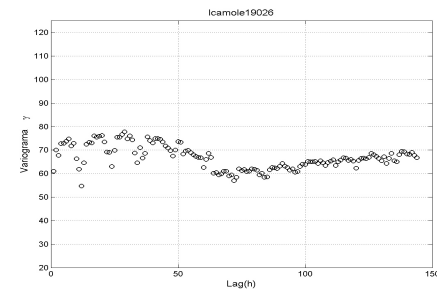
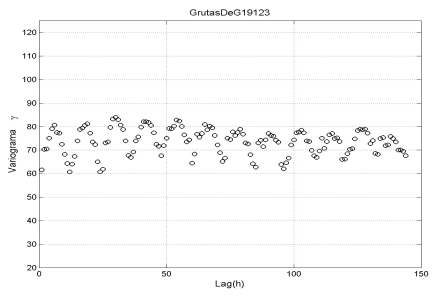
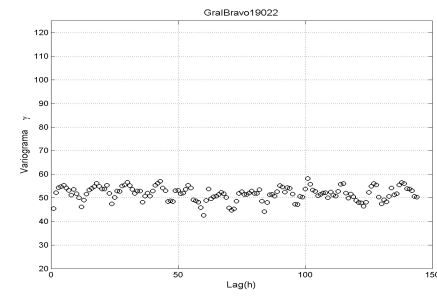
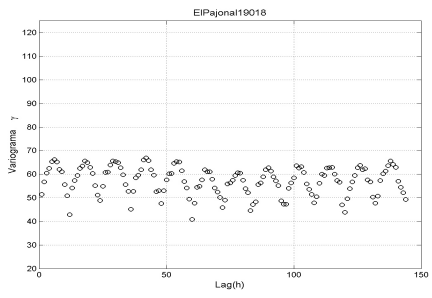
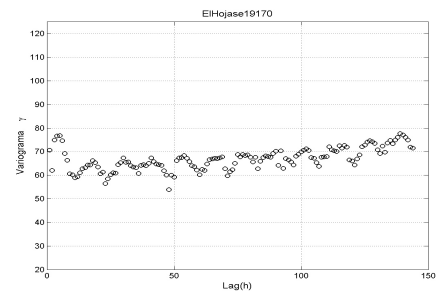
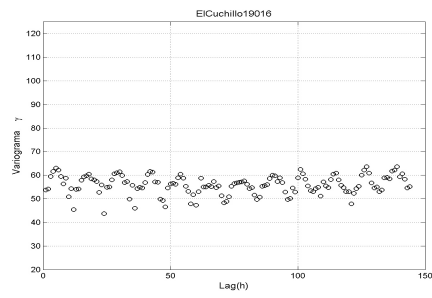
$$\gamma(h) = \frac{1}{2n(h)} \sum_{t=1}^{n(h)} (x_{t+h} - x_t)^2 \quad (6.9)$$

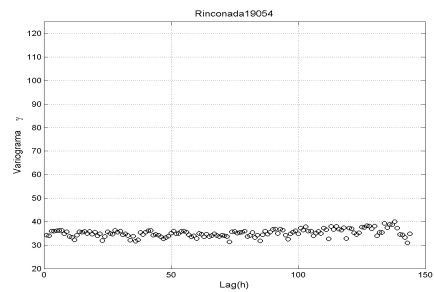
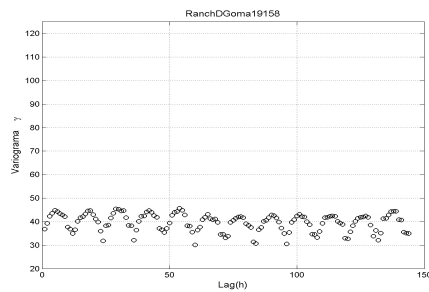
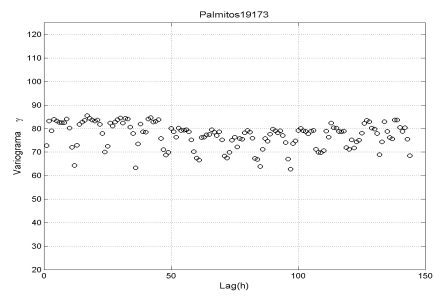
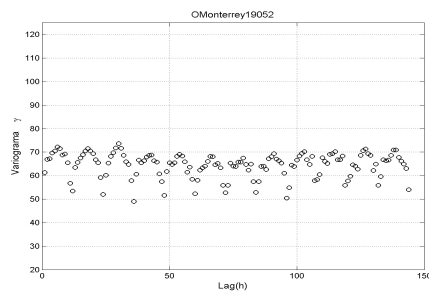
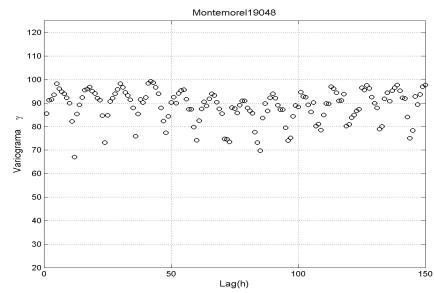
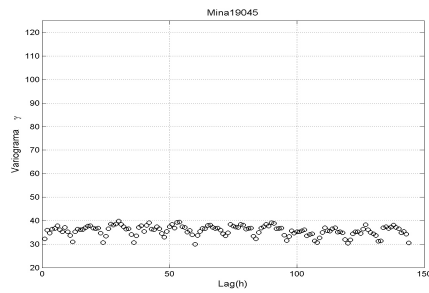
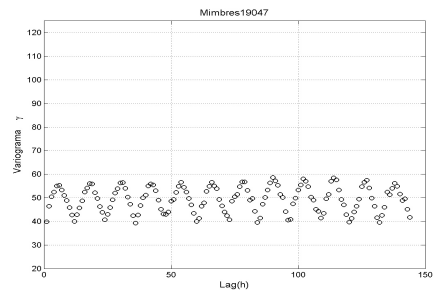
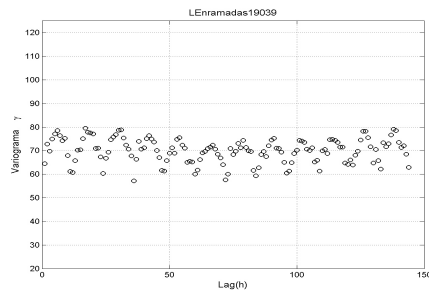
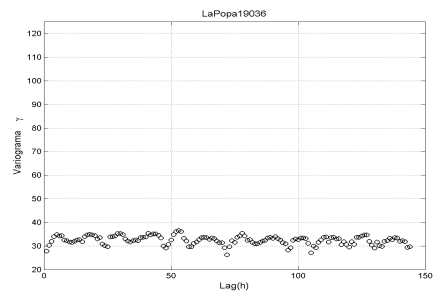
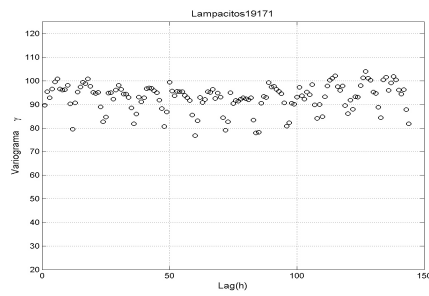
Este estimador de momentos, es un promedio de diferencias al cuadrado, que puede ser influenciado por un número pequeño de valores que ocasionan discrepancias al final de los cálculos debido a las particiones realizadas. Pero, se considera un estimador robusto, ya que disminuye la importancia de las diferencias grandes y al cuadrado.

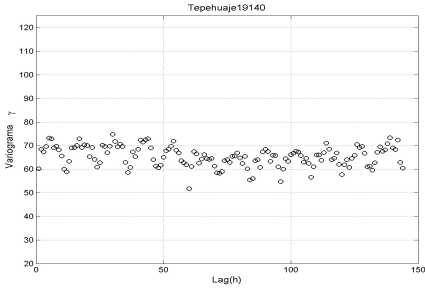
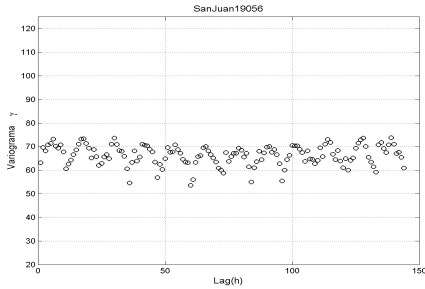
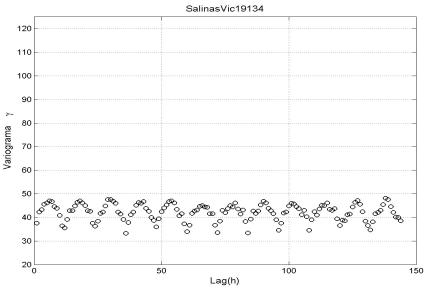
También se considera robusto en el sentido de que es resistente a distribuciones normales contaminadas y afloramientos posiblemente generados por distribuciones de colas pesadas. Esto puede verse por el uso de la raíz cuadrada de las diferencias, en lugar de diferencias de cuadrados, en el estimador insesgado.

La aplicación de variogramas a nuestros datos mostró periodos largos de comportamiento similar, con diferente duración, un comportamiento o patrón cíclico, pero no periódico.



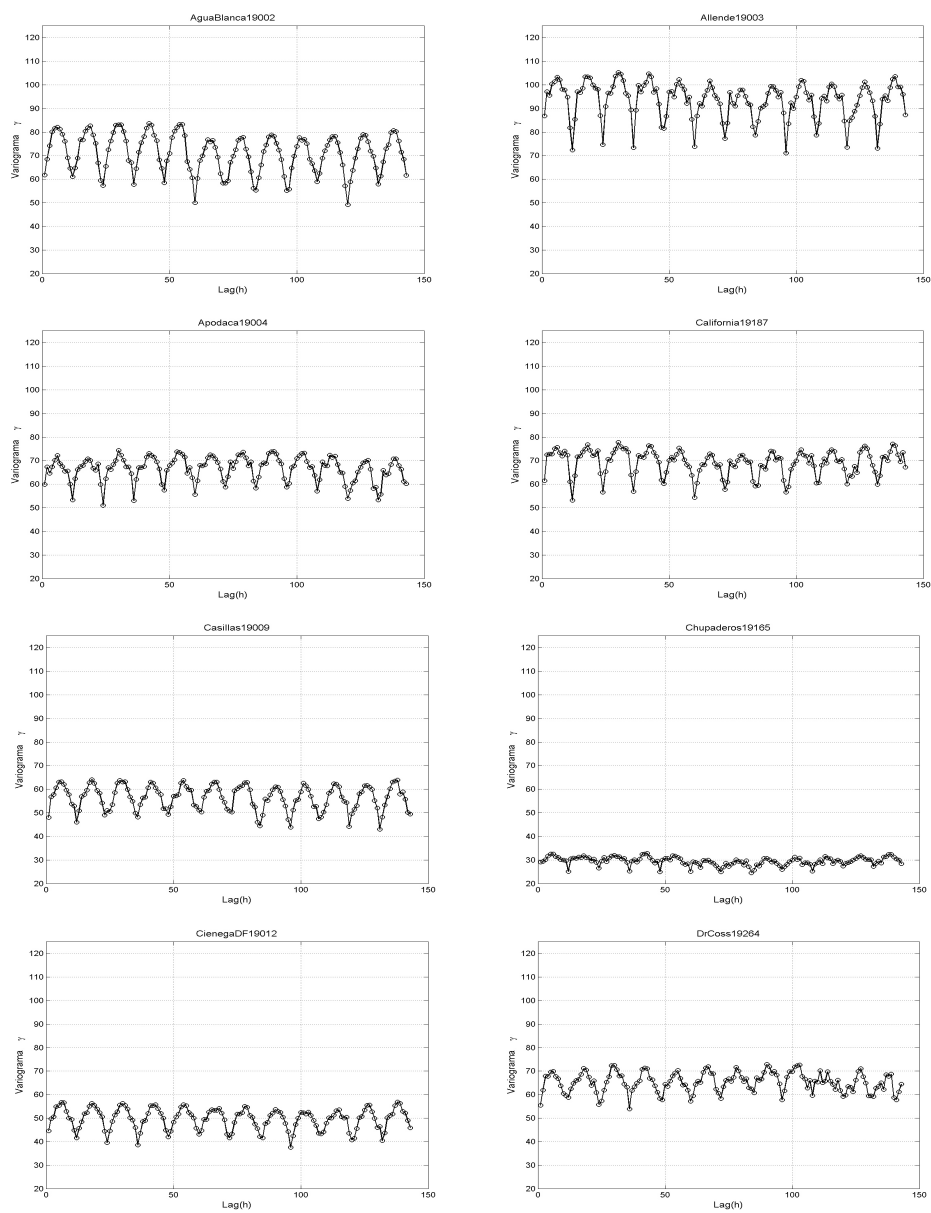


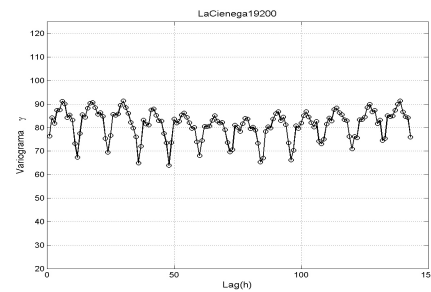
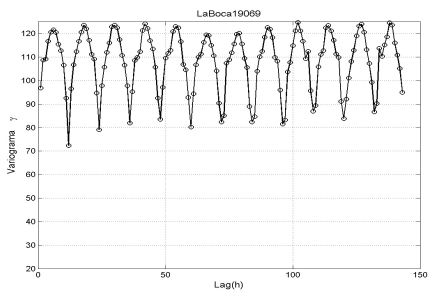
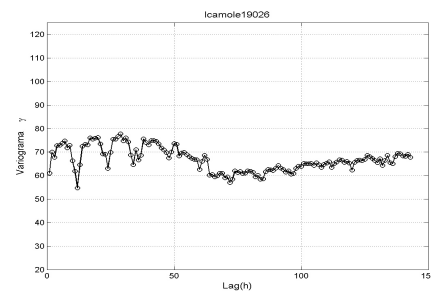
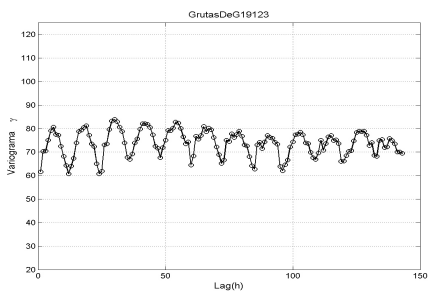
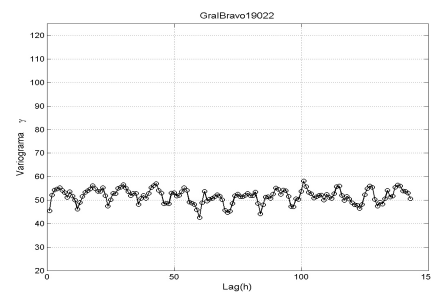
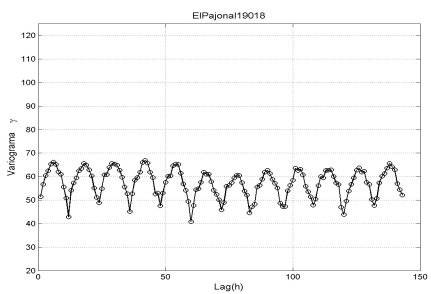
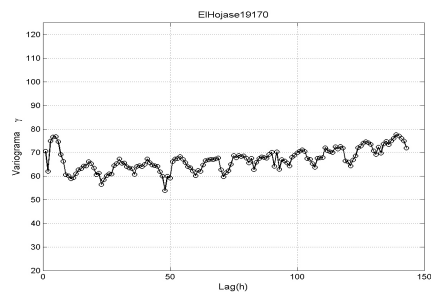
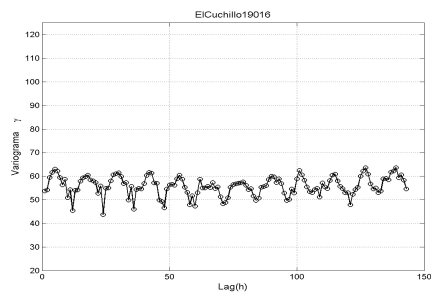
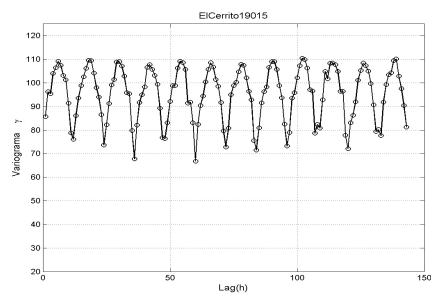
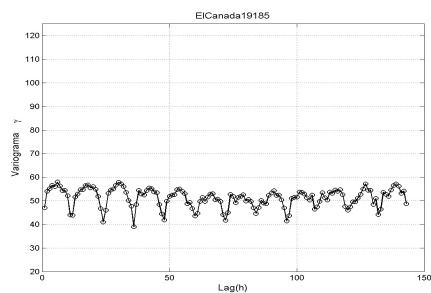


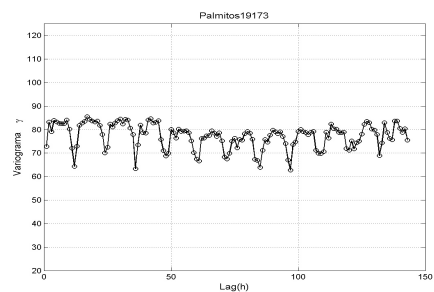
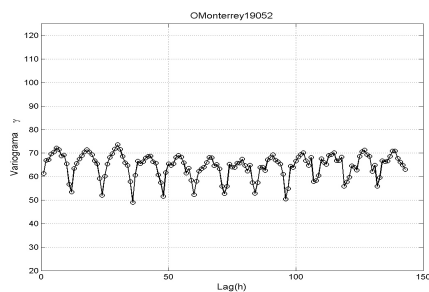
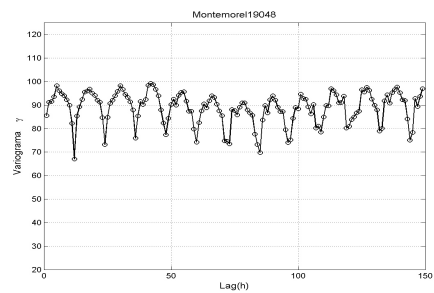
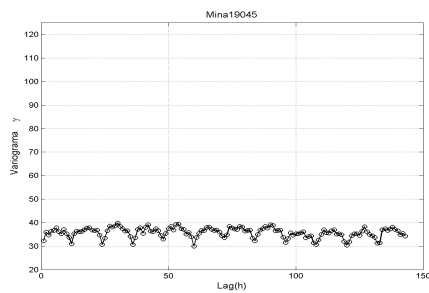
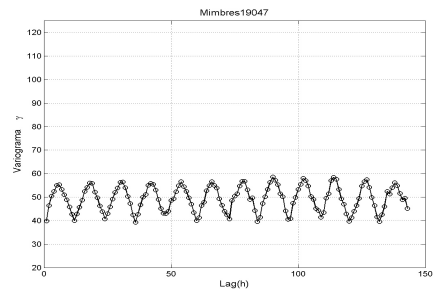
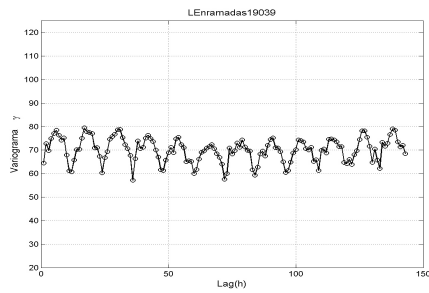
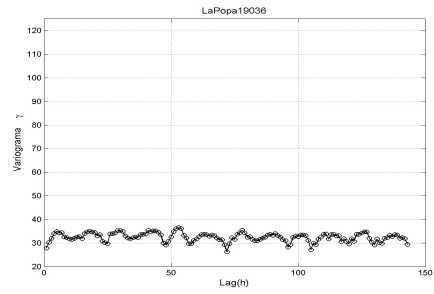
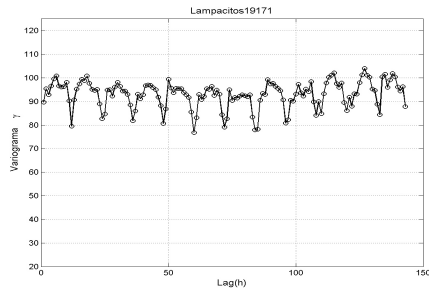
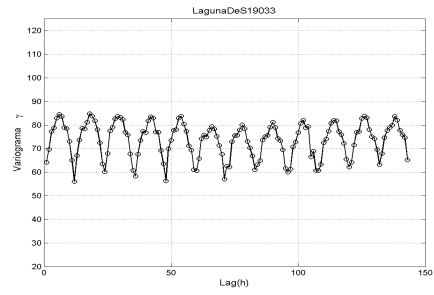
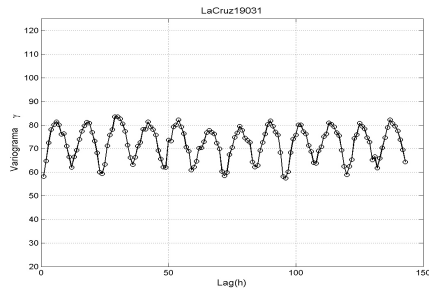


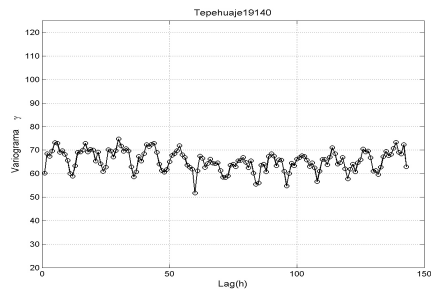
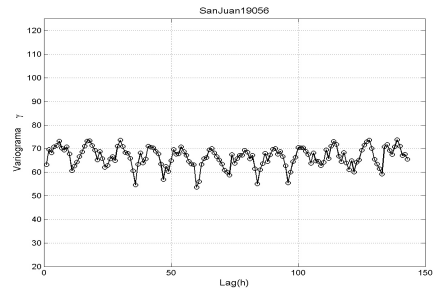
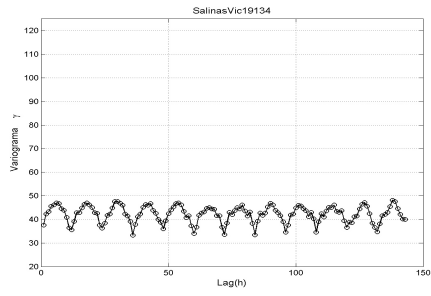
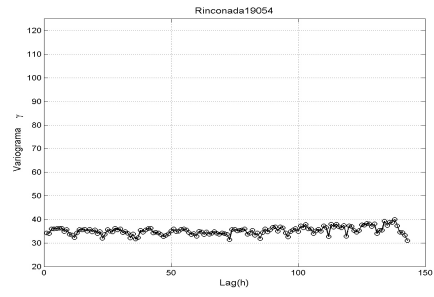
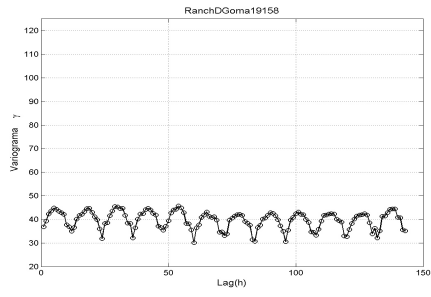
6.8. Gráficos de variogramas enlazados

En este anexo se muestran los gráficos de los 33 variogramas del anexo anterior, enlazando cada uno de los valores calculados de cada estación pluviométrica analizadas.









6.9. L-Momentos

Los datos disponibles a la hora de regionalizar se reducen a las observaciones registradas en las estaciones y a una serie de descriptores que tienen en cuenta sus características físicas (localización geográfica, altitud, longitud). Esta distinción se considera importante por Hosking y Wallis[5] ya que estos basan la definición de las diferentes regiones en los descriptores físicos, reservando las observaciones para testear la homogeneidad de la regionalización propuesta.

Diversos autores han establecido criterios alternativos para agrupar estaciones. La delimitación según áreas administrativas (McKerchar[20]) constituye una de las opciones, si bien a menudo resulta arbitraria y carente de justificación física. En ciertos estudios, sobretudo a pequeña escala, regionalizar por simple inspección subjetiva de las características físicas de las estaciones puede arrojar buenos resultados, tal y como muestran otros trabajos como los de Schaefer[21] para las precipitaciones máximas anuales en Washington.

El éxito depende de la destreza a la hora de seleccionar y comparar los descriptores físicos más relevantes. Por otro lado, métodos de división basados en criterios objetivos también son posibles; la asignación de una estación en una región u otra en función de si supera o no un umbral establecido para un determinado parámetro físico, o los procesos iterativos hasta conseguir un número de regiones aceptablemente homogéneas, constituyen dos de los casos más habituales.

No obstante, Hosking y Wallis[5] consideran el análisis de clúster como la alternativa más apropiada. Se trata de un método estándar de análisis multivariante por el que se asocia un vector a cada una de las estaciones, que son divididas y agrupadas atendiendo precisamente a la similitud entre vectores.

Los clústers se forman por agrupación de sitios cuyas características físicas son parecidas, y la mayoría de los algoritmos de clustering miden la similitud de dichas características atendiendo a sus distancias euclídeas en el espacio. Al verse afectadas estas distancias por la escala de medida, es habitual estandarizar cada variable física con el objetivo de conseguir un mismo escenario de dispersión para todas ellas.

Sin embargo, es de esperar que ciertas variables condicionen más que otras la forma de la función de distribución, por lo que, pese a la dificultad que ello conlleva, se aconseja ponderar y otorgar diferentes pesos a cada una de ellas. No hay que pasar por alto la idea de que no existe un número de clusters “correcto”, sino que es el área estudiada la que marca la necesidad de agrupar en más ó menos regiones.

En este sentido, es evidente que clústers con pocas estaciones corren el riesgo de mejorar muy poco la precisión en las estimaciones con respecto a los métodos de análisis local, mientras que regiones de gran tamaño pueden vulnerar el criterio de homogeneidad.

Por ello, la elección de algoritmos que tienden a formar clusters de tamaño y variabilidad parecida, como el método de Ward o el enlazado promedio, parece la opción más razonable.

Se destaca además otro aspecto importante. El resultado que arroja un análisis de clúster, no tiene por qué ser definitivo. Son muchos los ajustes de tipo subjetivo que pueden ser empleados para mejorar la coherencia física del reparto e incluso reducir la heterogeneidad de ciertas regiones.

Entre otros:

- Mover estaciones de una región a otra;
- Suprimir o prescindir de ciertas estaciones;
- Subdividir regiones resultantes;
- Romper regiones mediante el traslado de estaciones de un clúster a otro;
- Combinar regiones entre sí; y
- Obtener nuevos datos y redefinir la regionalización.

Establecidas y ajustadas las diferentes regiones, llega el momento de emplear las observaciones registradas en las estaciones para comprobar la hipótesis de homogeneidad.

Diferentes tipos de tests han sido propuestos por Dalrymple[4], Acreman y Sinclair[22], Wiltshire (1986 a,b), Buishand(1989), Chowdhury et al.(1991), Lu y Stedinger (1992), Hosking y Wallis[5], y Fill y Stedinger(1995). La mayoría de ellos implican una cantidad δ que mide un determinado aspecto de la distribución de frecuencias y que es constante para una región homogénea: δ puede ser, por ejemplo, el evento de 10 años escalado por la media (casos de Dalrymple[4]; Lu y Stedinger, 1992; Fill y Stedinger, 1995), el coeficiente de variación (Wiltshire, 1986 a), una combinación de $L - C_v$ y L-asimetría (Chowdhury et al., 1991), o el $L - C_v$ o la combinación de $L - C_v$, L-asimetría y L-curtosis (Hosking y Wallis, 1993).

Se calculan las estimaciones de $\delta^{(i)}$ donde δ es la estimación local de la estación i basada en sus observaciones, y $\delta^{(R)}$ que es la estimación regional calculada a partir de los datos de todas las estaciones de la región asumiendo homogeneidad. Se construye entonces un estadístico S que mide la diferencia entre las estimaciones locales y la estimación regional; una posible elección es:

$$S = \sum_{i=1}^N (\widehat{\delta}^{(i)} - \widehat{\delta}^{(R)})^2 \quad (6.10)$$

El valor observado de S es comparado con la distribución que S tendría si la región fuera verdaderamente homogénea, y este cálculo a menudo implica el asumir una determinada forma para la distribución de frecuencias de la región: desde la Gumbel empleada por Dalrymple[4] y Fill y Stedinger(1995), por ejemplo, hasta la General de Valores Extremos asumida por Chowdhury et al. (1991).

Si el valor observado de S queda en la cola de la distribución, se rechaza la hipótesis de homogeneidad por considerarse poco probable que un valor tan extremo de S pueda ser debido al azar.

A pesar de las diferentes alternativas para proceder, en el contexto de un análisis regional de frecuencias, Hosking y Wallis[5] recomiendan el empleo de tests basados en los L-momentos.

Una vez definidas las regiones homogéneas, se procede a determinar la función de distribución más apropiada para cada una de ellas. De forma general podemos decir que, la homogeneidad total es imposible ya que toda región es siempre, ligeramente heterogénea, y no existe una distribución “verdadera” que la pueda caracterizar perfectamente. Por tanto, el objetivo se centra en encontrar aquélla que proporciona las estimaciones más precisas para cada estación.

Es importante remarcar que la distribución elegida no tiene por qué ser la que más se aproxime a las observaciones. El que exista una función con buen ajuste a los datos no garantiza que, en ella, los valores futuros de la variable vayan a ser coherentes con los del pasado, pues a menudo derivan de procesos físicos propensos a originar outliers alejados del resto.

En este sentido, es preferible anteponer la elección de un modelo robusto en estimaciones a la selección de la distribución que mejor ajusta. Por lo general, existe un rango de periodos de retorno de interés para los cuales se requiere estimar cuantiles.

En el estudio de eventos extremos, como precipitaciones, avenidas o sequías, son de particular relevancia las estimaciones que recaen en la cola de la distribución de frecuencia. Ésta y otras muchas consideraciones han de ser tomadas en cuenta a la hora de seleccionar.

Lo anterior nos hace decir que, el enfoque de análisis regional de frecuencia es eficaz en ampliar la información sobre un sitio o en sitios dentro de una región homogénea.

6.9.1. Estimación de parámetros

En general en muchas áreas no solo de ingeniería los problemas se analizan y sintetizan a través del uso de modelos matemáticos. Estos últimos pueden ser del tipo determinístico, paramétrico o estadístico (o bien estocástico).

Un modelo completamente determinístico es aquel que se obtiene a través de relaciones físicas y no requiere de datos experimentales para su aplicación, es un modelo matemático donde las mismas entradas producirán invariablemente las mismas salidas, no contemplándose la existencia del azar ni el principio de incertidumbre.

Un modelo paramétrico puede ser considerado como un determinístico en el sentido de que una vez que se estiman los parámetros del modelo, este siempre genera la misma salida a partir de la información de entrada. Por otro lado, un modelo paramétrico es estadístico en el sentido de que los parámetros estimados dependen de los datos observados y aquellos cambiarán cuando los datos observados también lo hagan.

Un modelo estadístico es aquel en el cual las salidas son predecibles en el sentido estadístico. En este modelo, dado un grupo de entradas que se emplea repetidamente genera salidas que no son las mismas pero siguen cierto patrón (Escalante Sandoval y Reyes Chávez, 2005).

Todo modelo estocástico es por naturaleza estadístico. Sin embargo, suele emplearse la denominación “estocástico” cuando existe una dependencia de las variables que intervienen con el tiempo (Clarke, 1988).

La estimación de los parámetros es importante antes de hacer inferencias de cualquier modelo. Cada estimador de un parámetro es una función definida sobre valores numéricos de la muestra, las cuales caracterizan una población o modelo. Así, el propio parámetro estimado es una variable aleatoria que tiene su propia distribución muestral. Un estimador que se obtiene a partir de un grupo de valores puede considerarse como un valor observado de una variable aleatoria. Por lo cual, la bondad de un estimador puede ser juzgada a partir de su distribución.

Independientemente de la técnica que se use para la estimación de los parámetros se deben cumplir las siguientes propiedades:

Sesgo nulo o insesgado: Un estimador $\hat{\theta}$ de un parámetro θ se dice que tiene sesgo nulo cuando $E(\hat{\theta}) = \theta$. De lo contrario diremos que es sesgado. El sesgo se obtiene como $B = E(\hat{\theta}) - \theta$, que es la esperanza de $\hat{\theta}$.

Eficiencia: Un estimador $\hat{\theta}$ es más eficiente (más preciso) que otro estimador, si la varianza del primero es menor que la del segundo. Esto es si $\hat{\theta}_1$ y $\hat{\theta}_2$ son ambos estimadores de θ . Si $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$. Un estimador es más eficiente, cuanto menor es su varianza. Generalmente es posible obtener más de un estimador no sesgado para el mismo parámetro θ . En ocasiones también es utilizada la denominada eficiencia relativa de la siguiente forma, $\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$.

Consistencia: Un estimador $\widehat{\theta}$ de un parámetro θ se dice que es consistente si para cualquier número positivo ε , el $\lim_{n \rightarrow \infty} P(|\widehat{\theta} - \theta| > \varepsilon) = 0$, n es el tamaño de muestra.

Suficiencia: $\widehat{\theta}$ es un estimador suficiente para θ , si $\widehat{\theta}$ emplea toda la información relevante contenida en la muestra.

Los L-momentos se definen para distribuciones de probabilidad, pero en la práctica para una muestra de datos finita generalmente se obtiene una estimación. El cálculo de los L-momentos se obtiene de una muestra de tamaño n ordenada ascendentemente (Landwehr et al., 1979) (esto es por el hecho de que al ordenar los datos y establecer las formulas, los datos más pequeños no afecten tanto el cálculo numérico dentro de las estimaciones).

Los L-momentos son análogos a los momentos convencionales, sin embargo, tienen cierta ventaja sobre ellos, ya que son capaces de caracterizar a un mayor número de distribuciones, además de estar virtualmente libres de sesgo aún para muestras pequeñas (Hosking, 1990, Escalante Sandoval y Reyes Chávez, 2005).

Sea $x_{i:n}$ la i -ésima observación en una muestra de tamaño n , ordenada de mayor a menor. Para cualquier distribución de probabilidad se define de la siguiente manera los L-momentos:

$$\lambda_n = \frac{1}{n} \sum_{j=0}^{n-1} (-1)^j \binom{n-1}{j} E[X_{n-j:n}] \quad n = 1, 2, \dots \quad (6.11)$$

De esta forma:

$$\lambda_1 = \frac{1}{1} \sum_{j=0}^0 (-1)^j \binom{0}{j} E[X_{1-j:1}] = E[X_{1:1}] = E[X] \quad (6.12)$$

Que sería la media.

De la misma manera para cualesquier distribución de probabilidad, el segundo L-momento es una descripción de escala basada en la diferencia esperada entre dos observaciones seleccionadas de forma aleatoria:

$$\lambda_2 = \frac{1}{2} \sum_{j=0}^1 (-1)^j \binom{1}{j} E[X_{2-j:2}] = \frac{1}{2} (E[X_{2:2}] - E[X_{1:2}]) = \frac{1}{2} E[X_{2:2} - X_{1:2}] \quad (6.13)$$

Similarmente los L-momentos, tres y cuatro quedarían definidos de la siguiente forma:

$$\begin{aligned} \lambda_3 &= \frac{1}{3} \sum_{j=0}^2 (-1)^j \binom{2}{j} E[X_{3-j:3}] = \frac{1}{3} (E[X_{3:3}] - 2E[X_{2:3}] + E[X_{1:3}]) \\ &= \frac{1}{3} E[X_{3:3} - 2X_{2:3} + X_{1:3}] \end{aligned} \quad (6.14)$$

$$\begin{aligned} \lambda_4 &= \frac{1}{4} \sum_{j=0}^3 (-1)^j \binom{3}{j} E[X_{4-j:4}] = \frac{1}{4} (E[X_{4:4}] - 3E[X_{3:4}] + 3E[X_{2:4}] - E[X_{1:4}]) \\ &= \frac{1}{4} E[X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}] \end{aligned} \quad (6.15)$$

La L en “L-momentos” hace hincapié en si λ_r es una función lineal de los momentos esperados de orden estadístico. Más aún, un estimador natural de λ_r basado en un muestra observada de datos es una combinación lineal de los valores de datos ordenados, es decir, L-estadística. La esperanza matemática también la podemos escribir como:

$$E[X_{j:r}] = \frac{r!}{(j-1)!(r-j)!} \int x(F) \{F(x)\}^{j-1} \{1-F(x)\}^{r-j} dF(x) \quad (6.16)$$

Sustituyendo esta expresión en la definición de los L-momentos λ_r , expandiendo los binomios en $F(x)$ y sumando los coeficientes de cada potencia de $F(x)$ se puede obtener la siguiente expresión de L-momentos(David (1981), p.33).

$$\lambda_n = \int_0^1 x(F) P_n^* \quad n = 1, 2, \dots \quad (6.17)$$

En donde $P_n^*(F) = \sum_{k=1}^n p_{n,k}^* F^k$ con, $p_{n,k}^* = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k}$

$P_n^*(F)$ son los polinomios de Legendre trasladados los cuales se pueden obtener de fórmulas recursivas como:

$$\begin{aligned} P_0^*(F) &= \sum_{k=0}^0 p_{0,k}^* F^k = p_{0,0}^* F^0 \\ &= 1 \end{aligned} \quad (6.18)$$

$$\begin{aligned} P_1^*(F) &= \sum_{k=0}^1 p_{1,k}^* F^k = p_{1,0}^* F^0 + p_{1,1}^* F^1 \\ &= 2F - 1 \end{aligned} \quad (6.19)$$

$$\begin{aligned} P_2^*(F) &= \sum_{k=0}^2 p_{2,k}^* F^k = p_{2,0}^* F^0 + p_{2,1}^* F^1 + p_{2,2}^* F^2 \\ &= 6F^2 - 6F + 1 \end{aligned} \quad (6.20)$$

$$\begin{aligned} P_3^*(F) &= \sum_{k=0}^3 p_{3,k}^* F^k = p_{3,0}^* F^0 + p_{3,1}^* F^1 + p_{3,2}^* F^2 + p_{3,3}^* F^3 \\ &= 20F^3 - 30F^2 + 12F - 1 \end{aligned} \quad (6.21)$$

Otra forma escrita para los L-momentos en términos de los Momentos ponderados probabilísticos α_k y β_k (Hosking 1990, 1993) es la siguiente:

$$\lambda_{r+1} = (-1)^r \cdot \sum_{k=0}^r p_{r,k}^* \cdot \alpha_k = \sum_{k=0}^r p_{r,k}^* \cdot \beta_k \quad (6.22)$$

Donde:

$$p_{r,k}^* = (-1)^{r-k} \cdot \binom{r}{k} \cdot \binom{r+k}{k} = \frac{(-1)^{r-k} \cdot (r+k)!}{(k!)^2 \cdot (r-k)!} \quad (6.23)$$

Desarrollando algebraicamente estas ecuaciones, se obtienen los Momentos L-momentos. Cabe señalar que los L-momentos son realmente lineales a los momentos PWM.

$$\begin{aligned} \lambda_1 &= \alpha_0 = \beta_0 \\ \lambda_2 &= \alpha_0 - 2\alpha_1 = 2\beta_1 - \beta_0 \\ \lambda_3 &= \alpha_0 - 6\alpha_1 + 6\alpha_2 = 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \end{aligned} \quad (6.24)$$

Una vez conocidos los valores de λ_1 , λ_2 , λ_3 y λ_4 , podemos definir las razones o ratios de los L-momentos que son medidas de ubicación (τ) de L-momentos, comenzando con $L-C_v$, que es análogo al coeficiente de variación y después los de similitud con los coeficientes de asimetría (C_s) y de curtosis (C_k) los cuales son respectivamente:

- Coeficiente de Variación $L - C_v$:

$$\tau_2 = \lambda_2/\lambda_1 \quad (6.25)$$

- Coeficiente de Asimetría L-Asimetría:

$$\tau_3 = \lambda_3/\lambda_2 \quad (6.26)$$

- Coeficiente de Curtosis L-curtosis:

$$\tau_4 = \lambda_4/\lambda_2 \quad (6.27)$$

- Coeficiente de orden superior:

$$\tau_r = \lambda_r/\lambda_2 \quad (6.28)$$

Los L-momentos son un sistema alternativo para caracterizar una función de distribución de probabilidad (FDP). Históricamente aparecen como modificaciones de los momentos de probabilidad pesada (MPP)(Probability Weigthed Moments (PWM)) desarrollados por Greenwood et al., (1979), pero en la práctica a menudo son estimados a partir de una muestra finita, (Hosking y Wallis, 1997) los definen de la siguiente manera:

Dada una muestra de tamaño n , con sus elementos dispuestos de forma ordenada. Sea $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ esta muestra ordenada. Un estimador insesgado del momento ponderado de probabilidad β_r es (Landwehr et al., 1979).

$$b_r = \frac{1}{n \binom{n-1}{r}} \sum_{j=r+1}^n \binom{j-1}{r} x_{j:n} \quad (6.29)$$

El cual se puede escribir de manera alternativa:

$$b_0 = \frac{1}{n} \sum_{j=1}^n x_{j:n} \quad (6.30)$$

$$b_1 = \frac{1}{n} \sum_{j=2}^n \frac{(j-1)}{(n-1)} x_{j:n} \quad (6.31)$$

$$b_2 = \frac{1}{n} \sum_{j=3}^n \frac{(j-1)(j-2)}{(n-1)(n-2)} x_{j:n} \quad (6.32)$$

$$b_3 = \frac{1}{n} \sum_{j=4}^n \frac{(j-1)(j-2)(j-3)}{(n-1)(n-2)(n-3)} x_{j:n} \quad (6.33)$$

Y en general:

$$b_r = \frac{1}{n} \sum_{j=r+1}^n \frac{(j-1)(j-2)\dots(j-r)}{(n-1)(n-2)\dots(n-r)} x_{j:n} \quad (6.34)$$

Los estimadores muestrales de β_r serán los b_r los cuales están definidos por las ecuaciones 6.30 a 6.33 y los de los cocientes serán t_2, t_3 y t_4 , según las ecuaciones 6.25 a 6.27.

$$\begin{aligned} \lambda_1 &= l_1 = b_0 \\ \lambda_2 &= l_2 = 2b_1 - b_0 \\ \lambda_3 &= l_3 = 6b_2 - 6b_1 + b_0 \\ \lambda_4 &= l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \end{aligned} \quad (6.35)$$

En general:

$$\lambda_{r+1} = l_{r+1} = \sum_{k=0}^r p_{r,k}^* b_k ; r = 0, 1, \dots \quad (6.36)$$

Los coeficiente están definidos como en las ecuaciones 6.35. La modelo de L-momentos l_r es un estimador insesgado de β_r . De las ecuaciones. 6.34 y 6.36, l_r es una combinación lineal de la muestra ordenada de valores $X_{1:n}, X_{2:n}, \dots, X_{n:n}$ y podemos escribir:

CUADRO 6.4: Notación para Momentos y L-Momentos

	Momento Poblacional	Momento Muestral	L-momento Poblacional	L-momento Muestral
Ubicación (media)	μ	\bar{x}	λ_1	l_1
Escala	σ	s	λ_2	l_2
CV	C_v	\widehat{C}_v	τ	l
Asimetría	γ	g	τ_3	l_3
Curtosis	κ	k	τ_4	l_4

Cuadro 6.4 Notación para Momentos y L-Momentos

$$l_r = \frac{1}{n} \sum_{i=0}^n w_{i:n}^{(r)} x_{i:n}; \quad r = 0, 1, \dots \quad (6.37)$$

En la notación de Neuman y Schonbach (1974), los pesos $w_{i:n}^{(r)}$ son los polinomios discretos de Legendre $(-1)^{r-1} P_{r-1}(j-1, n-1)$

Análogamente a las Ecs. 6.25 y 6.28, la proporción muestral de L-momentos se definen por:

$$t_r = \frac{l_r}{l_2} \quad r = 3, 4, \dots \quad (6.38)$$

y se define $L - C_v$ por:

$$t_2 = \frac{l_2}{l_1} \quad (6.39)$$

Estas cantidades son análogas al coeficiente de variación ordinario, $C_v, L - C_v$ no es una abreviatura de “L-coeficiente de variación”, sino de manera más apropiadamente debiese ser descrita como un “coeficiente de L-variación”. Estos son considerados, estimadores naturales de τ_r y τ , respectivamente.

La tabla 6.4 muestra la relación que existe entre los momentos convencionales y los L-momentos tanto muestrales como poblacionales.

Por lo tanto, una distribución puede ser caracterizada por sus L-momentos, incluso si algunos de sus momentos convencionales no existen. Además, tal especificación es siempre única: esto no es por supuesto cierto para los momentos convencionales.

Los L-momentos caracterizan una distribución, mientras que los momentos convencionales, en general no lo hacen.

Para abordar el problema de encontrar muestras homogéneas, un análisis de conglomerados se lleva a cabo utilizando estimaciones del exponente de Hurst. Los agrupamientos y el resultado de las distribuciones de probabilidad de ajuste a través de los datos de cada estación de lluvias se muestran en la tabla 3.4.

Los L-momentos $\lambda_1, \lambda_2, \dots, \lambda_r$, y las relaciones de L-momentos τ_3, \dots, τ_r , son cantidades útiles para describir una distribución. Los L-momentos son, de alguna manera, análogos a los momentos centrales (convencionales) y las relaciones de L-momento son análogos a las relaciones de momento. En particular, $\lambda_1, \lambda_2, \tau_3$, y τ_4 pueden considerarse como medidas de ubicación, de escala, la asimetría y la curtosis, respectivamente.

Teniendo en cuenta las ecuaciones (6.24), se definen las λ_r como las esperanzas matemáticas de combinaciones lineales de las estadísticas de orden. Es evidente que λ_1 , la media, es una medida de la ubicación.

λ_2 es la medición de la escala o la dispersión de la distribución.

λ_3 es la asimetría, aunque no de forma independiente de la escala.

λ_4 , es la curtosis.

Una justificación alternativa de las interpretaciones de la L-momentos puede estar basada en la obra de Oja (1981). Ampliando el trabajo de Bickel y Lehmann (1975, 1976) y van Zwet (1964), Oja define intuitivamente criterios razonables para una distribución de probabilidad sobre la recta real que se encuentra más a la derecha (más dispersa, más inclinación, más curtosis) que otro. Una funcional de valores reales de una distribución que preserva el orden parcial de las distribuciones que suponen esos criterios, pueden entonces, razonablemente ser llamado una "medida de la ubicación" (dispersión, asimetría, curtosis).

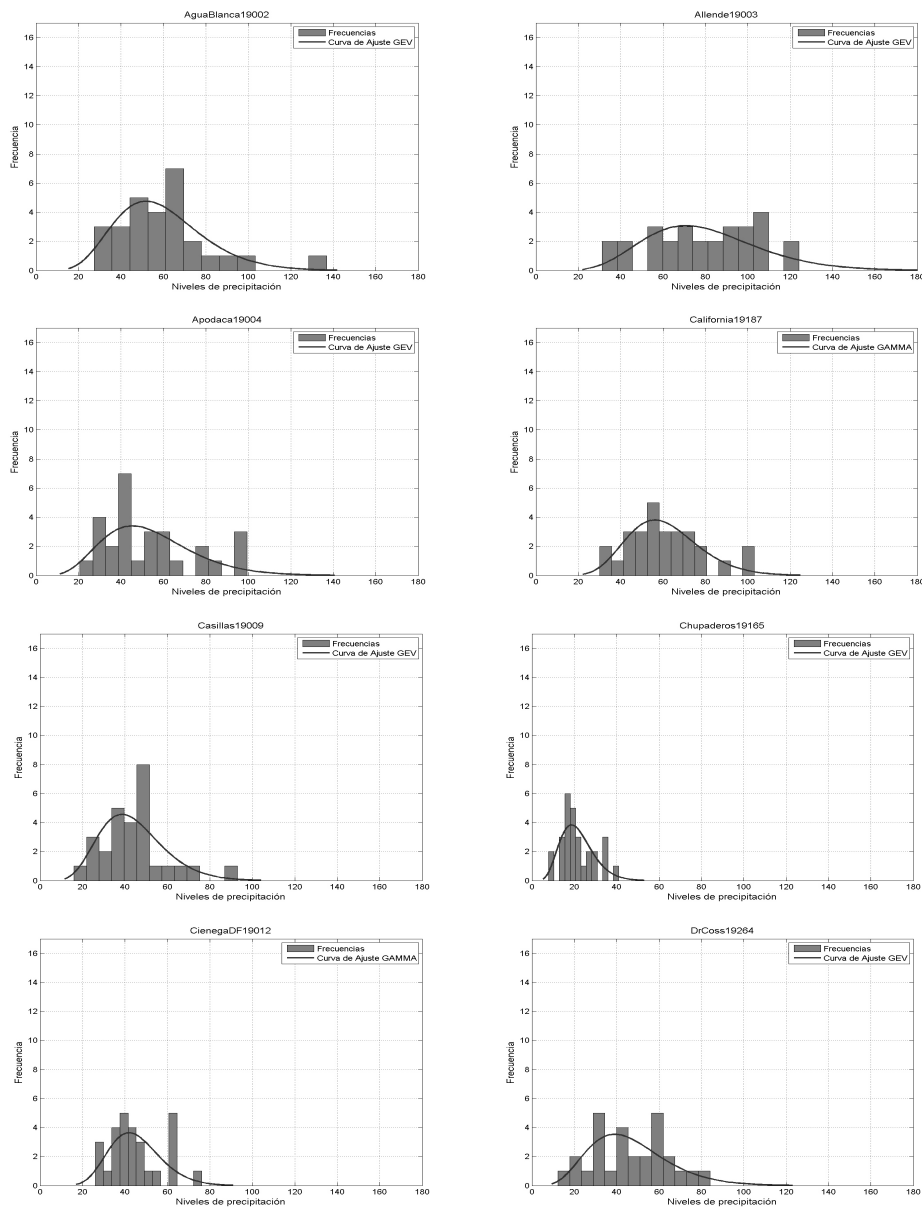
Esto se deduce inmediatamente a partir de la obra de Oja que λ_1 y λ_2 , en la notación de Oja, $\mu_1(F)$ y, una $(\sigma_1(F))$, son medidas de ubicación y de escala, respectivamente. Hosking (1989) muestra que el τ_3 y τ_4 son, según los criterios del Oja, medidas de asimetría y curtosis, respectivamente.

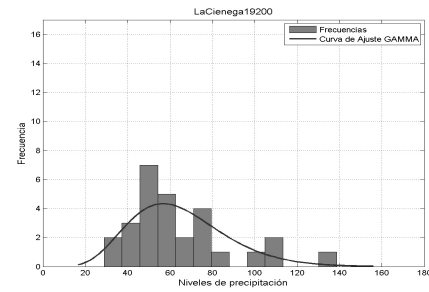
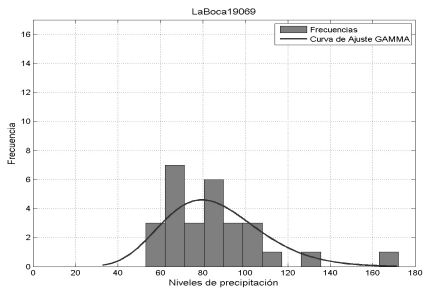
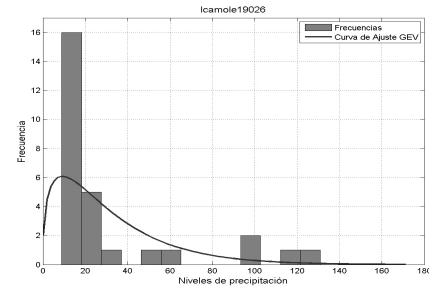
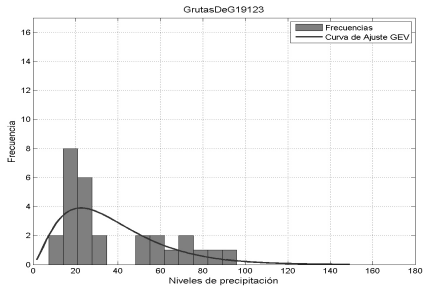
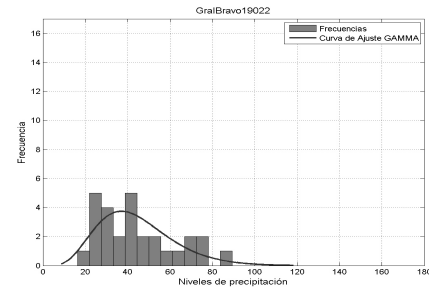
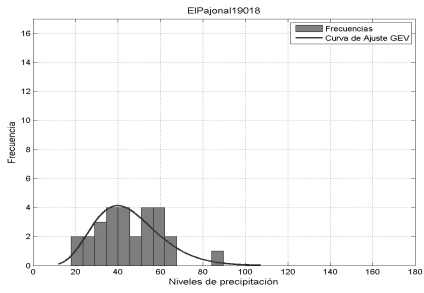
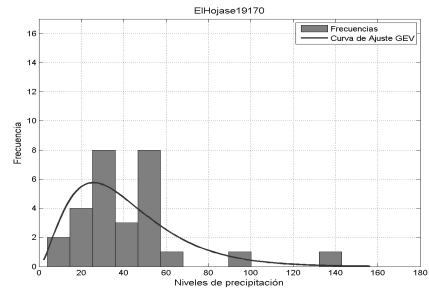
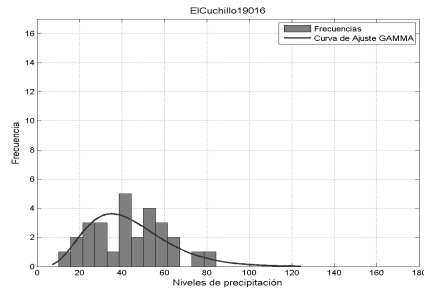
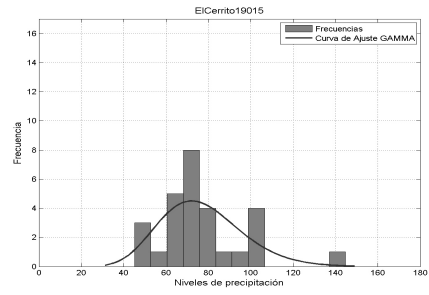
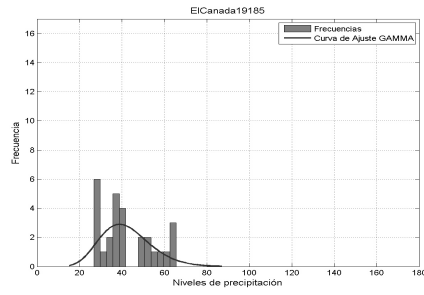
Lo anterior implica que las principales características de una distribución de probabilidad deben estar bien resumidos por los siguientes cuatro medidas: la media o L-ubicación, λ_1 ; la L-escala, λ_2 ; la L-asimetría, τ_3 ; la L-curtosis, τ_4 .

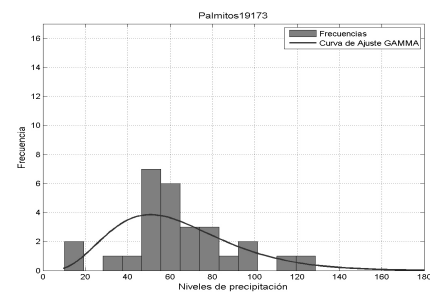
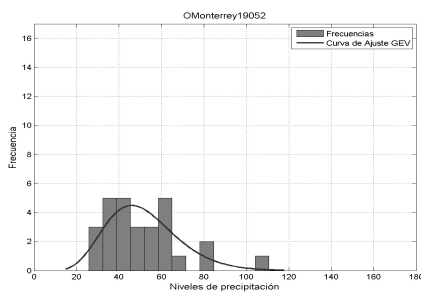
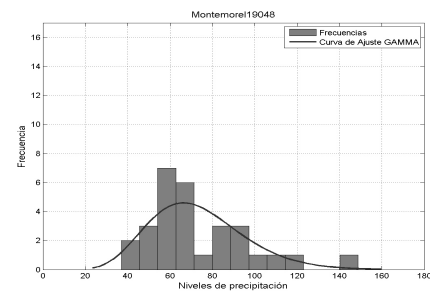
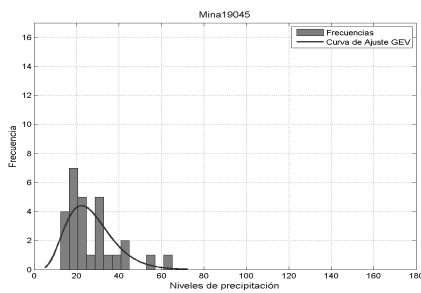
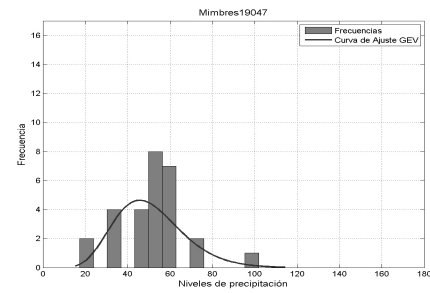
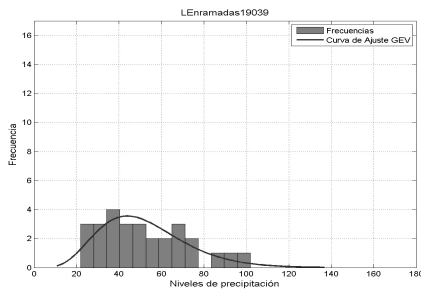
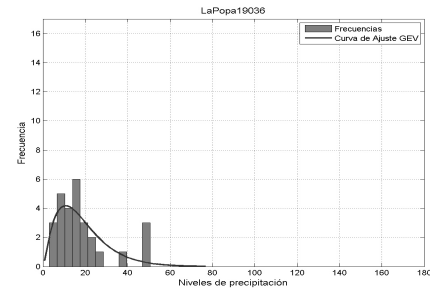
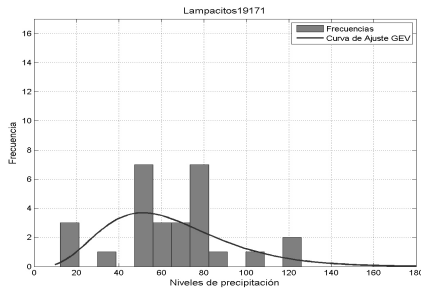
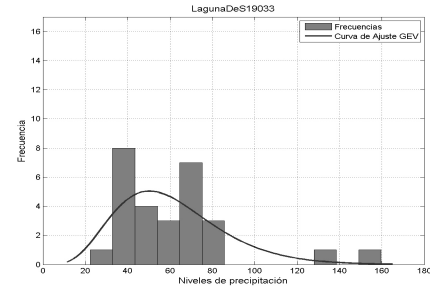
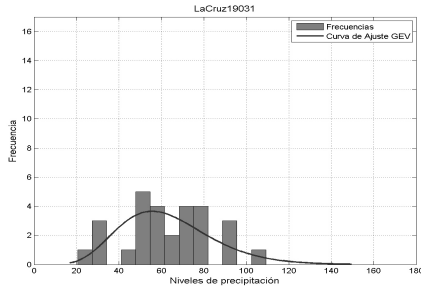
Para encontrar las funciones de distribución de probabilidad se calcularon los cuantiles de cada base de datos y una vez realizado esto se utilizó la librería de Matlab para asignar el tipo de distribución, entre varias que pone a disposición se comparó con las que excell, minitab y easyfit proporcionan y se asignó la distribución en base a las que se estudian en hidrología. Las mas comunes son: Distribución Gamma y General de valores extremos entre otras. Estas librerías también realizan pruebas de bondad de ajuste como: Anderson Darling, Smirnov Kolmogorov y Chi-cuadrado.

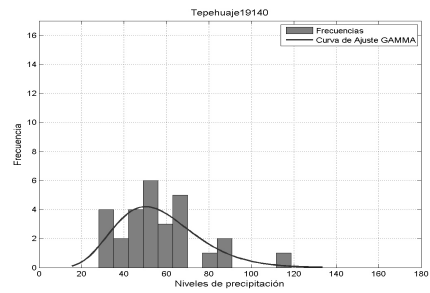
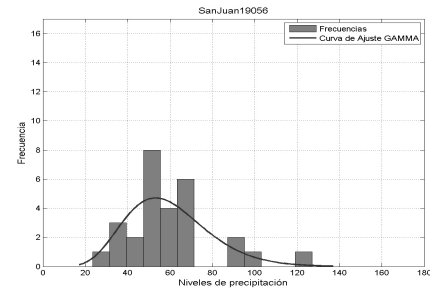
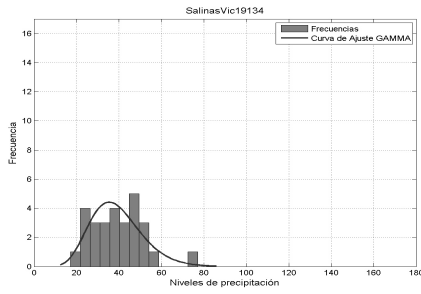
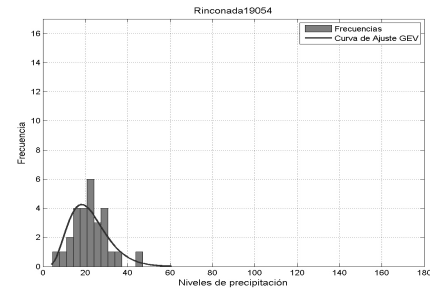
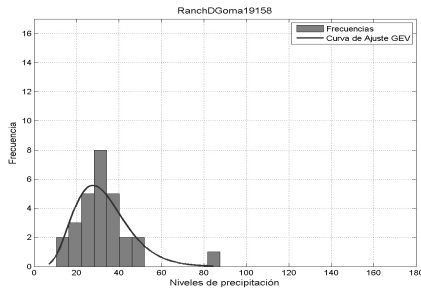
6.10. Histogramas y Distribuciones de probabilidad

En este anexo se muestran los gráficos de histogramas y ajuste de distribución de las 33 estaciones pluviométricas analizadas y de la cual se calcularon los parámetros de ubicación, dispersión, asimetría y curtosis, mediante L-momentos, para ajustar la curva que se muestra en una de ellas.









Research Article

Clustering of Rainfall Stations in RH-24 Mexico Region Using the Hurst Exponent in Semivariograms

Francisco Gerardo Benavides-Bravo,¹ F-Javier Almaguer,² Roberto Soto-Villalobos,² Víctor Tercero-Gómez,³ and Javier Morales-Castillo²

¹Instituto Tecnológico de Nuevo León, 67170 Guadalupe, NL, Mexico

²Universidad Autónoma de Nuevo León, 66451 San Nicolás de los Garza, NL, Mexico

³Tecnológico de Monterrey, 64849 Monterrey, NL, Mexico

Correspondence should be addressed to Francisco Gerardo Benavides-Bravo; fgbenavid@gmail.com

Received 15 August 2015; Revised 14 November 2015; Accepted 23 November 2015

Academic Editor: Costas Panagiotakis

Copyright © 2015 Francisco Gerardo Benavides-Bravo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An important topic in the study of the time series behavior and, in particular, meteorological time series is the long-range dependence. This paper explores the behavior of rainfall variations in different periods, using long-range correlations analysis. Semivariograms and Hurst exponent were applied to historical data in different pluviometric stations of the Río Bravo-San Juan watershed, at the hydrographic RH-24 Mexico region. The database was provided by the Water National Commission (CONAGUA). Using the semivariograms, the Hurst exponent was obtained and used as an input to perform a cluster analysis of rainfall stations. Groups of homogeneous samples that might be useful in a regional frequency analysis were obtained through the process.

1. Introduction

When limited observations of hydrological events are available, the ability to provide appropriate characterization, analysis, and predictions of a phenomenon gets compromised. However, the analysis can be improved by identifying homogeneous samples that can be used in combination to make better estimates of a probability model. This is one of the major concerns within the practice of regional frequency analysis (RFA), where the final output is the estimation of extreme events in a geographical area that can be used as input in risk analysis, water management, zoning, and land use applications, Hosking and Wallis [1]. However, the estimation of extreme events is considered a complex problem, mostly because the information is usually limited, serial correlation exists, multiple change-points might be present, and observations follow trends and seasonal patterns. To address these issues, hydrological time series studies have been successfully applied in the past, Machiwal and Jha [2]; however, most research efforts have been focused on trend detection tests, leaving aside other important properties such

as stationarity, homogeneity, periodicity, and persistence. By addressing these properties, a better selection of homogeneous samples might be possible, and as a consequence, practitioners might achieve better predictions.

Previous works on time series analysis in climatology with applications in precipitation go back to Bhuiya [3], with the development of a test for stationarity after periodic and trend components were subtracted from hydrologic series. Buishand [4] used trend tests to evaluate the difference in precipitation between rural and urban areas of Amsterdam and Rotterdam. Buishand [5, 6] constructed several tests of homogeneity in the mean of series with the use of cumulative sums, likelihood tests, and Bayesian inference. Kothyari et al. [7] evaluated three stations in India, Agra, Dehradun, and Delhi, to test for changes in rainfall and temperature, providing evidence of a change in the number of rainy days during monsoon season and an increment in temperature. Giakoumakis and Baloutsos [8] performed a trend analysis on historical series of annual precipitations from the basin of the Evinos Riven in Greece. By applying different tests of randomness, decreasing trends were found in the rainfall

records. Other authors dealing with trend analysis, homogeneity, and change-points found in the literature are Angel and Huff [9], Mirza et al. [10], Tarhule and Woo [11], Luís et al. [12], Kripalani and Kulkarni [13], Adamowski and Bougadis [14], Yu et al. [15], and Kumar et al. [16]. A comprehensive review of these works can be found in Machiwal and Jha [2] with descriptions of related developments in hydrological time series analysis.

Recent developments in hydrological analysis include the works of Golian et al. [17] with a classification and clustering approach of rainfall data using the natural-breaks classification method and the fuzzy c-means (FCM) algorithm. Shi et al. [18] analyzed variations in trends for precipitation data using a linear regression method, the Mann-Kendall test, and the Hurst exponent. The Hurst exponent, as part of a fractal analysis, was used to evaluate long-range dependence and the possibility of trends in the data. The following works around the Hurst exponent include the developments of Golder et al. [19], where the Hurst exponent is also used to explore long-term correlations, and cumulative rainfall observations were modeled using the alpha-stable probability law to deal with heavy-tailed distributions. Chang [20] extended the application of the Hurst exponent by developing a computation approach to estimate the exponent over time series that fits a discrete time fractional Brownian motion and fractional Gaussian noise. Yu et al. [21] also studied long-term correlations using the Hurst exponent and performed a multifractal analysis of rainfall series (see Kantelhardt [22]) based on a multiplicative cascade model and a multifractal detrended fluctuation analysis. Other recent works on time series analysis can be found in Carbone et al. [23] with the construction of a simulation model of storms using a double exponential distribution. Chou [24] investigated the complexity at different temporal scales of rainfall and runoff time series using the sample-entropy method, and finally, García-Marín et al. [25] performed a regional frequency analysis over rainfall data from Málaga, Spain, where the grouping of stations into homogenous regions has been done by following a cluster analysis with multifractal values of the different series.

In this paper, a clustering approach is used to group stations into homogeneous samples after summarizing the results of semivariogram analysis into a Hurst exponent. As a case study, a sample of pluviometric stations, the Río Bravo-San Juan watershed, at the RH-24 Mexico region was analyzed (Figure 1). A map of the Río Bravo-San Juan watershed is shown in Figure 2. This region is located in Mexico between the states of Nuevo León, Coahuila, and Tamaulipas covering an approximate area of 29420 km². Some of the rainfall stations are shown in Figure 3. The data used has been provided by CONAGUA, the local institution responsible for water management in the country.

2. Problem Description

In practice, to perform RFA, it is required to identify homogeneous regions where data follow similar patterns that can be analyzed together to improve the identification of probability models that in turn can be used to estimate extreme events



FIGURE 1: The watersheds of Mexico with Google Earth. In black edge, the Río Bravo-San Juan watershed. Source of the database: <http://www.conagua.gob.mx/>.

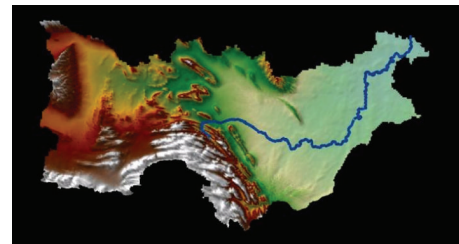


FIGURE 2: Río Bravo-San Juan watershed (San Juan river in blue). Image source: “Water Management in the Río San Juan Watershed, in the Southern Río Bravo Hydrologic Region of Mexico” at <http://earthzine.org/2012/08/13/>.

and their frequency in terms of return periods. This analysis is usually executed when dealing with droughts, pollution, wind movement, temperature, atmospheric pressure, and rainfall observations, to name a few. These researches deal with the problem of finding groups of rainfall stations that create homogeneous regions by considering fractal structures captured through semivariograms and Hurst exponents. Rainfall data from a sample of the hydrographic region RH-24 Mexico, the Río Bravo-San Juan watershed, are used as a case study to evaluate the proposed approach.

3. Methodology

Semivariograms, in the present study, are used to quantify long-range correlations of data from different pluviometric stations using monthly records. By considering the analysis of semivariograms of historical series, a rescaled range analysis R/S is performed to obtain a measure of the Hurst exponent [26]. The Hurst exponent is used as a metric of a particular pluviometric station. The process is repeated over each pluviometric station within the region under analysis. Hurst exponents are used as a reference to identify stations that exhibit similar patterns. As a consequence, a cluster analysis is applied to identify homogeneous samples. An advantage of the Hurst exponent is the simplicity of its algorithm that can be used to measure the condition of persistence or antipersistence of a process, and it provides a metric that can be used to classify different time series.

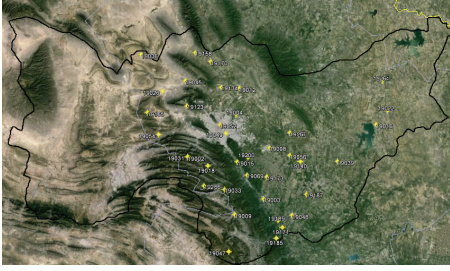


FIGURE 3: Geographical locations (from Google Earth) of the pluviometric stations at the Río Bravo-San Juan watershed. Source of the database: <http://www.conagua.gob.mx/>.

3.1. Semivariogram. The semivariogram or variogram $\gamma(h)$ is used to describe the relationship of paired observations separated by a distance h . It is a geostatistical technique that allows a quantitative measure of the long-range persistence in nonstationary time series Witt and Malamud [27], Haslett [28], and Dmowska and Saltzman [29]. Correlations over time and space create patterns that can be used to describe the behavior of a set of observations. Mathematically, the variogram estimates the expected squared difference between neighboring random variables. This calculation is performed over different h values. Given a time series or stochastic processes $\{X_t, t \geq 0\}$, the autocovariance function at the point $(t, t + h)$ is defined as $C_X(t, t + h) = E[X_t X_{t+h}] - E[X_t]E[X_{t+h}]$, with $E[X_t]$ as the mean of the process at time t . The semivariogram $\gamma(h)$ is given by half of the variance of the difference between pairs of observations at different “locations” in time:

$$\begin{aligned} \gamma(h) &= \frac{1}{2} \text{Var}(X_{t+h} - X_t) \\ &= \frac{1}{2} E[(X_{t+h} - X_t - E[X_{t+h} - X_t])^2] \\ &= \frac{\text{Var}(X_t) + \text{Var}(X_{t+h})}{2} - \text{Cov}(X_t, X_{t+h}) \quad (1) \\ &= \frac{\text{Var}(X_t) + \text{Var}(X_{t+h})}{2} \\ &\quad - \sqrt{\text{Var}(X_t) \text{Var}(X_{t+h})} \rho_X(t, t + h), \end{aligned}$$

where $-1 \leq \rho_X(t, t + h) \leq 1$ is the autocorrelation function (the autocovariance function normalized).

In the special case when $\rho(x_{t+h}, x_t) = 0, \forall(t, h)$, it is said that the stochastic processes $\{X_t, t \geq 0\}$ are uncorrelated and the semivariogram is reduced to the arithmetic mean of the variance of processes at times t and $t + h$:

$$\gamma(h) = \frac{\text{Var}(X_t) + \text{Var}(X_{t+h})}{2}. \quad (2)$$

If the random field $\{X_t, t \geq 0\}$ has constant mean $= E[X_t] = E[X_{t+h}] = \mu \forall t$, the semivariogram (1) adopts the simple form:

$$\gamma(h) = \frac{1}{2} E[(X_{t+h} - X_t)^2]. \quad (3)$$

If $X(t)$ and $X(t + h)$ are independent random variables $\forall t$, again, though for a different reason, the semivariogram is reduced to the special case (2).

In principle, given a stochastic process $\{X_t, t \geq 0\}$, the expected value of differences $E[X_{t+h} - X_t]$ at time t and with lag h is empirically estimated by the average over a “large enough” ensemble of realizations or paths in time. However, for a single time series $\{X_n, n = 1, 2, \dots, n\}$, the expected value can be estimated assuming an ergodic hypothesis, that is, a statistical principle of equivalence according to which the average over time and the average over the ensemble are the same, Lefebvre [30]. Thereby the differences $X_{t+h} - X_t$, which would be obtained with an infinitely reproducible process, are “simulated” or “cloned” from the “mother series.” Thus, the average value of differences $X_{t+h} - X_t$ is estimated by

$$E[X_{t+h} - X_t] = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (x_{i+h} - x_i), \quad (4)$$

where $n(h)$ is the number of differences with a lag h . When $h = 1, 2, 3, \dots$, the averages (4) are, respectively,

$$\begin{aligned} E[X_{t+1} - X_t] &= \frac{x_n - x_1}{n - 1}, \\ E[X_{t+2} - X_t] &= \frac{x_{n-1} + x_n - (x_1 + x_2)}{n - 2}, \\ E[X_{t+3} - X_t] &= \frac{x_{n-2} + x_{n-1} + x_n - (x_1 + x_2 + x_3)}{n - 3}, \\ &\vdots \\ E[X_{t+h} - X_t] &= \frac{1}{n(1 - h/n)} \left(\sum_{j=0}^{k-1} x_{n-j} - \sum_{l=1}^k x_l \right). \end{aligned} \quad (5)$$

According to (5) for a maximum value of h “relatively moderate” or $h/n < 1$, except in the presence of isolated extreme outliers, the two summations in (5) are roughly of the same order, such that the empirical average value (4) can be approximated by $E[X_{t+h}] \approx E[X_t] = m = \text{constant}$. This is an observed characteristic in the time series of the pluviometric stations. Therefore, the corresponding estimator of (3) is simply

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} (x_{i+h} - x_i)^2. \quad (6)$$

3.2. Measurement of the Hurst Exponent H . To estimate the Hurst exponent from a temporal series $\{X_k\}$, with $k \in 1, 2, \dots, N$, the series is divided in a group of d -subseries of length m . Really, the size m is an average number. A standard way, though not the only, to obtain the m size of the subseries is partitioning the original series in powers of base 2. In doing so, in each of the successive partitions, the approximate value

of m is as follows: $N, N/2, N/2^2, N/2^3, \dots$. For each subseries $n = 1, 2, \dots, d$, do the following:

- (1) Calculate the mean E_n and the standard deviation S_n .
- (2) Calculate the deviation with respect to the mean by subtracting the mean of each element using

$$Z_{in} = X_{in} - E_n, \quad i = 1, 2, \dots, m. \quad (7)$$

- (3) Get the partial sums:

$$Y_{in} = \sum_{j=1}^i Z_{jn}, \quad i = 1, 2, \dots, m. \quad (8)$$

- (4) Calculate the range:

$$R_n = \max_{i=1:m} (Y_{in}) - \min_{i=1:m} (Y_{in}). \quad (9)$$

- (5) Normalize the range:

$$\frac{R_n}{S_n}. \quad (10)$$

- (6) For each subseries of length m take the average:

$$\left\langle \frac{R}{S} \right\rangle_m = \frac{1}{d} \sum_{n=1}^d \frac{R_n}{S_n}. \quad (11)$$

- (7) Hurst [26] found the relation of the statistical $\langle R/S \rangle_m$ given by the following power law:

$$\left\langle \frac{R}{S} \right\rangle_m \approx cm^H, \quad (12)$$

where H is the Hurst exponent and c is a positive constant.

Two factors involved in the determination of the Hurst coefficient are the way time series is divided into a group of subseries and the asymptotic behavior of the rescaled range. First, the range of values m are used to calculate the slope of $\log(\langle R/S \rangle_m)$ given the relationship

$$\log \left(\left\langle \frac{R}{S} \right\rangle_m \right) = \log(c) + H \log(m). \quad (13)$$

Second, the determination of H is the result of the asymptotic behavior of the rescaled range, that is, when the value m tends to infinity. The analysis of the rescaled $\langle R/S \rangle_m$ over some values of m is estimated using a log/log expression given in (13). To obtain H coefficient, the least-squares method is used. The slope of this line is the Hurst coefficient H .

This exponent is considered a fractal index, Mandelbrot and Wallis [31], and provides information about long-term correlations exhibited by a series of observations; for a theoretical review of the Hurst exponent, see Mandelbrot [32]. In practice, the Hurst exponent can take values between 0 and 1, where

- (i) $0 < H < 0.5$ indicates nonpersistence in a series; that is, an increment is more likely to be followed by a decrement and vice versa;
- (ii) $H = 0.5$ indicates lack of serial correlation (Gaussian white noise);
- (iii) $0.5 < H < 1$ indicates persistency; that is, an increment is likely to be followed by an increment and a decrement by another decrement.

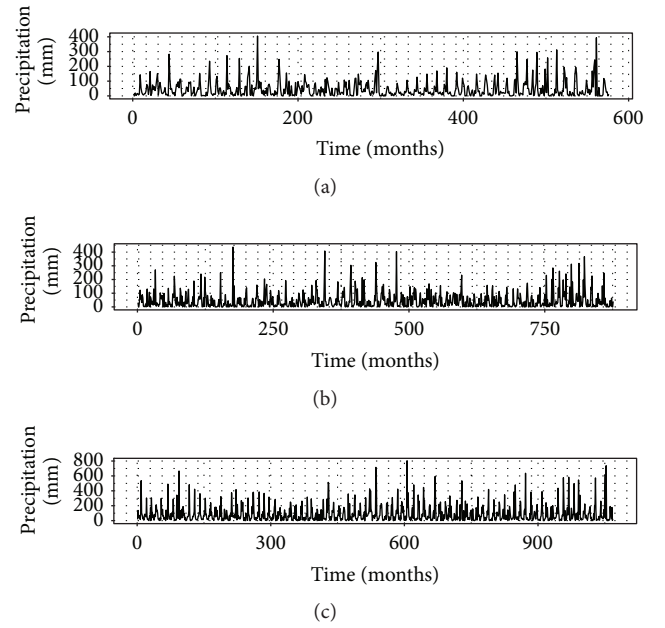


FIGURE 4: Time series for rainfall measurements from three stations at Río Bravo-San Juan watershed, Mexico. From the top to the bottom, respectively: Apodaca, station number 19004, 1940 January–2012 December; El Cuchillo, station number 19016, 1939 January–2012 December; La Boca, station number 19069, 1923 January–2012 December.

4. Results

To illustrate the procedure, only the analysis of three rainfall stations was selected to be presented in this section. The measured values of monthly precipitation in millimeters from the stations Apodaca, El Cuchillo, and La Boca are displayed in Figure 4, and results obtained taking into account all stations (following the same process) are presented at the end of this section.

As can be seen, patterns and relationships between different stations are difficult to assess only through “eyeball” analysis. However, when semivariograms are obtained, as shown in Figure 5, a footprint of the data becomes more evident. A closer inspection in every station shows a seasonal pattern that repeats every 12 observations in the semivariogram. This can be explained due to the fact that monthly observations were used in the analysis.

Once the semivariograms were obtained, the Hurst exponent for the series of the γ values was calculated. It can be seen that Hurst coefficients are close to 1, which indicates a positive long dependency of the data in the semivariograms. Hurst exponents from all stations under analysis are presented in Table 1, where the long dependency in all variograms becomes clear. Some of the coefficients that appear in the table exceed the interval established of possible values of the Hurst exponent, $0 < H < 1$; this is a known error due to estimation bias or a possible linear retrogression.

To address the issue of finding homogeneous samples, a cluster analysis is performed using estimates of the Hurst exponent. As shown in Figure 6, a histogram of frequencies

TABLE 1: Hurst coefficients for semivariogram (6) from pluviometric stations at Río Bravo-San Juan watershed.

Station	Name	Latitude	Longitude	Data	Hurst
19015	El Cerrito	25 30 36	100 11 36	1939–2012	0.71245
19039	Las Enramadas	25 30 05	099 31 17	1940–2012	0.78917
19018	El Pajonal	25 29 23	100 23 20	1955–2012	0.79764
19012	Ciénega de Flores	25 57 08	100 10 20	1940–2012	0.80106
19003	Allende	25 17 01	100 01 13	1940–2012	0.82492
19069	La Boca	25 25 46	100 07 44	1923–2013	0.8387
19189	El Pastor	25 09 06	099 55 36	1987–2011	0.84065
19009	Casillas	25 11 47	100 12 51	1956–2012	0.84258
19187	California	25 18 23	099 44 02	1982–2011	0.84406
19173	Palmitos	25 25 02	099 59 50	1982–2012	0.85304
19002	Agua Blanca	25 32 39	100 31 23	1958–2012	0.86171
19134	Salinas Victoria	25 57 33	100 17 34	1979–2011	0.86501
19031	La Cruz	25 32 47	100 31 23	1955–2011	0.86751
19036	La Popa	26 09 50	100 49 40	1956–2011	0.87324
19008	Cadereyta Jiménez	25 35 25	099 58 30	1995–2012	0.87335
19185	El Canada	25 02 48	099 56 29	1982–2011	0.87351
19056	San Juan	25 32 36	099 50 25	1944–2012	0.87678
19016	El Cuchillo	25 43 05	099 15 21	1960–2012	0.87778
19052	Monterrey(Obs)	25 44 01	100 16 01	1986–2008	0.87802
19026	Icamole	25 56 28	100 41 13	1954–2012	0.88481
19200	La Cienega	25 32 10	100 07 15	1984–2011	0.89216
19004	Apodaca	25 47 37	100 11 50	1964–2012	0.89506
19047	Mimbres	24 58 26	100 15 31	1957–2011	0.89837
19140	Tepehuaje	25 30 19	099 46 15	1979–2012	0.91092
19267	Santa Ma. La Floreña	25 10 59	99 46 00	1984–2011	0.91686
19048	Montemorelos	25 10 55	099 49 56	1940–2012	0.91836
19040	Los Aldama	26 03 52	099 11 48	1942–1994	0.92157
19022	General Bravo	25 48 05	099 10 32	1927–2012	0.92301
19158	Rancho de Gomas	26 10 11	100 27 52	1981–2011	0.93016
19045	Mina	26 00 08	100 32 00	1953–2012	0.93308
19054	Rinconada	25 40 52	100 43 03	1945–2012	0.93977
19165	Chupaderos del Indio	25 48 49	100 47 24	1982–2011	0.94288
19266	San Jose de Barranquillas	26 32 41	100 28 21	1978–2011	0.94911
19171	Lampacitos	25 06 38	099 53 57	1982–2011	0.95188
19264	Dr. Coss	25 51 16	099 56 36	1982–2011	0.96461
19170	El Hojase	26 06 55	100 21 38	1982–2011	1.00650

was used to separate stations into 6 clusters. These clusters and the result of fitting probability distributions over the data of each rainfall station are shown in Table 2.

Distributions were selected based on a goodness of fit analysis. After identifying a set of feasible distributions with p values bigger than a significant level of 0.05, in every case, the distribution with the highest average p value was selected for each station. Gamma and Generalized Extreme Value were the distributions that gave the best fit over the data analyzed. These functions are

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta}, \tag{14}$$

$$f(x; k, \sigma, \mu) = \frac{1}{\sigma} e^{-(1+kz)^{-1/k}} (1+kz)^{-1-1/k}, \tag{15}$$

$$z = \frac{x - \mu}{\sigma},$$

respectively.

As a result of the analysis, estimated distributions do not mix within cluster. This input can be used in a posterior analysis to obtain maximum likelihood estimators of the distribution parameters using all data concurrently.

5. Conclusions

Variograms followed by analysis of R/S or Hurst exponent estimation were used as input of a cluster analysis. Hurst

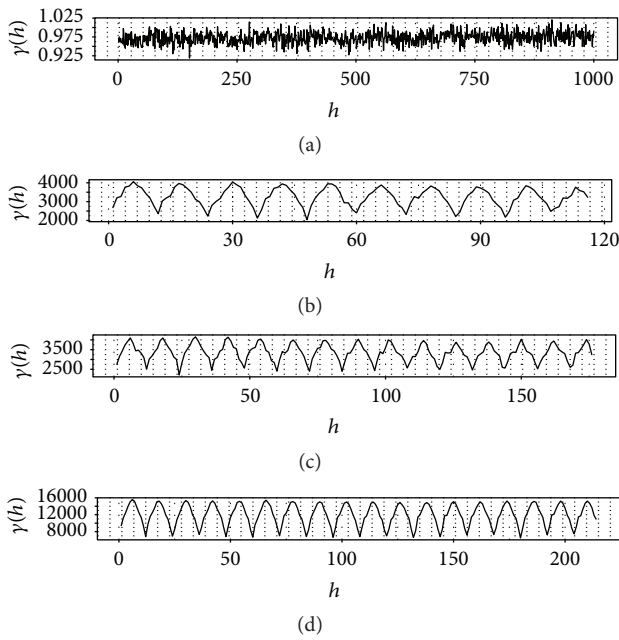


FIGURE 5: Semivariogram (6) corresponding to the time series of rainfall measurements from the three stations of the Río Bravo-San Juan watershed shown in Figure 4. In each case the maximum lag $h_{\max} = 0.20 * N$ was used, with N indicating the size of the time series. For comparison, the semivariogram corresponding to Gaussian white noise is included. From top to bottom, respectively: Gaussian white noise, Apodaca (station number 19004), El Cuchillo (station number 19016), and La Boca (station number 19069).

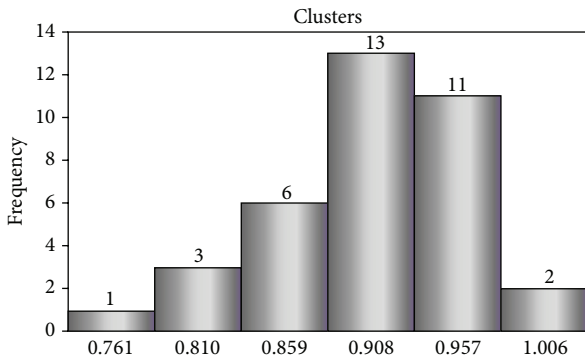


FIGURE 6: Histogram for the Hurst exponents of the analyzed pluviometric stations.

exponent provides a measure to determine if a time series is like a Gaussian white noise or has underlying trends, and it can be used to cluster the pluviometric stations according to the values of their semivariograms. As a case study to evaluate this approach, a sample of rainfall stations, those included in the Río Bravo-San Juan watershed from the hydrographic region RH-24 Mexico, was used in the analysis. Long-range dependency was found in every variogram evaluated with the Hurst exponent; however, it was still found useful as an input of a cluster analysis. A goodness of fit process was executed with every series, and the results showed that

TABLE 2: Clustering of pluviometric stations by the Hurst exponent of the semivariogram.

Cluster	Station	Hurst	Distribution	Parameters
1	19015	0.71245	Gamma(α, β)	(0.7022, 121.06)
	19039	0.78917		(0.8822, 101.75)
3	19018	0.79764	Gamma(α, β)	(0.5731, 78.07)
	19012	0.80106		(0.5402, 119.7)
6	19003	0.82492	GEV(k, σ, μ)	(0.349, 43.609, 38.143)
	19069	0.83870		(0.389, 43.321, 33.483)
	19189	0.84065		(0.453, 35.036, 22.915)
	19009	0.84258		(0.376, 25.309, 17.281)
	19187	0.84406		(0.377, 31.132, 24.640)
	19173	0.85304		(0.379, 32.826, 25.056)
	19002	0.86171		(0.351, 27.643, 20.471)
13	19134	0.86501	GEV(k, σ, μ)	(0.319, 21.703, 16.845)
	19031	0.86751		(0.313, 33.237, 23.081)
	19036	0.87324		(0.506, 9.5213, 5.0588)
	19008	0.87335		(0.407, 28.933, 22.421)
	19185	0.87351		(0.379, 22.415, 16.03)
	19056	0.87678		(0.372, 30.176, 22.56)
	19016	0.87778		(0.366, 24.295, 16.683)
	19052	0.87802		(0.429, 25.864, 19.559)
	19026	0.88481		(0.418, 8.7536, 5.424)
	19200	0.89216		(0.407, 33.194, 23.434)
11	19004	0.89506	GEV(k, σ, μ)	(0.368, 23.938, 17.563)
	19047	0.89837		(0.168, 33.305, 28.427)
	19140	0.91092		(0.369, 30.367, 23.162)
	19267	0.91686		(0.407, 20.642, 15.472)
	19048	0.91836		(0.356, 37.599, 30.006)
	19040	0.92157		(0.356, 20.614, 13.77)
	19022	0.92301		(0.383, 25.054, 16.587)
	19158	0.93016		(0.421, 16.537, 10.853)
	19045	0.93308		(0.457, 12.116, 7.9187)
	19054	0.93977		(0.485, 9.1504, 5.2866)
	19165	0.94288		(0.434, 10.814, 6.4381)
2	19266	0.94911	GEV(k, σ, μ)	(0.369, 32.361, 22.162)
	19171	0.95188		(0.435, 31.356, 22.367)
	19264	0.96461		(0.401, 20.168, 1.2395)
	19170	1.00650		(0.480, 20.046, 12.109)

existing dominant distributions within a feasible set (found independently in each station) do not overlap over clusters. The probability distributions were found nested within each cluster. This is indicative that homogeneous patterns were identified within groups, and groups were heterogeneous between themselves.

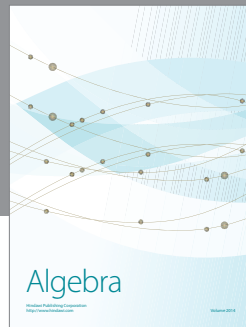
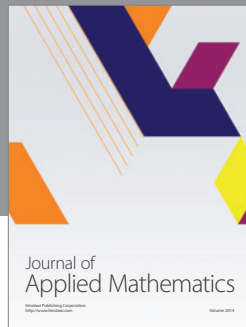
The study of the rainfall stations with semivariograms and R/S analysis provides a powerful tool that allows practitioners to analyze long-term correlations and clustering in hydrological time series. In future work, L -moments and spectral and wavelets analysis will be used to improve understanding of complex time series of pluviometric rainfall levels.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] J. R. M. Hosking and J. R. Wallis, *Regional Frequency Analysis: An Approach Based on L-Moments*, 2005.
- [2] D. Machiwal and M. K. Jha, "Current status of time series analysis in hydrological sciences," in *Hydrologic Time Series Analysis: Theory and Practice*, pp. 96–136, Springer Netherlands, 2012.
- [3] R. K. Bhuiya, "Stochastic analysis of periodic hydrologic process," *Journal of the Hydraulics Division*, vol. 97, no. 7, pp. 949–962, 1971.
- [4] T. A. Buishand, "Urbanization and changes in precipitation, a statistical approach," *Journal of Hydrology*, vol. 40, no. 3-4, pp. 365–375, 1979.
- [5] T. A. Buishand, "Some methods for testing the homogeneity of rainfall records," *Journal of Hydrology*, vol. 58, no. 1-2, pp. 11–27, 1982.
- [6] T. A. Buishand, "Tests for detecting a shift in the mean of hydrological time series," *Journal of Hydrology*, vol. 73, no. 1-2, pp. 51–69, 1984.
- [7] U. C. Kothyari, V. P. Singh, and V. Aravamuthan, "An investigation of changes in rainfall and temperature regimes of the Ganga Basin in India," *Water Resources Management*, vol. 11, no. 1, pp. 17–34, 1997.
- [8] S. G. Giakoumakis and G. Baloutsos, "Investigation of trend in hydrological time series of the Evinos River basin," *Hydrological Sciences Journal*, vol. 42, no. 1, pp. 81–88, 1997.
- [9] J. R. Angel and F. A. Huff, "Changes in heavy rainfall in Midwestern United States," *Journal of Water Resources Planning and Management*, vol. 123, no. 4, pp. 246–249, 1997.
- [10] M. Q. Mirza, R. A. Warrick, N. J. Ericksen, and G. J. Kenny, "Trends and persistence in precipitation in the Ganges, Brahmaputra and Meghna river basins," *Hydrological Sciences Journal*, vol. 43, no. 6, pp. 845–858, 1998.
- [11] A. Tarhule and M.-K. Woo, "Changes in rainfall characteristics in northern Nigeria," *International Journal of Climatology*, vol. 18, no. 11, pp. 1261–1271, 1998.
- [12] M. D. Luís, J. Raventós, J. C. González-Hidalgo, J. R. Sánchez, and J. Cortina, "Spatial analysis of rainfall trends in the region of Valencia (East Spain)," *International Journal of Climatology*, vol. 20, no. 12, pp. 1451–1469, 2000.
- [13] R. H. Kripalani and A. Kulkarni, "Monsoon rainfall variations and teleconnections over South and East Asia," *International Journal of Climatology*, vol. 21, no. 5, pp. 603–616, 2001.
- [14] K. Adamowski and J. Bougadis, "Detection of trends in annual extreme rainfall," *Hydrological Processes*, vol. 17, no. 18, pp. 3547–3560, 2003.
- [15] P.-S. Yu, T.-C. Yang, and C.-C. Kuo, "Evaluating long-term trends in annual and seasonal precipitation in Taiwan," *Water Resources Management*, vol. 20, no. 6, pp. 1007–1023, 2006.
- [16] V. Kumar, S. K. Jain, and Y. Singh, "Analysis of long-term rainfall trends in India," *Hydrological Sciences Journal*, vol. 55, no. 4, pp. 484–496, 2010.
- [17] S. Golian, B. Saghafian, S. Sheshangosht, and H. Ghalkhani, "Comparison of classification and clustering methods in spatial rainfall pattern recognition at Northern Iran," *Theoretical and Applied Climatology*, vol. 102, no. 3, pp. 319–329, 2010.
- [18] P. Shi, X. Ma, X. Chen, S. Qu, and Z. Zhang, "Analysis of variation trends in precipitation in an upstream catchment of Huai River," *Mathematical Problems in Engineering*, vol. 2013, Article ID 929383, 11 pages, 2013.
- [19] J. Golder, M. Joelson, M.-C. Neel, and L. Di Pietro, "A time fractional model to represent rainfall process," *Water Science and Engineering*, vol. 7, no. 1, pp. 32–40, 2014.
- [20] Y.-C. Chang, "Efficiently implementing the maximum likelihood estimator for Hurst exponent," *Mathematical Problems in Engineering*, vol. 2014, Article ID 490568, 10 pages, 2014.
- [21] Z.-G. Yu, Y. Leung, Y. D. Chen, Q. Zhang, V. Anh, and Y. Zhou, "Multifractal analyses of daily rainfall time series in Pearl River basin of China," *Physica A: Statistical Mechanics and Its Applications*, vol. 405, pp. 193–202, 2014.
- [22] J. W. Kantelhardt, "Fractal and multifractal time series," in *Encyclopedia of Complexity and Systems Science*, pp. 3754–3779, Springer, New York, NY, USA, 2009.
- [23] M. Carbone, M. Turco, G. Brunetti, and P. Piro, "A cumulative rainfall function for subhourly design storm in Mediterranean Urban areas," *Advances in Meteorology*, vol. 2015, Article ID 528564, 10 pages, 2015.
- [24] C.-M. Chou, "Complexity analysis of rainfall and runoff time series based on sample entropy in different temporal scales," *Stochastic Environmental Research and Risk Assessment*, vol. 28, no. 6, pp. 1401–1408, 2014.
- [25] A. P. García-Marín, J. Estévez, M. T. Medina-Cobo, and J. L. Ayuso-Muñoz, "Delimiting homogeneous regions using the multifractal properties of validated rainfall data series," *Journal of Hydrology*, vol. 529, pp. 106–119, 2015.
- [26] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, vol. 116, no. 1, pp. 770–799, 1951.
- [27] A. Witt and B. D. Malamud, "Quantification of long-range persistence in geophysical time series: conventional and benchmark-based improvement techniques," *Surveys in Geophysics*, vol. 34, no. 5, pp. 541–651, 2013.
- [28] J. Haslett, "On the sample variogram and the sample autocovariance for non-stationary time series," *Journal of the Royal Statistical Society—Series D: The Statistician*, vol. 46, no. 4, pp. 475–485, 1997.
- [29] R. Dmowska and B. Saltzman, Eds., *Advances in Geophysics, Long-Range Persistence in Geophysical Time Series*, Academic Press, 1999.
- [30] M. Lefebvre, *Applied Stochastic Processes*, Springer, New York, NY, USA, 2007.
- [31] B. B. Mandelbrot and J. R. Wallis, "Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence," *Water Resources Research*, vol. 5, no. 5, pp. 967–988, 1969.
- [32] B. Mandelbrot, "Statistical methodology for nonperiodic cycles: from the covariance to R/S analysis," *Annals of Economic and Social Measurement*, vol. 1, no. 3, pp. 259–290, 1972.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

