**PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR**

**FACULTAD DE INGENIERÍA**

**INGENIERÍA CIVIL**

**TRABAJO DE DISERTACIÒN PREVIO A LA OBTENCIÓN DEL**

**TÍTULO DE INGENIERO CIVIL**

# DESARROLLO DE UN PROGRAMA EN MATLAB PARA LA GENERACIÓN DE SERIES SINTÉTICAS DE HIDROGRAMAS PARA PROYECTOS HIDROELÉCTRICOS. APLICACIÓN AL CASO DEL PROYECTO HIDROELÉCTRICO TOPO

**AUTOR:**

CAICEDO PÉREZ DARWIN GERMÁN

**DIRECTOR:**

CARLOS LUIS NAVAS

**Quito, febrero del 2020**

**Agradecimientos**

A mi familia y amigos por haberme apoyado de alguna u otra manera para lograr mis objetivos.

A los profesores a lo largo de la carrera por enseñarme las técnicas, herramientas y prácticas de la profesión.

## Índice de figuras

## Índice de tablas

# 1. INTRODUCCIÓN

## 1.1. Antecedentes

El modelo para la generación de curvas sintéticas, desarrollado por Thomas y Fiering en 1962, es el primer modelo para la predicción de caudales en ríos que ha sido ampliamente aceptado y validado para el planeamiento en la construcción y operación de obras hidráulicas. Esto se debe a que logra conservar la estadística de los registros históricos a la vez que añade un componente de incertidumbre, lo cual es algo inherente en los procesos hidrológicos. Este es un modelo univariable que utiliza las observaciones de un solo río a través del tiempo para inferir sus características en el futuro sin tomar en cuenta otras variables como humedad, temperatura y precipitaciones.

En 1967, Matalas propuso un modelo que fue desarrollado posteriormente por Young y Pisano en 1968, y que, partiendo de la similitud o correlación entre los caudales de diferentes ríos en una cuenca, utiliza ecuaciones matriciales para generar múltiples curvas sintéticas de una manera correspondiente a la ecuación de Thomas y Fiering.

La familia de modelos de medias condicionales, Box-Jenkins, (ARIMA) también es utilizada con frecuencia en hidrología, a pesar de que fue desarrollada para distintas ramas de la ciencia y en especial para la econometría, ha producido resultados satisfactorios en el pronóstico de los flujos de ríos en numerosos estudios desde su aparición en 1976, sus fundamentos, además, han servido para la generación de otra clase de modelos más recientes.

A parte de estas aproximaciones, modelos más complejos, que incluso logran modelar a la par, otras características de la cuenca, como la temperatura, la humedad, la velocidad del viento y la precipitación, y otros algoritmos más desarrollados, como los árboles de decisiones y las redes de neuronas artificiales, también son utilizados en la actualidad, y comprenden la amplia gama de modelos disponibles, la elección del modelo más adecuado de acuerdo al proyecto es difícil, particularmente porque cada investigador que crea y desarrolla estos modelos tiende a promover sus propios méritos y a mantener indiferencia o desacreditar otros puntos de vista, por tanto, lo factible es que cada modelo en específico debe ser analizar la factibilidad de aplicar alguno de estos modelos de manera individual o al menos como un estudio regional.

## 1.2. Justificación

La producción de energía en centrales hidroeléctricas concuerda con las políticas del país de impulsar un desarrollo económico, social y ambiental sostenible, en cuanto a que crea numerosas plazas de trabajo y es menos nociva con el medio ambiente en comparación a otras alternativas de generación.

Actualmente, alrededor de un 76% de la energía eléctrica producida en el Ecuador viene de una fuente hidráulica (ARCONEL, 2019) y con la implementación del Plan Maestro de Electrificación vigente se espera crezca hasta un 90%. (PME 2016-2025)

Las centrales hidroeléctricas se dimensionan en base a registros que posibilitan una estimación de la capacidad de la fuente de energía, en este caso, del río. Las decisiones tomadas durante la fase de operación se basan en datos proyectados, los cuales tienen un grado inherente de incertidumbre (Medina, 2003). Las series sintéticas son prácticas que reducen esta incertidumbre y que buscan generar una situación virtual, que, en lo posible, sea estadísticamente indistinguible de la situación real.

El estudio de los procesos que experimenta el agua, como es su circulación en los ríos, es competencia de la hidrología y de la misma manera, las series sintéticas y otras aproximaciones matemáticas que buscan pronosticar esos procesos, son parte de la rama de la hidrología conocida como modelación hidrológica. (Quispe, 2017)

Los modelos Thomas-Fiering y ARIMA son los utilizados en el presente documento para estimar los caudales del Río Topo, que aporta el caudal turbinable al proyecto hidroeléctrico Topo. Para estimar la bondad que los modelos tienen para ajustarse a las futuras observaciones, se usan los registros de los caudales del 2011 al 2018, datos proporcionados por la empresa ECUAGESA. Los caudales de los primeros 7 años se utilizan para calcular los parámetros de los modelos y con ello se los proyecta para el 2018 para distintos intervalos de tiempo y con ello se concluye si la capacidad predictiva de estas aproximaciones es lo suficientemente buena.

Los cálculos, en su totalidad, se realizan en el software MATLAB, una aplicación con orientación educativa y profesional, que permite el manejo de grandes cantidades de datos, y que tiene una interfaz gráfica para su mejor visualización, además, se hablan de otras aplicaciones que posee la aplicación para el manejo de series en el tiempo.

## 1.3. Objetivos

### 1.3.1. Objetivo general

Generar series sintéticas de Hidrogramas con base en registros del río Topo, que es aprovechado por la Central Hidroeléctrica Topo.

### 1.3.2. Objetivos específicos

- Ilustrar los principios en los que se basa la estimación de los procesos hidrológicos y relacionar estos procesos con el manejo de obras hidráulicas, haciendo énfasis en las centrales hidroeléctricas.
- Crear una serie sintética basada en el modelo Thomas-Fiering en el entorno MATLAB, para el pronóstico de caudales en base a registros diarios, mensuales o anuales.
- Analizar los resultados generados por la serie y compararlos con los registros mensuales de los flujos que llegan a la Central Hidroeléctrica Topo para comprobar su aplicabilidad en el caso.

## 1.4. Alcance

El presente trabajo presenta algunos conceptos hidrológicos y de centrales hidroeléctricas, así como las bases sobre las que se cimienta la modelación hidrológica univariable, profundizando en los mecanismos de los modelos ARIMA y Thomas-Fiering para su posterior implementación y validación en el registro de 8 años de los caudales de un río que sustenta a la central hidroeléctrica Topo, en la provincia de Tungurahua.

# 2. MARCO TEÓRICO

## 2.1. Nociones de probabilidad y estadística

### 2.1.1. Conceptos de Probabilidad

#### *2.1.1.1. Aleatoriedad*

Aquello que se define como aleatorio es algo que resulta demasiado complejo o del cual no se conocen muy bien los mecanismos bajo los que se fundamenta, por lo que es simplemente una limitación del conocimiento. Para el caso del lanzamiento de una moneda, la capacidad humana, por si sola, no permite predecir el resultado y, por lo tanto, es algo que conocemos como aleatorio, pero suponiendo se tiene una gran cantidad de información acerca de las variables que conllevan a un acto, puede ser posible estimar con bastante certeza su desenlace.

Una secuencia numérica es estadísticamente aleatoria cuando no contiene patrones reconocibles o regularidades, es decir algo como los resultados de un lanzamiento de dados o los dígitos de un número irracional. Esta propiedad no tiene que venir necesariamente de un proceso aleatorio o impredecible. (Purdue University, 2005)

#### *2.1.1.2. Espacio Muestral*

Se denomina con este nombre a los posibles resultados de un experimento, y generalmente es denotado por la letra griega omega(Ω). Este puede estar compuesto por números, palabras, letras o símbolos.

Dependiendo de la naturaleza de las *variables aleatorias* que contiene, este puede ser finito o infinito, aquellos subconjuntos de un espacio muestral se conocen como eventos aleatorios.

#### *2.1.1.3. Variables Aleatorias*

Al contrario que las variables trabajadas en álgebra, las variables aleatorias sirven para cuantificar todos los posibles valores de un experimento o fenómeno contable. Se llaman aleatorias debido a que su magnitud no es fija, sino que puede tomar algunos valores dentro

de un rango, la frecuencia con la que toma distintos valores está definida por una distribución de probabilidad. Un ejemplo de variable algebraica es la altura de una persona, mientras que, una variable aleatoria, es la altura de todos los adultos en el país, cantidad que no puede ser expresada como un solo valor sino como un conjunto de valores.

A partir de una misma muestra se pueden medir distintas variables aleatorias, por ejemplo, de un conjunto de empleados podrían encontrarse distintas características como su gasto mensual en comida, en vivienda, en transporte y en intereses, su nivel de satisfacción en el trabajo en una escala cuantitativa, la edad, el sueldo, entre otros.

Las variables aleatorias cuando se miden en números enteros se conocen como variables aleatorias discretas (el grado de educación de una persona, el número de estudiantes presentes en un aula), mientras que cuando pueden tomar cualquier cantidad sobre un intervalo específico se denominan variables aleatorias continuas (el tiempo necesario en llegar a la universidad, la distancia de un lanzamiento de jabalina), también puede existir el caso de que las variables se deban expresar como una mezcla de ambas, a través de variables aleatorias mixtas (el tiempo que espera una persona para ser atendida por un cajero, este puede tomar un valor discreto de 0 con cierta probabilidad, o un valor continuo, si hay gente que está siendo atendida).

Las mediciones de algunos fenómenos, por ejemplo, los caudales, obligan a que estas cantidades, las cuales son variables continuas, sean trabajadas como variables discretas, debido a que los instrumentos tienen una precisión limitada y debe asignar resultados similares en los mismos grupos, la magnitud de un caudal con sus 4 primeros decimales .6788 y otro con .6841, el aparato guarda a ambos como .68, suponiendo que este tiene una precisión de 2 decimales.

Dentro de la hidrología estocástica, debido a que se combina una parte determinística, que es fija o certera, con una componente incierta que fluctúa entre un rango, o en otras palabras, una variable aleatoria, tiene ecuaciones que tendrán como resultados variables aleatorias.

Otro concepto importante es el del experimento aleatorio, y se define como la observación de un proceso en el cual se observa algo incierto, es decir, una variable aleatoria, ejemplos de esto pueden ser, el número de celulares vendidos por una compañía en un mes, el resultado del lanzamiento de un dado o de una moneda.

#### *2.1.1.4. Probabilidad*

Es la proporción de la ocurrencia de todos los resultados en un experimento o situación observada en la vida real. Para asignar una probabilidad a estos eventos, estos deben ser en conjunto:

- Mutuamente excluyentes: solo un resultado sucede a la vez
- Colectivamente exhaustivos: el conjunto contempla todos los posibles resultados en el experimento.

En estadística, la lista de los posibles resultados de un experimento se conoce como espacio muestral. La suma de las posibilidades de todos los posibles subeventos que conlleven a una realización debe ser igual a 1 o 100%, las probabilidades 0 y 1 corresponderían a la imposibilidad y certeza total de un evento respectivamente.

- *Probabilidad condicional*

Dentro de un espacio muestral, la probabilidad condicional $P(A \mid B)$, es la probabilidad de que un evento A ocurra, dado que B ocurrió. Se supone entonces, que para el cálculo de estas posibilidades existe un nuevo universo o espacio muestral donde las condiciones cambian debido a que B ocurrió.

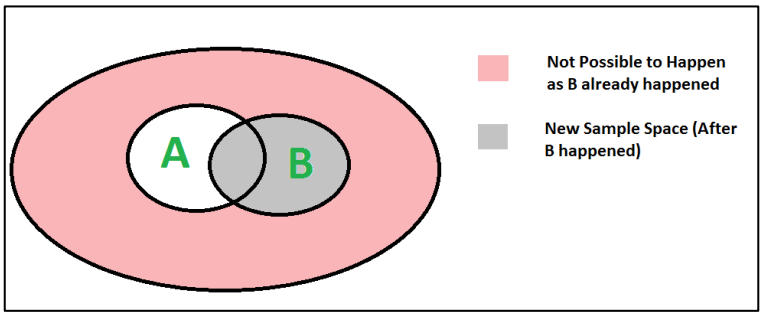


Figura 2.1-1. Ilustración a través de conjuntos de las probabilidades condicionales.

- *Probabilidad conjunta*

Para el caso de un par de variables aleatorias, es la probabilidad de que ambos eventos se presenten uno a continuación del otro. Puede también expresarse como la probabilidad de que la intersección de los conjuntos de los eventos de ambas variables suceda. Siendo "X" y "Y" las variables aleatorias, y la ocurrencia de dos eventos posibles "x" y "y" para cada una, la probabilidad conjunta de que ambos sucedan se denota como $P(X = x, Y = y)$

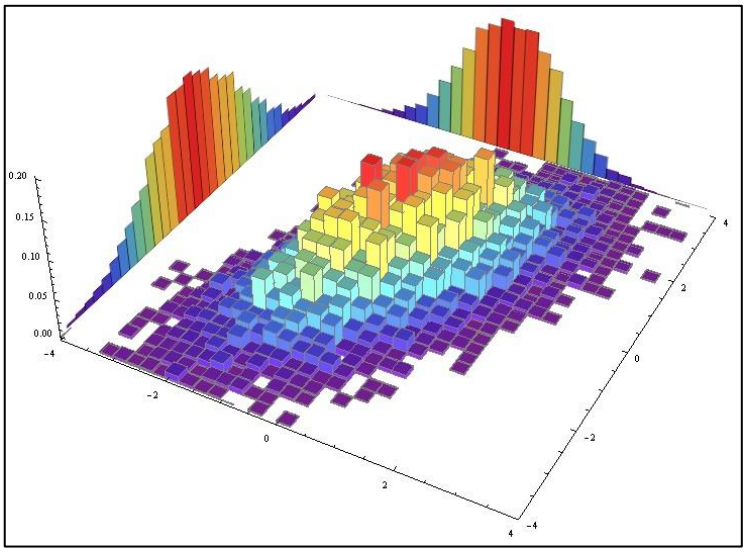


Figura 2.1-2. Distribución de probabilidad conjunta de dos variables. Copyright 2018 por Stack Overflow. *Reimpreso con permiso*

- *Funciones de probabilidad*

Debido a que existirán varios puntos en un espacio muestral, para fines estadísticos, se puede expresar la probabilidad de encontrarse sobre cierto valor, dependiendo de la naturaleza de la variable, a través de *funciones de probabilidad*. Ciertas consideraciones a tomar serán:

- La suma de las probabilidades debe ser 1 o 100%.
- No pueden existir probabilidades negativas.

- *Función de masa de probabilidad (PMF)*

Es una relación que expresa la probabilidad de que una variable aleatoria discreta sea igual a cierto valor $P(X = x) = f_X(x)$.

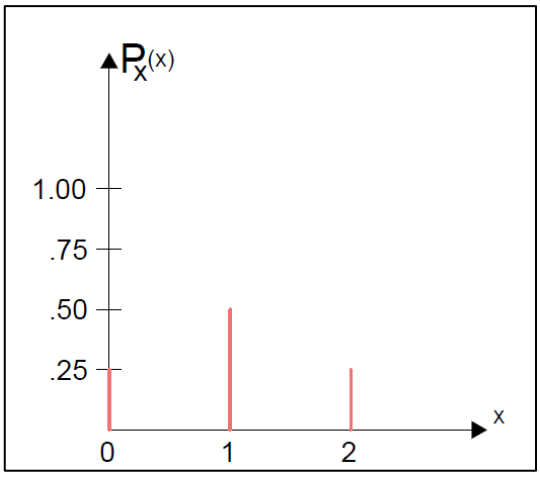


*Figura 2.1-3. PMF de la obtención de caras en dos lanzamientos de monedas.*

- *Función de densidad de probabilidad (PDF)*

Es una relación que expresa la probabilidad de que una variable aleatoria continua se encuentre dentro de un cierto rango de valores, para encontrar dicha probabilidad, es necesario integrar la función dentro de este rango ya que cada ordenada de esta función no expresa propiamente probabilidades absolutas sino relativas a los demás puntos de la gráfica. $P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$.

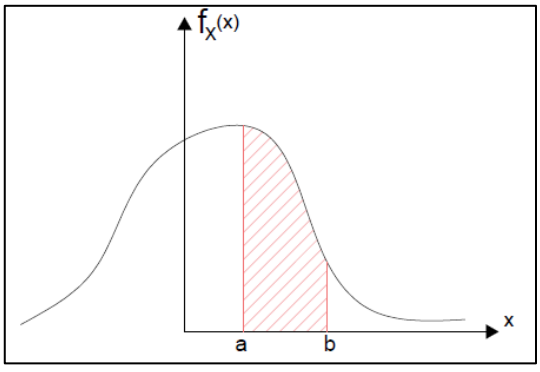


*Figura 2.1-4. PDF y rango de integración arbitrario de una variable aleatoria continua.*

- *Función de probabilidad acumulada (CDF)*

Es una relación que expresa la probabilidad de que una variable aleatoria sea menor a un cierto valor sobre el que se evalúa la función. De esta manera, esta función es capaz de relacionar variables que en ciertos casos toman valores discretos y en otros continuos. La función de masa de probabilidad y la integral de la función de densidad representan la derivada de esta función. $f_X(x) = P(X \leq x)$.

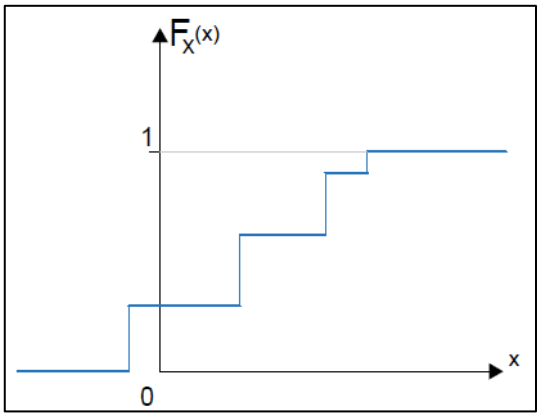


*Figura 2.1-5. CDF de una variable aleatoria discreta.*

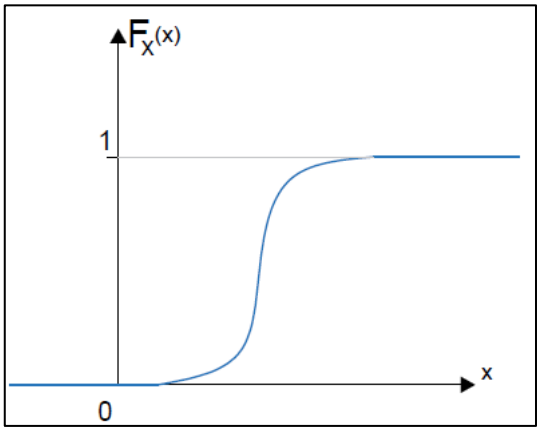


*Figura 2.1-6. CDF de una variable aleatoria continua. (Distribución normal)*

### *2.1.1.5. Método de la transformada inversa*

Las computadoras son instrumentos determinísticos que actúan de acuerdo a las instrucciones con las que fueron fabricadas y por lo tanto no pueden generar por sí solas números aleatorios, los números pseudoaleatorios siguen un conjunto de reglas que los hace

lucir aleatorios ya que no siguen patrones que permita pronosticarlos. El estándar en este tipo de generadores y el que MATLAB utiliza, a menos de que se le indique lo contrario, es *Mersenne Twister* **y** está basado en el Generador lineal congruencial, el cual tiene 4 datos de entrada:

- El módulo: m
- El multiplicador: a
- El incremento: c
- La semilla: el valor inicial de la secuencia

La fórmula para definir los datos de la secuencia es:

$$X_{n+1} = (a * X_n + c) mod\ m$$

Donde "mod" implica que el siguiente dato de la serie es el residuo de la división de lo que se encuentra dentro del paréntesis y "m", el divisor. Cada número que se genera $(X_{n+1})$ es una función de un número producido anteriormente por el generador $(X_n)$.

Para mejorar la calidad de este proceso, es posible buscar el valor de la semilla a partir de mediciones de fenómenos que son muy difícilmente predecibles:

- Ruido atmosférico
- Radiactividad
- Electricidad estática en la atmosfera
- Tomando un libro de números aleatorios y escogiendo uno arbitrariamente
- La temperatura en el Monte Everest o en el Sahara dividido para la fecha actual en segundos.

Debido a que se basan en una fórmula, estos algoritmos eventualmente vuelven a generar la misma secuencia de números después de un tiempo suficiente, la ventaja de usar *Mersenne Twister* es que el periodo con el que las iteraciones se vuelven a repetir es aproximadamente "$2^{19937}$-1" números. Se muestran los algoritmos con los que MATLAB puede producir números aleatorios:

| Keyword | Generator |
|---|---|
| mt19937ar | Mersenne twister (used by default stream at MATLAB® startup) |
| dsfmt19937 | SIMD-oriented fast Mersenne twister |
| mcg16807 | Multiplicative congruential generator |
| mlfg6331_64 | Multiplicative lagged Fibonacci generator |
| mrg32k3a | Combined multiple recursive generator |
| philox4x32_10 | Philox 4x32 generator with 10 rounds |
| threefry4x64_20 | Threefry 4x64 generator with 20 rounds |
| shr3cong | Shift-register generator summed with linear congruential generator |
| swb2712 | Modified subtract with borrow generator |

*Figura 2.1-7. Algoritmos de generación de números aleatorios utilizados por MATLAB. Copyright 2019 por MathWorks. Reimpreso con permiso.*

Los números aleatorios, generados por cualquier algoritmo de computadora, se encontrarán dentro de un rango definido y cada uno de ellos tendrá igual probabilidad que el otro. Es decir, si existe "n" enteros que el algoritmo es capaz de generar, entonces la probabilidad para cada número "i" de que suceda es:

$$P(X_i) = \frac{1}{n}$$

Para satisfacer la propiedad de que la suma de todas las probabilidades sea igual a la unidad. Debido a esto, los números generados tienen una distribución uniforme, sin embargo, en la generación de modelos estocásticos, suele ser necesario generar números aleatorios con otro tipo de distribuciones.

Para ello existe el Método de la transformada inversa, que trabaja con la CDF perteneciente a la distribución de interés, para ello, primero es necesario generar la cantidad requerida de números distribuidos uniformemente y dividirlos para su máximo, de modo que tengamos números entre 0 y 1.

Cada uno de los números generados, se coloca en el eje de las "y" de la CDF de la distribución requerida, se halla el valor en "x" correspondiente a la función al intersecar la curva y con ello es posible generar números que sigan la distribución necesaria.

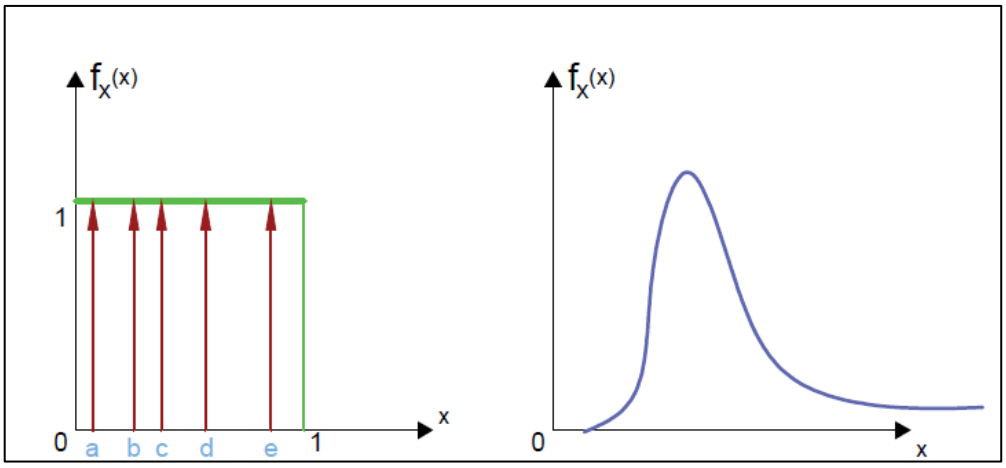


*Figura 2.1-8. Generación de números aleatorios en una distribución uniforme y PDF de la distribución objetivo.*

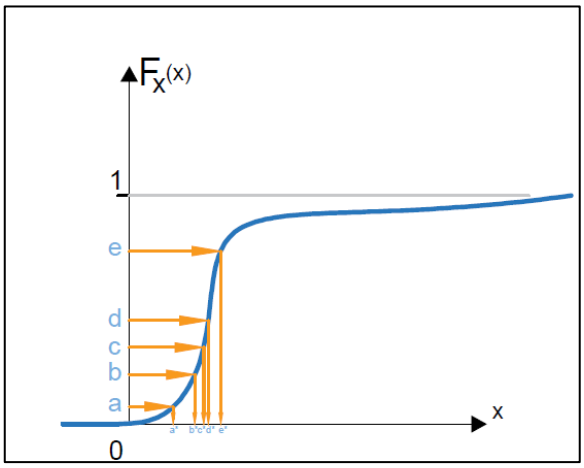


*Figura 2.1-9. Obtención de los valores dentro de la distribución requerida*

### *2.1.1.6. Distribución Normal*

Es una función de densidad de probabilidad definida por la ecuación:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Una curva en forma de campana definida por los parámetros:

$\mu$: *la media de la distribución*

$\sigma$: *la desviación estándar*

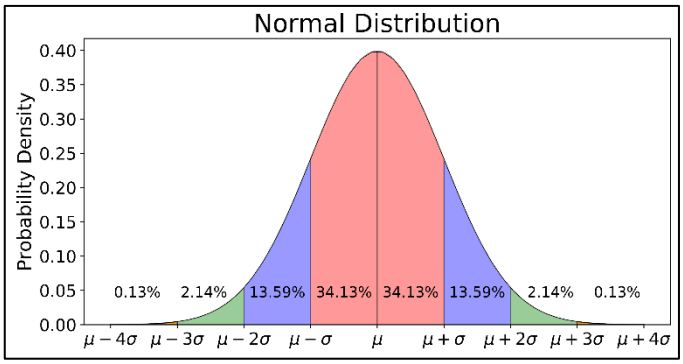


Figura 2.1-10. Distribución normal. Copyright 2018 por Towards Data Science. *Reimpreso con permiso*

Se muestra en esta figura el porcentaje de los datos que se encuentran en las áreas delimitadas horizontalmente por la media y 1,2 y 3 distribuciones estándar respectivamente.

La distribución normal tiene gran importancia en la estadística debido al Teorema del Límite Central que establece que, en un experimento donde cada ensayo es independiente el uno del otro, si "n" es el número de ensayos y "n" es un número lo suficientemente grande, al dibujar el histograma de las pruebas que resultaron en éxito, su distribución tenderá a la normalidad.

Utilizando el software MATLAB, se simulan lanzamientos de monedas con probabilidades de caer en cara de 25%, 50% y 75 %, se grafica el histograma de este proceso realizado una cierta cantidad de veces.

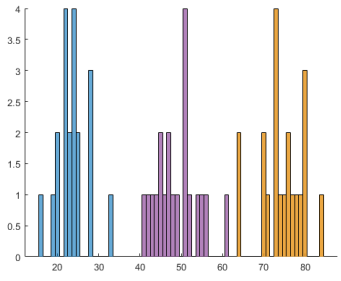

*Figura 2.1-11. Demostración del teorema del límite central con 20 lanzamientos de monedas.*

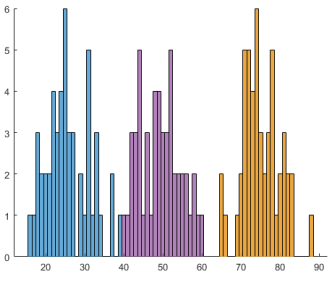

*Figura 2.1-12. Demostración del teorema del límite central con 50 lanzamientos de monedas.*

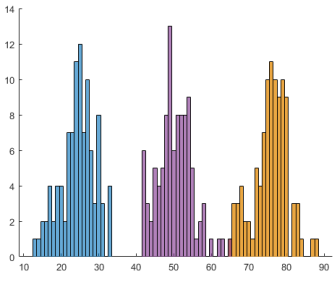

*Figura 2.1-13. Demostración del teorema del límite central con 100 lanzamientos de monedas.*

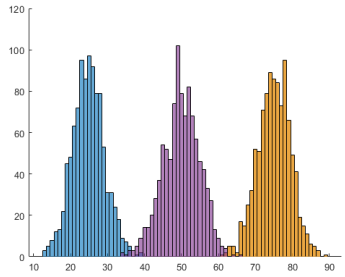


*Figura 2.1-14. Demostración del teorema del límite central con 1000 lanzamientos de monedas.*

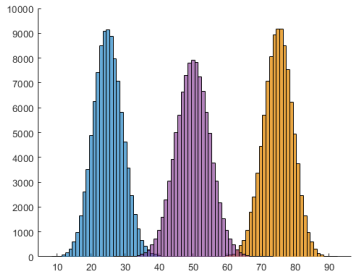


*Figura 2.1-15. Demostración del teorema del límite central con 100000 lanzamientos de monedas.*

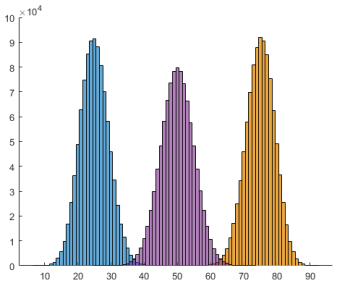


*Figura 2.1-16. Demostración del teorema del límite central con 1000000 lanzamientos de monedas.*

Para el caso más general, para una suma de variables dependientes y aleatorias:

$$S_n = X_1 + X_2 + \cdots X_n$$

Debido a la linealidad de la varianza y la esperanza:

$$var(S_n) = var(X_1) + var(X_2) + \cdots var(X_n)$$

$$E[S_n] = E[X_1] + E[X_2] + \cdots E[X_n]$$

La suma de las variables aleatorias estandarizada tenderá a aproximarse a la función normal estándar, aquella que tiene varianza y desviación estándar igual a 1 y media 0.

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{var(S_n)}} \sim N(0,1)$$

Lo cual es aplicable, independientemente de las distribuciones de probabilidad de las variables $X_i$.

Para el caso de los caudales, debido a que existe una alta dependencia de un momento al otro, la suma de sus distribuciones a lo largo del tiempo no colapsa sobre una forma de campana. Por otra parte, los residuos de los modelos, al ser agrupados en un intervalo de tiempo lo suficientemente largo, se deberían asemejar a una distribución normal, debido a que se suponen independientes y distribuidos idénticamente, como se explica posteriormente en la teoría de las regresiones.

#### *2.1.1.7. Cadenas de Markov*

Es un proceso utilizado para describir una secuencia en la que cada estado:

$$estado\ nuevo: f(estado\ viejo, aleatoriedad)$$

Es una función del valor precedente de ese experimento y una componente aleatoria relacionada con dicho estado. Funciona para expresar la relación entre variables discretas y continuas. Los principios de esta teoría son utilizados en la meteorología, la hidrología, la economía y la teoría de colas o las líneas de espera en un sistema.

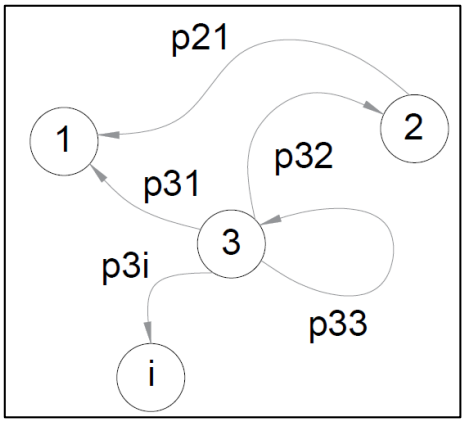


*Figura 2.1-17. Cadena de Markov de 3 estados y sus probabilidades de transición*

En este ámbito, espacio estado, se conoce como el conjunto de todos los estados que una variable que sigue este proceso puede tomar. Las probabilidades de transición, por su parte, son las probabilidades que tienen cada estado de saltar a otro, tomando en cuenta que la suma de cada una de ellas, para cada estado, debe sumar 100% de probabilidad. Para el caso de variables discretas, estas pueden ser mejor visualizadas a través de un diagrama o una matriz de transición, en los a que cada columna le corresponde un estado y a cada fila, sus posibles transiciones. Un ejemplo de matriz de transición para 3 diferentes estados se presenta:

$$T = \begin{matrix} \\ s1 \\ s2 \\ s3 \end{matrix} \begin{matrix} \begin{matrix} s1 & s2 & s3 \end{matrix} \\ \begin{bmatrix} 0{,}5 & 0{,}1 & 0{,}8 \\ 0{,}2 & 0{,}3 & 0{,}2 \\ 0{,}3 & 0{,}6 & 0 \end{bmatrix} \end{matrix}$$

En un proceso que se lleva en intervalos de tiempo discretos, generalmente la variable independiente en un tiempo "n" $X_n$ , dependerá de valores que tomó la variable

anteriormente $X_{n-1}, X_{n-2}\ldots$, por lo tanto, ocurridos estos eventos, la probabilidad condicional de que $X_n$ tome un valor específico viene dado por:

$$P(X_n \mid X_{n-1}, X_{n-2} \ldots\ldots X_1)$$

Una propiedad que tienen en común las cadenas de Markov, y por lo tanto todo modelo hidrológico basado en ellas, es que, en su forma más básica, se cumple:

$$P(X_n \mid X_{n-1}, X_{n-2} \ldots\ldots X_1) = P(X_n \mid X_{n-1})$$

Es decir, el valor que tome la variable aleatoria en cualquier tiempo solamente dependerá del valor que tomó un instante antes, esto se conoce como la propiedad de Markov.

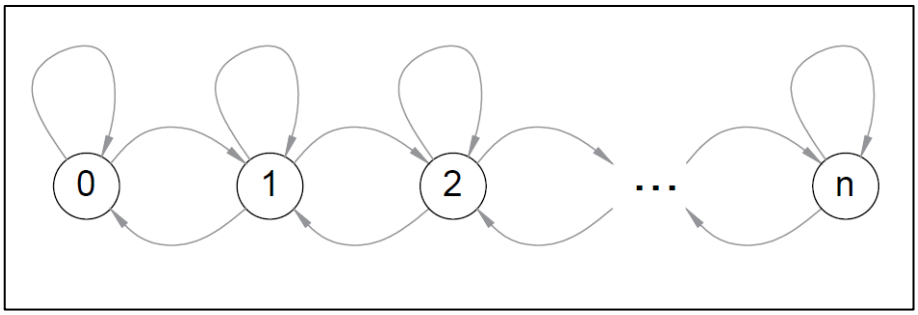


*Figura 2.1-18. Proceso de nacimiento-muerte , cadena de Markov donde todos los estados se encuentran comunicados con su estado inmediatamente anterior y posterior, con excepción de su estado inicial y final que se comunican con un estado más*

Un ejemplo de modelación a través de cadenas de Markov es el algoritmo PageRank para la presentación de resultados de Google. Las probabilidades de transición, hacia páginas con resultados similares, se determinan mediante el número de vínculos que dirigen hacia cada una de estas páginas o estados, siendo ésta, proporcional a la importancia relativa de cada página.

Ejemplos de modelos hidrológicos que pueden ser modelados a partir de cadenas de Markov pueden ser:

- Proporción de días que llueven y días que no llueven y las veces que cada uno de estos estados pasa hacia el otro o se mantiene.
- En reservorios, al tener un 0 al 30% de su capacidad, un 30-60%, un 60-100% o exceder su capacidad y sus transiciones.
- En caudales, aquellos que se encuentra en ciertos rangos pueden ser agrupados a través de estados (ej: 100-150$m^3$/s o 25-30$m^3$/s).

- En general, para el estudio de las series en el tiempo en la hidrología y otras ramas de la ciencia, las series markovianas se refieren a aquellos procesos en los que su probabilidad de transición depende únicamente del estado anterior en el que se encontraban.

### 2.1.2. Conceptos de estadística

#### *2.1.2.1. Población y muestra*

La población, en estadística, es un conjunto de objetos o eventos de interés de los que se desea extraer información, en general se refiere a todos los elementos que forman un conjunto, como pueden ser:

- La población del Ecuador
- Las notas de los estudiantes de una región para el ingreso a la universidad
- Las personas que votan en las elecciones
- Los caudales en un río

Las mediciones o variables aleatorias que podamos obtener de una población suelen resultar demasiado numerosas para ser realizadas en su totalidad, por lo que deben limitarse a una muestra o subconjunto que sea más manejable por el instrumento que realiza la medida.

Las conclusiones que se realicen de una población en función de una muestra deben tener en consideración que a aquello que se infiera a partir de una población se lo conoce como parámetro y a lo relacionado a una muestra, como estadístico, que solo es una aproximación al parámetro de la población que contiene a dicha muestra.

Por ejemplo, en la varianza, para un cálculo más realista y aproximado, se utiliza en su denominador un número más pequeño en el cálculo de su estadístico que para el cálculo de su parámetro.

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$var(X) = \frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2$$

Al valor $S_{n-1}^2$ se lo conoce como el estimador insesgado de la varianza, además de esto, existen ajustes en el cálculo de la covarianza y otros parámetros, realizados con el fin de obtener valores más cercanos a las características de la población, en función de muestras más pequeñas.

En las series en el tiempo, los parámetros de un mes, por ejemplo, deben calcularse como estadísticos de una muestra, tomando en cuenta que los registros no pueden contener las medidas de una variable para el mismo mes en todos los años precedentes, especialmente cuando la longitud de estos registros es corta.

#### *2.1.2.2. Esperanza matemática*

La esperanza es un promedio de los posibles valores que una variable aleatoria podría tomar suponiendo que se realiza un gran número de observaciones. También es posible definirla como el promedio ponderado del producto de los posibles resultados de un experimento por sus probabilidades.

$$E[X] = \sum_{i=1}^{N} x_i * P(X = x_i)$$

Para estimar cual es este valor, conocidos los resultados de una muestra, se calcula el estadístico conocido como la media:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n}$$

La media, junto a la mediana y la varianza, comparten el concepto de medidas de tendencia central y sirven para establecer un valor central o típico de un conjunto de datos.

#### *2.1.2.3. Varianza*

Es una medida de dispersión de una variable aleatoria que hace énfasis en valores demasiado alejados de la media. Es igual al valor promedio de la diferencia al cuadrado de las observaciones con respecto a su media.

$$var(x) = E[(X - E[X])^2]$$

Su valor se estima, a partir de los datos de la muestra como:

$$var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Debido a que esta medida tiene como unidades el cuadrado de las unidades de la variable, (si la medición se encuentra en metros, la varianza se encontraría en metros al cuadrado, lo cual es una medida no tan intuitiva) es común expresar la dispersión de los datos como la raíz cuadrada de la varianza o la desviación estándar $\sigma$, la cual si comparte las mismas unidades de la variable medida.

#### *2.1.2.4. Covarianza*

Es una medida de la dependencia de dos variables expresada como el producto de la desviación esperada de ambas variables con respecto a sus medias respectivas.

$$cov(X, Y) = E[(X - E[X]) * (Y - E[Y])]$$

Para una muestra de valores se calcula como:

$$cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})$$

La covarianza, al igual que la varianza no tiene unidades muy interpretables, ya que calcula valores que tienen las unidades de la variable aleatoria al cuadrado, el coeficiente de correlación de Pearson, por otra parte, estandariza esta relación y permite tener una cantidad adimensional.

$$\rho(X, Y) = E\left[\frac{E[(X - E[X]) * (Y - E[Y])]}{\sigma_x \sigma_y}\right]$$

Es posible calcular este valor a partir de un conjunto de observaciones como:

$$\rho(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right) * \left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

La ventaja de este valor es que es adimensional y que no depende de la magnitud de las variables, un valor de -1 y 1 implican una relación lineal negativa y positiva perfecta entre las variables respectivamente, 0 significa que existe independencia entre ellas.

## 2.2. Modelación matemática

### 2.2.1. Generalidades

Un modelo es una equivalencia de un fenómeno real, aunque no idéntica, en cuanto a propiedades, utilizado para describir un problema y que puede ser tan complejo o puede omitir tantos detalles irrelevantes como sea necesario. (World Meteorological Organization,Commission for Hydrology (CHy)., 2009)

Su propósito es el de predecir como los sistemas se comportarán bajo ciertas condiciones, en ciertos casos estos pueden ser puestos a prueba mediante experimentos para comprobar que lo supuesto por el modelo es veraz. En el caso de los modelos hidrológicos, los registros futuros son los que determinarán la certeza de las predicciones del modelo.

Todo modelo es una aproximación de la realidad y por lo tanto no es exacto, sin embargo, la búsqueda y comprobación de diferentes alternativas de modelación debe ser una tarea de los profesionales encargados de los proyectos hidráulicos, debido a que su implementación puede traer beneficios para la operación de la obra.

### 2.2.2. Regresiones

#### *2.2.2.1. Definición*

Es una herramienta que permite modelar la relación matemática entre dos variables, al dejar expresada una de las variables como una función de la otra. Una de las formas más básicas de la regresión es la regresión lineal, de forma:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Donde $Y$ se la conoce como la variable dependiente o de respuesta, y a $X$ como la variable independiente o explicativa. Aquella función mostrada es correspondiente a una función lineal, teniendo en cuenta que nunca será posible crear una relación perfecta entre las variables, existirán residuos o componentes impredecibles denotados por $\epsilon$.

La aproximación más cercana teórica de una variable tiene la misma forma, pero sin esa componente de residuos, y se denota por $\hat{Y}$:

$$\hat{Y} = \beta_0 + \beta_1 X$$

El propósito de introducir más variables a la ecuación, será el reducir al mínimo los residuos, es decir, la diferencia existente entre lo predicho y lo observado. Si al realizar una gráfica de dispersión de una de las variables con respecto a la otra, se aprecia que sus puntos se encuentran sobre una línea vertical u horizontal, entonces las variables son independientes entre sí.

#### *2.2.2.2. Residuos*

Los errores o residuos se definen como la diferencia entre el valor medido y el valor predicho en la regresión. Existen debido a la heterogeneidad intrínseca del sistema sobre el que el fenómeno se desarrolla, ciertos parámetros como la rugosidad de la superficie, la conductividad hidráulica y la vegetación, por nombrar algunos, son muy variables en el espacio y también en el tiempo. (Farahmand, 2011). Aquellos fenómenos, vuelven a las componentes aleatorias más prominentes en la serie y la vuelven más difícil de proyectar en el futuro.

#### *2.2.2.3. Ruido Blanco*

Es el patrón que deben tener los residuos de un buen modelo, y tiene como suposiciones:

- *Los valores de la serie son aleatorios*

La magnitud y signo de un error no se encuentran influidos por los errores anteriores, por lo tanto, no pueden ser predichos.

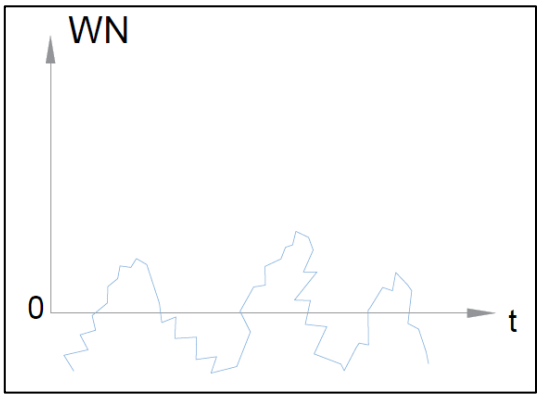


*Figura 2.2-1. Ruido blanco con autocorrelación , se puede predecir que un residuo será grande o pequeño conociendo el valor de su instancia anterior*

- *Homocedasticidad de los errores*

La varianza de los errores, registrada durante distintos intervalos debe ser la misma. Su infracción no es tan grave como la de las otras suposiciones, pero podría causar problemas en los test de hipótesis para la validación del modelo y, además, porque algunos métodos de regresión de las series de tiempo, suponen que la distribución de errores es constante en el tiempo.

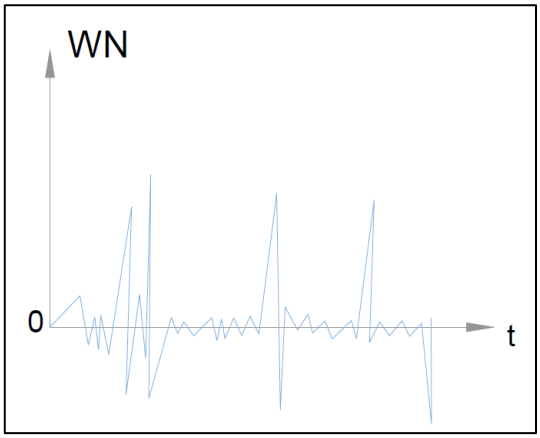


*Figura 2.2-2. Residuos con una desviación estándar que no es constante en el tiempo.*

- *Errores distribuidos normalmente y con media 0*

El teorema del límite central debería aplicarse al momento de graficar un histograma de los residuos, si de verdad son eventos independientes. Si la distribución de errores no toma una forma de campana, en especial cuando la serie tiene una gran cantidad de observaciones, significa que la distribución de los errores no es consistente sobre todo el rango en el que se predice, y el modelo tiene una exactitud diferente en diferentes puntos de su pronóstico. Así mismo, el pronóstico, de preferencia, no debe estar inclinado hacia la sobreestimación o subestimación y en general, debería tener una media de 0.

De igual manera que la homocedasticidad, esta no es una regla muy estricta, porque incluso modelos que son muy eficientes en cuanto a los pronósticos podrían presentar residuos que no están distribuidos normalmente.

El tener residuos que siguen este patrón, implica que son completamente aleatorios y no existe variable alguna que permita su explicación, por lo tanto, el modelo ya ha utilizado las variables explicativas suficientes.

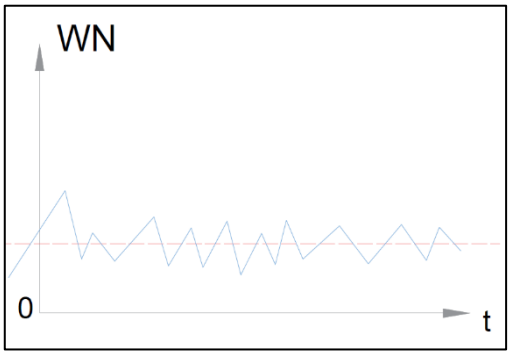


*Figura 2.2-3. Residuos con una desviación estándar que no es constante en el tiempo.*

#### *2.2.2.4. Exogeneidad*

Además de que los valores pasados pueden explicar las ocurrencias en el presente, puede ser que, al comparar las mediciones en una serie con las de otra, exista una gran correlación entre estos dos conjuntos, que permita la generación de regresiones más exactas.

Antes de añadir una de estas variables para explicar la serie, es necesario haber reconocido y haber tomado las medidas necesarias para remover su tendencia y estacionalidad. Si de alguna manera, esta variable logra explicar los residuos que el mejor modelo estocástico no pudo explicar, puede ser incluida en la ecuación a través de variables ficticias o de pesos.

$$S_t = \beta_1 S_{t-1} + \beta_2 S_{t-2} + \cdots + \beta_m S_m + \epsilon_t$$

Siendo $S_m$ el valor disponible en un instante "t" de la variable externa. Si existe proporción entre los valores de los residuos y dicha variable, entonces $\beta_m$es un número entero, si por otra parte $S_m$ solo se activa cuando existe algún evento extremo que es posible conocerlo antes del pronóstico, entonces $\beta_m$ puede asumir la forma de una variable ficticia que toma el valor de 0 o 1, de modo que este factor se encuentre en la ecuación pero no siempre permanezca activo.

### 2.2.3. Modelos hidrológicos

#### *2.2.3.1. Generalidades*

El cambio del ciclo hidrológico debido a variaciones en la demografía, el uso de la tierra, el clima o simplemente el paso del tiempo puede ser simulado a través de modelos hidrológicos. Dentro de un ámbito operacional, permiten observar cómo algunas acciones pueden intervenir en el balance y distribución del agua en una región, y gracias a ello, adoptar las medidas necesarias para distribuir recursos hídricos adecuadamente, crear obras y mejorar las prácticas agrícolas y de riego. (FutureWater, 2019)

Las actividades humanas como la urbanización, la deforestación y los desvíos artificiales de los cursos de agua, modifican la ya entramada interacción entre el suelo, la vegetación, la temperatura, el viento y el agua, que se encuentra constantemente cambiando entre estados físicos, todo esto vuelve a la circulación en los ríos, un proceso muy complejo. Por esto, si no se dispone de la suficiente información acerca de su cuenca, resulta más conveniente tratarla con modelos estadísticos que capturen sus propiedades más importantes matemáticamente, y que permitan deducir datos de interés para su explotación.

Además de cantidades, también podría ser posible pronosticar variables cualitativas, por ejemplo, la calidad del agua.

#### *2.2.3.2. Modelos determinísticos y estocásticos*

En un mundo ideal, y donde se conozca cada detalle de las partículas a nuestro alrededor, absolutamente todas las acciones podrían ser descritas y predichas por ecuaciones físicas. Sin embargo, la realidad no es así, debido a que es prácticamente imposible, entender y registrar la interrelación y características de los causantes de algunos fenómenos, como en este caso, los hidrológicos.

Los modelos que usan una ecuación física determinística para predecir un resultado son denominados de caja blanca (Leyes de Newton, Stokes). Los modelos que usan ecuaciones estocásticas (o parcialmente aleatorias) pueden ser agrupados dentro de los llamados de caja gris y de caja negra, los modelos de caja negra utilizan relaciones matemáticas calibradas para que los valores de entrada correspondan a los valores de salida (como la ecuación de Thomas-Fiering), debido a esto la ecuación encontrada no explica la relación física entre sus variables y, por tanto, es generalmente independiente de la rama de la ciencia que trata. Los modelos de caja gris combinan una estructura teórica como los de caja blanca y de parámetros estadísticos como los de caja negra y por lo tanto la mayoría de ecuaciones pueden ser descritas en este grupo.

Los procesos que pueden ser descritos a través de estos modelos también se observan en otras disciplinas:

- Tasas de cambio
- Dispersión de enfermedades
- Terremotos
- Crecimiento de colonias de bacterias
- Recesión de genes

Las variables hidrológicas tienen una variación intrínseca que vuelve a sus magnitudes oscilantes sobre valores específicos y que, por lo tanto, puede ser descompuesta en patrones invariables y en una componente incierta, por lo que, en muchos casos, la modelación estocástica suele resultar apropiada.

#### *2.2.3.3. Series de tiempo*

Muchas variables en hidrología son observadas en intervalos de tiempo más o menos regulares, ejemplos de ello son registros de lluvia, escorrentía superficial y niveles de agua

subterránea. La observación en gráficas o ecuaciones de estas variables aleatorias en el tiempo, obtenida mediante estaciones de monitoreo y durante intervalos regulares se conocen las series en el tiempo. Dentro de la hidrología estocástica se usan técnicas, provenientes en gran parte de la econometría, para analizar y modelar series hidrológicas en el tiempo. (van Geer & P. Bierkens, 2012)

El análisis de las series en el tiempo puede ayudar a:

- Encontrar datos estadísticos como la probabilidad de tener valores atípicos y las tendencias de la serie.
- Pronóstico de datos en el futuro y completado de datos en el pasado.

- *Regresión de series de tiempo*

El propósito de las regresiones en la estadística descriptiva es la de utilizar los datos acerca de los individuos en una población para poder estimar su comportamiento, no necesariamente a través del tiempo, y con ello, hacer inferencias acerca del resto de la población.

Para su utilización en las series de tiempo, los datos de entrada suelen ser los patrones observados en la serie, aquí se estimará una ecuación o modelo en base a registros pasados que serán comparados con mediciones más recientes para su validación.

#### *2.2.3.4. Series de tiempo sintéticas*

El conocer los parámetros o el rango que ocupan los parámetros que gobiernan un sistema, permite modificarlos de cierta manera, esto permite, a través de los modelos matemáticos mencionados, crear situaciones que muy probablemente se presenten. Los datos que componen a estos nuevos escenarios se denominan datos sintéticos. La elaboración de datos sintéticos, en el dominio del tiempo, conforman las series de tiempo sintéticas.

A pesar de que se espera que las propiedades de cada registro varíen individualmente, uno de los supuestos en la generación de estas series, es que las propiedades estadísticas, en

conjunto, no varíen, es decir, que la media, la varianza y la correlación, una vez descontados los efectos de la tendencia y la estacionalidad, sean conformes a los valores históricos.

#### *2.2.3.5. Estimación de Parámetros de un modelo*

- *Mínimos cuadrados (OLS)*

La estimación de parámetros se realiza con este método para una gran cantidad de aplicaciones de la regresión y algunas veces, en series de tiempo.

Si al tener un gráfico de dispersión de 2 variables, se desea obtener una línea que estime su relación de la mejor manera posible, su valor estimado en cada observación tendrá, en una de sus formas más simples:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

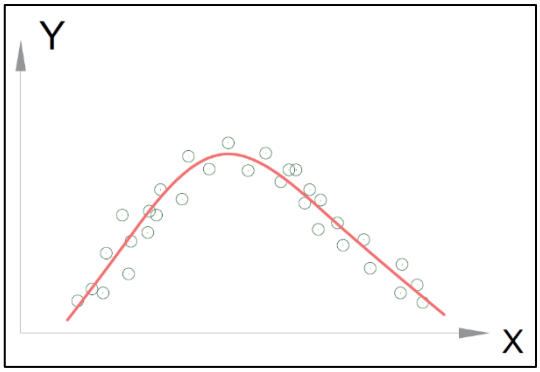


*Figura 2.2-4. Gráfico de dispersión y línea de mejor ajuste, relación cuadrática entre variables*

Para cada abscisa $X_i$, existirá un valor estimado $\hat{Y}_i$ y un valor real $Y_i$, los residuos en cada punto se definen como la diferencia:

$$Y_i - \hat{Y}_i$$

La suma de los cuadrados de esta diferencia, considerando todos los puntos, se denomina RSS (Residual Sum of Squares):

$$\sum_{i=1}^{n} \left(\hat{Y}_i - \hat{Y}_i\right)^2$$

De manera similar a la varianza, los términos que se encuentren muy alejados a la estimación perjudicarán de mayor manera a este término, independientemente de si son errores positivos o negativos.

Este método de estimación de parámetros consiste en encontrar, por ejemplo, para el caso de la función lineal, los valores de $\beta_0$y $\beta_1$, mediante diferenciaciones parciales, que provoquen que RSS sea el mínimo posible para una serie de datos.

- *Máxima verosimilitud (MLE)*

A partir de una recolección de datos $x_1, x_2, \dots, x_n$, que siguen una supuesta distribución de probabilidad, es posible encontrar el parámetro o conjunto de parámetros $\theta$, de un modelo, que maximicen la verosimilitud de estos datos en la función de densidad de probabilidad. La función de verosimilitud se estima como una función de probabilidad de $\theta$, dadas las ocurrencias de $x_i$:

$$f_\theta = \mathrm{L}\,(\theta \mid x_1, x_2, \dots, x_n)$$

El principio en el que se basa este método es el de encontrar el conjunto de parámetros de un modelo que vuelvan más probable la ocurrencia de los valores ya observados (Xu, 2009). La probabilidad conjunta de que todos estos eventos de interés sucedan, se denota como:

$$\mathrm{L}\,(\theta \mid x_1, x_2, \dots, x_n) = f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid \theta) = \mathrm{f}\,(x_1, x_2, \dots, x_n \mid \theta)$$

Debido a que cada uno de estos valores tiene una distribución que se asume independiente de la anterior, entonces:

$$\mathrm{f}\,(x_1, x_2, \dots, x_n \mid \theta) = \mathrm{f}\,(x_1 \mid \theta) * \mathrm{f}\,(x_2 \mid \theta) * \dots * \mathrm{f}\,(x_n \mid \theta)$$

$$\mathrm{f}\,(x_1, x_2, \dots, x_n \mid \theta) = \prod_{i=1}^{n} \mathrm{f}\,(x_i \mid \theta)$$

La probabilidad de que todos eventos ocurran es el producto de sus probabilidades individuales. Los parámetros que provoquen que la función de verosimilitud sea máxima son los estimadores de máxima verosimilitud.

$$\theta\text{: } \mathrm{argmax}_\theta\ \mathrm{L}\,(\theta \mid x_1, x_2, \dots, x_n)$$

De la parte derecha de la ecuación, generalmente se obtiene el logaritmo, que debido a que es una transformación monótona y por lo tanto mantiene el orden en cualquier conjunto de datos, entonces los componentes que provocan que una función sea máxima, también provocaran que su logaritmo lo haga. Esta transformación se realiza para mejorar la interpretación y para facilitar los cálculos en la computadora.

#### *2.2.3.6. Sobreajuste*

Al intentar encontrar un modelo que logre acomodarse a una serie de datos determinada y que a la vez, permita hacer pronósticos en otros intervalos de tiempo, puede creerse que es buena idea generar un modelo que logre ajustar todos los registros, por ejemplo, a través de un mayor grado en los exponentes en sus variables.

Sin embargo, al momento de querer inferir nuevos datos sobre dicho registro o población, debido a que el modelo no tomó en cuenta las variaciones que existen entre los datos, al compararlos con los datos verdaderos, estos tendrán errores más grandes en comparación a modelos más simples. Al contrario que estos, los modelos simples, poseen una holgura que les permite tomar en cuenta volatilidades en los datos.

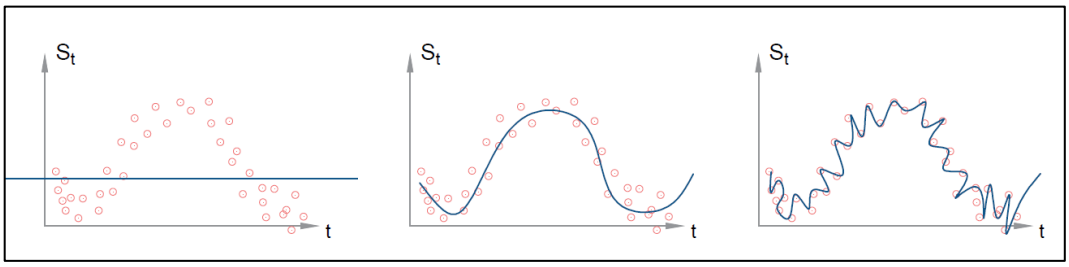


*Figura 2.2-5. Regresiones con subjasute, ajuste adecuado y sobreajuste.*

#### *2.2.3.7. Suavizado de curvas*

Debido a que existen fluctuaciones que pueden ser difíciles de prever a través de modelos, existen ciertos métodos que permite remover esta volatilidad en los datos sin perder los patrones significativos en la serie.

Es posible promediar los resultados de una serie para franjas de tiempo de una longitud mayor a la de la muestra para lograr este propósito, por ejemplo, convertir una serie diaria en semanal al promediar los registros de cada uno de los días. Sin embargo, al realizar esto,

se trabaja con una cantidad menor de datos, y por lo tanto puede que se pierdan algunas de las características de la muestra y los pronósticos que se deriven de ellos. El trabajar con intervalos de tiempo más grandes suele resultar en una mejora en las predicciones, pero a cambio se pierde precisión en mostrar la distribución de los valores de la serie.

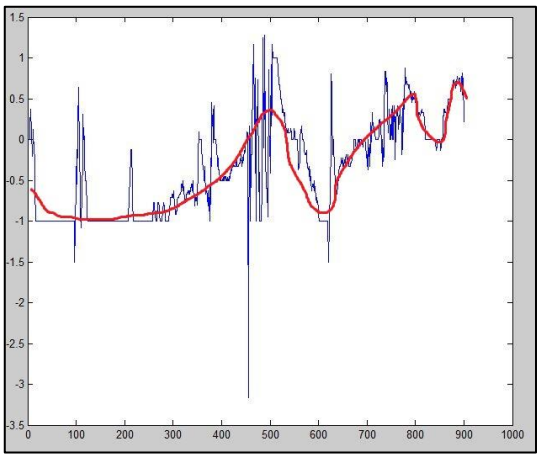


*Figura 2.2-6. Serie de tiempo suavizada. Copyright 2018 por Stack Overflow.* Reimpreso con permiso

### *2.2.3.8. Validación cruzada*

Es una técnica utilizada para evaluar el comportamiento de un modelo en comparación a una serie de datos, consiste en dividir a los datos en una parte de "entrenamiento" o prueba y otra de "validación". El conjunto de entrenamiento suele ser más grande y con este se estiman los parámetros del modelo, para ellos, se pretende que los datos del conjunto de validación, el cual corresponde a los valores más futuros de la serie, no existen. El modelo se utiliza para estimar los valores que la serie tomaría durante el intervalo de tiempo correspondiente al periodo de validación y luego se compara dichos valores con los verdaderamente observados, con ello, comprobando la capacidad predictiva del modelo.

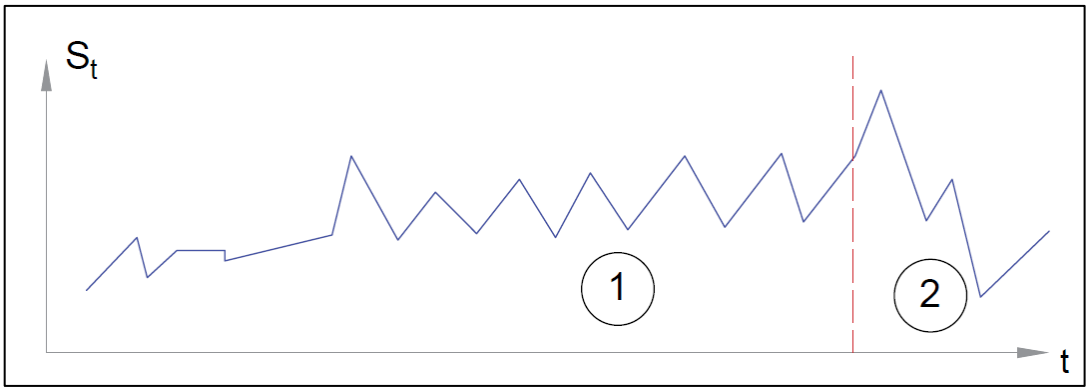


*Figura 2.2-7. Periodos de entrenamiento y validación , se representa a ambos con el número (1) y (2) respectivamente*

### *2.2.3.9. Intervalos de Predicción*

El componente aleatorio en los modelos estocásticos provocará que, a pesar de partir del mismo registro histórico, cada serie generada será diferente, por lo que puede resultar necesario expresar a estos resultados como una distribución probabilística para cada punto.

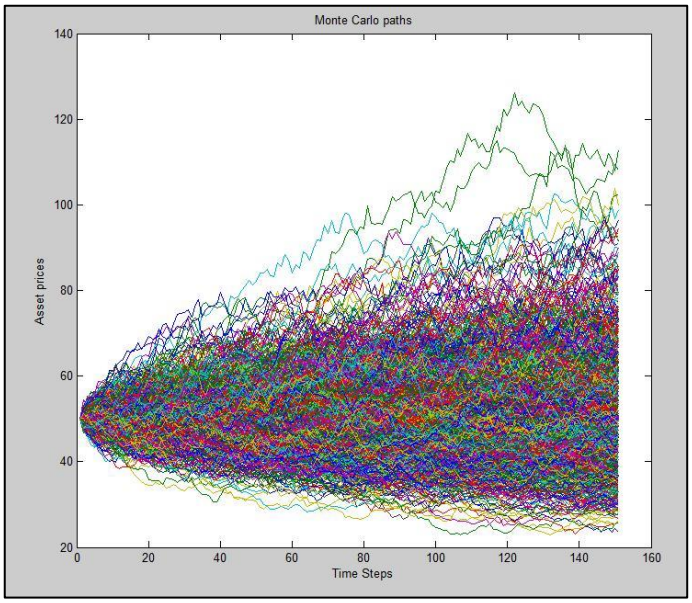


Figura 2.2-8. Series sintéticas generadas por procesos estocásticos. Copyright 2016 por MathWorks. *Reimpreso con permiso*

A medida que aumente el número de generaciones de series sintéticas, los valores para cada instante tendrán una distribución de probabilidad definida, por lo tanto, su media y distribuciones estándar convergen en un único valor, de igual manera lo harán los valores que dividen a la serie en partes iguales, es decir los cuantiles.

En los intervalos de predicción, es posible formular una expresión, dentro de la cual, con un cierto grado de certeza (90%,95% o 99% generalmente), se asegure que el valor pronosticado se encuentre dentro de un rango. Conociendo la distribución de errores del periodo de validación, es posible generar un rango:

$$S_k \pm k\sigma_e$$

Donde k es el número de desviaciones estándar de los errores necesarias para contener cierto porcentaje de dichos errores. Por ejemplo, para el caso de tener predicciones que en cada instante siguen una distribución normal, k tomaría el valor de 2 si se requiere una certeza de 95.45% debido a que, en diferentes escalas, una distribución normal mantiene esa propiedad.

Debido a que aún existe una posibilidad mínima de obtener valores atípicos que se desvíen demasiado de la tendencia de los valores generados, para cada instante, se puede obtener los cuantiles extremos (por decir el percentil 2.5 y 97.5) como límites inferior y superior respectivamente para mostrar márgenes dentro de los cuales los valores son generados. (Salazar & Cadavid, 2008).

La generación de varias rutas que puede tomar la variable a lo largo del tiempo se conoce como generación de Montecarlo y se explicará en la sección de los métodos de generación de pronósticos para los modelos Box-Jenkins.

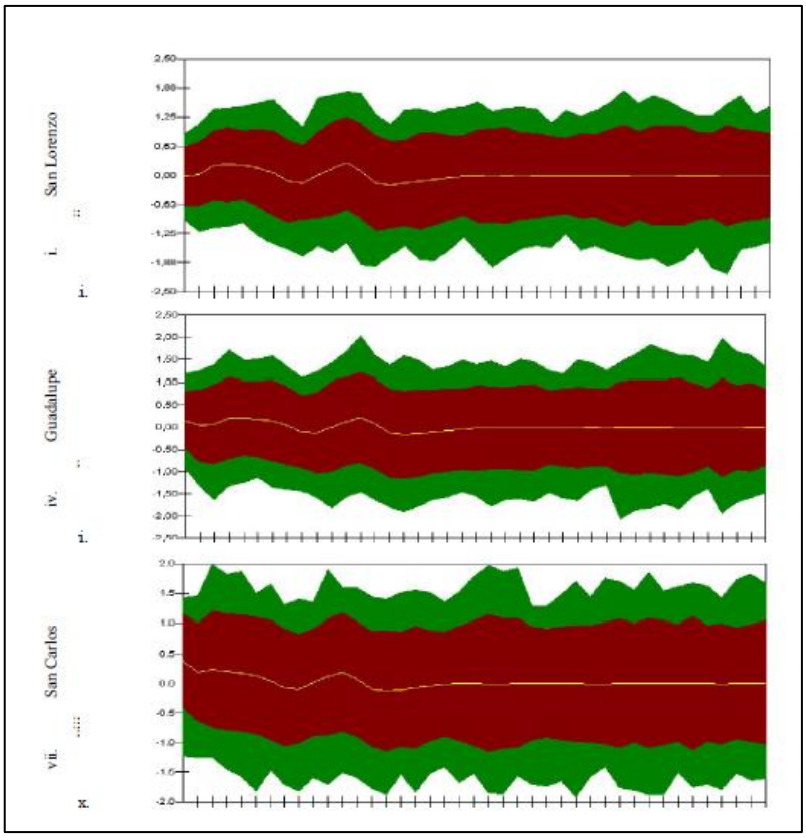


Figura 2.2-9. *Presentación de los resultados en percentiles*. (*Salazar & Cadavid, 2008*)

### *2.2.3.10. Estacionariedad*

Una serie de tiempo se puede considerar estacionaria cuando su media y desviación estándar se mantienen constantes independientemente del tiempo, es decir, si se toman intervalos de tiempo aleatorios e iguales de la serie, la distribución de probabilidad de los datos debería ser la misma.

Sin embargo, además del componente estacionario de la serie, los datos en la vida real se encuentran gobernados por otros componentes (Buteikis, 2018), estos pueden ser:

- Tendencia: El incremento o decremento que se experimenta a lo largo de la serie.

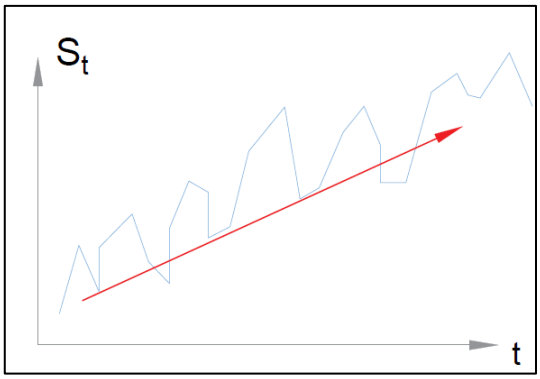


*Figura 2.2-10.Serie de tiempo con tendencia creciente*

- Ciclos: Son caídas y subidas experimentadas por los datos que no se presentan durante periodos bien definidos.

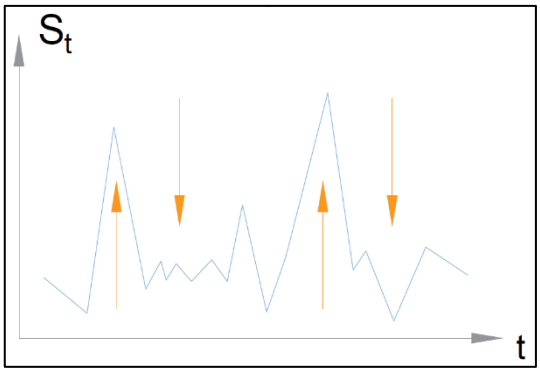


*Figura 2.2-11. Serie de tiempo con ciclos, apreciable por los incrementos y los decrementos no periódicos*

- Estacionalidad: A pesar de la similitud con la palabra estacionariedad, corresponde a las fluctuaciones regulares de las series en épocas específicas, por ejemplo, la crecida en los caudales durante marzo, el cual es un mes con mucha precipitación.

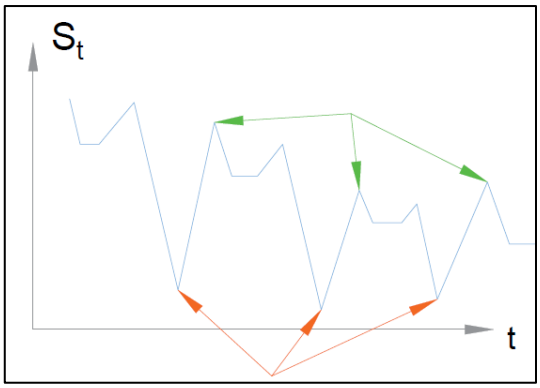


*Figura 2.2-12. Serie de tiempo con periodos, los patrones de la serie se presentan en intervalos regulares.*

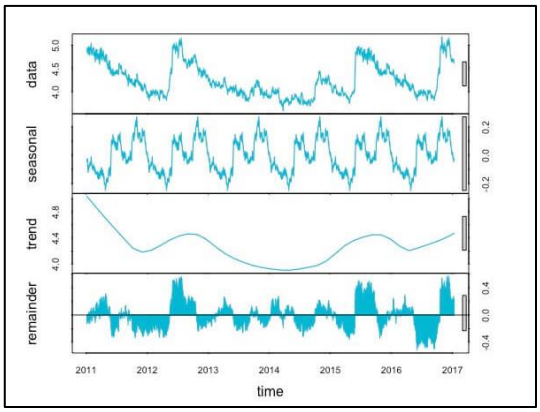


Figura 2.2-13. Descomposición de una serie de tiempo en sus componentes estacionales, de tendencia y un restante. Copyright 2016 por Medium. Reimpreso con permiso

#### *2.2.3.11. Autocorrelación y memoria*

La autocorrelación es una magnitud que mide la relación lineal existente entre los valores presentes de una serie y valores en el pasado en diferentes desfases. Existen coeficientes de autocorrelación en una serie, casi tanto como valores en la misma.

La memoria de una serie de tiempo se refiere a lo mucho que es posible ir atrás en el tiempo y aun así poder predecir los valores presentes de manera certera. Una serie con memoria larga tiene coeficientes significativos, aun realizados varios desfases, mientras que una serie con memoria corta decae rápidamente después de pocos desfases. Esto puede ser mejor explicado al observar las gráficas de autocorrelación, explicadas en el siguiente tema.

#### *2.2.3.12. Función de autocorrelación (ACF)*

En toda serie existirá un grado de dependencia entre valores presentes y valores pasados, es posible explicar su grado de dependencia a través de dos gráficas, la de la función de autocorrelación y la función de autocorrelación parcial.

Por intuición, es posible notar que, por ejemplo, valores negativos y pequeños estarán a continuación de valores de igual magnitud y signo, lo mismo con valores que sean positivos y estén en algún orden de magnitud determinado.

Este gráfico es posible realizarlo al agrupar, en pares ordenados, los valores que toma una serie en instantes desfasados en el tiempo y calcular el coeficiente de correlación de Pearson para distintos desfases. Suponiendo los registros de los meses hasta septiembre, se muestra un ejemplo de distintos desfases:

| **Desfase:1** | | **Desfase:2** | | **Desfase:3** | |
|---|---|---|---|---|---|
| Enero | Febrero | Enero | Marzo | Enero | Abril |
| Febrero | Marzo | Febrero | Abril | Febrero | Mayo |
| Marzo | Abril | Marzo | Mayo | Marzo | Junio |
| Abril | Mayo | Abril | Junio | Abril | Julio |
| Mayo | Junio | Mayo | Julio | Mayo | Agosto |
| Junio | Julio | Junio | Agosto | Junio | Septiembre |

*Tabla 2.2-1. Ejemplo de diferentes desfases para datos mensuales*

La ACF tendrá unas bandas de significancia (aquí representadas en azul) con un ancho inversamente proporcional al número de muestras, dentro de las cuales, los valores son estadísticamente iguales a 0, por lo que no deberían ser tomados en cuenta al momento de generar modelos de predicción. La correlación en el desfase 0 será igual a 1 debido a que estamos relacionando un parámetro con sí mismo.

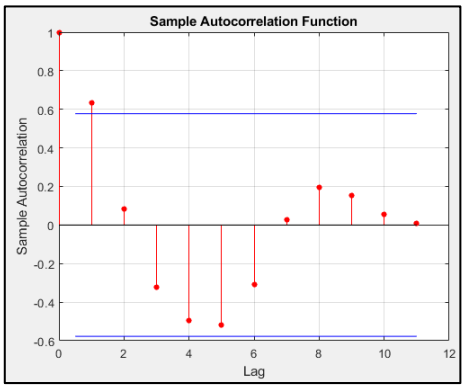


*Figura 2.2-14. Gráfico del ACF de una serie de tiempo creado por MATLAB.*

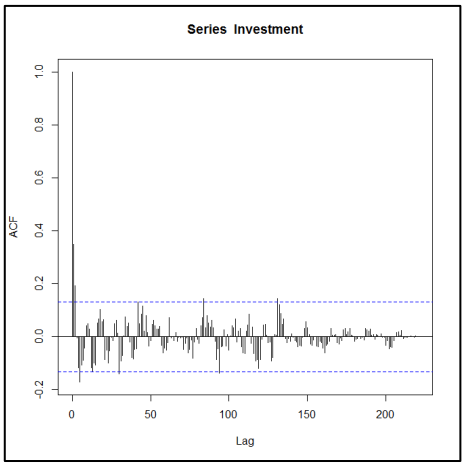


Figura 2.2-15. ACF de una serie con memoria corta. Copyright 2014 por Stack Exchange. *Reimpreso con permiso*

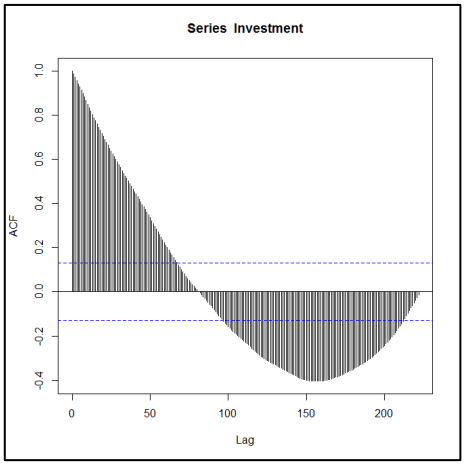


Figura 2.2-16. ACF de una serie con memoria larga. Copyright 2014 por Stack Exchange. *Reimpreso con permiso*

Las series de tiempo con estacionalidad suelen tener una correlación significativa para desfases similares a la duración de su estación.

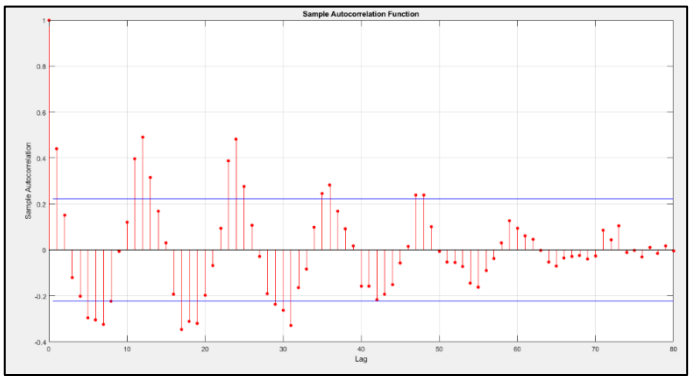


*Figura 2.2-17. ACF de una serie con estacionalidad.*

En MATLAB, la función autocorr (y) genera la gráfica de la ACF para una serie en el tiempo contenida en el vector y.

#### *2.2.3.13. Función de autocorrelación parcial (PACF)*

De manera similar, la PACF grafica los efectos que tienen los datos en el pasado, pero librándose de la influencia de los registros más cercanos. Si un valor en el pasado es capaz de ajustar una regresión para explicar un valor en el presente, entonces quedarán residuos que probablemente puedan ser explicados por valores aún más lejanos en el tiempo. Teniendo una regresión:

$$S_t = \phi_1 S_{t-1} + \phi_2 S_{t-2} + \cdots + \phi_i S_{t-i} + \cdots + \phi_n S_{t-n} + \epsilon_t$$

Para cada retraso "i" en la PACF, se grafica el valor de $\phi_i$.Los coeficientes en esta ecuación pueden ser hallados con el método de los mínimos cuadrados. De igual manera, existirán unas bandas de significancia, que permiten conocer que desfases son relevantes en los modelos.

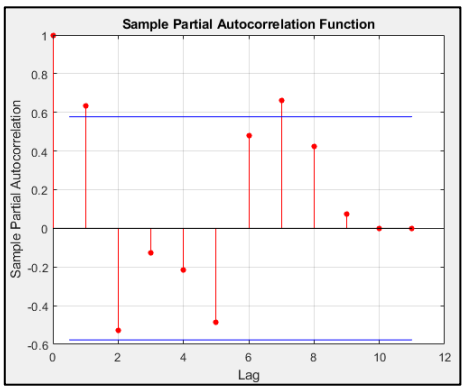


*Figura 2.2-18. Gráfico del PACF de una serie de tiempo creado por MATLAB.*

En MATLAB, la función parcorr(y) gráfica la PACF de una serie en el tiempo contenida en el vector y.

#### *2.2.3.14. Diferenciación*

Es un tipo de transformación que se encarga de la estacionalidad y tendencia de una serie de tiempo, como un requisito previo para aplicar algunas técnicas de pronóstico. Consiste en

utilizar otra variable que guarda las diferencias entre valores consecutivos o distantes de una serie.

Una diferenciación de retraso 1 implica sustraer dos valores de la serie sucesivos:

$$Z_t = S_{t+1} - S_t$$

Esto ayudar a eliminar la tendencia lineal en la serie. Ya que la tasa de cambio de una función lineal es constante, la media de la variable $Z_t$ será más o menos constante, si con ello no se logra estacionariedad en la serie, la variable transformada $Z_t$ puede seguirse diferenciando.

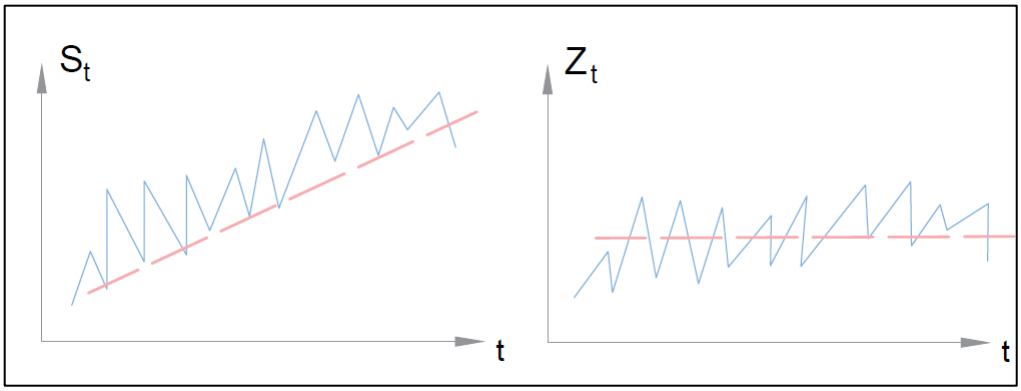


*Figura 2.2-19. Serie de tiempo con tendencia y serie de tiempo diferenciada*

Una diferenciación con respecto a valores más distantes es útil para remover la estacionalidad de una serie de tiempo, el retraso en este caso, es igual a la cantidad de datos que se encuentran en un ciclo.

$$Z_t = S_{t+m} - S_t$$

No existe ninguna restricción para realizar ambas diferenciaciones a una serie de datos, en caso de que sea necesario remover su tendencia y estacionalidad.

Una vez que se realice la predicción de la variable transformada, puede utilizarse la misma ecuación para volver a la variable original. Por ejemplo, para el caso de una diferencia con respecto a un retraso, suponiendo que $S_k$ es el primer valor pronosticado de la serie original, $S_{k-1}$ es el último valor conocido y $Z_{k-1}$ la primera variable transformada pronosticada, entonces es posible obtener los nuevos pronósticos de la forma:

$$S_k = Z_{k-1} + S_{k-1}$$

El siguiente valor de la serie será:

$$S_{k+1} = Z_{k-1} + Z_{k-2} + S_{k-1}$$

#### *2.2.3.15. Medidas numéricas del desempeño de modelos*

Para este estudio, se evaluará el desempeño de los modelos en función de lo pequeños que son los errores de sus pronósticos.

- *Errores dependientes de la escala*

Son valores que no sirven para medir la exactitud de los métodos aplicados entre series que tienen diferentes magnitudes o que corresponden a una versión transformada de ellas, por ejemplo, por logaritmos o estandarizaciones, debido a que esta medida, se encuentra en la misma escala que las mediciones.

  - *Error medio absoluto (MAE)*

Es el promedio de los valores absolutos de los errores en cada punto, tomando en cuenta que, si se realiza un promedio aritmético de ellos, los errores positivos cancelarían a los negativos.

$$MAE = \frac{\sum_i^n |e_i|}{n}$$

  - *Error de la raíz cuadrada de la media (RMSE)*

Es otra medida que hace más énfasis en valores atípicos y que se encuentran, por lo tanto, muy lejanos a lo pronosticado para esas observaciones. Su desventaja se encuentra en que, a pesar de que puede que un modelo capture correctamente los patrones de una serie, el énfasis que hace sobre errores muy grandes puede perjudicar la valoración de esta aproximación.

$$RMSE = \sqrt{\frac{\sum_i^n (e_i)^2}{n}}$$

- *Errores Porcentuales*

Se generan a partir de la relación entre el error para cada intervalo de tiempo y el valor de la observación respectiva $p_i = \frac{e_i}{y_i}$. Tiene la ventaja de que es mejor para comparar en diferentes escalas, por ejemplo, el flujo de dinero en diferentes épocas del año, donde es notoria la mayor demanda durante ciertos periodos. Debido a que son valores adimensionales, permiten verificar el desempeño de los nuevos modelos en varias series de datos.

  - *Error medio absoluto porcentual (MAPE)*

Expresado como un promedio de los errores porcentuales en cada punto. El único inconveniente con esta clase de métricas es que el valor del error puede tender a infinito en caso de observaciones muy pequeñas.

$$MAPE = \frac{\sum_i^n |p_i|}{n}$$

- *Coeficiente de Nash-Sutcliffe modificado*

Es un valor utilizado frecuentemente para evaluar la capacidad predictiva de modelos hidrológicos. Es correspondiente al coeficiente de determinación ($R^2$), utilizado en la validación de regresiones. Tiene la forma:

$$NSE = 100 * \left(1 - \frac{\sum(\hat{y} - y)^2}{\sum(\bar{y} - y)^2}\right)$$

Tiene una modificación para el caso de variables que contienen una gran asimetría en su distribución, como es el caso de los caudales, y que utiliza los logaritmos para no impactar tanto en los errores de los pronósticos de valores grandes, que generalmente tienen mayor volatilidad y pueden impactar gravemente en la validación del modelo.

$$log\ NSE = 100 * \left(1 - \frac{\sum[\log(\hat{y}) - \log(y)]^2}{\sum[\log(\bar{y}) - \log(y)]^2}\right)$$

El valor de este coeficiente puede variar de 100 a $-\infty$, siendo los valores a 100 los más acertados. Se muestran algunos valores de referencia (Rojas et al., 2007) (Ritter y Muñoz, 2012):

| Valor | Significado |
|---|---|
| **100** | Modelo ajustado perfectamente con respecto a lo observado |
| **50-65** | Modelos suficientemente buenos |
| **0-50** | Modelos que varían entre buenos y regulares |
| **0** | La media de los valores del periodo de prueba es un pronóstico igual de bueno |
| **<0** | La media de los valores del periodo de prueba habría sido un mejor pronóstico |

*Tabla 2.2-2. Valores referenciales para el coeficiente de Nash-Sutcliffe modificado*

#### *2.2.3.16. Test de hipótesis*

Una hipótesis es una idea que puede ser puesta a prueba, en lo que respecta a estadística, esta idea se trata de los parámetros de una población, como, por ejemplo:

- La media del peso de sandías provenientes de una granja es 5kg.
- El número de turistas que visitan una ciudad diariamente, proviene de una distribución exponencial.
- Las ganancias mensuales de una tienda son menores que 10000$ para todos los meses.
- La probabilidad de que llueva un día cualquiera en Quito es mayor o igual al 40%.

El test de hipótesis es un cálculo realizado para comprobar la veracidad de estos supuestos acerca de una población mediante una muestra. Es un procedimiento que permite reducir la subjetividad de las conclusiones mediante procedimientos más formales.

Para las hipótesis en estadística se tiene una hipótesis determinada $H_0$ y una hipótesis alternativa $H_a$, las cuales son mutuamente excluyentes, ejemplos de hipótesis determinadas son las ideas mencionadas al principio de este tema, mientras que las hipótesis alternativas serían sus opuestos lógicos, que en conjunto deben comprender todos los posibles resultados:

- La media del peso de las sandías no es 5kg.
- El número de turistas que visitan la ciudad no provienen de una función exponencial.
- Las ganancias mensuales de una tienda son mayores o iguales que 10000$ para todos los meses.
- La probabilidad de que llueva un día cualquiera en Quito es menor al 40%.

Suposición sobre la cual se realiza una serie de cálculos, que servirá para establecer cuál de las hipótesis es más consistente con las posibles características de la población. El estadístico de la prueba es el producto de estos cálculos. Por ejemplo, existe la hipótesis de que las galletas con chocolate que salen de una fábrica, tienen en promedio 10gr de chocolate, se realizan durante distintos días, muestreos de 50 galletas y dan como resultado:

- 9,9gr
- 13gr
- 22gr

El estadístico es un valor más favorable para aceptar la hipótesis durante el primer día, ya que se asemeja más a lo supuesto, mientras que, para el segundo y tercer día, el estadístico es un valor que vuelve más propenso el rechazo de esta hipótesis debido a lo mucho que se aleja de lo supuesto.

Ya que siempre existe una incertidumbre inherente al momento de trabajar con una porción de la población y no con su totalidad, nunca existirá una certeza del 100%, por lo que se debe elegir niveles de confianza, de los cuales generalmente se utiliza un 95%, aunque también es común elegir un 90 o 99%.

- *Test de Dickey-Fuller aumentado*

Es un ensayo para comprobar el supuesto de que una muestra de datos proviene de una serie estacionaria. Sirve para determinar la estacionariedad requerida para el pronóstico de a través de los modelos Box-Jenkins.

La teoría detrás del test de Dickey-Fuller aumentado es relativamente compleja y no se la aborda, se utiliza MATLAB para realizar la comprobación de que la serie a trabajarse es estacionaria una vez realizadas las diferenciaciones respectivas previo al pronóstico a partir de los modelos ARIMA.

MATLAB permite hacer este test mediante la función adftest (y) para una serie ubicada en el vector “y”. El resultado “0” implica que se acepta la hipótesis determinada $H_0$ de no estacionariedad, mientras que el resultado “1” implica el rechazo de esta hipótesis y la aceptación de la hipótesis alternativa.

#### *2.2.3.17. Medidas subjetivas para el diagnóstico de residuos*

Son las utilizadas en el presente documento para comprobar la veracidad de los pronósticos, son descritas en la sección de econometría para MATLAB en la página oficial del software.

- *Gráfico Q-Q*

Es un método gráfico utilizado para comparar dos distribuciones de probabilidad al contrastar los cuantiles correspondientes de cada una. Sirve para comprobar ciertos supuestos de las regresiones, para el caso presente, se lo utiliza para comprobar que la distribución de los errores es más o menos normal, al comparar la distribución de los errores con una distribución normal estándar.

El principio en el que se basa este método es el comprobar la proporcionalidad entre dos series. Después de ordenar la serie, se determina a que cuantiles representa cada número de la serie, los cuantiles de la serie con la que se la desea comparar, se grafican en un plano cartesiano como las abscisas de los cuantiles de la otra serie, que corresponden a las ordenadas. Si ambas series son proporcionales, entonces el gráfico de dispersión de los puntos debería ajustarse a una línea recta.

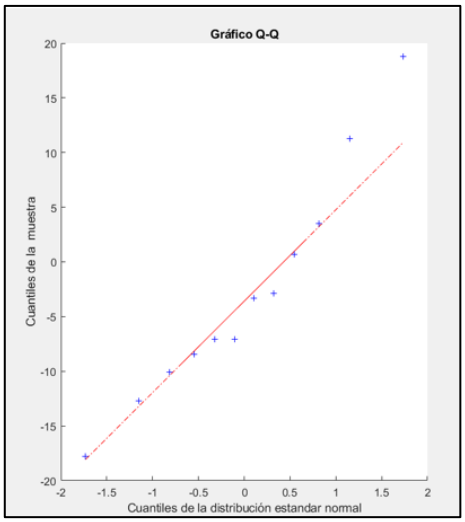


*Figura 2.2-20. Gráfico QQ de una distribución similar a la normal*

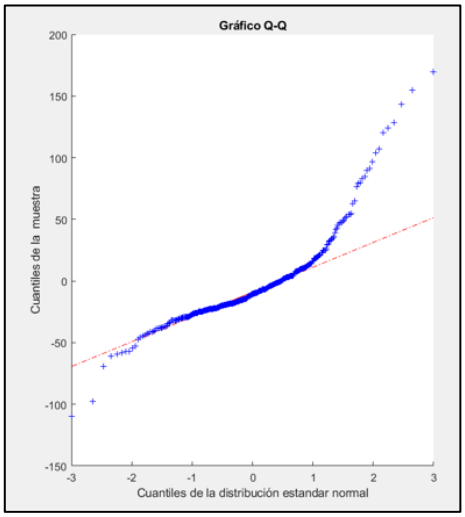


*Figura 2.2-21. Gráfico QQ de una distribución diferente a la normal*

- *ACF Y PACF de los residuos*

Para comprobar que los errores son realmente aleatorios y no pueden ser explicados por errores anteriores, una de las características del ruido blanco, puede graficarse las funciones de autocorrelación total y parcial como un procedimiento informal para comprobar este supuesto, la autocorrelación, en ambos casos, debería ser cercana a 0 para retrasos diferentes de 1.

- *ACF Y PACF de los residuos al cuadrado*

Mediante este procedimiento, se comprueba que la varianza de la serie de residuos es constante, en decir, los errores son homocedásticos, partiendo del supuesto de que el valor absoluto de los errores o el cuadrado de los errores puede expresar lo volátil que es una serie.

Se supone la serie estocástica:

$$y_t = \mu + \varepsilon_t$$

Compuesta por un valor medio $\mu$ y una variable aleatoria $\varepsilon_t$. Debido a la linealidad en la varianza, esta puede ser calculada a partir de:

$$var(y_t) = var(\varepsilon_t)$$

De la definición de la varianza se puede obtener que:

$$var(\varepsilon_t) = E[\varepsilon_t^2] - E[\varepsilon_t]^2$$

La media del proceso de los errores $\varepsilon_t$ se supone igual a 0 en la mayoría de suposiciones de los modelos, por lo que la ecuación mencionada se convierte en:

$$var(\varepsilon_t) = \varepsilon_t^2$$

Por lo que la varianza en cada instante está relacionada directamente con su error correspondiente al cuadrado. Para demostrar que la varianza de los errores no es dependiente de la varianza en periodos anteriores, es decir no puede ser expresada como una relación lineal de errores en el pasado, y se mantiene constante en el tiempo:

$$e_t^2 = \sum a_i e_i^2$$

Las gráficas de autocorrelación parcial y total de los residuos al cuadrado, igualmente como en el método anterior, deben ser valores cercanos a 0 para retrasos diferentes de 1.

### 2.2.4. Modelos para pronósticos de caudales

#### *2.2.4.1. Modelos de medias condicionales Box-Jenkins*

- *Generalidades*

Son modelos matemáticos univariables creados por los matemáticos George Box y Gwilym Jenkins en 1970 (Box & Jenkins, 1970), diseñados para el pronóstico de series estacionarias, pero que tienen una metodología bien definida para incluir patrones de la serie como la tendencia y estacionalidad y que además pueden añadir variables exógenas.

- *Modelos Autoregresivos (AR)*

Es un tipo de regresión que permite conocer valores de la magnitud de un fenómeno en un instante a través de la combinación lineal de registros en el pasado. Tiene la forma:

$$AR(p): S_t = \phi_1 S_{t-1} + \phi_2 S_{t-2} + \cdots + \phi_p S_{t-p} + \epsilon_t$$

El orden "p" de la ecuación, o el número de datos que se toman en cuenta al retroceder en el tiempo, depende de la función de autocorrelación parcial (PACF), debido a que en ella se puede observar las influencias parciales del registro en diferentes desfases.

- *Modelos de Media Móvil (MA)*

Otro modelo usado en hidrología y que permite estimar valores futuros, pero esta vez, utilizando una combinación lineal de los errores experimentados en el pasado, es el modelo de Media Móvil. En base a lo mucho que esta ecuación se equivoca prediciendo valores pasados, el modelo corrige las nuevas estimaciones en diferentes proporciones, su ecuación tiene la forma:

$$MA(q): S_t = \mu + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

La cual tiene un polinomio de errores que, de acuerdo a si la ecuación sobrestimó o subestimó los datos en el pasado, este corrige estos errores en futuros pronósticos a través de este sumando.

La ACF sirve para elegir el orden (q) en el modelo, el valor esperado de los residuos en intervalos anteriores se espera que siga teniendo una influencia en los errores posteriores, en cuanto a las ordenadas de ese retraso en la ACF se encuentren fuera de las bandas de significancia.

- *Modelos Autoregresivos de Media Móvil (ARMA)*

Estos modelos combinan los polinomios de los procesos estocásticos autoregresivos y de media móvil para generar la ecuación:

$$ARMA(p,q): S_t = \phi_1 S_{t-1} + \phi_2 S_{t-2} + \cdots + \phi_p S_{t-p}$$

$$+\theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Donde “p” y “q” son los órdenes de los componentes autoregresivos y de media móvil respectivamente, estos parámetros se estiman de igual manera que se hacía en los modelos AR Y MA. Al contener una mayor cantidad de parámetros, lo vuelve un modelo más preciso y que captura más patrones de la serie.

- *Modelos Autoregresivos Integrados de Media Móvil (ARIMA)*

Muchos de los registros de econometría y de los fenómenos climatológicos tendrán una tendencia y/o estacionalidad, que, si no se toman en cuenta al momento de realizar las regresiones, puede que muestren patrones en la serie de residuos, infringiendo una de las suposiciones para la validación de modelos.

El modelo ARIMA es una generalización del modelo ARMA que también contiene un parámetro “d” que es el grado de diferenciación que requiere la serie hasta que se la considere estacionaria. Debido a esta nueva condición, le permite tomar en cuenta la tendencia de la serie.

La forma del modelo es:

$$ARIMA(p,d,q): Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p}$$

$$+\theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Donde "Z" es la variable ya diferenciada, y todos los demás parámetros $\phi_i$ y $\theta_i$ , así como el orden del modelo se estiman para esta nueva variable transformada.

- *ARIMA estacional (SARIMA)*

Además de los patrones que logra capturar el modelo ARIMA por sí solo, es posible identificar estacionalidad en las series y convertir a la serie en una estacionaria al realizar una diferenciación con un retraso correspondiente al número de valores contenidos en una estación.

Estos fenómenos pueden ser identificados por inspección de la serie, por test de estacionalidad, o si al graficar la ACF, se observa una alta correlación entre un valor y su correspondiente una semana, un mes, un año o una temporada atrás.

La forma de este modelo es:

$$SARIMA: (p,d,q)(P,D,Q)_m$$

Donde "$m$" es el factor estacional que implica el retraso de los valores que se necesitan sustraer para obtener la estacionariedad. Los órdenes $P, D, Q$ implican aplicar el modelo ARIMA, pero sobre los datos que ya no son estacionales.

$$SARIMA: (p,d,q)(P,D,Q)_m = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p}$$

$$+\upsilon_1 Z_{t-1} + \cdots + \upsilon_P Z_{t-P}$$

$$+\theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

$$+\varphi_1 \epsilon_{t-1} + \cdots + \varphi_Q \epsilon_{t-Q} + \epsilon_t$$

En MATLAB, la función estmdl (Mdl,y) utiliza la máxima verosimilitud para la estimación de modelos del orden SARIMA a partir de las características del modelos guardadas en la

variable “Mdl” y la serie de datos pasados “y”. La estimación de parámetros por máxima verosimilitud es la utilizada por los modelos SARIMA y sus predecesores.

- *Modelos contiguos y no contiguos*

Puede darse el caso de que incluso cuando la serie se vuelve estacionaria, valores que no tan cercanos al presente puedan ayudar a explicarlo, un modelo autoregresivo de la forma:

$$S_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \epsilon_t$$

Sería un modelo contiguo, y, al contrario, un modelo de la forma:

$$S_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-8} + \epsilon_t$$

No sería contiguo porque está tomando en cuenta la influencia de los retrasos entre el 2 y el 7.

- *Parsimonia y criterio de información bayesiano*

El principio de la parsimonia en un principio implica que, para el esclarecimiento de un fenómeno, y al tener varias explicaciones a la disposición, es más conveniente y generalmente lo correcto elegir la más simple. Para el caso de los mecanismos explicados por la ciencia, esto implica que mientras menos pasos un proceso tenga, más fácil se vuelve para explicar.

Para el caso de las series en el tiempo, el elegir modelos que sean más simples, resultará en mayor eficiencia computacional y que, al depender de menos parámetros, las desviaciones que estos pueden tener son menores.

El criterio de información bayesiano es un método que ayuda a elegir un modelo de pronóstico de series en el tiempo, entre varios calibrados a partir de los mismos datos, en base a:

- El valor de máxima verosimilitud
- El número de parámetros

Y tiene la forma:

$$BIC = \ln(n)\, k - 2\ln\left(\hat{L}\right)$$

Donde:

$$n: número\ de\ datos\ observados$$

$$k: número\ de\ parámetros$$

$$\hat{L}: valor\ que\ toma\ la\ función\ de\ MLE$$

El mejor modelo, según este método, es el que vuelve mínimo a este coeficiente.

Se debe tener en cuenta que el valor obtenido al realizar esta operación no es una calificación absoluta del modelo, sino que es una calificación relativa frente a otra clase de modelos de la misma naturaleza y con parámetros calculados a partir de una misma cantidad de observaciones.

La función aicbic (logL,numParam,numObs), esta función utiliza los valores del logaritmo de la función de máxima verosimilitud, el número de parámetros (suma de los órdenes del modelo ARIMA) y el tamaño de la muestra utilizada para la calibración de los parámetros.

- *Métodos de generación de pronósticos*
    - *Método de Montecarlo*

Es un algoritmo que utiliza la aleatoriedad para dar estimados determinísticos acerca de los pronósticos de una serie en el tiempo.

Se trata de generar un gran número de "rutas" para expresar a cada pronóstico no como un valor puntual sino como una distribución de probabilidades, que permite observar todo el rango de posibilidades, con una esperanza matemática en cada uno de los intervalos, que en lo posible se acercara a un valor más "real" de la serie. Las componentes aleatorias se estiman a partir de la distribución de errores encontrada por máxima verosimilitud, se supone el caso de un modelo Box-Jenkins AR (2):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

Dado que la componente de errores es distribuida normalmente y tiene una varianza y media obtenida de la calibración, $\varepsilon_t \sim N(0, \sigma^2)$, esta puede ser generada a partir de un generador de números aleatorios, para los dos primeros pronósticos, siendo "0" el último tiempo registrado:

$$y_1 = c + \phi_1 y_0 + \phi_2 y_{-1} + \varepsilon_1$$

$$y_2 = c + \phi_1 y_1 + \phi_2 y_0 + \varepsilon_2$$

Este y el método de generación de pronósticos MMSE necesitan un mínimo número de observaciones necesarios para calibrase y proyectarse, para el caso de los modelos ARIMA este número corresponde al máximo orden del modelo. Además, solo deberían usarse para proyecciones al corto plazo porque después de 100 o 200 iteraciones, estos modelos convergen en la media de la serie estacionaria.

En MATLAB, la función simulate (Mdl, x,'NumPaths', y,'Y0', z); crea varias rutas para una simulación Montecarlo a partir de las características del modelo guardadas en la variable "Mdl", generando "y" rutas de "x" intervalos de tiempo partiendo de una serie de datos pasados "z".

- *Método MMSE*

Este método de generación tiene una naturaleza determinística en cuanto a que utiliza la suposición de que el mejor estimador de una serie es el que minimiza el valor esperado de los residuos al cuadrado en cada punto:

$$E(y_{t+1} - \hat{y}_{t+1})^2$$

Al realizar las operaciones correspondientes, se obtiene que, dado que la esperanza de los errores es igual a 0, en todas sus iteraciones, debido a su distribución de probabilidad $\varepsilon_t \sim N(0, \sigma^2)$, existe una única respuesta para cada instante, para el caso de una serie AR (2):

$$\hat{y}_{N+1} = c + \phi_1 y_N + \phi_2 y_{N-1}$$

$$\hat{y}_{N+2} = c + \phi_1 \hat{y}_{N+1} + \phi_2 y_N$$

$$\hat{y}_{N+3} = c + \phi_1 \hat{y}_{N+2} + \phi_2 \hat{y}_{N+1}$$

En MATLAB, la función forecast(Mdl,x,'Y0',y); crea una única ruta para la generación a través de MMSE a partir de las características del modelo guardadas en la variable "Mdl", generando "x" intervalos de tiempo partiendo de una serie de datos pasados "y".

#### *2.2.4.2. Modelos Autoregresivos con Heteroscedasticidad Condicional*

A pesar de que son menos usados en hidrología, esta clase de modelos pueden servir cuando se aprecia que existe alta volatilidad en los residuos y por lo tanto su varianza está condicionada a su posición en el tiempo. En otras palabras, los errores de los pronósticos de algunos periodos son muy grandes en comparación con los de otros.

La mejor justificación para usar este modelo, es cuando existe un fenómeno que, al usar los métodos anteriores, tiene periodos que son fácilmente predecibles, pero en algún momento, debido a alguna causa no considerada, dejan de serlo.

Además de un componente autoregresivo, toma en cuenta un componente de error que no viene dado por un polinomio, como es el caso de los modelos de media móvil, sino por un componente aleatorio y distribuido normalmente multiplicado por un coeficiente que es función de anteriores errores y que define la magnitud de este valor.

$$\varepsilon_t = ruido * \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \cdots + \alpha_n \varepsilon_{t-n}}$$

$$ruido \sim N(0,1)$$

El ACF de los residuos de anteriores modelos puede dar una idea de hasta que cierto punto los errores se encuentran relacionados.

#### *2.2.4.3. Neuronas Artificiales*

Es una serie de algoritmos que se basan en la identificación de los patrones en conjuntos de datos y que se encuentra compuesto por “capas” de entrada y salida, así como capas ocultas que les permiten capturar más detalles y aumentan la complejidad de los modelos. Son utilizados para generar modelos de energía renovable, reconocimiento facial y de caracteres en libros, filtrado de datos, economía y últimamente en la hidrología. Son especialmente útiles cuando se tiene registros que llegan con una alta frecuencia debido a la capacidad de patrones que es capaz de reconocer. A pesar de su utilidad, son modelos de caja negra que tienen interrelaciones que resultan difíciles de interpretar.

Cada una de estas capas contiene “neuronas” que son funciones matemáticas que se calculan en base a la salida de neuronas anteriores.

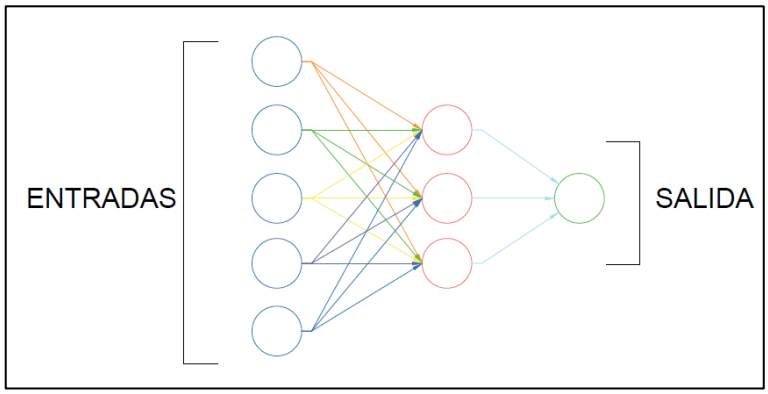


*Figura 2.2-22. Capas y neuronas en una red.*

Cada salida de una neurona es una función del peso relativo de los aportes de sus neuronas anteriores, además de un sesgo que el programa sobre el que se realice este proceso considere adecuado.

Las salidas de cada neurona $a_i$ y sus pesos $w_i$ además del sesgo $b_j$ dan lugar a una expresión $z$, de la que se deriva una función $g(z)$, esta función puede ser una transformación logarítmica, una curva S, una regresión lineal u otra operación. El programa itera constantemente los pesos relativos y el sesgo de cada una de estas neuronas para obtener un resultado más similar al deseado.

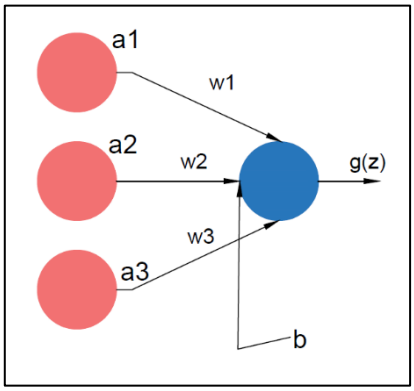


*Figura 2.2-23. Dependencia de las neuronas con sus precedentes.*

El proceso de validación cruzada es un análogo, al “entrenamiento” o cálculo de parámetros de la red, en el cual, el rango de prueba corresponde a la neurona de entrada y el rango de validación corresponde a la neurona de salida. Este método puede aplicarse para calcular no uno, sino varios valores en una serie, por ejemplo, podría colocarse en la neurona de entrada 10 años de registros y en la neurona de salida 1 año o un registro de similar longitud al

horizonte del pronóstico, para con ello, por ejemplo, obtener un algoritmo que permita conocer cuál será el valor de todas las observaciones en un año a partir de lo sucedido en los 10 años anteriores.

En MATLAB, las aplicaciones Neural Network Time Series APP y Neural Network Fitting APP pueden ser utilizadas para generar una simple red de neuronas utilizando los valores por defecto en el programa para pronosticar una serie en el tiempo de la preferencia del usuario.

#### *2.2.4.4. Ecuación de Thomas-Fiering*

Uno de los mayores pasos en el desarrollo de las secuencias sintéticas de flujos de ríos lo hizo el grupo interdisciplinario Harvard Water Program, en sus inicios con más enfoque en el saneamiento, que se encargaba del desarrollo de técnicas para el manejo del recurso hídrico, desde hidroeléctricas hasta piscinas públicas. (Lawrance & Kottegoda, 1977) Los profesores de esta universidad y parte de este grupo, Myron Fiering y su educador Harold Thomas, introducen una ecuación generadora de caudales mensuales, con componentes estacionales y de naturaleza autoregresiva, y que fue en sus inicios aplicada al manejo de numerosos pequeños reservorios en la cuenca del río Meramec, en Missouri.

Debido a la necesidad ya explicada de añadir un parámetro aleatorio a las ecuaciones de generación de parámetros hidrológicos, toma la forma:

$$Q_i = d_i + e_i$$

La parte determinística de la ecuación "d" servirá para mantener los patrones del registro histórico, mientras que la parte aleatoria "e", añadirá un componente de incertidumbre al modelo. Aquí se realiza la suposición de que cada flujo en un instante de tiempo "t" será una función únicamente dependiente del valor de los flujos de instantes anteriores, una autocorrelación:

$$d_i = \beta_0 + \beta_1 q_{t-1} + \beta_2 q_{t-2} + \cdots + \beta_m q_{t-m}$$

Por lo que la parte determinística es una combinación lineal de los caudales "m" instantes antes de "t", la ecuación de Thomas-Fiering, en su forma más básica considera m=1 para lo cual la ecuación del caudal:

$$Q_i = \beta_0 + \beta_1 q_{t-1} + e_i$$

Con estos componentes se realiza las siguientes suposiciones:

- $\beta_0$ es la media de los valores observados para ese intervalo de tiempo (mismo mes en anteriores años, misma hora en anteriores días).
- $\beta_1 q_{t-1}$ es la desviación del flujo anterior en comparación a la media de los datos pasados en $\beta_0$ y es proporcional al coeficiente de correlación entre la serie de datos actuales y del periodo anterior.
- La varianza ($\sigma^2$) y media( $\mu$) de cada grupo de datos es igual.

$$Q_i = \mu + \rho(Q_{i-1} - \mu) + e_i$$

La naturaleza de este modelo también puede explicarse por un proceso de Markov, que toma estados en un espacio continuo y que, al realizar la diferenciación con respecto a los valores medios en cada instante, toma en cuenta la estacionalidad en alguna medida. La validez de usar un proceso de Markov, según uno de sus autores (Fiering & Jackson, 1971), no tiene una explicación particular más que la persistencia estadística de los caudales, de que el orden de magnitud de los flujos se mantenga durante ciertos periodos, épocas con grandes flujos, formaran niveles de agua subterránea, los cuales tenderán a ser liberados durante el siguiente periodo, incluso si no hay una cantidad importante de lluvias. Por otra parte, se asume que, si los flujos en un periodo anterior fueron bajos, incluso si hay una gran cantidad de lluvia, esta será absorbida y por lo tanto los flujos subsiguientes serán igualmente bajos.

La varianza de la componente aleatoria $e_i$ puede ser explicada a partir de las propiedades de la varianza del caudal en el instante "$i$":

$$var(Q_i) = \sigma^2 = var(\mu + \rho(Q_{i-1} - \mu) + e_i)$$

$$\sigma^2 = var(\mu) + \rho^2 var(Q_{i-1} - \mu) + var(e_i)$$

$$\sigma^2 = \rho^2 var(Q_{i-1}) - \rho^2 var(\mu) + var(e_i)$$

La varianza de la variable $e_i$ se la describe como $\sigma_e^2$.

$$\sigma^2 = \rho^2 var(Q_{i-1}) - \rho^2 var(\mu) + \sigma_e^2$$

Tomando en cuenta que $\sigma^2$ se supone igual para todos los periodos:

$$\sigma^2 = \rho^2 \sigma^2 + \sigma_e^2$$

Se despeja y se obtiene:

$$\sigma_e^2 = \sigma^2(1-\rho^2)$$

Se supone que esta componente asume una distribución normal, así que a partir de un número generado aleatoriamente $t_i$ , dentro de una función normal estándar $N(0,1)$, se escala este número por la desviación estándar de dicho componente y se obtiene:

$$e_i = t_i\sigma\sqrt{(1-\rho^2)}$$

La ecuación generadora de caudales en su forma más simple y sin tener en cuenta las tendencias durante distintos periodos (estacionalidad), es entonces:

$$Q_i = \mu + \rho(Q_{i-1} - \mu) + t_i\sigma\sqrt{(1-\rho^2)}$$

Este modelo solo sirve para cuando los caudales son idénticamente distribuidos, es decir, tienen igual varianza y media, lo cual solo es posible observar en los caudales medios anuales, cuando se necesita generar flujos para estaciones, meses u otras subdivisiones, se requiere añadir algunos términos más al modelo.

$$Q_i = \bar{Q}_j + \rho_j \frac{\sigma_j}{\sigma_{j-1}}\left(Q_{i-1} - \bar{Q}_{j-i}\right) + t_i\sigma_j\sqrt{1-\rho_j^2}$$

Cada caudal producido en un intervalo "$i$" es correspondiente a la serie "j" y de igual manera, el caudal del intervalo "$i-1$" ocurre en la serie "$j-1$". Además, se añade las desviaciones estándar de las series a la gradiente de la ecuación para tomar en cuenta de mejor manera a esta característica.

- *Uso de logaritmos*

La suma de los componentes de la ecuación de Thomas y Fiering, para el caso de que exista bastante dispersión entre los resultados de un mismo día, mes, año o cualquier manera de discretización, pueden dar lugar a caudales negativos. Además, uno de los supuestos del modelo es que los caudales son normalmente distribuidos, los cuales, en la realidad y para intervalos de tiempo cortos, generalmente se asemejan más a una función exponencial con gran asimetría hacia la derecha. Para evitar estos problemas, puede hacerse uso de la transformación:

$$Y = \ln(Q - \tau)$$

La ecuación de Thomas y Fiering se convierte en:

$$Y_i = \bar{Y}_j + \rho_j \frac{\sigma_j}{\sigma_{j-1}} \left(Y_{i-1} - \bar{Y}_{j-i}\right) + t_i \sigma_j \sqrt{1 - \rho_j^2}$$

Para lo cual se debe calcular nuevos parámetros estadísticos para la nueva variable transformada "Y" $(\bar{x}, \sigma, \rho)$ y donde incluso se puede establecer un caudal mínimo $(\tau)$, a partir del cual los resultados nunca serán menores. La ecuación se aplica de la misma manera y al final se despeja de la ecuación que relaciona ambas variables los caudales:

$$Q = \tau + e^Y$$

Por el motivo de que la variable "Y" está expresada en números reales,

$$-\infty < Q < \infty$$

A pesar de obtener valores negativos de "Y", al ser el exponente de un número positivo como "e", los caudales no pueden ser negativos.

$$0 < e^Q < \infty$$

Otra ventaja del uso de logaritmos es que proporcionan mayor sensibilidad frente a caudales bajos, debido a que cambios en estos valores son cambios relativamente grandes en sus logaritmos. Por ello puede resultar útil en sitios que tengan sequías frecuentes.

- *Variaciones*
    - *Matalas*

La mayoría de modelos existentes utilizan la información de un solo río, pero en el caso de tener los datos de un sistema de ríos efluentes, la opción más lógica es usar modelos multisitio, debido a que las condiciones, durante el mismo instante de tiempo en sitios cercanos, generalmente están altamente relacionadas. (Fiering & Jackson, 1971)

Matalas fue el primero en proponer esta idea en 1967, para que posteriormente en 1968 Young y Pisano la desarrollaran (Young & Pisano, 1968). Al calcular los parámetros estadísticos y las interrelaciones entre cada una de las series históricas, se crean matrices que están en función de estos valores, además de números aleatorios correspondientes a cada

caudal a generarse. Con esto se genera un proceso de Markov similar al modelo de Thomas y Fiering pero que genera simultáneamente los caudales de varios ríos.

- *Aumento en la persistencia*

Con el argumento de que el agua absorbida por la superficie también es liberada durante ciertas temporadas, se puede aludir que los residuos del modelo pueden ser explicados al añadir otro parámetro que no se encuentre relacionado con el dato inmediatamente anterior, como puede ser:

- El promedio de los caudales del anterior mes
- El caudal medio diario del mismo día en el año anterior
- El caudal medido a la misma hora en el día anterior

Por lo que se asume que el modelo tiene una memoria más larga y la parte determinística de la ecuación, para el caso de tomar en cuenta también los datos de 2 meses atrás, la ecuación tendría la forma:

$$Q_i = \bar{Q}_j + \rho_{j\_j-1}\frac{\sigma_j}{\sigma_{j-1}}\left(Q_{i-1} - \bar{Q}_{j-i}\right) + \rho_{j\_j-2}\frac{\sigma_j}{\sigma_{j-2}}\left(Q_{i-2} - \bar{Q}_{j-2}\right) + t_i\sigma_j\sqrt{1-\rho_j^2}$$

- *Necesidad del componente aleatorio*

Por si sola, la parte determinística es inadecuada para describir la correlación entre dos conjuntos de valores. Por ejemplo, al momento de generar el caudal medio mensual de febrero, si el caudal de enero $Q_{i-1}$ coincide o es muy parecido con el promedio de los datos para el mes de enero $\bar{Q}_{j-i}$**,** la ecuación por sí sola, generaría de vuelta el promedio del mes de febrero $\bar{Q}_j$, esto a su vez, generaría el promedio de marzo y así hasta terminar la secuencia, por ello es necesario un componente que simule las variaciones naturales de los flujos que no se toman en cuenta por la regresión. (Arselan, 2012)

- *Limitaciones*

    - *Frecuencia y horizonte de pronóstico*

Si es necesario realizar pronósticos dentro de un horizonte muy grande, incluso si se cuenta con un modelo estocástico muy complejo, el ruido blanco introducido en él volverá casi imposible pronosticar los últimos valores de la serie. Conforme avanzamos en el tiempo, la diferencia entre el registro y el modelo se volverá más notoria. (Hyndman & Athanasopoulos, 2018)

Es una mejor alternativa que el modelo se vaya actualizando en el tiempo, esto es, que los parámetros cambien, en lo posible, de manera persistente, debido a que es poco realista pensar que las características de la serie se mantengan en el tiempo.

Los modelos de medias condicionadas y todas sus variaciones, una vez removidos los componentes de tendencia y estacionalidad, sirven para capturar las fluctuaciones estacionarias en el corto plazo, por lo que su uso a largo plazo causará pronósticos poco realistas.

    - *Convergencia*

Para series realmente estacionarias, por ejemplo, los modelos autoregresivos de orden 1 como la ecuación de Thomas-Fiering, los datos tienden a converger en la media de la serie.

Este fenómeno, probablemente no pueda ser apreciado si el software realiza diferenciaciones por estacionalidad y tendencia, para el caso del modelo Thomas-Fiering, al realizarse varias iteraciones, los pronósticos convergerán en la media de los meses, días o semanas.

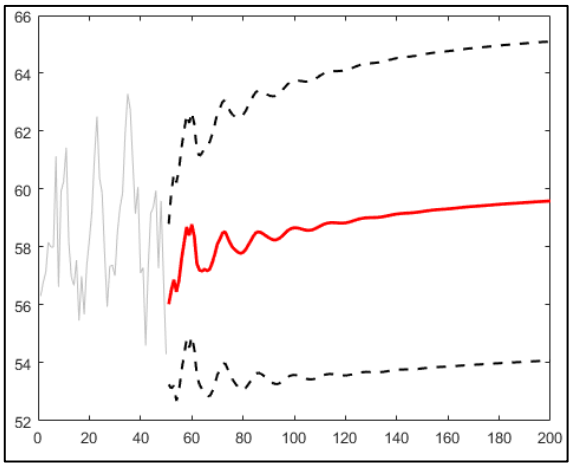


Figura 2.2-24. Convergencia de una serie estacionaria en su media y sus límites inferior y superior. Copyright 2019 por MathWorks. *Reimpreso con permiso*

## 2.3. Nociones de hidrología

### 2.3.1. Hidrología

Es la ciencia que estudia el agua, su ocurrencia, circulación, distribución, sus propiedades, sus relaciones con el medio ambiente y con los seres vivos. (Monsalve Sáenz, 1999)

**Aplicaciones de la hidrología**

- Elección de las fuentes de abastecimiento de agua
- Diseño y construcción de obras hidráulicas
- Estudio del nivel freático
- Irrigación
- Regulación de cursos de agua y control de inundaciones
- Control de polución debido a aguas servidas y control de erosión
- Navegación
- Aprovechamiento hidroeléctrico
- Recreación y conservación del medio ambiente
- Preservación de la vida acuática.

### 2.3.2. Cuenca hidrográfica

Es un área definida topográficamente, de la cual la mayor parte del agua proveniente de las precipitaciones es drenada por un curso de agua o sistema de cursos de agua, la parte restante se estanca en las depresiones de la cuenca, o eventualmente se infiltra o evapora. (Monsalve Sáenz, 1999)

- **Características físicas de una cuenca**

    - **Área de drenaje:** proyección horizontal definida dentro de su perímetro topográfico.
    - **Forma de la cuenca:** se refiere al relieve que se encuentra conformado por montañas, quebradas, llanuras y mesetas, y que, dependiendo del tipo de suelo, determina el tiempo de concentración, el tiempo que les toma a todos los puntos de una cuenca en contribuir al sistema de drenaje y llegar a la salida de la cuenca.
    - **Tipo de suelo:** los suelos arcillosos tienden a ser casi impermeables, a diferencia de los suelos arenosos, que absorben gran parte del agua, el conocer este detalle permite estimar el volumen de agua de la superficie de la cuenca que terminará en los ríos.
    - **Sistema de drenaje:** conformado por los diferentes afluentes y la corriente principal que, en conjunto, drenan el agua de una cuenca.
    - **Orden de las corrientes de agua:** grado o nivel de las ramificaciones de los cursos de agua dentro de la cuenca. Ordenes de menor valor implican pequeños tributarios, mientras que los de mayor valor se refieren a las corrientes principales.

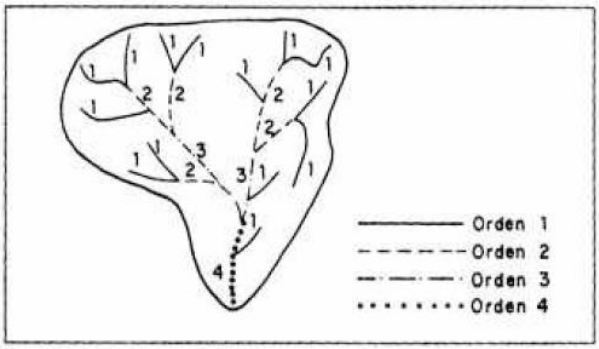


Figura 2.3-1. Designación del orden de las corrientes de una cuenca. (Monsalve, 1999)

- **Pendiente de una cuenca:** es una magnitud que es calculada de acuerdo a la distancia relativa general entre las curvas de nivel de una cuenca, es un factor que controla la velocidad y el tiempo de drenaje de los volúmenes de escorrentía en una cuenca.
- **Pendiente de la corriente principal:** es proporcional a la velocidad del flujo sobre este curso, en conjunto con la rugosidad y la forma del lecho.

El conocer cuál es la cuenca que contiene a un río, puede dar una idea de cómo las acciones antrópicas y naturales que sucedan sobre esta, afectan en a la cantidad y calidad al flujo de agua. (USGS, 2019)

### 2.3.3. Escorrentía

Durante las precipitaciones, se forman riachuelos, canales o flujos sobre pequeñas grietas en el terreno, que transportan agua y que a su vez alimentan a cuerpos de agua más grandes como ríos y lagos. En hidrología, se suele expresar como una lámina de agua equivalente y de cierta altura, que ocuparía este volumen de agua al extenderlo en la superficie de la cuenca. A pesar de que las lluvias aportan con la mayoría del flujo, la escorrentía, también se abastece, en gran parte, por el constante intercambio que tiene ésta con el agua subterránea que se encuentra dentro del suelo en contacto.

La construcción de obras cercanas a las cuencas, además de implicar la eliminación de la vegetación en las inmediaciones al río, significan el reemplazo de las superficies naturales con superficies artificiales más impermeables, que alteran el ciclo natural del agua, y que, para el caso de la escorrentía, esta se ve aumentada, debido a que una fracción menor de la precipitación se infiltra en el terreno.

Los ríos son cursos de agua naturales, generalmente de agua dulce, que tienen su origen en otros cursos más pequeños, el agua subterránea y/o depósitos de nieve y que terminan su recorrido en otros ríos y últimamente en los océanos o en lagos. También puede suceder el caso de que estos cuerpos longitudinales de agua se sequen antes de conectarse con otros cuerpos de agua.

El origen del movimiento y, por lo tanto, la energía contenida en los ríos, se encuentra en la gravedad, el agua en la tierra terminará ocupando valles y depresiones en las zonas bajas, debido a que todas las partículas son atraídas hacia el centro de la tierra, pero como no pueden tomar una trayectoria directa, se desplazan con pendientes negativas.

El principal fenómeno que influye sobre el nivel del agua en un río es la precipitación, que incluso si sucede muy lejos, aguas arriba de cierta sección transversal del río, tendrá una incidencia sobre él.

Contribuyen con el ciclo natural del agua, al rellenar depósitos subterráneos, alimentar otros ríos y contribuir con la humedad atmosférica debido a que gran cantidad de agua también se evapora y es contenida en el ambiente.

Los ríos son invaluables para la vida animal y vegetal, el transporte de personas y recursos, la generación de electricidad, el manejo de desechos y riego de cultivos, pero principalmente para el suministro de agua de las poblaciones.

La planificación del manejo de los ríos, dirigidos hacia sectores urbanos y rurales, es necesario para obtener el máximo beneficio de este limitado recurso. Los modelos de pronóstico de caudales de ríos, inicialmente desarrollados para la optimización en el manejo de reservorios, también sirven para la administración de otras actividades como:

- Control de inundaciones y medidas de la seguridad de obras
- Suministro de agua
- Riego de cultivos
- Ámbitos ambientales

De acuerdo con su permanencia Monsalve (Monsalve Sáenz, 1999), los clasifica como:

- Perennes: existen constantemente, el agua subterránea en contacto con estos cursos mantiene niveles invariables, debido a que tiene volúmenes de agua que infiltran en todo momento.

- Intermitentes: estos cursos y los niveles de los depósitos de agua subterránea que se encuentran comunicados con ellos, solo se dan en épocas lluviosas.
- Efímeros: similares a los intermitentes, pero solo se presentan durante fuertes tormentas.

### 2.3.4. Hidrograma

Es una representación que muestra el cambio de una variable hidrometeorológica a través del tiempo. Es usualmente utilizada para representar la evapotranspiración, la infiltración, la escorrentía y el flujo del agua subterránea, entre otra clase de fenómenos. (Woodard & Curran, 2006)

Al tomar los registros de los hidrogramas durante periodos de lluvia, es posible relacionarlos con las mediciones pluviométricas para separar al caudal en los ríos en sus componentes: aquella que tiene como fuente a la lluvia y aquella que no, ambos elementos tienen leyes físicas y mecánicas diferentes que los gobiernan y su estudio por separado sirve para caracterizar de mejor manera a la cuenca.

Este tipo de descomposición permite hacer inferencias de los caudales que pasan por las secciones de los ríos estudiadas, debido a que, si no existen mediciones directas de los caudales, se puede hacer un estimado a partir de los registros pluviométricos.

Otra manera de estudiar los hidrogramas es hacer un análisis de las frecuencias de las diferentes magnitudes que han tomado los caudales durante su historia registrada, esto permite conocer el valor de los máximos y mínimos observados, la proporción con la que se presentan y de esta manera, tomar decisiones con respecto al aprovechamiento de este recurso. (Monsalve Sáenz, 1999)

### 2.3.5. Hidroelectricidad

Es electricidad creada al aprovechar el agua en movimiento en centrales hidroeléctricas, es la fuente más grande de energía renovable y es un mercado que en los últimos años ha visto un enorme crecimiento. La hidroelectricidad conforma actualmente alrededor del 17% de la

producción mundial y así como un 70% de las fuentes renovables, se espera que su aporte aumente en un 3,1% de la producción mundial cada año en los próximos 25 años. (REN21, 2019)

Las centrales han sido usadas principalmente en Europa, Norteamérica y Asia, pero la proporción de las construcciones que aprovechan este recurso se ha visto incrementada en América Latina debido a su conveniencia; el uso de las hidroeléctricas a pequeña escala es una de las maneras de producción de electricidad con una mejor relación costo-beneficio para abastecer a zonas rurales y semirrurales.

Una desventaja de construir hidroeléctricas a gran escala es su impacto sobre el ambiente, al inundar grandes áreas y cambiar el curso de los sistemas naturales, además de posibles daños sociales al desplazar a las personas y afectar a sus bienes. A pesar de que, además de esto, las centrales hidroeléctricas necesitan de emplazamientos que cumplan con condiciones muy específicas, al contrario de las plantas que utilizan combustibles fósiles, la hidroelectricidad es la fuente más eficiente de generación de electricidad con los medios actuales, alcanzando una eficiencia de hasta un 95%, mientras que las demás solo pueden alcanzar un máximo de 45%. (Electropedia, 2005)

#### *2.3.5.1. Manejo de modelos hidrológicos por parte de centrales hidroeléctricas*

Las predicciones en las centrales hidroeléctricas pequeñas suelen referirse a pronosticar los niveles y caudales del río con el que se encuentran en contacto, debido a que estas dos variables son proporcionales a la cantidad de energía generada y son fenómenos naturales que son difícilmente modificables por las personas. La cantidad de agua en un río se encuentra influenciada, como ya ha sido explicado, por un fuerte componente de aleatoriedad proveniente de la gran heterogeneidad de la cuenca.

Es un proceso importante en el manejo de las centrales de toda magnitud y en especial en aquellas que no tienen embalse de reserva, esto se debe a que la producción de energía en estas obras es intermitente. La malla de electricidad de una región requiere certidumbre durante todo momento, debido a que debe satisfacer los requerimientos del mercado

eléctrico, un componente esencial en el desarrollo del país sobre el que se basan las actividades de la población y las industrias.

Las predicciones realizadas por la intuición, hechas por una gran proporción de las centrales pequeñas en zonas rurales y que pronostican los caudales en base a días anteriores y lo observado en los correspondientes meses, tendrán una gran fluctuación sobre los registros, que se traducirá en una inestable generación de electricidad si no se realizan las acciones adecuadas.

Al contrario, al tener redes hidrometeorológicas públicas y privadas, que logren tomar varios registros y permitan generar modelos compuestos y multivariables, que sería lo esencial en estos casos, o al menos contar con una metodología que logre predecir caudales al corto y mediano plazo, será posible mejorar los ingresos netos de los proyectos al planificar las actividades con antelación.

Si no se cuenta con mecanismos que logren medir correctamente los caudales correspondientes a una central, también se ha apreciado en la literatura (Cheng , Liao, Liu, & Li , 2015), que es posible realizar análisis de correlación, de lo apreciado en otras estaciones de la región, para contar con un buen estimador de los flujos.

## 2.4. MATLAB

### 2.4.1. Generalidades

Es un entorno y lenguaje de programación desarrollado por Mathworks y que se basa en la computación de matrices. El programa también tiene adjunto un paquete de aplicaciones útiles para ingenieros y otros investigadores en diversas áreas de la ciencia, entre ellas:

- Aprendizaje de maquinas
- Optimización de funciones
- Sistemas de control de los dispositivos y maquinas autómatas en industrias
- Procesamiento de señales e imágenes
- Finanzas

- Biología
- Desarrollo de otras aplicaciones

Su principal ventaja, y el motivo de que sea útil para motivos académicos, es el que permite al usuario programar de una manera intuitiva y entender de esta manera, los conceptos sobre los que se fundamentan las teorías que permiten aplicar su profesión en un entorno laboral. Así, librándolo de utilizar programas costosos que realizan todo de manera automatizada y no dejan conocer sus supuestos o los algoritmos que utilizan y que pueden llevar a cálculos erróneos, debido a que el usuario no se encuentra obligado a adquirir esa capacidad de interpretación.

La exportación y el procesamiento de una gran cantidad de datos se manejan con las funciones incorporadas en el programa, además de tener una interfaz gráfica que sirve para visualizarlos con una gran cantidad de preferencias.

### 2.4.2. Aplicaciones en hidrología

- Resolución de problemas de programación lineal para la estimación de parámetros hidrológicos que tienen condiciones de desigualdades y restricciones.
- A través de la función fitdist () y otras, puede encontrar la mejor distribución de probabilidad teórica que se acople a los diagramas de frecuencias de distintas variables.
- Integrar archivos *raster*, como los trabajados por el programa ARCGIS, para la generación de modelos hidrológicos complejos, debido a que cada uno de estos está conformado por pixeles que contienen información georreferenciada como la altitud y muestreos cercanos.

### 2.4.3. Econometrics Toolbox App


*Figura 2.4-1. Ubicación de la herramienta en la pestaña de aplicaciones de MATLAB*

Para el cálculo de demandas de productos, flujo de turistas, pasajeros en los sistemas de transporte, los mercados de valores y en el campo de la hidrología, se ha apreciado como ciertos modelos, que en un principio fueron desarrollados por la econometría, la rama de la economía conectada con la estadística para la predicción de sistemas económicos, han sido bastante exitosos al momento de aplicarlos a otras disciplinas.

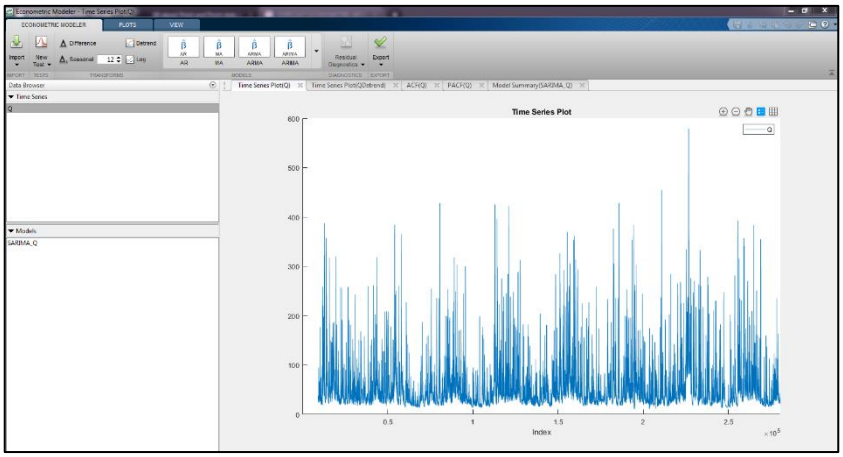


*Figura 2.4-2. Visualización de una serie de tiempo en la aplicación.*

Es un conjunto de herramientas integrado a MATLAB que contiene funciones para el ajuste de series en el tiempo por medio de los modelos ya explicados, además de otros, con sus respectivas variantes, y además de ello, le permite hacer diagnósticos como test de hipótesis y estacionariedad. Sin necesidad de tener un conocimiento previo de programación.

### 2.4.4. Neural Network Time Series App


*Figura 2.4-3. Icono en la tabla de aplicaciones de MATLAB*

Es una herramienta incluida en MATLAB para la generación de pronósticos usando neuronas artificiales a través de modelos no lineales. Existen 3 alternativas posibles:

a) NARX: son modelos autoregresivos no lineales que pronostican los valores de una serie "y", a partir de valores pasados de "y" y de otra variable externa "x".

$$y_t = f(y_{t-1} \dots y_{t-d}, x_{t-1} \dots x_{t-d})$$

b) NAR: son modelos autoregresivos no lineales que pronostican los valores de una serie "y", a partir de valores pasados de "y".

$$y_t = f(y_{t-1} \dots y_{t-d})$$

c) Modelo no lineal de entrada-salida: que pronostican los valores de una serie "y", a partir de valores pasados de una serie externa "x". Son menos recomendados que los anteriores 2.

$$y_t = f(x_{t-1} \dots x_{t-d})$$

# 3. APLICACIÓN AL CASO EN ESTUDIO

## 3.1. Localización y Descripción del Proyecto hidroeléctrico Topo

La central hidroeléctrica se encuentra ubicada en la provincia de Tungurahua, cantón Baños, parroquia Río Negro y que recibe la corriente del río con el mismo nombre que la central. La empresa Ecuagesa S.A. es la concesionaria designada por la CONELEC y es la que obtuvo y proporcionó los registros de los últimos 8 años, que han sido procesados y proyectados en un futuro en el presente documento.

Las coordenadas de la captación y la casa de máquinas respectivamente son: N 9849200, E 810400 y N 9847800, E 810000.

El proyecto tiene un reservorio y genera una potencia de 29 MW a partir de una captación de alrededor de 20$m^3$/s del caudal del Río Topo, la central comenzó sus operaciones el 21 de enero del 2017.

Figura 3.1-1 Reservorio y presa de la central. Copyright 20202 por Ecuagesa. *Reimpreso con permiso*


*Figura 3.1-2. Localización del proyecto*

## 3.2. Descripción de la cuenca

La cuenca sobre la que se encuentra el río Topo, según la subdivisión oficial del INAMHI, es la cuenca del río Pastaza. Es la tercera cuenca más grande del Ecuador y una de las fuentes de agua dulce más grande del mundo, ocupa un área de 32182,34 km2 y dentro de ella se encuentra alrededor de un 11,28% de la población del Ecuador (Yépez, 2015). Tienen un gran potencial hídrico que puede ser aprovechado para la construcción de obras de este tipo

y que actualmente, el gobierno nacional, ha aprovechado y ha decidido construir proyectos en varios sitios adecuados de esta cuenca.

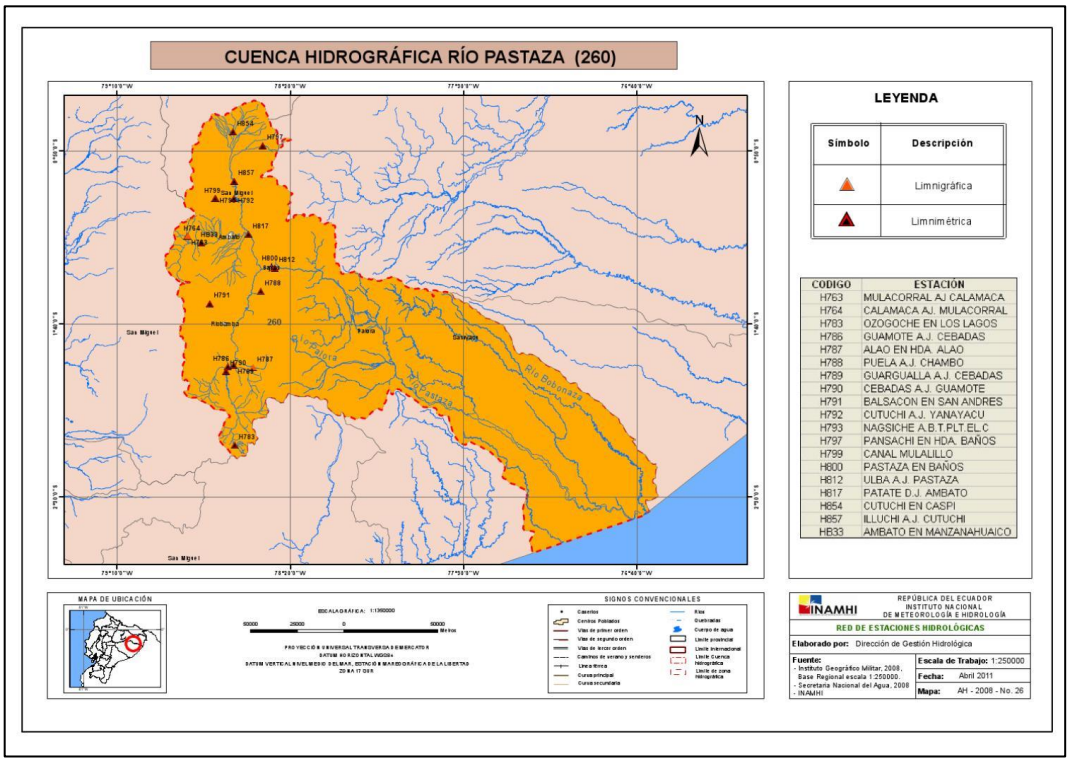


Figura 3.2-1. Cuenca sobre la que se ubica la central. Copyright 2011 por INAMHI. *Reimpreso con permiso*

## 3.3. Flujograma para la aplicación del modelo Thomas-Fiering

Es necesario mencionar que, para el caso de los años bisiestos, se consideró necesario omitir los datos del 29 de febrero de estos años al momento de calibrar el modelo; los coeficientes de correlación, media y desviación estándar en la ecuación generadora de caudales, no serían muy significativos, debido a que solo existen 2 años bisiestos disponibles que contienen registros para este día.

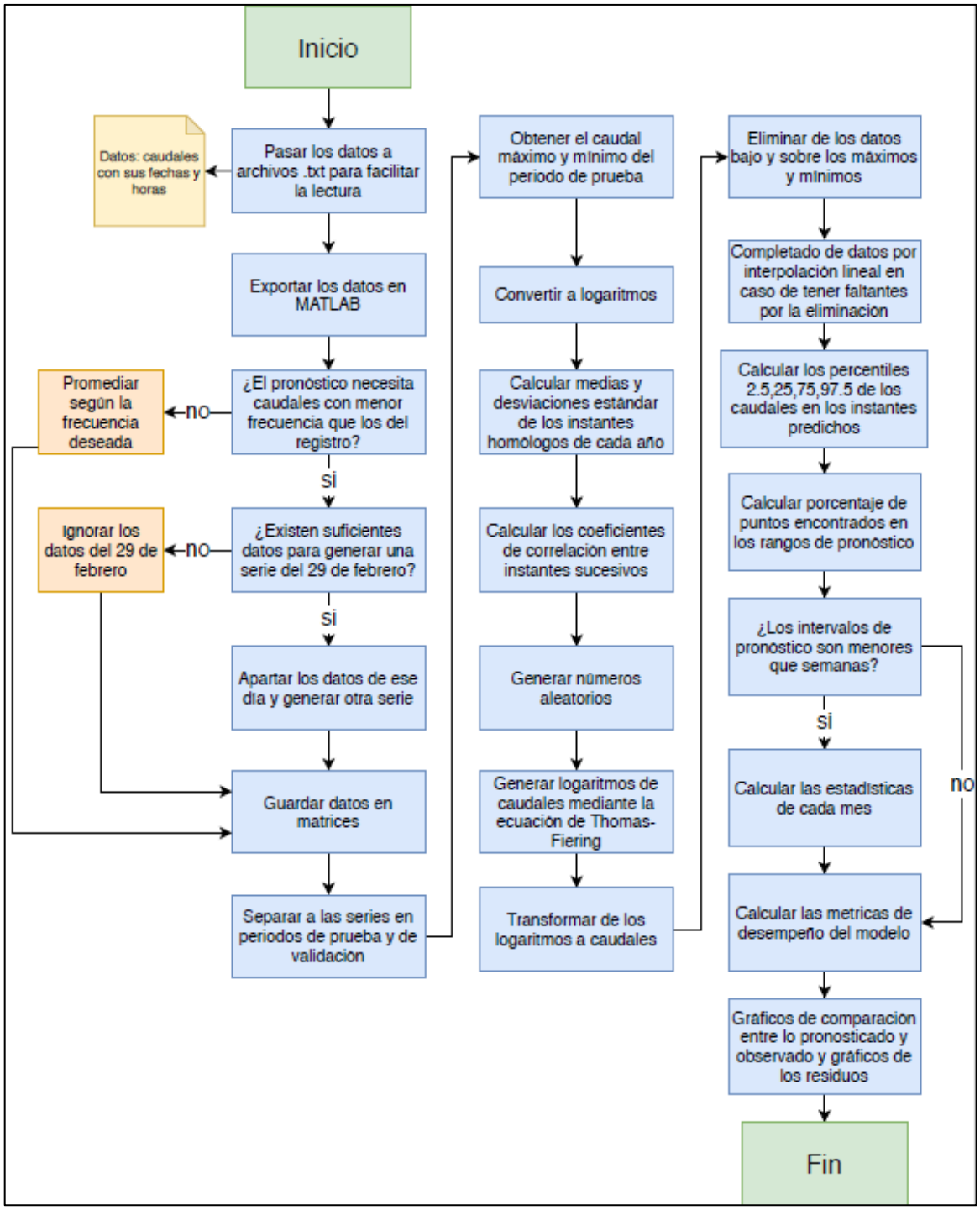


*Figura 3.3-1. Pasos para la aplicación del modelo Thomas-Fiering*

## 3.4. Flujograma para la aplicación de modelos Box-Jenkins

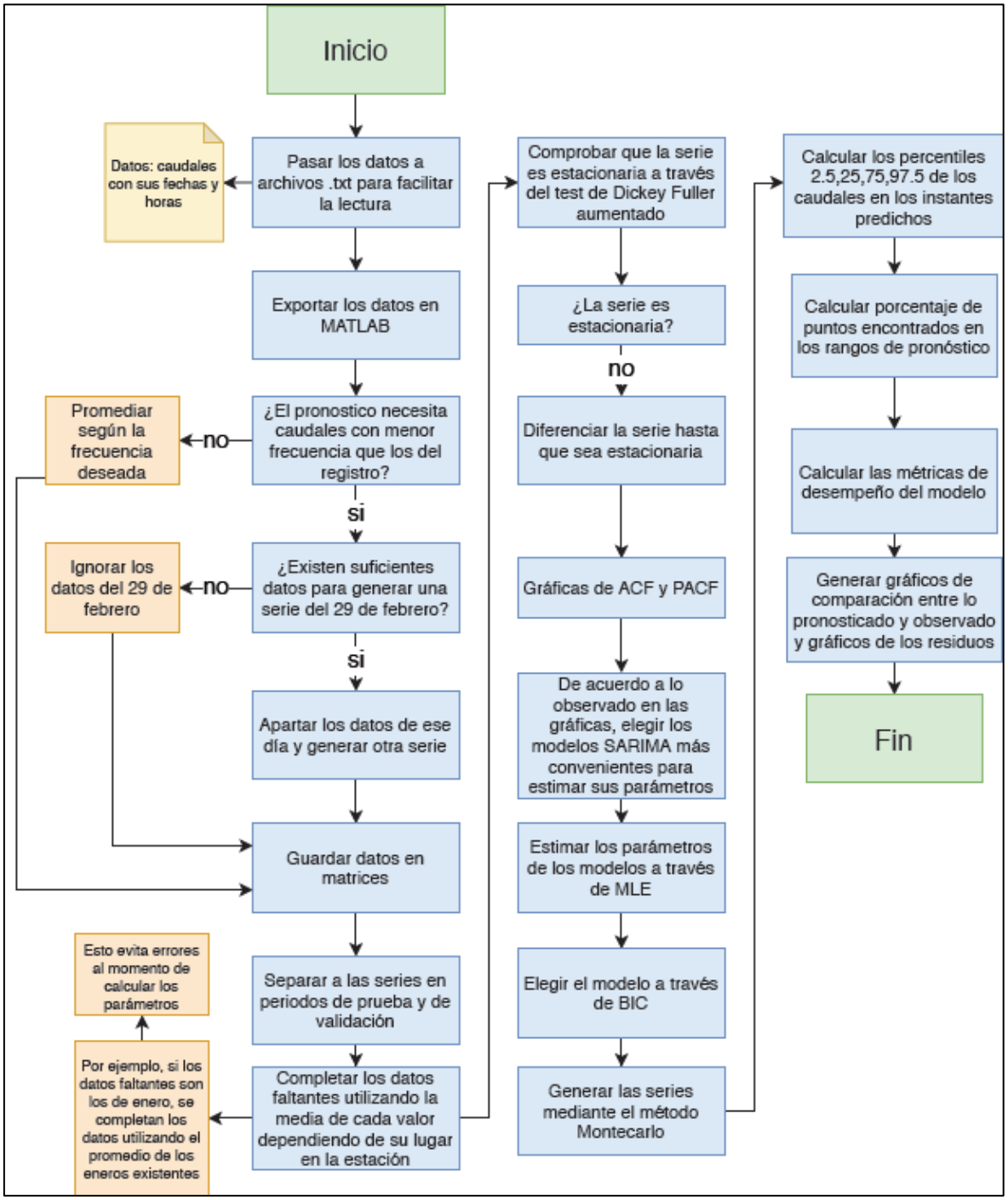


*Figura 3.4-1. Pasos para la aplicación de modelos ARIMA*

## 3.5. Series trabajadas

A partir de las observaciones con intervalos de 15minutos, se promedian los valores de la serie para observar si intervalos más largos vuelven a la predicción más precisa. La cantidad de datos respectiva para la serie original y las promediadas es:

| **Intervalo de tiempo** | Número de datos | |
|---|---|---|
| | **Periodo de prueba (Años 2011-2017)** | **Periodo de validación (Año 2018)** |
| 15 minutos (original) | 235486 | 35040 |
| 1 hora | 58871 | 8760 |
| 1 día | 2454 | 365 |
| 1 semana | 358 | 53 |
| 2 semanas | 183 | 27 |
| 1 mes | 81 | 12 |

*Tabla 3.5-1. Número de datos en los periodos de validación y prueba para distintos intervalos*

Cada uno de los resultados mostrados para diferentes intervalos y diferentes metodologías fue realizado usando para cada una, una rutina diferente de MATLAB (6 para Thomas-Fiering, 3 para ARIMA), esto se debe a que cada uno de ellos requiere un procesamiento ligeramente diferente de datos debido a que la cantidad de datos y el tamaño de las estaciones varía entre cada una.

La función histogram () se utiliza para graficar la distribución de los caudales para los diferentes intervalos, confirmado o no que se cumplen los supuestos de normalidad en los datos de entrada del modelo Thomas-Fiering.

### 3.5.1. Intervalos de 15 minutos

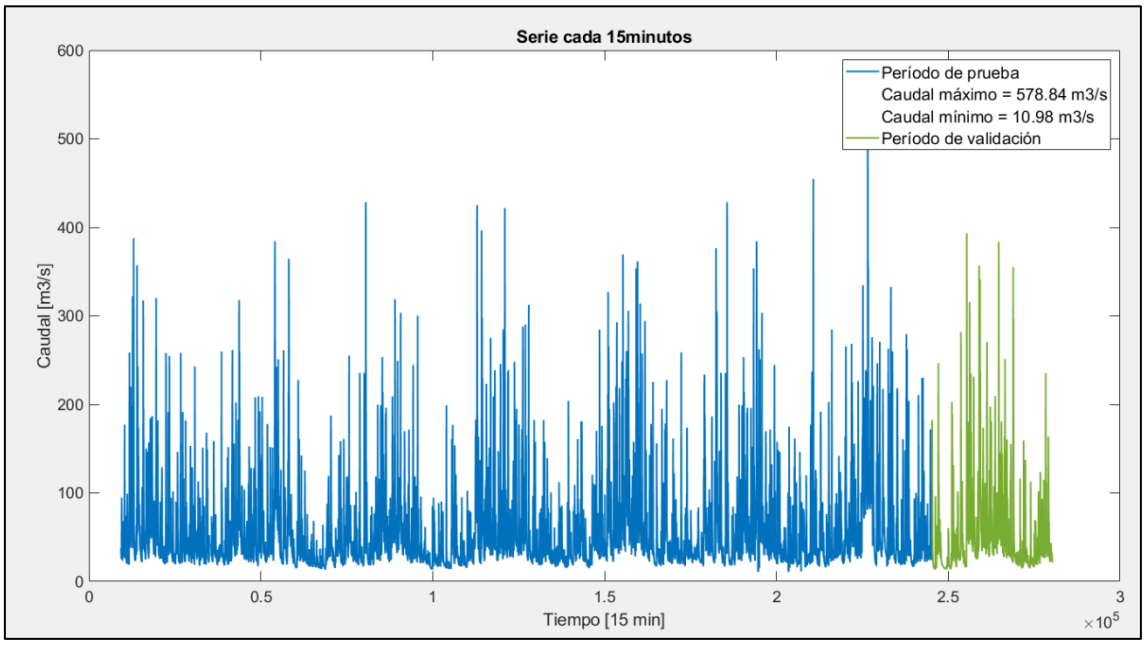


*Figura 3.5-1. Periodo de prueba y validación de la serie cada 15minutos.*

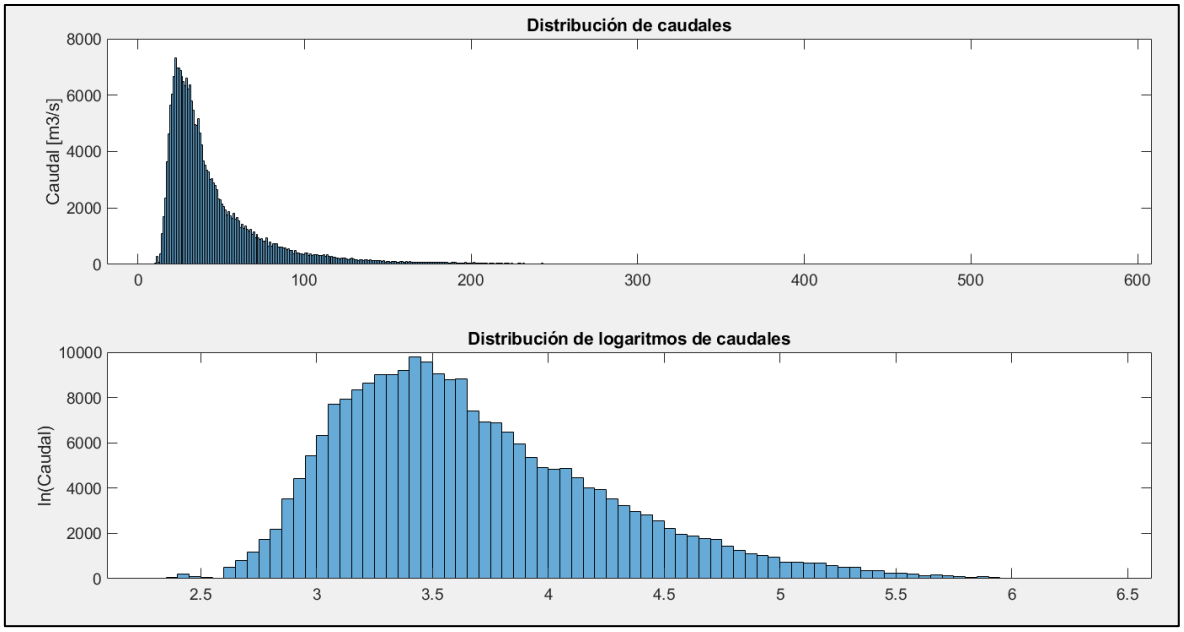


*Figura 3.5-2. Distribución de la serie y de los logaritmos de la serie cada 15minutos*

### 3.5.2. Intervalos de 1 hora

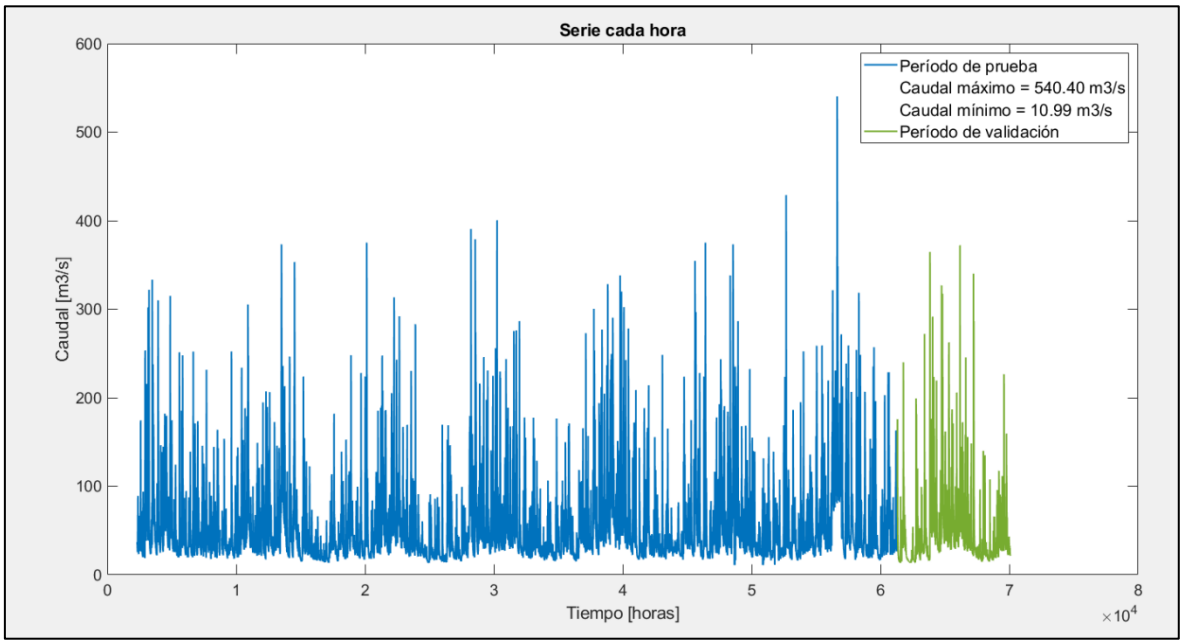


*Figura 3.5-3. Periodo de prueba y validación de la serie cada hora.*

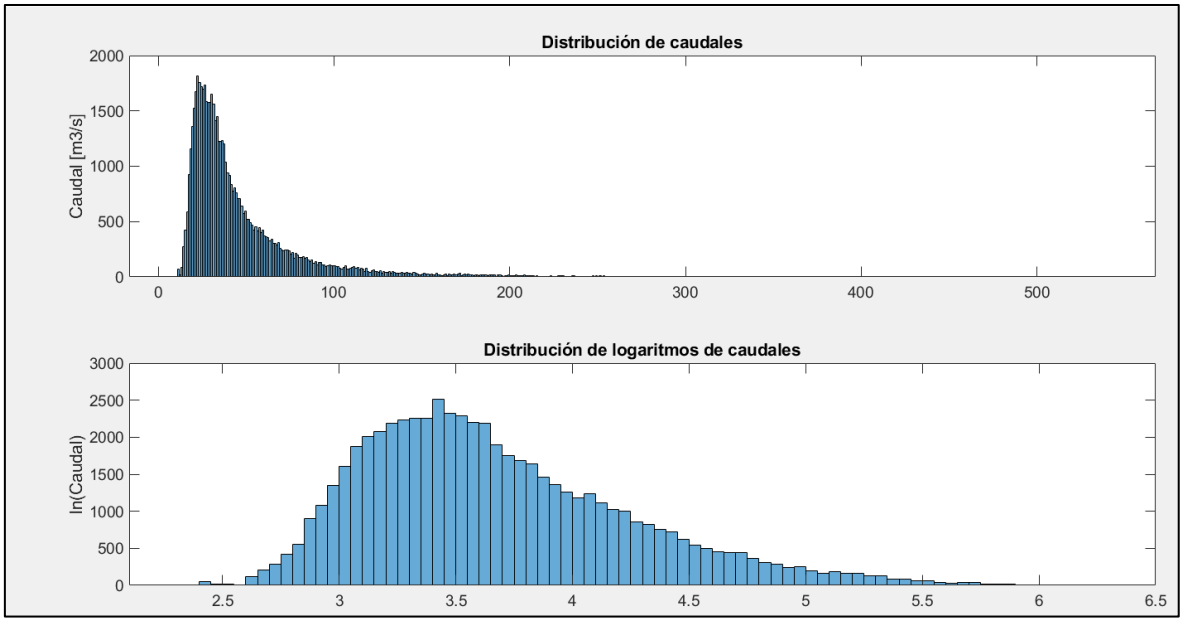


*Figura 3.5-4. Distribución de la serie y de los logaritmos de la serie cada hora*

### 3.5.3. Intervalos de 1 día

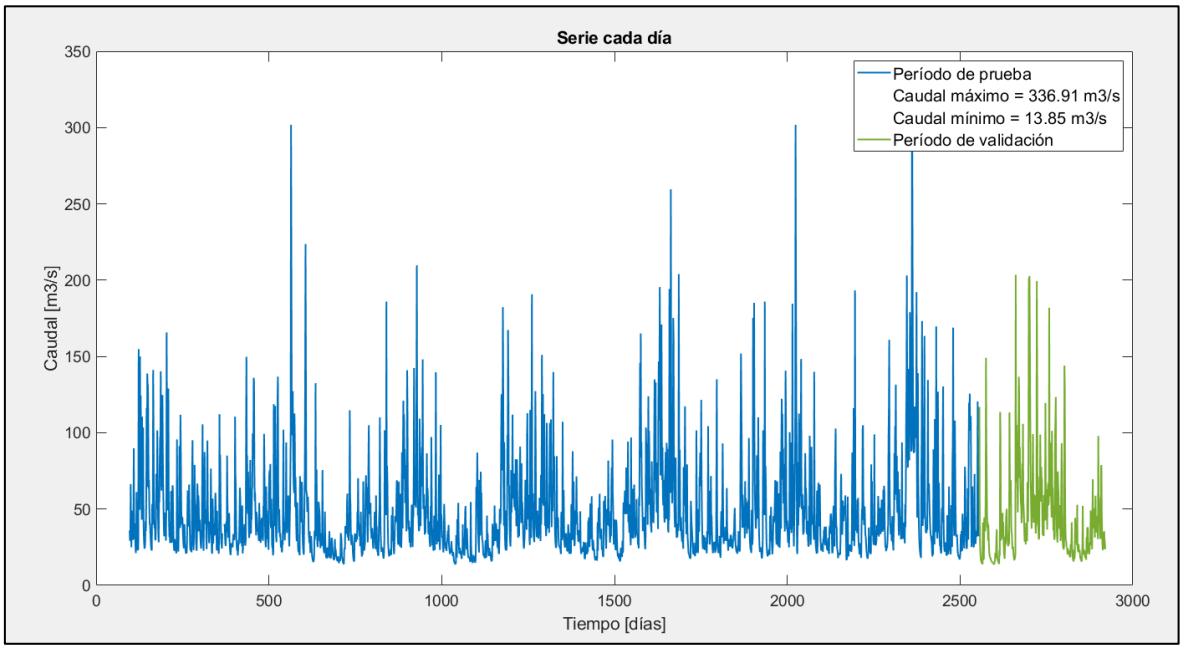


*Figura 3.5-5. Periodo de prueba y validación de la serie cada día.*

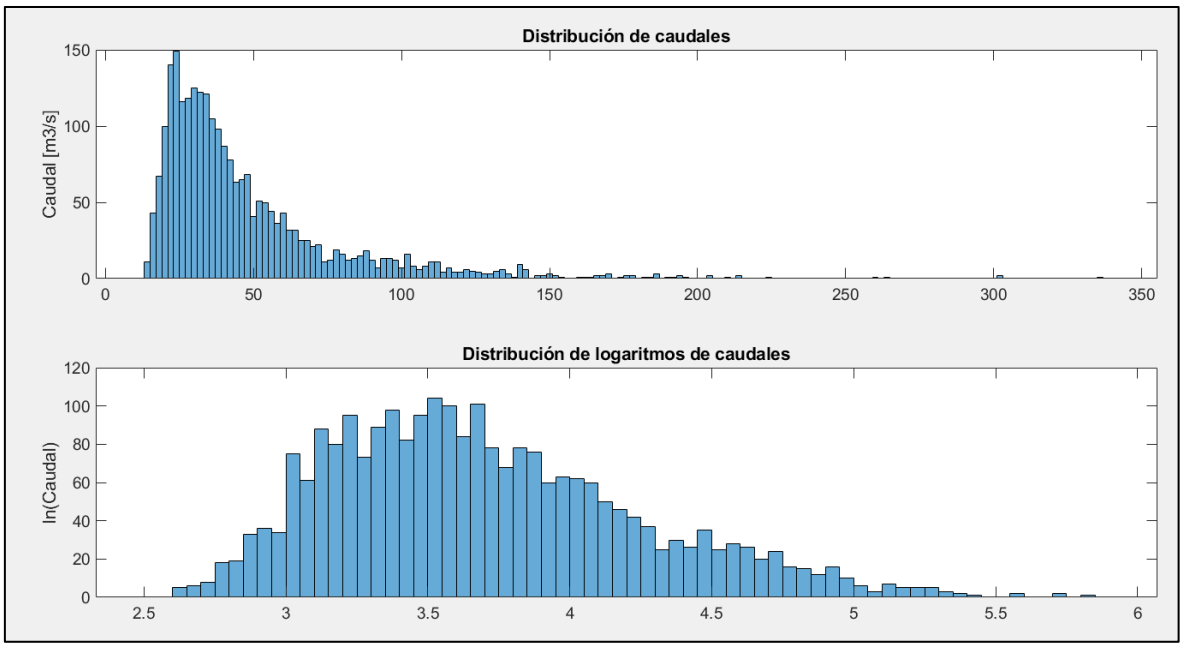


*Figura 3.5-6. Distribución de la serie y de los logaritmos de la serie cada día*

Es posible observar que los caudales, hasta este nivel de discretización, tienen una distribución similar a una función exponencial, lo cual no concuerda con la suposición de normalidad de Thomas-Fiering, para intervalos mayores a estos, los caudales, así como los logaritmos de los caudales tienen una distribución similar a la normal.

### 3.5.4. Intervalos de 1 semana

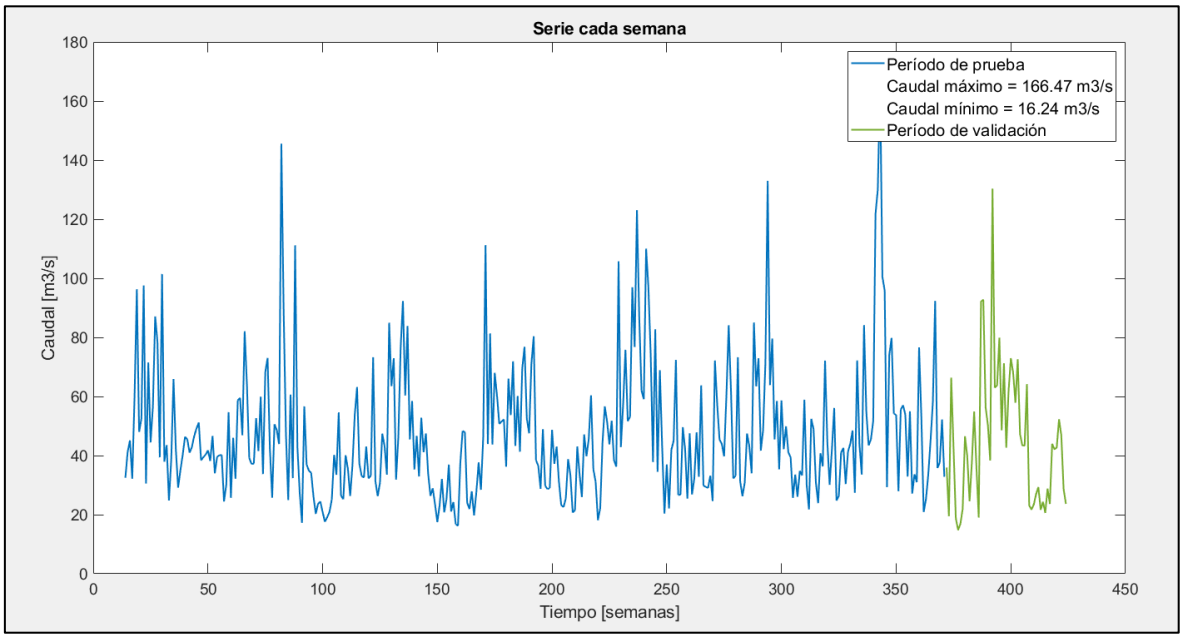


*Figura 3.5-7. Periodo de prueba y validación de la serie cada semana.*

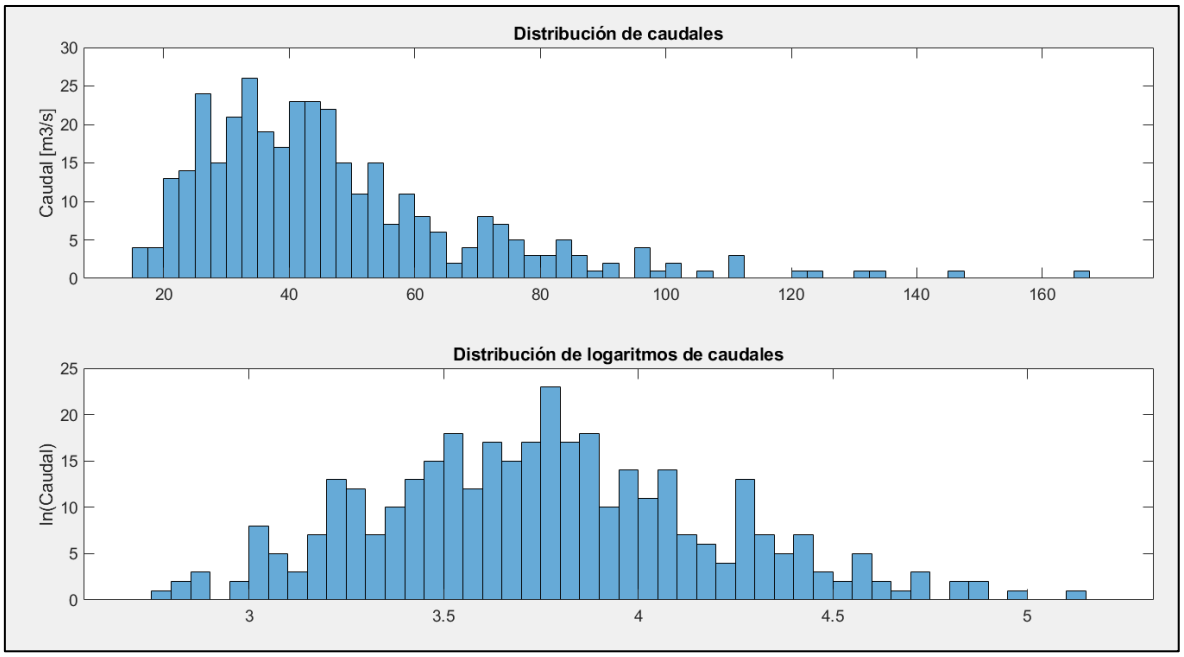


*Figura 3.5-8. Distribución de la serie y de los logaritmos de la serie cada semana*

### 3.5.5. Intervalos de 2 semanas

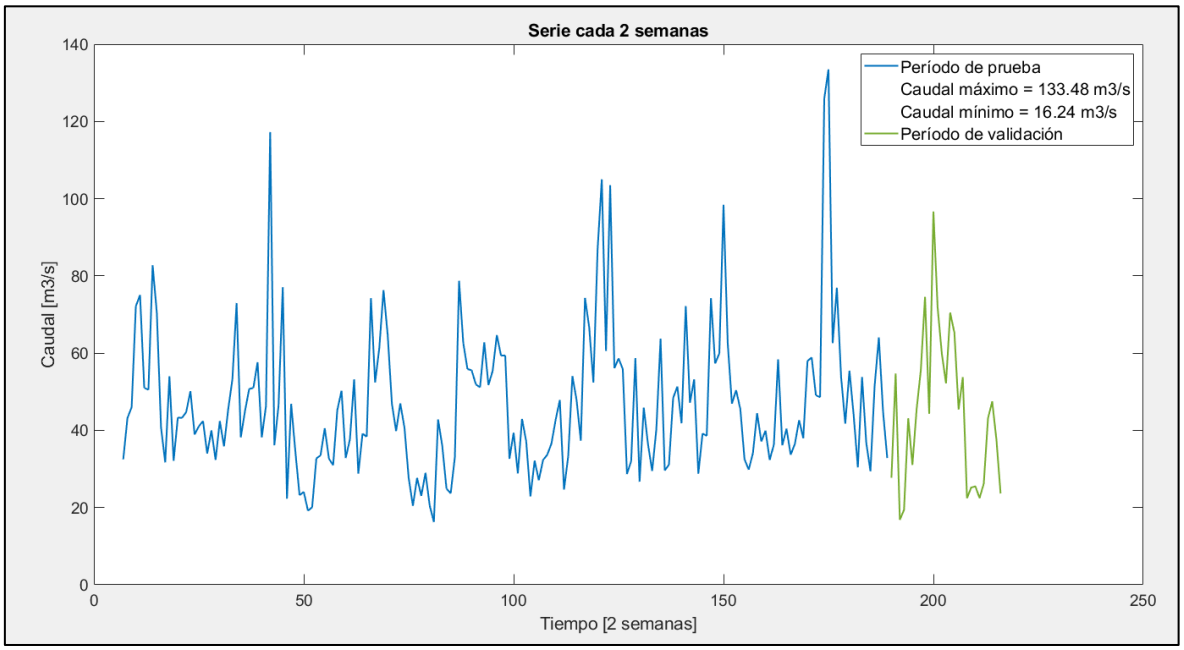


*Figura 3.5-9. Periodo de prueba y validación de la serie cada 2 semanas.*

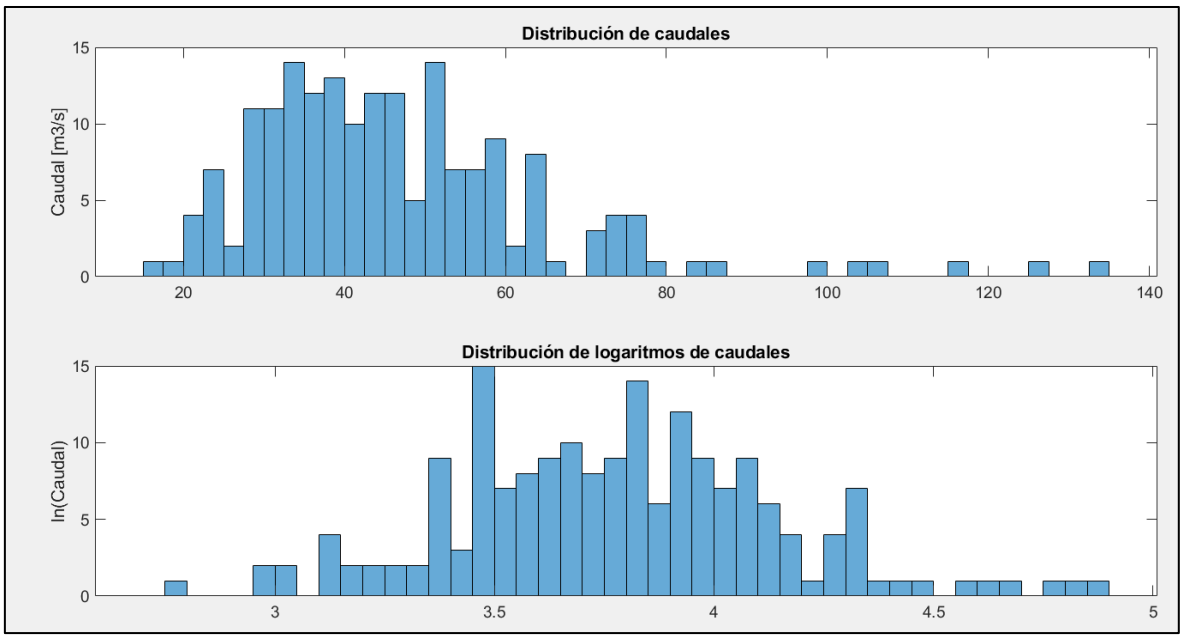


*Figura 3.5-10. Distribución de la serie y de los logaritmos de la serie cada 2 semanas*

### 3.5.6. Intervalos de 1 mes

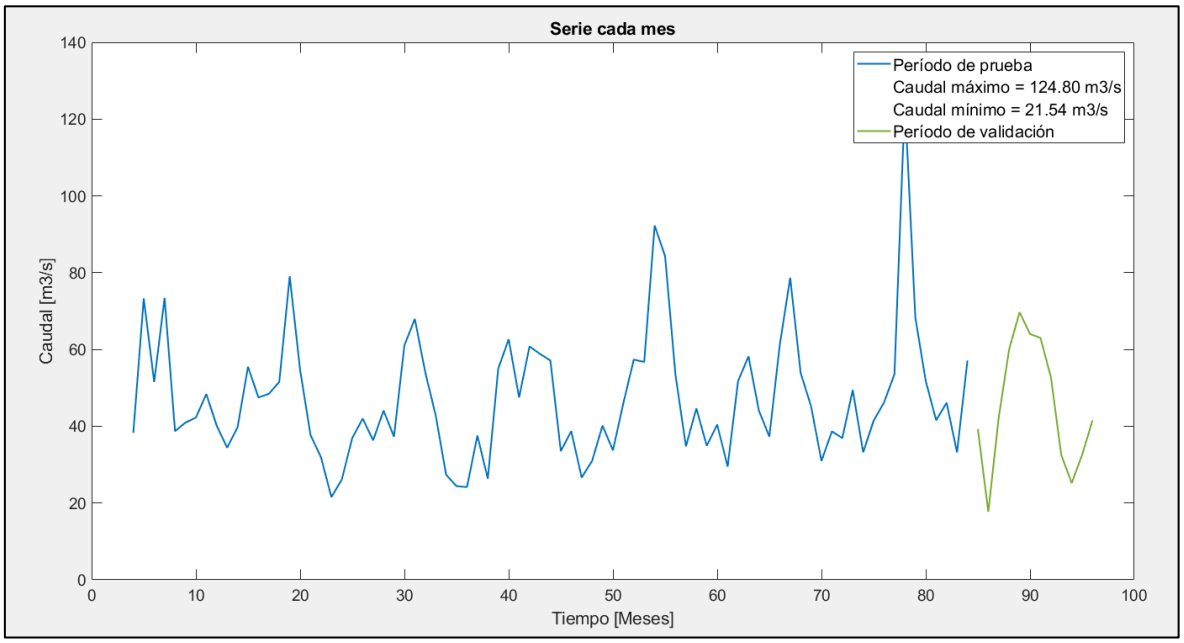


*Figura 3.5-11. Periodo de prueba y validación de la serie cada 15minutos*

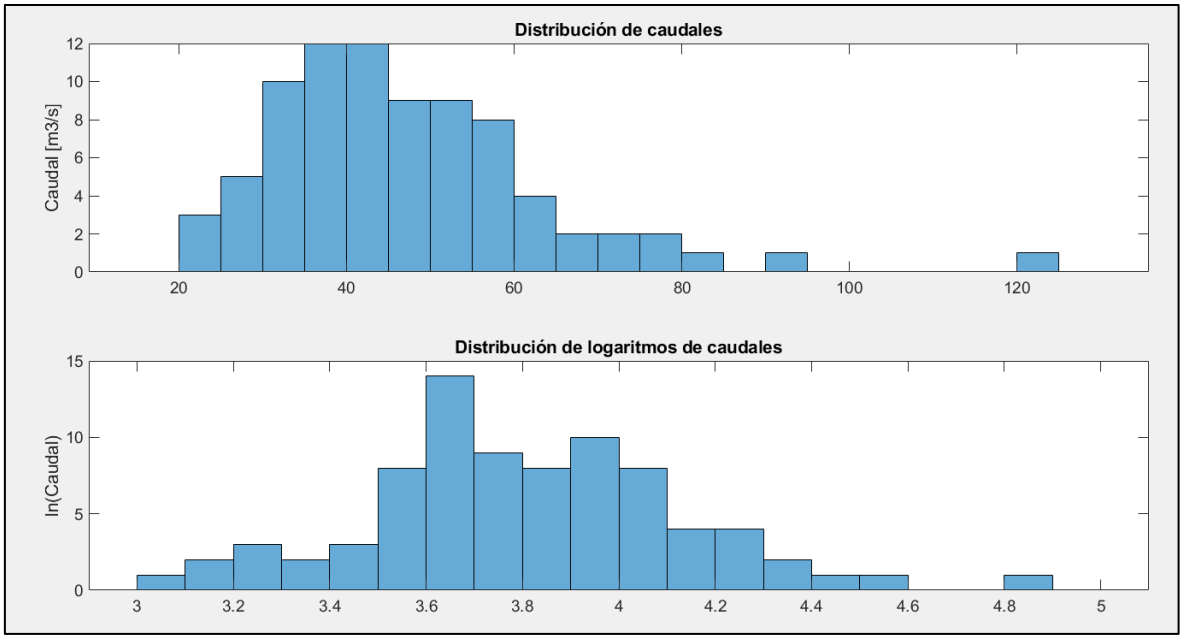


*Figura 3.5-12. Distribución de la serie y de los logaritmos de la serie cada mes*

### 3.5.7. Caudales máximos y mínimos de los periodos de prueba

| Discretización de la serie | Qmáx [m3/s] | Qmín [m3/s] |
|---|---|---|
| 15minutos | 578,84 | 10,98 |
| 1 hora | 540,40 | 10,99 |
| 1 día | 336,91 | 13,85 |
| 1 semana | 166,47 | 16,24 |
| 2 semanas | 133,48 | 16,24 |
| 1 mes | 124,80 | 21,54 |

*Tabla 3.5-2. Limites inferior y superior para el dominio de los caudales en diferentes intervalos*

## 3.6. Modelos ARIMA considerados para el pronóstico

Se utiliza el criterio de información Bayesiano para elegir el modelo SARIMA para las diferentes series, que posteriormente es calibrado en base al periodo de prueba y utilizado para los pronósticos. La calibración de modelos con una frecuencia mayor a las indicadas aquí no fue posible debido a que la versión de MATLAB utilizada (2018a), no permite utilizar suficiente memoria para el cálculo de parámetros por máxima verosimilitud, los cuales necesitan la aplicación de métodos numéricos, y por lo tanto, gran capacidad de cálculo en series muy largas.

| Serie semanal | | Serie cada dos semanas | | Serie mensual | |
|---|---|---|---|---|---|
| Orden módelos (p,d,q)(P,D,Q)m | BIC | Orden módelos (p,d,q)(P,D,Q)m | BIC | Orden módelos (p,d,q)(P,D,Q)m | BIC |
| (1,0,0)(1,1,1)52 | 3209,24 | (1,0,0)(1,1,1)27 | 1555,51 | **(1,0,0)(1,1,1)12** | **635,46** |
| (2,0,0)(1,1,1)52 | 3221,63 | (2,0,0)(1,1,1)27 | 1565,74 | (2,0,0)(1,1,1)12 | 639,62 |
| (3,0,0)(1,1,1)52 | 3227,66 | (3,0,0)(1,1,1)27 | 1565,48 | (3,0,0)(1,1,1)12 | 642,58 |
| (1,0,0)(2,1,1)52 | 3174,94 | (1,0,0)(2,1,1)27 | 1523,71 | (1,0,0)(2,1,1)12 | 636,98 |
| (2,0,0)(2,1,1)52 | 3188,36 | (2,0,0)(2,1,1)27 | 1547,71 | (2,0,0)(2,1,1)12 | 641,43 |
| (3,0,0)(2,1,1)52 | 3195,77 | (3,0,0)(2,1,1)27 | 1549,81 | (3,0,0)(2,1,1)12 | 643,88 |
| (1,0,0)(1,1,2)52 | 3174,14 | (1,0,0)(1,1,2)27 | 1548,18 | (1,0,0)(1,1,2)12 | 636,29 |

| | | | | | |
|---|---|---|---|---|---|
| (2,0,0)(1,1,2)52 | 3188,30 | (2,0,0)(1,1,2)27 | 1555,88 | (2,0,0)(1,1,2)12 | 639,98 |
| (3,0,0)(1,1,2)52 | 3194,99 | (3,0,0)(1,1,2)27 | 1556,79 | (3,0,0)(1,1,2)12 | 642,83 |
| **(1,0,0)(2,1,2)52** | **3173,27** | **(1,0,0)(2,1,2)27** | **1506,57** | (1,0,0)(2,1,2)12 | 639,01 |
| (2,0,0)(2,1,2)52 | 3174,47 | (2,0,0)(2,1,2)27 | 1524,12 | (2,0,0)(2,1,2)12 | 643,06 |
| (3,0,0)(2,1,2)52 | 3195,03 | (3,0,0)(2,1,2)27 | 1528,79 | (3,0,0)(2,1,2)12 | 645,84 |

*Tabla 3.6-1. Modelos ARIMA elegidos para las diferentes series*

# 4. RESULTADOS

## 4.1. Generación de predicciones Montecarlo

Se grafican 1000 rutas de la generación de caudales durante distintos intervalos: el color celeste expresa el rango intercuartil, mientras que el azul oscuro, el rango entre los percentiles 2,5 y 97,5. Como se ha mencionado, la parte estocástica de las ecuaciones provoca que los pronósticos nunca sean iguales, por lo que cada vez se generan diferentes valores en el mismo instante, se hace uso de cuantiles para mostrar esta distribución. El registro se muestra a través de una línea roja.

### 4.1.1. Modelo Thomas Fiering

#### *4.1.1.1. Intervalos de 15 minutos*

Los pronósticos con los tres primeros intervalos no son muy exactos en cuanto a que producen valores dentro de un rango de alrededor de 100$m^3$/s, considerando que los caudales registrados tienen una media de 45,25$m^3$/s.

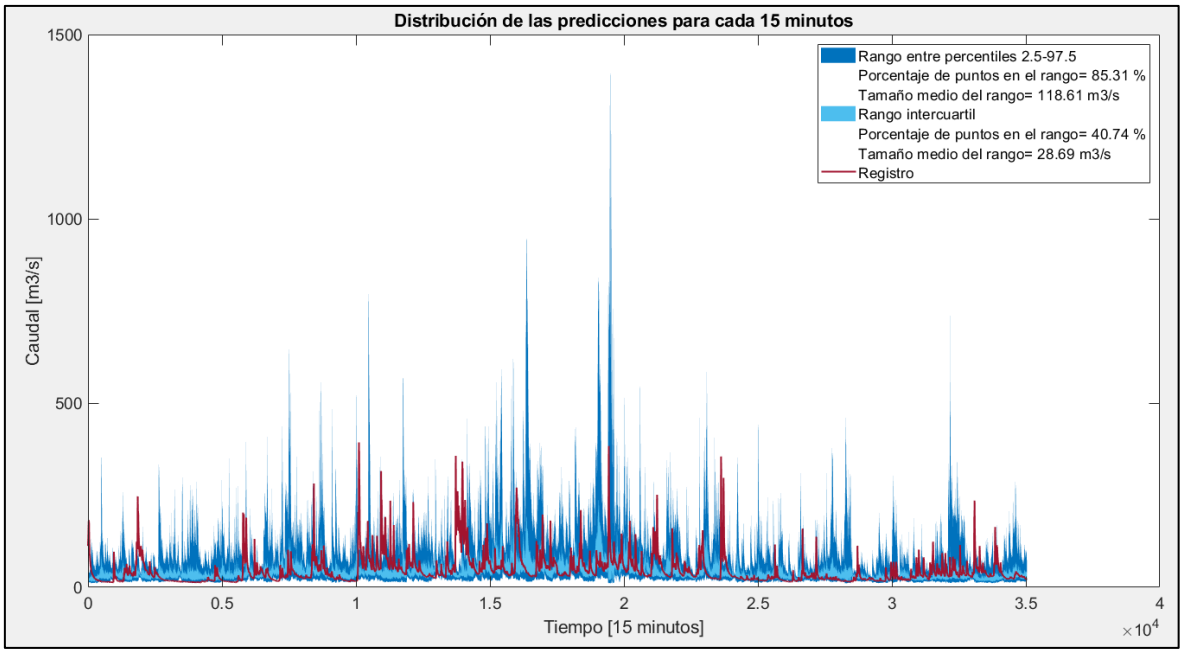


*Figura 4.1-1. Generación Montecarlo de la serie cada 15minutos, modelo Thomas-Fiering*

#### *4.1.1.2. Intervalos de 1 hora*

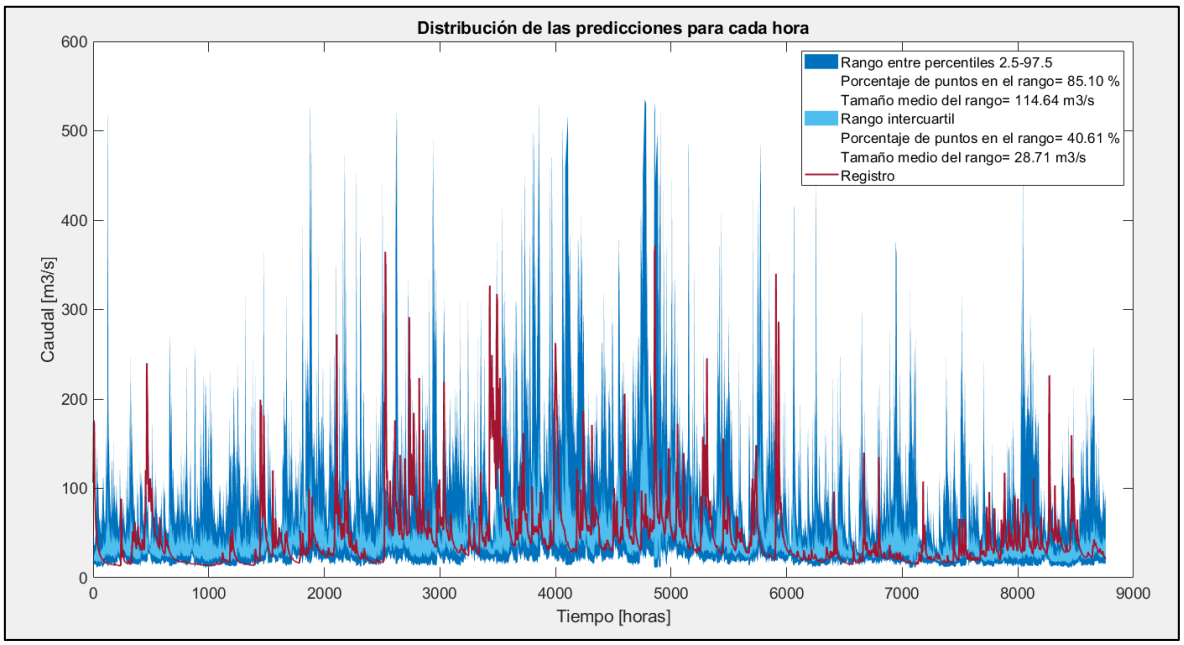


*Figura 4.1-2. Generación Montecarlo de la serie cada hora, modelo Thomas-Fiering*

#### *4.1.1.3. Intervalos de 1 día*

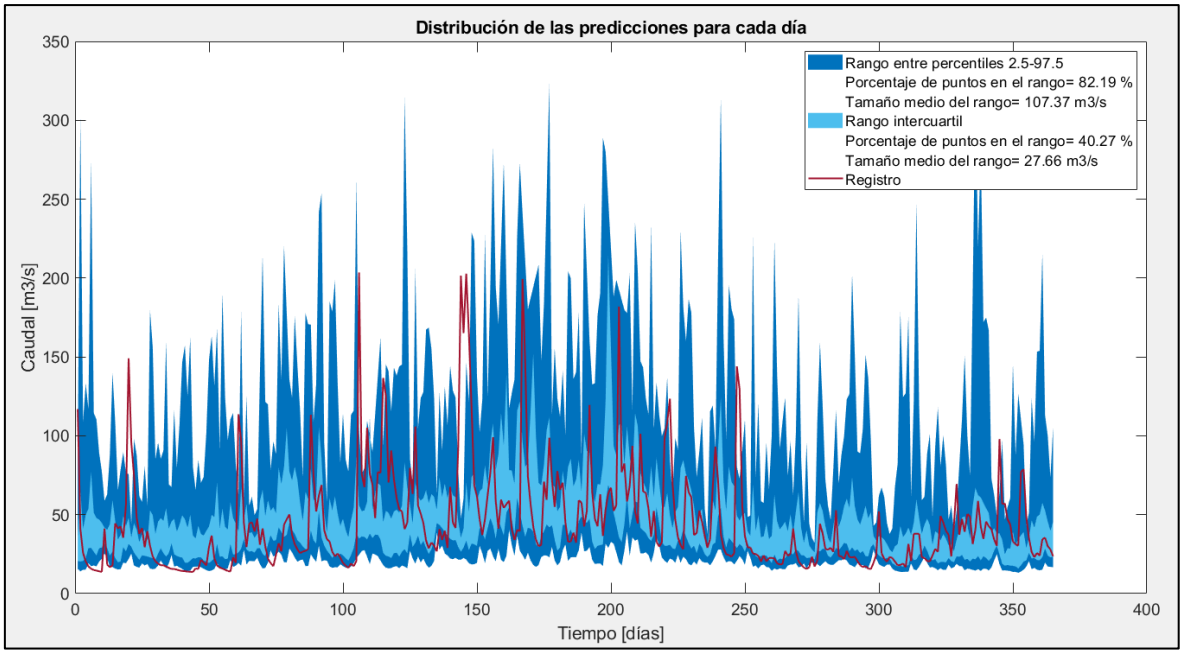


*Figura 4.1-3. Generación Montecarlo de la serie cada día, modelo Thomas-Fiering*

#### *4.1.1.4. Intervalos de 1 semana*

Se aprecia una reducción de los tamaños de los rangos sobre los que se generan las predicciones, a pesar de esto, se conserva el porcentaje de puntos que se encuentran dentro de los rangos de predicción: alrededor de un 80% en el rango del 95% y un 40% en el rango intercuartil. La razón de tener rangos más pequeños y haber mejorado la precisión, es producto de tener una menor dispersión en las series usadas para la calibración, lo cual puede ser observado en la desviación estándar de las series, al aumentar el tamaño de los intervalos, se reduce la desviación estándar.

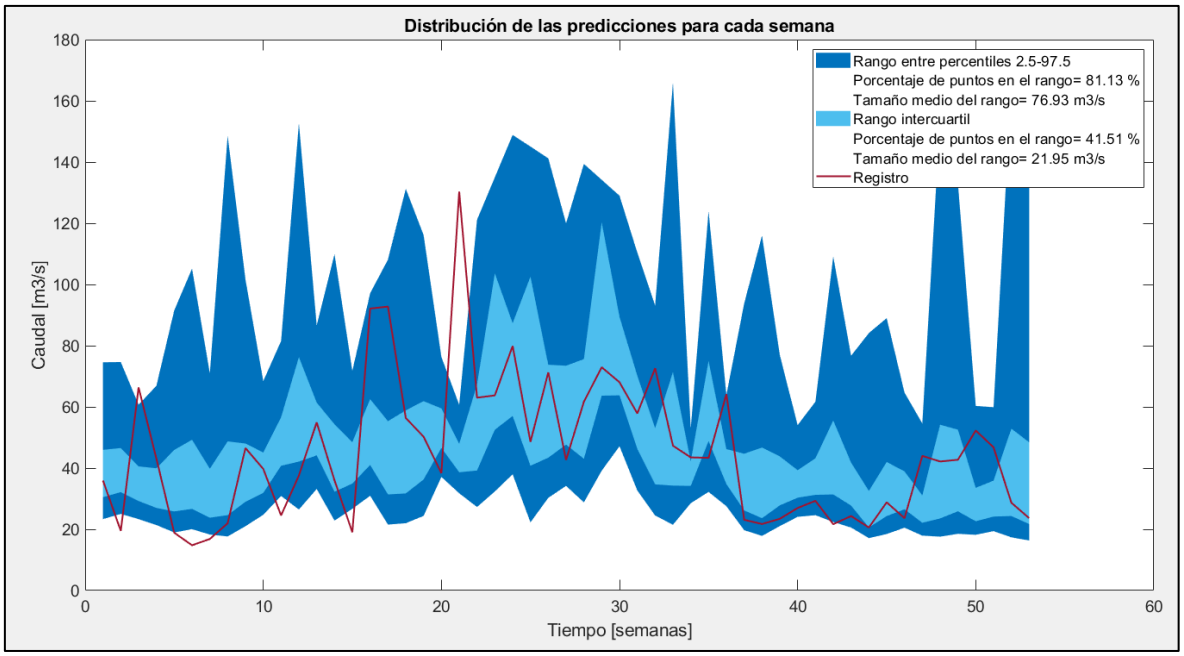


*Figura 4.1-4. Generación Montecarlo de la serie cada semana, modelo Thomas-Fiering*

#### *4.1.1.5. Intervalos de 2 semanas*

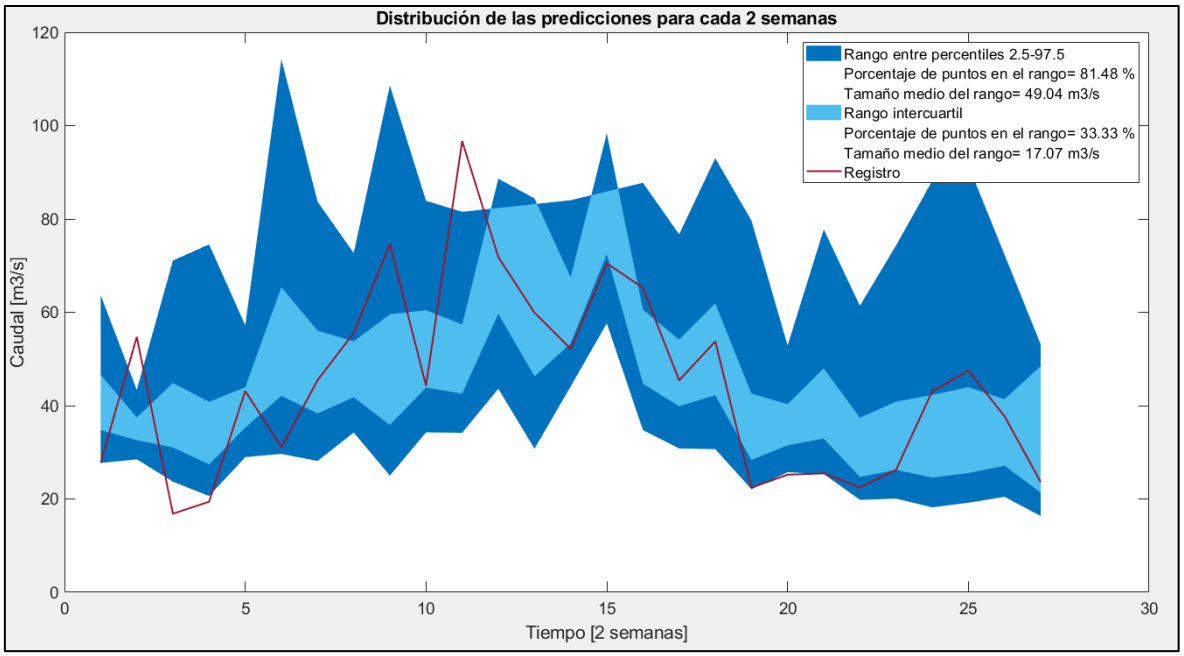


*Figura 4.1-5. Generación Montecarlo de la serie cada 2 semanas, modelo Thomas-Fiering*

#### *4.1.1.6. Intervalos de 1 mes*

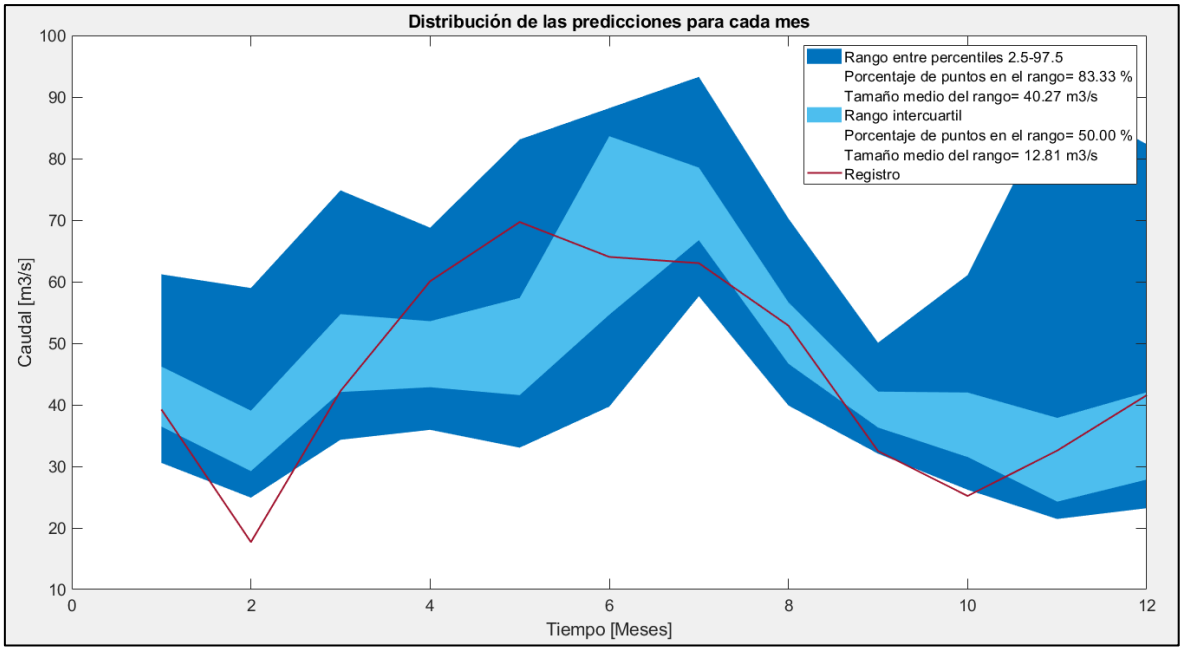


*Figura 4.1-6. Generación Montecarlo de la serie cada mes, modelo Thomas-Fiering*

### 4.1.2. Modelo ARIMA

El mismo procedimiento se utiliza para la metodología Box-Jenkins, para la cual se generan 1000 rutas por medio del método Montecarlo a través del comando simulate (), de las cuales se obtienen los rangos del 95% centrales e intercuartil para todos los instantes pronosticados.

#### 4.1.2.1. *Intervalos de 1 semana*

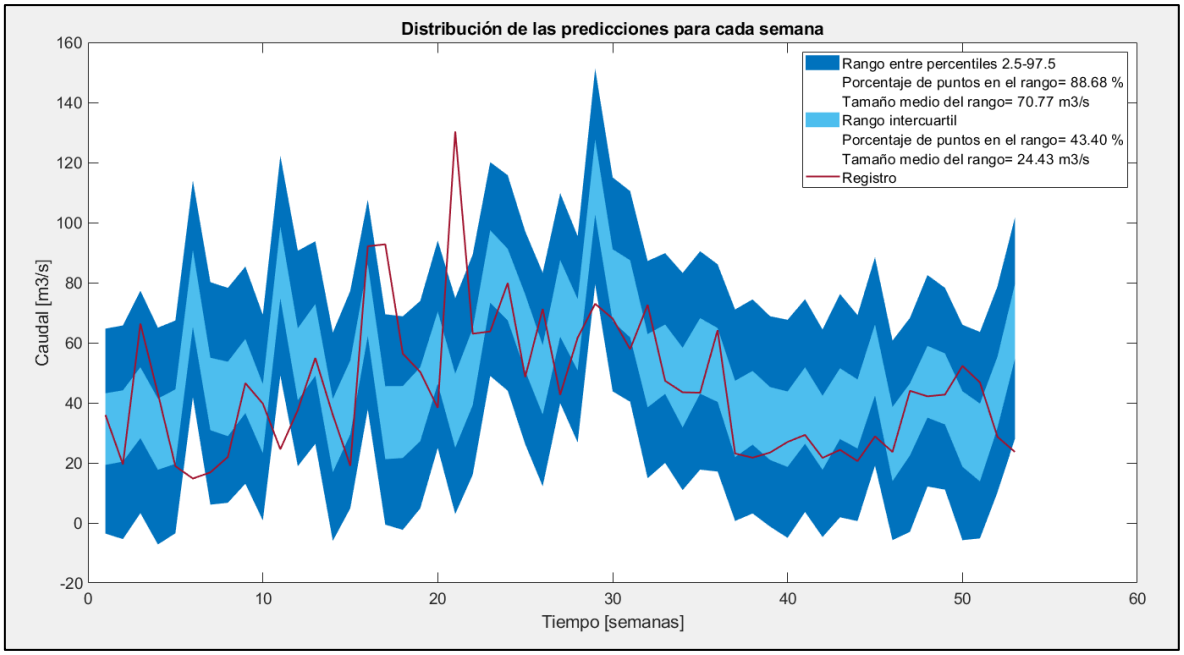


*Figura 4.1-7. Generación Montecarlo de la serie cada semana, modelo ARIMA*

#### 4.1.2.2. *Intervalos de 2 semanas*

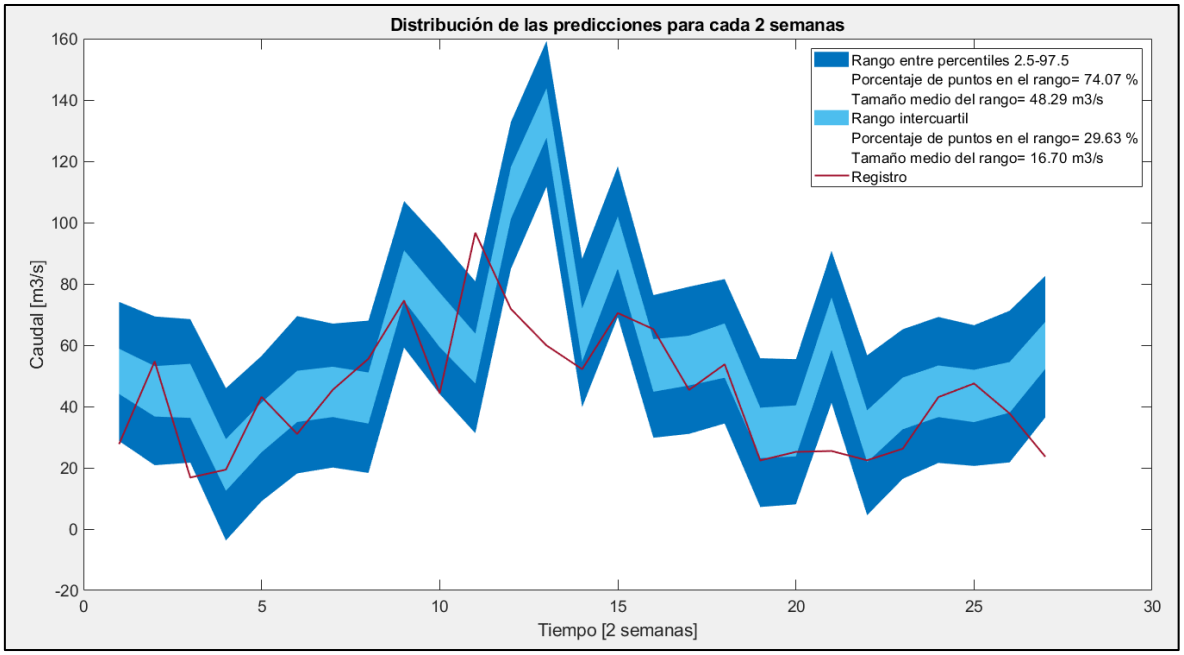


*Figura 4.1-8. Generación Montecarlo de la serie cada 2 semanas, modelo ARIMA*

#### *4.1.2.3. Intervalos de 1 mes*

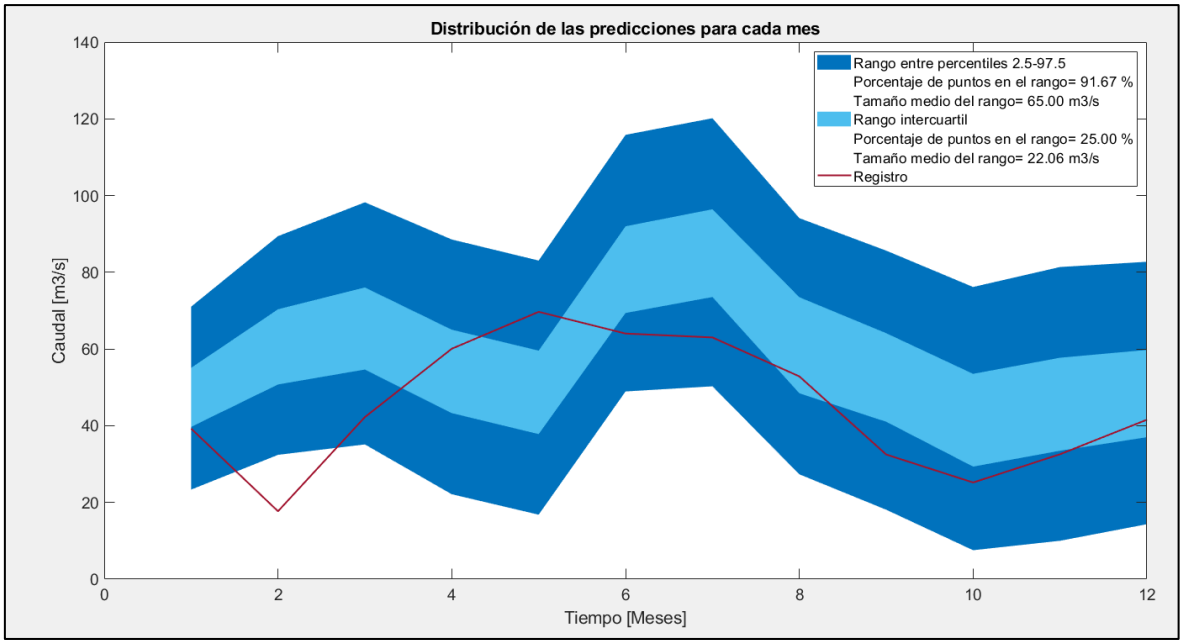


*Figura 4.1-9. Generación Montecarlo de la serie cada mes, modelo ARIMA*

### 4.1.3. Resumen de los datos observados capturados por los intervalos de predicción

Una tendencia que se puede observar en las generaciones es que, en general, conforme aumenta el tamaño de los intervalos entre pronósticos, los rangos de generación se vuelven más pequeños y el porcentaje de pronósticos que se encuentran dentro de estos rangos se mantiene.

| Discretización de la serie | Tamaño medio del rango intercuartil [m3/s] | % de puntos en el rango intercuartil | Tamaño medio del rango del 95% central [m3/s] | % de puntos en el rango del 95% central |
|---|---|---|---|---|
| **Modelo de Thomas-Fiering** | | | | |
| 15minutos | 28,69 | 40,74 | 118,61 | 85,31 |
| 1 hora | 28,71 | 40,61 | 114,64 | 85,10 |
| 1 día | 27,66 | 40,27 | 107,37 | 82,19 |
| 1 semana | 21,95 | 41,51 | 76,93 | 81,13 |
| 2 semanas | 17,07 | 33,33 | 49,04 | 81,48 |

| | | | | |
|---|---|---|---|---|
| 1 mes | 12,81 | 50,00 | 40,27 | 83,33 |
| **Modelo ARIMA** | | | | |
| 1 semana | 24,43 | 43,40 | 70,77 | 88,68 |
| 2 semanas | 16,70 | 29,63 | 48,29 | 74,07 |
| 1 mes | 22,06 | 25,00 | 65,00 | 91,67 |

*Tabla 4.1-1. Resultados de las distribuciones generadas para cada intervalo*

## 4.2. Comparación de la media de las predicciones con la serie observada

Se asume que la media es el valor más representativo de la serie en cada instante y se la utiliza para calcular magnitudes basadas en los errores, así como el coeficiente de eficiencia de Nash-Sutcliffe modificado, que sirve para establecer una calificación de la bondad del ajuste del modelo.

### 4.2.1. Modelo Thomas Fiering

Desde la serie cada 15 minutos hasta la serie semanal, los errores absolutos RMSE son muy grandes en comparación a la magnitud de la serie, y de igual manera son los errores porcentuales MAPE, los cuales son mayores al 37%.

#### 4.2.1.1. *Intervalos de 15minutos*

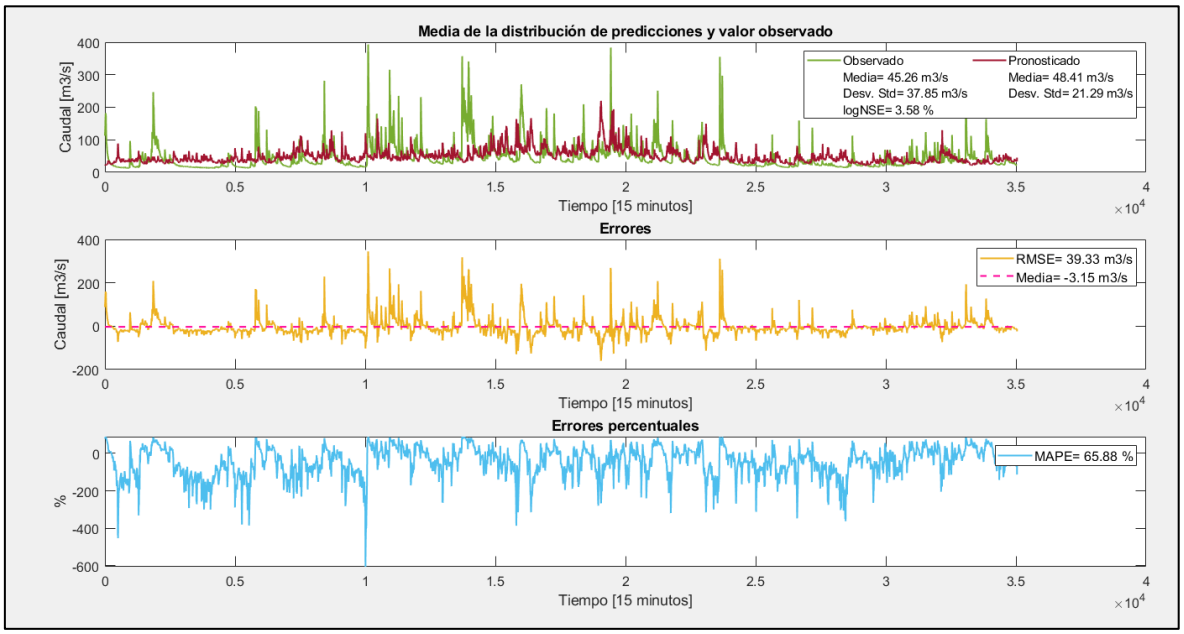


*Figura 4.2-1. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada 15minutos, modelo Thomas-Fiering.*

#### 4.2.1.2. *Intervalos de 1 hora*

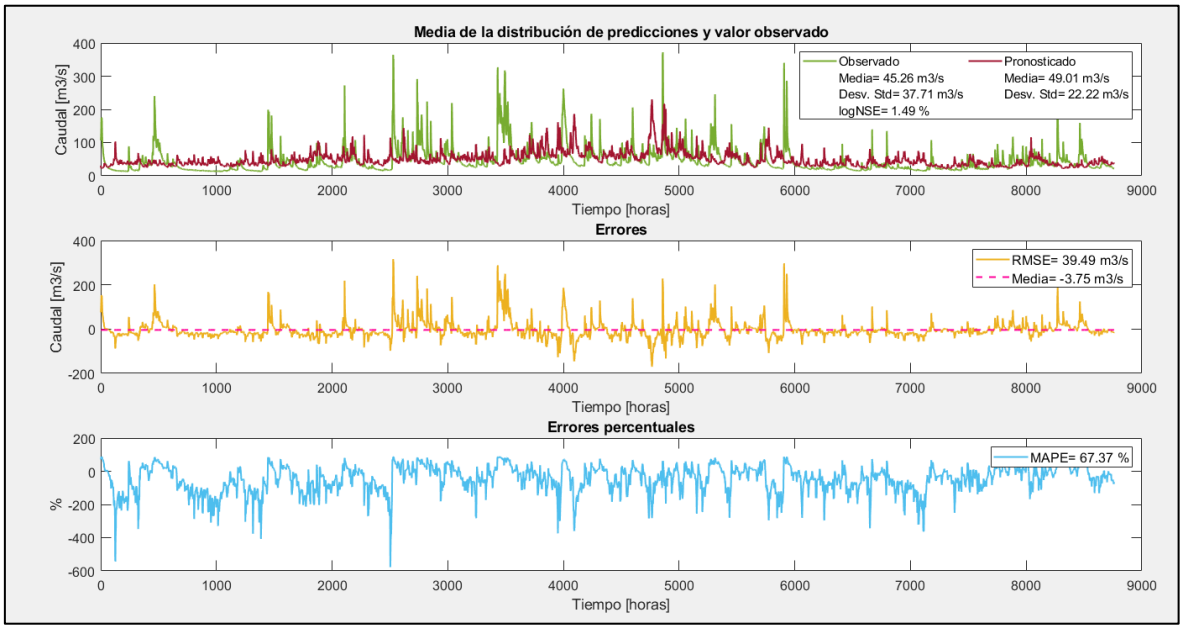


*Figura 4.2-2. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada hora, modelo Thomas-Fiering.*

#### 4.2.1.3. *Intervalos de 1 día*

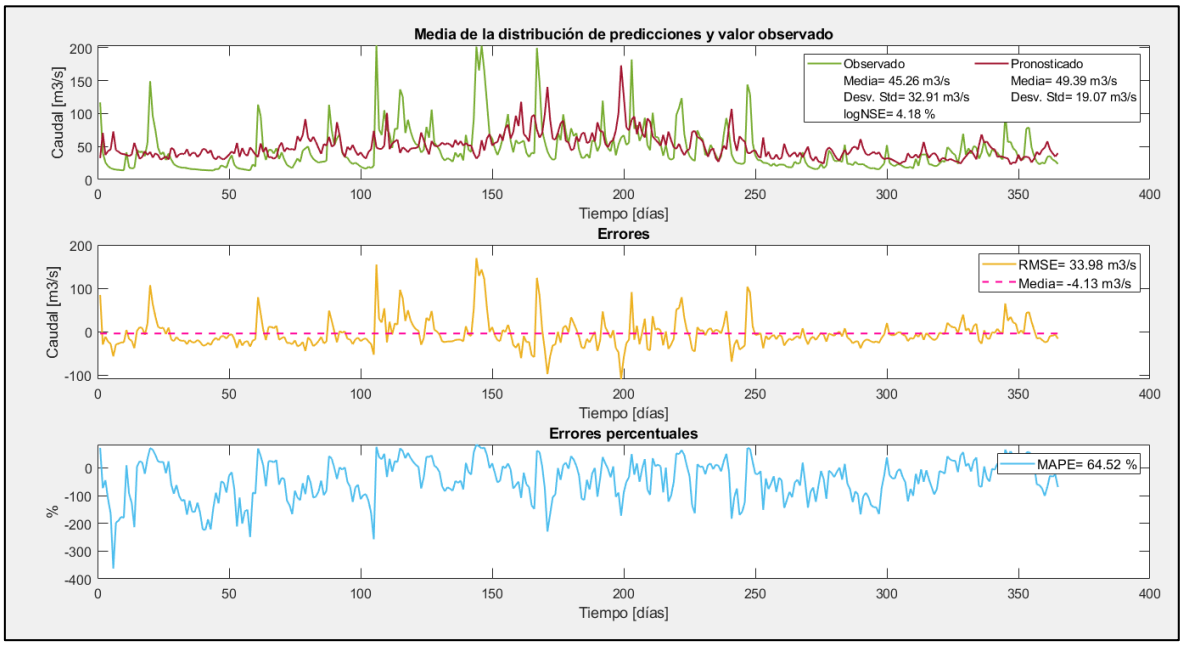


*Figura 4.2-3. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada día, modelo Thomas-Fiering.*

#### 4.2.1.4. *Intervalos de 1 semana*

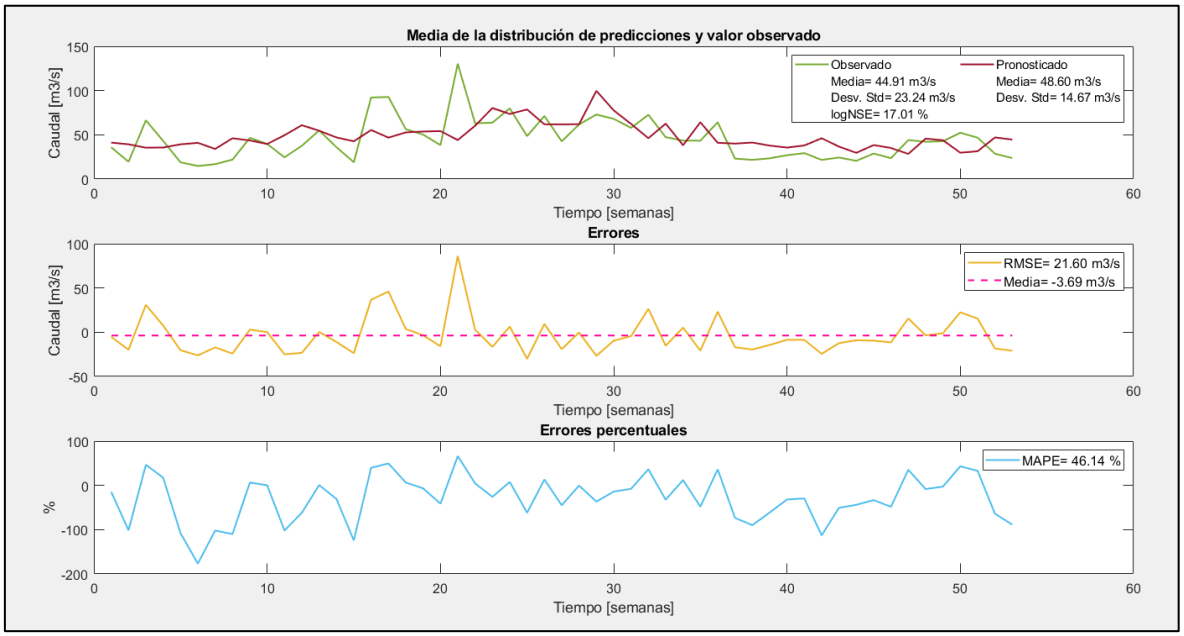


*Figura 4.2-4. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada 2 semanas, modelo Thomas-Fiering.*

#### *4.2.1.5. Intervalos de 2 semanas*

La media de los errores, que en teoría debe ser muy cercana a 0, lo es en las últimas dos series, y a la vez, sus errores absolutos y porcentuales son relativamente más pequeños en comparación a las anteriores series. A pesar de una mejora en los resultados del modelo, se pierde exactitud en expresar la distribución de los valores de la serie, porque se utilizan intervalos más grandes y que son menos "reales" ya que utilizan el promedio de algunos registros.

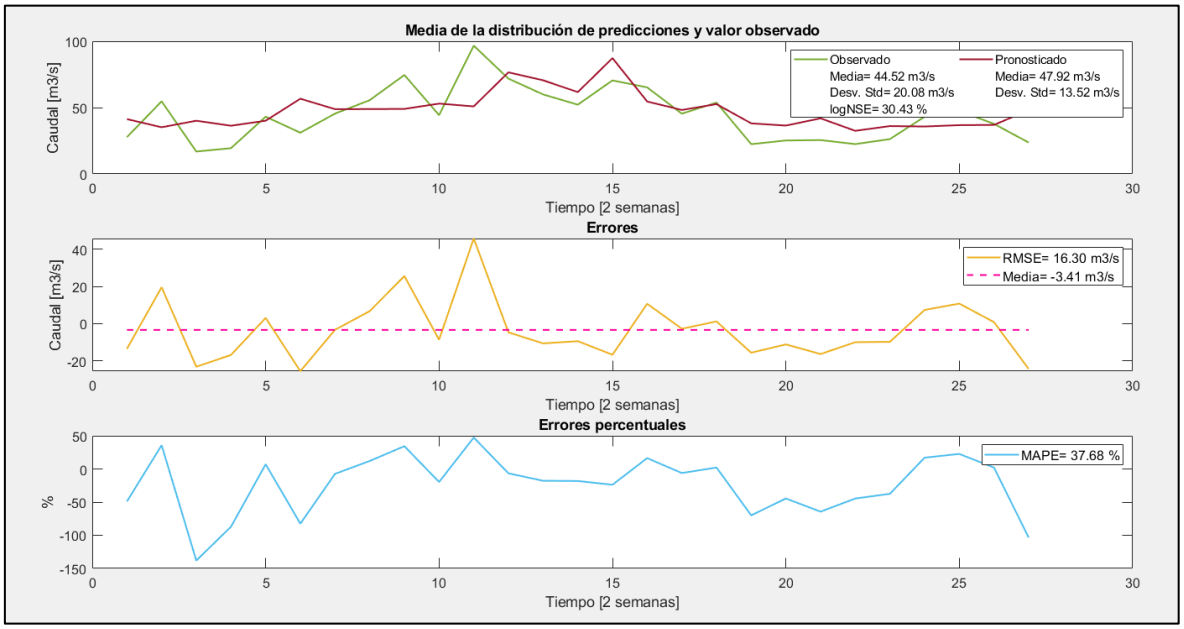


*Figura 4.2-5. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada 2 semanas, modelo Thomas-Fiering.*

#### *4.2.1.6. Intervalos de 1 mes*

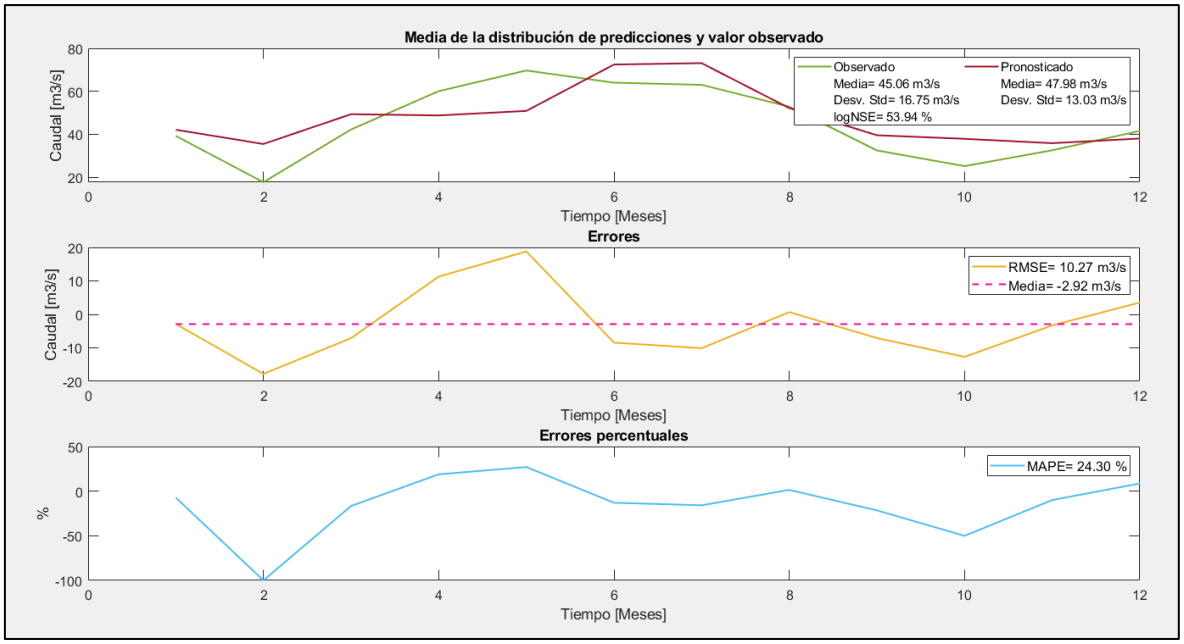


*Figura 4.2-6. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada mes, modelo Thomas-Fiering.*

### 4.2.2. Modelo ARIMA

El coeficiente de eficiencia de Nash-Sutcliffe modificado, que expresa lo muy buenos que son los modelos al comparar su capacidad predictiva con la media del periodo de prueba, tiene valores muy bajos para los modelos ARIMA, incluso un valor negativo para la serie semanal, además de esto, tiene mayores errores para los mismos intervalos en comparación a lo pronosticado por el modelo Thomas-Fiering.

#### 4.2.2.1. *Intervalos de 1 semana*

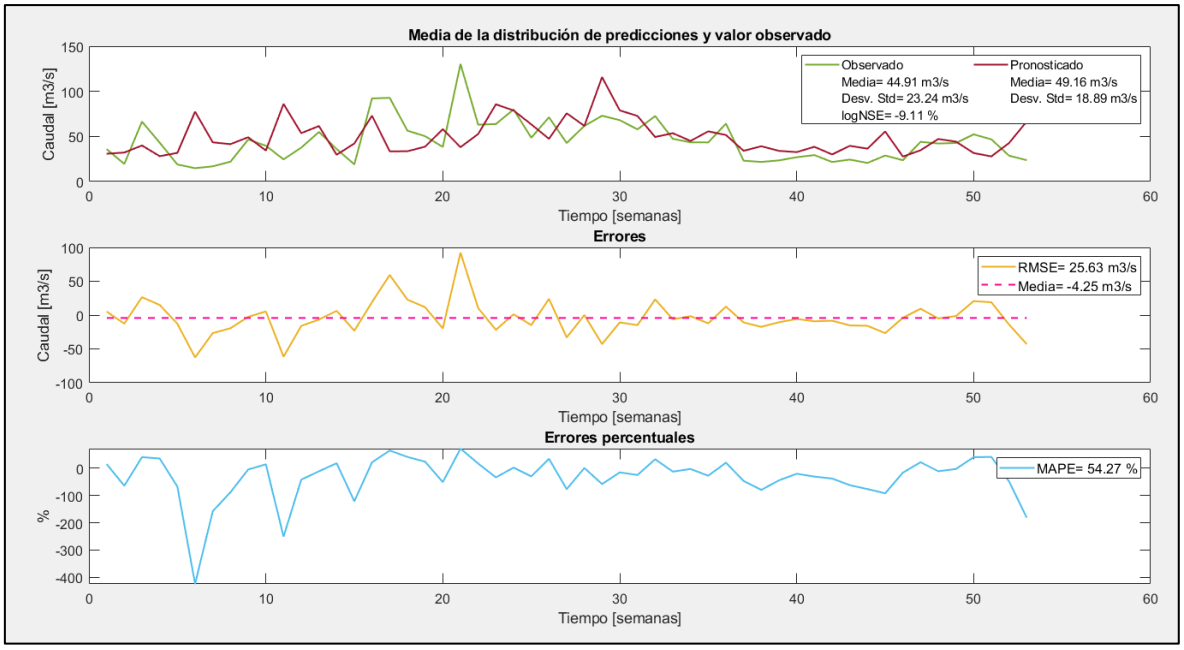


*Figura 4.2-7. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada semana, modelo ARIMA.*

#### 4.2.2.2. *Intervalos de 2 semanas*

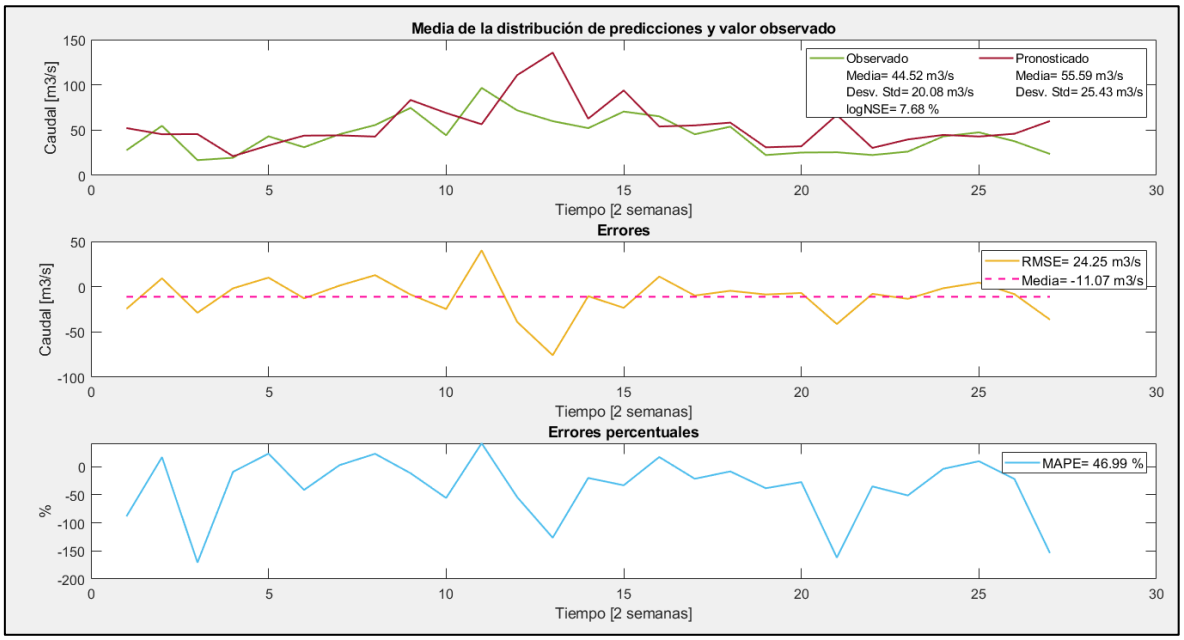


*Figura 4.2-8. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada 2 semanas, modelo ARIMA.*

#### *4.2.2.3. Intervalos de 1 mes*

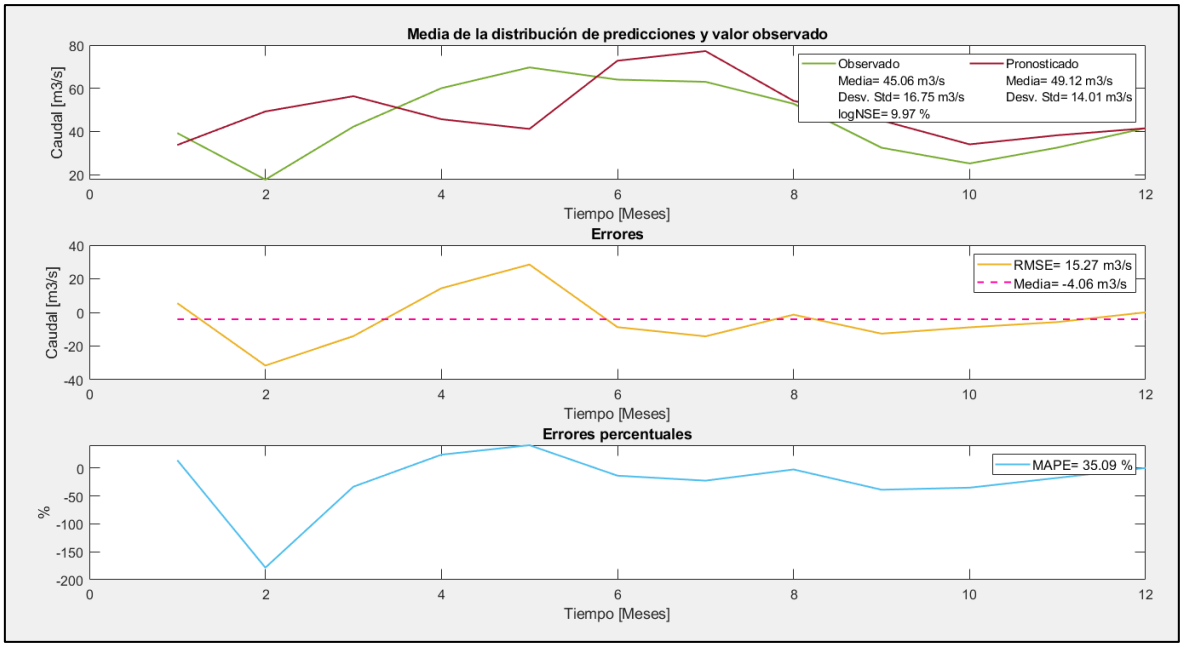


*Figura 4.2-9. Media de las generaciones Montecarlo, registro, errores y errores porcentuales a través del tiempo en la serie cada mes, modelo ARIMA.*

### 4.2.3. Métricas del desempeño de los modelos

Se muestra un resumen de las diferentes maneras utilizadas para expresar el desempeño de los modelos, los errores absolutos y porcentuales se vuelven menores conforme aumenta el tamaño de los intervalos, la media de los errores es cercana a 0 en todos los casos, como se supone al momento de realizar regresiones. El desempeño de los modelos ARIMA es, en general, peor que el modelo Thomas-Fiering para todos los intervalos donde fue posible aplicar este método.

| Discretización de la serie | RMSE [m3/s] | MAPE [%] | log NSE [%] | Media de los errores [m3/s] |
|---|---|---|---|---|
| **Modelo de Thomas-Fiering** | | | | |
| 15minutos | 39,33 | 65,88 | 3,58 | -3,15 |
| 1 hora | 39,49 | 67,37 | 1,49 | -3,75 |
| 1 día | 33,98 | 64,52 | 4,18 | -4,13 |

| | | | | |
|---|---|---|---|---|
| 1 semana | 21,60 | 46,14 | 17,01 | -3,69 |
| 2 semanas | 16,30 | 37,68 | 30,43 | -3,41 |
| 1 mes | 10,27 | 24,30 | 53,94 | -2,92 |
| **Modelo ARIMA** | | | | |
| 1 semana | 25,63 | 54,27 | -9,11 | -4,25 |
| 2 semanas | 24,25 | 46,99 | 7,68 | -11,07 |
| 1 mes | 15,27 | 35,09 | 9,97 | -4,06 |

*Tabla 4.2-1. Diferentes métricas para el desempeño de los modelos*

### 4.2.4. Estadísticos da la serie observada y pronosticada

| Discretización de la serie | Desviación estándar observada [m3/s] | Desviación estándar pronosticada [m3/s] | Media observada [m3/s] | Media pronosticada [m3/s] |
|---|---|---|---|---|
| **Modelo de Thomas-Fiering** | | | | |
| 15minutos | 37,85 | 22,26 | 45,26 | 49,06 |
| 1 hora | 37,71 | 22,53 | 45,26 | 49,13 |
| 1 día | 32,91 | 19,33 | 45,26 | 49,49 |
| 1 semana | 23,24 | 15,18 | 44,91 | 48,86 |
| 2 semanas | 20,08 | 13,66 | 44,52 | 47,70 |
| 1 mes | 16,75 | 13,03 | 45,06 | 47,97 |
| **Modelo ARIMA** | | | | |
| 1 semana | 23,24 | 18,89 | 44,91 | 49,46 |
| 2 semanas | 20,08 | 25,56 | 44,52 | 55,59 |
| 1 mes | 16,75 | 13,93 | 45,06 | 49,16 |

*Tabla 4.2-2. Características de la distribución de la serie sintética y la observada*

## 4.3. Diagnóstico de los errores

Debido a los pocos años que se posee registros, a diferencia de lo recomendado para estos modelos, que es por lo menos de 10 a 15 años, el modelo no se analizó como un generador de caudales que logre conservar la media, la desviación estándar y la autocorrelación serial, motivos para lo que fue concebido en un principio por sus autores, sino como una regresión, en la que se analizan sus residuos y la bondad del ajuste con sus observaciones.

Se dibuja el histograma de los errores, además de una distribución normal ajustada y el gráfico Q-Q para verificar la normalidad de la serie, la gráfica del ACF Y PACF para demostrar la aleatoriedad de los errores y la gráfica del ACF Y PACF de los errores al cuadrado para verificar la homocedasticidad.

### 4.3.1. Modelo Thomas-Fiering

#### *4.3.1.1. Intervalos de 15minutos*

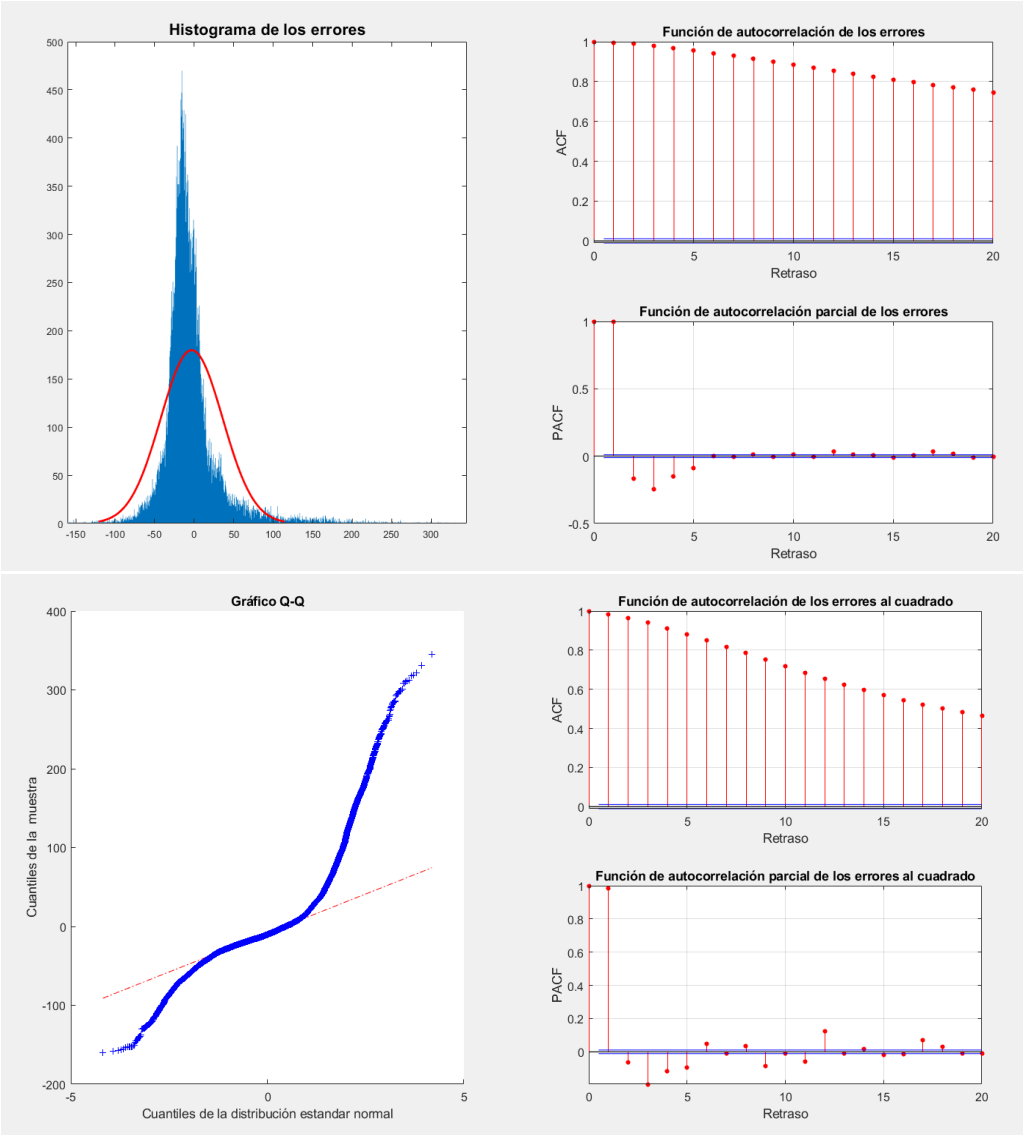


*Figura 4.3-1. Gráficos de los errores en el pronóstico de la serie cada 15minutos, modelo Thomas-Fiering.*

#### *4.3.1.2. Intervalos de 1 hora*

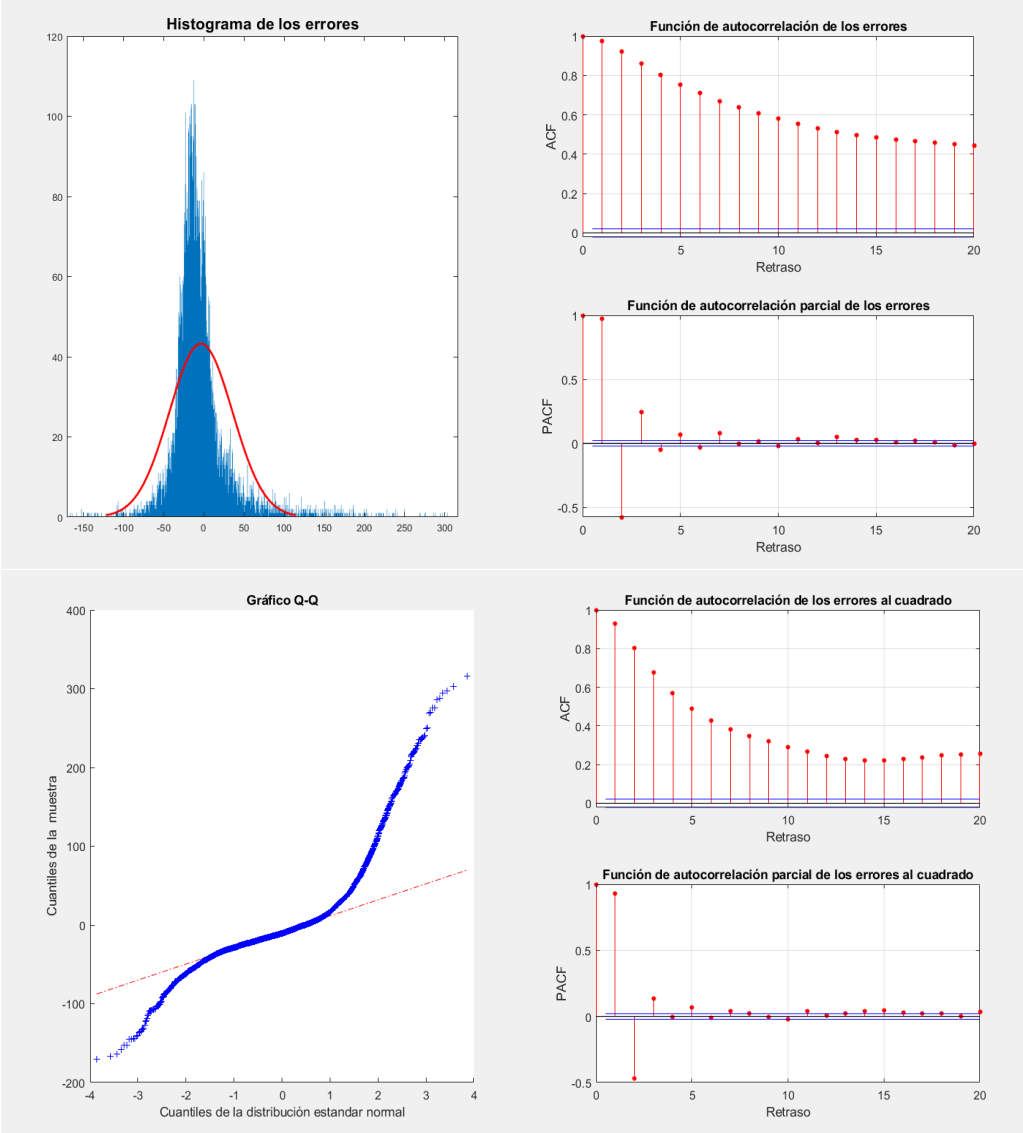


*Figura 4.3-2. Gráficos de los errores en el pronóstico de la serie cada hora, modelo Thomas-Fiering.*

La función histfit () es utilizada para dibujar un histograma de los errores en cada serie con una función de densidad normal ajustada, en comparación a esta curva, las 4 primeras series parecen tener en común las características de media y desviación estándar, pero no la curtosis, que es lo pronunciado que es el "pico" de la distribución, la cual se aprecia que es mucho mayor en la distribución de errores.

#### *4.3.1.3. Intervalos de 1 día*

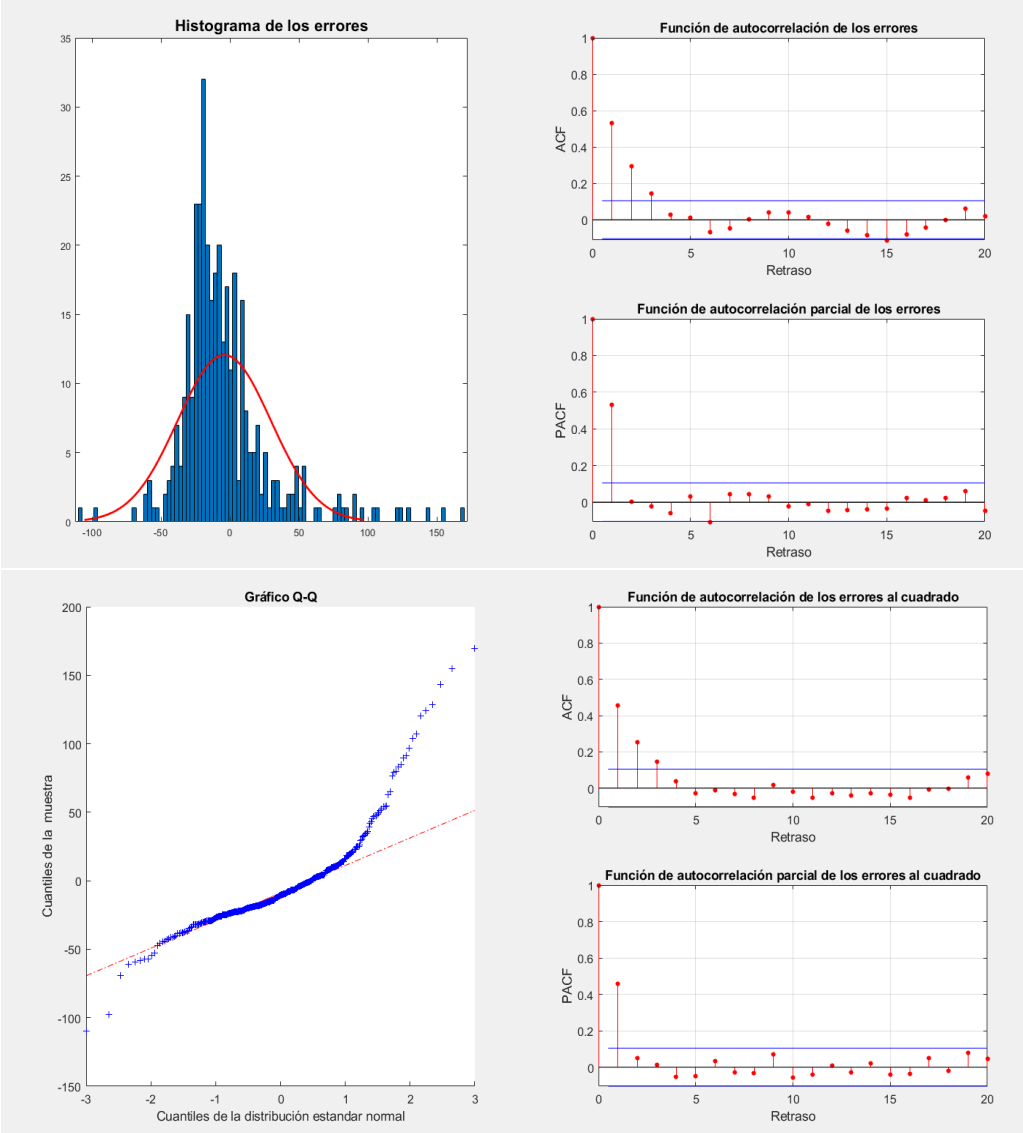


*Figura 4.3-3. Gráficos de los errores en el pronóstico de la serie cada día, modelo Thomas-Fiering*

Hasta este punto, los errores de la serie tienen una autocorrelación significativa para retrasos mayores a 0, por lo que hay lugar para mejorar el modelo, quizás con otra metodología o con la utilización de otra variable. De igual manera, los gráficos Q-Q de la serie muestran que los errores tienen una distribución alejada de la normal y los gráficos de ACF y PACF de los errores al cuadrado, que la serie no es homocedástica ya que si existe autocorrelación entre ellos.

#### *4.3.1.4. Intervalos de 1 semana*

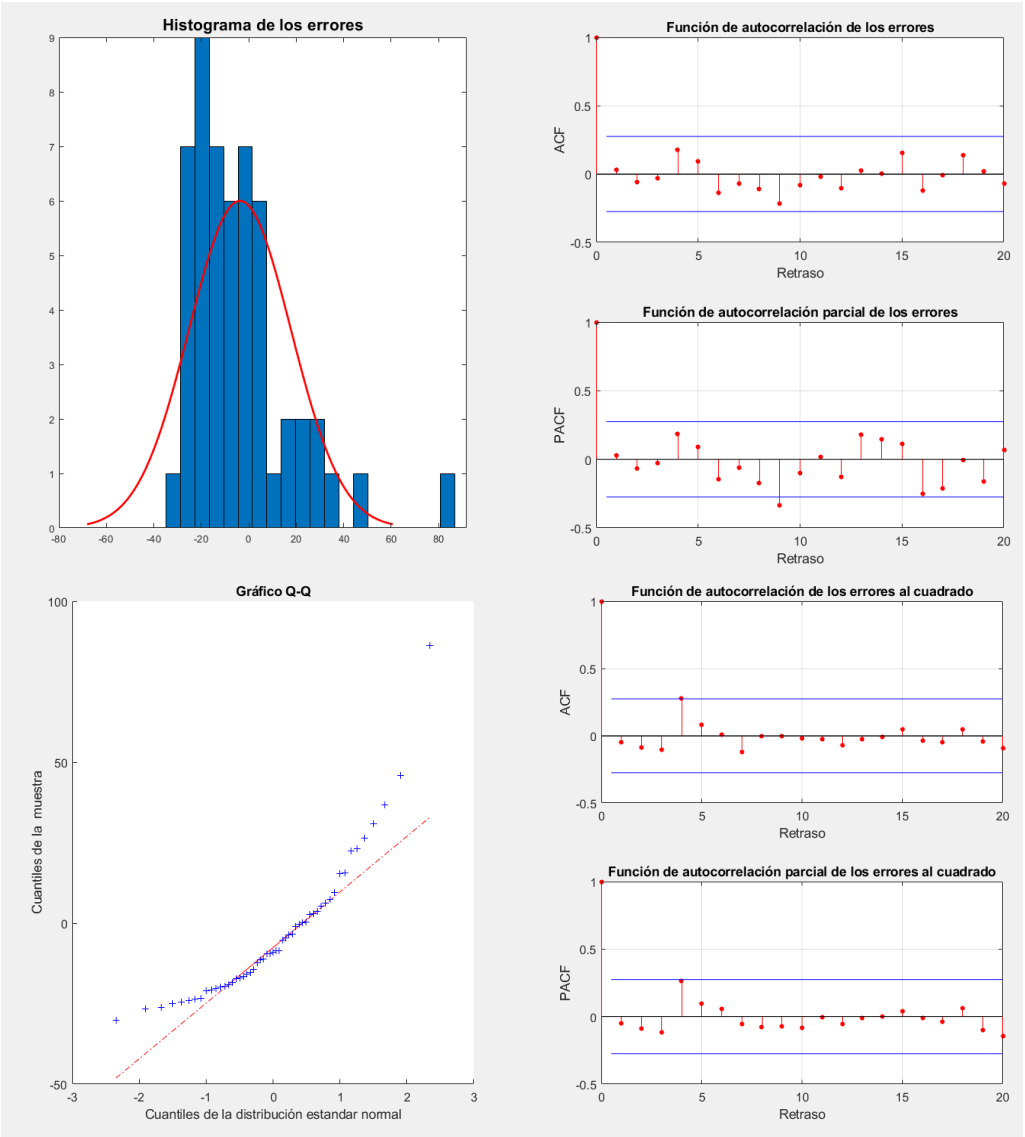


*Figura 4.3-4. Gráficos de los errores en el pronóstico de la serie cada semana, modelo Thomas-Fiering.*

#### *4.3.1.5. Intervalos de 2 semanas*

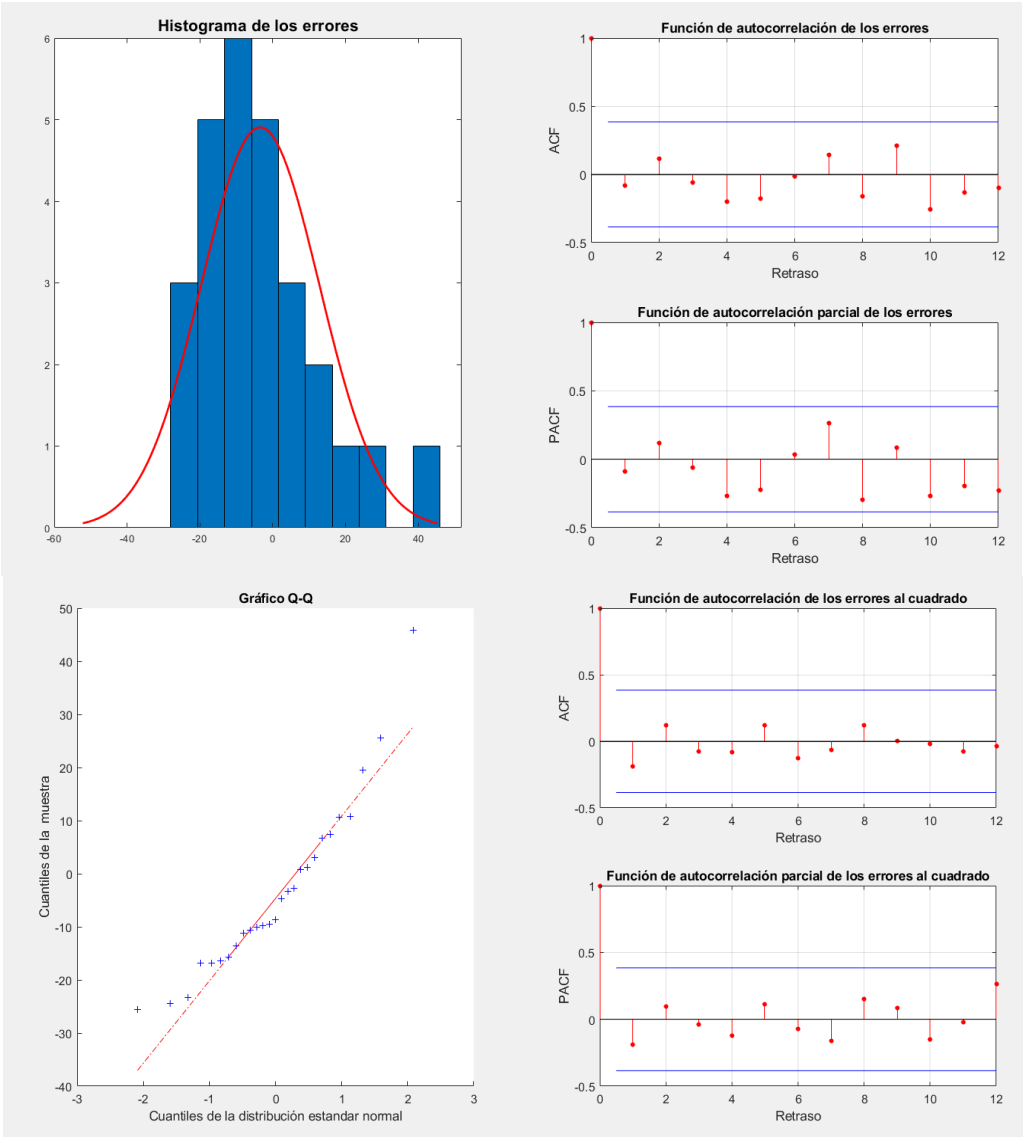


*Figura 4.3-5. Gráficos de los errores en el pronóstico de la serie cada 2 semanas, modelo Thomas-Fiering.*

#### *4.3.1.6. Intervalos de 1 mes*

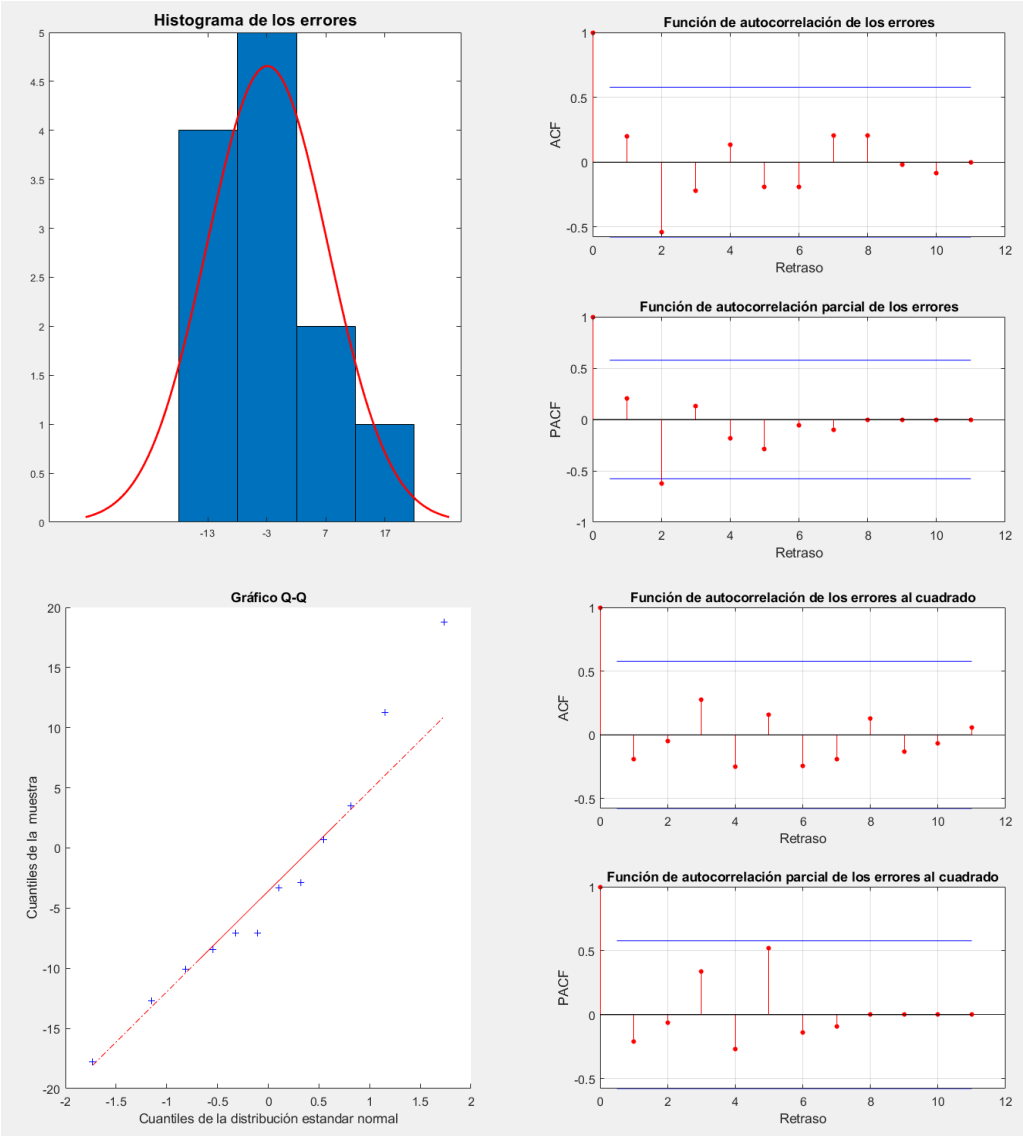


*Figura 4.3-6. Gráficos de los errores en el pronóstico de la serie cada mes, modelo Thomas-Fiering.*

Los errores en las últimas 3 series muestran que se cumplen los supuestos de los errores, de aleatoriedad, media igual a cero, homocedasticidad y normalidad, que implican que la serie muy probablemente no puede ser perfeccionada incluso al agregar otra variable más a la ecuación.

### 4.3.2. Modelo ARIMA

#### *4.3.2.1. Intervalos de 1 semana*

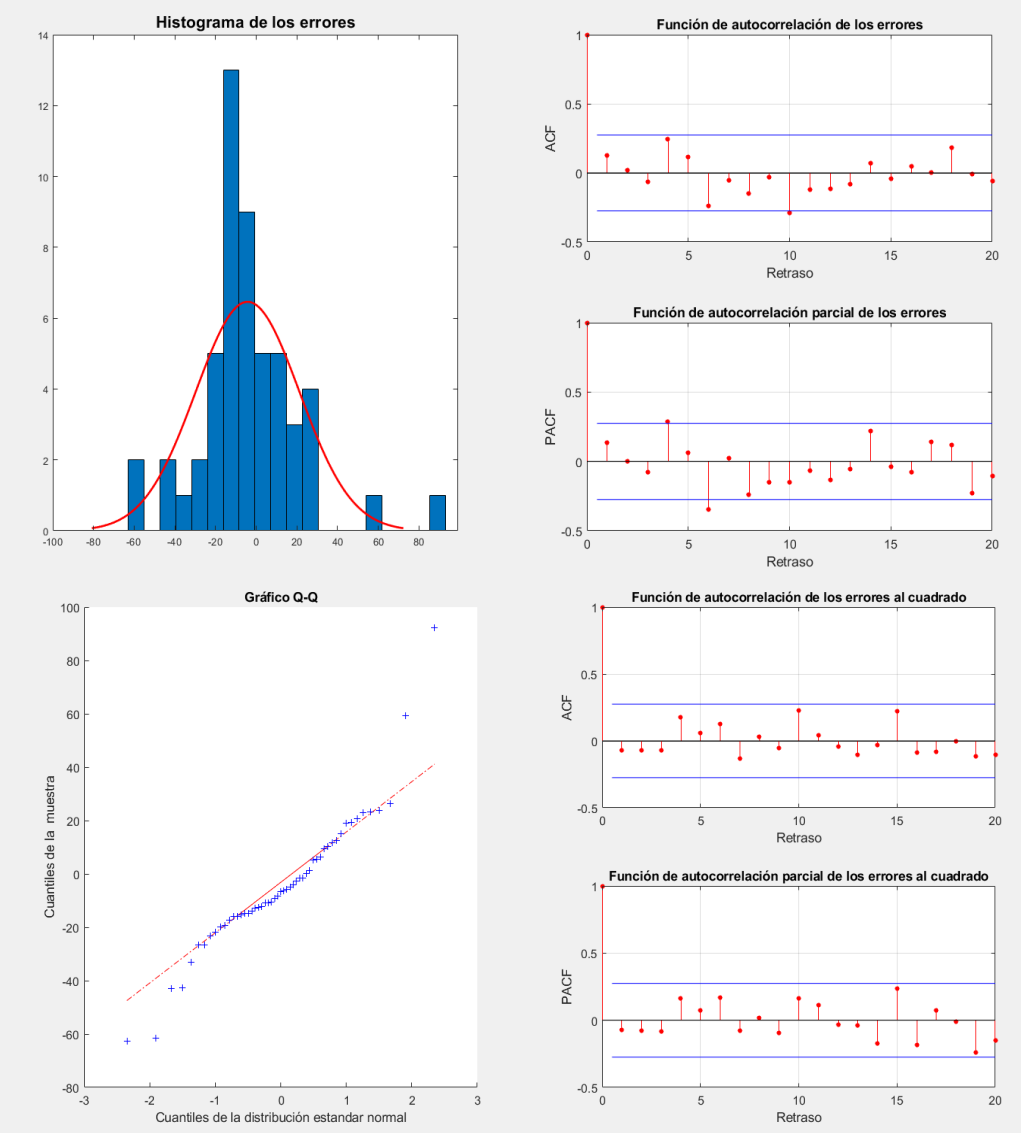


*Figura 4.3-7. Gráficos de los errores en el pronóstico de la serie cada semana, modelo ARIMA.*

#### *4.3.2.2. Intervalos de 2 semanas*

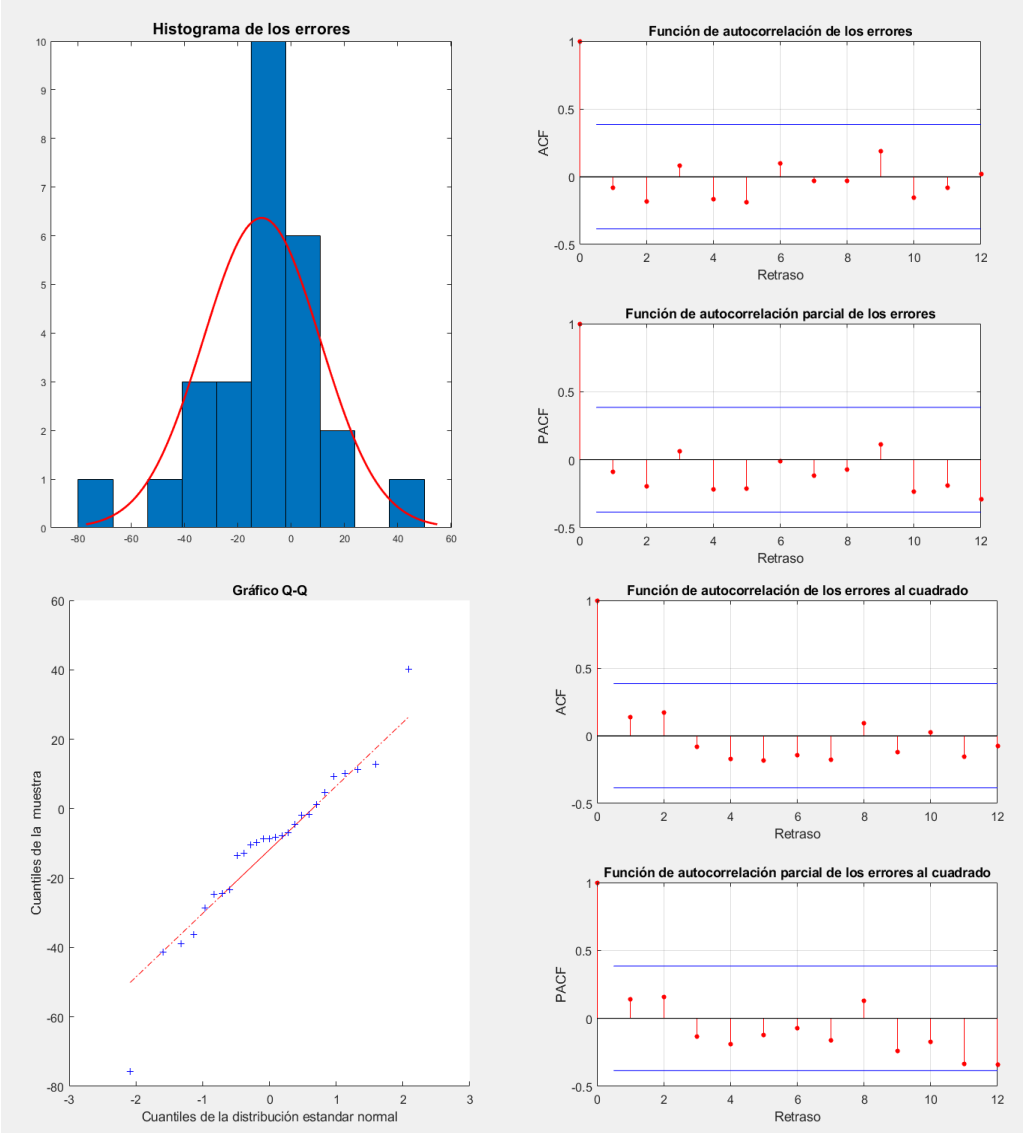


*Figura 4.3-8. Gráficos de los errores en el pronóstico de la serie cada 2 semanas, modelo ARIMA*

#### *4.3.2.3. Intervalos de 1 mes*

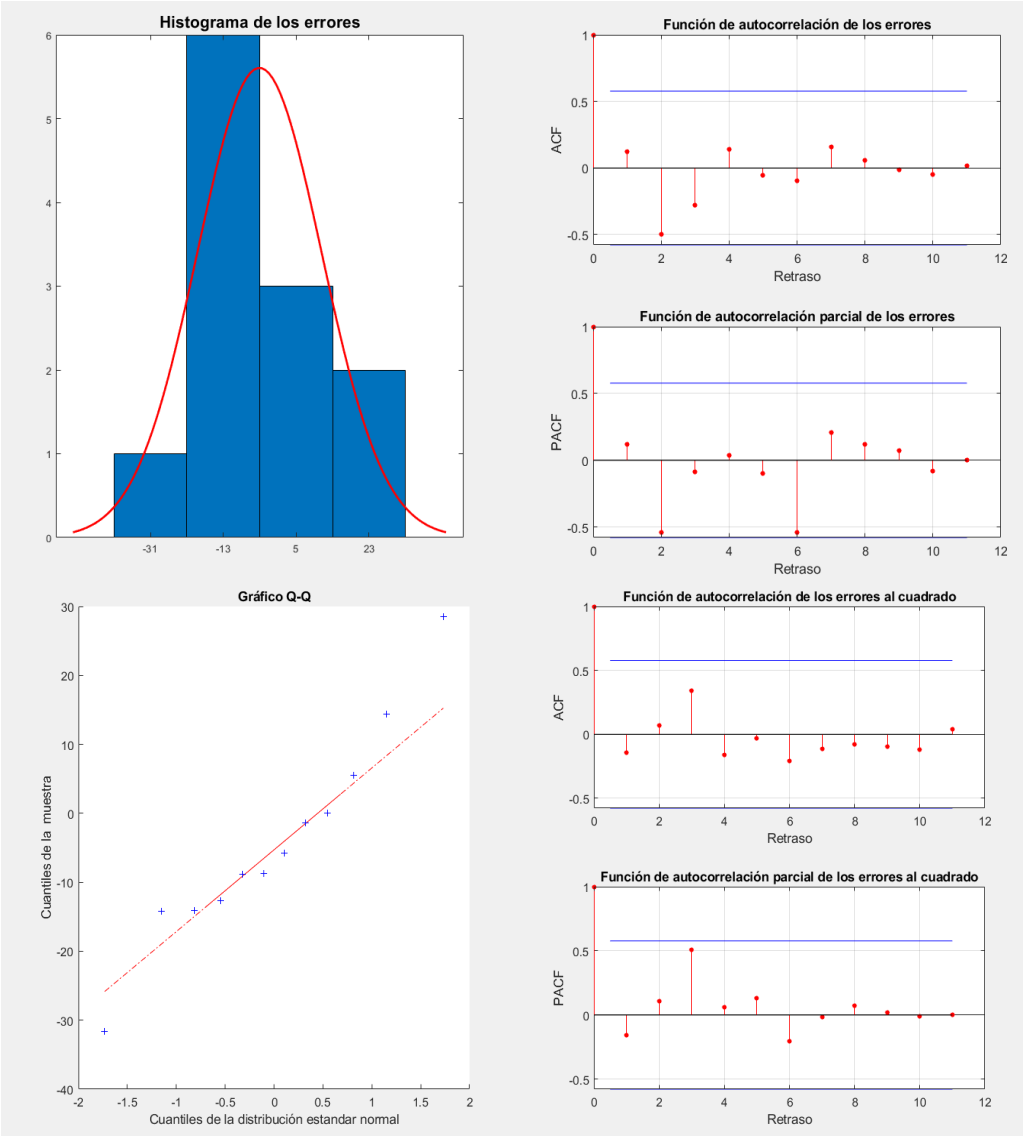


*Figura 4.3-9. Gráficos de los errores en el pronóstico de la serie cada mes, modelo ARIMA.*

De igual manera los modelos ARIMA, para los intervalos con los que fue posible predecir, cumplen con todos los supuestos de los errores, no obstante, las magnitudes de sus errores son mayores que los del modelo de Thomas-Fiering como se observó en la anterior sección.

### 4.3.3. Comprobación de los supuestos de los errores

En color rojo se representa aquellas características de las series que no cumplen con lo supuesto para los errores en las regresiones, mientras que en verde aquellos que si lo hacen.

| Discretización de la serie | ¿Residuos Aleatorios? | ¿Homocedasticidad en los residuos? | ¿Normalidad en los residuos? |
|---|---|---|---|
| **Modelo de Thomas-Fiering** | | | |
| 15minutos | | | |
| 1 hora | | | |
| 1 día | | | |
| 1 semana | | | |
| 2 semanas | | | |
| 1 mes | | | |
| **Modelo ARIMA** | | | |
| 1 semana | | | |
| 2 semanas | | | |
| 1 mes | | | |

*Tabla 4.3-1.Verificación de los supuestos de los errores de las series generadas*

## 4.4. Comparación de la media y desviación estándar de la distribución mensual de caudales

Se comprueba la hipótesis planteada de que las distribuciones mensuales de los caudales tienen los mismos parámetros de media y desviación estándar. Dichos estadísticos son encontrados al agrupar los pronósticos hallados mediante el método de Thomas-Fiering en los meses correspondientes.

### 4.4.1. Pronóstico cada 15minutos

- Media

| | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Observado** | 39,26 | 15,99 | 42,25 | 58,14 | 69,69 | 61,96 | 63,02 | 52,86 | 31,45 | 25,20 | 31,52 | 41,57 |
| **Pronosticado** | 30,40 | 27,85 | 42,45 | 39,06 | 42,90 | 57,51 | 60,45 | 43,99 | 32,14 | 32,34 | 26,50 | 29,52 |

*Tabla 4.4-1. Comparación de las medias de los registros de cada mes para las series cada 15minutos.*

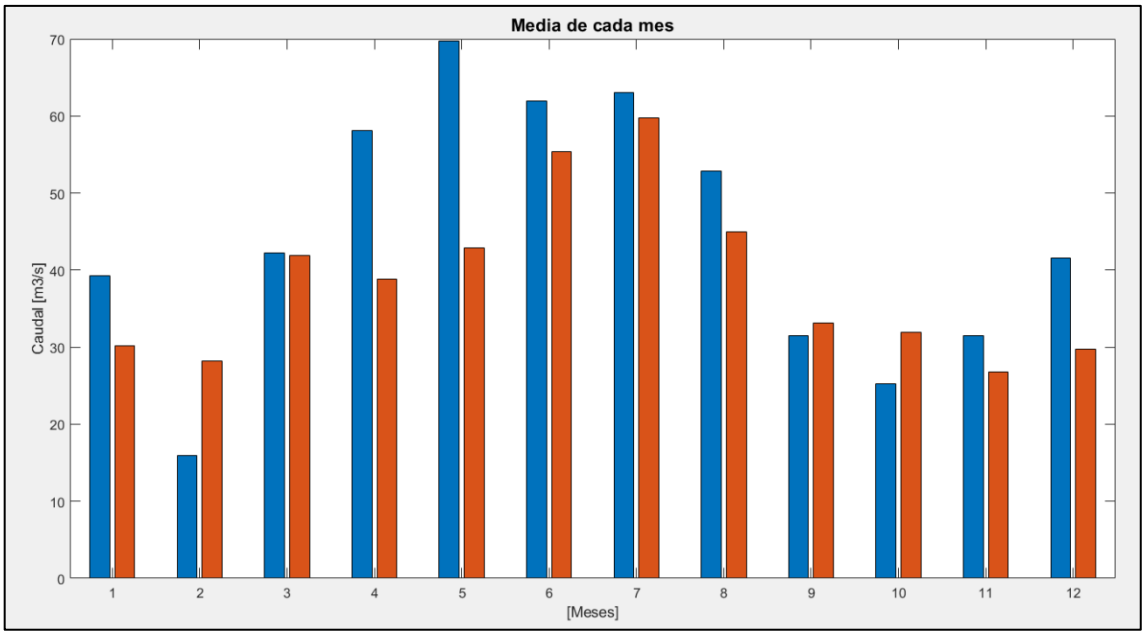


*Figura 4.4-1. Media de cada mes para la serie cada 15minutos.*

- Desviación estándar

| | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Observado** | 39,26 | 15,99 | 42,24 | 58,14 | 69,70 | 61,96 | 63,02 | 52,87 | 31,45 | 25,20 | 31,51 | 41,57 |
| **Pronosticado** | 31,08 | 28,50 | 40,93 | 38,93 | 42,76 | 55,85 | 59,33 | 44,95 | 32,58 | 32,37 | 26,66 | 29,86 |

*Tabla 4.4-2. Comparación de las desviaciones estándar de los registros de cada mes para las series cada 15minutos.*

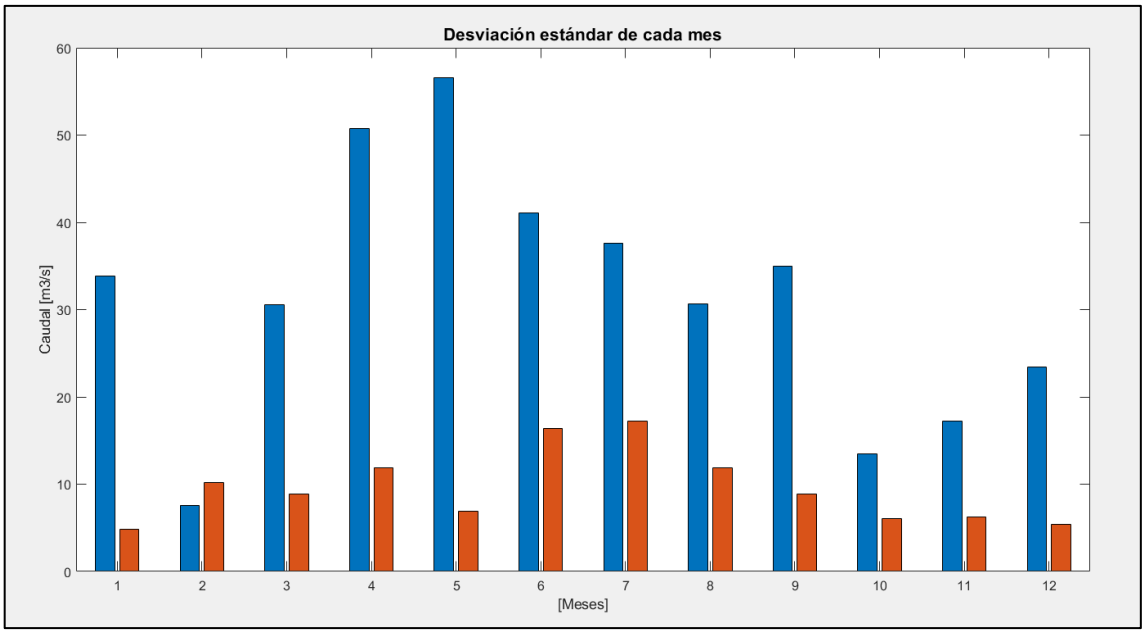


*Figura 4.4-2. Desviación estándar de cada mes para la serie cada 15minutos.*

### 4.4.2. Pronóstico cada hora

- Media

| | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Observado** | 39,26 | 15,99 | 42,25 | 58,14 | 69,69 | 61,96 | 63,02 | 52,86 | 31,45 | 25,20 | 31,52 | 41,57 |
| **Pronosticado** | 32,73 | 28,55 | 42,38 | 40,32 | 43,10 | 57,43 | 61,90 | 45,55 | 32,45 | 32,69 | 26,46 | 29,96 |

*Tabla 4.4-3. Comparación de las medias de los registros de cada mes para las series cada hora.*

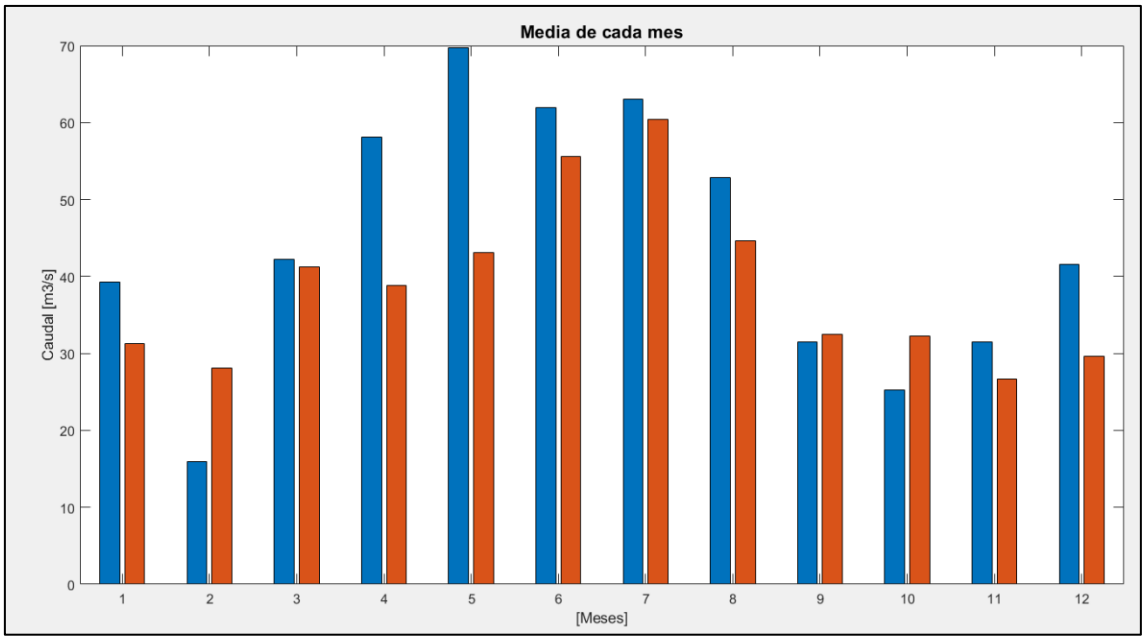


*Figura 4.4-3. Media de cada mes para la serie cada hora.*

- Desviación estándar

| | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Observado** | 33,81 | 7,60 | 30,56 | 50,77 | 56,62 | 41,05 | 37,64 | 30,67 | 35,01 | 13,45 | 17,27 | 23,42 |
| **Pronosticado** | 4,76 | 10,07 | 9,91 | 12,17 | 6,93 | 17,36 | 17,78 | 12,21 | 8,77 | 6,38 | 6,43 | 5,26 |

*Tabla 4.4-4. Comparación de las desviaciones estándar de los registros de cada mes para las series cada hora.*

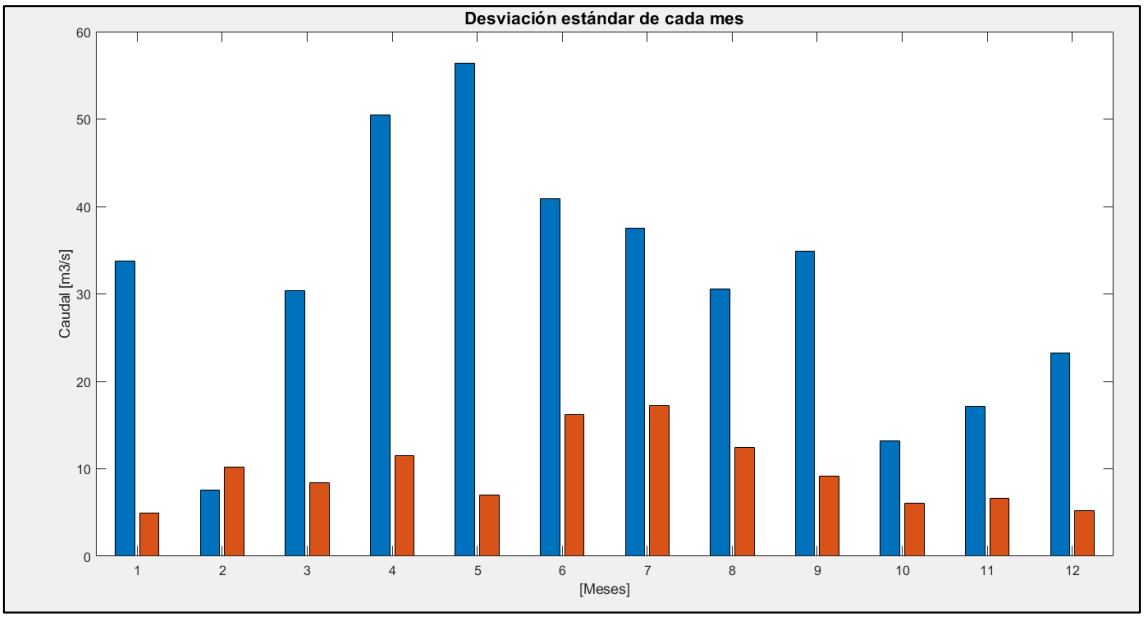


*Figura 4.4-4. Desviación estándar de cada mes para la serie cada hora*

### 4.4.3. Pronóstico cada día

- Media

| | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Observado** | 33,74 | 7,58 | 30,38 | 50,50 | 56,42 | 40,94 | 37,48 | 30,56 | 34,87 | 13,15 | 17,17 | 23,27 |
| **Pronosticado** | 4,84 | 10,28 | 8,46 | 12,00 | 6,94 | 16,35 | 15,80 | 12,48 | 9,22 | 6,31 | 6,45 | 5,46 |

*Tabla 4.4-5. Comparación de las medias de los registros de cada mes para las series cada día.*

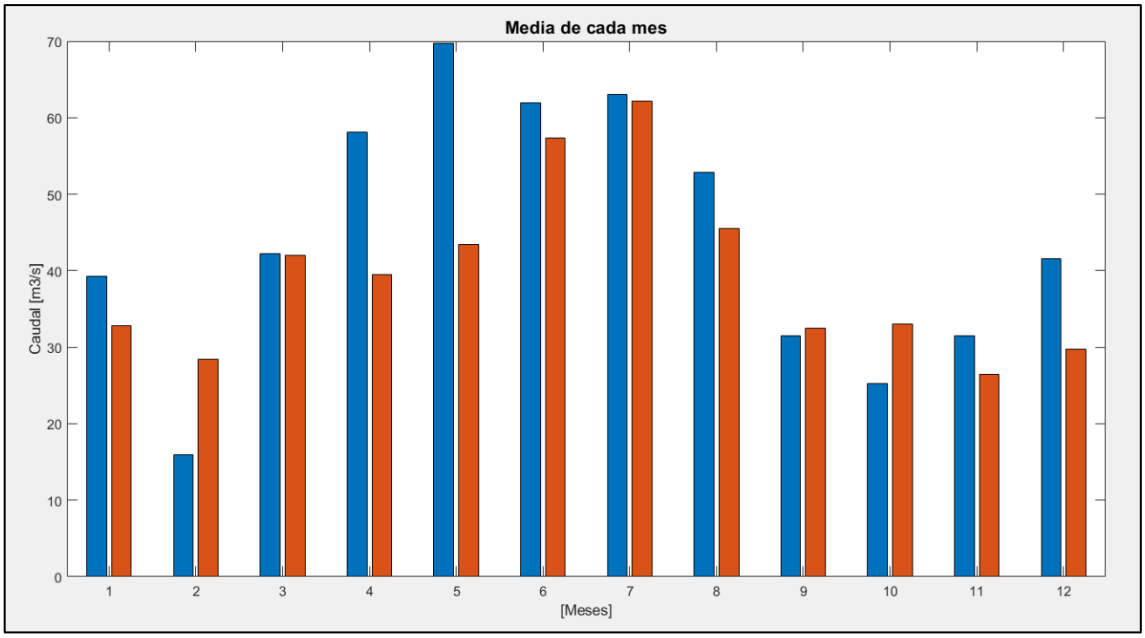


*Figura 4.4-5. Media de cada mes para la serie cada día*

- Desviación estándar

| | Enero | Febrero | Marzo | Abril | Mayo | Junio | Julio | Agosto | Septiembre | Octubre | Noviembre | Diciembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Observado** | 30,94 | 7,10 | 24,68 | 42,56 | 49,18 | 36,03 | 29,38 | 25,21 | 29,06 | 9,34 | 13,81 | 17,27 |
| **Pronosticado** | 4,33 | 10,26 | 8,33 | 12,93 | 6,03 | 15,41 | 16,79 | 11,50 | 8,80 | 5,21 | 6,34 | 5,48 |

*Tabla 4.4-6. Comparación de las desviaciones estándar de los registros de cada mes para las series cada día.*

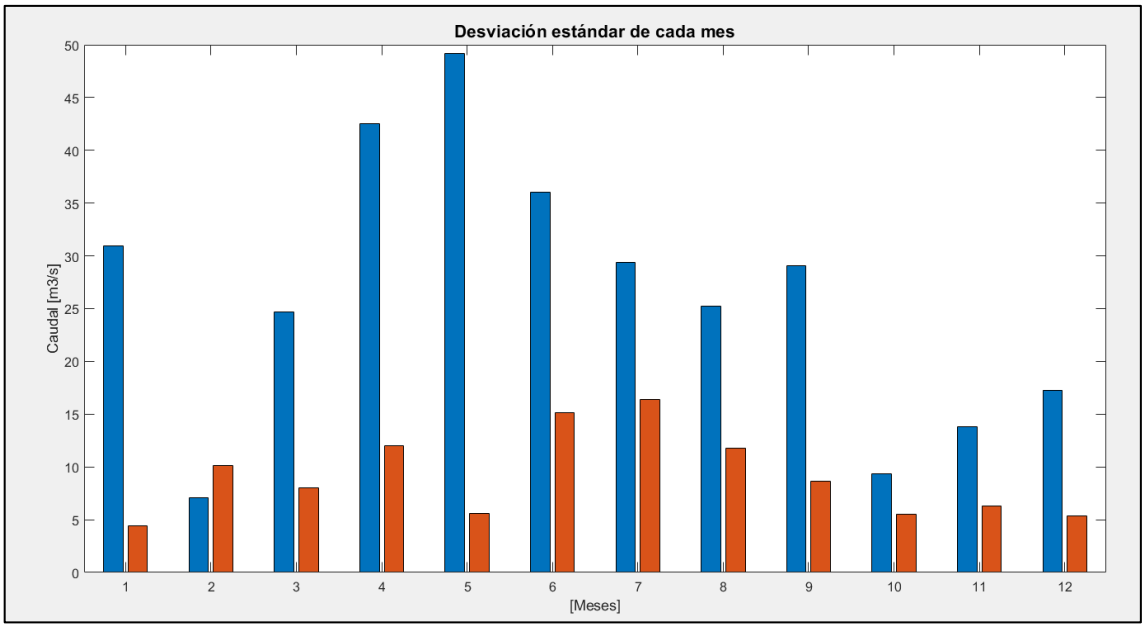


*Figura 4.4-6. Desviación estándar de cada mes para la serie cada hora.*

La serie, en estos tres intervalos, conserva adecuadamente las medias de los meses, pero no lo hacen con las desviaciones estándar, la cual en general es menor, probablemente porque la componente estacional de la serie:

$$Q_{i_1} = \bar{Q}_j$$

Es mucho más significativa que los demás sumandos de la ecuación.

# 5. CONCLUSIONES

De acuerdo a los cálculos realizados para comprobar la eficiencia de los modelos, el modelo de Thomas y Fiering, propuesto en un principio para este proyecto, es aplicable para las características y el número de datos disponibles del río en intervalos mayores o iguales a una semana.

A partir de los resultados obtenidos en este estudio, es posible notar que a medida que se utilizan intervalos de tiempo más cortos, la distribución de los datos y de los logaritmos de los datos se asemeja más a una distribución exponencial, mientras que, con intervalos más grandes, estos tienden más a la normalidad. Esta es una de las razones para que los resultados obtenidos con una discretización menor o igual a la diaria tengan errores tan grandes, ya que una de las suposiciones del modelo de Thomas-Fiering es la normalidad en los datos de calibración, propiedad que si se logra conservan en los caudales y logaritmos de las series semanales, cada 2 semanas y mensuales.

Luego de analizar las predicciones a lo largo de un año, con intervalos de 15minutos, 1 hora y 1 día, se aprecia que éstas logran conservar la media de cada mes, pero no sucede lo mismo con la desviación estándar, la cual en general, es mucho menor. Además, para estas series, los residuos tienen alta autocorrelación para distintos retrasos, debido a la acumulación de errores en el pronóstico, que se debe a que existe un número significativo de valores a predecir.

En base a lo evidenciado, series con mayor frecuencia logran capturar de mejor manera la realidad de los fenómenos, para el caso presente, se observa que para intervalos más cortos, los caudales mínimos de las series de prueba son de menor magnitud, mientras que los máximos, de mayor magnitud, además, la distribución de los datos son más afines a lo observado en el río, que en general, tienen una asimetría positiva, es decir los caudales bajos se presentan con mayor frecuencia, esto se presenta también en datos de caudales disponibles al público como es el caso de los anuarios hidrológicos del INAMHI.

De entre los modelos generados, el que tiene mejor desempeño es el de Thomas-Fiering para la serie mensual, sin embargo, los modelos de las series semanales y de cada 2 semanas también cumplen los supuestos de las regresiones acerca de los errores y tienen métricas relativamente buenas. Los modelos SARIMA (1,0,0)(2,1,2)52; (1,0,0)(2,1,2)27; (1,0,0)(2,1,2)12, utilizados para el pronóstico semanal, de cada dos semanas y mensual

respectivamente, presentan un peor rendimiento en cuanto al pronóstico de los datos futuros, ya que tienen errores mayores y un coeficiente de eficiencia de Nash-Suttclife modificado menor que sus análogos de Thomas-Fiering para dichas series.

Se puede observar en éste y otros estudios, que la diferenciación estacional en los modelos ARIMA, vuelven a los nuevos datos predichos altamente dependientes de sus correspondientes en la estación anterior, por lo que sistemas hidrológicos muy cambiantes en el tiempo pueden tener problemas para prever el futuro con esta metodología.

Es muy posible que el usar un registro más grande para la generación de un modelo se traduzca en una mejora en la eficiencia de este, muestras más grandes de una población, en este caso, de los caudales, reducen la incertidumbre al momento de inferir características sobre ella, ya que la distribución de datos es más acorde y, por tanto, se comprende mejor el sistema.

# 6. RECOMENDACIONES

Deben ser comprobadas, en lo posible, al menos 2 alternativas para el pronóstico de los caudales en una estación y ser comparadas mediante las mismas métricas, para conocer cuál de ellas es la mejor objetivamente. La generación de datos sintéticos debe realizarse en un horizonte igual al periodo de validación con el que fue verificado el modelo y en lo posible, este horizonte no debe ser muy grande.

Es muy probable que la precisión del modelo incremente conforme aumente la cantidad de información para la calibración, por lo que será prudente actualizar las ecuaciones conforme se obtengan nuevos datos; igualmente se deberá comprobar, en la misma medida, la bondad del modelo actualizado.

Para la utilización de estos modelos en otras centrales hidroeléctricas, se sugiere que se realicen estudios similares para analizar su validez en cuencas con diferentes características hidrológicas.

Las transformaciones de los datos de las series en hidrología son frecuentes y tienen el objetivo de lograr las condiciones necesarias para la aplicación de los modelos, y, por lo tanto, deberían considerarse previa la conformación de las ecuaciones generadoras de datos sintéticos, como es el caso de los logaritmos utilizados en el presente caso, que ayudan de cierta manera a reducir la asimetría en la distribución de los datos.

Para el caso del modelo de Thomas-Fiering, los periodos de tiempo entre datos generados no deberían ser más pequeños que las semanas, debido a que esto infringe con la suposición en la normalidad en los datos. El utilizar estos intervalos tiene como resultado mejores pronósticos, a pesar de la pérdida de la precisión de la variable aleatoria caudales que supone promediar los datos cada 15 minutos para pasar a periodos más largos.

Comprobar si existen registros de estaciones cercanas y realizar estudios para el análisis de correlaciones, ya que pueden servir en el caso de que los registros de la estación de interés no se encuentren disponibles o también para expandir el estudio hacia predicciones multivariables, usando métodos existentes como el de Matalas.

En lo posible, el número de pronósticos, para los dos modelos utilizados en este estudio, no debería exceder de 50, debido a que se presentará convergencia en la serie estacionaria y solo permanecerá la parte estacional en el modelo.

La utilización del modelo Thomas-Fiering debe ser considerada para el pronóstico de caudales en el río Topo en la operación de la referida central, debido a que este no requiere gran capacidad computacional, ni tampoco conocimientos avanzados en matemática o estadística para entender sus mecanismos, además de un corto tiempo de cálculo. Asimismo, mediante una actualización constante de los datos que generan el modelo, se puede obtener buenas predicciones al corto plazo, debido a que estas no tienen errores acumulados de predicciones en instantes anteriores, sino que dependen de observaciones.

# 7. BIBLIOGRAFIA


Alfa, M., & Adie, D. (2018). Reliability assessment of Thomas Fiering's method of stream flow prediction. Plateau, Nigeria: University of Jos, Department of civil Engineering.

arima. (25 de Noviembre de 2019). Obtenido de https://www.mathworks.com/help/econ/arima.html;jsessionid=24261ef91caad110eb008565f03a

Arselan, C. (August de 2012). Stream flow Simulation and Synthetic Flow Calculation by Modified Thomas Fiering Model. Kirkuk, Iraq: College Of Engineering, University of Kirkuk.

Balance Nacional de Energía Eléctrica – ARCONEL. (23 de enero de 2020). Obtenido de https://www.regulacionelectrica.gob.ec/balance-nacional/

Beven, K. (2018). WIREs Water. Obtenido de On hypothesis testing in hydrology: Why falsification of models is still a really good idea.: https://doi.org/10.1002/wat2.1278

Box, G., & Jenkins, G. (1970). Time Series Analysis Forecasting and Control. San Francisco: Holden-Day.

Buteikis, A. (Febrero de 2018). Vilnius University. Obtenido de Time series with trend and seasonality: http://web.vu.lt/mif/a.buteikis/wp-content/uploads/2018/02/Lecture_03.pdf

Cheng , C.-T., Liao, S.-L., Liu, C.-X., & Li , G. (2 de Septiembre de 2015). Applying a Correlation Analysis Method to Long-Term Forecasting of Power Production at Small Hydropower Plants. Dalian, Liaoning, China: Institute of Hydropower and Hydroinformatics, Dalian University of Technology, Dalian 116024, China; .

Engle’s ARCH Test. (25 de Noviembre de 2019). Obtenido de https://www.mathworks.com/help/econ/engles-arch-test.html

Estrategia Nacional para el cambio de la matriz productiva. (23 de enero de 2020). Obtenido de https://www.vicepresidencia.gob.ec/wp-content/uploads/2013/10/ENCMPweb.pdf

Farahmand, T. (16 de Septiembre de 2011). Aquatic Informatics: Water Data Management Software. Obtenido de https://aquaticinformatics.com/blog/hydrology/hydrological-measurements-part-1/

Fiering , M., & Jackson, B. (1971). Synthetic Streamflows. Washington, D.C., Estados Unidos: American Geophysical Union.

FutureWater. (2019). Obtenido de Modelos hidrológicos: https://www.futurewater.es/metodos/modelos-hidrologicos/

Galit Shmueli. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/channel/UCiO5ShvaPDsbw-3_zR3Ht6w

Hydroelectric Power Generation. (23 de enero de 2020). Obtenido de https://www.mpoweruk.com/hydro_power.htm

Hydrograph - an overview | ScienceDirect Topics. (s.f.). Obtenido de Sciencedirect.com: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/hydrograph

Hyndman, R., & Athanasopoulos, G. (2018). Forecasting: principles and practice (Segunda ed.). Melbourne, Australia: OTexts. Recuperado el 3 de Noviembre de 2019, de OTexts.com/fpp2

Jonsdottir, H. (2006). Stochastic Modelling of Hydrologic Systems. Lyngby, Denmark: Technical University of Denmark.

Lawrance, A., & Kottegoda, N. (1977). Stochastic Modelling of Riverflow Time Series. Journal of the Royal Statistical Society. Series A (General), Vol. 140, No. 1, pp.1-47.

Medda, S., & Bhar, K. (4 de April de 2019). Comparison of single-site and multi-site stochastic models. Shibpur, West Bengal, India: Civil Engineering Department, Indian Institute.

MIT OpenCourseWare. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/channel/UCEBb1b_L6zDS3xTUrIALZOw

MMSE Forecasting of Conditional Mean Models. (25 de Noviembre de 2019). Obtenido de https://www.mathworks.com/help/econ/mmse-forecasting-for-arima-models.html

Monsalve Sáenz, G. (1999). Hidrología en la ingenieriá. México, D.F: AlfaOmega.

Monte Carlo Forecasting of Conditional Variance Models. (25 de Noviembre de 2019). Obtenido de https://www.mathworks.com/help/econ/monte-carlo-forecasting-of-conditional-variance-models.html

Monte Carlo Simulation of Conditional Mean Models. (25 de Noviembre de 2019). Obtenido de https://www.mathworks.com/help/econ/monte-carlo-simulation-of-arima-models.html

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., & Veith, T. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions Of The ASABE, 50(3), 885-900. doi: 10.13031/2013.23153

Nau, R. (2 de Enero de 2019). Statistical forecasting: notes on regression and time series analysis. Obtenido de http://people.duke.edu/~rnau/411home.htm

nptelhrd. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/channel/UC640y4UvDAlya_WOj5U4pfA

Pi seems a good random number generator, but not always the best. (23 de enero de 2020). Obtenido de https://www.purdue.edu/uns/html4ever/2005/050426.Fischbach.pi.html

Plan Maestro de Electrificación 2016-2025. (23 de enero de 2020). Obtenido de https://www.celec.gob.ec/hidroagoyan/images/PME%202016-2025.pdf

Powell, V. (25 de Noviembre de 2019). Markov Chains. Obtenido de http://setosa.io/ev/markov-chains/

Professor Knudson. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/channel/UCjknLK_siVSCY14qfDu-f-w

Proyectos Energéticos Ecuagesa S.A. (23 de enero de 2020). Obtenido de http://ecuagesa.com.ec/

Rasmus Pedersen. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/channel/UCwpqtj2ztnVDc7yw49zfC8A

RENEWABLES 2019: Global Status Report. (23 de enero de 2020). Obtenido de https://www.ren21.net/wp-content/uploads/2019/05/gsr_2019_full_report_en.pdf

Residual Diagnostics. (25 de Noviembre de 2019). Obtenido de https://www.mathworks.com/help/econ/residual-diagnostics.html

Ritter, A., & Muñoz-Carpena, R. (2012). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. Retrieved 26 January 2020, from https://www.sciencedirect.com/science/article/pii/S0022169412010608

Ritvikmath. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/channel/UCUcpVoi5KkJmnE3bvEhHR0Q

Salazar, J., & Cadavid, J. (2008). Generación de series sinteticas de caudales usando un Modelo Matalas con medias condicionadas. Avances en Recursos Hidráulicos., 17-24.

simulate:. (25 de Noviembre de 2019). Obtenido de Monte Carlo simulation of ARIMA or ARIMAX models: https://www.mathworks.com/help/econ/arima.simulate.html

Stack Exchange. (2019). Obtenido de Cross Validated: https://stats.stackexchange.com/

Thelin, S. (25 de Noviembre de 2019). Forecaster's Toolbox: How to Perform Monte Carlo Simulations. Obtenido de Forecaster's Toolbox: How to Perform Monte Carlo Simulations

USGS. (25 de Noviembre de 2019). United States Geological Survey. Obtenido de Rivers, Streams, and Creeks: https://www.usgs.gov/special-topic/water-science-school/science/rivers-streams-and-creeks?qt-science_center_objects=0#qt-science_center_objects

van Geer, F. C., & P. Bierkens, M. F. (2012). Stochastic Hydrology. Netherlands: Department of Physical Geography, Utrecht University.

Vetter, T., Huang, S., Aich, V., Yang, T., Wang, X., Krysanova, V., & Hattermann, F. (2015). Multi-model climate impact assessment and intercomparison for three large-scale river basins on three continents. Obtenido de Earth Syst. Dynam., 6, 17–43: https://doi.org/10.5194/esd-6-17-2015, 2015.

Wasserman, L. (2013). All of Statistics. New York, NY: Springer.

Woodard & Curran, I. (2006). En Industrial Waste Treatment Handbook (Second Edition), (págs. 127-147). Butterworth-Heinemann. Obtenido de http://www.sciencedirect.com/science/article/pii/B9780750679633500084

World Meteorological Organization,Commission for Hydrology (CHy). (2009). Guide to Hydrological Practices. Hydrological Systems Modeling. Geneva,, Switzerland.

Xu, C. (2009). Statistical Methods In Hydrology (Lecture notes). Uppsala, Sweden: Uppsala University.

Yépez, M. (2015). "Los recursos naturales y el manejo de cuencas hidrográficas" (Pregrado). PUCE.

Young, G., & Pisano, W. (Julio de 1968). Operational hydrology using residuals. Journals of the Hydraulic Division

zedstatistics. (s.f.). Inicio [Canal de Youtube]. Obtenido de https://www.youtube.com/user/zedstatistics

# 8. ANEXOS

- **Estimación de parámetros de un modelo MA(q) por máxima verosimilitud**

A partir de un modelo de media móvil de orden "q":

$$MA(q): S_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

De donde los residuos se esperan independientes entre ellos y están distribuidos normalmente:

$$\epsilon_t \sim N(0, \sigma^2)$$

La varianza y la esperanza de cada nuevo valor de la serie son:

$$var(S_t) = var\left(\mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}\right) = var(\epsilon_t) = \sigma^2$$

$$E[S_t] = E\left[\mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}\right] = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Debido a que, al momento de pronosticar, los residuos experimentados por anteriores iteraciones, $\epsilon_{t-i}$, y la media de la muestra, $\mu$, son fijas en el tiempo.

La distribución de los pronósticos, debido a la linealidad de la distribución normal al momento de ser operada es:

$$S_t \sim N(\mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \sigma^2)$$

Ya que la distribución de probabilidad a diferentes instantes, es independiente de distribuciones anteriores, el vector de parámetros que vuelve al conjunto de los pronósticos, más probable a partir de lo ya observado, es el estimador de máxima verosimilitud:

$$\theta: (\mu, \theta_i, \sigma^2)$$

Para cualquier tiempo t, condicionado para los errores que sucedieron hasta un retraso "q":

$$S_t \mid \epsilon_{t-i}, \epsilon_{t-2} \dots \epsilon_{t-q} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-(S_t - E[S_t])^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-{\epsilon_t}^2}{2\sigma^2}\right]$$

Que corresponde, para cada instante, la componente de la función de máxima verosimilitud:

$$\mathrm{f}\,(\,x_i \mid \theta\,)$$

Se conoce que el primer residuo, $\epsilon_0$, es igual a 0, porque coincide con el último conocido de la serie que se quiere pronosticar:

$$\epsilon_0 = 0$$

Despejando de la definición del modelo de media móvil, es posible realizar una aproximación de los errores, a partir de los valores ya registrados:

$$\epsilon_1 = S_1 - \mu - \sum_{i=1}^{q} \theta_i \epsilon_{1-i} = S_1 - \mu$$

$$\epsilon_2 = S_2 - \mu - \sum_{i=1}^{q} \theta_i \epsilon_{2-i}$$

$$\vdots$$

$$\epsilon_n = S_n - \mu - \sum_{i=1}^{q} \theta_i \epsilon_{n-i}$$

- **Estimación de parámetros de un modelo AR(p) por máxima verosimilitud**

De la misma manera, de un modelo autoregresivo de orden “p”:

$$AR(p): S_t = \sum_{i=1}^{p} \phi_i S_{t-i} + \epsilon_t$$

Los residuos se esperan independientes entre ellos y distribuidos normalmente:

$$\epsilon_t \sim N(0, \sigma^2)$$

La varianza y la esperanza de cada nuevo valor de la serie son:

$$var(S_t) = var\left(\sum_{i=1}^{p} \phi_i S_{t-i} + \epsilon_t\right) = var(\epsilon_t) = \sigma^2$$

$$E[S_t] = E\left[\sum_{i=1}^{p} \phi_i S_{t-i} + \epsilon_t\right] = \sum_{i=1}^{p} \phi_i S_{t-i}$$

La distribución de probabilidad de cada pronóstico, está definida por una curva normal:

$$S_t \sim N(\mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \sigma^2)$$

De donde el vector estimador de máxima verosimilitud es igual a:

$$\theta: (\mu, \theta_i, \sigma^2)$$

Para cualquier tiempo t, condicionado para los errores que sucedieron hasta un retraso "p":

$$S_t \mid \epsilon_{t-i}, \epsilon_{t-2} \dots \epsilon_{t-q} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-(S_t - E[S_t])^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-\left(S_t - \sum_{i=1}^{p} \phi_i S_{t-i}\right)^2}{2\sigma^2}\right]$$

Que corresponde, a los multiplicandos de la función de máxima verosimilitud:

$$\mathrm{f}\,(\,x_i \mid \theta\,)$$

- **Operador de Retardo**

Es un operador utilizado en la literatura de las series de tiempo para simplificar la redacción de las ecuaciones de distintos modelos. El operador de retardo, "retrasa" a los registros una cierta cantidad de intervalos:

$$By_t = y_{t-1}$$

Al aplicarlo por una segunda vez:

$$By_{t-1} = B.By_t = B^2 y_t$$

Para hacer referencia a un dato "k" intervalos atrás:

$$B^k y_t = y_{t-k}$$

Por ejemplo, para un modelo ARMA (p,q):

$$ARMA(p,q): S_t = \phi_1 S_{t-1} + \phi_2 S_{t-2} + \cdots + \phi_p S_{t-p}$$

$$+\theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Se puede reorganizar sus elementos de la manera:

$$S_t - \phi_1 S_{t-1} - \phi_2 S_{t-2} - \cdots - \phi_p S_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

$$S_t(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_P B^P) = \epsilon_t(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)$$

De aquí se pueden encontrar los polinomios de retraso correspondiente a los residuos y valores de la serie, definidos por las letras phi y theta mayúsculas:

$$\Phi(B) S_t = \Theta(B) \epsilon_t$$