UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMATICA
ESCUELA DE MATEMATICA

ANÁLISIS E IDENTIFICACIÓN DEL PATRÓN DE COMPORTAMIENTO DE LAS LLUVIAS EN EL SALVADOR (1971 – 2012) EMPLEANDO DIVERSAS TÉCNICAS ESTADÍSTICAS PARA LA PREVENCIÓN DE EVENTOS EXTREMOS

Trabajo de Graduación presentado por:
Br. José Alexander Avalos Fuentes

Para Optar al grado de:
Licenciado(a) en Estadística

Asesor:
Msc. Rolando Lemus Gómez

Mayo, 2017

San Salvador **El Salvador** **Centroamérica**

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMATICA
ESCUELA DE MATEMATICA

Trabajo de Graduación:

ANÁLISIS E IDENTIFICACIÓN DEL PATRÓN DE COMPORTAMIENTO DE LAS LLUVIAS EN EL SALVADOR (1971 – 2012) EMPLEANDO DIVERSAS TÉCNICAS ESTADÍSTICAS PARA LA PREVENCIÓN DE EVENTOS EXTREMOS

Presentado por:
Br. José Alexander Avalos Fuentes

_____
Msc. Rolando Lemus Gómez
Asesor

Ciudad Universitaria, San Salvador, Mayo de 2017

Universidad de El Salvador

Lic Luis Argueta Antillón
Rector

Doctora Ana Leticia Zavaleta de Amaya
Secretario General

Facultad de Ciencias Naturales y Matemática

M.Sc. Martín Enrique Guerra Cáceres
Decano

Lic. Carlos Antonio Quintanilla Aparicio
Secretario

Escuela de Matemática
M.Sc. José Nerys Funes Torres
Director

M.Sc Alba Idalia Córdoba Cúellar
Secretaria

Contenido

Lista de Figuras y Gráficos

Figuras

Gráficos

Tablas

1. Introducción

Sabemos que para la planeación y construcción de toda obra arquitectónica se toman en cuenta muchos factores como por ejemplo el tipo de suelo, la humedad ambiental, la cercanía de otras estructuras y muchos otros factores a considerar.

Cuando se trata de la construcción de obras hidráulica, se toma en cuenta lo que se llama periodo de retorno. Período de retorno es uno de los parámetros más significativos a ser tomado en cuenta en el momento de dimensionar una obra hidráulica destinada a soportar avenidas, como por ejemplo: el vertedero de una presa, los diques para control de inundaciones; o una obra que requiera cruzar un río o arroyo con seguridad, como por ejemplo un puente.

En varias áreas de la ingeniería, el período de retorno es el tiempo esperado o tiempo medio entre dos sucesos improbables y con posibles efectos catastróficos. Así, en ingeniería hidráulica es el tiempo medio entre dos trombas de agua por encima de un cierto caudal.

En hidrología es frecuente considerar zona inundable a aquella zona que es cubierta por las aguas en tormentas de hasta quinientos años de periodo de retorno.
Esto significa que la cantidad de lluvia caída en un sólo día para ese periodo de retorno solamente se iguala o supera, estadísticamente, una vez en el período de 500 años.
En términos numéricos se expresa que la probabilidad de que se presente una precipitación superior en un determinado año es

$$p = \frac{1}{500} = 0.002 = 0.2\ \%$$

O bien, la probabilidad de que no se presente es la complementaria:

$$1 - p = 1 - \frac{1}{500} = 0.998 = 99.8\ \%$$

Sin embargo, eso no implica que no puedan producirse dos tormentas de igual o superior intensidad en dos años consecutivos, o incluso en un mismo año.

En el ámbito de la meteorología, el período de retorno es de vital importancia y generalmente es expresado en años, es un indicador que puede ser entendido como el número de años en que se espera que en promedio se repita un cierto caudal, o un caudal mayor.

Por lo cual, la probabilidad de superar k veces un caudal determinado en un período de tiempo T, viene dada por una distribución de Poisson:

$$Prob(x = k, t) = \frac{\left(\frac{t}{t_p}\right)^k}{k!} e^{-\left(\frac{t}{t_p}\right)}$$

Donde:

$$E(x) = \lambda = \left(\frac{t}{t_p}\right)$$

t : Es la unidad de tiempo o espacio con que se está trabajando.

t_p : Es la cantidad total de periodos de tiempo o espacio que se este manejando.

Por otro lado, si un evento tiene un periodo de retorno real de x cantidad de años, el número medio de eventos que se pueden presentar en un año determinado es:

$$E(x) = \frac{t}{t_p}$$

Así podemos decir como un ejemplo que el período de retorno de un caudal de 100 m^3/s, para una sección específica de un río determinado, es de 20 años, si, caudales iguales o mayores de 100 m^3/s se producen, en promedio cada 20 años.

El período de retorno para lo cual se debe dimensionar una obra varía en función de la importancia de la obra (interés económico, socio-económico, estratégico, turístico), de la existencia de otras vías alternativas capaces de reemplazarla y de los daños que

implicaría su ruptura: pérdida de vidas humanas, costo y duración de la reconstrucción, costo del no funcionamiento de la obra, etc.

En muchos lugares, se podría por ejemplo proponer la construcción de badenes en vez de un puente, derivando los esfuerzos financieros hacia otras zonas, donde se estima necesaria mayor seguridad.

Al contrario de otras obras, se tiene a veces la posibilidad de sobredimensionar una obra sin mayor costo adicional (por ejemplo en el caso de un valle estrecho, se puede, sin mayor costo sobre-elevar el puente), permitiendo así prevenir repuntas y aluviones cuya descarga pico es imprevisible.

La idea es evitar el súper dimensionamiento de toda la obra, concentrando los esfuerzos en algunas partes definidas como vitales o esenciales, y adoptar disposiciones constructivas permitiendo minimizar los daños en caso de eventos excepcionales.

Períodos de retorno generalmente aceptados:

- Obras hidráulicas para canalización de aguas de lluvia en ciudades de mediano porte o grandes: de 20 a 50 años;
- Obras hidráulicas para canalización de aguas de lluvia en ciudades de pequeño porte: de 5 a 10 años;
- Puentes importantes: 100 años;
- Vertederos para presas que posea poblaciones que se encuentran aguas abajo del lugar de ubicación de la presa: 1.000 a 10.000 años. Evidentemente en estos casos se trata de estimaciones basadas en procedimientos estadísticos. En algunos casos para obras hidráulicas cuya ruptura significaría un riesgo muy elevado de pérdidas de vidas humanas, estos valores son corroborados también con el método de la "Precipitación Máxima Probable".

El cálculo del periodo de retorno para algunas zonas de nuestro país se ha mantenido invariante por las condiciones climáticas más o menos estables; pero en años anteriores se han tenido temporadas lluviosas un poco fuera de lo normal, por lo cual se ve la necesidad de hacer un estudio para determinar si este comportamiento es solo un caso atípico o si se repetirá en el futuro; no es que vayamos a predecir el futuro comportamiento de las lluvias a largo plazo, sino analizar el comportamiento que se ha tenido de las lluvias en los últimos cuarenta años y ver si hay una tendencia del incremento o modificación del patrón de las lluvias para diferentes zonas de nuestro país, así como del comportamiento de los casos extremos.

Es por ello que se hará uso de una base de datos proporcionada por el Ministerio del Medio Ambiente y Recursos Naturales (MARN) y que consiste en la información de las precipitaciones pluviales captadas y registradas por las estaciones de monitoreo en diferentes puntos de nuestro territorio (ver Figura 1).

Figura 1. Red de estaciones meteorológicas en El Salvador. Fuente SNET.

2. Planteamiento del problema

Es bien sabido que el agua es uno de los recursos vitales en toda región o país, no solo por el hecho de que nos es útil en la vida diaria sino que a corto y largo plazo afecta a la economía, la industria y por sobre todo las vidas humanas y recursos materiales.

Esa agua de la que nos servimos proviene en gran medida de la lluvia, es por ello que estamos constantemente monitoreando las precipitaciones pluviales debido no solo por su escasez sino también por su abundancia.

Es entonces que surgen las preguntas:
¿En qué mes del año es más adecuado iniciar una construcción?
¿En qué tiempo es más adecuado sembrar las cosechas de frijol y/o de maíz?
¿Soportará la cárcava el aumento del cauce del río este año?

Estas y otras preguntas que hemos omitido se podrían contestar si conociéramos de antemano el comportamiento del ciclo de la época lluviosa, claro que con el conocimiento empírico hemos descubierto que en nuestra región las lluvias llegan a mediados de abril y terminan a principios de octubre, pero también sabemos que actualmente el calentamiento global está alterando las condiciones climáticas de todo el planeta y es cuando surge la inquietud:

¿El comportamiento de las precipitaciones pluviales se ha visto afectado por el calentamiento global en nuestro país?

Podemos decir que no lo sabemos con certeza, solo podemos suponer que el cambio climático nos está afectando, pero no podemos decir que las lluvias torrenciales están incrementándose o disminuyendo de frecuencia.

Es por esa incertidumbre que debemos investigar cómo se está viendo afectado el

patrón de comportamiento de las precipitaciones pluviales y con esa información prevenir las dificultades que se puedan presentar en el futuro próximo.

El agua es muy importante en nuestra vida diaria, una de las más importantes para la economía de nuestro país es que nos provee energía eléctrica por medio de las presas hidroeléctricas que se han construido en nuestro país.

Figura 2. Presa 5 de Noviembre

Si las lluvias se ven incrementadas no padeceremos de racionamiento energético pero deberemos estar preparados para prevenir inundaciones, derrumbes y deslaves de tierra ya que nuestros suelos son susceptibles a estos eventos.

En cambio si hay una disminución de las lluvias o incluso sequia, podría haber racionamientos energéticos y de agua potable e incremento en los precios de la energía eléctrica y otros servicios que se ven afectados directa o indirectamente por la escases de agua.

3. Antecedentes

Anteriormente ya se ha analizado la serie de precipitaciones pluviales de las captaciones registradas en nuestro país, pero únicamente para la región oriental y las predicciones obtenidas solo incluían periodos de tiempo en que la lluvia era normal, es decir, que no se toman en cuenta los periodos en los cuales los eventos climáticos como son el del niño y de la niña afectan la climatología de nuestro país se ve afectada de manera significativa.

Es un trabajo de grado para aspirar al título de maestro(a) en estadística y se llamó: "Ajuste de un modelo ARIMA para la precipitación diaria en la zona oriental de El Salvador", el cual fue desarrollado por Daysi Maribel Renderos y Mario Giovanni Molina Masferrer.

En dicho trabajo se menciona que se emplearon los datos de las estaciones de monitoreo de la región oriental y efectúan predicciones empleando modelos ARIMA para la lluvia diaria en periodo normal, es decir, que no intervienen los eventos climáticos del niño ni de la niña, se debe mencionar además que no se hicieron predicciones para este tipo de eventos extremos.

4. Justificación

La utilización de técnicas de modelado y pronóstico en el área de medio ambiente reviste importancia en cuanto nos permite conocer el comportamiento de un determinado evento a través del tiempo, además podemos realizar estimaciones de la posible ocurrencia de extremos y cuantificación de los mismos, en cuanto aumento o disminución de estos o ocurrencia de fenómenos con comportamiento cíclico o estacional.

El modelar las precipitaciones pluviales en nuestro caso nos permitirá construir un modelo que se adecue lo suficientemente bien al comportamiento de las lluvias, lo cual nos ayudará a la estimación de la correntia o cantidad de agua que se transporta sobre la superficie en un punto específico, este dato nos será de mucha utilidad en la estimación del periodo de retorno, el cual se utiliza en la construcción de obras hidráulicas como lo son los puentes, presas, bóvedas, etc.

Todo lo anterior permitirá estudiar históricamente las precipitaciones pluviales y determinar las zonas propensas a sufrir derrumbes, deslaves o inundaciones a través de la formulación de un modelo adecuado para los pronósticos basado en las precipitaciones pluviales a través del tiempo.

El uso de la técnica en las instituciones tiene amplias aplicaciones entre las más importantes está el hecho de permitir a las instituciones prepararse anticipadamente en diversos aspectos tales como atención sanitaria, construcciones, planificación estratégica en cuanto la asignación de recurso humano y material para enfrentar los diversos fenómenos naturales.

Cada año en nuestro país se enfrentan problemas debido a las altas precipitaciones pluviales, siendo algunas zonas más afectadas que otras, y por ende los grupos humanos más vulnerables económicamente son los que sufren las consecuencias.

Es por eso que la utilización de modelados y pronósticos tiene un fuerte impacto para apalear o disminuir los efectos que esta situación tiene sobre la población.

Sabemos que en nuestro país, el comportamiento climático está influenciado por la ocurrencia de fenómenos provenientes de la zona norte de América, como son las corrientes de aire frio que bajan de la zona norte, el análisis de las precipitaciones permitirá identificar estos e incorporarlos en el modelado de las series para el pronóstico como un modelo de intervención o un modelo ARIMA estacional.

Todo lo anterior permitirá estudiar históricamente las precipitaciones pluviales y ayudaría a determinar las zonas más propensas a sufrir derrumbes, deslaves o inundaciones a través de la formulación de un modelo adecuado, esto se haría empleando los modelos de pronóstico y creando con ello mapas de susceptibilidad.

Se debe aclarar que hay diversos trabajos en los que se han aplicado el análisis de series temporales por medio de los modelos ARIMA, así como el de redes neuronales artificiales, pero no hay uno que abarque ambos aspectos así como el tratamiento de datos faltantes en series climáticas y la estimación de un modelo para los valores extremos de una serie climatológica como lo es la precipitación pluvial.

5. Objetivos

Objetivo general

Efectuar un análisis de la serie histórica de las lluvias captadas en El Salvador aplicando diversas técnicas estadísticas para series de tiempo y valores extremos; para diferentes puntos de nuestro país, en el período de tiempo 1971- 2012 para determinar si existen señales de cambio en su comportamiento normal.

Objetivos específicos

- Determinar un modelo para la serie de tiempo precipitación para diferentes regiones de nuestro país El Salvador aplicando la metodología Box-Jenkins.

- Determinar algún patrón recurrente en la misma serie de tiempo empleando técnicas de aprendizaje automático.

- Determinación de señales de cambio en los patrones de comportamiento mensual de las precipitaciones en algunos puntos importantes de nuestro país

- Generación de predicciones utilizando los modelos encontrados, para luego realizar una comparación por tramos de 10 años y determinar si ha habido un efecto del cambio climático sobre las precipitaciones pluviales en nuestro país

- Realizar una comparación por tramos de 10 años y determinar si ha habido un efecto del cambio climático sobre las precipitaciones pluviales en nuestro país

- Realizar un análisis de valores extremos, e identificar si existe una tendencia a corto o largo plazo

6. Fundamento teórico

6.1. Tratamiento de datos ausentes

En esta sección abordaremos el problema de rellenar los espacios en blanco dentro de una serie de tiempo climatológica.

Es de vital importancia realizar este proceso antes de hacer cualquier tipo de análisis, ya que en la mayoría de métodos de análisis, no es posible prescindir de ningún dato y es que si se omitiera alguno, no se podría obtener un modelo estocástico adecuado que se ajuste a los datos.

Para la estimación de los datos ausentes se pueden aplicar diferentes técnicas, entre ellas están las que a continuación se enumeran:

1. Método por análisis de regresión
2. Método de la razón
3. Método de interpolación con otras estaciones

Método por análisis de regresión:

Para la aplicación de éste método se requiere seleccionar una serie de datos con un comportamiento similar, esto es, dentro de la misma área de influencia climática (referencia), a la serie que tiene los datos faltantes (estudio). La serie de datos de la referencia, debe contener el registro de datos completo en los períodos para los cuales faltan datos en la serie de estudio.
Los valores de la serie de referencia se denotan como x_i y los valores de la serie de estudio, cuyos datos no están completos, se denotan como y_i.

Para caracterizar los registros de las series, se toman aquellos períodos en los cuales los datos en ambas series están presentes, obteniéndose la media y desviación

estándar para cada serie. Luego se estiman los coeficientes de la regresión de Y con respecto a X para los períodos donde los datos en ambas series están completos, es decir, usando la ecuación de regresión lineal:

$$\hat{y}_i = a + bx_i$$

Dónde:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{Cov_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Donde:

$a \;\; y \;\; b$ Son las constantes de regresión,

Cov_{xy} Es la covarianza entre “x” y “y”.

S_x^2 Es la varianza de los valores observados

$\bar{x}$ Es la media de los valores observados

$\bar{y}$ Es la media de los valores estimados

Si el coeficiente de regresión es diferente de cero estadísticamente, según la prueba t, al 95% de confianza, con un coeficiente de determinación mayor que 75% y se cumplen los siguientes supuestos:

- Linealidad del modelo,
- varianza constante,
- independencia y
- normalidad de los errores.

Entonces se tendría un método estadístico para estimar los datos faltantes de la serie de estudio, simplemente reemplazando en la expresión obtenida, una vez identificado el tiempo en el cual falta el dato en la serie de tiempo, el valor correspondiente al mismo tiempo de ocurrencia del dato faltante de la serie de referencia.

Método de la Razón

Este método tiene una aplicación específica para estimar datos faltantes en series de lluvia y consiste en obtener la razón q (Barger, 1960, WMO 1966 y WMO 1983), a partir de pares de estaciones meteorológicas, de tal manera que sus valores mensuales, anuales o medios, tienden a ser constantes. Es decir, si se tienen dos estaciones (A y B), el procedimiento consiste en:

1. Una vez identificado en cada estación los datos comunes en ambas, obtener q como el cociente entre la sumatoria de los datos de la estación B (datos faltantes), con la sumatoria de los datos de la estación A (con todos los datos), es decir:

$$q = \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$$

Donde:

N: Número de registros en cada estación.

b_i: Registro de precipitación i de la estación B

a_i: Registro de precipitación i de la estación A

Lo cual indica que N es el número de registros de la estación B y que éste a su vez, es menor al número de registros de la estación A.

2. Luego de estimado q, se obtiene el valor faltante de la precipitación en la estación B, como

$$b_j = q * a_j$$

Donde:

b_j Es la precipitación estimada para el día faltante j

a_j Es la precipitación registrada en la estación de referencia el día j

Este método es utilizado, además, para valores mensuales y anuales.

Método de interpolación con otras estaciones

El método de interpolación fue propuesto por Paulhus y Kohler (1952), este método estima el dato de lluvia faltante, como el promedio de la precipitación ocurrida en tres estaciones adyacentes que están bajo la misma influencia topo-climática en el tiempo referente (dato faltante), siempre y cuando la precipitación anual de cada una de las tres estaciones adyacentes difiera descriptivamente, en menos del 10% de la precipitación anual de la estación con el dato faltante.

Existen otros métodos bivariantes y multivariantes, pero esos métodos no se abordarán debido a que no es parte de los objetivos de este trabajo.

Cabe mencionar que las técnicas de estimación descritas anteriormente ya han sido implementadas en algunas librerías de R entre las que se puede mencionar RClimTools; también podemos mencionar el paquete estadístico CHAC (Cálculo Hidrometeorológico de Aportaciones y Crecidas), el cual es un software especializado en el procesamiento de variables meteorológicas y entre sus aplicaciones podemos mencionar que CHAC permite la creación de cronogramas de los datos y completar datos faltantes para series climatológicas con ciertas limitaciones, esto se debe a que CHAC no puede trabajar con una gran cantidad de variables a la vez, y al decir gran cantidad nos referimos a mas de 100.

6.2 Análisis factorial múltiple como técnica de análisis descriptivo

Antes de abordar lo que es el análisis factorial múltiple hablaremos un poco sobre lo que es el análisis factorial en sí.

El análisis factorial es un nombre genérico que se le da a una clase de métodos estadísticos multivariante, cuyo propósito principal es definir la estructura subyacente de una matriz de datos.

De manera generalizada podemos decir que el análisis factorial aborda el problema de cómo analizar las interrelaciones (correlaciones) existentes entre un gran número de variables con la definición y ubicación de una serie de dimensiones subyacentes conocidas como factores.

El análisis factorial es básicamente una técnica de reducción de datos para encontrar grupos homogéneos de variables a partir de un numeroso grupo de variables observadas.

Esos grupos homogéneos están conformados por aquellas variables que tengan mayor correlación entre sí, lo que se espera además es que esos grupos sean independientes con otros.

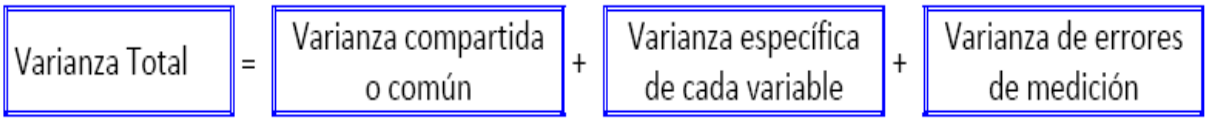
Así, el análisis factorial sirve para explicar el comportamiento de un conjunto de variables observadas a partir de un pequeño grupo de variables latentes no observadas, a estas variables no observadas se les denomina factores.

Los autores suelen mencionar que la técnica de análisis factorial debe ser una de las primeras en ser utilizadas, por el hecho de que puede tener un papel único en el uso de otras técnicas multivariantes.

Es necesario destacar que el análisis factorial es diferente de las técnicas de dependencia como lo son la regresión múltiple, el análisis discriminante, el análisis multivariante de varianza o la correlación canónica, la diferencia se da en que estas técnicas toman en cuenta una o más variables explícitamente como de criterio o dependientes y todas las demás son consideradas como independientes o de predicción.

Debemos decir además que en el análisis factorial se pueden emplear variables cualitativas y cuantitativas; para poder tener una métrica entre ellas se emplean la distancia euclidiana para las variables cuantitativas y la distancia chi-cuadrado para las cualitativas.

Debemos decir que al analizar un conjunto de variables, estas poseen una variabilidad total que se pretende explicar, esta variabilidad se puede descomponer en:



La variabilidad compartida es la que se da entre dos o más variables, mientras que la variabilidad de cada variable es la que posee cada una de ellas como es de esperarse.

Por otra parte, también se da la varianza de errores de medición, la cual es normal encontrar en el mundo real, ya que es casi imposible no cometer un error al medir un dato sin importar si es cuantitativo o cualitativo.

Es necesario decir que no nos extenderemos más en el estudio de los diversos tipos de análisis factoriales, debido a que no es de nuestro interés en profundizar en las diversas técnicas factoriales sino simplemente en la que emplearemos en el presente trabajo y más específicamente emplearemos el análisis factorial múltiple aplicado a variables cuantitativas ya que en nuestra base no disponemos de variables cualitativas. El Análisis Factorial Múltiple (AFM) tiene la característica de que muestra información en forma de una tabla de datos, donde un conjunto de Individuos están descritos por

un conjunto de variables agrupadas en diferentes sub-tablas (grupos) en función de: su naturaleza, tipología, características, número, correlaciones, etc. La única limitación radica en no mezclar en un mismo grupo, variables de diferente naturaleza.

Podemos formar grupos de variables cuantitativas o cualitativas pero de ninguna forma se puede configurar un grupo de variables entremezclando ambos tipos.

Esta importancia, sin embargo, no es imprescindible si el grupo de variables a formar no actúa en la formación de los factores sino sólo como descriptoras de la situación. En este sentido, ilustran el análisis y se les denomina conjuntamente grupo ilustrativo.

La tabla de datos X tiene I filas y K columnas, estructurada en J conjuntos.
Lo que nos servirá para identificar la posible relación entre las variables será la distancia entre dos individuos-grupos (i, g), esta es una medida de inercia que se define mediante la fórmula:

$$d^2(i,g) = \sum_{i=1}^{n} (X_{i,k} - X_{g,k})^2 * m_k$$

El AFM se sustenta en el concepto de inercia como una medida de dispersión y en la utilización de una distancia que determinará la métrica a seguir.

La metodología del AFM analiza individualmente cada grupo de variables o sub-tablas en diferentes fases a través de un análisis de componentes principales, ponderando cada sub-tabla por el inverso del primer valor propio del ACP, así equilibra el peso de los grupos para la obtención del primer factor de manera que el peso de cada grupo sea el mismo y no dependa del número y calidad de las variables empleadas.

Esto quiere decir que primero vamos a realizar una transformación de la matriz de los datos, por lo cual se realiza un ACP normado (es decir, centrado y reducido) de las tablas escogidas por el analista obteniendo los primeros valores propios de cada grupo.

Luego, se forma la tabla X, de la siguiente forma:

$$X_J = (\lambda_{11}^{-1/2} X_1 \quad \lambda_{12}^{-1/2} X_2 \quad \lambda_{13}^{-1/2} X_3 \quad \cdots \quad \lambda_{1J}^{-1/2} X_J)$$

Donde λ_{1J} es el primer valor propio que proveniente de un análisis de componentes principales realizado sobre la *j-ésima* tabla. Posteriormente se calculan las coordenadas de las variables de cada uno de los grupos respecto a los factores para obtener la representación global.

Como en toda técnica factorial, su objetivo inicial es estudiar las semejanzas existentes entre los individuos y las relaciones entre las variables.

Sin embargo, esta técnica incorpora al estudio una nueva dimensión. Contempla las relaciones entre grupos de variables, grupos de individuos y los factores encontrados.

Se puede detectar qué grupo de variables es el responsable de situar a un individuo o grupo en una u otra posición ante el factor y por tanto qué variable actúa directamente mejorando su posición mediante ese factor y/o frente a los individuos de su entorno.

Otra de las ventajas de esta técnica es la realización de comparaciones de los individuos en diferentes momentos en el tiempo. El procedimiento será posible si, tras detectar para cada momento las variables más representativas de los individuos a efectos de la investigación, dichas variables tienen el mismo comportamiento en la formación de cada uno de los factores que configuran los diferentes espacios.

En ese caso, las variables comunes en cada momento del tiempo y con una mayor correlación con cada uno de los factores determinarán el sentido o carácter de cada uno de los factores, planos y espacios factoriales.

Por otro lado, esta técnica incorpora con los Gráficas factoriales un gran potencial explicativo e interpretativo sin perder consistencia y rigurosidad en el carácter matemático.

Debemos aclarar que este y otros procedimientos multivariantes no es posible realizarlos manualmente, es decir, es muy difícil efectuar este procedimiento empleando una hoja de cálculo como Excel, ya que la cantidad de operaciones matriciales es exageradamente elevada, por lo cual se emplean paquetes estadísticos especializados como R.

6.3 Aplicación de metodología Box-Jenkins y los modelos ARIMA para series temporales

Iniciamos este apartado explicando que una serie temporal univariante consiste en un conjunto de observaciones de la variable de interés.

Para poder obtener un modelo de series temporales se debe tener en cuenta que este debe reproducir las características de la serie. Si contamos con T observaciones $Y_t\ ,\ t=1,\ 2,\ \cdots\ ,\ T$, el modelo univariante de series temporales se formularía en términos de los valores pasados de Y_t y/o su posición en relación con el tiempo. Se debe aclarar que no existe un único modelo para conseguir el resultado deseado.

Proceso estocástico y series temporales.

Un proceso estocástico es una familia de variables aleatorias que están relacionadas entre sí y siguen una ley de distribución conjunta.

Una serie temporal es una sucesión de observaciones en la que cada una de ellas corresponde a una variable aleatoria distinta, y la ordenación de la sucesión de observaciones es esencial para el análisis de la misma, no se puede alterar, porque se cambiaran las características de la serie que se quiere estudiar.

Características de un proceso estocástico

Un proceso estocástico se puede caracterizar bien por su función de distribución o por sus momentos.

Función de distribución.

Para conocer la función de distribución de un proceso estocástico es necesario conocer las funciones de distribución univariante de cada una de las variables aleatorias del proceso $F\left[Y_{t_i}\right], \forall\, t_i$, y las funciones bivariantes correspondientes a todo

par de variables aleatorias del proceso $F\left[Y_{t_i}, Y_{t_j}\right], \quad \forall (t_i, t_j)$, y todas las funciones trivariantes $F\left[Y_{t_i}, Y_{t_j}, Y_{t_k}\right], \quad \forall (t_i, t_j, t_k)$,...

En resumen, la función de distribución de un proceso estocástico incluye todas las funciones de distribución para cualquier subconjunto finito de variables aleatorias del proceso:

$$F\left[Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\right], \ \forall (t_1, t_2, \dots, t_n)$$

Siendo n un número finito de observaciones.

Momentos del proceso estocástico.

Como suele ser muy complejo determinar las características de un proceso estocástico a través de su función de distribución se suele recurrir a caracterizarlo a través de los dos primeros momentos.

El primer momento de un proceso estocástico viene dado por el conjunto de las medias de todas las variables aleatorias del proceso:

$$E(Y_t) = \mu_t, \qquad t = 0, \pm 1, \pm 2, \cdots$$

El segundo momento centrado del proceso viene dado por el conjunto de las varianzas de todas las variables aleatorias del proceso y por las covarianzas entre todo par de variables aleatorias:

$$V(Y_t) = E(Y_t - \mu_t)^2 = \sigma_t^2, \qquad t = 0, \pm 1, \pm 2, \cdots$$

$$Cov(Y_s, Y_t) = E(Y_s - \mu_s)(Y_t - \mu_t) = \gamma_{s,t}, \forall s, t \quad (s \neq t)$$

Si la distribución del proceso es normal y se conocen sus dos primeros momentos (medias, varianzas y covarianzas), el proceso está perfectamente caracterizado y se conoce su función de distribución.

Procesos estocásticos estacionarios

En el análisis de series temporales el objetivo es utilizar la teoría de procesos estocásticos para determinar qué proceso estocástico ha sido capaz de generar la serie temporal bajo estudio con el fin de caracterizar el comportamiento de la serie y efectuar predicciones a futuro.

Si se quieren conseguir métodos de predicción consistentes, no se puede utilizar cualquier tipo de proceso estocástico, sino que es necesario que la estructura probabilística del mismo sea estable en el tiempo.

Lo que se hace básicamente es aprender de las regularidades del comportamiento pasado de la serie y luego proyectarse hacia el futuro. Por lo tanto, es preciso que los procesos estocásticos generadores de las series temporales tengan algún tipo de estabilidad.

Si, por el contrario, en cada momento de tiempo presentan un comportamiento diferente e inestable, no se pueden utilizar para predecir. A estas condiciones que se les impone a los procesos estocásticos para que sean estables para predecir, se les conoce como estacionariedad.

El concepto de estacionariedad se puede caracterizar en términos de la función de distribución o a partir de los momentos del proceso. En el primer caso, se hablará de estacionariedad en sentido estricto y, en el segundo, de estacionariedad de segundo orden o en covarianza.

Estacionariedad estricta.

Un proceso estocástico Y_t, es estacionario en sentido estricto si la función de distribución de cualquier conjunto finito de n variables aleatorias del proceso no se altera si se desplaza k periodos en el tiempo.

Es decir:

$$F\left[Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}\right] = F\left[Y_{t_1+k}, Y_{t_2+k}, \ldots, Y_{t_n+k}\right], \ \forall (t_1, t_2, \ldots, t_n) \text{ y } \forall k.$$

Estacionariedad en covarianza.

Un proceso estocástico Y_t , es estacionario en covarianza si cumple con las siguientes condiciones:

a) Es estacionario en media, es decir, todas las variables aleatorias del proceso tienen la misma media y es finita:

$$E(Y_t) = \mu, \quad \forall t.$$

b) Todas las variables aleatorias tienen la misma varianza y es finita, es decir, la dispersión en torno a la media constante a lo largo del tiempo es la misma para todas las variables del proceso.

$$V(Y_t) = \sigma_t^2, \quad \forall t.$$

c) La covarianza lineal entre dos variables aleatorias del proceso que disten k periodos de tiempo es la misma que existe entre cualesquiera otras dos variables que estén separadas también k periodos, independientemente del momento concreto de tiempo al que estén referidas, es decir, un proceso es estacionario en covarianza si se cumplen:

$$Cov(Y_t, Y_s) = \gamma_k.$$

Si un proceso estocástico es estacionario en covarianza y su distribución es Normal, entonces es estacionario en sentido estricto.

Función de autocovarianzas y de autocorrelación

En principio, si se considera el proceso estocástico teórico Y_t, que comienza en algún momento del pasado lejano y acaba en un futuro indeterminado, se pueden calcular un número indefinido de autocovarianzas, por lo cual es necesario definir una función que las agrupe a todas.

Función de autocovarianzas (FACV)

La función de autocovarianzas de un proceso estocástico estacionario es una función de k (numero de periodos de separación entre las variables) que recoge el conjunto de las autocovarianzas del proceso y se denota por:

$$\gamma_k = E[Y_t - \mu][Y_{t-k} - \mu], \quad para \quad k = 0, 1, 2, 3, \cdots$$

Características de la función de autocovarianzas:

- Incluye la varianza del proceso.

$$\gamma_0 = E[Y_t - \mu][Y_t - \mu] = \sigma_t^2$$

- Es una función simétrica:

$$\gamma_k = E[Y_t - \mu][Y_{t-k} - \mu] = E[Y_{t-k} - \mu][Y_t - \mu] = \gamma_{-k}$$

La función de autocovarianzas de un proceso estocástico recoge toda la información sobre la estructura dinámica lineal del mismo, pero depende de las unidades de medida de la variable, por lo que, en general, se suele utilizar la función de autocorrelación.

Función de autocorrelación (FAC)

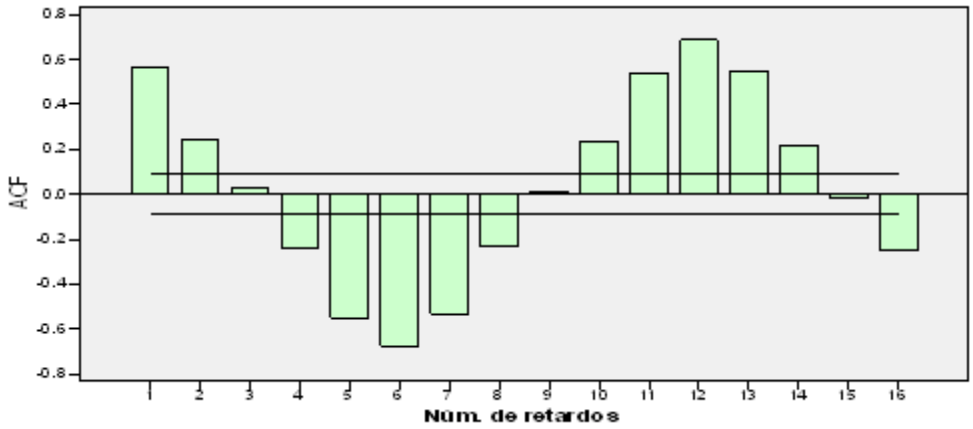
También conocida en algunos libros de series temporales como función de autocorrelación simple (FAS) y corresponde al coeficiente de autocorrelación de orden k de un proceso estocástico estacionario y mide el grado de asociación lineal existente entre dos variables aleatorias del proceso, dichas variables están separadas k periodos:

$$\rho_k = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{V(Y_t)\ V(Y_{t+k})}} = \frac{\gamma_k}{\gamma_0} \qquad , k = 0, \pm 1, \pm 2, \cdots$$

Por ser un coeficiente de correlación, no depende de unidades y $|\rho_k| \leq 1, \forall k$. .

La función de autocorrelación de un proceso estocástico estacionario es una función que depende del valor de k, que recoge el conjunto de los coeficientes de autocorrelación del proceso.

La función de autocorrelación se suele representar gráficamente por medio de un Gráfica de barras denominado correlograma, que en su eje “x” se encuentran los valores del tiempo de manera discreta y en su eje “y” los valores de autocorrelación existentes entre las observaciones. Ejemplo de un correlograma se muestra a continuación en el Gráfica 1.



Gráfica 1: Ejemplo de correlograma de la FAS

Las características de la función de autocorrelación de un proceso estocástico estacionario son:

- El coeficiente de autocorrelación de orden 0 es, por definición, 1. Por eso, a menudo, no se le incluye explícitamente en la función de autocorrelación.
- Es una función simétrica:

  $\rho_k = \rho_{-k}$.

  Por ello, en el correlograma se representa la función de autocorrelación solamente para los valores positivos del retardo k.
- La función de autocorrelación de un proceso estocástico estacionario tiende a cero rápidamente cuando k tiende al infinito.

La función de autocorrelación va a ser uno de los principales instrumentos utilizados para recoger la estructura dinámica lineal del modelo.

Aunque para el proceso estocástico teórico se cuenta con un número indefinido de autocovarianzas y coeficientes de autocorrelación, cuando se dispone de una serie temporal finita de tamaño T, como máximo se pueden estimar T−1 coeficientes de autocorrelación, pero, en la práctica, se van a estimar muchos menos. Se recomienda un máximo de $T/3$. Esto es debido a que cuanto mayor sea k menos información hay para estimar ρ_k y la calidad de la estimación es menor.

Función de autocorrelación parcial (FAP)

La función de autocorrelación parcial mide la correlación existente entre dos momentos en el tiempo, luego de haber eliminado el efecto de los momentos intermedios.

$$\phi_{1,1} = \rho_1 = \phi_1 \qquad , \phi_{j,j} = \frac{\rho_j - \sum_{i=1}^{j-1} \phi_{j-1,i}\ \rho_{j-i}}{1 - \sum_{i=1}^{j-1} \phi_{j-1,i}\ \rho_i} \qquad , \mathrm{j}=2,3,\cdots$$

Proseguimos mostrando los distintos modelos asociados a las series de tiempo.

Proceso Ruido Blanco

El proceso estocástico más sencillo es el denominado Ruido Blanco que es una secuencia de variables aleatorias de media cero, varianza constante y covarianzas nulas.

Se denotara habitualmente por a_t.

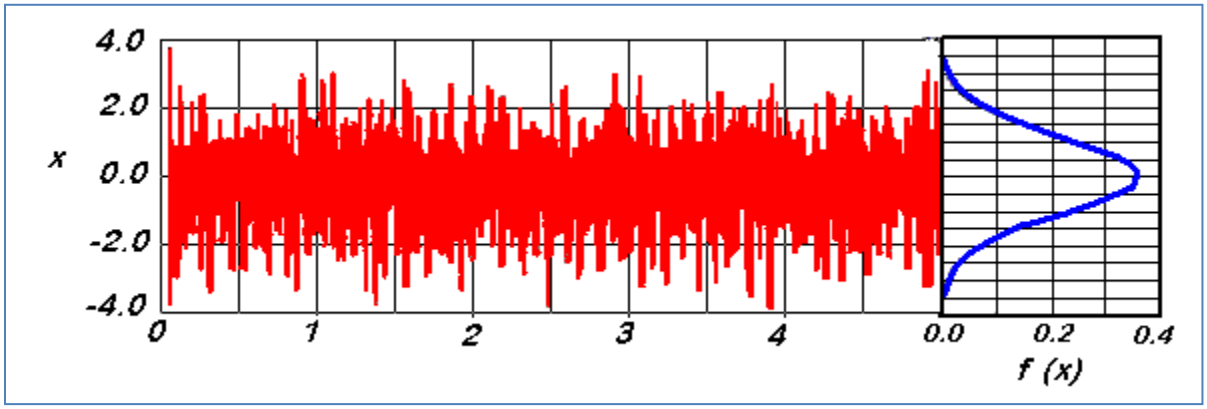
$$E(a_t) = 0, \qquad \forall t.$$
$$V(a_t) = \sigma^2, \qquad \forall t.$$
$$Cov(a_i, a_j) = 0, \qquad \forall i \neq j.$$

Así, un proceso ruido blanco a_t, es estacionario si la varianza es finita.

Dado que la característica fundamental de las series temporales es la dependencia temporal entre sus observaciones, presentan un comportamiento que responde a una ausencia de correlación temporal en el sentido de que lo que ocurre hoy no tiene

relación lineal con lo sucedido en el pasado. Aun así, el proceso ruido blanco es muy útil en el análisis de series temporales porque es la base para la construcción de los modelos ARIMA (p, d, q).

Un Gráfica de secuencia de un proceso de ruido blanco se muestra a continuación:



Gráfica 2. Ejemplo de una serie ruido blanco.

Modelos lineales estacionarios

La metodología de la modelización univariante es sencilla. Dado que el objetivo es explicar el valor que toma en el momento t una variable que presenta dependencia temporal, una forma de trabajar es recoger información sobre el pasado de la variable, observar su evolución en el tiempo y explotar el patrón de regularidad que muestran los datos.

La estructura de dependencia temporal de un proceso estocástico está recogida en la función de autocovarianzas (FACV), en la función de autocorrelación (FAC) y función de autocorrelación parcial (FAP), por lo cual, se trata de utilizar la información de estas funciones para extraer un patrón sistemático, y a partir de este, un modelo que reproduzca el comportamiento de la serie y se pueda utilizar para predecir. Este procedimiento se hará operativo mediante los modelos ARMA y ARIMA que son una aproximación a la estructura teórica general.

En un modelo de series temporales univariante se descompone la serie Y_t en dos partes, una que recoge el patrón de regularidad o parte sistemática, y otra parte puramente aleatoria, así:

$$Y_t = PS_t + a_t \ , \quad t = 1, \ 2, \ \ldots$$

La parte sistemática (PS) es la parte predecible con el conjunto de información que se utiliza para construir el modelo y la parte aleatoria (a_t) son valores que no tienen ninguna relación o dependencia entre sí.

A la hora de construir un modelo estadístico para una variable temporal, el problema es formular la parte sistemática de tal manera que el elemento residual sea una variable que tenga una distribución normal con media cero y varianza constante.

Dada una serie temporal de media cero, como el valor de Y en el momento t depende de su pasado, un modelo teórico capaz de describir su comportamiento sería:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \cdots) + a_t \ , \quad t = 1, \ 2, \ \ldots$$

Donde se exige que el comportamiento de Y_t sea una función de sus valores pasados, posiblemente infinitos.

Dentro de los procesos estocásticos estacionarios se considerara únicamente la clase de procesos lineales que se caracterizan porque se pueden representar como una combinación lineal de variables aleatorias. De hecho, en el caso de los procesos estacionarios con distribución normal y media cero, la teoría de procesos estocásticos señala que Y_t se puede expresar como una combinación lineal de los valores pasados infinitos de Y_t más una perturbación aleatoria ó ruido blanco:

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + a_t$$

Las condiciones generales que ha de cumplir el proceso son:

a) Que el proceso sea no anticipante, es decir, que el presente no venga determinado por el futuro, luego el valor de Y en el momento t no puede depender de valores futuros de la serie o de las perturbaciones aleatorias.

b) Que el proceso sea invertible, es decir, que el presente dependa de forma convergente de su propio pasado lo que implica que la influencia de Y_{t-k} en Y_t ha de ir disminuyendo conforme nos alejemos en el pasado.

El modelo general se puede escribir de forma más compacta en términos del operador de retardos, de la siguiente forma:

$$Y_t = \left(\pi_1 B + \pi_2 B^2 + \pi_3 B^3 + \cdots\right) Y_t + a_t$$

Donde el operador de retardo B lo que hace es lo siguiente:

$$B^k Y_t = Y_{t-k}$$

Otra forma alternativa de escribir el modelo lineal general es:

$$Y_t = \frac{1}{\pi_1 B + \pi_2 B^2 + \pi_3 B^3 + \cdots} a_t = (1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots) * a_t$$

Esto significa que el valor de Y_t se puede representar como la combinación lineal del ruido blanco a_t y su pasado infinito.

Acudiendo a la teoría de polinomios, bajo condiciones muy generales, se puede aproximar un polinomio de orden infinito mediante un cociente de polinomios finitos:

$$\pi_1 B + \pi_2 B^2 + \pi_3 B^3 + \cdots \approx \frac{1 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 + \cdots + \psi_p B^p}{1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q}$$

Por lo tanto, el modelo lineal general admite tres representaciones y todas ellas son igualmente validas bajo los supuestos señalados:

- Representación puramente autorregresiva ó $AR(\infty)$:

El valor presente de la variable se representa en función de su propio pasado más una perturbación aleatoria.

- Representación puramente de medias móviles ó $MA(\infty)$:

El valor presente de la variable se representa en función de todas las innovaciones presentes y pasadas.

- Representación finita:

En este modelo finito, el valor de Y_t depende de los valores pasados hasta el momento t − p (parte autorregresiva), de la perturbación aleatoria y su pasado hasta el momento t − q (parte medias móviles).

Este modelo se denomina proceso Autorregresivo de Medias Móviles de orden (p, q), y se denota por $ARMA(p,q)$.

Dos casos particulares del modelo $ARMA(p,q)$ de gran interés son:

- $AR(p)$. Modelo que solo presenta parte autorregresiva, es decir, el polinomio de medias móviles es de orden 0 ($q = 0$),

- $MA(q)$. Modelo que solo presenta la parte medias móviles, es decir, el polinomio autorregresivo es de orden 0 ($p = 0$)

Cuando el modelo es conocido se puede utilizar cualquiera de las tres representaciones dependiendo de los intereses. Si el modelo no es conocido y hay que especificarlo y estimarlo a partir de una serie temporal concreta, hay que utilizar necesariamente la formulación finita.

Cuando se construye un modelo de series temporales univariante el objetivo no es conseguir el “verdadero” modelo. Es preciso ser conscientes de que estamos tratando de modelar una realidad compleja y el objetivo es lograr un modelo parsimonioso y suficientemente preciso que represente adecuadamente las características de la serie recogidas fundamentalmente en la función de autocorrelación.

Los modelos $ARMA(p,q), AR(p)$ y $MA(q)$ son aproximaciones al modelo lineal general.

A continuación abordaremos los diferentes modelos de forma resumida y se omitirán las demostraciones debido a que estas se pueden encontrar en cualquier libro de serie de tiempo y además porque no es parte de los objetivos el profundizar en los modelos ARIMA.

Procesos autorregresivos.

El modelo o proceso autorregresivo es aquel en el cual los valores presentes pueden ser expresados en términos de los valores de su pasado más una perturbación aleatoria, cabe mencionar que dichos modelos pueden o no incluir una constante de nivel.

Debemos mencionar que todos los procesos autorregresivos son invertibles pero no todos los procesos autorregresivos son estacionarios.

Proceso autorregresivo de orden uno ($\boldsymbol{AR(1)}$).

En el proceso autorregresiva de orden uno, Y_t viene determinado únicamente por su valor pasado un periodo atrás Y_{t-1} multiplicado por una taza de transferencia más una perturbación aleatoria:

$$Y_t = \phi Y_{t-1} + a_t$$

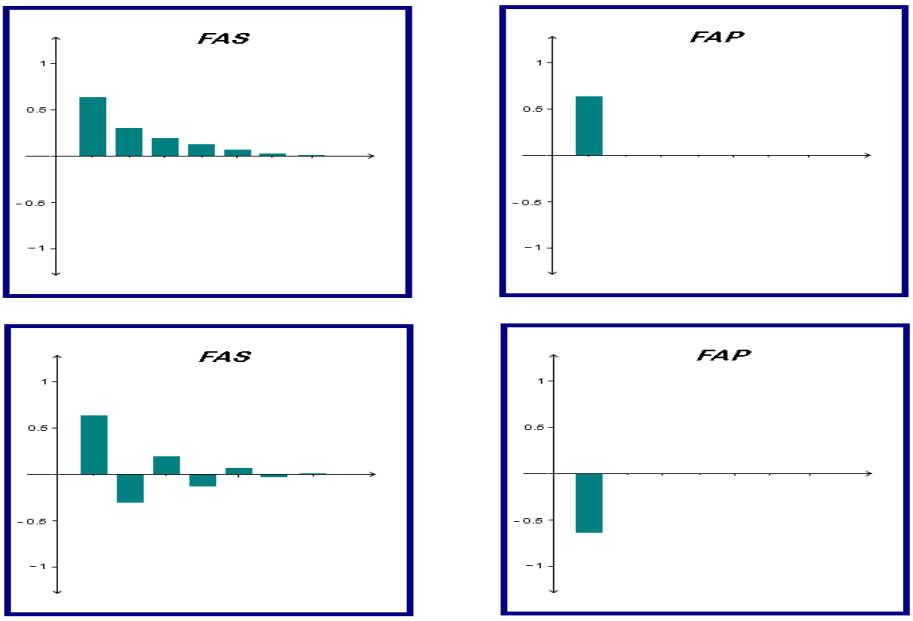
Los valores más alejados en el pasado tienen poca influencia en el valor presente y valores futuros de Y_t.

El parámetro del modelo autorregresivo $AR(1)$ debe cumplir que $\phi < 1$ para que el proceso sea estacionario e invertible; el valor de la varianza y de las autocovarianzas deben ser finitas, y además las autocovarianzas deben depender únicamente de los períodos de separación entre las variables y no del tiempo.

Ahora bien, el comportamiento de su FAS tiene una forma exponencial o sinusoidal descendente y su velocidad de descendencia depende del valor de ϕ, ya que si este valor es cercano a uno en valor absoluto, el descenso será lento y si es cercano a cero su descenso será acelerado.

Debemos mencionar también que si el valor de ϕ es mayor que la unidad, entonces el proceso será explosivo, lo que significa que los valores de la serie crecen de manera exponencial y en consecuencia el proceso no es estacionario.

El comportamiento de su FAP tiene un único valor el cual es el coeficiente ϕ.
A continuación se muestran los posibles correlogramas de la FAS y la FAP de un $AR(1)$.



Gráfica 3. FAS y FAP teóricas de un proceso AR(1).

Proceso autorregresivo de orden dos ($AR(2)$).

En el proceso autorregresivo de orden dos viene determinado únicamente por su valor pasado uno y dos periodos atrás más una perturbación aleatoria, así:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t$$

Los valores más alejados en el pasado tienen poca influencia en el valor presente y valores futuros de Y_t.

Los parámetros del modelo autorregresivo $AR(2)$ deben cumplir con varias condiciones para que este sea estacionario e invertible; estas son:

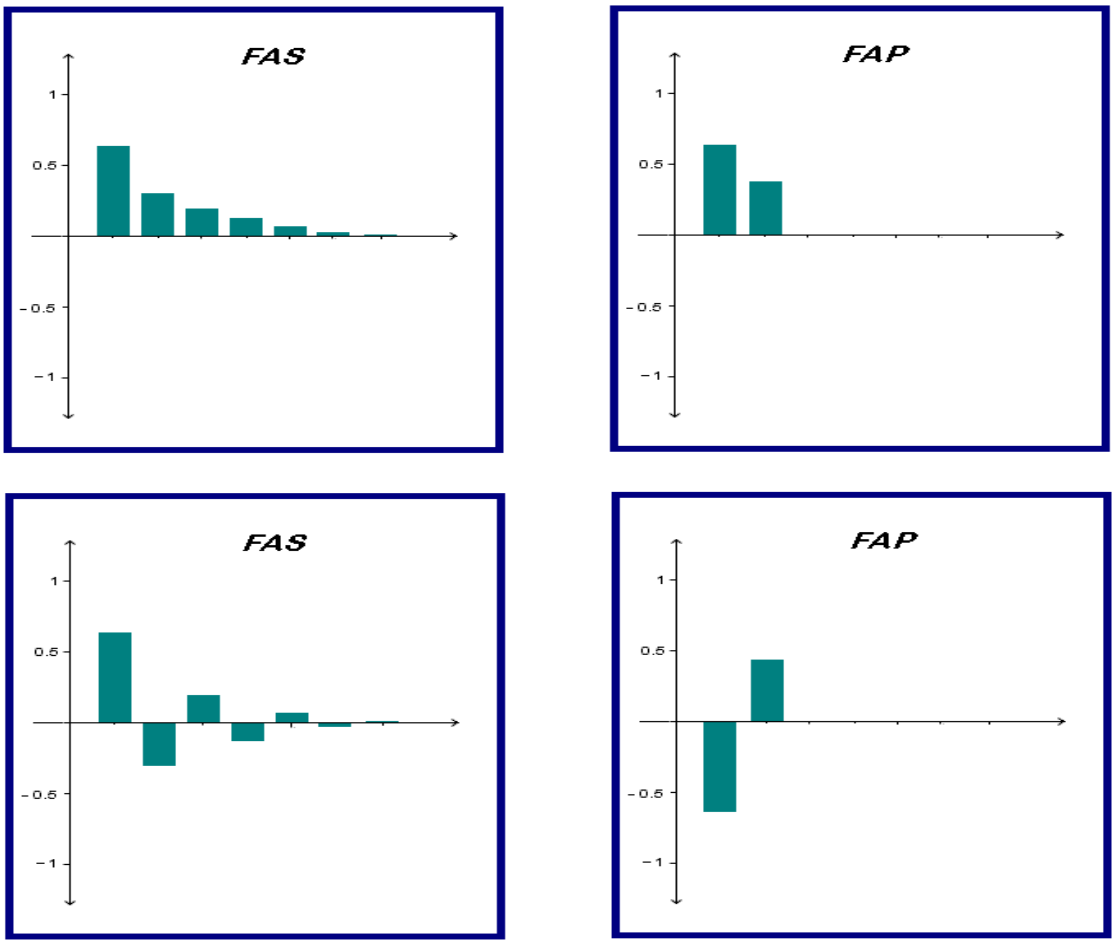
$$|\phi_1 + \phi_2| < 1$$
$$|\phi_2 - \phi_1| < 1$$
$$|\phi_2| < 1$$

La varianza y las autocovarianzas son finitas, y además dependen únicamente de los periodos de separación entre las variables y no del tiempo.

El comportamiento de su FAC tiene una forma exponencial o sinusoidal descendente y su velocidad de descendencia depende de los valores de los coeficientes ϕ, si el proceso cumple con las condiciones descritas, entonces el descenso puede tener diversas variantes dependiendo del signo de los coeficientes, pero en resumen será exponencial o sinusoidal. Debemos mencionar también que si los valores de ϕ_1 y ϕ_2 no cumplen con las condiciones descritas, entonces el proceso será explosivo y por lógica no estacionario.

El comportamiento de su FAP tiene los primeros dos valores los cuales son los coeficientes ϕ_1 y ϕ_2 .

A continuación se muestran los posibles correlogramas de la FAS y la FAP de un $AR(2)$.



Gráfica 4. FAS y FAP teóricas de un proceso AR(2).

Proceso autorregresivo de orden p ($AR(p)$)

El proceso autorregresivo de orden p, expresa Y_t en función de su pasado hasta el retardo t − p más una perturbación aleatoria:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + a_t$$

Poseen media y varianza constantes, sus covarianzas son finitas y dependen de la distancia entre las variables.

Las condiciones de estacionariedad se cumplen si: $\sum_{i=1}^{p} \phi_i < 1$, $\sum_{i=1}^{p} |\phi_i| < 1$

La primera de las condiciones es necesaria pero no suficiente para que el modelo sea estacionario, con la segunda condición nos aseguramos que el modelo sea definitivamente estacionario, aunque debemos mencionar que es una condición suficiente pero no necesaria.

Ahora bien, para valores de p > 1 al ser el modelo más complejo, la estructura de la FAS también puede presentar una gran variedad de formas.

En general, el comportamiento de su FAS tiene una forma exponencial o sinusoidal descendente y su velocidad de descendencia depende de los valores de los coeficientes ϕ; si el proceso cumple con las condiciones de estacionariedad anteriormente descritas, entonces el descenso será exponencial o sinusoidal acelerado.

El comportamiento de su FAP tiene los primeros p valores diferentes de cero, los cuales son los coeficientes ϕ y el resto de valores aparecerán bajo las bandas de confianza indicando que no son significativamente diferentes de cero.

Procesos de Medias Móviles.

El modelo de medias móviles consiste en expresar el valor presente de la serie como una combinación lineal de las perturbaciones aleatorias

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Debemos mencionar que todos los procesos de medias móviles son estacionarios, ya que en el futuro lejano se estabilizan en un valor constante y por lo cual, solo es necesario revisar la característica de invertibilidad, es decir, ver si se pueden expresar como un modelo autorregresivo y esto se debe a que es más adecuado predecir los valores futuros empleando los modelos autorregresivos que los de medias móviles, ya que presentan memoria más larga.

Proceso de medias móviles de orden uno ($MA(1)$).

El modelo $MA(1)$ determina el valor de Y en el momento t en función de la perturbación aleatoria y su primer retardo:

$$Y_t = a_t - \theta a_{t-1}$$

En principio, un modelo de medias móviles no parece ser adecuado para predecir ya que depende de valores aleatorios y no depende directamente de las observaciones pasadas, pero es de notar que los modelos de medias móviles también se pueden escribir de forma autorregresiva.

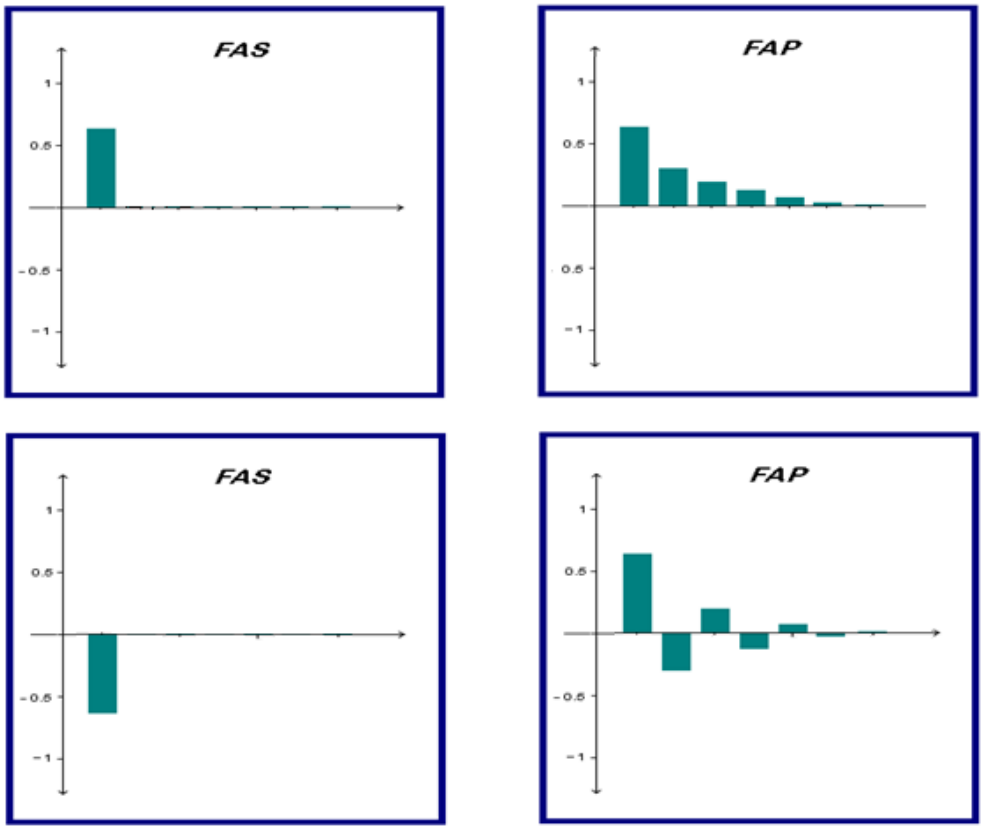
Esta representación será un $AR(\infty)$ restringido ya que todos los parámetros autorregresivos dependen del único parámetro de medias móviles del modelo, θ.

Para que el modelo $MA(1)$ sea invertible es preciso que su representación autorregresiva sea convergente, esto significa que la influencia de Y_{t-k} vaya siendo menor conforme nos alejamos en el futuro.

Para que se cumpla la condición anterior es necesario y suficiente que $|\theta| < 1$, por lo que el modelo $MA(1)$ no es siempre invertible. Bajo las condiciones de invertibilidad, el modelo $MA(2)$ y cualquier modelo de medias móviles, se puede escribir en forma autorregresiva, por lo que si deseamos una representación autorregresiva para el modelo, no hace falta empezar por un $AR(p)$.

La media, la varianza y las covarianzas del proceso son finitas, la función FAS muestra un único valor significativamente diferente de cero y la función FAP muestra un decrecimiento acelerado dependiendo del valor del parámetro θ, ya que si este es cercano a uno en valor absoluto, el decrecimiento es lento y si es cercano a cero el decrecimiento es acelerado.

A continuación se muestran las funciones de autocorrelación simple y parcial teóricas de un $MA(1)$



Gráfica 5. FAS y FAP teóricas de un proceso MA(1)

Proceso de medias móviles de orden dos ($\boldsymbol{MA(2)}$).

El modelo $\boldsymbol{MA(2)}$ determina el valor de Y en el momento t en función de la perturbación aleatoria, así como de su primer y segundo retardo:

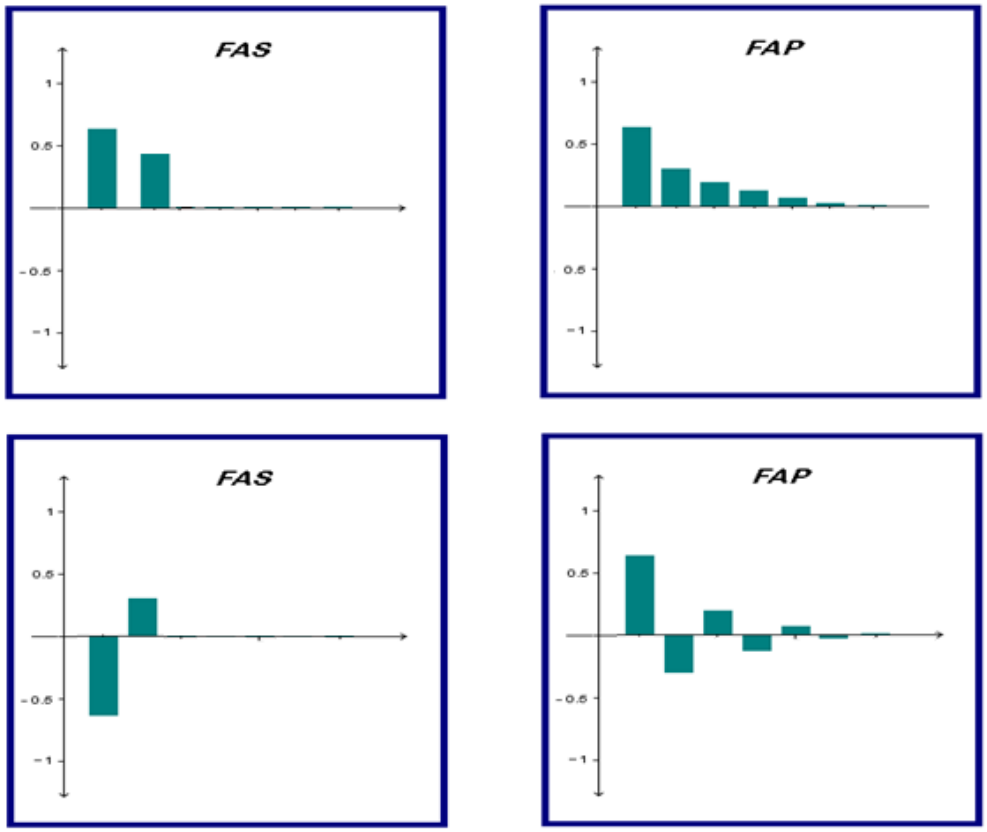
$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

Para que el modelo $\boldsymbol{MA(2)}$ sea invertible debe cumplir las siguientes condiciones:

$$|\theta_1 + \theta_2| < 1$$
$$|\theta_2 - \theta_1| < 1$$
$$|\theta_2| < 1$$

La media, la varianza y las covarianzas del proceso son finitas, la función FAC muestra dos valores significativamente diferentes de cero y la función FAP muestra un decrecimiento que puede ser exponencial, sinusoidal o alternado.

A continuación se muestran las funciones de autocorrelación simple y parcial teóricas de un $MA(2)$



Gráfica 6. FAS y FAP teóricas de un proceso MA(2)

Proceso de medias móviles de orden q ($MA(q)$).

El modelo medias móviles de orden q expresa el valor de Y_t en función de la perturbación aleatoria y de su pasado hasta el retardo q:

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Como el número de parámetros de un $MA(q)$ es finito, la condición de estacionariedad se cumple para cualquier valor de sus parámetros.

Un proceso de medias móviles finito $\boldsymbol{MA(q)}$ es invertible si y solo si el modulo de las raíces del polinomio de medias móviles $\theta q\,(L)$ está fuera del circulo unidad.

Por lo general, para evitar cálculos engorrosos se emplean las siguientes condiciones:

$$\sum_{i=1}^{q} \theta_i < 1 \qquad , \qquad \sum_{i=1}^{q} \left|\theta_i\right| < 1$$

La primera condición es necesaria pero no suficiente para que el modelo sea invertible, con la segunda condición nos aseguramos que el modelo es definitivamente invertible.

Procesos autorregresivos de medias móviles (ARMA)

Un proceso autorregresivo de medias móviles finito ARMA (p, q) es estacionario si y solo si el modulo de las raíces del polinomio autorregresivo está fuera del circulo unidad.

Las condiciones de estacionariedad del modelo $ARMA(p,q)$ vienen impuestas por la parte autorregresiva, dado que la parte de medias móviles finita siempre es estacionaria.

Un proceso autorregresivo de medias móviles finito $ARMA(p,q)$ es invertible si y solo si el modulo de las raíces del polinomio medias móviles está fuera del circulo unitario.

Las condiciones de invertibilidad del modelo $ARMA(p,q)$ vienen impuestas por la parte de medias móviles, dado que la parte autorregresiva finita siempre es invertible porque está directamente escrita en forma autorregresiva.

El modelo $ARMA(p,q)$ va a compartir las características de los modelos $AR(p)$ y $MA(q)$, ya que contiene ambas estructuras a la vez. El modelo $ARMA(p,q)$ tiene media y varianza constantes y finitas y una función de autocovarianzas infinita. La función de autocorrelación es infinita decreciendo rápidamente hacia cero pero sin truncarse.

Modelos lineales no estacionarios

Los modelos presentados anteriormente se basaban en el supuesto de estacionariedad en covarianza, es decir, en procesos donde la media y la varianza son constantes y finitas y las autocovarianzas no dependen del tiempo sino solo del número de periodos de separación entre las variables. Pero la mayoría de las series del mundo real no siempre se comportan de forma estacionaria, bien porque suelen ir cambiando de nivel en el tiempo o porque la varianza no es constante.

No estacionariedad en varianza

Cuando una serie no es estacionaria en varianza, significa que no se puede sostener el supuesto de que ha sido generada por un proceso con varianza constante en el tiempo, la solución es transformar la serie mediante algún método que estabilice la varianza.

El comportamiento habitual en las series de tiempo suele ser que la varianza cambie conforme el nivel de la serie cambia. La transformación más adecuada es la logarítmica y nos proporcionará una varianza constante. Pero si la varianza de la serie es proporcional a su nivel, entonces habría que tomar la raíz cuadrada de la serie para obtener una varianza constante. En general, para estabilizar la varianza se utilizan las transformaciones Box-Cox.

Transformaciones de Box y Cox

Las transformaciones de Box-Cox son una familia de transformaciones potenciales usadas en estadística para corregir sesgos en la distribución de errores, para corregir varianzas desiguales y principalmente para corregir la no linealidad en la relación mejorando con ello la correlación entre las variables. Esta transformación recibe el nombre de los estadísticos George E. P. Box y David Cox.

Estas transformaciones están definidas como sigue:

$$Y_t^{(\lambda)} = \begin{cases} K_1 * (Y_t^{\lambda} - 1) & , \lambda \neq 0 \\ k_2 * Log(Y_t) & , \lambda = 0 \end{cases}$$

Donde λ es el parámetro de transformación, K_2 es la media geométrica de los valores de la serie y K_1 depende de K_2 y del valor de λ, de la siguiente forma:

$$k_2 = \left(\prod_{i=1}^{n} Y_i \right)^{1/n} \qquad k_1 = \frac{1}{\lambda * k_2^{\lambda - 1}}$$

De aquí surge la pregunta: ¿Cómo elegir el valor de lambda (λ) más adecuado? Para responder a la pregunta, se siguen los siguientes pasos:

1- Primero se debe seleccionar el rango de valores de lambda de donde se quiere elegir el que logra que la transformación se acerque lo más posible a los datos.
2- Luego, para cada valor de lambda se realiza la transformación del paso anterior.
3- Finalmente se sustituyen los valores de la o las variables explicativas en las diferentes funciones y se calculan los cuadrados de los residuales estadísticos. Aquel valor de lambda (λ) que tenga el menor valor de la suma de los cuadrados de los residuales será la mejor opción de lambda (λ).

Nótese que K_2 es un valor fijo para todos los casos y que sólo hay que calcular de nuevo el valor de K_1.

No se profundiza mucho en este tema ya que el procedimiento de elegir lambda está mecanizado en diferentes paquetes estadísticos y aun cuando se tuviera el mejor valor de lambda (λ), esto no garantiza una ecuación que nos sea de utilidad, por lo cual en la mayoría de ocasiones nos valemos de la experiencia y de la ayuda visual para realizar la transformación de Box-Cox más adecuada.

No estacionariedad en media.

Cuando la serie presenta no estacionariedad en media, lo más recomendable es aplicar diferenciación, la diferenciación consiste en restarle a los valores de la serie actual los valores de la serie k periodos atrás, de la siguiente forma:

$$Y_t^* = Y_t - Y_{t-k} \quad , \quad k \geq 1 .$$

Se toman diferencias hasta que la serie sea estacionaria, es decir, aplicar el método de prueba y error, esto significa que antes de diferenciar la serie debemos probar si posee raíces unitarias en su polinomio autorregresivo, además nos valemos de los Gráficas de la serie y de los Gráficas de las funciones FAS y FAP.

Sobrediferenciación

Se dice que la serie está sobrediferenciada si se toman más diferencias de las necesarias, por ejemplo, si se elige un orden de diferenciación d cuando la serie diferenciada d-1 veces ya es estacionaria.

En este punto conviene recordar que si se diferencia un proceso estacionario sigue siendo estacionario. Por lo tanto, en principio, el hecho de que la serie diferenciada d veces sea estacionaria, no significa necesariamente que la serie diferenciada d-1 veces no lo sea, por lo que hay que tener cuidado ya que el objetivo es determinar el menor número de diferencias d capaz de convertir a una serie en estacionaria.

Para decidir cuál es valor de “d” más apropiado para la serie bajo estudio utilizaremos los siguientes instrumentos:

a) Gráfica de la serie original y las transformaciones correspondientes, para observar si se cumple o no la condición de estacionariedad que consiste en oscilar en torno a un nivel constante.

b) Correlograma estimado de la serie original y de las transformaciones correspondientes, para comprobar si decrece rápidamente hacia cero o no.

c) Contrastes de raíces unitarias.

Las pruebas de raíces unitarias proporcionan contrastes estadísticos que permiten, a partir del conjunto de información, hacer inferencia sobre la existencia o no de una raíz unitaria en una serie, es decir, sobre la no estacionariedad en media de la serie. Si se rechaza la hipótesis nula de existencia de raíz unitaria, no se diferenciara más la serie.
En caso contrario, si no se rechaza la hipótesis nula se tomara una diferencia más de orden 1.

De entre las pruebas de raíces unitarias se encuentran las siguientes:

- Contraste de Dickey-Fuller.
- Contraste de Dickey-Fuller aumentado.
- Contraste de Phillips – Perron
- Contraste de Kwiatkowski–Phillips–Schmidt–Shin (KPSS)

Modelo de paseo aleatorio

El modelo de paseo aleatorio es simplemente un modelo AR (1) con parámetro $\phi = 1$:

$$Y_t = Y_{t-1} + a_t$$

Donde a_t es una perturbación aleatoria o ruido blanco.

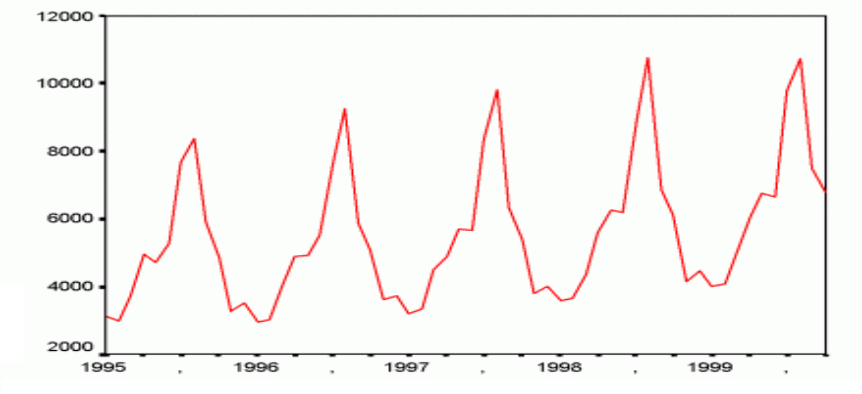
El modelo de paseo aleatorio no es estacionario porque la raíz del polinomio autoregresivo no tiene modulo mayor que la unidad.

En lo que se refiere a la FAC del modelo paseo aleatorio, se caracteriza porque sus coeficientes decrecen muy lentamente.

Modelos estacionales

Cuando se habla de series estacionales de entrada se puede decir que la serie no es estacionaria y por lo tanto antes de elegir un modelo adecuado hay que diferenciarla para poder estabilizarla.

Un ejemplo de este tipo de series se muestra en la gráfica 7 a continuación:



Gráfica 7. Ejemplo de una serie estacional.

Esto conlleva que a la hora de elaborar el modelo ARIMA adecuado para una serie temporal de este tipo se ha de tener en cuenta el comportamiento estacional, si lo hubiere, esto se debe a que la observación de un periodo y una observación del mismo periodo pero del año anterior tienen una pauta de comportamiento similar por lo que estarán temporalmente correlacionadas.

Por lo tanto, el modelo de series temporales ARIMA apropiado para este tipo de series deberá recoger las dos clases de dependencia temporales que presentan, estas son:

- La relación lineal existente entre observaciones sucesivas (comportamiento tendencial o regular) y

- La relación lineal existente entre observaciones del mismo mes en años sucesivos (comportamiento estacional).

Nótese que se habla de modelos ARIMA y no ARMA y esto es porque para estabilizar la serie por lo general se diferencia d veces en la parte regular y/o D veces en la parte estacional.

Antes de especificar un modelo apropiado, dentro del marco ARIMA, para una serie con tendencia y estacionalidad comenzaremos por estudiar las características de la dependencia lineal estacional a través de unos modelos muy sencillos, los modelos estacionales puros.

Modelos estacionales puros

Se entiende por modelo estacional puro aquel que recoge únicamente relaciones lineales entre observaciones del mismo mes para años sucesivos, es decir, entre observaciones separadas s periodos o múltiplos de s, donde s = 4 si la serie es trimestral, s = 12 si la serie es mensual, etc.

En teoría, “D” es el número de las diferencias estacionales necesarias que se han de aplicar para convertir a la serie en estacionaria; “D” puede tomar cualquier valor en los números naturales dependiendo de las características de la serie, aunque en la práctica nunca es superior a uno.

El modelo estacional puro que se denomina $ARIMA(P,D,Q)_S$,

Donde

P: es el orden del polinomio autorregresivo estacional,

Q: es el orden del polinomio medias móviles estacional y

D: es el número de diferencias estacionales que es necesario aplicar a la serie Y_t para que sea estacionaria.

Este proceso no es estacionario, no tiene una media constante y, al igual que para los modelos lineales no estacionarios en la parte regular, su función de autocorrelación no va a decrecer rápidamente hacia cero, sino lentamente.

Los modelos estacionales multiplicativos $ARIMA(p,d,q)*(P,D,Q)_s$ **.**

Estos modelos son flexibles en el sentido de que especifican estacionalidades estocásticas, tendencias estocásticas y además recogen la posible interacción entre ambos componentes.

Esta clase de modelos se basa en la hipótesis central de que la relación de dependencia estacional es la misma para todos los periodos. Este supuesto no se tiene porque cumplir siempre, aun así estos modelos son capaces de representar muchos fenómenos estacionales que encontramos en la práctica de una forma muy simple.

Para completar la descripción de las características de los procesos estacionales multiplicativos, se va a tratar la estructura de la función de autocorrelación parcial (FAP) de un modelo $ARMA(p,q)*(P,Q)_s$:

a) La FAP del proceso multiplicativo es una mezcla de las FAP correspondientes a su estructura regular y a su estructura estacional por separado, con un componente de interacción porque ambas no son independientes.

b) En los retardos más bajos, $k = 1, 2, \ldots,$ que recogen la estructura regular se reproduce la estructura marcada por los polinomios autoregresivos y de medias móviles de la parte regular. Si es un $MA(q)$ o un $ARMA(p,q)$ presentaría el decaimiento rápido y continuado característico de estos modelos, mientras que si fuera un $AR(p)$ se truncarla en el retardo p.

c) En los retardos estacionales, k = s, 2s, 3s, ..., se representa la estructura estacional generada por los polinomios autoregresivos y de medias móviles de la parte estacional. Si es un $MA(Q)_s$ o un $ARMA(P,Q)_s$ presentará el decaimiento rápido y continuado característico de estos modelos, mientras que si fuera un $AR(P)_s$ se truncarla en el retardo $k = Ps$.

d) Alrededor de los retardos estacionales se observa la interacción entre la parte regular y la estacional

   - A la derecha de cada coeficiente estacional, en los órdenes $k = js + 1, js + 2, \cdots$ se reproduce la FAP de la parte regular.
   - A la izquierda de los coeficientes estacionales, en los órdenes $k = js - 1, js - 2, \cdots$ se reproduce la FAS de la parte regular.

Luego de haber visto los posibles modelos, proseguimos a explicar la metodología Box-Jenkins.

Metodología Box - Jenkins.

En la aplicación de la metodología Box-Jenkins el punto de partida es el de conocer los valores de la serie temporal $Y_1, Y_2, \ldots, Y_T$ y se tratará de determinar la estructura $ARIMA(p,d,q)$ que la ha podido generar.

La construcción de los modelos ARIMA se lleva a cabo de forma iterativa mediante un proceso en el que se pueden distinguir cuatro etapas:

a) Identificación,
b) Estimación,
c) Validación y
d) Predicción.

A este proceso se le denomina metodología Box-Jenkins y continuación se describe con detalle cada una de las etapas.

a) Identificación

Esta etapa consiste en utilizar los datos y/o cualquier tipo de información disponible sobre cómo ha sido generada la serie, con ello se intentar sugerir una subclase de modelos $ARMA(p,d,q)$ que merezca la pena ser investigada.

El objetivo es determinar los órdenes p, d y q que parecen apropiados para reproducir las características de la serie bajo estudio y si se incluye o no la constante de nivel. En esta etapa es posible identificar más de un modelo candidato que haya podido generar la serie.

A. Análisis de estacionariedad, en el que se determinan las transformaciones que son necesarias aplicar para obtener una serie estacionaria.

Incluye, a su vez, dos apartados:

- Estacionariedad en varianza: transformaciones estabilizadoras de varianza.
- Estacionariedad en media: número de diferencias d que hay que tomar para lograr que la serie sea estacionaria en media.

B. Elección de los ordenes p y q. Una vez obtenida la serie estacionaria, el objetivo es determinar el proceso estacionario $ARMA(p,q)$ que la haya generado.

Los instrumentos que se van a utilizar en estas dos fases de la identificación del modelo son fundamentalmente los siguientes:

- Gráfica y correlogramas muestrales de la serie original.
- Gráfica y correlogramas muestrales de determinadas transformaciones de la serie: logaritmos, diferencias,...
- Contrastes de raíces unitarias.

Para identificar los órdenes p y q, se compararan las funciones de autocorrelación muestral con las FAS y FAP teóricas de los modelos ARMA cuyas características conocemos:

Modelo	FAS	FAP
AR(p)	Decrecimiento rápido y no se anula	Se anula para j > p
MA(q)	Se anula para j > q	Decrecimiento rápido y no se anula
ARMA(p, q)	• Decrecimiento rápido y no se anula • q valores significativamente diferentes de cero	• Decrecimiento rápido y no se anula • p valores significativamente diferentes de cero

Tabla 1. Comparación de las características de algunos modelos ARIMA.

Si el correlograma muestral de la serie presenta un corte a partir de un retardo finito j, la identificación del proceso adecuado para la misma es sencilla, ya que se correspondería con la FAS teórica de un $MA(j)$.

Pero si el correlograma muestral no presenta ningún corte, sino que parece decrecer rápidamente siguiendo una estructura exponencial o de onda seno-coseno, la identificación no es tan clara, ya que basándose únicamente en la FAS podría corresponder a un modelo teórico AR o $ARMA$ de cualquier orden.

Para ayudarnos en la identificación de los modelos $ARMA(p,q)$ nos valemos de los Gráficas de la serie y de las funciones de correlación y correlación parcial, así como de las pruebas estadísticas de raíces unitarias y de estacionariedad.

b) Estimación.

Usando de forma eficiente los datos, se realiza inferencia sobre los parámetros condicionada a que el modelo investigado sea apropiado.

Dado un determinado proceso propuesto, se trata de cuantificar las estimaciones de los parámetros del mismo.

c) Validación.

Se realizan contrastes de diagnostico para comprobar si el modelo se ajusta a los datos, si no es así, revelar las posibles discrepancias del modelo propuesto para poder mejorarlo.

Se toma en cuenta lo siguiente:

a) Si las estimaciones de los coeficientes del modelo son significativas y cumplen las condiciones de estacionariedad e invertibilidad que deben satisfacer los parámetros del modelo.

b) Si los residuos del modelo tienen un comportamiento similar a las perturbaciones, es decir, si son ruido blanco.

Análisis de coeficientes estimados

En primer lugar, se han de realizar los contrastes habituales de significación individual de los coeficientes AR y MA para comprobar si el modelo propuesto está sobre-identificado, es decir, si se ha incluido un parámetro en el modelo que no es relevante.

En el caso más general de un $ARMA(p,q)$ con constante se plantean los siguientes contrastes:

$$H_0 : \delta = 0 \quad frente\ a \quad H_1 : \delta \neq 0$$

$$H_0 : \phi_i = 0 \quad frente\ a \quad H_1 : \phi_i \neq 0 \quad , \ i = 1, 2, \dots , p$$

$$H_0 : \theta_j = 0 \quad frente\ a \quad H_1 : \theta_j \neq 0 \quad , \ j = 1, 2, \dots , q$$

Y el modelo que estamos asociando en la parte regular de la serie es el siguiente:

$$Y_t = \delta - [\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}] + [a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}]$$

Por otro lado, es importante comprobar si las condiciones de estacionariedad e invertibilidad se satisfacen para el modelo propuesto. Las condiciones de estacionariedad son obligatorias mientras que las de invertibilidad son optativas ya que se emplean únicamente si se desea convertir el modelo $ARMA(p,q)$ en un $AR(p)$ ó un $MA(q)$.

Prueba de correlación entre los residuos

Luego siguen las pruebas de significancia sobre un conjunto de coeficientes de autocorrelación, siendo la hipótesis nula,

$$H_0 : \rho_1(a) = \rho_2(a) = \cdots = \rho_M(a) = 0$$

Y la hipótesis alternativa que algún coeficiente ρ_k sea no nulo, para $k = 1, 2, \cdots, M$.

El contraste más utilizado para probar esta hipótesis es el propuesto por Ljung-Box(1978).

Si existe correlación entre los residuos del modelo, se concluye que el modelo no ha sido capaz de reproducir el patrón de comportamiento sistemático de la serie y habría que reformularlo.

Contraste de normalidad.

Luego le sigue el contraste de normalidad de los residuos, para ello podemos valernos del Gráfica de P-P y poder ver si los residuos tienen un cierto comportamiento normal o no.

Para verificar la hipótesis de normalidad nos valemos de las pruebas estadísticas, el contraste más utilizado es el de Jarque - Bera,

d) Predicción.

Esta etapa consiste en obtener pronósticos en términos probabilísticos de los valores futuros de la variable. Además se tratará de evaluar la capacidad predictiva del modelo.

Validación del modelo

En la etapa de validación, diagnosis o chequeo se procede a evaluar la adecuación de los modelos estimados a los datos.

En este punto, si se tiene un único modelo, no hay problema de elegir el mejor; pero en la realidad siempre habrá más de uno y a simple vista parecen todos adecuados, por lo que debemos encontrar una manera de elegir el mejor.

No debemos elegir un modelo sobre otro simplemente tomando como criterio la suma de los cuadrados de los errores, esto es debido a que toda variable que incluyamos en la regresión reducirá la suma de los cuadrados de los residuos ya sea que tenga mucha o poca relación con la variable respuesta.
Existen diversos criterios de elección y a continuación mostramos algunos de ellos.

- R cuadrado corregido
- Cp de Mallows
- Criterio de información de Akaike

Al final, elegimos el modelo que minimice el criterio o criterios elegidos

6.4 Redes neuronales artificiales

El cerebro humano es el sistema de cálculo más complejo que conoce el hombre. El ordenador y el hombre realizan bien diferentes clases de tareas; así la operación de reconocer el rostro de una persona resulta una tarea relativamente sencilla para el hombre y difícil para el ordenador, mientras que la contabilidad de una empresa es tarea costosa para un experto contable y una sencilla rutina para un ordenador básico.

La capacidad del cerebro humano de pensar, recordar y resolver problemas ha inspirado a muchos científicos a intentar o procurar modelar en el ordenador el funcionamiento del cerebro humano.

Los profesionales de diferentes campos como la ingeniería, filosofía, fisiología y psicología han unido sus esfuerzos debido al potencial que ofrece esta tecnología y están encontrando diferentes aplicaciones en sus respectivas profesiones.

Un grupo de investigadores ha perseguido la creación de un modelo en el ordenador que iguale o adopte las distintas funciones básicas del cerebro. El resultado ha sido una nueva tecnología llamada Computación Neuronal o también Redes Neuronales Artificiales.

El resurgimiento del interés en esta nueva forma de realizar los cálculos tras dos décadas de olvido se debe al extraordinario avance y éxito tanto en el aspecto teórico como de aplicación que se está obteniendo estos últimos años.

Características de las redes neuronales artificiales

Las Redes Neuronales Artificiales, las que de aquí en adelante las nombraremos por sus siglas en español RNA, están inspiradas en las redes neuronales biológicas del cerebro humano. Están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes. Estos elementos están organizados de una forma parecida a la que presenta el cerebro humano.

Las RNA al margen de "parecerse" al cerebro presentan una serie de características propias del cerebro. Por ejemplo las RNA aprenden de la experiencia, generalizan de ejemplos previos a ejemplos nuevos y abstraen las características principales de una serie de datos.

Aprender: adquirir el conocimiento de una cosa por medio del estudio, ejercicio o experiencia. Las RNA pueden cambiar su comportamiento en función del entorno. Se

les muestra un conjunto de entradas y ellas mismas se ajustan para producir unas salidas consistentes.

Generalizar: extender o ampliar una cosa. Las RNA generalizan automáticamente debido a su propia estructura y naturaleza. Estas redes pueden ofrecer, dentro de un margen, respuestas correctas a entradas que presentan pequeñas variaciones debido a los efectos de ruido o distorsión.

Abstraer: aislar mentalmente o considerar por separado las cualidades de un objeto. Algunas RNA son capaces de abstraer la esencia de un conjunto de entradas que aparentemente no presentan aspectos comunes o relativos.

Estructura básica de una red neuronal

Analogía con el cerebro.

La neurona es la unidad fundamental del sistema nervioso y en particular del cerebro. Cada neurona es una simple unidad procesadora que recibe y combina señales desde y hacia otras neuronas. Si la combinación de entradas es suficientemente fuerte la salida de la neurona se activa. La Figura (1.1) muestra las partes que constituyen una neurona.

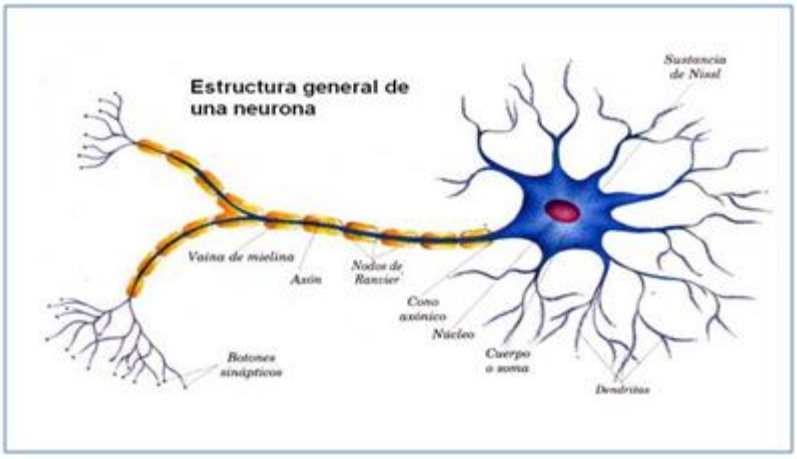


Figura 3. La neurona y sus partes.

El cerebro consiste en uno o varios billones de neuronas densamente interconectadas.

El axón (salida) de la neurona se ramifica y está conectada a las dendritas (entradas) de otras neuronas a través de uniones llamadas sinapsis. La eficacia de la sinapsis es modificable durante el proceso de aprendizaje de la red.

RNA.

En las RNA, la unidad análoga a la neurona biológica es el elemento procesador (EP). Un elemento procesador tiene varias entradas y las combina, normalmente con una suma básica. La suma de las entradas es modificada por una función de transferencia y el valor de la salida de esta función de transferencia se pasa directamente a la salida del elemento procesador.

La salida del EP se puede conectar a las entradas de otras neuronas artificiales mediante conexiones ponderadas correspondientes a la eficacia de la sinapsis de las conexiones neuronales.

La Figura 4 representa un elemento procesador de una red neuronal artificial implementada en un ordenador.

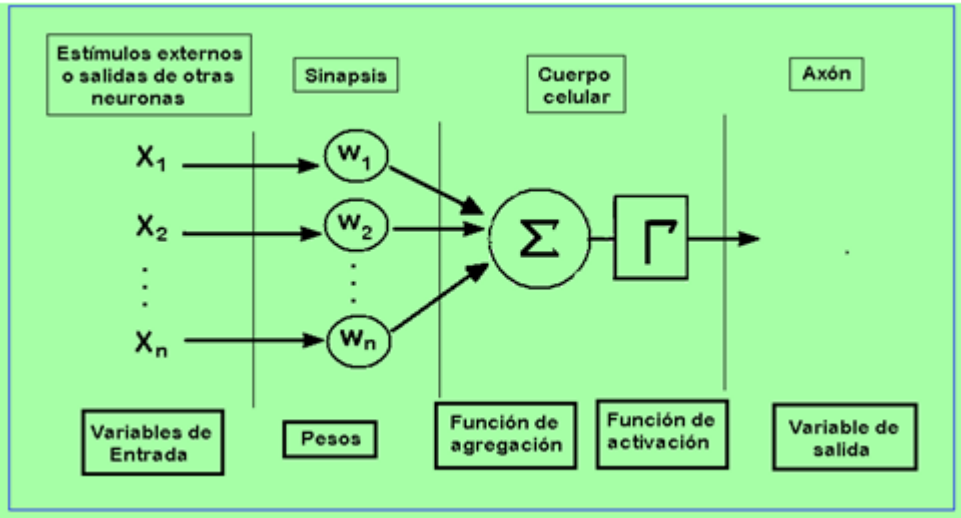


Figura 4. Modelo de una red neuronal artificial.

Una red neuronal consiste en un conjunto de unidades elementales EP conectadas de una forma concreta. El interés de las RNA no reside solamente en el modelo del elemento EP, sino en las formas en que se conectan estos elementos procesadores.

Generalmente los elementos EP están organizados en grupos llamados niveles o capas. Una red típica consiste en una secuencia de capas con conexiones entre capas adyacentes consecutivas.

Existen dos capas con conexiones con el mundo exterior. Una capa de entrada, donde se presentan los datos a la red, y una capa de salida que mantiene la respuesta de la red a una entrada. El resto de las capas reciben el nombre de capas ocultas.

La Figura 4 muestra el aspecto de una Red Neuronal Artificial y las partes que la componen.

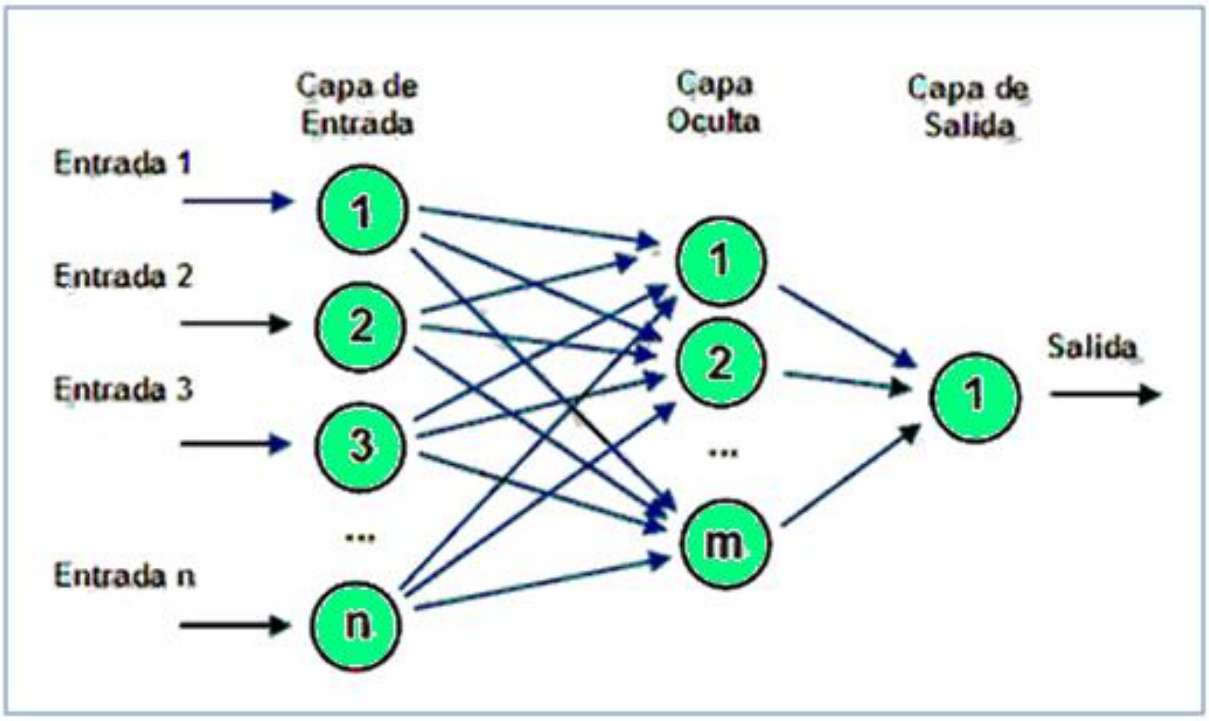


Figura 5. Red neuronal artificial y sus partes.

Programación / Entrenamiento.

Las técnicas tradicionales de programación utilizadas para la solución de un problema requieren la creación de un algoritmo.

Un algoritmo consiste en una secuencia de instrucciones que indica el modo en el que debe proceder el sistema basado en un ordenador para lograr el fin perseguido que es la resolución del problema.

El diseño de una secuencia de instrucciones para resolver por ejemplo, un problema de contabilidad es relativamente sencillo, mientras que existen muchos problemas del mundo real en los que resulta difícil realizar un algoritmo que resuelva dichos problemas.

Por ejemplo, imaginemos desarrollar un programa para cualquiera de los problemas de reconocimiento de imágenes como el rostro de una persona. Hay muchas variaciones de la imagen de una persona, como que presente un rostro serio o un rostro alegre, variaciones en general que deben tenerse en cuenta a la hora de diseñar el algoritmo.

Las RNA, a diferencia de los algoritmos que son instrucciones previamente programadas, deben ser previamente entrenadas. Esto significa que a la red se le muestra en su capa de entrada unos ejemplos y ella misma se ajusta en función de alguna regla de aprendizaje.

Clasificación de las redes neuronales artificiales de acuerdo a su arquitectura.

Las RNA presentan una arquitectura totalmente diferente de los ordenadores tradicionales de un único procesador. Las máquinas tradicionales basadas en el modelo de Von Neuman tienen un único elemento procesador, la CPU (del inglés

Control Process Unit) que realiza todos los cálculos ejecutando todas las instrucciones de la secuencia programada en el algoritmo.
Cualquier CPU realiza más de cien comandos básicos, incluyendo sumas, restas, y desplazamientos entre otros.

Los comandos o instrucciones se ejecutan secuencialmente y sincronizadas con el reloj del sistema. Sin embargo en los sistemas de computación neuronal cada EP sólo puede realizar uno, o como mucho, varios cálculos. La potencia del procesado de las RNA se mide principalmente por el número de interconexiones actualizadas por segundo durante el proceso de entrenamiento o aprendizaje.

La arquitectura de las RNA parte de la organización de los sistemas de procesado en paralelo, es decir, sistemas en los que distintos procesadores están interconectados. No obstante los procesadores son unidades procesadoras simples, diseñadas para la suma de muchas entradas y con un ajuste automático de las conexiones ponderadas.

Las redes neuronales se pueden clasificar conforme a su arquitectura dependiendo del número de capas que la componen o de la forma en que fluye la información entre las neuronas que conforman la red.

Dependiendo del número de capas se clasifican en:

- Redes monocapas
- Redes multicapas

Explicaremos a continuación en qué consiste cada una de ellas.

Redes monocapa

Las redes monocapa son redes con una sola capa. Para que las neuronas de este tipo de red puedan comunicarse entre ellas, crean conexiones laterales, es decir, que las neuronas crean conexiones con las neuronas de su misma capa.

En la figura 6 podemos apreciar una red monocapa en la que cada neurona se conecta con todas las demás.

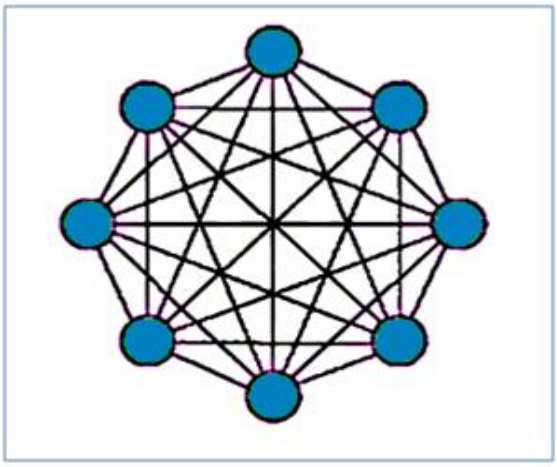


Figura 6. Ejemplo de una red neuronal monocapa.

Las redes más representativas de este tipo son la red de Hopfield, la red BRAIN-STATE-IN-A-BOX o memoria asociativa y las maquinas estocásticas de Botzmann y Cauchy.

Entre las redes neuronales monocapa, existen algunas que permiten que las neuronas tengan conexiones a sí mismas y se denominan autorecurrentes.

Las redes neuronales artificiales monocapa han sido ampliamente utilizada en la creación de circuitos eléctricos, esto se debe a su topología, son adecuadas para ser implementadas mediante hardware, usando matrices de diodos que representan las conexiones de las neuronas.

Redes multicapa

Las redes multicapa están formadas por varias capas de neuronas. Estas redes pueden a su vez clasificarse atendiendo a la manera en que se conectan sus capas.

Usualmente, las capas están ordenadas por el orden en que reciben la señal desde la entrada hasta la salida y están unidas en ese orden. Ese tipo de conexiones se denominan conexiones feed forward o hacia delante.

Un ejemplo Gráfica de este tipo de redes se muestra a continuación (figura 7).

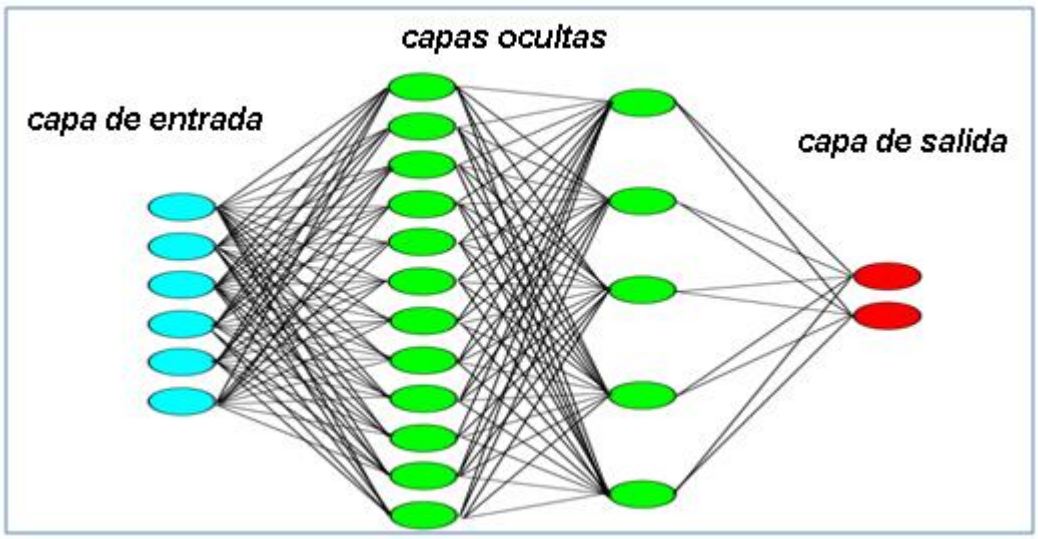


Figura 7. Ejemplo de una red neuronal artificial multicapa.

Por el contrario existen algunas redes en que las capas, aparte del orden normal también están unidas desde la salida hasta la entrada en el orden inverso en que viajan las señales de información. Las conexiones de este tipo se llaman conexiones hacia atrás, feed back o retroalimentación.

Redes con conexiones hacia adelante (Feed forward)

Este tipo de redes contienen solo conexiones entre capas hacia delante. Esto implica que una capa no puede tener conexiones a una que reciba la señal antes que ella en la dinámica de la computación.

Ejemplos de estas redes son Perceptron, Adaline, Madaline, Backpropagation y los modelos LQV y TMP de Kohonen.

Redes con conexiones hacia atrás (Feed back)

Este tipo de redes se diferencia en las anteriores en que si pueden existir conexiones de capas hacia atrás y por tanto la información puede regresar a capas anteriores en la dinámica de la red.

Este Tipo de redes suelen ser de dos capas. Ejemplos de estas redes son las redes ART, Bidirectional Associative Memory (BAM) y Cognitron.

Existen otros tipos de clasificación como por ejemplo respecto a la forma de aprendizaje, el cual puede ser supervisado o no supervisado o incluso semisupervisado.

La clasificación anterior solo la mencionamos, ya que sería muy extenso el profundizar más en esta temática y es por eso que nos limitaremos a mencionar a continuación los diferentes tipos de redes neuronales y luego describiremos un poco la red neuronal artificial que emplearemos.

Existe una serie de modelos que aparecen en la mayoría de estudios académicos y la bibliografía especializada. Entre ellos:

- Perceptron
- Adaline
- Perceptron multicapa
- Memorias asociativas
- Máquina de Boltzmann
- Máquina de Cauchy
- Propagación hacia atrás (backpropagation)
- Redes de Elman
- Redes de Hopfield
- Red de contrapropagación
- Redes de neuronas de base radial
- Redes de neuronas de aprendizaje competitivo
- Mapas Autoorganizados (RNA) (Redes de Kohonen)
- Crecimiento dinámico de células
- Gas Neuronal Creciente
- Redes ART (Adaptative Resonance Theory)

Cada uno de estos modelos de redes neuronales artificiales se emplean para diferentes propósitos, las neuronas y las capas que las conforman tiene una manera de interconectarse, así como un método de aprendizaje y funciones de activación, etc.

Pero el modelo de red neuronal artificial que nos interesará conocer es el perceptron multicapa, debido a que este tipo de red es adecuado para la identificación de patrones y ha sido implementado en código R para la estimación y predicción en series de tiempo empleando modelos autorregresivos.

Red Neuronal artificial Perceptron

El perceptron simple tiene una serie de limitaciones muy importantes. La más importante es su incapacidad para clasificar conjuntos que no son linealmente independientes.

El modelo perceptron multicapa es una ampliación del perceptron creado originalmente, el cual añade una serie de capas ocultas que, básicamente, hacen una transformación sobre las variables de entrada, esto le permite eludir el problema anterior.

Esto acaba con el problema del perceptron, convirtiendo las funciones linealmente no independientes en linealmente independientes gracias a la transformación de la capa oculta; además, el perceptron multicapa admite valores reales. Podemos decir que el perceptron multicapa es un modelador de funciones universal.

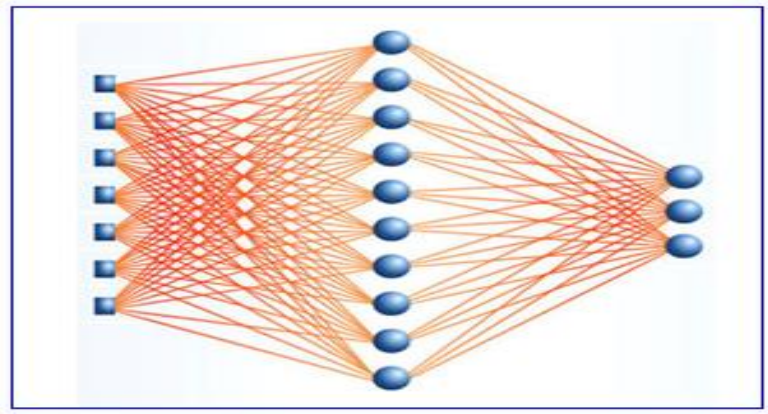


Figura 8. Ejemplo de una red Perceptron multicapa.

El perceptron multicapa consta de una capa de entrada, una capa de salida y una o más capas ocultas. Dichas capas se unen de forma total hacia delante, esto significa que cada neurona de entrada se une con la primera capa oculta y esta con la siguiente y finalmente cada neurona artificial de la última capa oculta se une con la capa de salida.

En nuestro caso, emplearemos el modelo perceptron multicapa aplicado a una serie de tiempo mediante la función ARNN implementada en el paquete estadístico R, el cual describimos a continuación:

Consideramos que los valores Y_t de la serie dependen de los p valores pasados mediante una función no lineal conforme a la siguiente ecuación:

$$Y_t^* = \eta + \sum_{n=1}^{p} \varphi_n Y_{t-n} + \sum_{h=1}^{H} \beta_h G\left(\omega_h + \sum_{n=1}^{p} \alpha_{n,h} Y_{t-n} \right)$$

Donde $\underline{G()}$ es una función adaptativa sinusoidal determinada por la siguiente ecuación:

$$G(u) = \left(\frac{1}{1 + \exp(-u)} \right)^M$$

Las constantes del modelo η , φ_n , β_h , ω_h , $\alpha_{n,h}$ y M para $h = 1, 2, \cdots, H$; $p = 1, 2, \cdots, P$ son estimados minimizando el error de regulación dado:

$$\lambda E_*$$

El valor de λ es definido por el usuario, mientras que E_* es una función definida únicamente por los parámetros del modelo, de la siguiente manera;

$$E_* = |\eta| + \sum_{h=1}^{H} (|\beta_h| + |\omega_h|) + \sum_{n=1}^{p} |\varphi_n| + \sum_{n=1}^{p} \sum_{h=1}^{H} |\alpha_{n,h}|$$

El modelo escrito anteriormente se reduce a una red perceptron multicapa imponiendo la restricción $\varphi_1 = \varphi_2 = \varphi_3 = \cdots = \varphi_p$ y al imponer la restricción de que $H = 0$ se logra obtener un modelo autorregresivo.

Es de mucha importancia hacer notar que el paquete ARNN es una herramienta de usuario final, por lo cual el usuario no debe preocuparse por la implementación interna de los elementos de la red neuronal.

El paquete ARNN consta de dos funciones principales que nos serán de utilidad, las cuales son:

- ✓ "arnn", función que nos permite la creación y estimación del modelo matemático
- ✓ "forecast", función que efectúa predicciones varios periodos hacia adelante empleando el modelo estimado con la función "arnn".

6.5 Análisis de Valores Extremos

No se puede concebir un entendimiento completo de las distribuciones asintóticas de extremos sin un conocimiento previo de los estadísticos de orden. Por ello, se comienza este apartado dando las funciones de densidad y distribución de un estadístico de orden aislado y los conjuntos de varios de ellos. Seguidamente se analiza el caso de muestras no aleatorias simples o dependientes.

Definición (estadístico de orden).
Sea $(X_1, X_2, \cdots, X_n)$ una muestra procedente de una población. Si los valores de la secuencia $X_1, X_2, \cdots, X_n$, se ordenan en orden creciente $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$, entonces el miembro r-ésimo de esta nueva secuencia se denomina estadístico de orden r de la muestra dada.

Entre los estadísticos de orden destacan el primero y el último, que son el mínimo, $X_{1n} = Min(X_1, X_2, \cdots, X_n)$ y el máximo, $X_{nn} = Max(X_1, X_2, \cdots, X_n)$ de la muestra, respectivamente, y que juegan un papel importante en las aplicaciones.

Estadísticos de orden procedentes de muestras aleatorias simples

Anteriormente habíamos supuesto que la muestra era no aleatoria o de valores dependientes, ahora suponemos que $X_1, X_2, \cdots, X_n$ son independientes e idénticamente distribuidos con función de distribución $F(x)$.

Entonces tenemos que la función de densidad para el máximo se deduce de la siguiente manera:

$$\begin{aligned} F_{X_n}(x) &= P[X_n \le x] = P[X_1 \le x, X_2 \le x, \ \ldots, X_n \le x] \\ &= P[X_1 \le x].P[X_2 \le x]. \ldots . \mathrm{P}[X_n \le x] && \text{esto es por el supuesto de independencia} \\ &= F_X(\mathrm{x}).F_X(\mathrm{x}). \ \ldots \ F_X(\mathrm{x}) && \text{esto es por estar identicamente distribuidos} \\ &= \left(F_X(\mathrm{x})\right)^n \end{aligned}$$

Cuando n tiende al infinito se tiene:

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \left(F_X(\mathrm{x})\right)^n = \begin{cases} 1 \ , \ \text{si } F_X(\mathrm{x}) = 1 \\ 0 \ , \ \text{si } F_X(\mathrm{x}) < 1 \end{cases}$$

Esto significa que la distribución del máximo es degenerada. Para saber que significa el término “degenerada” agregamos su definición a continuación:

Definición) de una variable o función degenerada.

Una variable aleatoria X es degenerada en un valor $a \in R$ si toma dicho valor con probabilidad 1, es decir, $P(x = a) = 1$; su media y varianza son entonces obvias a partir de este resultado:

$$E\,X = a; var\,X = 0$$

Para evitar que la función de distribución del máximo sea degenerada, lo que se hace es tipificarla, es decir, que debemos encontrar sucesiones de constantes $a_n > 0$ y $b_n, n = 1, 2, \ldots$ de forma que la expresión $\frac{M_n - b_n}{a_n}$ tenga una

distribución no degenerada, esto significa que podamos encontrar una función $G(z)$ tal que

$$\lim_{n\to\infty} F^{M}(a_n z + b_n) = G(z)$$

Así, se tiene el siguiente teorema:

Teorema de valores extremos (teorema de Fisher- Tippett - Gnedenko)

Si existen sucesiones de constantes $\{a_n > 0\}$ y $\{b_n\}$ $(n = 1, 2, \ldots)$ de forma que

$$P\left[\frac{M_n - b_n}{a_n} \le z\right] \to G(z) \qquad \text{cuando } n \to \infty$$

Donde G es una función de distribución no degenerada, entonces G pertenece a una de las siguientes familias de funciones:

$$Gumbell:\ G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \qquad -\infty < z < \infty;$$

$$\text{Frechét:}\quad G(z) = \begin{cases} 0 & , z \le b \\ \exp\left[-\left(\frac{z-b}{a}\right)^{-a}\right] & , z > b \end{cases}$$

$$\text{Weibull:}\quad G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right]\right\} & , z \le b; \\ 1 & , z > b. \end{cases}$$

Se omite la demostración del teorema por estar fuera del alcance de nuestro conocimiento actual. Adicionalmente se debe mencionar lo que es el dominio de atracción de una función mostrando el siguiente teorema:

Teorema - (Dominio de atracción para máximos de una distribución dada).

La distribución $G(z)$ pertenece al dominio de atracción para máximos de alguno de los casos siguientes:

i) $G(z)$ Si y solamente si $w(F) = \infty$ y $\lim_{t\to\infty} \frac{1-F(tx)}{1-F(x)} = x^{-\gamma}$; $\gamma > 0$

ii) $G(z)$ Si y solamente si $w(F) < \infty$ y la función $F^*(x)$ satisface el literal i), donde:

$$F^*(x) = F\left[w(F) - \frac{1}{x}\right] ;\ x > 0$$

iii) $G(z)$ Si y solamente si

$$\lim_{n\to\infty} n\left\{1 - F\left[X_{1-\frac{1}{n}} + x\left(X_{1-\frac{1}{ne}} - X_{1-\frac{1}{n}}\right)\right]\right\} = \exp(-x)$$

Donde: X_α es el percentil 100α de $F(x)$.

Además, las constantes de normalización $a_n\ y\ b_n$ pueden ser elegidas según el caso:

i) $a_n = 0\ ;\ b_n = \inf\left\{x: 1 - F(x) \le \frac{1}{n}\right\}$

ii) $a_n = w(F)\ ;\ b_n = w(F) - \inf\left\{x: 1 - F(x) \le \frac{1}{n}\right\}$

iii) $a_n = \inf\left\{x: 1 - F(x) \le \frac{1}{n}\right\}$; $\mathrm{b_n} = [1 - F(a_n)]^{-1} \int_{a_n}^{w(F)} [1 - F(y)]dy$

ó $b_n = \inf\left\{x: 1 - F(x) \le \frac{1}{ne}\right\} - a_n$

Algunas indicaciones prácticas del teorema anterior son:

- Solo tres distribuciones pueden ocurrir como distribuciones limites de máximos, estas son: Fréchet, Weibull y Gumbel.
- Se dan reglas para determinar si una distribución dada F(x) pertenece al dominio de atracción de esas tres distribuciones.
- Se dan reglas para determinar sucesiones $\{a_n\}\ y\ \{\mathrm{b_n}\}$ que verifican esas condiciones.
- Una distribución con límite no finito en la cola de interés no puede pertenecer a una distribución de Weibull.
- Una distribución con límite finito en la cola de interés no puede pertenecer a una distribución de Fréchet.

Tomando en consideración los resultados del teorema de dominios de atracción se genera la siguiente tabla de resumen únicamente para el máximo ya que es nuestro tema de interés en este momento:

Distribución	Dominio de atracción
Normal	Gumbel
Exponencial	Gumbel
Log-normal	Gumbel
Gamma	Gumbel
Gumbel	Gumbel
Rayleigh	Gumbel
Uniforme	Weibull
Weibull	Weibull
Cauchy	Fréchet
Pareto	Fréchet
Fréchet	Fréchet

Tabla 2. Dominio de atracción del máximo para algunas distribuciones conocidas

Funciones de distribución asociadas a los valores extremos

Anteriormente se ha hecho mención de la distribuciones de Gumbel, Frechet y Weibull, pero es importante saber que no son las únicas asociadas a los valores extremos aunque si las más importantes. Debemos mencionar también que existe una distribución que está relacionada con las tres mencionadas con anterioridad y es la distribución de valores extremos generalizada.

La distribución de valores extremos generalizada, (en inglés, Generalized Extreme Value distribution, cuyas siglas son GEV), que también es conocida como la

distribución de Fisher-Tippett, la distribución tipo von Mises-Jenkinson o la distribución de valores extremos tipo von Mises y que tiene por ecuación la siguiente:

$$H\left(x,\alpha,\beta,\mathrm{c}\right)=\exp\left[-\left(1+c\,\frac{x-\beta}{\alpha}\right)^{-1/c}\right] \qquad \text{Definida en } \left\{x:1+c\,\frac{x-\beta}{\alpha}>0\right\}$$

Según Kotz y Nadarajah, dicha distribución fue inicialmente introducida por Jenkinson en 1955.

Esta función de distribución es la que se emplea a la hora de la estimación de un modelo matemático que se adecue a los datos observados y explicaremos el porqué.

El señor Von-Mises encontró que las funciones de Gumbel, Weibull y Frechet están encapsuladas en la distribución de valores extremos generalizada al observar el comportamiento del parámetro c. el notó que para $c>0, c<0\ y\ c=0$ se obtienen las familias de distribuciones de distribuciones de Fréchet, Weibull y Gumbell respectivamente, con la observación de que en el caso de $c=0$ se debe entender en el sentido límite de la expresión y no de manera literal.

Es por esta razón que al emplear la distribución de valores extremos generalizada se evita el estar probando cual de las tres distribuciones se asocia mejor a los datos, solo realizamos la estimación de los parámetros de la distribución de la GEV y analizamos el valor del parámetro c para asi decir a que tipo de familia de distribuciones encaja mejor.

A continuación mostraremos algunas de las características de las distribuciones antes mencionadas.

- **Distribución de valores extremos generalizada**

  Su funcion de densidad de probabilidad es:

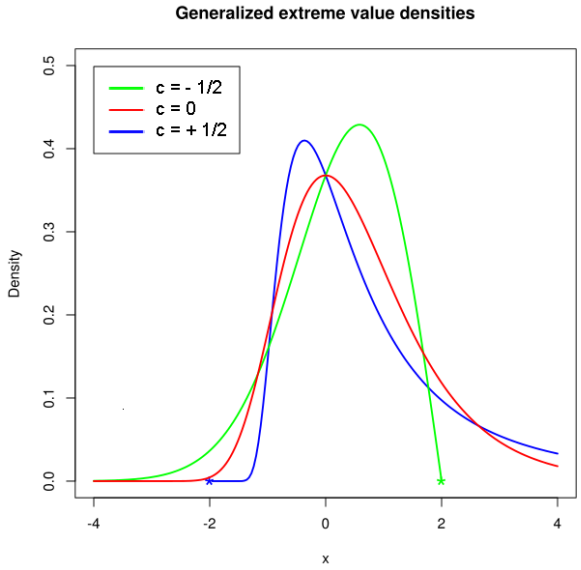
$$g(x)=\frac{1}{\sigma}\left[1+c\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{c}-1}\exp\left\{-\left[1+c\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{c}}\right\} \text{ Definida en } \left\{x:\left[1+c\left(\frac{x-\mu}{\sigma}\right)\right]>0\right\}$$

Su esperanza matemática es: $\begin{cases} \mu + \sigma\frac{\Gamma(1-c)-1}{c} & si\ \ c < 1\,, c \neq 0 \\ \mu + \sigma\gamma & si\ \ c = 0 \\ no\ existe & si\ \ c \geq 1 \end{cases}$

Donde γ es la constante de Euler y $\Gamma(*)$ es la función gamma.

Su varianza es: : $\begin{cases} \sigma^2\frac{g_2-g_1^2}{c^2} & si\ \ c < \frac{1}{2}\,, c \neq 0 \\ \sigma^2\frac{\pi^2}{6} & si\ \ c = 0 \\ no\ existe & si\ \ c \geq \frac{1}{2} \end{cases}$ donde $g_k = \Gamma(1-kc)$

Un ejemplo de su comportamiento Gráfica es el siguiente:



Gráfica 8. Ejemplo de la distribución de valores extremos generalizada

- **Distribución de Gumbel**

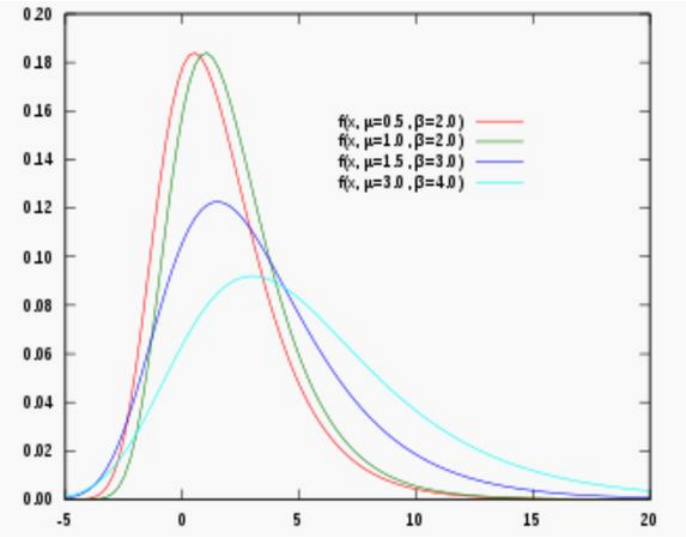
Su funcion de densidad de probabilidad es:

$$g(x) = \frac{1}{\sigma}\exp\left[\frac{-x-\beta}{\alpha} - \exp\left(\frac{-x-\beta}{\alpha}\right)\right] \quad \text{Definida en } x\epsilon(-\infty, +\infty)$$

Su esperanza matemática es: $\mu + \gamma\beta$, donde γ es la constante de Euler.

Su varianza es: $\beta^2\frac{\pi^2}{6}$ donde $g_k = \Gamma(1-kc)$

Un ejemplo de su comportamiento Gráfica es el siguiente:



Gráfica 9. Ejemplo de la distribución de Gumbel

- **Distribución de Weibull**

Su funcion de densidad de probabilidad es:

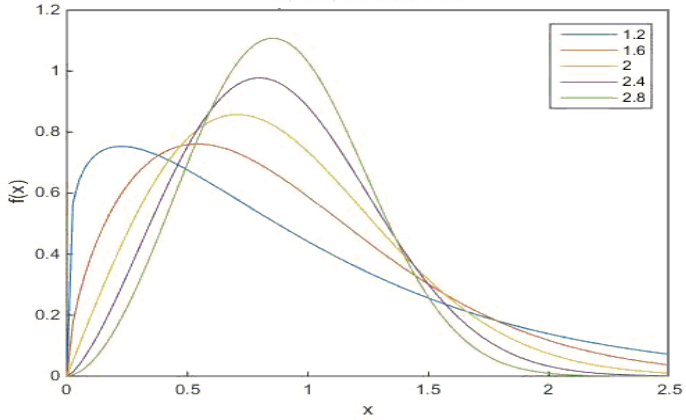
$$g(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp -\left(\frac{-x}{\lambda}\right)^{k} \qquad \text{Definida para } \; x \geq 0$$

Donde k es el parámetro de forma y $\lambda > 0$ es el parámetro de escala de la distribución.

Su esperanza matemática es: $\lambda\Gamma\left(1+\frac{1}{\mathrm{k}}\right)$.

Su varianza es: $\lambda^2\left[\Gamma\left(1+\frac{2}{\mathrm{k}}\right)-\Gamma^2\left(1+\frac{1}{\mathrm{k}}\right)\right]$ donde $\Gamma(*)$ es la función gamma

Un ejemplo de su comportamiento Gráfica es el siguiente:



Gráfica 10. Ejemplo de la distribución de Weibull

- **Distribución de Fréchet**

Su funcion de densidad de probabilidad es:

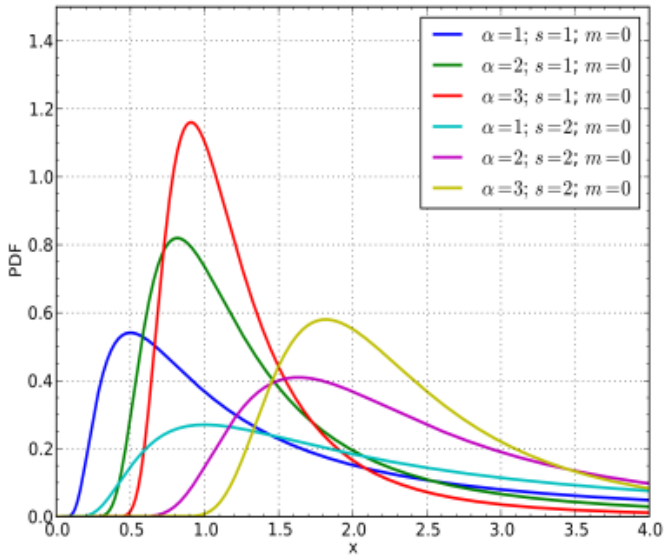
$$g(x) = \frac{\alpha}{\delta}\left(\frac{x-\lambda}{\delta}\right)^{-1-\alpha} \exp -\left(\frac{x-\lambda}{\delta}\right)^{-\alpha} \qquad \text{Definida para } x > \lambda$$

Donde $\alpha > 0, \lambda > 0, \delta > 0$ son el parámetro de forma, el parámetro de localización y el parámetro de escala de la distribución respectivamente.

Su esperanza matemática es: $\lambda + \delta\Gamma\left(1 - \frac{1}{\alpha}\right)$. siempre que $\alpha > 1$.

Su varianza es: $\delta^2\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \left(\Gamma\left(1 - \frac{1}{\alpha}\right)\right)^2\right]$ donde $\Gamma(*)$ es la función gamma y siempre que $\alpha > 2$.

Un ejemplo de su comportamiento Gráfica es el siguiente:



Gráfica 11. Ejemplo de la distribución de Fréchet

7. Metodología

7.1 Preparación de la base de datos

Iniciaremos el análisis preparando la base de datos, esto debido a que los datos se nos entregaron en una hoja de cálculo de Excel en la cual se muestran los datos ordenados por fechas diarias, y separadas por décadas mediante hojas en el mismo libro de trabajo; sin embargo, si un dato no se pudo recolectar simplemente no aparece en la base de datos.

Además, no es posible realizar un análisis tal como está la base de datos en algún paquete estadístico de los que pretendemos emplear como SPSS o R, esto debido a que hay que ordenarlo como una sola columna o variable para cada estación de monitoreo.

Es de hacer notar que existe la posibilidad de que los cálculos del valor acumulado mensual y otros valores simplemente se digitaron y no se calcularon por medio de fórmula, lo cual llevaría a no tener un dato real, por lo cual es necesario revisar las fórmulas para obtener estos valores antes de realizar cualquier procedimiento.

Adicionalmente deberemos preparar la base para poderla emplear en el análisis descriptivo por medio del paquete CHAC (Cálculo Hidrometeorológico de Aportaciones y Crecidas).

Para los análisis posteriores se emplearán los datos de la precipitación acumulada mensuales desde 1971 al 2012, excepto para el análisis de valores extremos, en ese tipo de análisis se emplearan los valores máximos mensuales de algunas series, por lo cual es necesario la revisión de las formulas de la hoja de cálculo para obtener el verdadero valor máximo ocurrido en cada mes.

7.2 Análisis descriptivo

En el análisis descriptivo pretendemos mostrar la tendencia de los datos así como la posible relación existente entre las estaciones de monitoreo o variables, esto mediante el análisis factorial múltiple, el cual nos permitirá obtener Gráficas con los cuales se evidenciara una realidad empírica y es que se menciona que la cantidad de lluvia debida a precipitaciones pluviales están relacionadas con la altura y la ubicación geográfica.

Es en este momento que se empleará CHAC para la elaboración de cronogramas de las series de tiempo así como un mapa de ubicación de las estaciones de manera geo-referenciada, esto no solo nos permitirá descartar algunas variables innecesarias sino ubicar las estación y verificar si son actas para emplearse en el rellenado de datos faltantes si es que fuera necesario.

Es de notar que este último software será útil únicamente pare el análisis exploratorio y no se empleará en los demás tipos de análisis del proyecto.

7.3 Análisis descriptivo por medio del análisis factorial múltiple

Mediante el análisis factorial múltiple se obtendrán factores que permitirían la comparación entre los puntos de monitoreo y así reducir el número de estaciones a analizar en las etapas posteriores del estudio.

Este tipo de análisis se realizará mediante el paquete estadístico R que es un software libre de mucha utilidad, con el cual se podrá obtener Gráficas adecuados para mostrar la estructura existente en la base de datos, esto se hará por medio de la librería FactomineR y de otras librerías existentes del entorno del mismo R,.

7.4 Tratamiento de datos ausentes

Como en la mayoría de estudios, casi siempre es necesario disponer de algún procedimiento para solventar la ausencia de datos, por lo cual se pretenderá emplear técnicas de rellenado de datos en caso de necesitarlo y para ello será de mucha utilidad el paquete CHAC, el cual dispone de métodos implementados para rellenar los datos de varias estaciones de monitoreo a la vez, debemos mencionar que el paquete que es de uso libre posee limitantes con respecto a la cantidad de variables a manejar a la vez, por lo cual no será posible ingresar muchas variables al mismo tiempo.

7.5 Aplicación de metodología Box-Jenkins y los modelos ARIMA para series temporales

Esta es una de las etapas principales del proyecto, ya que es en esta etapa que se empleará el software SPSS para la aplicación de la metodología Box-Jenkins y con ello determinar el comportamiento de las series de tiempo seleccionadas en la etapa descriptiva anterior, de manera paralela se usará el paquete estadístico R para realizar las pruebas de contraste de estacionariedad, tendencia, raíces unitarias y demás pruebas necesarias en el análisis de series de tiempo.

Debemos decir que el uso del SPSS no es exclusivo, podríamos emplear el paquete R u otro software disponible para dicha tarea, pero el SPSS es más fácil de utilizar en el análisis de series de tiempo.

Debemos detallar que se escogerá una de las variables con la mayor cantidad de datos, ya que con ella se identificará el patrón de comportamiento y definirá si ha habido algún cambio en el tiempo, luego se elegirán dos grupos tomando el mismo criterio para extraer el modelo ARIMA que mejor se les ajuste a todas y realizar el contraste de comparación de medias.

7.6 Análisis de los datos por medio de Redes Neuronales Artificiales

Para el análisis mediante redes neuronales artificiales se empleara el paquete estadístico R, el cual permite implementar dicha técnica aplicada a una serie de tiempo mediante las librerías del paquete ARNN.

Dicho paquete no fue fácil de encontrar debido a que no se encuentra entre las librerías del CRAN disponibles en la página de R-Project, sino que fue necesario conseguirla por medio de correo electrónico.

En este paquete se encuentran librerías para la estimación del modelo autorregresivo, así como para poder obtener predicciones varios pasos hacia adelante.

7.7 Análisis de valores extremos

Como ya se mencionó anteriormente, en esta técnica se emplearan los datos máximos mensuales de algunas de las series y empleando la librería extreme del paquete R para el análisis de series climatológicas, el cual nos proveerá de Gráficas y resultados a partir de los datos de la muestra máxima mensual.

8. Aplicación práctica

Iniciamos este apartado comentando que disponemos de la información siguiente:

- Lluvia diaria captada en diferentes puntos de nuestro país El Salvador, 292 estaciones de monitoreo en total, pero actualmente no todas se encuentran en funcionamiento y esto se debe a diversas causas, las cuales no son de interés para el presente trabajo. Podría especificar cuantas estaciones de monitoreo se encuentran en funciones pero esa información no fue proporcionada por el ministerio del medio ambiente y recursos naturales debido a que no forma parte de los propósitos de este proyecto.

- Además se dispone de la ubicación geográfica de todas y cada una de las estaciones de monitoreo presentes en la base de datos (longitud y latitud), dadas en grados y minutos, así como la altura del terreno sobre el cual están posicionadas.

En un primer momento procedemos a preparar la base de datos para el análisis descriptivo y demás análisis que pretendemos realizar en el presente trabajo, pero nos damos cuenta que la base está incompleta, esto se debe a que las personas encargadas de recolectar la información solo ingresan la información si disponen de ella, sino es así, entonces simplemente el dato del día no es registrado como ausente, lo mismo sucede si una estación deja de estar en servicio.

Los datos disponibles se encuentran en cinco archivos de Excel, los cuales detallamos a continuación.

- Formato Ayala lluvia 1971
- Lluvia General 2010
- Lluvia General 2011
- Lluvia General 2012
- Red de Estaciones actualizada

El primer archivo que es “Formato Ayala lluvia 1971” nos muestra los datos de año, mes, departamento, código y nombre del lugar donde se ubica la estación, la lluvia captada diariamente medida en centímetros cúbicos, así como fórmulas para obtener el acumulado y el máximo de cada mes y separadas por décadas.

Los archivos “Lluvia General 2010”, “Lluvia General 2011” y “Lluvia General 2012” nos muestran la misma información que el primer archivo excepto que solo para los años 2010, 2011 y 2012 respectivamente.

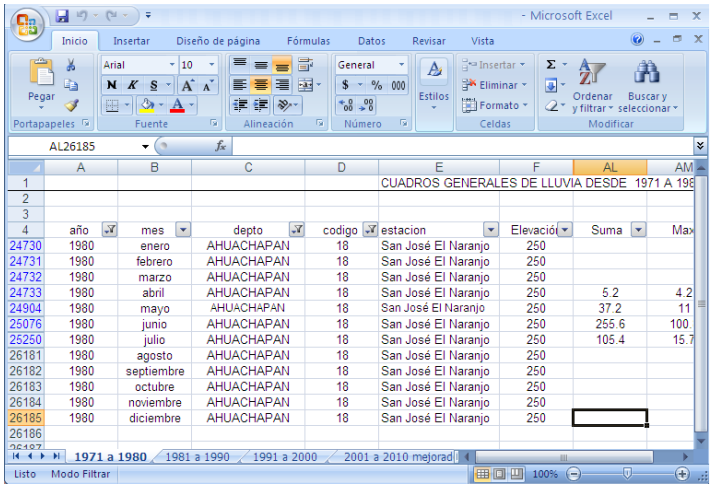


Figura 9. Base de datos lluvia general 1971-2010.

El quinto archivo nos muestra las estaciones separadas por departamento, su código asignado, el nombre completo del lugar de ubicación así como sus coordenadas geográficas y adicionalmente nos muestra las estaciones activas en el año 2012.

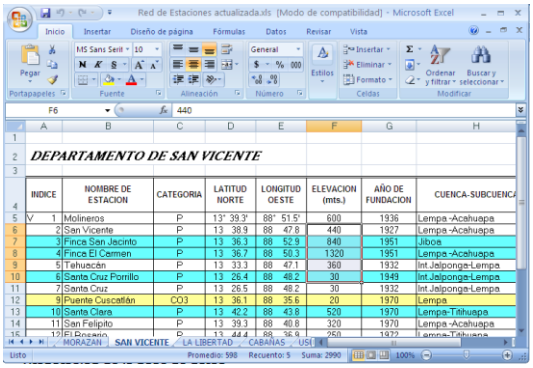


Figura 10. Estaciones de monitoreo actualizadas.

Se podía solicitar al Ministerio del Medio Ambiente y Recursos Naturales los datos de los últimos años, pero no era necesario ya que la información proporcionada anteriormente era suficiente para cumplir con los objetivos propuestos.

8.1. Preparación de la base de datos

Lo primero que debemos hacer antes que cualquier cosa es preparar la base de datos, convertirla en el formato que necesitamos para los posteriores análisis a realizar.

Por medio de un algoritmo aplicado a las estaciones de cada departamento, que se implementa en R podemos conseguir que cada estación de monitoreo muestre los datos disponibles así como los datos ausentes, obteniéndose de esta manera series de tiempo que se pueden analizar en SPSS o en R.

Ya habiendo preparado la base de datos, nuestro deseo es proseguir con la fase del análisis descriptivo aplicando para ello la técnica del análisis factorial múltiple, pero tenemos una dificultad y es la de los valores faltantes, para saber donde se encuentran estos huecos hacemos uso del paquete estadístico CHAC.

Se tomo la decisión del uso del paquete CHAC ya que no solo nos permite ver el cronograma de las series temporales hidrológicas, sino que tiene la opción de completar datos faltantes, obtención de estadísticas y otras opciones que no mencionaremos debido a que solo lo emplearemos para obtener la estadística de cada variable y si fuera necesario para el completado de datos faltantes.

Con este software no solo podemos obtener los cronogramas de cada variable temporal sino también nos permitirá entre otras cosa el rellenado de datos ausentes y descartar algunas variables que no posean suficiente información para el análisis posterior.

Una vista previa de la ubicación de las estaciones nos muestra lo siguiente.

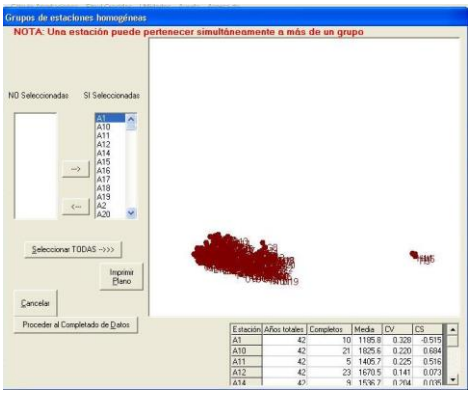


Figura 11. Datos meteorológicos ingresados al CHAC junto con sus coordenadas UTM.

En esta imágenes podemos tener un primer acercamiento con respecto a la ubicación de cada estación de monitoreo y podemos notar que no toda la información con respecto a la ubicación geográfica es correcta ya que algunas de las estaciones se ubican fuera de nuestro territorio nacional, lo cual significa nada mas que no es posible utilizarlas para el rellenado de datos faltantes debido a esta discrepancia.

Luego de haber eliminado las estaciones con anomalías en su ubicación geográfica tenemos la siguiente imagen:

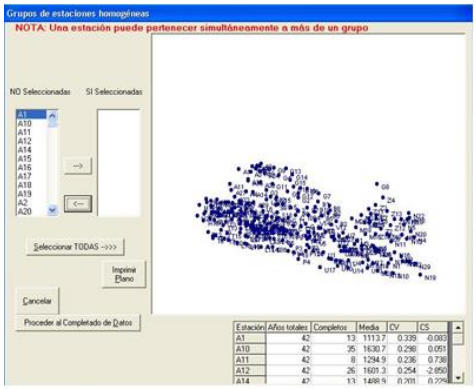


Figura 12. Datos meteorológicos ingresados al CHAC excluyendo variables con error en las coordenadas UTM.

Como se puede observar, ya se muestra una silueta más parecida al mapa de nuestro territorio, lo que significa que ya podemos aplicar el método de rellenado de datos ausentes empleando todas las variables existentes en esta base de datos restringida.

Debemos aclarar que las estaciones descartadas anteriormente puedan ser utilizadas en análisis posteriores, no son útiles para complementar datos faltantes, pero para los demás análisis posiblemente si puedan ser utilizadas.

Procedemos a activar la función de rellenado de datos, no solo para completar los huecos sino también para descartar algunas estaciones que carezcan de los datos suficientes para aplicar un modelado de series de tiempo, es decir, tener estaciones con al menos 5 años de valores continuos sin datos faltantes o que puedan ser completados por medio de las técnicas de rellenado implementadas en el software.

Como uno de nuestros objetivos es determinar el patrón de comportamiento en diferentes décadas para determinar si ha habido cambio en su comportamiento, es por eso que al tener series de tiempo con pocos datos, nos impediría completar este objetivo.

Antes de proceder al completado de datos ausentes, el software CHAC descarta las variables que no cumplen con las condiciones necesarias para este fin, es decir, aquellas variables que posean menos de 60 datos de manera continua, esto nos servirá como un criterio de exclusión.

8.2. Análisis descriptivo

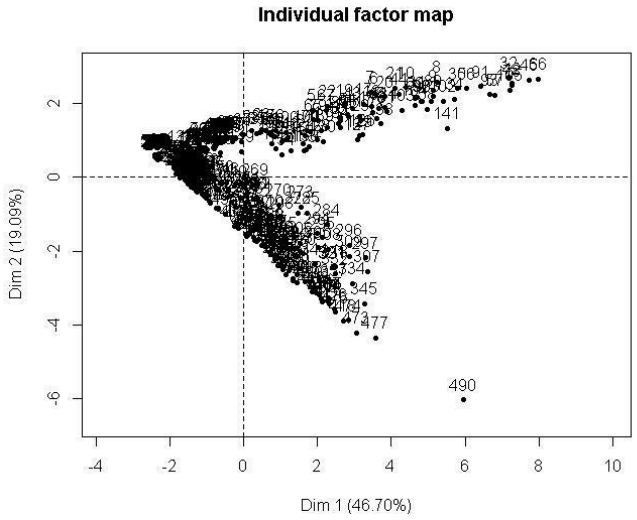
8.2.1. Análisis descriptivo por medio del análisis factorial múltiple

El objetivo de este método es la reducción del número de variables a emplear en procedimientos futuros.

En un primer momento, como no tenemos ni idea de la estructura de la base de datos, ni de la cantidad de grupos posibles, vamos a utilizar el menor número de grupos admisible para el AFM que es dos.

Debido a que las variables poseen datos faltantes, se decidió rellenar los huecos con una constante, la misma que se emplea en las base de datos utilizadas en el paquete CHAC que es -100 y luego aplicar el método de análisis factorial múltiple en R.

El resultado obtenido se puede ver en los siguientes Gráficas:



Gráfica 12. Mapa de factores individuales.

El Gráfica 12 nos muestra dos o tres grupos posibles que se están formando y una variable que se aleja de todas las demás, la apariencia del Gráfica se

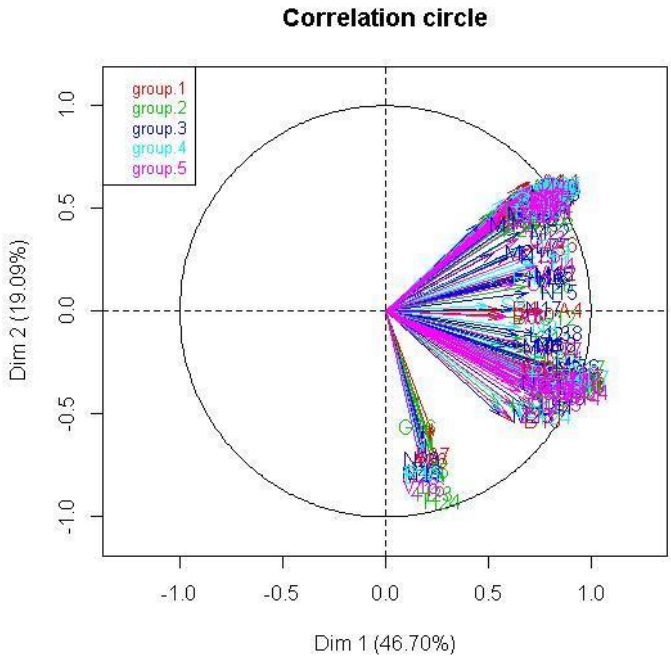
asemeja mucho a la de una parábola o una función de valor absoluto orientada horizontalmente.

Lo que se debe observar al realizar el análisis de dicho Gráfica es la distancia entre las variables y también la distancia entre los grupos de variables que se formen.

En este caso podemos notar que la distancia es bastante reducida entre las variables, lo cual nos hace pensar que su correlación es lineal, la cual puede ser directa o indirectamente proporcional, esto significa que se pueden tomar dos de variables cualesquiera y crear un modelo de regresión lineal.

En los diferentes grupos que son observables, la distancia entre esos grupos es corta y esto nos indica que los grupos están estrechamente relacionados entre sí.

A continuación el Gráfica de correlación circular del análisis anterior:

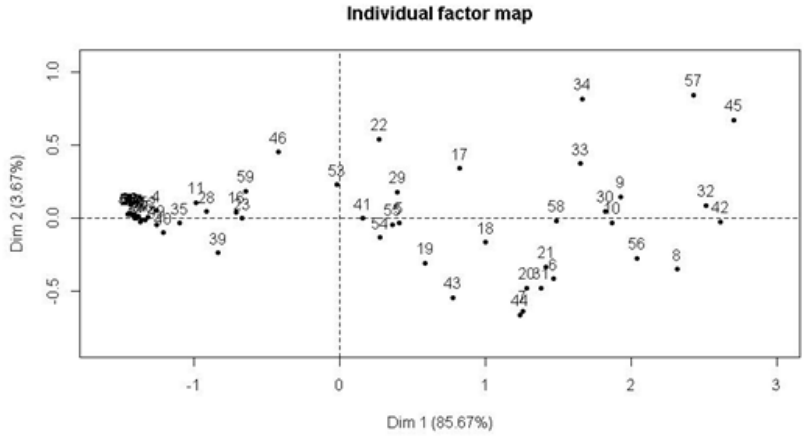


Gráfica 13. Correlaciones circular de todas las variables.

En el Gráfica 13 se observa que la correlación de los individuos es bastante alta, mayor que 0.5, esto se deduce conforme a la longitud de los vectores que parten del origen; esto significa que las variables tienen una correlación directamente proporcional o inversamente proporcional, dependiendo de la ubicación geográfica o la distancia que exista entre ellas. Claro que esto último no es una conclusión ya que no es posible por el momento demostrarlo de manera directa, solo es una especulación que se obtiene observando el Gráfica.

El haber realizado el AFM con esta base nos hace pensar que quizá el resultado obtenido no es por completo aceptable, por lo cual se hará el AFM restringiendo la base a un periodo de tiempo de cinco años y tomaremos las variables que tengan información en esos años, procurando que sean muchas.

Luego de aplicar el AFM bajo las condiciones descritas anteriormente, se obtuvieron los siguientes resultados:

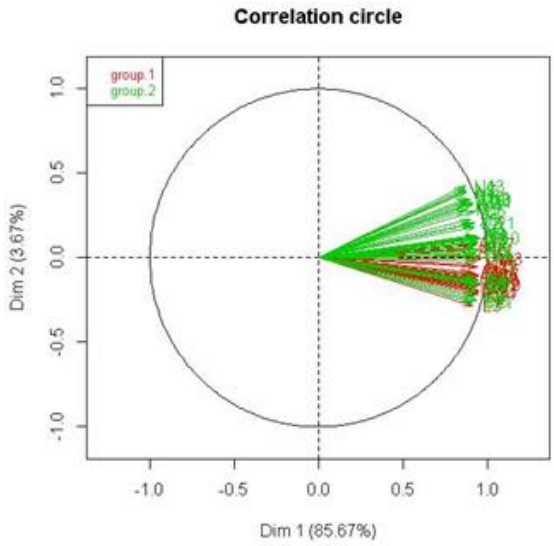


Gráfica 14. Mapa de factores individuales de grupo seleccionado

El Gráfica 14 es una porción del Gráfica que obtuvimos con la base completa y hasta nos veríamos tentados a aceptarlo como el definitivo, pero luego de meditarlo llegamos a la conclusión de que nos quedamos con el primero ya que es

el más completo, nos muestra que los datos faltantes también son una característica inherente de la base de datos y que tienen su razón de ser, ya que un dato faltante en una de las variables indica que la estación de monitoreo en cuestión dejo de funcionar y no se pudo registrar el dato de la cantidad de lluvia caída en un día especifico.

Veamos ahora el Gráfica circular de correlaciones:



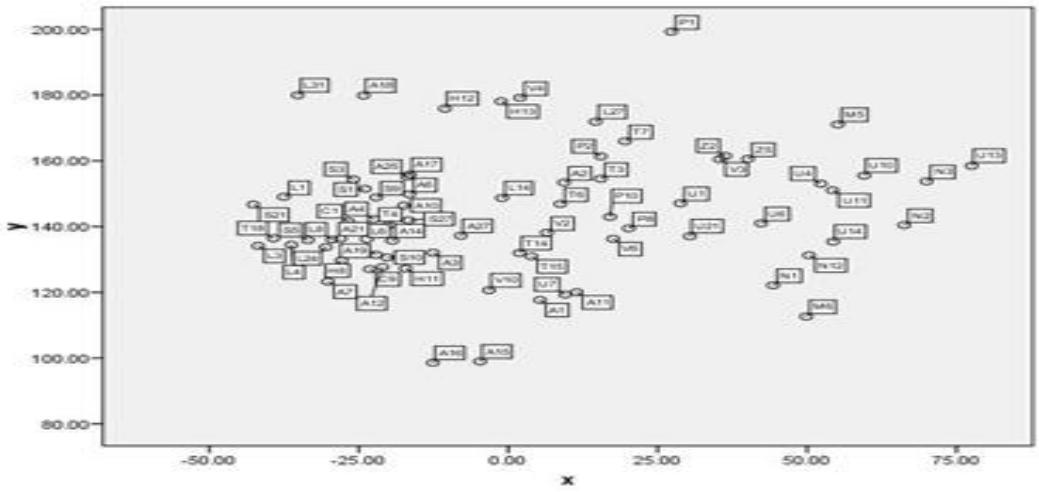
Gráfica 15. Correlaciones circular del grupo seleccionado de variables.

Como podemos apreciar en el Gráfica, todos los vectores se acercan bastante a la frontera del circulo unitario, lo que nos indica que la lluvia captada en un punto puede estar inversa o directamente correlacionada con la de otro punto y a causa de esta deducción podríamos suponer que se puede crear un modelo lineal entre dos variables cualesquiera.

Esto es similar al resultado que obtuvimos con la base de datos completa, por lo cual nos confirma que no es necesario tomar tantas variables para analizar el comportamiento de la lluvia en el tiempo.

A partir del análisis con la base de datos reducida, tomaremos las coordenadas obtenidas por medio del análisis factorial múltiple y le agregaremos las variables de altitud y código asignado a cada estación, con ello generaremos el mismo Gráfica obtenido pero nos permitirá además identificar las estaciones más relevantes en cada grupo así como la altitud a la que se encuentran instaladas las estaciones de monitoreo respectivo.

Esto no lo haremos en un solo Gráfica sino en dos, ya que de otra manera nos confundiría tanta información presente en el Gráfica. Primero la altitud a la que se encuentran las estaciones de monitoreo se muestra a continuación:



Gráfica 16. Mapa de factores individuales de grupo seleccionado agregando el dato de la altura

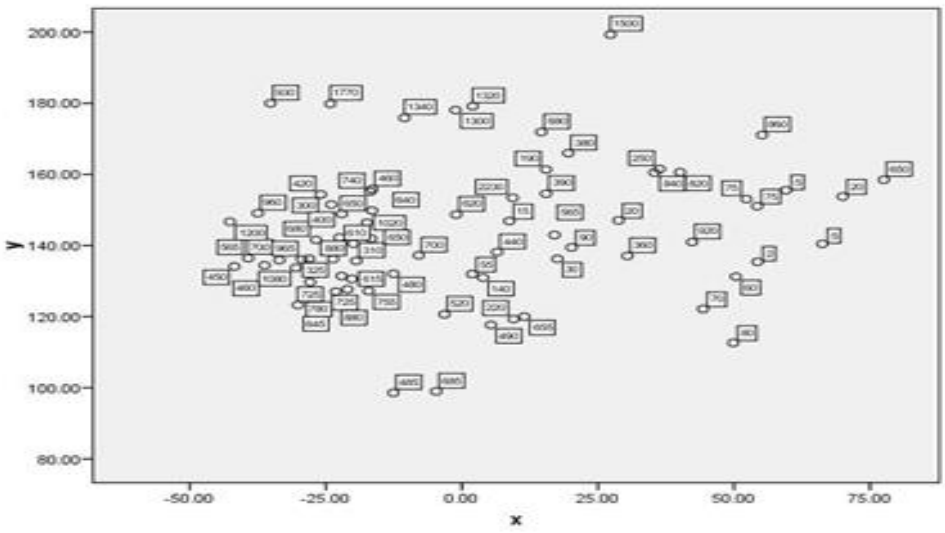
Podemos apreciar que el grupo de la izquierda se conforma de estaciones con diversa altitud, van desde los 300 metros hasta los 1200 metros sobre el nivel del mar.

En el grupo de la derecha se observa un aglutinamiento de estaciones que se encuentran arriba de los 900 metros

Tambien se puede apreciar un grupo un poco reducido de estaciones que oscilan entre los 2 y 90metros de altitud asi como tambien se encuentran estaciones con una elevacion de 650 y 850 metros, pero no todas entran en el patrón porque tambien se muestarn altitudes variadas de entre 220 metros y 550 metros.

Esto tiende a confundir un poco ya que se esperaria que siguieran un patrón de agrupamiento mas notorio.

Veamos ahora el mismo Gráfica pero en lugar de la altitud se muestran los codigos asignados, esto es para ver de que departamento se encuentran en cada grupo.



Gráfica 17. Mapa de factores individuales de grupo seleccionado agregando el dato código asignado.

En este mapa podemos observar que las estaciones están mezcladas con respecto a su ubicación, ya que aparecen estaciones del mismo departamento pero también de otros departamentos, lo que hace pensar que la lluvia captada en un punto específico esta correlacionada con las de su entorno, pero también con las de otros lugares por encontrarse en condiciones de altitud y ubicación geográfica similares.

La idea que dan estos resultados la muestro en la siguiente imagen:

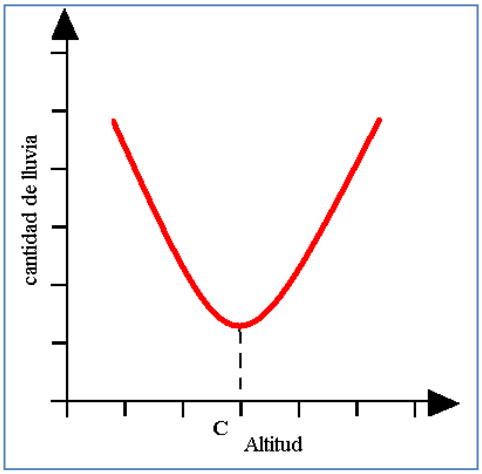


Figura 13. Idea del comportamiento de las precipitaciones pluviales

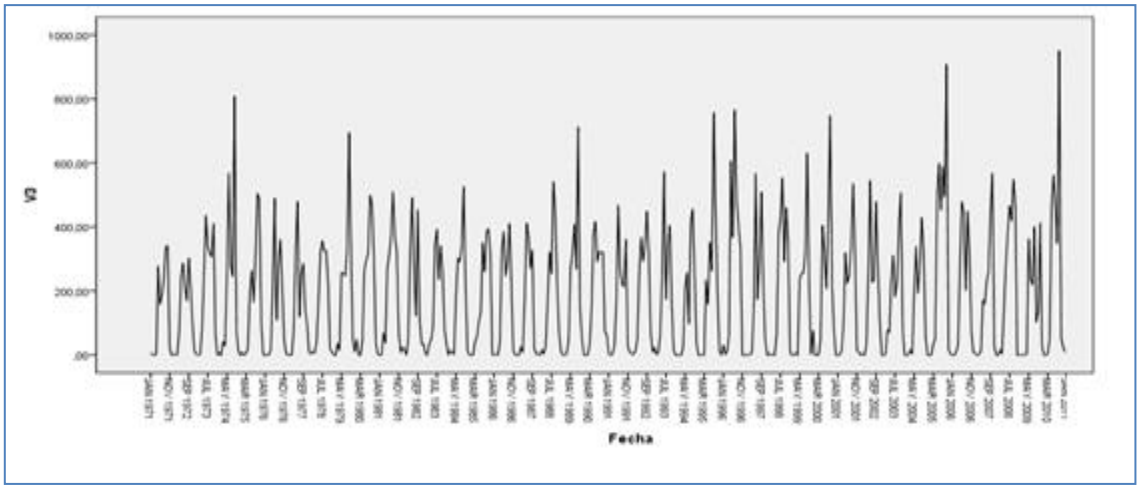
Considero que a medida que la altitud aumenta desde 0 metros hasta un punto "C", la cantidad de lluvia captada va disminuyendo; si seguimos subiendo de altitud en el terreno a partir del punto "C", la cantidad de lluvia va aumentando.

Esta es la conclusión a la que he llegado y el porqué en los resultados obtenidos aparecen las variables mezcladas. Estos resultados nos permiten tener una idea de cómo están relacionadas las variables en estudio y a la vez reducir el número de estaciones a analizar.

Para continuar con el análisis descriptivo tomaremos la serie de datos de la estación con código v3 correspondiente a la estación localizada en la Finca San Jacinto en San Vicente.

Como los datos de la base original están diarios, tomaremos el acumulado mensual para formar nuestra serie V3.

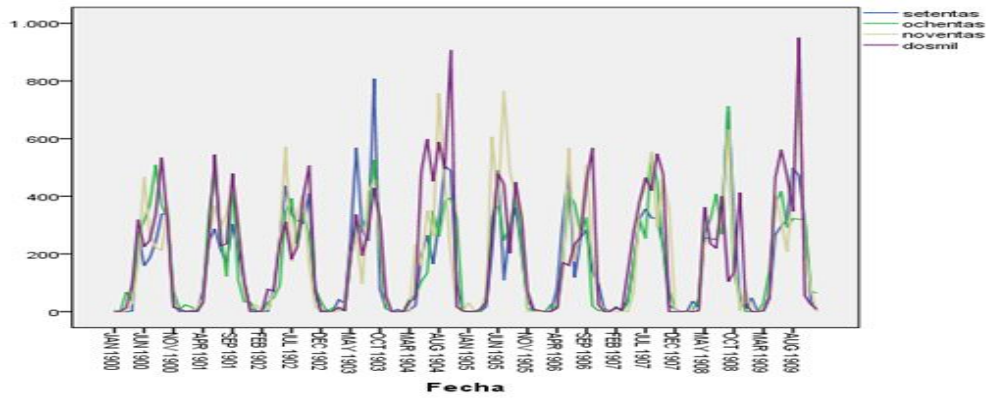
Primeramente se obtendrá lo que llamamos Gráfica de secuencia:



Gráfica 18. Gráfica de secuencia de la serie V3.

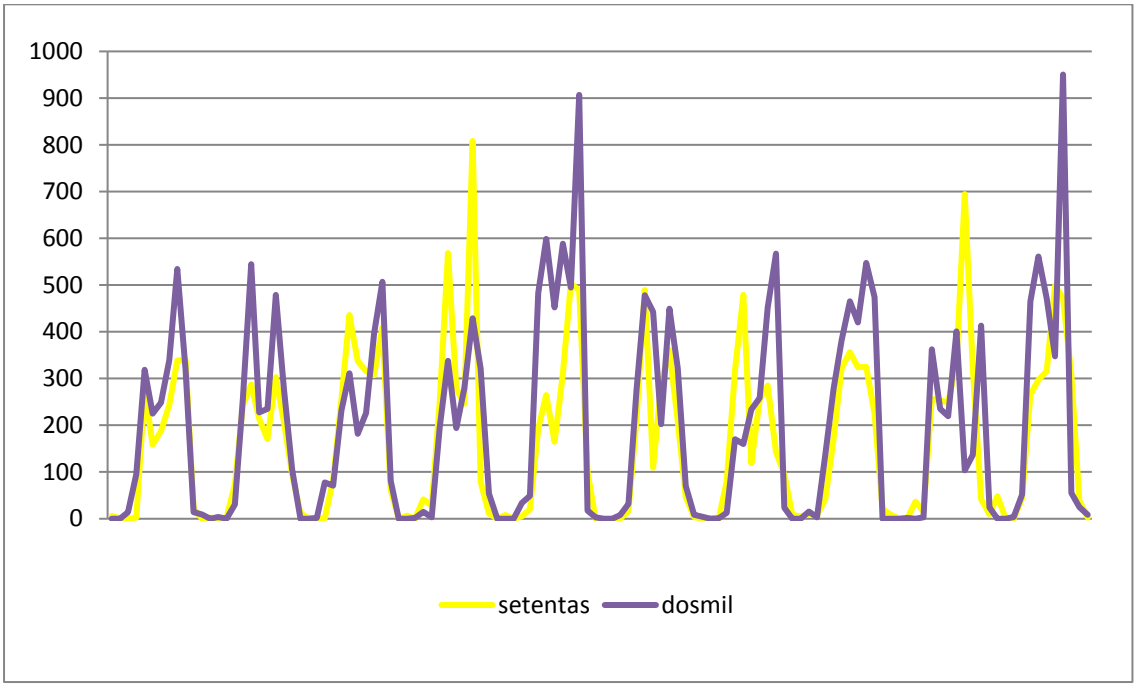
Al observar el Gráfica 18, se nota un comportamiento que es estacional, es decir, que se repite periódicamente un comportamiento similar periódicamente y que en nuestro caso ese periodo es de un año; con lo cual se justifica la aplicación de los modelos ARIMA haciendo uso de la metodología Box-Jenkins, de igual manera para los modelos autorregresivos que emplearemos en el análisis por medio de redes neuronales artificiales.

Partiremos esta serie en segmentos de diez años y veremos su desenvolvimiento en el siguiente Gráfica:



Gráfica 19. Gráfica de secuencia de la serie V3 particionada en décadas.

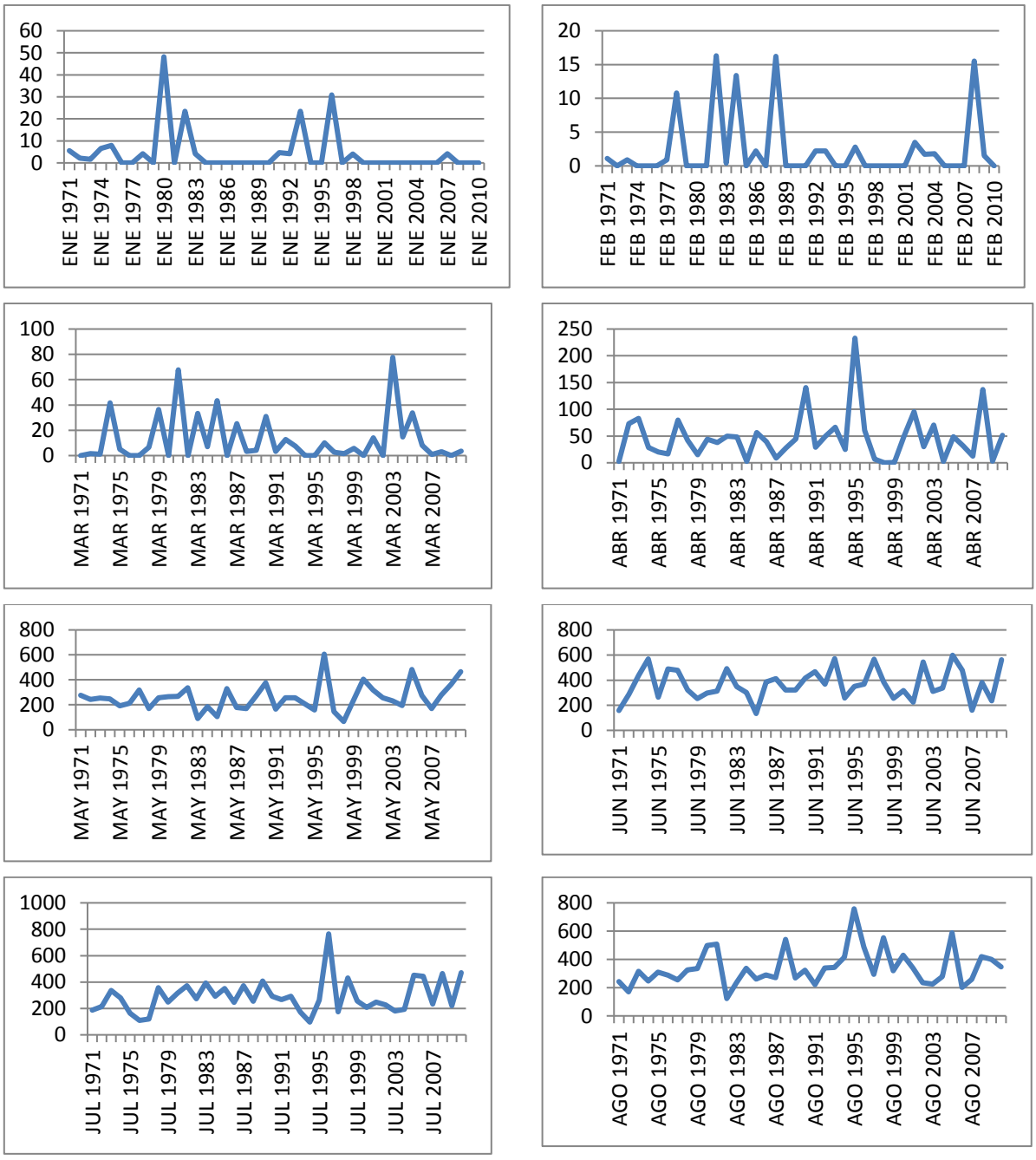
En el Gráfica 19 se ha dividido la serie V3 por décadas y podemos notar que en cada década, el comportamiento es similar aunque con un incremento leve de la cantidad de lluvia en relación a décadas anteriores, como por ejemplo si comparamos la década de los setenta con la década del dos mil, se tendrá lo siguiente:



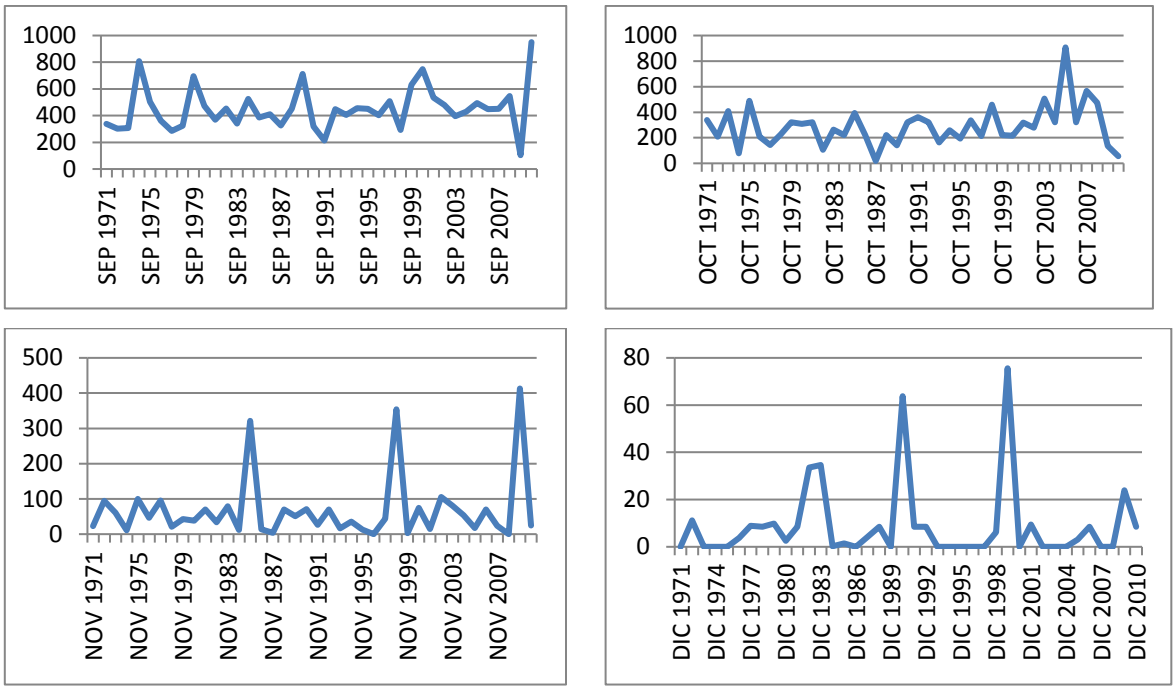
Gráfica 20. Gráfica de secuencia de la serie V3
Décadas de los 70's y 2000.

Se nota claramente un incremento en algunos ciclos, lo cual hace suponer que el ciclo del comportamiento normal de las precipitaciones pluviales se ha visto afectado de alguna manera.

Ahora veamos el comportamiento de la lluvia por cada mes:



Gráfica 21. Gráfica de secuencia de la serie V3 restringida para cada mes Meses de enero hasta agosto.



Gráfica 22. Gráfica de secuencia de la serie V3 restringida para cada mes Meses de septiembre hasta diciembre.

En este conjunto de Gráficas podemos notar que en algunos meses se muestra un incremento y una tendencia aunque no muy notoria ya que su comportamiento no parece ser cíclico.

Pero se ve perfectamente que el patrón de comportamiento ha sufrido un cambio y es que debemos analizar con más detalle los meses de la época lluviosa que va desde abril hasta octubre, en algunos de esos meses se nota un considerable aumento en relación al tiempo como por ejemplo para el mes de octubre, en los últimos años se ha notado un considerable aumento de la cantidad de lluvia en relación a las décadas de los setenta y ochentas.

Podemos ver que en los meses de noviembre y diciembre que son los meses de finalización de la época lluviosa y comienzo de la época seca, la lluvia se ha incrementado ligeramente y mostrando una tendencia a la alza.

8.3. Análisis de la serie cronológica aplicando la metodología Box-Jenkins y los modelos ARIMA

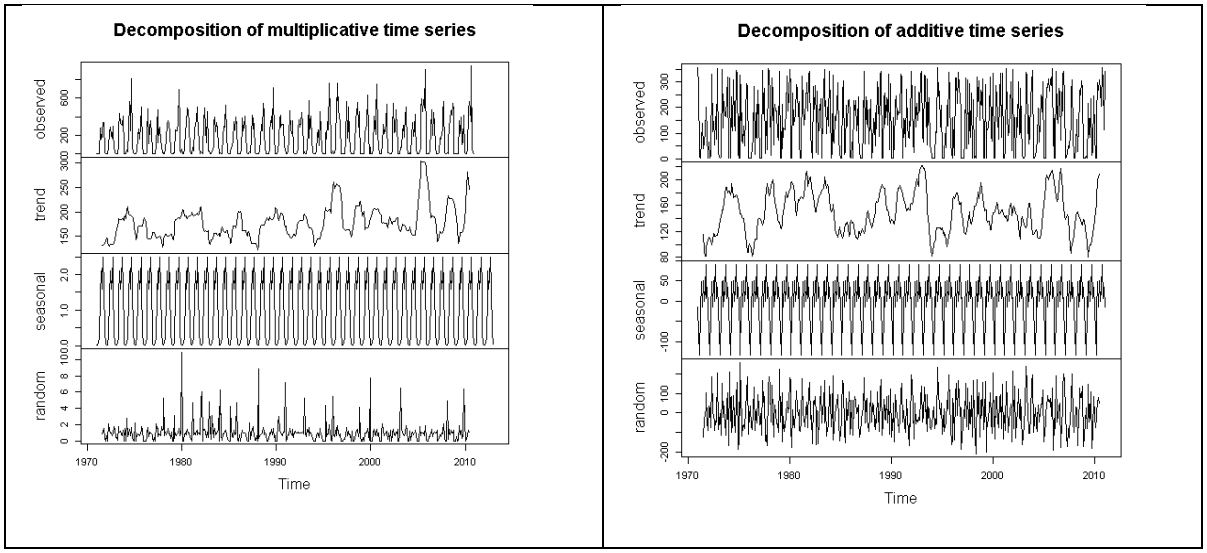
Para empezar, nuestro análisis, hemos elegido una de las variables, V3 que es la que mayor cantidad de años completos posee, obtenemos el Gráfica de secuencia lo cual ya se hizo en el análisis descriptivo (Gráfica 18).

Al observar el Gráfica de secuencia podemos notar una ligera tendencia a incrementarse en los picos de mayor cantidad de lluvia captada, además de un ciclo estacional, con lo que se concluye que la serie no es estacionaria pero emplearemos las pruebas para verificar si la serie es estacionaria o no.

Antes de hacer la prueba de estacionariedad, vamos a descomponer la serie en sus elementos básicos que son tendencia, estacionalidad y componente aleatoria.

Como no estamos seguros si el modelo buscado es aditivo o multiplicativo, realizaremos la descomposición considerando ambos casos.

Por medio del paquete R se tienen los siguientes resultados:

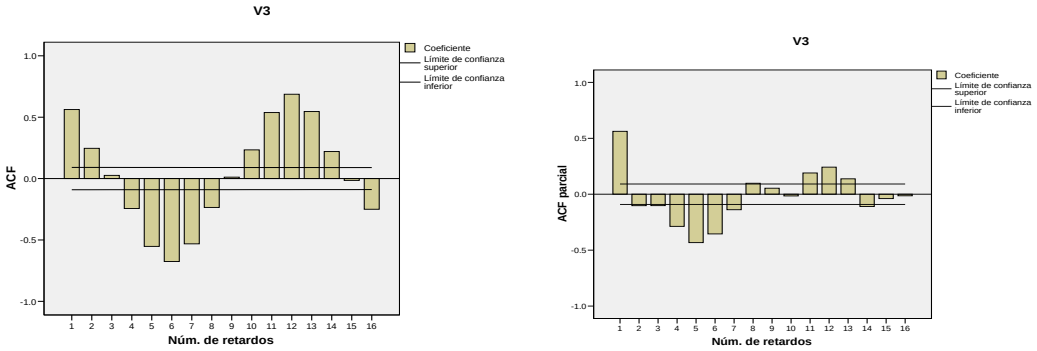


Gráfica 23. Descomposición de la serie V3 en sus elementos.

En ambos Gráficas se observa que la serie no posee una tendencia definida, pero si se muestra un componente estacional.

Al observar la variabilidad en ambos Gráficas, se observa que esta es menor en el modelo multiplicativo, lo cual indica que este es el más indicado de utilizar.

Lo que sigue es obtener los Gráficas de las funciones de autocorrelación y autocorrelación parcial conocidos como FAS y FAP respectivamente.



Gráfica 24. Gráficas de la funciones FAS y FAP serie V3.

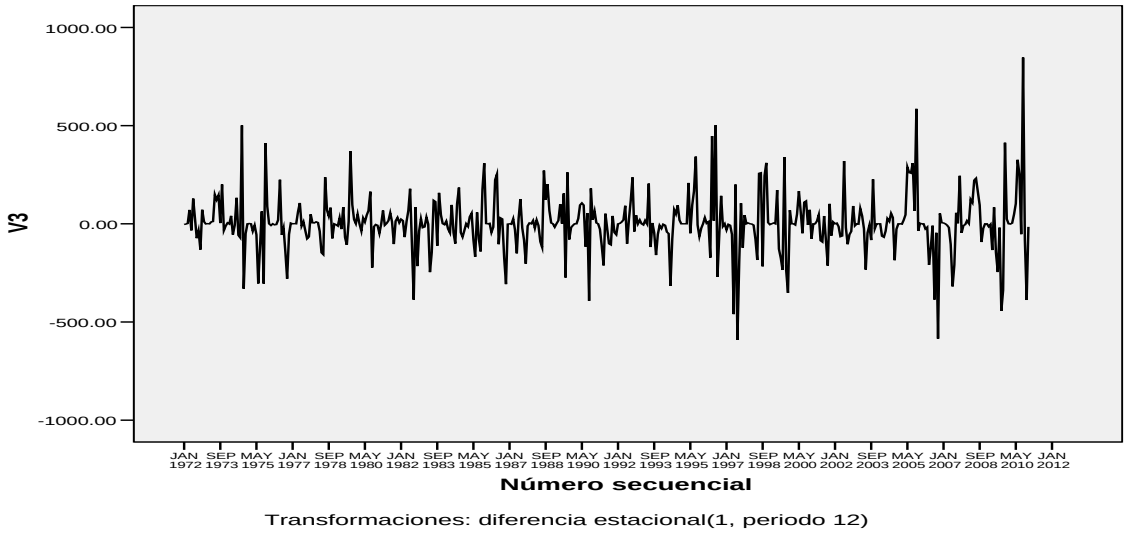
En ambas funciones, la FAS y la FAP (Gráfica 24) se nota un decaimiento lento sinusoidal, lo cual muestra que la serie no es estacionaria y necesita diferenciación.

A continuación la prueba de estacionariedad de Dickey-Fuller aumentada.

```
> adf.test(z.ts)
        Augmented Dickey-Fuller Test
data:  z.ts
Dickey-Fuller = -9.3057, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(z.ts) : p-value smaller than printed p-value
```

Podemos apreciar los resultados de la prueba de estacionariedad, dichos resultados confirman que la serie es estacionaria, lo cual es cierto ya que no se observa ninguna tendencia sino simplemente el comportamiento cíclico; por lo visto en los Gráficas de secuencia, de descomposición y correlaciones, la serie en estudio necesita ser diferenciada estacionalmente.

Luego de aplicar diferenciación de orden estacional un periodo para eliminar el comportamiento cíclico, se tiene el siguiente Gráfica de secuencia:



Gráfica 25. Gráfica de la serie V3 diferenciada un periodo estacional

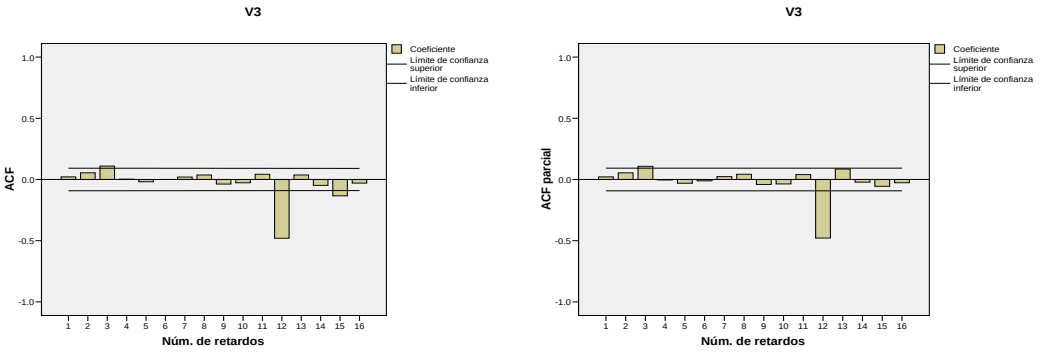
Al aplicar diferenciación estacional de orden uno, se muestra una serie más estable, aun cuando su variabilidad aumenta en los últimos periodos, lo cual hace sospechar un incremento en los eventos extremos como son sequias y temporales.

Antes de proseguir con la siguiente etapa vamos a realizar nuevamente la prueba de estacionariedad.

```
> adf.test(z.dif)
        Augmented Dickey-Fuller Test
data:  z.dif
Dickey-Fuller = -39.169, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(z.dif) : p-value smaller than printed p-value
> pp.test(z.dif)
        Phillips-Perron Unit Root Test
data:  z.dif
Dickey-Fuller Z(alpha) = -860.78, Truncation lag parameter = 5, p-value
= 0.01
alternative hypothesis: stationary
Warning message:
In pp.test(z.dif) : p-value smaller than printed p-value
```

Las pruebas nuevamente concluyen que la serie es estacionaria, pero en este caso ya no es necesario diferenciar la serie ni en la parte regular ni en la parte estacional, debido a que no tiene tendencia ni comportamiento estacional.

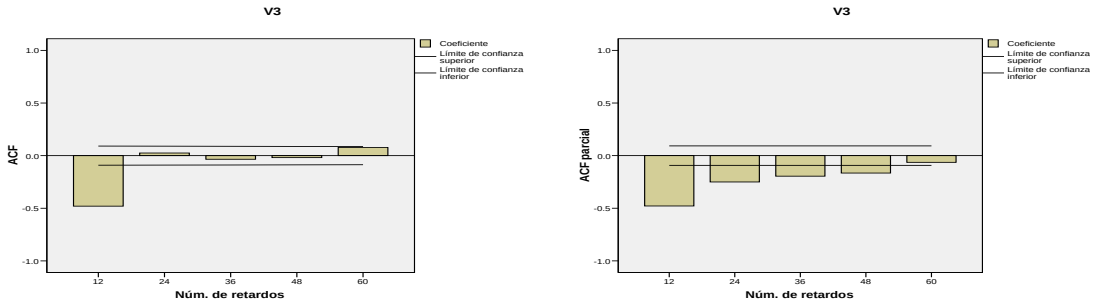
Observemos los Gráficas de la FAS y FAP de la serie diferencia.



Gráfica 26. Gráficas de la funciones FAS y FAP serie V3 Diferenciada un periodo estacional.

Al observar el Gráfica 25 de la FAS y FAP de la serie V3 diferenciada estacionalmente un periodo, se concluye que la mayoría de valores están bajo los límites y ya es posible identificar uno o más modelos que se le pueden ajustar a los datos de la serie.

Pero como se puede observar, estos modelos asociados son del tipo estacional y no incluyen ningún coeficiente en la parte regular, por lo que antes es necesario ver los Gráficas de la FAS y FAP de los periodos estacionales para la serie diferenciada.



Gráfica 27. Gráficas de la FAS y FAP estacionales de la serie V3 diferenciada un periodo estacional.

De acuerdo al Gráfica de la FAS de los órdenes estacionales, en el modelo puede existir la presencia de un $MA(1)_S$ ya que todos los coeficientes se encuentran bajo las bandas de confianza exceptuando el primer valor.

Conforme al Gráfica de la FAP podríamos suponer la presencia de un modelo $AR(4)_S$ o la confirmación de un modelo MA estacional. Esto debido al comportamiento en descenso de los coeficientes.

Juntando los resultados de ambos Gráficas también podríamos suponer la existencia de un $ARMA(4,1)_S$

Por todo lo anterior, podemos identificar los posibles modelos, estos son:

- ARIMA (0,1,1)s
- ARIMA (4,1,0)s
- ARIMA (4,1,1)s

- **Modelo a analizar: ARIMA (0,1,1)s**

Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		.574	.412	1.393	.164
			Diferenciación estacional		1			
			MA, estacional	Retardo 1	1.000	25.399	.039	.969

Tabla 3. Parámetros estimados de un ARIMA (0,1,1)s para la serie V3.

Este modelo es rechazado ya que todos los coeficientes son rechazados, es decir, el nivel de significancia de cada uno de ellos es mayor que 0.05 con lo cual se rechaza la hipótesis de que sean diferentes de cero.

- **Modelo a analizar: ARIMA (4,1,0)s**

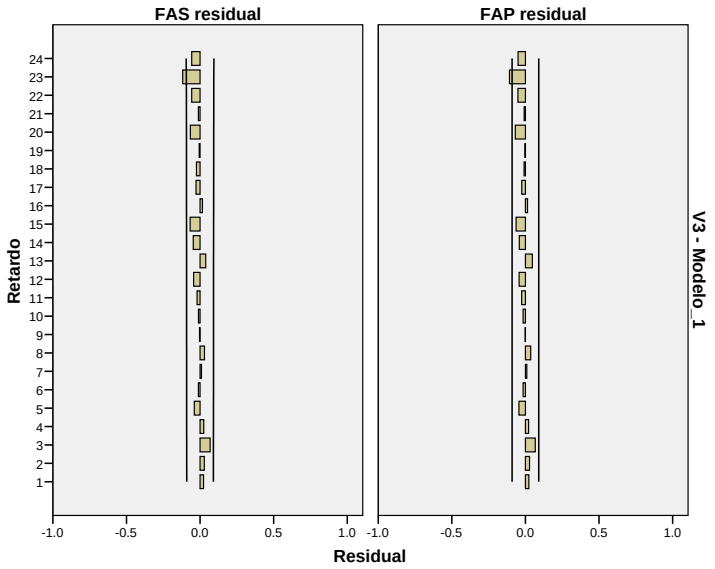
Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		1.256	1.590	.790	.430
			AR, estacional	Retardo 1	-.846	.047	-17.871	.000
				Retardo 2	-.646	.060	-10.858	.000
				Retardo 3	-.509	.060	-8.544	.000
				Retardo 4	-.284	.050	-5.714	.000
			Diferenciación estacional		1			

Tabla 4. Parámetros estimados de un ARIMA (4,1,0)s para la serie V3.

Este modelo es aceptado, excepto por la constante que no es diferente de cero, ya que su nivel de significancia es mayor que 0.05.

Acompañado a los coeficientes estimados viene el Gráfica de la FAS y FAP de los residuos del modelo anterior.



Gráfica 28. Funciones FAS y FAP de los residuales del modelo ARIMA (4,1,0)s

Se puede notar que la mayoría de los valores residuales caen bajo las bandas de confianza, por lo que a simple vista este modelo parece ser el adecuado.

Ajuste del modelo

Estadístico de ajuste	Media	ET	Mínimo	Máximo	Percentil 5	10	25	50	75	90	95
R-cuadrado estacionaria	.413	.	.413	.413	.413	.413	.413	.413	.413	.413	.413
R-cuadrado	.674	.	.674	.674	.674	.674	.674	.674	.674	.674	.674
RMSE	109.735	.	109.735	109.735	109.735	109.735	109.735	109.735	109.735	109.735	109.735
MAPE	193.046	.	193.046	193.046	193.046	193.046	193.046	193.046	193.046	193.046	193.046
MaxAPE	18201.104	.	18201.104	18201.104	18201.104	18201.104	18201.104	18201.104	18201.104	18201.104	18201.104
MAE	65.521	.	65.521	65.521	65.521	65.521	65.521	65.521	65.521	65.521	65.521
MaxAE	580.903	.	580.903	580.903	580.903	580.903	580.903	580.903	580.903	580.903	580.903
BIC normalizado	9.462	.	9.462	9.462	9.462	9.462	9.462	9.462	9.462	9.462	9.462

Tabla 5. Estadísticos de ajuste del modelo ARIMA (4,1,0)s

Agregamos la tabla de los estadísticos de ajuste por si aparece otro modelo que se le pueda ajustar a los datos y elegir el mejor de ellos.

- **Modelo a analizar: ARIMA (4,1,1)s**

Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		1.234	.372	3.315	.001
			AR, estacional	Retardo 1	-.075	.057	-1.307	.192
				Retardo 2	-.037	.059	-.619	.536
				Retardo 3	-.089	.058	-1.546	.123
				Retardo 4	-.005	.058	-.093	.926
			Diferenciación estacional		1			
			MA, estacional	Retardo 1	.960	.053	18.154	.000

Tabla 6. Parámetros estimados de un ARIMA (4,1,1)s para la serie V3.

Este modelo es rechazado (tabla 4), porque la mayoría de los coeficientes, exceptuando el coeficiente de la parte de medias móviles son rechazados como diferentes de cero, y es raro ya que anteriormente verificamos un modelo de medias móviles estacional y este se rechazaba.

Al final nos quedamos solo con un modelo: ARIMA (4,1,0)s.

Ahora lo que sigue es la verificación de los supuestos con respecto a los residuos del modelo.

Se puede notar que la mayoría de valores residuales (Gráfica 31), tanto en la FAS como en la FAP están bajo las bandas de confianza, lo cual nos confirma que los residuos no poseen correlación entre sí.

A continuación las pruebas de estacionariedad

```
> adf.test(z.ts)
        Augmented Dickey-Fuller Test
data:  z.ts
Dickey-Fuller = -7.0999, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(z.ts) : p-value smaller than printed p-value
> pp.test(z.ts)
        Phillips-Perron Unit Root Test
data:  z.ts
Dickey-Fuller Z(alpha) = -480.5, Truncation lag parameter = 5, p-value
= 0.01
alternative hypothesis: stationary
Warning message:
In pp.test(z.ts) : p-value smaller than printed p-value
```

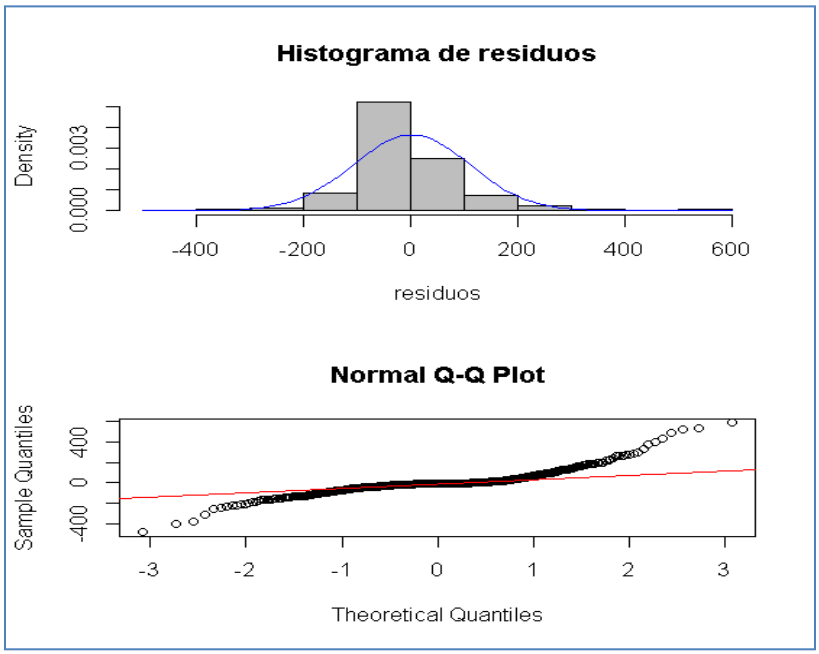
Las pruebas de Dickey-Fuller y Philip-Perron confirman que los residuos no poseen ninguna tendencia.

Lo que sigue son las pruebas de normalidad de los residuos:

```
> shapiro.test(z.ts)

        Shapiro-Wilk normality test

data:  z.ts
```

```
W = 0.85462, p-value < 2.2e-16
> jarque.bera.test(z.ts)
        Jarque Bera Test
data:  z.ts
X-squared = 898.49, df = 2, p-value < 2.2e-16
> ks.test(z.ts,pnorm)
        One-sample Kolmogorov-Smirnov test
data:  z.ts
D = 0.60674, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(z.ts, pnorm) :
  ties should not be present for the Kolmogorov-Smirnov test
```

Las pruebas de normalidad arrojan que los residuos no tienen una distribución normal por lo cual observaremos gráficamente que es lo que sucede con los residuos del modelo.



Gráfica 29. Histograma de los residuales y Gráfica Q-Q.

Al obtener un histograma podemos apreciar que los residuos poseen valores extremos, lo cual hace que se presente una asimetría y en el Gráfica q-q se nota además que muchos valores están sobre la recta, pero no todos los puntos caen sobre ella y esa es la razón de que las pruebas rechacen que los residuos tengan una distribución normal.

Para finalizar con esta sección, analizaremos la variable V3 por décadas y extraeremos el modelo en cada una de ellas, teniéndose los siguientes resultados:

Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		3.139	2.816	1.115	.267
			AR, estacional	Retardo 1	-.751	.095	-7.935	.000
				Retardo 2	-.625	.109	-5.718	.000
				Retardo 3	-.609	.108	-5.621	.000
				Retardo 4	-.388	.101	-3.853	.000
			Diferenciación estacional		1			

Tabla 7. Parámetros estimados serie V3 parcial, 1971-1980.

Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		.842	2.353	.358	.721
			AR, estacional	Retardo 1	-1.050	.099	-10.637	.000
				Retardo 2	-.884	.135	-6.569	.000
				Retardo 3	-.791	.133	-5.963	.000
				Retardo 4	-.394	.110	-3.575	.001
			Diferenciación estacional		1			

Tabla 8. Parámetros estimados serie V3 parcial, 1981-1990.

Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		4.875	4.462	1.093	.277
			AR, estacional	Retardo 1	-.816	.106	-7.679	.000
				Retardo 2	-.673	.147	-4.584	.000
				Retardo 3	-.448	.154	-2.907	.004
				Retardo 4	-.253	.156	-1.626	.107
			Diferenciación estacional		1			

Tabla 9. Parámetros estimados serie V3 parcial, 1991-2000.

Parámetros del modelo ARIMA

					Estimación	ET	t	Sig.
V3-Modelo_1	V3	Sin transformación	Constante		3.474	5.373	.647	.519
			AR, estacional	Retardo 1	-.915	.118	-7.754	.000
				Retardo 2	-.549	.182	-3.019	.003
				Retardo 3	-.285	.192	-1.482	.141
				Retardo 4	-.111	.151	-.736	.463
			Diferenciación estacional		1			

Tabla 10. Parámetros estimados serie V3 parcial, 2001-2010.

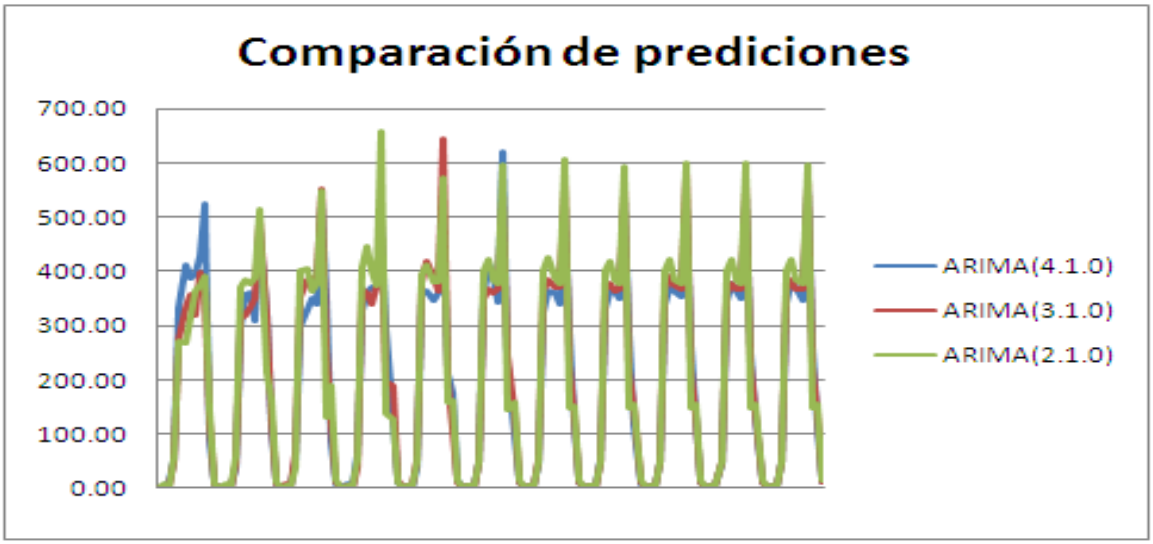
Anteriormente habíamos encontrado un modelo $ARIMA(4,1,0)_S$ para la serie completa V3, y es por eso que aplicamos dicho modelo a cada década.

Lo que se concluye de las tablas 4, 5, 6 y 7 es que el comportamiento ha cambiado, en las décadas de los setenta y ochenta presenta un modelo $ARIMA(4,1,0)_S$ con modificaciones en los coeficientes, lo que nos indica que el clima no variaba mucho en esos días, pero en la década de los noventa se vio modificado a un $ARIMA(3,1,0)_S$ y ya para la década de dos mil en adelante se tiene un $ARIMA(2,1,0)_S$.

Con esto se confirma de alguna manera el cambio climático está afectando de manera brusca nuestro entorno y en especial las precipitaciones pluviales.

Pero antes de proseguir veamos el comportamiento de estos modelos al realizar predicciones 10 años adelante.

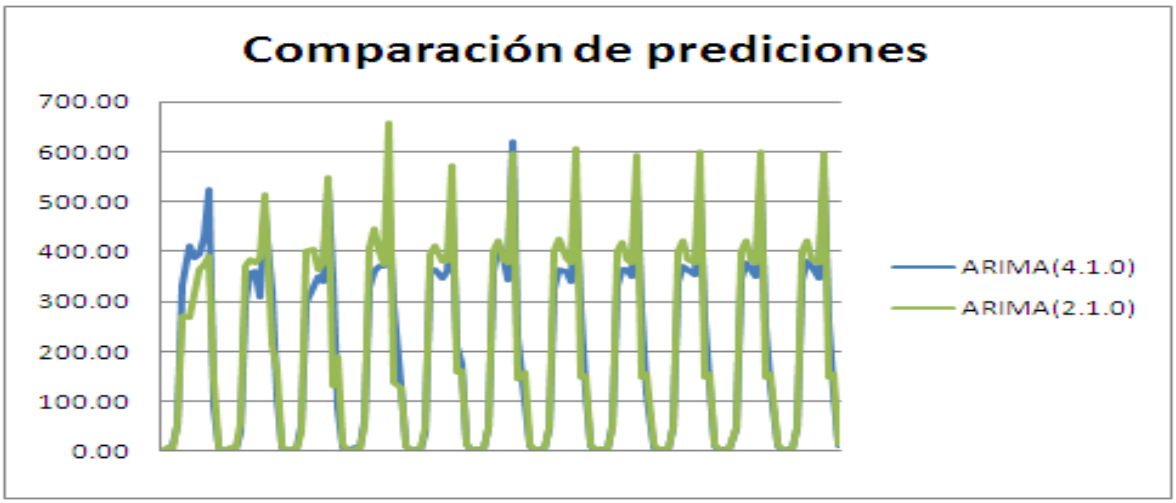
Haciendo uso del SPSS obtenemos las predicciones aplicando los respectivos modelos encontrados y llevando estos a Excel para tener siguiente Gráfica:



Gráfica 30. Predicciones 10 años adelante empleando los tres modelos encontrados

Notamos que el modelo $ARIMA(4,1,0)$ se estabiliza menos rápido que los otros dos modelos, esto es porque los coeficientes lo indican, la correlación se va perdiendo al avanzar al futuro y las series se estabilizan.

Esto no solo nos indica que la lluvia en el pasado era más estable y fácil de predecir, en cambio las condiciones han cambiado y ahora los periodos de lluvia son más impredecibles.



Gráfica 31. Comparación de las predicciones de los modelos ARIMA(4,1,0)s y ARIMA(2,1,0)s.

Hay una cosa más que los Gráficas anteriores nos dicen y es que si se observa con detenimiento, podemos apreciar que la cantidad de lluvia va disminuyendo en relación a la década anterior, ya que las predicciones del $ARIMA(4,1,0)$ son inferiores a las del modelo $ARIMA(3,1,0)$ y estas a su vez son inferiores a las del modelo $ARIMA(2,1,0)$, puede que esta diferencia sea mínima pero está ahí y confirma lo que mencionamos anteriormente sobre el cambio en el comportamiento del ciclo lluvioso en nuestro país.

Obsérvese que dijimos “cantidad de lluvia” y es que nuestra variable es lluvia acumulada mensual.

Finalmente y debido a los resultados del análisis factorial múltiple y los obtenidos anteriormente, se puede concluir que el comportamiento de las lluvias no se puede generalizar a todo el territorio nacional, solamente a zonas cercanas a la estación V3, esto se debe a que las lluvias varían dependiendo de la altitud del terreno y de otros factores. Lo que sí se puede hacer es identificar la correlación existente entre la estación V3 y cualquier otra de interés y con ello efectuar estimaciones empleando represión lineal o modelos de series de tiempo bivariante o multivariante.

8.4. Análisis de la serie cronológica aplicando la metodología de redes neuronales artificiales

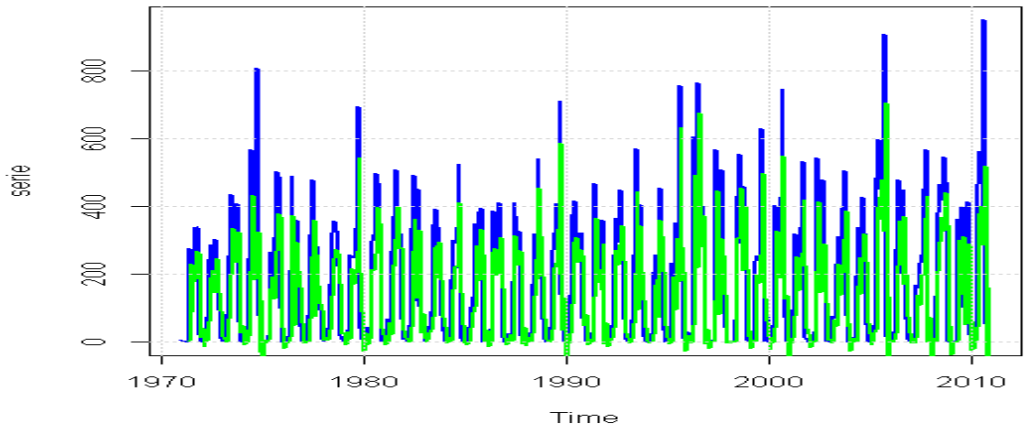
Para esta sección utilizaremos los mismos datos de la variable V3 empleados en el análisis de series de tiempo, separamos una porción de los datos para el entrenamiento de la red neuronal y la estimación del modelo planteado.

A continuación se presenta el código en R y los resultados obtenidos:

```
> ### Separando la muestra de estimación ###
> y <- ts(serie[1:400],s=1,f=1)
> fit <- arnn(x=y,lags=1:4,H=2)
> fit

Method: arnn
Call:
 arnn(x = y, lags = 1:4, H = 2)
Parameters:
        M       Wio[1]       Wio[2]       Wio[3]       Wio[4]       Wih[1,1]
-0.00958145  0.82600944  -0.15442710  0.16765690  -0.03426351  -0.39362899

   Wih[2,1]     Wih[3,1]     Wih[4,1]     Wih[1,2]     Wih[2,2]     Wih[3,2]
 0.17466466  0.27494961  -0.89031854  -0.30553358  0.80835041  -0.27835045

   Wih[4,2]           Wbh[1]       Wbh[2]          Who[1]       Who[2]          Wbo
 0.88927249  0.26431881  0.92489943  -0.39096871  0.96776018  0.77315371

Sigma^2 estimated as:  23593.1917735596
LogL:  -2531.19174419993

Information Criteria:
     AK          AKc                HQ                SC
5098.383    5100.217    5126.714    5169.866
```

Luego obtendremos los pronósticos del modelo estimado incluyendo un Gráfica de secuencia de esos valores.

```
> fit1 <- arnn(x=serie,model=fit)
> plot(serie,lwd=2,col="blue")
> lines(fit1$fitted,col="green",lwd=2)
> grid()
```



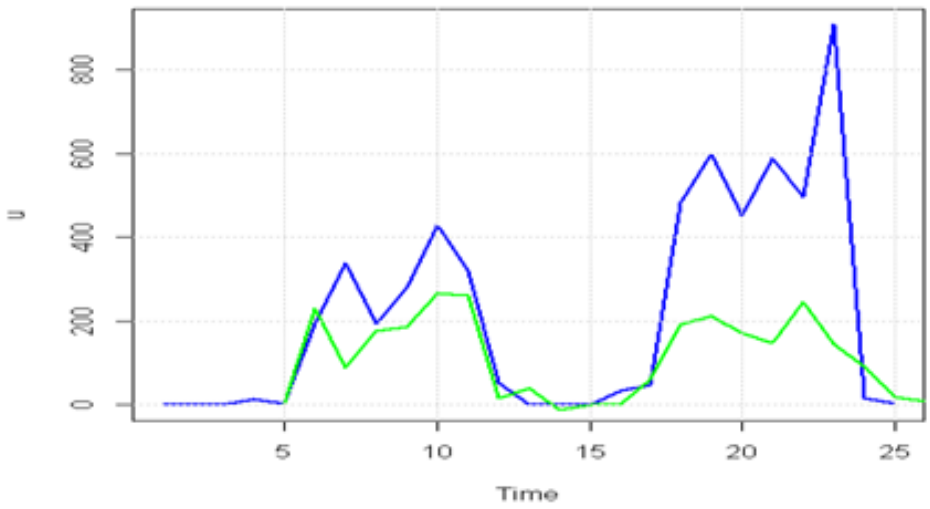
Gráfica 32. Datos reales y estimados por medio de Redes Neuronales Artificiales

Se muestra en azul los valores observados y en verde las estimaciones; cómo podemos apreciar en este Gráfica, la red neuronal logro identificar el patrón de comportamiento de manera exitosa, aunque no pudo ajustar muy bien los valores extremos que se presentan en la serie, esto se concluye porque las estimaciones no alcanzan los picos más elevados de la gráfica.

Como en el principio hicimos una separación de los valores reservando datos de la serie, esos se utilizarán para verificar gráficamente las predicciones del modelo de red neuronal, la corrida del código en R se muestra a continuación:

```
> k <-forecast.arnn(fit,h=20)
> u <- ts(serie[396:420],s=1,f=1)
> plot(u,lwd=2,col="blue")
> lines(k$fitted,col="green",lwd=2)
> grid()
```

El Gráfica resultante es el que se muestra a continuación:



Gráfica 33. Predicciones dos años adelante usando RNA.

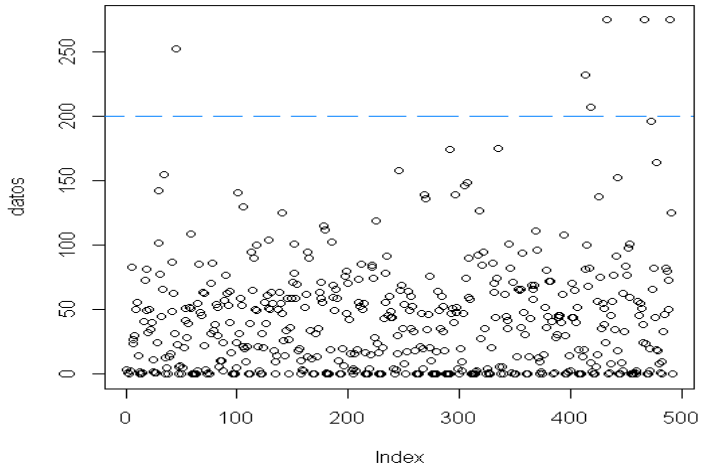
Se hicieron predicciones dos años adelante para poder compararlos con los datos que se habían apartado de 2011 y 2012. Como podemos apreciar, los valores estimados (color verde) están por debajo de los valores observados (color azul) y esto confirma que la red neuronal es buena estimando el patrón de comportamiento, pero no es muy adecuada para la predicción de valores extremos de una serie de tiempo climática como lo es la precipitación pluvial.

8.5. Análisis de la serie cronológica aplicando la metodología de valores extremos

Para este tipo de análisis pensábamos en un principio utilizar un software especializado, pero tomamos la decisión de emplear el paquete R y la librería extremes que nos permite obtener los mismos resultados y de manera más rápida.

Para este análisis empleamos los máximos valores mensuales de las precipitaciones pluviales de la serie V3, ya que esta serie es la más completa que tenemos a disposición.

Pero no solo emplearemos los datos máximos mensuales sino también obtendremos los máximos acumulados de dos, tres y cuatro días consecutivos, ya que si recordamos, en nuestro país no nos afectan mucho las lluvias fuertes sino más bien la acumulación de agua debido a la susceptibilidad del suelo.
Iniciamos con una representación gráfica de los datos:



Gráfica 34. Gráfica de dispersión de los valores máximos mensuales de la variable V3.

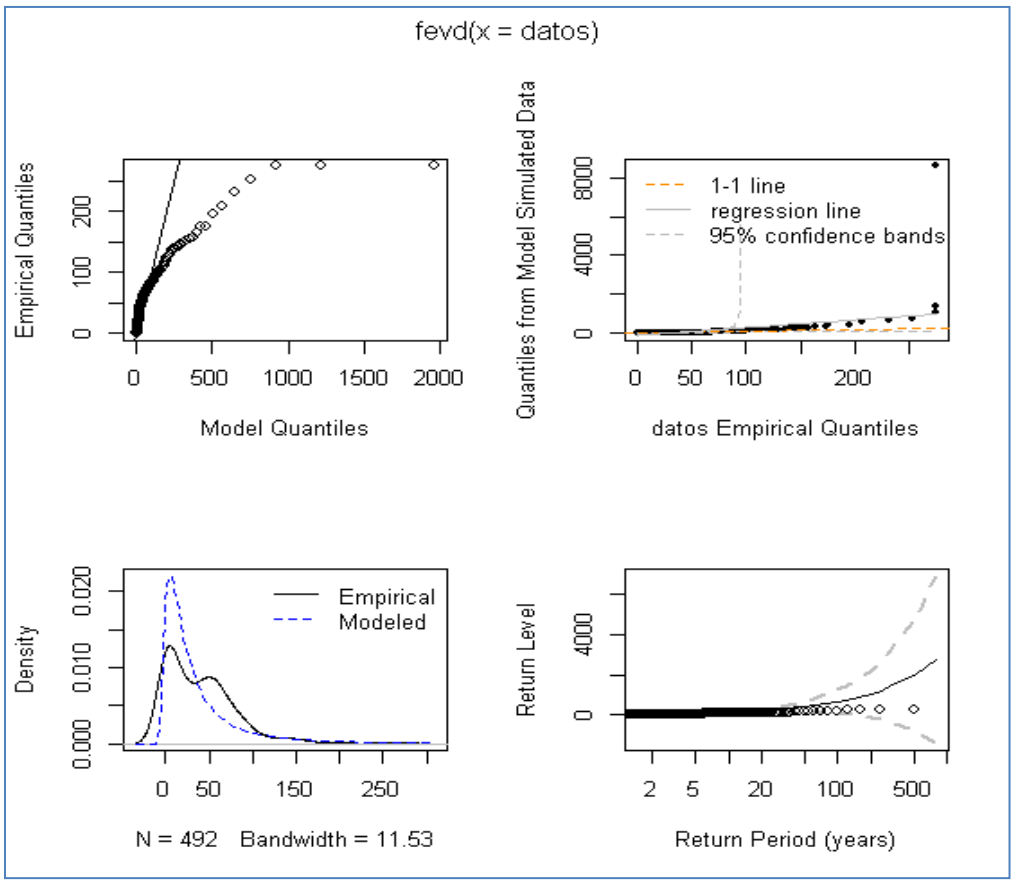
Se puede notar que son pocos los valores mensuales que sobrepasan los 200 milímetros cúbicos de agua, esta cota sólo es un punto arbitrario para describir los valores máximos extremos de este Gráfica.

```
> fit <- fevd(datos)
> fit

fevd(x = datos)

[1] "Estimation Method used: MLE"

 Negative Log-Likelihood Value:  2426.472

 Estimated parameters:
  location      scale      shape
14.2737837 20.0763318  0.6782141
```

```
 Standard Error Estimates:
 location     scale     shape
1.9152202 1.8076426 0.1595191

 Estimated parameter covariance matrix.
          location      scale       shape
location  3.6680684  3.2140407 -0.25394567
scale     3.2140407  3.2675717 -0.20303165
shape    -0.2539457 -0.2030317  0.02544636

 AIC = 4858.943

 BIC = 4871.539
```

Vamos a analizar los resultados obtenidos diciendo que los primeros dos valores estimados son la media y la desviación estándar, ya que son los parámetros de localización y de escala. El tercer término estimado es de forma y nos dice que tipo de familia de distribuciones que mejor se ajusta corresponde a los datos y en este caso el valor es mayor que cero, lo cual nos indica que es de la familia de funciones de Fréchet.

De lo que se debe hablar del análisis anterior es de la matriz de varianzas covarianzas, esta muestra que los coeficientes estimados de localización y escala están correlacionados entre si, mientras que los coeficientes de localización y escala no tiene mucha correlación con el coeficiente de forma o “shape” como aparece en los resultados; esta matriz también nos indica que hay una varianza diferente de cero en relación al valor estimado de la muestra y el valor estimado por el modelo.



Gráfica 35. Gráficas resultantes de la estimación de la función de valores extremos generalizada asociada a la serie V3 máximos mensuales.

Podemos ver que el modelo estimado es bastante bueno, se nota que no hay un ajuste perfecto pero que hay valores observados que se salen de las curvas estimadas.

La forma del Gráfica de densidad hace suponer que el modelo estimado no es una distribución de Fréchet sino de Gumbel, por lo cual se hará una estimación forzando el tipo de modelo a una distribución de Gumbel.

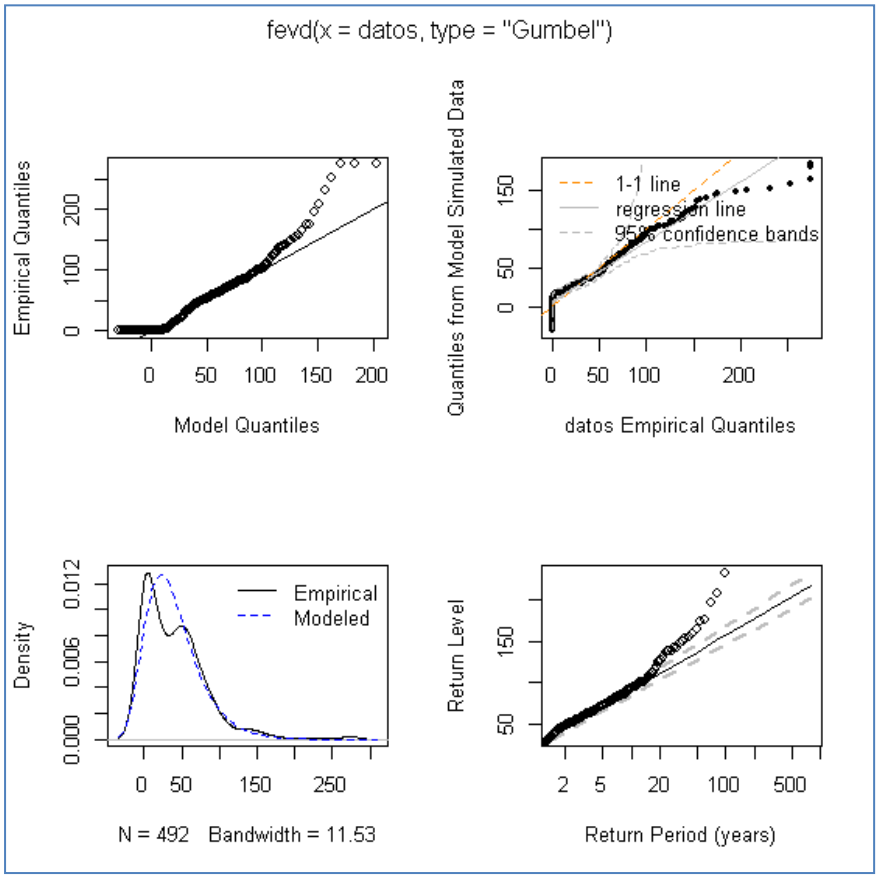
```
> fit <- fevd(datos,type="Gumbel")
> fit
```

```
fevd(x = datos, type = "Gumbel")
[1] "Estimation Method used: MLE"

 Negative Log-Likelihood Value:  2461.295

 Estimated parameters:
location    scale
23.18422 29.11002

 Standard Error Estimates:
location    scale
1.372961 1.094642

 Estimated parameter covariance matrix.
          location     scale
location 1.8850210 0.4415136
scale    0.4415136 1.1982406

 AIC = 4926.589
 BIC = 4934.986
```

Se estimó el modelo de una distribución de Gumbel que mejor se ajusta a los datos y al compararlo con el modelo anterior, se nota un aumento en el coeficiente de AIC lo cual nos hace ver que el modelo de Fréchet es mejor pero la diferencia es nada más de 32 puntos, eso da una pauta para tomar la decisión de cual modelo es mejor.

Lo que nos indica que este modelo es mejor que el anterior, ya que en la matriz de varianzas-covarianzas, podemos observar que la varianza disminuyo bastante en relación a la varianza de los parámetros estimados, más específicamente, en el modelo anterior la varianza era mayor que 3 y en este modelo es cercana a dos.

Veamos a continuación los resultados Gráficas del modelo de Gumbel estimado:



Gráfica 36. Gráficas resultantes de la estimación de la función de Gumbel asociada a la serie V3 máximos mensuales.

Se puede observar que al igual que el modelo anterior de Fréchet, este modelo no se ajusta perfectamente pero es bastante bueno para poder hacer predicciones.

El mejor ajuste se puede ver en el Gráfica de densidad, ya que a diferencia del de Fréchet, este se ajusta mejor a su forma empírica.

Ahora, analizaremos el modelo que se ajuste mejor a los datos máximos acumulado de dos días consecutivos de lluvia.

```
> fit <- fevd(datos)
Warning messages:
1: In log(z) : NaNs produced
2: In log(z) : NaNs produced
3: In log(z) : NaNs produced
4: In log(z) : NaNs produced
> fit
fevd(x = datos)

[1] "Estimation Method used: MLE"

 Negative Log-Likelihood Value:  2211.921

 Estimated parameters:
 location     scale     shape
 4.579103 25.483479  5.564944

 Standard Error Estimates:
   location       scale       shape
0.0001474781 0.0000000200 0.0001474781

 Estimated parameter covariance matrix.
              location        scale        shape
location  2.174979e-08 -2.165685e-18 -2.174979e-08
scale    -2.165685e-18  4.000001e-16  2.165685e-18
shape    -2.174979e-08  2.165685e-18  2.174979e-08

 AIC = 4429.841
 BIC = 4442.215
```

En la corrida de este código nos dio varias advertencias y es debido a que en algunos valores de la muestra se tiene que el dato es cero y sabemos que log(0) no está definida.

El tercer término estimado es de forma y nos dice que tipo de familia de distribuciones que mejor se ajusta corresponde a los datos y en este caso el valor es mayor que cero, lo cual nos indica que es de la familia de funciones de Fréchet.

La matriz de varianzas y covarianzas muestra que este modelo se ajusta bastante bien ya que todos los coeficientes de la matriz son cercanos a cero

Ahora, analizaremos el modelo que se ajuste mejor a los datos máximos acumulado de tres días consecutivos de lluvia.

```
fevd(x = datos)

[1] "Estimation Method used: MLE"

 Negative Log-Likelihood Value:  2226.792

 Estimated parameters:
 location     scale     shape
 4.707814 28.152907  5.979946

 Standard Error Estimates:
location    scale    shape
   2e-08    2e-08    2e-08

 Estimated parameter covariance matrix.
              location        scale        shape
location  4.000001e-16 4.956909e-25 -2.670206e-24
```

```
scale     4.956909e-25 4.000001e-16  5.868993e-25
shape    -2.670206e-24 5.868993e-25  4.000001e-16

 AIC = 4459.584
 BIC = 4471.958
```

El tercer término estimado es de forma y nos dice que tipo de familia de distribuciones es el que mejor se ajusta a los datos y en este caso el valor es mayor que cero, lo cual nos indica que es de la familia de funciones de Fréchet.

La matriz de varianzas y covarianzas muestra que este modelo se ajusta bastante bien ya que todos los coeficientes de la matriz son cercanos a cero

Finalmente, analizaremos el modelo que se ajuste mejor a los datos máximos acumulado de cuatro días consecutivos de lluvia.

```
fevd(x = datos)
[1] "Estimation Method used: MLE"

 Negative Log-Likelihood Value:  2274.635

 Estimated parameters:
 location     scale     shape
 5.957545 35.935499  6.031848

 Standard Error Estimates:
  location      scale      shape
0.00331508 0.00000002 0.00331508

 Estimated parameter covariance matrix.
              location         scale         shape
```

location 1.098976e-05 -1.076115e-16 -1.098976e-05

scale -1.076115e-16 4.000001e-16 1.076115e-16

shape -1.098976e-05 1.076115e-16 1.098976e-05

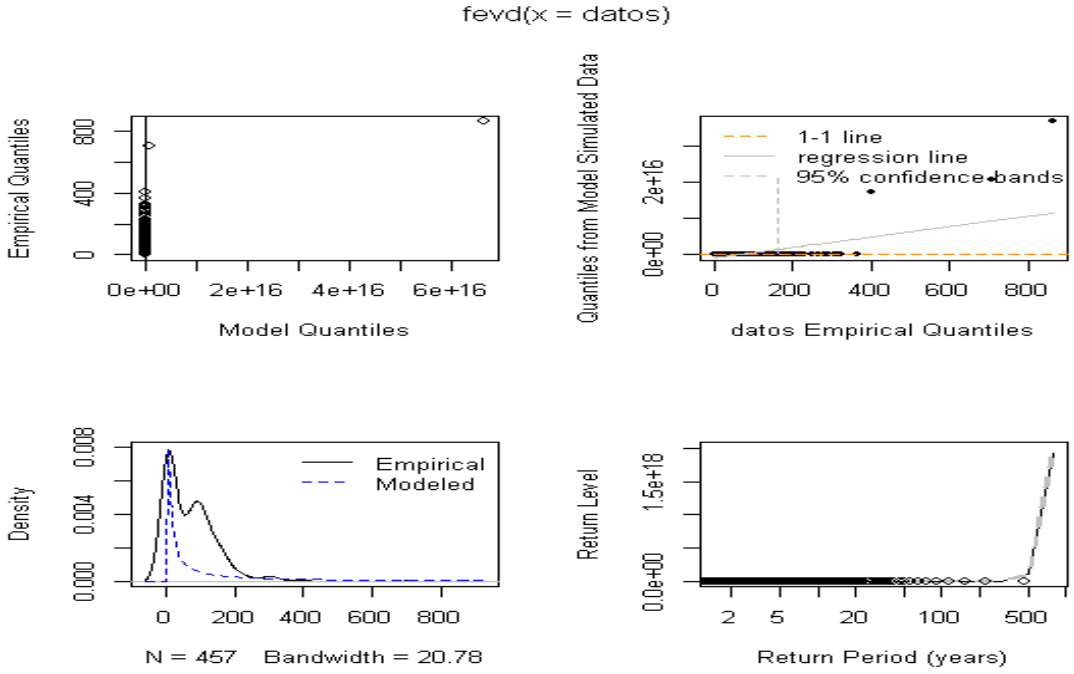
AIC = 4555.27

BIC = 4567.644

El tercer término estimado es de forma y nos dice que tipo de familia de distribuciones es el que mejor se ajusta a los datos y en este caso el valor es mayor que cero, lo cual nos indica que es de la familia de funciones de Fréchet.

La matriz de varianzas y covarianzas muestra que este modelo se ajusta bastante bien ya que todos los coeficientes de la matriz son cercanos a cero

Los resultados Gráficas del modelo estimado para cuatro días consecutivos de lluvia se muestran a continuación:



Gráfica 37. Gráficas resultantes de la estimación de la función asociada a la serie V3 lluvia máxima mensual acumulada cuatro días consecutivos.

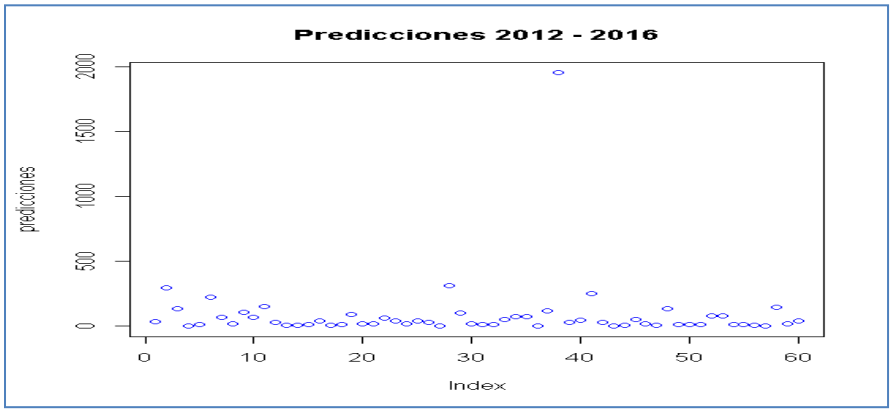
No parece ser un buen modelo ya que en el Gráfica de densidad, la figura de los datos empíricos se muestran más parecidos a una distribución de Gumbel.

Si efectuamos la estimación del modelo de Gumbel para los datos anteriores, se tiene lo siguiente:

```
fevd(x = datos, type = "Gumbel")
[1] "Estimation Method used: MLE"

 Negative Log-Likelihood Value:  2556.905
 Estimated parameters:
location    scale
38.97125 52.23975

 Standard Error Estimates:
location    scale
2.553588 2.053761
Estimated parameter covariance matrix.
         location    scale
location 6.520809 1.522105
scale    1.522105 4.217936

 AIC = 5117.811
 BIC = 5126.06
```

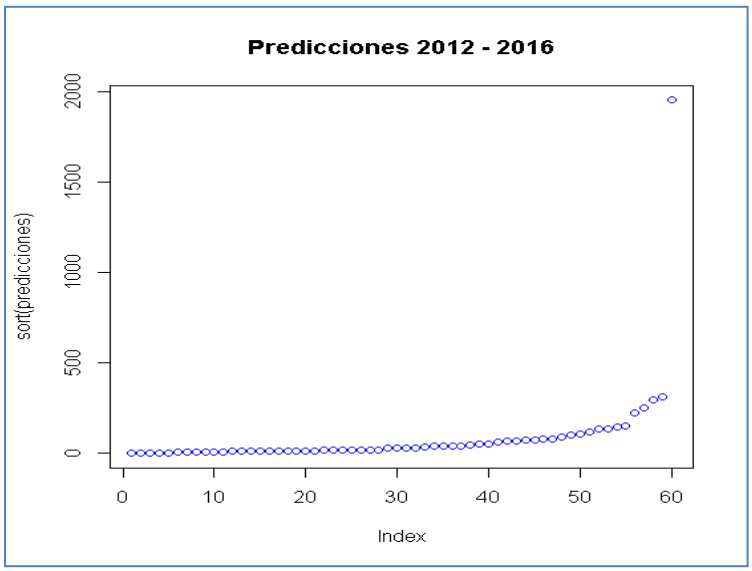
Se observa que el error estándar y los coeficientes de AIC aumentaron lo cual indica que el modelo de Fréchet es mejor, pero la diferencia no es muy significativa, lo que nos da la pauta de usar cualquiera de los dos

Empleando el mejor modelo estimado para la serie de datos máximos en cada mes se tienen las predicciones resumidas en el siguiente Gráfica:



Gráfica 38. Gráfica de dispersión de las predicciones empleando una función generalizada de valores extremos.

Al ordenar los datos de menor a mayor se tiene el siguiente Gráfica, esto se hace para tener una idea más clara de las estimaciones.



Gráfica 39. Gráfica de dispersión de las predicciones Ordenadas de menor a mayor

Como puede observarse, la cantidad de lluvia extrema es muy rara, es decir, que se presenta un mes con una lluvia máxima estimada cercana a los 2000 milímetros cúbicos en cinco años, el resto del tiempo la cantidad de lluvia no superara los cuatrocientos milímetros cúbicos.

Con esto se concluye el análisis de resultados del presente documento, no sin antes aclarar que durante el proceso de análisis empleando la metodología Box – Jenkins se empleo el modelizador experto del SPSS para encontrar el mejor modelo que se ajustara a los datos y en todos los resultados se obtuvieron modelos estacionales de medias móviles estacionales con una diferenciación, es decir, ARIMA $(0, 1, 1)^{12}$, pero se decidió expresar estos modelos en la forma autoregresiva debido a que no nos permitían la comparación clara entre diferentes modelos y porque a la hora de obtener predicciones no son de mucha utilidad porque es un modelo que resulta de la combinación lineal de perturbaciones aleatorias y no emplea en ningún momento los datos disponibles.

9. Conclusiones y recomendaciones

- De conformidad con los resultados del AFM podemos decir que si se desea analizar la cantidad de lluvia acumulada o el comportamiento climático de una zona en especial de nuestro país, no es necesario instalar una nueva estación de monitoreo en ese lugar, solo se debe hacer uso de las que ya están en funcionamiento, ya que se puede hacer estimaciones a partir de la información disponible hasta el momento tomando en consideración la altura y la correlación entre las zonas estudiadas.

- De los resultados obtenidos de las técnicas de series de tiempo empleando metodología Box-Jenkins y de estimación empleando redes neuronales artificiales, se puede concluir que el patrón de las precipitaciones pluviales se ha ido modificando poco a poco en el transcurso del tiempo de manera imperceptible y posiblemente esa alteración se deba en parte al calentamiento global del planeta que se está manifestando en las ultimas décadas.

- Debido al cambio en el comportamiento del ciclo pluvial, se recomienda realizar estimaciones de valores extremos con los datos más actuales, debido a que los datos de décadas anteriores modificarían los modelos estimados y se obtendrían predicciones con mucha variación y muy alejados de la realidad.

- Las metodología Box-Jenkins y los modelos ARIMA así como las redes neuronales artificiales autoregresivas son buenas estimadoras del patrón de comportamiento de series climatológicas y en general de cualquier serie de tiempo que no tenga cotas superiores o inferiores, ya que en el caso de la lluvia, la cota inferior es cero y al realizar estimaciones de los modelos antes mencionados se tienen predicciones negativas, lo cual no es posible tener en la realidad,

- La técnica RNA permite obtener las mejores estimaciones siempre y cuando la variación de los datos no sea muy grande, ya que esa variación hace que las predicciones se ubiquen muy por debajo de la realidad.

- Debido a que los datos de la serie lluvias de diferentes lugares geográficos presentan una elevada correlación entre ellos, podrían emplearse modelos multivariantes para series temporales como una continuación del presente trabajo.

10. Fuentes bibliográficas

- Box, G.E.P., Jemkin, G.M., Reinsel, G.C. (1994); Tome Series Analisys – Forecasting and Control (3rd edition), Prentice Hall.

- Freeman J.A., Skapura D.A.; Neural Networks, Algorithms, Applications and Programming Techniques, Addison-Wesley Publishing Company (1991).

- Harris, R., Sollis, Robert; Applied Time Series Modelling and Forecasting; books in a variety of electronic formats. (2003)

- Kirchgässner, G., Wolters J.; Introduction to Modern Time Series Analysis; Springer-Verlag Berlin Heidelberg (2007).

- Liu, Puyin, Li, Hongxing; Fuzzy Neural Network Theory and Application; World Scientific Publishing Co. Pte. Ltd. (2004).

- Medina R., Rubén; Estimación Estadística de Valores Faltantes en Series Históricas de Lluvias, Tesis de Maestría, Universidad Tecnológica de Pereira, 2008.

- Peña, Daniel (2005); Análisis de Series Temporales, Alianza.

- Rosales, Alejandro I; Análisis Estadístico de Valores Extremos; Tesis de Maestría; Universidad de Granada; España (2011)

- Serrano, Antonio J., Soria, Martín, José E.; Redes Neuronales Artificiales, Universidad de Valencia, Open Course Ware (2009-2010).

Anexos

Descripción del software CHAC

La aplicación CHAC ha sido desarrollada por el Centro de Estudios HidroGráficas del CEDEX con metodologías propias con el fin de proporcionar una herramienta útil para el desarrollo de trabajos hidrológicos dentro del Curso Internacional de Hidrología General y Aplicada del CEDEX.

Se trata de una aplicación desarrollada en Visual Basic para MS WINDOWS, con subrutinas de cálculo en Fortran 77, de fácil manejo a través de una interfaz gráfica.

Esta aplicación es de libre distribución, respondiendo a uno de los fines del CEDEX, como es la transferencia tecnológica a la sociedad.

La aplicación se entrega en un CDROM o bien puede descargarse desde Internet en la dirección http://hercules.cedex.es/Chac. El archivo *zip* contiene otros tres dentro de sí (*setup.exe, setup.lst* y *chac.cab*).

Tras descomprimir, debe ejecutarse *setup.exe* y seguir las instrucciones de instalación. Después de comprobar si hay suficiente espacio en el disco duro, el programa propone la carga de la aplicación en la carpeta "*Archivos de programa**chac*" (recomendada), pudiéndose cargar en otro lugar si el usuario lo prefiere.

Aceptado el directorio de trabajo de la aplicación, se procede a su carga pulsando el botón correspondiente.

Figura 14. Pantalla de inicio del Paquete CHAC.

Para poder trabajar con el CHAC se deben crear los ficheros de datos de acuerdo al formato que se detalla en su manual de usuario, a continuación se muestra una imagen del formato:

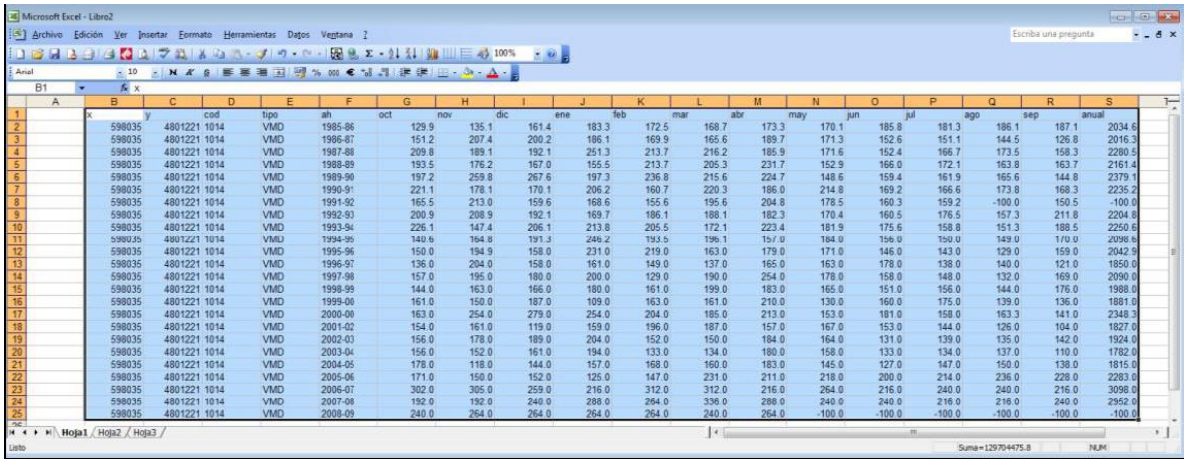


Figura 15. Presentación de los datos en Excel del formato usado por CHAC.

Los encabezados son todos iguales para las distintas variables, sólo va a cambiar el código asignado a cada tipo de variable, a continuación se describe el orden de las columnas:

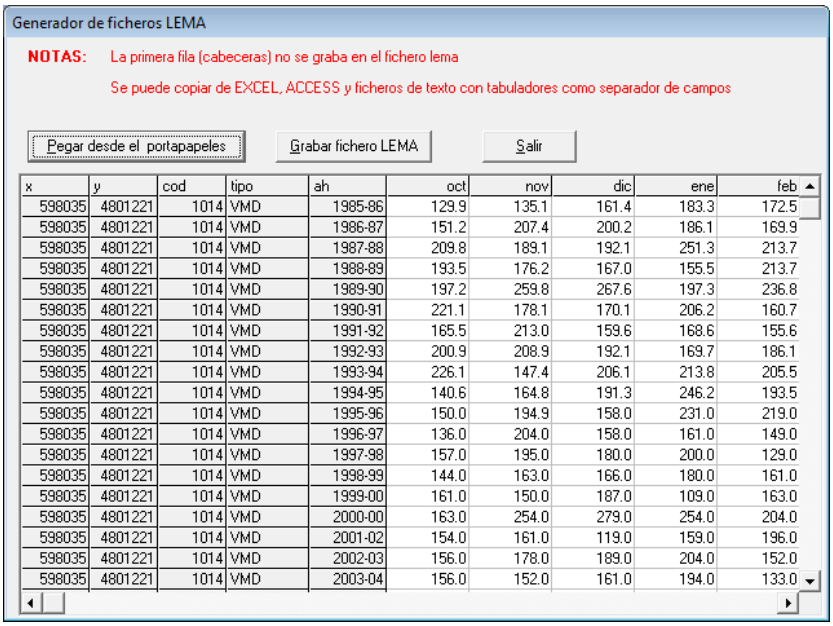
- Las primeras dos columnas contienen las coordenadas x e y en el sistema UTM.
- La tercera columna los códigos asignados a cada estación o variable
- La cuarta columna el tipo de variable
- La quinta contiene el año de inicio y fin por ejemplo 1980 – 81
- Las siguientes 12 columnas contienen los datos para cada mes si es que se trabaja con datos mensuales

Los archivos se pueden crear en un documento .txt o copiarse directamente de Excel, teniendo el cuidado del formato, ya que CHAC es un paquete que diferencia entre un dato numérico y un dato en formato texto.

Cuando ya se tengan los datos, se debe generar el archivo en formato CHAC antes de poder empezar a trabajar en este paquete estadístico.

Si los datos no cumplen con las especificaciones predefinidas, se verán en color rojo y el botón de “grabar fichero” estará desactivado. Es importante saber que en este punto de generación de ficheros CHAC, no se permite modificar los datos que tengan errores, se deben modificar en un editor de texto ASCII o en EXCEL.

Si todos los datos cumplen con las especificaciones del software, entonces ya se puede generar el fichero CHAC y se verá la siguiente imagen con el botón de “grabar fichero” activo.



x	y	cod	tipo	ah	oct	nov	dic	ene	feb
598035	4801221	1014	VMD	1985-86	129.9	135.1	161.4	183.3	172.5
598035	4801221	1014	VMD	1986-87	151.2	207.4	200.2	186.1	169.9
598035	4801221	1014	VMD	1987-88	209.8	189.1	192.1	251.3	213.7
598035	4801221	1014	VMD	1988-89	193.5	176.2	167.0	155.5	213.7
598035	4801221	1014	VMD	1989-90	197.2	259.8	267.6	197.3	236.8
598035	4801221	1014	VMD	1990-91	221.1	178.1	170.1	206.2	160.7
598035	4801221	1014	VMD	1991-92	165.5	213.0	159.6	168.6	155.6
598035	4801221	1014	VMD	1992-93	200.9	208.9	192.1	169.7	186.1
598035	4801221	1014	VMD	1993-94	226.1	147.4	206.1	213.8	205.5
598035	4801221	1014	VMD	1994-95	140.6	164.8	191.3	246.2	193.5
598035	4801221	1014	VMD	1995-96	150.0	194.9	158.0	231.0	219.0
598035	4801221	1014	VMD	1996-97	136.0	204.0	158.0	161.0	149.0
598035	4801221	1014	VMD	1997-98	157.0	195.0	180.0	200.0	129.0
598035	4801221	1014	VMD	1998-99	144.0	163.0	166.0	180.0	161.0
598035	4801221	1014	VMD	1999-00	161.0	150.0	187.0	109.0	163.0
598035	4801221	1014	VMD	2000-00	163.0	254.0	279.0	254.0	204.0
598035	4801221	1014	VMD	2001-02	154.0	161.0	119.0	159.0	196.0
598035	4801221	1014	VMD	2002-03	156.0	178.0	189.0	204.0	152.0
598035	4801221	1014	VMD	2003-04	156.0	152.0	161.0	194.0	133.0

Figura 16. Pantalla de Generación del fichero de datos CHAC.

Mas detalles de su funcionamiento se encuentran en la página de descargar así como en el manual de usuario.

Código en R de la librería “ARNN”.

Nota: el código no es de mi creación, lo anexo porque considero de importancia compartirlo con las personas que tengan acceso a este documento, ya que no se encuentra en el CRAN del proyecto R.

```
#####################################################################
#                                                                   #
#       Autoregressive and multilayer perceptron artificial neural networks         #
#                                                                   #
#####################################################################
.onAttach <- function(...)
{
   version = library(help = arnn)$info[[1]]
   version = version[pmatch("Version",version)]
   um = strsplit(version, " ")[[1]]
   version = um[nchar(um) > 0][2]
   #
   cat(paste("This is arnn package", version, "\n"))
}
#
.onLoad <- function(lib,pkg){ library.dynam("arnn","arnn") }
#
#####################################################################
#
coef.arnn      <- function(object, ...) { object$par       }
fitted.arnn    <- function(object, ...) { object$fitted    }
residuals.arnn <- function(object, ...) { object$residuals }
logLik.arnn    <- function(object, ...)
{ structure( object$loglik,df=length(object$par),class="logLik") }
#
```

```
summary.arnn   <- function(object, ...)
{
   require(forecast)
   print(object)
   cat("\nIn-sample error measures:\n")
   print(accuracy(object))
}
#
print.arnn <-
function (x, ...)
{
   cat(paste("\nMethod:", x$method, "\n\n"))
   cat(paste("Call:\n", deparse(x$call), "\n\n"))
   cat("Parameters:\n\n")
   print(coef(x))
   cat("\n\n")
   cat("Sigma^2 estimated as: ", as.character(x$sigma^2))
   cat("\n\n")
   cat("LogL: ", as.character(x$loglik))
   cat("\n\n")
   cat("Information Criteria:\n")
   print(x$IC)
   cat("\n\n")
}
#-----------------------------------------------------------------------------#
arnn <- function(x, lags = NULL, isMLP = FALSE, H = 1, w.max = 1.0, restarts = 1,
seed = NULL, lambda = 0, model = NULL,  optim.control = list())
{
 #
###########################################################################
 #                                                         #
```

```
#                    Funciones auxiliares                    #
#                                                            #
##############################################################################
#
lagvector <- function(x, lags)
{
        z = embed(x, max(lags)+1)
        return (z[,lags+1])
}
#
#---------------------------------------------------------------------
#
    object2par <- function(object)
{
        if (isMLP == TRUE)
        {
            return(c(  object$M,  c(object$Wih),  c(object$Wbh),
c(object$Who), object$Wbo))
        }
        return(c(  object$M,  c(object$Wio),  c(object$Wih),  c(object$Wbh),
c(object$Who), object$Wbo))
    }
    #
#---------------------------------------------------------------------------
#
    par2object <- function(object, par)
{
        k = 2
  ###
  object$M    = par[1]
  ###
```

```
if (isMLP == FALSE)
{
  object$Wio = matrix( data = par[k:(k + object$nlags - 1)],
                       nrow = object$nlags,
                       ncol = 1)
  k          = k + object$nlags
}
else
{
  object$Wio = matrix( 0, nrow = object$nlags, ncol = 1)
}
###
if( object$H > 0)
{
  ###
                object$Wih = matrix( data = par[k:(k + object$H * object$nlags - 1)],
                  nrow = object$nlags,
                               ncol = object$H)
                k  = k + object$H * object$nlags
  ###
                object$Wbh = matrix( data = par[k:(k+object$H-1)],
                   nrow = object$H,
                   ncol = 1)
                k  = k + object$H
                ###
  object$Who = matrix( data = par[k:(k+object$H-1)],
                   nrow = object$H,
                   ncol = 1)
                k  = k + object$H
}
###
```

```
  object$Wbo = par[k]
  ###
                return(object)
        }
        #
 #-------------------------------------------------------------------------
 #
        fn.foptim = function(w)
 {
                object = par2object(object, w)
                object = arnn(x = object$x, model = object)
                return((1 - object$lambda) * object$sigma ^ 2 + object$lambda * sum(
abs(w[-1])))
        }
        #
 ######################################################################
 #                                                                    #
 #                    Cuerpo de la funcion                            #
 #                                                                    #
  ####################################################################
 #
 if (!exists(".Random.seed", envir = .GlobalEnv, inherits = FALSE)) { runif(1) }
 if (is.null(seed))
 {
  RNGstate <- get(".Random.seed", envir = .GlobalEnv)
 }
 else
 {
  R.seed <- get(".Random.seed", envir = .GlobalEnv)
  set.seed(seed)
  RNGstate <- structure(seed, kind = as.list(RNGkind()))
```

```
  on.exit(assign(".Random.seed", R.seed, envir = .GlobalEnv))
}
#
#
#
        if (is.null(model))
{
              object = list( x = x, lags = lags, nlags = length(lags),
               maxlag = max(lags), call = match.call(),
               method   = "arnn",
               restarts = restarts,
               H        = H,                # número de neuronas en la capa oculta

               lambda   = lambda,           # parámetro de regularización
               numpar   = 1 + 1 + 2 * H + length(lags) * H)
                 #
  if( isMLP == FALSE )
  {
    object$numpar = object$numpar + object$nlags
  }
  #
                object = structure(object, class = "arnn")
                runopt = TRUE
                #
       }
else
{
                object      = model
                object$x    = x
                object$call = match.call()
                runopt      = FALSE
```

```
        }
        #
#-------------------------------------------------------------------------
#
        X.lagged = lagvector(object$x, object$lags)
y = x[(max(object$lags)+1):length(x)]
L = nrow(X.lagged)
        #
        #-------------------------------------------------------------------------
        #
        if (runopt == TRUE)
{
  for( irestart in 1:restarts)
  {
    wopt = optim( par     = runif(object$numpar, min = -w.max, max = w.max),
                  fn      = fn.foptim,
                  method  = "BFGS",
                  control = optim.control )

    u =  fn.foptim(wopt$par)

    if (irestart == 1 || u < u.opt )
    {
      w.gopt = wopt
      u.opt  = u
    }
  }
  object = par2object(object, w.gopt$par)
                #
  #-------------------------------------------------------------------------
  #
```

```
  names(object$M   ) = "M"
  names(object$Wbo ) = "Wbo"
  names(object$Wio ) = paste("Wio[", object$lags, "]", sep = "")
  if(object$H > 0)
  {
    names(object$Wbh ) = paste("Wbh[", 1:object$H,  "]", sep = "")
    names(object$Who ) = paste("Who[", 1:object$H,  "]", sep = "")
    names(object$Wih ) = paste("Wih[",
              matrix( rep(1:object$nlags, object$H),  object$nlags, object$H),
                      ",",
               t(matrix( rep(1:object$H, object$nlags), object$H, object$nlags)),
                      "]", sep = "" )

    #
    object$par = c( object$M, c(object$Wio), c(object$Wih), c(object$Wbh),
                  c(object$Who), object$Wbo)
  }
  else
  {
    object$par = c( object$M, c(object$Wio), object$Wbo)
  }
       }
       #
       #-------------------------------------------------------------------------
       #
 if( object$H > 0)
 {
        f = X.lagged %*% object$Wih + t(matrix(rep(object$Wbh, L), nrow =
object$H, ncol = L))
        f = (1 / (1 + exp(-f))) ^ object$M
        f = f %*% object$Who + object$Wbo + X.lagged %*% object$Wio
```

```
 }
 else
 {
  f = object$Wbo + X.lagged %*% object$Wio
 }
        #
        #---------------------------------------------------------------------------
        #
 response        = y
 residuals       = y - f
 object$sigma    = sqrt(sum(residuals ^ 2) / (length(residuals) - 1))
 if(L >= object$maxlag)
 {
  object$loglik   = -0.5 * (L - object$maxlag) * (log(2 * pi) + log(object$sigma^2)) -
                    0.5 * sum(residuals^2) / object$sigma^2
  p              = 1 + object$nlags
        N              = L - object$maxlag
  AK             = -2 * object$loglik  + 2 * object$numpar
  AKc            = -2 * object$loglik  + 2 * object$numpar * (N / (N - object$numpar -
1))
  HQ                      =  -2  *  object$loglik   +  2  *  object$numpar  *  log(log(L  -
object$maxlag))
  SC             = -2 * object$loglik  +    object$numpar * log(N)
  object$IC       = c(AK, AKc, HQ, SC)
  names(object$IC) = c("AK", "AKc", "HQ", "SC")
 }
 else
 {
  object$loglik = AK = AKc = HQ = SC = NULL
 }
 #
```

```
  #-------------------------------------------------------------------------
  #
  frequency       = attributes(object$x)$tsp[3]
        start           = attributes(object$x)$tsp[1] + object$maxlag / frequency
  object$response  = ts(response,  start = start, frequency = frequency)
  object$fitted    = ts(f,         start = start, frequency = frequency)
  object$residuals = ts(residuals, start = start, frequency = frequency)
  #
        return(object)
}
#-------------------------------------------------------------------------#
simulate.arnn <-
function (object, nsim = 1000, seed = NULL, h = length(object$x),
    bootstrap = FALSE, ...)
{
    #-------------------------------------------------------------------------
    if (!exists(".Random.seed", envir = .GlobalEnv, inherits = FALSE)) { runif(1) }
    if (is.null(seed))
    {
        RNGstate <- get(".Random.seed", envir = .GlobalEnv)
    }
    else
    {
        R.seed <- get(".Random.seed", envir = .GlobalEnv)
        set.seed(seed)
        RNGstate <- structure(seed, kind = as.list(RNGkind()))
        on.exit(assign(".Random.seed", R.seed, envir = .GlobalEnv))
    }
    #-------------------------------------------------------------------------
    if (bootstrap)
    {
```

```
    e = matrix(sample(object$residuals, h * nsim, replace = TRUE),
        nrow = h, ncol = nsim)
}
else
{
    e = matrix(rnorm(h * nsim, 0, object$sigma),
        nrow = h, ncol = nsim)
}
#-------------------------------------------------------------------------
To = length(object$x)
s = matrix(NA, nrow = h, ncol = nsim)
#
      for (iserie in 1:nsim)
{
            x.sim = c(object$x, rep(NULL, times = h))
   #
            for (k in 1:h)
   {
                  X = x.sim[(To + k) - object$lags]
     ###############################################################
                  f = matrix(NA, nrow = 1, ncol = object$H)
                  for(i in 1:object$H) {
                        f[,i] = X %*% object$Wih[,i]+ object$Wbh[i]
                        f[,i] = (1 / (1 + exp(-f[,i])))^object$M
                  }
                  f = f %*% object$Who + object$Wbo + X %*% object$Wio
     ###############################################################
                  x.sim[To + k] = f + e[k, iserie]
            }
            s[, iserie] = x.sim[(To + 1):(To + h)]
      }
```

```
        return(s)
}
#-------------------------------------------------------------------------------#
forecast.arnn <-
function (object, h = 10, level = c(80, 95), fan = FALSE, bootstrap = FALSE,
   seed = 1234, npaths = 5000, ...)
{
   if (fan)
      level <- seq(51, 99, by = 3)
   else
   {
      if (min(level) > 0 & max(level) < 1)
         level <- 100 * level
      else if (min(level) < 0 | max(level) > 99.99)
         stop("Confidence limit out of range")
   }
   nconf = length(level)
   lower = upper = matrix(NA, nrow = h, ncol = nconf)
   m = rep(0, h)
       ###################################################################
   paths = simulate(object = object, npaths = npaths, seed = seed, h = h, bootstrap
= bootstrap)
       ###################################################################
   for (k in 1:h) {
      m[k] = quantile(paths[, k], 0.5)
      lower[k, ] = quantile(paths[, k], (1 - level/100))
      upper[k, ] = quantile(paths[, k], level/100)
   }
   colnames(lower) = colnames(upper) = paste(level, "%", sep = "")
   #
   #
```

```
    #
    result          = list()
    result$model    = object
    result$method   = object$method
    result$level    = level
    result$x        = object$x
    result$residuals = residuals(object)
    result$fitted   = fitted(object)
    #
    frequency       = attributes(object$x)$tsp[3]
    start           = attributes(object$x)$tsp[1] + length(object$x)/frequency
    result$mean     = ts(data = m, start = start, frequency = frequency)
    result$lower    = ts(data = lower, start = start, frequency = frequency)
    result$upper    = ts(data = upper, start = start, frequency = frequency)
    #
    result = structure(result, class = "forecast")
    return(result)
}
#-------------------------------------------------------------------------------#
#
#### First 80 points are used to fit the model
#x <- ts(WWWusage[1:80], s = 1, f = 1)
#
#### A mlp neural network is fitted
#fit <- arnn(x=x, lags=1:4, isMLP=FALSE, H=2, w.max=1e-3,
#   restarts=10, seed = 1234, lambda=0, optim.control=list(maxit=200))
#
#### information about the fitted model
#summary(fit)
#
#### in-sample errors
```

```
#accuracy(fit)
#
#### out-of-sample errors
#fit1 <- arnn(x = WWWusage, model = fit)
#accuracy( fitted(fit1)[76:96], WWWusage[81:100] )
#
#### one-step forecasts plot
#plot(WWWusage)
#lines(fitted(fit1), col = 'red')
#grid()
#
#### multi-step forecast plot
#plot(forecast(fit, h=20, level=90, fan=FALSE, bootstrap=FALSE,
#    seed=1234, npaths=1000))
#grid()
#
####------------------------------------------------------------####
```

Hasta aquí llega el código de la librería, el código necesario para la extracción de los resultados es el siguiente:

```
##  iniciando código de implementación ##
##cargando la base de datos ##
x<- read.csv(file="V3.csv",head=TRUE,sep=",")
serie <- ts(x$V3,start=1971,freq=12)

### separando la muestra de estimación ###
y <- ts(serie,s=1,f=1)
## obteniendo el modelo a partir de los datos ###
fit <- arnn(x=y)
fit
```

```
## pronosticos ##
fit1 <- arnn(x=y,model=fit)
plot(y,lwd=2,col="blue")
lines(fitted(fit1),col="green",lwd=2)
grid()
forecast.arnn(fit1)
plot(fitted(fit1),col="green",lwd=2)

estimado <- ts(fitted(fit1),start=1971,freq=12)
plot(estimado,lwd=2,col="blue")
```