
# Modelo Estocástico a partir de Razonamiento Basado en Casos para la Generación de Series Temporales

Por
**José Alfredo Herrera Quispe**

Tesis presentada en el
Doctorado en Ciencias de la Computación
de la
UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

Universidad Nacional de San Agustín
Facultad de Ingeniería de Producción y Servicios
Doctorado en Ciencias de la Computación

# Modelo Estocástico a partir de Razonamiento Basado en Casos para la Generación de Series Temporales

Presentado por el Magister José Alfredo Herrera Quispe

Arequipa, 22 de octubre de 2013

Aprobado por:

---

Prof. Dr. Yvan Tupac Valdivia
Orientador

---

Prof. Dr. Jose Eduardo Ochoa
Co-revisor

---

Prof. Dr. Luis Alfaro Casas
Co-orientador

*A Mamá*

# Agradecimientos


- Al Consejo Nacional de Ciencia Tecnología e Innovación Tecnológica, CONCYTEC; por el otorgamiento de una Beca de estudios y las acciones de apoyo en el Doctorado en Ciencia de la Computación de la UNSA.
- Al Investigador principal de la CATEDRA CONCYTEC en TICs Dr. Luis Alfaro Casas y todos los profesores del doctorado, por las acciones de seguimiento de esta Tesis.
- Al Profesor Asesor Dr. Yvan Tupac Valdivia por todo su apoyo académico en el presente trabajo de Tesis.
- Al equipo de Investigación del CIDES-UNSA, Christian Portugal, Herbert Chuctaya, Jorge Suaña, Julio Vera y Edson Luque por sus apoyo complementario en las tareas de corrección e impresión de esta Tesis.
- A mi señor padre, Don Alejandro Herrera por su apoyo emocional, incondicional y estímulo para la finalización de la presente.
- A mi familia, Giovanna, Henry, Antonio, Matías y Fabiola por su cariño, una motivación para seguir adelante.

# Índice general

# Índice de figuras

# Índice de cuadros

# Glosario

- PE: Proceso Estocástico
- PEN: Proceso Estocástico Neuronal
- PERBC: Proceso Estocástico a partir de Razonamiento Basado en Casos
- RBC: Razonamiento Basado en Casos
- TF: Thomas Fiering
- MSE: Error medio estándar
- MRSE: Raiz del error medio estándar

# Resumen

Se propone un nuevo modelo estocástico a partir del Razonamiento Basado en Casos para la generación de series temporales, esta propuesta extiende los modelos con memoria auto-regresiva, cambiando los parámetros del componente determinístico por una función de similitud que usa la distancia euclidiana multidimensional ponderada y retardos de tiempo; se adjunta un componente aleatorio heredado del modelo de Thomas-Fiering con manejo de umbrales; la propuesta se clasifica como un modelo estocástico periódico auto-regresivo genérico.

El modelo se aplica en la generación de escenarios climáticos en el ámbito de la cuenca del Chili-Arequipa, los resultados muestran que la propuesta genera razonablemente realizaciones que reproducen las características de la serie, particularmente para el caso de valores mínimos extremos, representando una mejora complementaria a los esfuerzos previos de (Campos, 2010) y Taymoor (Awchi, Srivastava, y cols., 2009); luego el uso de casos multidimensionales y de grados superiores genera series leptocúrticas, lo que en ciertos casos no reproduce las características de los datos, pero reduce la incertidumbre. Computacionalmente una estructura de datos de acceso secuencial permite la indexación en memoria de todos los casos facilitando las tareas de búsqueda de relaciones ocultas.

Finalmente, luego de la revisión de los resultados, el modelo se vislumbra como un prometedor complemento en la simulación de escenarios y la modelación de eventos extremos, con un potencial interesante en la toma de decisiones vinculadas al desarrollo de acciones técnicas de previsión, que permitan reducir pérdidas económicas, sociales; dimensionando y escenificando el impacto de una sequía, inundación, helada sobre un área cultivable, sobre la producción hidro-energéticas, la producción minera y la demanda poblacional.

**Palabras clave:** Procesos Estocásticos, Razonamiento Basado en Casos, Series Temporales, Minería de datos.

# Abstract

We propose a new stochastic model from Case-Based Reasoning for generating time series, this proposal extends the autoregressive memory models, changing the deterministic component to a similarity function using the Euclidean distance with weighted multidimensional time delays, we attach a random component inherited from Thomas-Fiering model with threshold management, the proposal is classified as a generic periodic autoregressive stochastic model.

The model is applied to generate climate scenarios in Chili-Arequipa's basin, the results show that the proposal generate realizations that reproduce the characteristics of the series, particularly in the case of minimum values; representing an improvement to previous efforts of (Campos, 2010), Taymoor (Awchi y cols., 2009), and Thomas (Fiering, 1967); finally, multidimensional cases generates leptokurtic series, which in some cases not have the characteristics of analized data, but reduces uncertainty. Computationally, a data structure allows sequential access to memory, indexing all cases and facilitating task search from hidden relationships.

Finally, the model is seen as a promising addition to the scenario simulation and modeling of extreme events, with an interesting potential in the decision-making activities related to development of forecasting techniques; that can reduce economic losses, social, sizing and staging the impact of drought, flood, frost on a cultivable area on hydroenergetic production, mining and population demand.

**Keywords:** Time Series, Stochastic Proces, Case Based Reasoning, Datamining.

# Capítulo 1

# Introducción

---

Muchas variables aleatorias son funciones cuyos valores cambian con el tiempo, se tienen fenómenos climatológicos (Loucks, Van Beek, Stedinger, Dijkman, y Villars, 2005), fenómenos económicos (Hochreiter y Pflug, 2007), fenómenos biológicos (Wilkinson, 2009); un conjunto de estas observaciones son llamadas series temporales a partir de los cuales se generan sintéticamente realizaciones estocásticas utilizadas en tareas de modelado, pronóstico, planificación y toma de decisiones.

Los primeros modelos para la generación sintética de series temporales ensayaron, de manera consistente, la regresión lineal simple, usando modelos Auto-regresivos (AR) y algunas variaciones con Medias Móviles ($ARMA$); con variable exógena ($ARMAX$) (Wei, 1994) entre otros; En todos estos modelos, la relación lineal entre las variables relevantes es asumida, producto de su popularidad, muchos estudios emplean estos modelos con bajo orden para la generación estocástica de series temporales, reproduciendo satisfactoriamente las características analizadas (Salas, Tabios III, y Bartolini, 1985; Kjeldsen y Rosbjerg, 2004). Sin embargo no siempre producen los mejores resultados, apareciendo entonces los modelos multivariados(Raman y Sunilkumar, 1995).Peng

muestra que no hay evidencias que estos modelos en grado AR(1) sean inadecuados (Peng y Buras, 2000); finalmente Thomas Fiering afirma, que un modelo AR1 con coeficientes que varían estacionalmente es ampliamente aceptado para la generación de series temporales de caudales (Brittan, 1961; Julian, 1961; Thomas y Fiering, 1962; Beard y Kubîk, 1967; Fiering, 1967) reproduciendo características especiales como la periodicidad y considerando los efectos de la correlación lineal.(Colston y Wiggert, 1970; Gangyan, Goel, y Bhatt, 2002).

Recientemente Luciana Conceicao (Campos, 2010) y Taymoor (Awchi y cols., 2009) proponen el uso de Redes Neuronales para la generación de series temporales estocásticas, ellos afirman que los modelos tradicionales (aproximaciones lineales) son modelos poco eficientes y de aplicabilidad limitada, luego los modelos no-lineales necesitan un conocimiento profundo del dominio para su construcción (Campos, 2010; Han y Wang, 2009; Kantz y Schreiber, 2004). Una de las características que hacen ventajoso el uso de Redes Neuronales, es la no necesidad de asumir un tipo de distribución a priori, aprenden la distribución a través de ejemplos y manejan datos de diversas fuentes con diferentes niveles de precisión y ruido (Vieira, de Carvalho Júnior, y Solos, s.f.; Prudencio, 2002). Estos modelos manejan fácilmente características complejas como la no-linealidad y el comportamiento caótico; sin embargo por su naturaleza tienden a ocultar características extremas, siendo estas últimas de interés en dominios donde los casos excepcionales deben ser modelados, un ejemplo importante son los fenómenos climáticos y el estudio de eventos extremos (Campos, 2010; Taylor, 2008; Meng, Somani, y Dhar, 2004; El-Shafie y El-Manadely, 2011; Ochoa-Rivera, 2008; Bao y Cao, 2011).

Áreas como el *Soft Computing* y el *Datamining* ofrecen técnicas donde los casos excepcionales son incorporados a la memoria de las generaciones, sin importar su baja

significancia, no es preciso como «Razonamiento Basado en Casos», «Razonamiento Basado en Instacias», «Inferencia a partir de ejemplos». Aquí, todos los registros son manejados por la memoria, las nuevas experiencias y excepciones son significativas y su nivel de importancia es determinado por el contexto, siendo una ventaja sobre los modelos lineales, inductivos, basados en reglas, basados en aprendizaje o abstracciones matemáticas; sus algoritmos de indexación, recuperación, adaptación y retención (De Mantaras y cols., 2005) presentan el marco ideal para implementarlo en ambientes automáticos de generación de series temporales con énfasis en el descubrimiento de características ocultas (Lee, Liu, y Huang, 2010; Lee, Cheng, y Liu, 2008; Loor, Bénard, y Chevaillier, 2011; He, Xu, Means, y Wang, 2009; Smyth y Champin, 2009; Lajmi, Ghedira, y Benslimane, 2006; Pal y Shiu, 2004). Se decidió entonces utilizarlo como base para un nueva forma de generar series temporales llamándose «Procesos Estocásticos a partir de Razonamiento Basado en Casos».

## 1.1. Definición del problema

El comportamiento caótico y la no-linealidad de los datos ha fomentado recientes investigaciones en la generación de series temporales (Kantz y Schreiber, 2004; Campos, 2010), los modelos tradicionales que hacen uso de aproximaciones lineales se han convertido en modelos poco eficientes y de aplicabilidad limitada, los modelos no-lineales, necesitan un conocimiento profundo del dominio para su construcción (Campos, 2010; Han y Wang, 2009). Recientemente se propuso el uso de Redes Neuronales (Campos, 2010), una de las características resaltantes es la no necesidad de asumir un tipo de distribución a priori, aprenden la distribución a través de ejemplos y manejan datos de diversas fuentes con diferentes niveles de precisión y ruido. (Vieira y cols., s.f.; Prudencio, 2002). Luego las nuevas propuestas necesitan hacer una suposición a priori sobre el comportamiento de la serie, algunas realizan una descomposición sobre

la estacionalidad, ciclo o tendencia (Campos, 2010); luego el espacio de generaciones suele ser amplio representando a las soluciones con mayor probabilidad ocultando las características extremas, siendo estas últimas de interés en dominios donde los casos excepcionales deben ser modelados, un ejemplo importante son los fenómenos climáticos y el estudio de eventos extremos (Campos, 2010; Taylor, 2008; Meng y cols., 2004; Tokdemir y Arditi, 1999; El-Shafie y El-Manadely, 2011; Ochoa-Rivera, 2008; Bao y Cao, 2011).

## 1.2. Justificación

Para el modelado de fenómenos climatológicos son ampliamente usados los *modelos auto-regresivos periódicos*, un ejemplo común es el *Modelo de Thomas Fiering*, usado para generar caudales sintéticos, y precipitaciones (Cheng y Bear, 2008; Singh y Yadava, 2003; Ünal, Aksoy, y Akar, 2004; Srikanthan, 2002; Brockwell y Davis, 2009; Jaeger, 2000; Brillinger, 2001); ahora bien existen recientes investigaciones que proponen el uso de Redes Neuronales (Campos, 2010; Kantz y Schreiber, 2004; Han y Wang, 2009; Vieira y cols., s.f.; Prudencio, 2002); todos ellos trabajan bajo dos supuestos, el primero: Existe cierta relación entre un evento y el inmediato próximo, el Segundo: los eventos son periódicos; para el modelado apelan a la generalización a partir de la serie histórica. Ahora bien existen otras técnicas como el *Razonamiento Basado en Casos* (Lee y cols., 2010, 2008; Loor y cols., 2011; He y cols., 2009; Smyth y Champin, 2009; Lajmi y cols., 2006; Pal y Shiu, 2004), que pueden mantener los mismos supuestos (heredarlos) y para el modelado, manejar toda la serie histórica, evitando la perdida de información por generalización, en este contexto una función de similitud permitirá explorar todas las relaciones de dependencia histórica en un evento específico para intentar reproducirlas en el evento inmediato próximo para generar nuevas realizaciones.

## 1.3. Objetivos

Proponer un modelo de Proceso Estocástico para la generación de series temporales con la capacidad de capturar detalles ocultos, con las siguientes características:

- Modelo genérico que puede ser implementado en una amplia gama de fenómenos no lineales de comportamiento estocástico.
- Modelo con la capacidad de manejar todos los casos incorporados a la memoria.
- Modelo auto-regresivo, en series temporales que presenten un fenómeno de persistencia observable.

### 1.3.1. Objetivos Específicos

1. Estudio de los modelos lineales, familia *ARMA* y los modelos *PAR*, junto con una revisión bibliográfica de modelos basados en aprendizaje: Redes Neuronales, luego revisión del Razonamiento Basado en Casos, y su capacidad para mostrar información oculta y manejo de casos en memoria.
2. Propuesta del nuevo modelo a partir de los indicios sobre minería de datos del RBC para encontrar información oculta, adaptación de modelos anteriores a la propuesta, refinamiento del modelo.
3. Aplicación del modelo propuesto en la generación de series temporales para un caso de estudio basado en variables hidrometeorológicas (Caudales, Evaporación, Precipitación) en la cuenca del Chili, en tres estaciones de medición: el Pañe, Aguada blanca y el Frayle, para periodos mensuales.
4. Evaluación del modelo propuesto con el modelo de Thomas Fiering y el Modelo Estocástico Neuronal de Luciana. los parámetros utilizados para evaluar a nivel

mensual son la media, desviación estándar, el coeficiente de asimetría, máximos y mínimos.

5. Análisis detallado de la media, desviación estándar, asimetría para todos los experimentos de los modelos TF, PEN y PERBC para establecer las conclusiones, contribuciones, limitaciones, y trabajo futuro del modelo.

### 1.3.2. Otras aplicaciones

Se espera, que este modelo pueda ser usado también en la generación series temporales financieras, económicas, biológicas, y procesos que presenten un fenómeno de persistencia observable y donde los modelos tradicionales no descubran características ocultas. También en fenómenos, que no requieran una formulación a priori ni procesos de adecuación de las distribuciones. Finalmente, se puede adaptar el modelo para la completación de datos de series temporales.

### 1.3.3. Posibles ventajas y desventajas de la propuesta

Por las características del problema, la ventaja en la generación de series esta ligada a la capacidad del modelo para descubrir características ocultas que los modelos tradicionales no consiguen y la no necesidad de una formulación a priori. Por la naturaleza de los algoritmos del *CBR*, la ventaja es la recuperación de consultas y un proceso automático. La desventaja esta ligada a necesidad de hacer un análisis sobre los residuos para ajustarlo formalmente a un Proceso Estocástico.

## 1.4. Contribuciones del trabajo

- Un modelo de Proceso Estocástico a partir de *Razonamiento Basado en Casos* con la capacidad de descubrir características ocultas, un nuevo modelo con memoria

auto-regresiva, con una función de similitud, y un método de acceso métrico para mejorar la velocidad de recuperación, y de proceso automático.

## 1.5. Descripción de capítulos

**Capitulo 2: Marco Teórico.** Se presenta los fundamentos teóricos de la investigación, se explica el concepto de Variable Aleatoria, modelos lineales ARMA, PARMA; ruido blanco, series temporales, finalmente el Razonamiento Basado en Casos, métodos de acceso métrico y álgebra relacional; todos estos conceptos serán de utilidad para comprender la propuesta.

**Capitulo 3: Estado del Arte.** Se presenta brevemente los modelos de Thomas Fiering, modelos no-lineales, modelos complejos basados en redes neuronales, luego trabajos donde se muestra la capacidad del Razonamiento Basado en Casos para descubrir información oculta.

**Capitulo 4: Propuesta.** Se presenta el nuevo modelo a partir del Razonamiento Basado en Casos; en la etapa de representación, un modelo con memoria a corto plazo, multidimensional; para la indexación una estructura de acceso secuencial; luego la etapa de recuperación, búsqueda y generación de un componente determinístico; finalmente en la etapa de reutilización se presenta una realización estocástica.

**Capitulo 5: Estudio de Caso.** En este capítulo se evalúa la propuesta mediante la generación de variables hidrometeorológicas (Caudales, Evaporación, Precipitación) en la cuenca del Chili, en tres estaciones de medición: el Pañe, Aguada blanca y el Frayle, por periodos mensuales. se comparan el Modelo de Thomas Fiering el Modelo Estocástico Neuronal y la propuesta mediante la media, desviación estándar, el coeficiente de asimetría, máximos y mínimos; finalmente se discute los resultados.

# Capítulo 2

# Marco Teórico

---

En este capítulo se presentará brevemente los fundamentos teóricos para comprender un Proceso Estocástico, se explica el concepto de Variable Aleatoria, modelos lineales ARMA, PARMA y otros; luego el ruido blanco, finalmente se definirá las series temporales y se describirá algunos estimadores usados para caracterizarlas, todos estos conceptos serán de utilidad para comprender las bases sobre la que se desarrollará la propuesta.

## 2.1. Proceso Estocástico

Es un conjunto de variables aleatorias que dependen de un parámetro o argumento. En el análisis de series temporales el argumento es el *Tiempo*. Formalmente es una familia de variables aleatorias $Y_t$ donde $t$ denota el tiempo, tales que para cualquier elección finita de valores de $t : t_1, t_2, ...t_n$ existe la distribución de probabilidad conjunta correspondiente a las variables aleatorias $Y_{t_1}, Y_{t_2}, ...Y_{t_n}$ (Ramirez, 2007)

Los Procesos Estocásticos (PE) es usado en fenómenos donde se contemplan variaciones aleatorias (Cadavid y Salazar, 2008; Wilkinson, 2009; Thomas y Fiering, 1962;

Jaeger, 2000).

### 2.1.1. Variable Aleatoria

Dada una determinada variable aleatoria $Y_t$, supóngase que fueron observadas $T$ muestras

$$\{y_1, y_2, \ldots, \ y_T\} \tag{2.1}$$

Un ejemplo sería tener una colección de $T$ variables $\varepsilon_t$ independientes e idénticamente distribuidas

$$\{\varepsilon_1, \varepsilon_2, \ldots, \ \varepsilon_T\} \tag{2.2}$$

donde:

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

Que se refiere a $T$ muestras de un proceso de *ruido blanco gaussiano*. El ruido blanco gaussiano es una señal aleatoria, caracterizada porque sus valores en instantes de tiempo distintos no tienen relación alguna entre sí, es decir, no existe correlación estadística entre sus valores.

Debemos diferenciar que las muestras de la Ecuación (2.1) son $T$ números que pueden ser una de las posibles generaciones (o realizaciones) del proceso estocástico que está por detrás de ellos. Aunque se puede pensar en generar estos datos hasta tiempo infinito, llegando a la siguiente secuencia:

$$\{y_t\}_{t=-\infty}^{\infty} = \{\ldots, y_{-1}, y_0, y_1, \ldots, \ y_T, y_{T+1}, y_{T+2}, \ldots\} \tag{2.3}$$

Esta secuencia infinita $\{y_t\}_{t=-\infty}^{\infty}$ se puede ver como una única realización de un proceso de serie temporal (en sentido amplio de un proceso estocástico). Si se genera una secuencia $\left\{\varepsilon_t^{(1)}\right\}_{t=-\infty}^{\infty}$ en una computadora, y luego se manda generar otra serie $\left\{\varepsilon_t^{(2)}\right\}_{t=-\infty}^{\infty}$, se puede afirmar que estas son dos realizaciones independientes de un

proceso de *Ruido blanco Gaussiano.*

De esta forma, suponiendo un conjunto de $I$ computadoras generando secuencias $\left\{\varepsilon_t^{(i)}\right\}_{t=-\infty}^{\infty}$, $1 \leq i \leq I$ y pudiendo seleccionar el conjunto de $I$ realizaciones en tiempo $t$ $\left\{\varepsilon_t^{(1)}, \varepsilon_t^{(2)}, \ \ldots \ \varepsilon_t^{(I)}\right\}$. Este conjunto se puede describir como una muestra de $I$ realizaciones de la variable aleatoria $Y_t$.

Esta variable aleatoria posee una densidad $f_{Y_t}(y_t)$ denominada la densidad incondicional de $Y_t$, que para el proceso de *RuidoBlancoGaussiano* se define:

$$f_{Y_t}(y_t) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-y_t^2}{2\sigma^2}} \tag{2.4}$$

**Varianza**

La varianza $\gamma_{0t}$ de una variable aleatoria $Y_t$ se define como

$$\gamma_{0t} \equiv E[Y_t - \mu_t]^2 = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t)\, dy_t \tag{2.5}$$

Para un proceso que representa una tendencia en el tiempo más un ruido gaussiano, la varianza es

$$\gamma_{0t} = E[Y_t - \mu_t]^2 = E\left[\varepsilon_t^2\right]^2 = \sigma^2$$

**Estacionaridad**

Si la media $\mu_t$ y las covarianzas no dependen del tiempo, se puede afirmar que el proceso $Y_t$ es estacionario en la covarianza o con estacionaridad débil, es decir que:

$E[Y_t] = \mu$ para todo $t$ y $E[(Y_t - \mu)(Y_{t-j} - \mu)]$ para todo $t$ y cualquier $j$

Si $\{Y_t\}_{t=-\infty}^{\infty}$, representa la suma de una constante $\mu$ más un ruido gaussiano $\{\varepsilon_t\}_{t=-\infty}^{\infty}$, es estacionario en la covarianza

$$E[Y_t] = \mu$$

$$E\left[\left(Y_t-\mu\right)\left(Y_{t-j}-\mu\right)\right]=\begin{cases}\sigma^2 & si\ j=0\\ 0 & si\ j\neq 0\end{cases}$$

En cambio, el proceso $Y_t{=}\beta t{+}\varepsilon_t$ no es estacionario, ya que su media $\beta t$ es dependiente del tiempo $t$.

Note que si un proceso es estacionario, la covarianza $Cov\left(Y_t,Y_{t-j}\right)$ sólo depende de $j$ que significa la "distancia temporal" entre las observaciones, y no de $t$ que es el tiempo de la observación. De esto se deduce que para un proceso estacionario, las covarianzas $\gamma_{-j}$ y $\gamma_j$ representan el mismo valor ya que no hay dependencia del tiempo $t$.

$$\begin{aligned}\gamma_j{=}&\ E\left[\left(Y_t{-}\mu\right)\left(Y_{t-j}{-}\mu\right)\right]\\ \gamma_j{=}&\ E\left[\left(Y_{t+j}{-}\mu\right)\left(Y_{[t+j]-j}{-}\mu\right)\right]\\ \gamma_j{=}&\ E\left[\left(Y_t{-}\mu\right)\left(Y_{t+j}{-}\mu\right)\right]\\ \gamma_j{=}&\ \gamma_{-j}\ ,\quad \forall\ j\ \in Z\end{aligned}\tag{2.6}$$

### 2.1.2. Ruido Blanco

Es el bloque más útil en los procesos ARMA(AutoRegressive Moving Average), es decir la secuencia $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ en la cual todos los elementos tienen media 0 y varianza $\sigma^2$, es decir

$$E\left[\varepsilon_t\right]=0 \tag{2.7}$$

$$E\left[\left(\varepsilon_t\right)^2\right]{=}\sigma^2 \tag{2.8}$$

además, los valores $\varepsilon_t$ no poseen correlación en el tiempo, esto quiere decir que:

$$E\left[\varepsilon_t,\varepsilon_\tau\right]=0,\forall t\neq\tau \tag{2.9}$$

El proceso que satisface estas condiciones se denomina un *proceso de ruido blanco.* Muchas veces, la condición (2.9) se cambia por una que es un poco más fuerte, que afirma que los valores $\varepsilon_t$ son independientes en el tiempo, es decir que:

$$\varepsilon_t, \varepsilon_\tau \quad \textit{son independientes para } t\tau \tag{2.10}$$

2.10 implica que 2.9 se cumpla, pero no lo contrario. Un proceso que satisface 2.10 se denomina *proceso de ruido blanco independiente.*

## 2.2. Modelos Lineales

### 2.2.1. Procesos de Medias Móviles (MA)

**Proceso de Medias Móviles de Primer orden**

Sea $\{\varepsilon_t\}$ un proceso de ruido blanco y considérese el siguiente proceso

$$Y_t{=}\mu{+}\varepsilon_t{+}\theta\varepsilon_{t-1} \tag{2.11}$$

donde $\mu$ y $\theta$ son constantes. Este proceso es conocido como proceso de medias móviles de 1er orden., MA(1). Este nombre se da porque $Y_t$ se construye a partir de una suma ponderada, similar al cálculo de la media aritmética de los dos más recientes valores de $\varepsilon$.

El valor esperado de $Y_t$ es:

$$\begin{aligned} E[Y_t] &= E[\mu + \varepsilon_t + \theta\varepsilon_{t-1}] \\ &= \mu + E[\varepsilon_t] + \theta E[\varepsilon_{t-1}] \\ &= \mu \end{aligned} \tag{2.12}$$

La varianza de $Y_t$ es:

$$\begin{aligned} E[Y_t-\mu]^2 &= E\left[\varepsilon_t+\theta\varepsilon_{t-1}\right]^2 \\ &= E\left[\varepsilon_t^2+2\theta\varepsilon_t\varepsilon_{t-1}+\theta^2\varepsilon_{t-1}^2\right] \\ &= \left(1+\theta^2\right)\sigma^2 \end{aligned} \tag{2.13}$$

La primera autocovarianza:

$$\begin{aligned} E\left(Y_t-\mu\right)\left(Y_{t-1}-\mu\right) &= E\left(\varepsilon_t+\theta\varepsilon_{t-1}\right)\left(\varepsilon_{t-1}+\theta\varepsilon_{t-2}\right) \\ &= E\left(\varepsilon_t\varepsilon_{t-1}+\theta\varepsilon_{t-1}^2+\theta\varepsilon_t\varepsilon_{t-2}+\theta^2\varepsilon_{t-1}\varepsilon_{t-2}\right) \\ &= \theta\sigma^2 \end{aligned} \tag{2.14}$$

Las autocovarianzas de mayor orden son todas = 0

Si la media y covarianzas no dependen del tiempo, un proceso MA(1) es estacionario en la covarianza sin importar el valor de $\theta$, así, se satisface que

$$\sum_{j=0}^{\infty}|\gamma_j| = \left(1+\theta^2\right)\sigma^2 + \left|\theta\sigma^2\right| \tag{2.15}$$

Si el proceso $\{\varepsilon_t\}$ es *ruido blanco gaussiano*, entonces el proceso MA(1) es ergódico [1] en todos sus momentos.

La autocorrelación $\rho_j$ se define como la j-esima autocovarianza, dividida entre la varianza.

$$\rho_j \equiv \gamma_j/\gamma_0 \tag{2.16}$$

es decir, que es la correlación entre $Y_t$ y $Y_{t-j}$

$$Corr\left(Y_t, Y_{t-j}\right) = \frac{Cov\left(Y_t, Y_{t-j}\right)}{\sqrt{Var\left(Y_t\right)}\sqrt{Var\left(Y_{t-j}\right)}} = \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \rho_j \tag{2.17}$$

[1] Se aplica a una función aleatoria cuyos valores medios temporales son idénticos a los valores medios estadísticos correspondientes.

Usando las Ecuaciones (26, 27) la primera autocorrelación $\rho_1$ es dada por

$$\rho_1 = \frac{\theta\sigma^2}{(1+\theta^2)\,\sigma^2} = \frac{\theta}{(1+\theta^2)} \tag{2.18}$$

Las correlaciones superiores son todas igual a cero, $\rho_j = 0, \forall j > 1$

### 2.2.2. Procesos Autorregresivos (AR)

**Proceso Autorregresivo de 1er orden**

Un proceso autorregresivo de orden 1, denotado como AR(1), satisface la siguiente Ecuación:

$$Y_t = c + \varphi Y_{t-1} + \varepsilon_t \tag{2.19}$$

donde $\{\varepsilon_t\}$ también es un proceso de ruido blanco tal como se vio en la Sección (2.1). La Ecuación 2.17 tiene la forma de una ecuación diferencial de 1er orden en la que la variable de entrada es un ruido blanco más una constante. En este modelo de la Ecuación 2.19 se debe cumplir que $|\varphi| < 1$ para garantizar la estacionaridad en la covarianza, lo que es dado por la siguiente solución

$$\begin{aligned} Y_t &= (c+\varepsilon_t) + \varphi\,(c+\varepsilon_{t-1}) + \varphi^2\,(c+\varepsilon_{t-2}) \;+\; \cdots \\ &= [c/(1-\varphi)] + \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-1} + \cdots \end{aligned} \tag{2.20}$$

Que puede verse como un proceso $MA\,(\infty)$ donde cada $\psi_j = \varphi^j$ en el cual se satisface la condición $|\varphi| < 1$ lo que hace que se cumpla:

$$\sum_{j=0}^{\infty} |\psi_j| = \sum_{j=0}^{\infty} |\varphi|^j \tag{2.21}$$

Al asumir que $|\varphi| < 1$ se garantiza que el proceso $MA\,(\infty)$ existe y que puede manipularse y además que el proceso $AR\,(1)$ es ergódico en la media.

Al tomar el valor esperado en 2.20 se observa:

$$E[Y_t] = [c/(1-\varphi)] \; + 0 \; + \; 0 + \; \cdots \tag{2.22}$$

Por lo tanto, la media de un proceso estacionario $AR(1)$ es

$$\mu = c/(1-\varphi) \tag{2.23}$$

La varianza de un proceso $AR(1)$ está dada por:

$$\begin{aligned} \gamma_0 &= E[Y_t - \mu]^2 \\ &= E[\varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \varphi^3\varepsilon_{t-3} + \cdots]^2 \\ &= (1 + \varphi + \varphi^2 + \varphi^3 + \cdots)\,\sigma^2 \\ &= \sigma^2/(1-\varphi^2) \end{aligned} \tag{2.24}$$

y la $j$-ésima autoconvarianza está definida como:

$$\begin{aligned} \gamma_j &= E[Y_t - \mu][Y_{t-j} - \mu] \\ &= E\,[\varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots + \varphi^j\varepsilon_{t-j} + \varphi^{j+1}\varepsilon_{t-j-1} + \cdots] \times \\ &\quad [\varepsilon_{t-j} + \varphi\varepsilon_{t-j-1} + \varphi^2\varepsilon_{t-j-2} + \cdots] \\ &= [\varphi^j + \varphi^{j+2} + \varphi^{j+4} + \cdots]\,\sigma^2 \\ &= [\varphi^j/(1-\varphi^2)]\,\sigma^2 \end{aligned} \tag{2.25}$$

Entonces, la función de autocorrelación es:

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \varphi^j \tag{2.26}$$

que considerando que $|\varphi| < 1$ la autocorrelación se comporta como un decaimiento exponencial a medida que aumenta la distancia temporal $j$.

Un proceso autoregresivo de orden 2 $AR\,(2)$ tendrá la siguiente Ecuación:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varepsilon_t \tag{2.27}$$

Y un proceso autorregresivo de orden p, $AR\,(p)$, satisface la siguiente Ecuación:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \tag{2.28}$$

Y se prueba que las raíces del polinomio:

$$1 + \varphi_1 z + \varphi_2 z^2 + \cdots + \varphi_p z^p = 0 \tag{2.29}$$

Están ubicadas fuera del círculo unitario.
La Ecuación 2.28 se puede reescribir como

$$Y_t - \mu = c + \varphi_1 (Y_{t-1} - \mu) + \varphi_2 (Y_{t-2} - \mu) + \cdots + \varphi_p (Y_{t-p} - \mu) + \varepsilon_t \tag{2.30}$$

Las autocovarianzas se encuentran multiplicando 2.30 por $(Y_{t-j} - \mu)$ y calculando los valores esperados:

$$\gamma_j = \begin{cases} \varphi_1 \gamma_{j-1} + \varphi_2 \gamma_{j-2} + \cdots + \varphi_p \gamma_{j-p} & para\ j = 1, 2, \cdots \\ \varphi_1 \gamma_1 + \varphi_2 \gamma_2 + \cdots + \varphi_p \gamma_p + \sigma^2 & para\ j = 0 \end{cases} \tag{2.31}$$

Si aplicamos la identidad $\gamma_{-j} = \gamma_j$ en el sistema de Ecuaciones de 2.31 se puede solucionar para encontrar $\gamma_0, \gamma_1, \ldots, \gamma_p$ en función de $\sigma^2, \varphi_1, \varphi_2, \ldots, \varphi_p$. Se demuestra que el vector de tamaño $(p \times 1)$ $(\gamma_0, \gamma_1, \ldots, \gamma_p)'$ es formado por los $p$ primeros elementos de la primera columna de la matriz de tamaño $(p^2 \times p^2)$ $\sigma^2 [I_{p^2} - (F \otimes F)]^{-1}$, donde $F$ es una matriz $(p \times p)$ y $\otimes$ denota al operador *producto de Kronecker.*
Si se divide la Ecuación 2.31 entre $\gamma_0$, se obtienen las Ecuaciones de Yule-Walker.

$$\rho_j = \varphi_1\rho_{j-1} + \varphi_2\rho_{j-2} + \cdots + \varphi_p\rho_{j-p} \quad para \quad j = 1, 2, \ldots \tag{2.32}$$

Así, las autocovarianzas y autocorrelaciones siguen el mismo orden de las Ecuaciones de diferencia como el propio proceso 2.28. Para distintas raíces, sus soluciones tienen la siguiente forma:

$$\gamma_j = g_1\lambda_1^j + g_2\lambda_2^j + \cdots + g_p\lambda_p^j, \tag{2.33}$$

Donde los autovalores $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ son las soluciones de:

$$\lambda^p = \varphi_1\lambda^{p-1} - \varphi_2\lambda^{p-2} - \cdots - \varphi_p = 0 \tag{2.34}$$

### 2.2.3. Procesos Autorregresivos con Medias Móviles (ARMA)

Más conocidos como **procesos ARMA** (*Auto-Regresive Moving Average*), que como su nombre indica, incluyen tanto procesos autorregresivos de orden $p$ como procesos de medias móviles de orden $q$, conformando el modelo $ARMA\,(p, q)$

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \tag{2.35}$$

La estacionaridad de un proceso ARMA depende totalmente de los parámetros autorregresivos $(\varphi_1, \varphi_2, \ldots, \varphi_p)$ y no depende de los parámetros $(\theta_1, \theta_2, \ldots, \theta_q)$ del proceso de medias móviles.

Para analizar el modelo $ARMA\,(p, q)$, conviene escribirlo como desviaciones de la media $\mu$:

$$Y_t - \mu = c + \varphi_1\,(Y_{t-1} - \mu) + \cdots + \varphi_p(Y_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} \tag{2.36}$$

Las autocovarianzas se obtienen multiplicando 2.36 por el término $(Y_{t-j} - \mu)$ y luego

calculando el valor esperado. Para valores $j > q$ se tienen resultados de la forma

$$\gamma_j = \varphi_1\gamma_{j-1} + \varphi_2\gamma_{j-2} + \cdots + \varphi_p\gamma_{j-p} \tag{2.37}$$

Para valores $j = q+1,\ q+2, \ldots$

## 2.3. Series Temporales

Una secuencia de datos, observaciones o valores, vinculados a una variable temporal, ordenados cronológicamente y espaciados de manera uniforme, se llama Serie Temporal. Ejemplos se presentan en las observaciones de variables climatológicas, fenómenos físicos, financieros, biológicos, por un determinado tiempo; es de resaltar que para un segmento del tiempo $t_1$ se tiene una curva que representa una realización. Si, bajo las mismas condiciones, se realizan mediciones en otro segmento de $t_2$, se obtiene otra curva que por lo general no es igual a la primera. Cada conjunto de medidas define una trayectoria o realización del proceso que esta siendo observado. Asumiendo algunas condiciones, como la ergodicidad, a partir de una realización (la serie histórica única que se tiene disponible en la práctica) es posible modelar este proceso físico usando un proceso estocástico. Con este modelo se abre la posibilidad de generar un conjunto de trayectorias que son posibles de ser observadas. En este contexto, cada una de estas trayectorias se denomina tambien una serie temporal.

Un proceso estocástico es descrito por el conjunto de todas las series temporales (o realizaciones) que lo componen, que son infinitas por lo general como la Ecuación 2.3, o también por la distribución de probabilidades conjunta de todas las variables aleatorias que están en juego. En la realidad no se tiene ninguna de estas formas, queda el modelo de series sintéticas que busca ajustar un modelo, que se cree es el que generó, a la serie histórica y a partir de éste, generar series sintéticas que representan las series temporales que podrían ser "muestreadas" del proceso que se está analizando

como un proceso estocástico (Cadavid y Salazar, 2008; Wilkinson, 2009; Thomas y Fiering, 1962; Jaeger, 2000).

### 2.3.1. Series Temporales Estacionales

Muchos procesos físicos (que ocurren aquí en nuestro planeta Tierra) presentan escala diaria o mensual con comportamiento periódico descrito por ciclos estacionales. Cada periodo presenta un conjunto de características estadísticas particulares que se describen usando la media $\mu_t$, la varianza $\gamma_{0t}$ y la estructura de correlaciones tomando especial interés en las correlaciones estacionales.

#### Media y varianza muestreadas en un periodo

La media muestreada de un periodo $m$ de 12 meses se da por:

$$\mu_m = \frac{1}{N} \sum_{i=1}^{N} z_{i-1}12 + m \tag{2.38}$$

Donde $m = 1, \ldots, n$ y $n$ es el ultimo periodo
Análogamente, la desviación estándar de cada mes es:

$$\widehat{\sigma}_m = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( z_{(i-1)12+m} - \widehat{\mu}_m \right)} \tag{2.39}$$

donde $m = 1, \ldots, 12$

### 2.3.2. Coeficiente de Correlación

Es el valor que determina el grado de relación que existe entre 2 o más variables. Los valores que puede tomar el coeficiente de correlación $r$ son: $-1 < r < 1$.

El signo indica la dirección de la correlación, positiva o directamente proporcional (a mayor A mayor B o a menor B menor A) y negativa o inversamente proporcional (a menor A mayor B o viceversa).

El valor te indica la *fuerza de la correlación*. Una correlación perfecta tendría un valor cercano al 1 o -1, mientras que una ausencia de correlación tendría un valor cercano al 0.

Entre los coeficientes de correlación más conocidos podemos encontrar:

- Coeficiente de Correlación Lineal

  Mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos se aproximaría a una recta).

  El coeficiente de correlación lineal se calcula aplicando la siguiente fórmula:

$$r = \frac{\frac{1}{n} * \sum((X_i - X_m) * (Y_i - Y_m))}{\sqrt{(\frac{1}{n} * \sum(X_i - X_m)^2) * (\frac{1}{n} * \sum(Y_i - Y_m)^2)}} \tag{2.40}$$

  Donde el numerador se denomina covarianza y se calcula de la siguiente manera: en cada par de valores $(x, y)$ se multiplica el valor de $x$ menos su media, multiplicado por el valor de $y$ menos su media. Se suma el resultado de todos los pares de valores y este resultado se divide por el tamaño de la muestra.

  El denominador se calcula como la raíz cuadrada del producto de las varianzas de $x$ y de $y$.

- Coeficiente de Correlación de Pearson

Arroja un producto conocido como $r$ de Pearson cuando se habla de muestras y como $\rho$ de Pearson cuando hablamos de poblaciones. Esta dado por la siguiente fórmula.

$$r = \frac{N * \sum(X * Y) - (\sum X)(\sum Y)}{\sqrt{(N * \sum X^2 - (\sum X)^2) * (N * \sum Y^2 - (\sum Y)^2)}} \tag{2.41}$$

donde:

$N$ es el número de sujetos a correlación.

$\sum$ **X y** $\sum$ **Y** es la suma de los datos de $X$ y de $Y$ respectivamente.

$\sum X^2$ **y** $\sum Y^2$ es la suma de los datos elevados al cuadrado de $X$ y $Y$ respectivamente.

**Estructura de correlaciones mensuales**

En procesos mensuales se puede definir valores que describen la estructura de correlación lineal de un mes con los meses anteriores, que puede ser de orden 1, que describe la dependencia de un mes con el inmediato anterior, o una correlación de orden 2 que describe la dependencia de los meses $m$ con respecto a los meses $m-2$, o generalizando, una correlación de orden $k$ que representa la dependencia del mes $k$ con respecto al mes $m-k$.

$$\widehat{\gamma}^{m_{(k)}} = \frac{1}{N} \sum_{i=1}^{N} \left(z_{(i-1)12+m} - \widehat{\mu}_m\right) \left(z_{(i-1)12+m-k} - \widehat{\mu}_m\right) \tag{2.42}$$

$$\widehat{\rho}^{m_{(k)}} = \frac{\gamma^{m_{(k)}}}{\widehat{\sigma}_m \widehat{\sigma}_{m-k}} \tag{2.43}$$

donde $m = 1, \ldots, 12$.

## 2.4. Razonamiento Basado en Casos

En esta sección se describe los fundamentos sobre el Razonamiento Basado en Casos (RBC), los cuales serán aplicados en la propuesta del nuevo modelo de Procesos Estocástico para la generación de series temporales. La presente se inicia con la Sección 2.4.1, Definición, donde se comenta los conceptos asociados al RBC, la definición de un caso, el método de aprendizaje, sus etapas, algunos ejemplos típicos y el contexto del RBC, luego se explica el ciclo de vida los cuales se extienden y detallan en las secciones 2.4.3 Representación e Indexación de casos, 2.4.4 Recuperación de casos, 2.4.5 Reutilización o adaptación de casos y 2.4.6 Retención y Mantenimiento de Casos; finalmente se presentan comparaciones entre el RBC con Sistemas Basados en Conocimiento, Reglas, el razonamiento humano, finalmente las ventajas y desventajas y algunos lineamientos para el uso correcto de esta técnica.

### 2.4.1. Definición

El Razonamiento Basado en Casos (RBC) es un cuerpo de conceptos y técnicas que tocan temas relacionados a la representación del conocimiento, razonamiento y aprendizaje a partir de la experiencia (Zadeh, 2003); está basado en *Soft Computing* [2]. Surge a partir de las ciencias cognitivas(Schank, Abelson, y cols., 1977; Schank, 1982). los primeros prototipos fueron: Cyrus (Kolodner, 1983a, 1983b), Mediator (Simpson, 1985), Persuader (Sycara, 1988), Chef (Hammond, 1989), Julia (Hinrichs, 1992) Casey, y protos (Bareiss, 1989).

La similitud es el concepto que juega un papel fundamental en RBC; esta se puede definir como una relación donde el numerador es el número de atributos que dos objetos tienen en común y donde el denominador es el número total de atributos, tal como se

[2]Colección de metodologías que proveen las bases para la concepción, diseño y utilización de sistemas inteligentes. (Lógica Difusa, Redes Neuronales, Computación Evolutiva, Computación Probabilística, Computación Caótica, Teoría de conjuntos aproximados, mapas auto-organizativos, aprendizaje máquina y minería de datos, (Zadeh, 2003).

ve en la ecuación 2.44 (Tversky, 1977).

$$similitud_{e_p,e_q} = \frac{\alpha(A)}{\alpha\,(A) + \beta(B)} \tag{2.44}$$

donde $A$ representa los atributos comunes, $B$ los atributos diferentes, $\alpha$ y $\beta$ los pesos determinados por un algoritmo de aprendizaje, un experto o la fuerza de la relación $e_p, e_q$ representan casos, vea la Sección (2.45).

Existen otras definiciones de similitud para casos multivalentes, y atributos ponderados (Pal y Shiu, 2004) que será analizados en la Sección 2.4.4, página 37.

Ademas, el Razonamiento Basado en Casos (RBC) o *Case Based Reasoning (CBR)*; en este contexto se define como un modelo de razonamiento que integra resolución de problemas, entendimiento y aprendizaje con procesos de memoria; estas tareas se realizan en base a situaciones típicas, llamadas *casos* (Pal y Shiu, 2004).

**Definición de un caso**

También conocido como instancia objeto o ejemplo. Puede ser definido como una pieza de conocimiento contextualizado que representa una experiencia significativa. Enseña una lección fundamental para el logro de un objetivo en un sistema(Pal y Shiu, 2004). Se puede representar un caso como:

$$e_{(i)} = \{a_{(i,1)}, a_{(i,2)}, ...a_{(i,n)}\} \tag{2.45}$$

donde $e_{(i)}$ es el $i$ caso indexado, con un esquema $e$, $a_{(i,1)}, a_{(i,2)}, ...a_{(i,n)}$ son instancias de $n$ atributos $a$ relacionados para el $i$ caso.

Correspondientemente la Base de Casos se define:

$$BC = \{e_1, e_2, ...e_m\} \tag{2.46}$$

donde $BC$ es la librería de $m$ casos.

**Aprendizaje** Como un subproducto de la actividad de razonamiento, el sistema aprende, evoluciona, mejora la competencia y eficiencia de los resultados como producto de almacenar la experiencia pasada y recupera los casos pasados en el razonamiento futuro (Pal y Shiu, 2004).

### Funcionamiento

El mecanismo básico de funcionamiento del RBC es la búsqueda por similitud. Para un caso problema, el motor busca en su memoria de casos anteriores (llamado Base de Casos) un caso que tiene el mismo problema que las especificaciones del caso bajo análisis, vea la Figura 2.1. Si el razonador no puede encontrar un caso idéntico en su base de casos, intentará encontrar un caso o casos que se acerquen más al caso problema. En situaciones en que un caso idéntico anterior se recupera, y bajo el supuesto de que su solución se ha realizado correctamente, se puede ofrecer como solución al problema actual. En la situación más probable que el caso recuperado no sea idéntico al caso actual, una fase de adaptación se produce. Durante la adaptación, las diferencias entre el caso actual y los casos recuperados se identifican y luego la solución asociada con el caso recuperado se modifica, teniendo en cuenta estas diferencias. La solución devuelta, en respuesta a la especificación del problema actual, puede ser juzgada en la configuración de dominio correspondiente (Pal y Shiu, 2004).

**Componentes** Los componentes de un sistema RBC suelen ser concebidos de manera que reflejen las cuatro etapas típicas separadamente (recuperación, reutilización, revisión y retención); véase la Figura 2.3. Sin embargo, tal como se ve en la Figura 2.1; a un nivel de abstracción mas alto, el RBC puede ser visto como un mecanismo de razonamiento, y sus tres componentes externos:

- El mecanismo de razonamiento.
- Condiciones de entrada o problema caso.

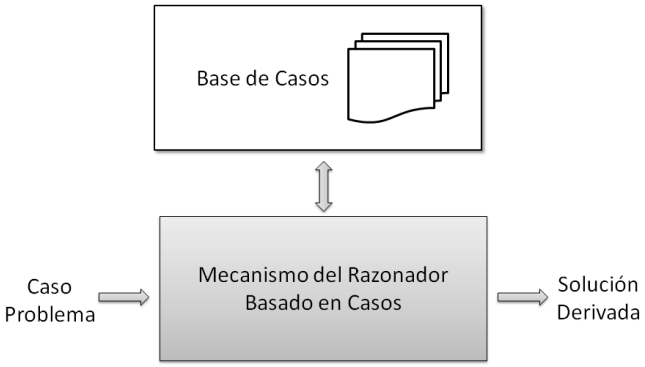



Figura 2.1: Esquema de un Sistema RBC
(Pal y Shiu, 2004).

- Salida que define una propuesta de solución al problema.
- La memoria de los casos anteriores.

En la mayoría de los sistemas RBC, el mecanismo de razonamiento se basa en casos, de forma alternativa es conocido como el solucionador de problemas o razonador. Su estructura interna, en un nivel abstracto, está dividida en dos partes principales: El recuperador de casos y el razonador (véase la Figura 2.2). La tarea del recuperador de casos es buscar el caso apropiado en la Base de Casos, mientras que el razonador utiliza los casos recuperados para encontrar una solución a un problema determinado. Este proceso de razonamiento en general, implica tanto la determinación de las diferencias entre los casos recuperados y el caso actual, y la modificación de la solución. El proceso de razonamiento puede, o no, implicar la recuperación de casos adicionales o partes de los casos de la base de casos.

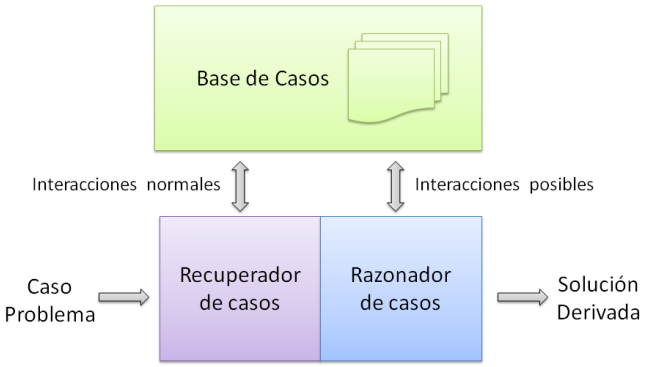



Figura 2.2: Componentes Internos del RBC
(Pal y Shiu, 2004).

### 2.4.2. Ciclo de vida del Razonamiento Basado en Casos

El ciclo de vida para la solución de problemas usando un sistema RBC consta de cuatro estados.

- Recuperación de casos similares de una base de experiencia.
- Reutilización de casos mediante copia o integración de soluciones desde los casos recuperados.
- Revisión o Adaptación de la solución(es) recuperada(s) para resolver el nuevo problema
- Retención de una nueva solución, una vez haya sido confirmada o validada.

En muchas aplicaciones prácticas, los estados de Reutilización y Revisión son difíciles de distinguir, y varios investigadores usan solo un estado de adaptación que reemplaza y combina ambos. Sin embargo la adaptación en los sistemas RBC es una pregunta aún abierta porque es un proceso complicado que intenta manipular los casos solución.

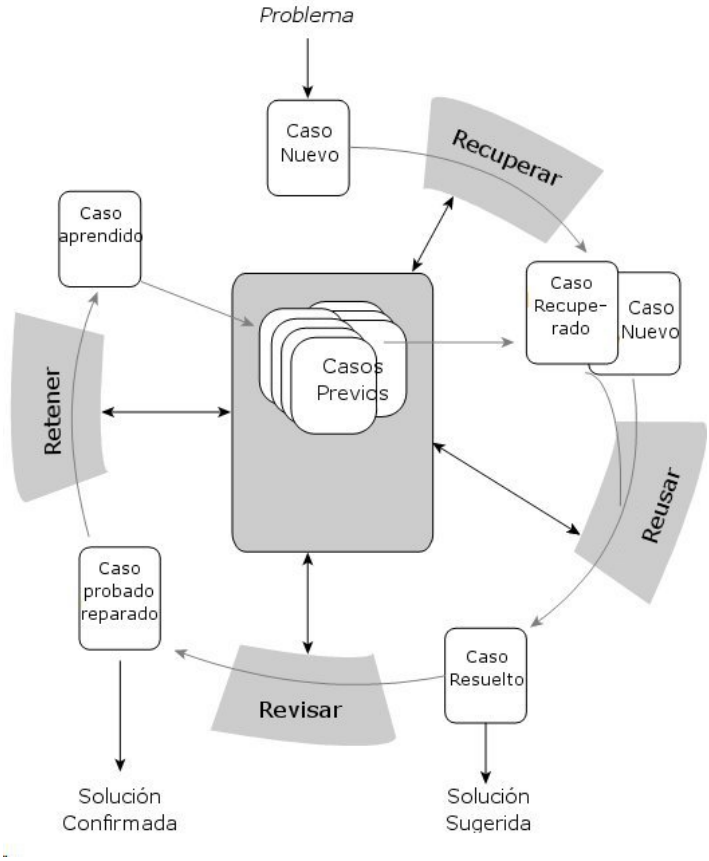



Figura 2.3: Ciclo de vida de RBC
(Pal y Shiu, 2004).

Generalmente, estos requieren el desarrollo de un modelo causal entre el espacio del problema y el espacio de la solución de los casos relacionados.

Como se aprecia en la Figura 2.3, los casos almacenados en la librería de casos, fueron complementados con el conocimiento general, que usualmente son dependientes del dominio. El soporte puede ser desde muy débil hasta muy fuerte, dependiendo del tipo de método RBC. Por ejemplo, en un sistema de diagnóstico un modelo causal de patología y anatomía pueden constituir el conocimiento general. Este conocimiento puede estar representado en la forma de un conjunto de reglas IF-THEN o algunas pre-condiciones. Cada estado en el ciclo de vida del RBC está asociado con algunas tareas de la Figura 2.4.

#### Vista orientada a tareas

Una visión orientada a tareas es buena para la descripción de los mecanismos internos del RBC, a comparación de la vista orientada a procesos o etapas del ciclo de vida del RBC que solo proporciona una visión global y externa de lo que esta pasando. Las tareas se establecen en función de los objetivos del sistema, y una tarea en particular se lleva a cabo mediante la aplicación de uno o más métodos (vea la Figura 2.4).

### 2.4.3. Representación e Indexación de casos

Para resolver algún problema en un sistema RBC los detalles usualmente están incluidos en la especificación del problema.

La base de casos en un sistema RBC es la memoria de todos los casos almacenados previamente, hay tres temas generales que se debe tener en cuenta a la hora de crear una base de casos:

- La estructura y representación de los casos.
- El modelo de memoria usado para organizar la base de casos completo.

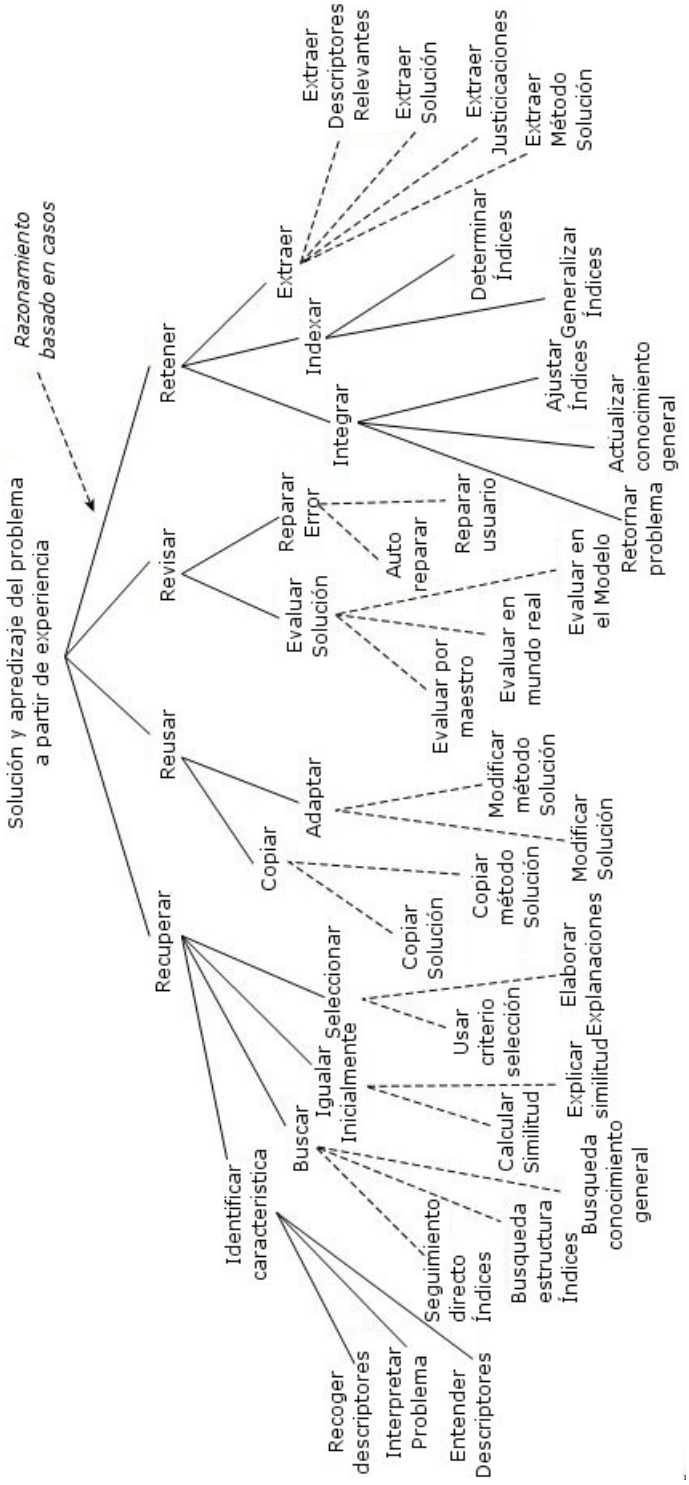



Figura 2.4: Descomposición de métodos y tareas del RBC (Pal y Shiu, 2004).

- La selección de los índices usados para identificar cada caso.

### Representación de Casos y almacenamiento

Los casos almacenados en una base de casos pueden representar una gran variedad de conocimiento que se pueden almacenar de distintas maneras. En cada tipo de sistema RBC, un caso puede representar a una persona, objeto, situación, diagnóstico, diseño, plano, y todas las entidades imaginables.

### Factores para la representación de un caso

Hay una serie de factores que deben considerarse para elegir un formato de representación de un caso.

- El formato elegido: Debe ser capaz de representar varias formas adoptadas para una estructura interna.
- Tipos y estructuras asociados con el contenido o las características que describen un caso: Estos tipos tienen que estar disponibles, o ser susceptibles de ser creados.
- El idioma o Shell elegido para implementar el sistema RBC: La elección de una Shell puede limitar los formatos que se pueden utilizar para la representación.
- El mecanismo de indexación y búsqueda planificada: Los casos tienen que estar en un formato que el mecanismo de recuperación de casos pueda tratar con eficacia.
- La forma en que los casos están disponibles: Por ejemplo, si una base de caso se forma a partir de una colección existente de las experiencias pasadas, la facilidad con que estas experiencias se pueden traducir a una forma apropiada para el sistema CBR puede ser importante.

### Modelo de memoria para representación de un caso

Independientemente del formato elegido para representar los casos, la colección de casos también tiene que estar estructurado de una manera que facilite su recuperación cuando se requiera. Una base de casos plana o *Flat Memory* es una estructura común. En este método los índices son elegidos para representar los aspectos importantes del caso, y la recuperación implica la comparación de las características, consultando cada uno con la base de casos, otra forma es agruparlos por categorías para reducir el número de casos que tienen que ser buscados durante la consulta. El modelo de memoria para la elección de una representación de casos dependerá de una serie de factores.

- La representación usada en la base de casos.
- El propósito del sistema RBC. Por ejemplo una estructura jerárquica es una elección natural para un sistema de resolución de problemas de clasificación.
- El número y la complejidad de los casos que van a ser almacenados. A medida que el número de casos crece en una base de casos, una estructura que busca secuencialmente consume más tiempo durante la recuperación.
- El número de características que se utilizan para la búsqueda de casos coincidentes.
- Si algunos casos son bastante similares estos deben agruparse.
- Cuanto se conoce sobre un dominio específico, esto influye en la capacidad de determinar si los casos son similares.

### Indexación de casos

La indexación de casos se refiere a la asignación de índices a los casos para futuras recuperaciones y comparaciones. La elección de los índices es importante para obtener

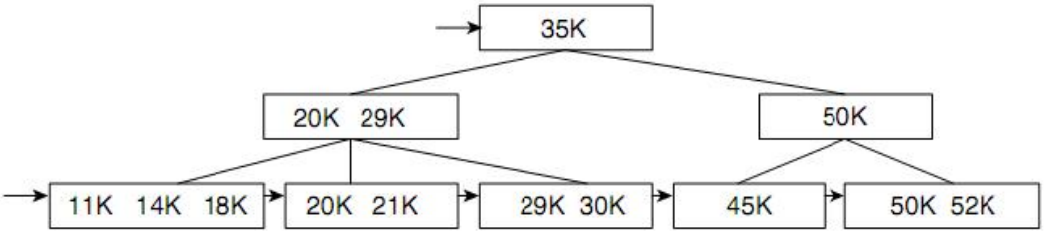



Figura 2.5: Ejemplo de $B^+$ para indexación de números (Pal y Shiu, 2004).

los casos similares en un tiempo rápido. Los índices deberán ser predictivos de una manera útil. Esto significa que los índices deben reflejar las características importantes de un caso y los atributos que influyen en el resultado de un caso, así como describir las circunstancias en las que se encuentra para ser recuperados en un futuro.

**Método de indexación tradicional** En los enfoques de base de datos relacionales tradicionales, índice se refiere a la clave primaria y secundaria de un registro, Indexación se refiere a la tarea de asignación de la clave a un registro para la ubicación de su almacenamiento. Esto se debería de hacer mediante el uso de métodos de acceso directo como son los *hash*; métodos indexados, como son la construcción de un $B^+tree$ o un $Rtree$ para la organización de los registros o metodos de acceso métrico como $Ommi-tree$ o secuenciales. La búsqueda y recuperación de los registros es para determinar su ubicación, es realizado ya sea mediante la asignación del árbol de índices o el uso de algoritmos *hashing*.

**Indexación vía** $B-Trees$ Se tienen diferentes estructuras de datos para indexación, esto dependerá mucho del problema a resolver. Para los $B-trees$ la forma de asignación de los registros puede explicarse por la Figura 2.5.

Aquí, los nodos de la capa inferior del árbol son los nodos hoja y las dos capas de arriba son los nodos intermedios. Los nodos intermedios contienen el valor o valores

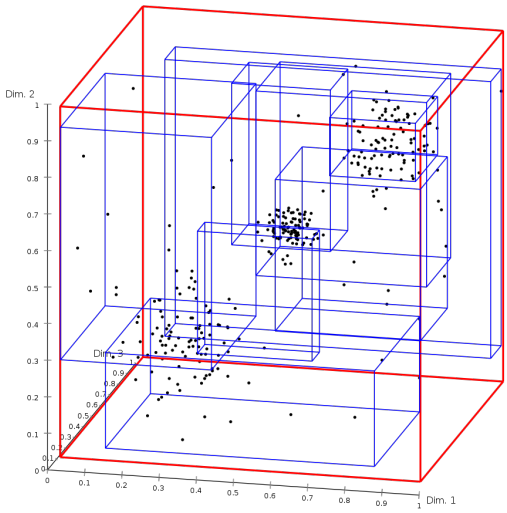



Figura 2.6: Indexación de datos en $R-tree$.
(Pal y Shiu, 2004)

de un intervalo de índice, y los nodos hoja contienen los punteros a los lugares de almacenamiento de los casos. Un nodo intermedio puede generar tres nodos secundarios.

El límite superior del intervalo en su nodo hijo izquierdo es mas pequeño que el límite inferior de su padre, y el límite inferior de la derecha del hijo es equivalente al mayor que el límite superior de su padre. El límite inferior del hijo medio es igual o mayor que el de su padre, y su límite superior es menor que el de su padre.

Otras estructuras de mejora de índices, como $R-tree$, $R^*-tree$, y $R-trees$ , soportan rangos y búsquedas multidimensionales de los registros, sin embargo, estos se basan en el concepto de concordancia exacta, es decir los objetos están dentro del rango o fuera de este, además la superposición de conceptos no está permitido, vea la Figura 2.6.

### 2.4.4. Recuperación de casos

La recuperación de casos es el proceso de encontrar, dentro de una base de casos, aquellos casos que son mas similares al caso actual. Para llevar a cabo la recuperación eficaz de los casos, hay criterios de selección que son necesarios para determinar cuál es el mejor de los casos para recuperar.

Los criterios de selección de los casos dependen en parte del caso que se va recuperar de la base de casos, a menudo se hace una búsqueda completa de las características de las cuales se comparan con el caso actual. Sin embargo, hay ocasiones en que solo una parte de un caso es la que se busca, esto puede deberse a que no existe un caso completo.

#### Técnicas de recuperación

La recuperación es un área de investigación importante en el RBC. Las técnicas de recuperación más investigadas, por el momento, son los k-vecinos más cercanos o *Nearest-neigbor retrieval (k-NN)*, árboles de decisión, y sus derivados. Estas técnicas implican el desarrollo de una métrica de similitud que le permite estar cerca entre los casos más parecidos.

- *K-vecinos más cercanos.* En la recuperación, el caso recuperado es elegido por la suma ponderada y la mínima distancia euclidiana de sus características, que coinciden con el caso actual. En términos sencillos, para todas las características el mismo peso, un caso que coincide o se parece con el caso actual.

- *Enfoque Inductivo.* Cuando los enfoques inductivos son utilizados para determinar la estructura del caso base, que determina la importancia de las características para discriminar entre los casos similares, la estructura jerárquica resultante de la base de casos ofrece un espacio de búsqueda reducido para recuperar un caso, el cual disminuye el tiempo de búsqueda.

- *Enfoque Conocimiento guiado.* Este enfoque es utilizado para determinar las características de un caso que son importantes para la recuperación de un caso futuro. En algunas situaciones las diferentes características de un caso tienen diferentes niveles de importancia o contribución a los niveles de éxito relacionados con ese caso.

- *Recuperación Validada.* Ha habido numerosos intentos de mejorar la recuperación. Uno de ellos es la recuperación validada propuesta por Simoudis (Simoudis, 1992), que consta de dos fases. La fase 1 consiste en la recuperación de todos los casos que parecen ser relevantes para un problema, sobre la base de las principales características del caso actual. La fase 2 implica derivar las características más exigentes del grupo inicial de casos recuperados, para determinar si estos casos son válidos en la situación actual.

**Factores para determinar el método de recuperación**

Los factores que se deben considerar para determinar el mejor método de recuperación son:

- El número de casos que se debe buscar.

- La cantidad de conocimiento del dominio disponible.

- La facilidad de determinar las ponderaciones de las características individuales

- Si todos los casos deben ser indexados por las mismas características o si cada caso tiene características que varían en importancia.

Una vez que un caso se ha recuperado, por lo general hay un análisis para determinar si este caso está lo suficientemente cerca al caso problema o si los parámetros de búsqueda deben ser modificado y llevar a cabo una nueva búsqueda. Si la opción

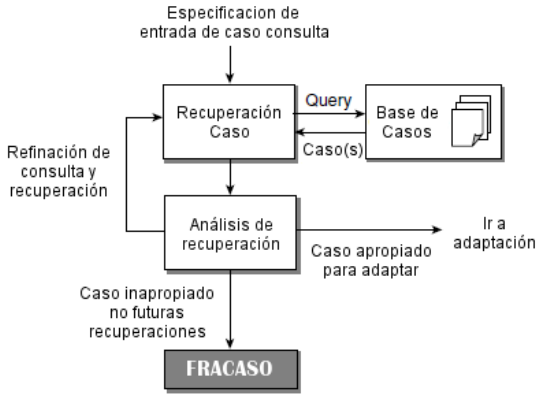



Figura 2.7: Procesos que involucra un RBC
(Pal y Shiu, 2004).

correcta se realiza durante este análisis, no puede haber un ahorro de tiempo significativo. Por ejemplo, el tiempo de adaptación necesario para un caso lejano podría ser significativamente mayor que buscar de nuevo.

Cuando consideramos un método de análisis para esta decisión, los siguientes puntos deben ser considerados:

- El tiempo y recursos requeridos para la adaptación.
- El número de casos en la base de casos.
- El tiempo y recursos requeridos para la búsqueda.
- Cuanto de la base de casos ya se ha buscado.

Si revisamos el proceso que involucra la recuperación en un RBC, se puede representar como se muestra en la Figura 2.7.

### Concepto de similitud

El significado de similitud depende en el contexto en el que se encuentra una aplicación en particular, y para cualquier contexto comparativo no expresa una característica fija.

En el RBC, calcular la similitud es un tema muy importante para el proceso de recuperación de los casos; la eficacia de una medida de similitud es determinada por la utilidad de un caso recuperado en resolver un nuevo problema. Se establece una función de similitud, apropiada al manejo de las relaciones escondidas y profundas entre los objetos más relevantes que están relacionados con los casos. Existen dos enfoques principales en la recuperación de casos:

- El primero está basado en el cálculo de la distancia, entre los casos en donde se determina el caso más similar por una medida (es decir métrica) de evaluación de similitud.
- El segundo enfoque está relacionado con las estructuras de representación/indexación de los casos, la cual la estructura de indexación puede recorrer en busca de un caso similar.

A continuación se describirá los conceptos básicos y características de algunas medidas de distancia que se utilizan en este sentido (Pal y Shiu, 2004).

**Distancia Euclidiana Ponderada**

Es el tipo mas común de medir una distancia y esta basado en la ubicación de los objetos en el espacio Euclidiano (es decir un conjunto ordenado de números reales). Formalmente los casos son expresados de la siguiente manera:

$$BC = (e_1, e_2, ...e_N) \tag{2.47}$$

donde $BC$ es una librería de casos y $e_N$ es el $N - esimo$ caso $e$.

Para la distancia Euclidiana se tiene que cada caso en esta librería está representado por un índice de su correspondiente característica, ademas cada caso esta asociado a una acción. Mas formalmente se usa una colección de características $\{F_j(j = 1, 2, ..., n)\}$

para indexar los casos y una variable $V$ que denota la acción. El i-ésimo caso $e_i$ en la librería puede ser representado por un vector $(n+1)-$dimensional que es, $e_i = (x_{i1}, x_{i2}, ..., x_{in}, \theta_i)$, donde $x_{ij}$ corresponde al valor de la característica $F_j (1 \leq j \leq n)$ y $\theta_i$ corresponde a los valores de la acción $V (i = 1, 2, ..., N)$.

Supongamos que para cada característica $\{F_j (j = 1, 2, ..., n)\}$, un peso $w_j (w_j \in [0, 1])$ ha sido asignado a la j-ésima característica para indicar la importancia de la característica. Entonces para un par de casos $e_p$ y $e_q$ en la librería, una distancia métrica ponderada puede ser definida como:

$$d_{pq}^{(w)} = d^{(w)}\left(e_p, e_q\right) = \left[\sum_{j=1}^{n} w_j^2 (x_{pj} - x_{qj})^2\right]^{1/2} = \left(\sum_{j=1}^{n} w_j^2 x_j^2\right)^{1/2} \tag{2.48}$$

donde $x_j^2 = (x_{pj} - x_{qj})^2$.

Cuando todos los pesos son iguales a 1 la distancia métrica ponderada definida anteriormente degenera a la medida Euclidiana $d_{pq}^1$ esto quiere decir que es denotada por $d_{pq}$ , usando la distancia ponderada una medida de similitud entre dos casos, $SM_{pq}^{(w)}$, puede ser definida como:

$$SM_{pq}^{(w)} = \frac{1}{1 + \alpha d_{pq}^{(w)}} \tag{2.49}$$

donde $\alpha$ es una constante.Cuanto más alto sea el valor de $d_{pq}^{(w)}$ , la similitud entre $e_p$ y $e_q$ será mas baja. Cuando todos los pesos toman valor de 1, la medida de similitud es denotada por $SM_{pq}^{(1)}$, $SM_{pq}^{(1)} \in [0, 1]$.

Las características del valor real mencionadas anteriormente, podrían extenderse sin dificultad a las características que tienen los valores en un espacio vectorial normalizado.

Por ejemplo: para cada característica una medida de distancia ha sido definida. La

medida de distancia para la j-ésima característica está denotada por $\rho_j$ ; que es , $\rho_j$ es un mapeo de $F_j$ x $F_j$ a $[0, \infty]$ (donde $F_j$ es denotado como el dominio de la j-esima característica) con las siguientes propiedades.

- $\rho_j(a, b) = 0$ si y solo si $a = b$.
- $\rho_j(a, b) = \rho_j(b, a)$.
- $\rho_j(a, b) \leq \rho_j(a, c) + \rho_j(c, b)$.

Para características numéricas y no numéricas, pueden ser usadas algunas fórmulas típicas para la medida de distancia; se muestran a continuación algunas:

- $\rho_j(a, b) =\mid a - b \mid$ si a y b son números reales.
- $\rho_j(A, B) = max_{a \in A, b \in B} \mid a - b \mid$ si A y B son intervalos.
- $\rho_j(a, b) = \begin{cases} 1 \; si \; a \neq b \\ 0 \; si \; a = b \end{cases}$ si a y b son símbolos.

En estas circunstancias, la distancia entre dos casos $e_p$ y $e_q$ pueden ser calculados por:

$$d_{pq}^{w} = \sqrt{\sum_{j=1}^{n} w_j^2 \rho_j^2(e_{pj}, e_{qj})} \tag{2.50}$$

**Medida de similitud de Tversky**

Mostraremos a continuación una medida de similitud usada comúnmente. Denotamos a $SM_{pq}$ como una medida de similitud entre dos casos; un nuevo caso consulta $e_p$ y un caso almacenado $e_q$. Una medida de similitud que está basada en el modelo de relación propuesto por Tversky (Tversky, 1977):

$$SM_{pq} = \frac{\alpha(comunes)}{\alpha\,(comunes) + \beta(diferentes)} \tag{2.51}$$

donde comunes y diferentes representan al número de atributos que son similares o diferentes, respectivamente entre el nuevo caso de consulta $e_p$ y el caso almacenado $e_q$. Por lo general, esta decisión implica considerar un valor umbral, para que las características se clasifiquen como similares si su similitud está por encima del umbral.

Los valores de $\alpha$ y $\beta$ son los pesos correspondientes, que pueden estar determinados por un experto o mediante el uso de técnicas de aprendizaje automático. Una medida de similitud, que se basa en el número de reglas de producción que se crea en una instancia, ha sido propuesto por Sebag y Schoenauer (Sebag y Schoenauer, 1994).

$$SM_{pq} = \sum_i w(r_i) \tag{2.52}$$

donde $(r_i)$ representa las reglas que son aprendidas desde el caso base y $w$ es el peso asignado. Una medida de similitud basado en el modelo de cambio propuesto por Weber (Weber, 1995).

$$SM_{pq} = \alpha f(e_p \cap e_q) - \beta f(e_p - e_q) - \gamma f(e_q - e_p) \tag{2.53}$$

La intersección $(e_p \cap e_q)$ describe aquellos atributos que son comunes a $e_p$ y $e_q$, y el conjunto de complementos $(e_p - e_q)$ y $(e_q - e_p)$ describe aquellos atributos que son observados solo en el caso consulta (mas no en el caso almacenado) y solo en el caso almacenado (mas no en el caso consulta), respectivamente. $f$ es denotado a algún operador o algoritmo para calcular su correspondiente calificación de la relación de conjuntos. $\alpha, \beta, \gamma$ son los pesos correspondientes.

Varias métricas de similitud son propuestas. Estas tienen en cuenta diferentes características comparativas. tales como el número consecutivo de aportes, el grado de normalización entre los atributos, la "tipicidad" de los casos, la relevancia de ciertos atributos entre un caso de una nueva consulta y un caso almacenado, el grado de similitud en las relaciones entre atributos, similitud en la estructura, similitud basada en la

jerarquía de clases orientada a objetos y medidas de similitud difusas supervisadas y no supervisadas (Pal y Shiu, 2004).

### 2.4.5. Reutilización o adaptación de casos

La adaptación de casos es el proceso de transformar una solución recuperada en una solución apropiada para un problema actual. Se ha argumentado que la adaptación es el paso más importante de un RBC, ya que añade inteligencia a lo que sería el cálculo de un patrón simple.

#### Enfoques para la adaptación

Una serie de enfoques se pueden tomar para llevar a cabo la adaptación de los casos:

- Las soluciónes devueltas (casos recuperados) podrían ser utilizados como una solución al problema actual sin modificaciones, o con las modificaciones donde la solución no es del todo apropiada para la situación actual.
- Los pasos o procesos que se siguieron para obtener la solución anterior, podría retornar sin modificaciones o con modificaciones que no son plenamente satisfactorios en la situación actual.
- Cuando más de un caso ha sido recuperado, una solución podría ser derivada a partir de varios casos, o varias soluciones podrían ser presentadas.

La adaptación puede usar varias técnicas, incluyendo las reglas o una iteración adicional de razonamiento basado en casos, en un espacio de recuperación mas similar para cada caso.

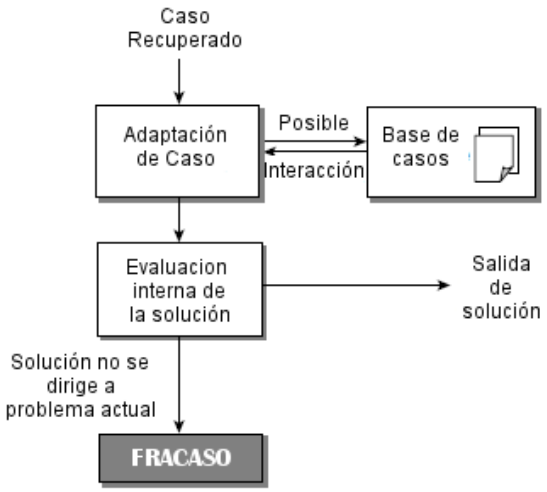



Figura 2.8: RBC dentro de un estado de aprendizaje
(Pal y Shiu, 2004).

**Consideraciones para escoger la estrategia de adaptación**

Cuando se elige una estrategia de adaptación de casos, puede ser útil considerar lo siguiente:

- En promedio, ¿cómo se cerrará el caso de ser recuperado?
- En general, ¿cómo muchas de las características difieren entre los casos?
- ¿Hay sentido común o reglas conocidas que se pueden utilizar en la realización de la adaptación?

Después que la adaptación se ha completado, es conveniente comprobar que la solución es adecuada y sí tiene en cuenta las diferencias entre el caso recuperado y el problema actual. En este punto, también hay una necesidad de considerar qué acción se debe tomar, si este control determina que la solución propuesta es poco probable que tenga éxito.

En esta etapa, la salida solución desarrollada está lista para las prueba en el mundo real de una aplicación, véase las Figuras 2.8 y 2.10, luego, muchos sistemas entran en

una fase de aprendizaje, tal como se explica en la siguiente sección.

### 2.4.6. Retención y Mantenimiento de Casos

En esencia, el mantenimiento de la base de casos es visto como un proceso de refinación del sistema RBC para mejorar el desempeño de los resultados (Craw, Jarmulak, y Rowe, 2001). Los resultados a obtener son definidos por el usuario de acuerdo al dominio del problema y el ambiente externo. Suelen haber dos tareas típicas en el mantenimiento: cuantitativas y cualitativas. las cualitativas se aseguran de la consistencia y las cuantitativas de la eficiencia, existen muchas técnicas para ambas tareas (Pal y Shiu, 2004).

#### Aprendizaje en sistemas RBC

Una vez que se genera una solución adecuada y da una salida, hay cierta expectativa de que la solución se ponga a prueba en la realidad, véase la Figura 2.8. Para probar una solución, tenemos que considerar tanto la forma en que puede ser probada y cómo los resultados de la prueba lo clasificará como un éxito o un fracaso. Usando esta evaluación en el mundo real, un sistema RBC puede ser actualizado para tener en cuenta cualquier nueva información descubierta en el procesamiento de la nueva solución.

#### Métodos de aprendizaje

El sistema evoluciona y mejora la competencia y eficiencia de los resultados como producto de almacenar la experiencia pasada en el sistema y recuperar los casos pasados en el razonamiento futuro (Pal y Shiu, 2004).

Se define un aprendizaje como:

$$BC = BC \bigcup \{e_{m+1}\} \tag{2.54}$$

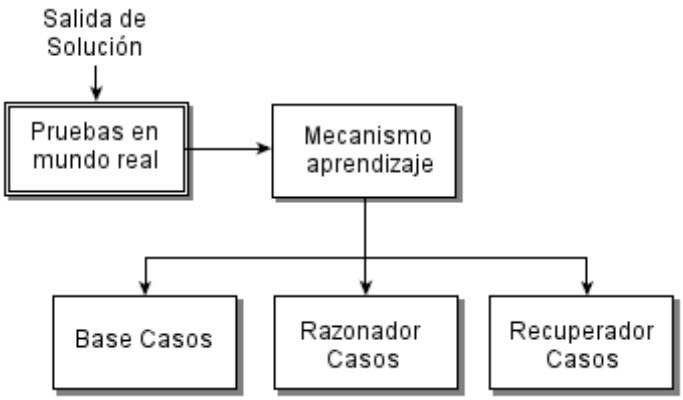



Figura 2.9: Mecanismo de aprendizaje en un RBC
(Pal y Shiu, 2004).

donde $\{e_{m+1}\}$ corresponde al caso $m+1$ producto del aprendizaje, representa una experiencia significativa con una nueva instancia sintética $\{a_{(m+1)}1, a_{(m+1)}2, ...a_{(m+1)}n\}$, vea las Ecuaciones (2.45), (2.46) en la página 24.

El aprendizaje puede ocurrir de varias maneras. Es un método común la adición de un nuevo problema, su solución, y el resultado a la base de casos. La base de casos incrementará la diversidad de situaciones cubiertas por los casos almacenados y reduce la distancia media entre un vector de entrada y el vector más cercano almacenado.

Otro método de aprendizaje en un sistema RBC es usar la solución evaluada para modificar los valores de los casos almacenados o modificar los criterios de recuperación de casos.

Se define un aprendizaje con modificación del caso $i$ como:

$$e_i \leftarrow \{a'_i1, a'_i2, ...a'_in\} \tag{2.55}$$

donde $\{e_i\}$ corresponde al caso a modificar, y $\{a'_i1, a'_i2, ...a'_in\}$ representa una nueva instancia sintética, vea las Ecuaciones (2.45), (2.46) en la página 24.

Si un caso tiene valores que no son relevantes para los contextos específicos en que debe ser recuperado, ajustamos los índices para que pueden aumentar la correlación

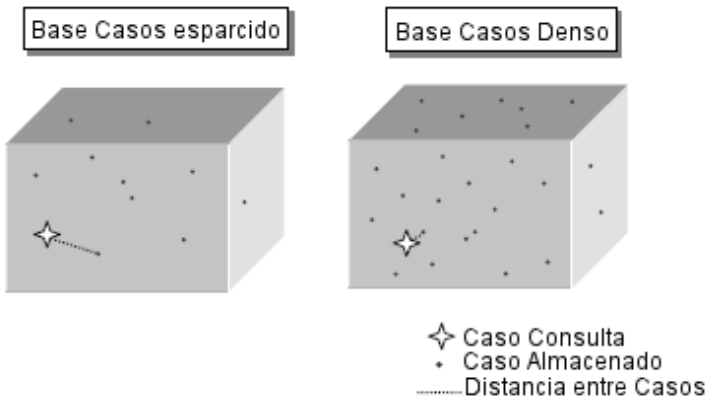



Figura 2.10: Distancia entre casos
(Pal y Shiu, 2004).

entre las ocasiones en que un caso es realmente recuperado y las ocasiones en las que debería haber sido recuperado.

### Consideraciones para agregar casos

Según Sankar Pal (Pal y Shiu, 2004), cuando el aprendizaje implica que se deben agregar nuevos casos a la base de casos, hay una serie de consideraciones:

- ¿En qué situaciones debe agregarse un caso a la base de casos, y en que situaciones se debe descartar? Tenemos que considerar el nivel de éxito de la solución, que tan similar es el caso actual con otros casos en la base de casos, y si hubiera importantes lecciones que se tuviera que aprender del caso.

- Si es que se añadiera el caso a la base de casos, los índices del nuevo caso debe ser determinadas cómo es que el caso se va agregar a la base de casos. Si la estructura de la base de casos y el método de recuperación son muy estructurados es decir usan estructuras jerárquicas determinadas por inducción o un conjunto de redes neuronales, la incorporación de un nuevo caso puede requerir una planificación y re-estructuración significativa de la base de casos.

**Lineamientos para el uso de RBC**

A pesar de que el RBC es útil en muchos dominios y problemas, hay ocasiones donde no es la mas apropiada metodología a utilizar. Los problemas candidatos y sus dominios deben reunir ciertas características que se mencionan a continuación (Pal y Shiu, 2004):

- *¿Se tiene un modelo de fondo?* Si el dominio es imposible de entender completamente o si los factores que determinan el éxito o fracaso de una solución no pueden ser modelados explicitamente; el RBC permite trabajar con la experiencia pasada sin comprender los mecanismos de fondo (Ejemplo Sistemas de pronostico financiero o de diagnóstico).

- *¿Hay casos nuevos o excepcionales?* Dominios sin casos nuevos o excepcionales pueden ser modelados con sistemas basados en reglas, las cuales se determinan inductivamente a partir de los datos históricos. Si embargo, en situaciones donde nuevas experiencias y excepciones son encontradas frecuentemente, harían difícil mantener la consistencia de las reglas del sistema. En este escenario las características de aprendizaje incremental convertirían a un sistema de RBC en una mejor alternativa a un sistema basado en reglas.

- *¿Existen Casos Recurrentes?* Si la experiencia de un caso no es probable de ser usada para un nuevo problema, por tener un bajo grado de similitud, hay poco valor en almacenar los casos. En otras palabras cuando las experiencias no son los suficientemente similares para ser comparados y adaptados, es mejor construir un modelo del dominio para derivar la solución.

- *¿Hay un beneficio significativo en adaptar una solución pasada?* Se debe considerar si hay un beneficio significativo en términos de recursos, tiempo de desarrollo, procesamiento al crear una solución a través de la modificación de una solución

similar en vez de crear una solución desde el principio.

- *¿Son relevantes los casos previos obtenibles?* ¿Es posible obtener datos que registren las características necesarias de los casos pasados? ¿Los casos registrados contienen las características relevantes del problema y su contexto influye en el resultado de la solución? ¿Tiene la solución guardada el suficiente detalle para ser adaptada en el futuro? si las respuestas son positivas permiten usar el marco del RBC(Pal y Shiu, 2004).

### Ventajas del uso de RBC

A continuación se resumen algunas de las ventajas en el uso del RBC (Pal y Shiu, 2004):

- Razonamiento a partir de datos incompletos o imprecisos: No es necesario tener toda la información para hacer inferencias, bastara con unos atributos relevantes.
- Aprendizaje interactivo: mientras el sistema crece, el sistema se entrena y aprende; utiliza los casos nuevos para trabajar con los nuevos; las redes neuronales tienen bien diferenciada una fase de entrenamiento que no lo hacen interactivo a las nuevas soluciones.
- Reducción de la tarea de adquisición de conocimiento: se elimina la necesidad de extraes un modelo formal o un conjunto de reglas.
- Evita repetir errores del pasado: Así como los casos de éxito, también se almacenan los errores, en sistemas de generalización como las redes neuronales simplemente solo se trabaja con casos exitosos.
- Extensible a un amplio rango de dominios: El RBC puede ser aplicado a un extremo, amplio y variado dominio de aplicaciones.

- Reflejan la forma de razonar humana: Los humanos no nos complicamos para la solución de problemas, buscamos a partir de la experiencia propia o ajena y plantemos soluciones rápidas y brillantes.

## 2.5. Métodos de acceso métrico

Los Métodos de Acceso Métrico (MAM) se enfocan en el problema de organización de datos para que, en base a un criterio de similitud, usado en la fase de recuperación del Razonamiento Basado en Casos, pueda facilitar la búsqueda de un conjunto de elementos que estén cerca de un elemento de consulta (Chávez, Navarro, Baeza-Yates, y Marroquín, 2001). Este problem está presente en un sinfin de aplicaciones que van desde escenarios de la vida cotidiana hasta las ramas de las ciencias de la computación, como el reconocimiento de patrones o la recuperación de información.

Tradicionalmente, las estructuras de datos han aplicado operaciones de búsqueda, donde se hace una coincidencia exacta. Por ejemplo, en las bases de datos donde se manejan registros, cada registro es comparado con los demás por medio de una clave y las búsquedas retornan los registros cuya clave coincida con la clave suministrada.

Tras la aparición de nuevos contextos, debido principalmente al desarrollo tecnológico, vienen surgiendo nuevos algoritmos y métodos de acceso más eficientes y veloces. En las búsquedas por similitud o proximidad, la similitud entre elementos es modelada a través de una función de distancia que satisfaga la desigualdad triangular, y un conjunto de objetos llamado espacio métrico.

### 2.5.1. Definiciones

Los Métodos de Acceso Métrico son estructuras ampliamente utilizadas en el campo de Recuperación de Información. Un MAM debe organizar un conjunto de datos en base a un criterio de similitud para responder eficientemente a consultas específicas de

proximidad.

Los Métodos de Acceso Métrico puedes ser descritos como una herramienta de organización de datos. Los MAMs trabajan sobre espacios métricos definidos por un conjunto de objetos y una función de distancia que mide la disimilitud entre los objetos del espacio métrico (Chávez y cols., 2001). Consideremos un conjunto $U$ que denota el universo de objetos válidos y la función $d : U \times U \longrightarrow R$ que mide la distancia entre objetos. Se define como espacio métrico al subconjunto $S \subseteq U$ de tamaño $n = |S|$ llamado diccionario o base de datos, que denota el conjunto de objetos de búsqueda, y a la función $d(x, y)$ que mide la disimilitud entre objetos y satisface las propiedades de:

- $\forall x, y \in U, d(x, y) \geq 0$, positividad;
- $\forall x, y \in U, d(x, y) = 0$, simetría;
- $\forall x \in U, d(x, x) \ = 0$, reflexibilidad;
- $\forall x, y \in U, x \neq y \Rightarrow d(x, y) \geq 0$, positividad estricta;
- $\forall x, y, z \in U, d(x, y) \leq d(x, z) + d(z, y)$, desigualdad triangular.

La desigualdad triangular es la propiedad más importante porque establece los límites de distancias que aún pueden no haberse calculado, generando algoritmos de búsqueda por similitud significativamente más rápidos.

Para los espacios vectoriales (un caso particular de espacios métricos) donde cada objeto es descrito como un vector de características $(x_1, x_2, x_3, \ldots, x_n)$ varios Métodos de Acceso Espacial (MAE) como Kd-Tree o R-Tree han sido propuestos para indexar este tipo de objetos multidimensionales. El problema principal de los espacios vectoriales está relacionado con las altas dimensiones de los datos, la también conocida maldición de la dimensionalidad (Chávez y cols., 2001).

### 2.5.2. Consultas de Proximidad

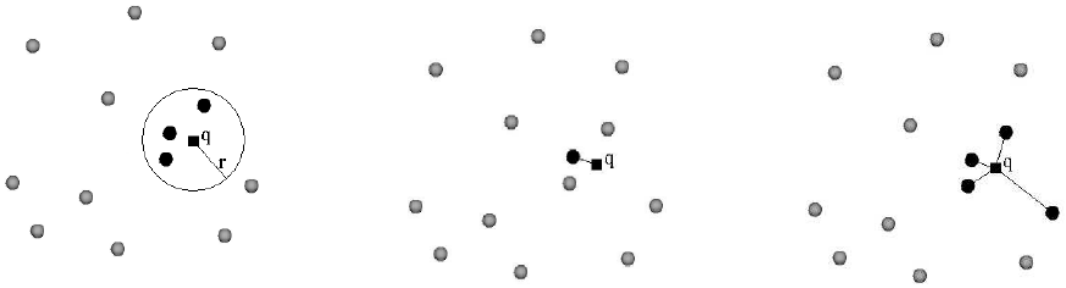
Dado un objeto de consulta $q \in U$, para poder recuperar los objetos similares a $q$, se definen los siguientes tipos básicos de consulta:

**Consultas de rango** $Rq(q, r)$. Recupera todos los elementos que se encuentran dentro de un radio $r$ de $q$. Esto es, $u \in U = d(q, v)/d(q, u) \leq r$.

**Consulta de vecino más cercano**$NN(q)$. Recupera el elemento en $U$ más cercano a $q$. Esto es $u \in U/\forall v \in U, d(q, u) \leq d(q, v)$. Adicionalmente se puede establecer un rango máximo $r$.

**Consulta de k-vecinos más cercanos** $NNk(q)$. Recupera los $k$ elementos en $U$ más cercanos a $q$. Esto es, $A \subseteq U/|A| = k \wedge \forall u \in A, v \in U - A, d(q, u) \leq d(q, v)$.

La Figura 2.11 muestra ejemplos de las consultas generadas.




(a) (b) (c)

Figura 2.11: Tipos básicos de consultas por proximidad:(a) Ejemplo de búsqueda por rango $r$ en un conjunto de puntos. (b) Ejemplo de búsqueda del vecino más cercano en un conjunto de puntos. (c) Ejemplo de búsqueda de los $k$-vecinos más cercanos en un conjunto de puntos con $k = 4$.

### 2.5.3. Algoritmos de Búsqueda

Los Métodos de Acceso Métrico son estructuras que trabajan sobre espacios métricos, organizando los datos para responder eficientemente a consultas por similitud. De

acuerdo con (Zezula, Amato, Dohnal, y Batko, 2006), los MAMs pueden ser clasificados en:

- Particionamiento de esferas: *Fixed Queries Tree (Baeza-Yates, Cunto, Manber, y Wu, 1994), Vantage Point Tree (Uhlmann, 1991).*
- Particionamiento de hiperplanos: *Generalized Hyper-plane Tree (Uhlmann, 1991).*
- Distancias Precomputadas: *Omni-Family (Filho, Traina, Jr., y Faloutsos, 2001), Approximating and Eliminating Search Algorithm (Ruiz, 1986).*
- Métodos híbridos: *GNAT (Brin, 1995), Spatial Approximation Tree (Navarro, 2002), Multi Vantage Point Tree (Bozkaya y Özsoyoglu, 1997).*
- Otros métodos: *M-Tree (Ciaccia, Patella, y Zezula, 1997), Slim-Tree (Jr., Traina, Seeger, y Faloutsos, 2000), DIndex (Dohnal, Gennaro, Savino, y Zezula, 2003).*

La Figura 2.12 muestra otra clasificación de los Métodos de Acceso Métrico propuesta en (Chávez y cols., 2001), aquí se clasifican a los métodos de búsqueda en: basados en agrupamiento y basados en pivotes. Los métodos basados en agrupamiento particionan el espacio en regiones representadas por un centroide o centro de grupo, para luego poder descartar regiones completas cuando se hace una búsqueda. Los métodos basados en pivotes seleccionan un conjunto de elementos como pivotes, y construyen un índice en base a las distancias entre cada elemento y los pivotes.

Se pueden encontrar buenas referencias sobre clasificación y definición de los MAMs en (?, ?) y (Hjaltason y Samet, 2003).

### 2.5.4. Omni-Secuencial

La técnica Omni (Filho y cols., 2001) hace uso de un conjunto de puntos de referencia llamados "focos" para reducir el número de cálculos de distancia. Cada vez que

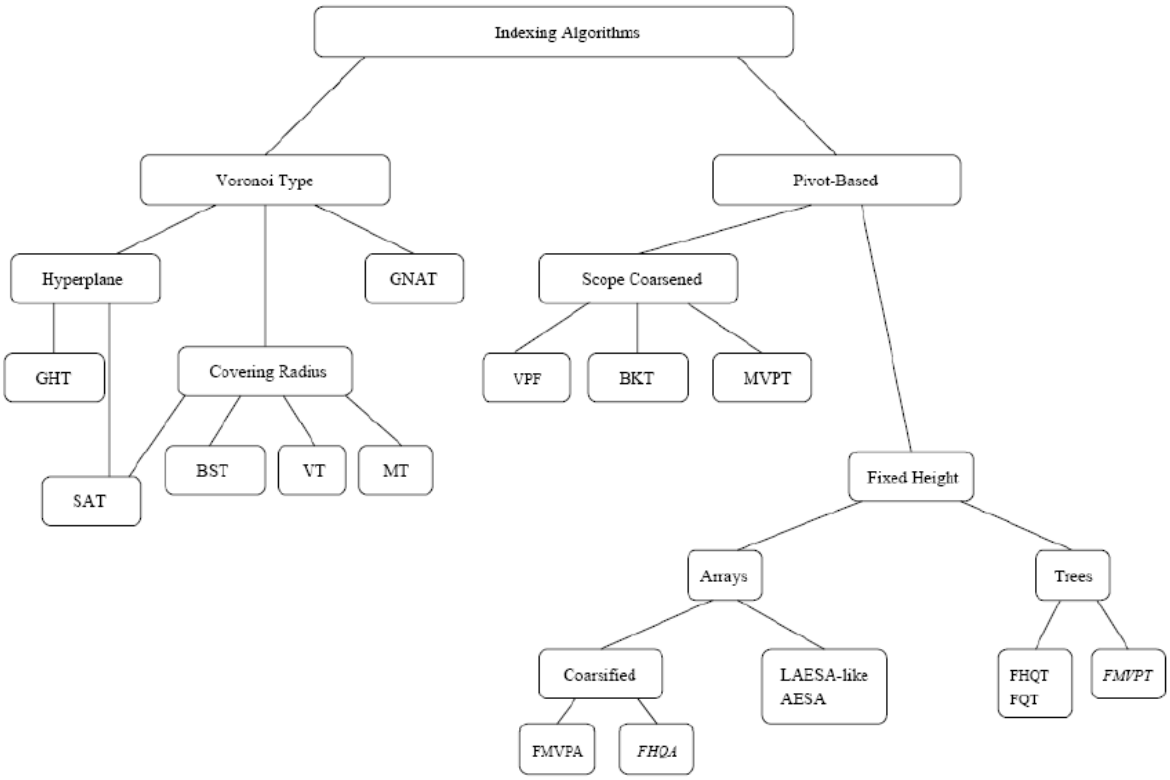
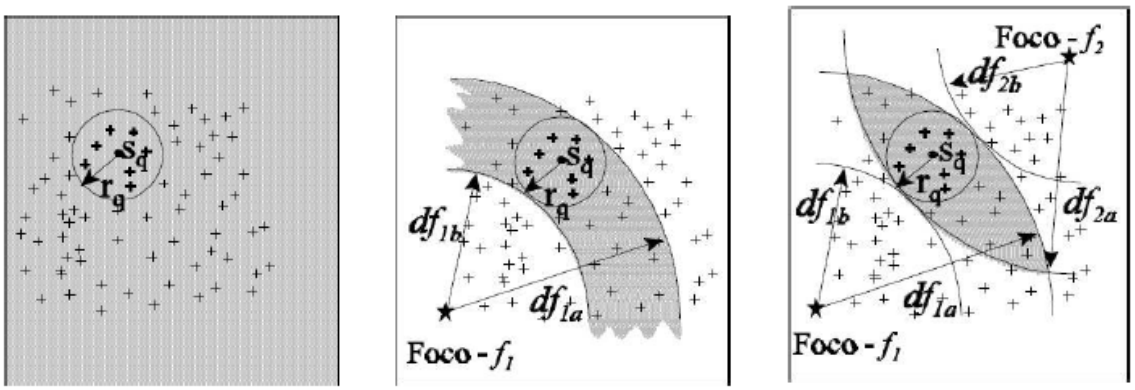



Figura 2.12: Taxonomía de algoritmos en base a sus características. (Chávez y cols., 2001)

se inserta un nuevo elemento se calculan las distancias de este elemento hacia cada uno de los focos, información que es luego utilizada en las consultas para reducir los cálculos de distancia haciendo uso de la propiedad de la desigualdad triangular vista anteriormente.

Esta técnica introduce los conceptos de Omni-focos y Omni-coordenadas. Los Omnifocos son definidos como el conjunto $F$ de distintos puntos que pertenecen al espacio métrico. Las Omni-coordenadas son definidas como el conjunto de distancias calculadas entre cada punto del espacio métrico y cada elemento de $F$, por lo tanto la cardinalidad de la coordenada es igual al número de focos. El costo adicional de calcular las Omni-coordenadas es compensado por el ahorro obtenido en las consultas.

Uno de los puntos críticos en esta técnica es la selección del conjunto de focos $F$ y su cardinalidad. Con respecto a la cardinalidad, y como puede ser observado en la Figura 2.13, con un mayor número de focos se puede reducir más el subconjunto de candidatos, acelerando el proceso de búsqueda, pero se requiere mayor espacio y tiempo para procesarlos. Los autores recomiendan una cardinalidad no mayor al doble de la dimensionalidad intrínseca de los datos ya que a un mayor número de focos se obtiene un pequeño o ningún beneficio.




(a) (b) (c)

Figura 2.13: Tipos básicos de consultas por proximidad:(a) Sin uso de focos todo el conjunto de datos es candidato. (b) Usando un foco el subconjunto de datos candidatos (área sombreada) se reduce. (c) Subconjunto de candidatos usando dos focos.

Para la selección de los focos se recomienda elegir puntos lo más separado posible y equidistantes. Se propone el algoritmo HF para esta tarea. Este algoritmo primero selecciona un objeto aleatoriamente y luego selecciona como primer foco al objeto más alejado de este. El segundo foco es el elemento más distante el primer foco. Por último se selecciona como siguiente foco a aquel objeto con distancias más similares a los focos anteriores, este proceso se repite hasta seleccionar todos los focos. El proceso completo está descrito en el Algoritmo 1 (Filho y cols., 2001).

**Algorithm 1** Algoritmo HF

1: Seleccionar aleatoriamente un elemento $s_0$ del conjunto de datos.
2: Encontrar el elemento $f_1$ más lejano a $s_0$ y seleccionarlo como foco.
3: Encontrar el elemento $f_2$ más lejano a $f_1$ y seleccionarlo como foco.
4: Encontrar el elemento $f_1$ más lejano a $s_i$ y seleccionarlo como foco.
5: Establecer $edge = d(f1, f2)$, variable usada para encontrar a los demás focos.
6: Mientras se necesiten encontrar más focos repetir los pasos 7 y 8.
7: Para cada punto si del conjunto de datos calcular:
$error_i = \sum_k^k esfoco|edge - d(f_k; s_i)|$.
8: Seleccionar como foco al elemento $s_i$ que posea el menor $error_i$ y que no haya sido seleccionado anteriormente como foco.

## 2.6. Álgebra relacional

Para poder expresar algunas operaciones sobre una Base de Casos multidimensional es necesario usar una notación matemática que permita incluir expresiones de consulta en una expresión matemática, se decide incorporar el Álgebra Relacional para mejorar la expresividad de la propuesta (Romero, Marcel, Abelló, Peralta, y Bellatreche, 2011), (Hajdinjak y Bierman, 2011)

### 2.6.1. Definición

El Álgebra Relacional (AR) es un lenguaje teórico abstracto con operaciones que trabajan sobre relaciones, para definir nuevas relaciones o subconjuntos de ellos sin cambiar las originales, la salida de una operación puede ser la entrada de otra operación (Sumathi y Esakkirajan, 2007), (Elmasri y Navathe, 2010).

Elmasri (Elmasri y Navathe, 2011) sugiere que cualquier modelo de datos debe incluir un conjunto de operaciones para manipularlos, además de conceptos para definir la estructura y las limitaciones del modelo de la base. Estas operaciones permiten al usuario especificar solicitudes de recuperación como expresiones matemáticas. El resultado es una nueva relación, la que se puede manipular adicionalmente usando los operadores del álgebra.

| Nombre | Operador |
|---|---|
| asignación | $\leftarrow$ |
| selección | $\sigma$ |
| proyección | $\Pi$ |
| inner join | $\bowtie$ |
| productor cartesiano | $\times$ |
| renombrado | $\rho$ |
| menor que | $<$ |
| mayor que | $>$ |
| menor que o igual | $\leq$ |
| mayor que o igual | $\geq$ |
| igual | $=$ |
| y | $\wedge$ |
| o | $\vee$ |
| no | $\neg$ |
| unión | $\cup$ |
| intersección | $\cap$ |
| división | $\div$ |
| diferencia | $-$ |

Cuadro 2.1: Operadores relacionales

El álgebra relacional es importante porque proporciona una base formal para mostrar las operaciones de relaciones como expresiones matemáticas. En segundo lugar, se utiliza como base para la aplicación y optimización de consultas.

El álgebra relacional se considera a menudo parte integral del modelo de datos relacional. Sus operaciones se pueden dividir en dos grupos. Un grupo incluye las operaciones previstas en la teoría matemática de conjuntos, aplicables por las definiciones de tuplas en el modelo relacional formal. Un juego de estas operaciones incluyen UNION, INTERSECCION, DIFERENCIA DE CONJUNTOS, y el PRODUCTO CARTESIANO (también conocido como producto vectorial). El otro grupo está formado por las operaciones desarrolladas específicamente para bases de datos relacionales, que incluyen SELECCION, PROYECCION, entre otros; se presenta un conjunto de ellos en el cuadro 2.1.

### 2.6.2. Operador relacional unario: Selección

La operación de selección se utiliza para elegir un subconjunto de registros de una relación que satisface una **condición de selección**[3]. Se puede considerar la operación SELECCION como un filtro que mantiene sólo las tuplas que satisfacen una condición de clasificación. Alternativamente, podemos considerar la operación de selección para restringir las tuplas de una relación a sólo aquellos tuplas que satisfacen la condición. La operación SELECCION también puede ser visualizada como una partición horizontal de la relación en dos conjuntos de tuplas, aquellas tuplas que satisfacen la condición y están seleccionadas, y las tuplas que no cumplen la condición y se descartan.

En general, la operación SELECCION se denota por:

$$\sigma_{<condicion\ de\ seleccion>}(R) \tag{2.56}$$

donde se utiliza el símbolo $\sigma$ (sigma) para denotar el operador SELECCION y la *condición de seleccion* es una expresión que especifica sobre que atributos de la relación $R$ se aplica. Nótese que $R$ es generalmente una expresión del álgebra relacional cuyo resultado es una relación mas simple. La relación resultante de la operación de selección tiene los mismos atributos que $R$.

En *condición de selección* se pueden usar los operadores de comparación $\{=, <, \leq, >, \geq, \neq\}$ aplicandose a atributos cuyos dominios son valores numéricos u otros.

### 2.6.3. Operador relacional unario: Proyección

Cuando se está interesado en seleccionar ciertos atributos de una relación, se utiliza la operación Proyección, el resultado de la operación se puede visualizar como una partición vertical de la relación en dos relaciones: uno tiene las columnas necesarias

[3]La operación de selección es diferente de la cláusula SELECT de SQL. La operación elige registros en una tabla, algunas veces se le llama restrición o FILTRO DE FUNCIONAMIENTO.

(atributos) y contiene el resultado de la operación, y el otro contiene las columnas desechadas.

En forma general de la operación PROYECCION se denota por:

$$\Pi< attribute\ list >(R) \tag{2.57}$$

donde $\pi$ (pi) es el símbolo usado para representar la operación PROYECCION, y <atributo de lista> es la lista secundaria deseada de los atributos de la relación R. Una vez más, cuenta de que R es, en general, una expresión de álgebra relacional cuyo resultado es una relación, que en el caso más simple es sólo el nombre de una relación de base de datos. El resultado de la operación PROYECCION sólo tiene los atributos especificados en atributo <list> en el mismo orden en que aparecen en la lista. Por lo tanto, su grado es igual al número de atributos en <atributo de lista>Ṡi la lista de atributos sólo incluye atributos sin clave de $R$, tuplas duplicadas son probables de ocurrir. La operación PROYECCION elimina las tuplas duplicadas, por lo que el resultado de la operación PROYECCION es un conjunto de tuplas distintas, y por lo tanto una relación válida. Esto se conoce como eliminación de duplicado.

El número de registros o tuplas en una relación resultante de una operación de proyección es siempre menor o igual al número de tuplas en R. Si la lista de proyección es una superclave de R es-que, que incluye alguna clave de R-la relación resultante tiene el mismo número de tuplas como R. Por otra parte, $\Pi_{<lista1>}(\Pi_{<lista2>}(R)) = \Pi_{<lista1>}(R)$ siempre y cuando <list2> contiene los atributos en <list1> de lo contrario, el lado izquierdo es una expresión incorrecta.

Se puede encontrar mas información sobre los operadores y el Álgebra Relacional en (Sumathi y Esakkirajan, 2007; Elmasri y Navathe, 2010).

## 2.7. Consideraciones finales

En este capítulo se ha descrito teóricamente los procesos estocásticos, conceptos de variable aleatoria, modelos lineales ARMA, PARMA, se ha visto la importancia del ruido blanco como un bloque que describe un Proceso Estocástico básico; Luego la definición de series temporales y algunos estimadores usados para describirlos, finalmente el Razonamiento Basado en Casos, Métodos de acceso métrico y el álgebra relacional; todos estos conceptos serán de utilidad para comprender las bases sobre la que se desarrolla la propuesta.

En el siguiente Capítulo se desarrollará el estado del arte, y como algunos modelos lineales (PAR1) son utilizados para la generación de series temporales (Modelo de Thomas Fiering). Luego las nuevas propuestas en el área basadas en Redes Neuronales, que reutilizan algunos conceptos aqui presentados.

# Capítulo 3

# Estado del Arte

---

Para el modelado de un Proceso Estocástico los modelos tradicionales (aproximaciones lineales) son modelos poco eficientes y de aplicabilidad limitada, los modelos no-lineales, requieren un conocimiento profundo del dominio para su construcción, siendo finalmente de formulación compleja (Campos, 2010; Han y Wang, 2009; Kantz y Schreiber, 2004), ahora bien existen trabajos que proponen la solución a este problema usando procesos estocásticos basado en redes neuronales, algunos especializados a fenómenos con características periódicas (Campos, 2010; El-Shafie y El-Manadely, 2011; Ochoa-Rivera, 2008; Bao y Cao, 2011); de las propuestas se destaca la contribución de Luciana Conceicao en su tesis doctoral *Modelo Estocástico Periódico baseado em Redes Neurais* (Campos, 2010), usada para generar series temporales de caudales el 2010. Luego existen otros trabajos, donde se muestra la capacidad del Razonamiento Basado en Casos para descubrir información oculta, se tiene los trabajos de Maria Malek en su tesis doctoral *Case-based Reasoning in Knowledge Discovery and Data Mining* (Malek y Kanawati, 2009), Ning Xiong (Funk y Xiong, 2006) que trabaja sobre series temporales el 2009; sobre la capacidad de pronóstico del RBC se tiene el trabajo de Pei-Chann Chang *Application of a Case Based Reasoning for Financial*

*Time Series Data Forecasting* (Chang, Tsai, Huang, y Fan, 2009) el 2009.

## 3.1. Modelo Estocástico de Thomas-Fiering

Un modelo para la generación de series temporales estocásticas fue desarrollado por Thomas y Fiering (Thomas y Fiering, 1962). Este modelo ademas de la media y la varianza, usa el coeficiente de correlación, pues se considera que los registros históricos de procesos hidrológicos presentan un fenómeno de persistencia observable (Cadavid y Salazar, 2008)

### 3.1.1. Descripción

$$Q_{j+1} = \bar{Q}_{j+1} + b_j \left(Q_j - \bar{Q}_j\right) + t_j.s_{j+1}\sqrt{1 - r_j^2} \tag{3.1}$$

donde:

$\bar{Q}j$ es el caudal en el mes j

$Qj$ es el caudal promedio en el mes j

$Bj$ es la pendiente de la recta de regresión entre el mes j y j+1

$Sj$ es la varianza de la distribución de los caudales en el mes j

$Rj$ es el coeficiente de correlación entre el mes j y j+1

$Tj$ es un número aleatorio que viene de una distribución normal de media nula y de varianza igual a uno.

Para calcular los promedios, la pendiente, la varianza y el coeficiente de correlación de los datos históricos.

El promedio:

$$\bar{Q}_j = \frac{1}{n}\sum_{i=1}^{n} Q_j \tag{3.2}$$

La varianza:

$$s_j = \sqrt{\frac{1}{n-1}\sum \left(Q_j - \bar{Q}_j\right)^2} \tag{3.3}$$

El coeficiente de correlación:

Para j mayor o igual a 2

$$r_j = \frac{\frac{1}{n-1}\sum \left(Q_j - \bar{Q}_j\right)\left(Q_{j-1} - \bar{Q}_{j-1}\right)}{s_j s_{j-1}} \tag{3.4}$$

Para j igual a 1

$$r_1 = \frac{\frac{1}{n-1}\sum \left(Q_1 - \bar{Q}_1\right)\left(Q_m - \bar{Q}_m\right)}{s_1 s_m} \tag{3.5}$$

La pendiente de la recta de correlación:

$$\begin{aligned} &b_j = \tfrac{r_j s_j}{s_{j-1}} \quad para \quad j \geq 2 \\ &para \quad j = 1 \quad b_1 = \tfrac{r_1 s_1}{s_m} \end{aligned} \tag{3.6}$$

**Para generar datos con una distribución log normal** Si el caudal mensual sigue una distribución log normal, su logaritmo sigue una distribución normal, se suele usar y:

$$y_{j+1} = \bar{y}_{j+1} + b_{yj}\left(y_j - \bar{y}_j\right) + t_j . s_{yj+1}\sqrt{1 - r_{yj}^2} \tag{3.7}$$

Para calcular los parámetros se tiene:

$$\begin{aligned} \bar{Q}_j &= e^{\bar{y}_j + \frac{s_{yj}^2}{2}} \\ s_j &= e^{2s_{yj}^2 + 2\bar{y}_j} - e^{s_{yj}^2 + 2\bar{y}_j} \\ r_j &= \frac{e^{s_{yj-1} s_{yj} r_{yj}} - 1}{\sqrt{e^{s_{yj-1}^2} - 1}\sqrt{e^{s_{yj}^2} - 1}} \\ b_{yj} &= \frac{r_{yj} s_{yj}}{s_{yj-1}} \end{aligned} \tag{3.8}$$

Se resuelve las dos primeras Ecuaciones 3.2 3.3 para calcular el promedio y la desviación estándar de la nueva variable. Se obtienen estas dos ecuaciones:

$$\begin{aligned} \bar{y}_j &= Ln\left(\bar{Q}_j\right) - \frac{s_{yj}^2}{2} \\ s_{yj}^2 &= Ln\left[1 + \left(\frac{s_j}{\bar{Q}_j}\right)^2\right] \end{aligned} \tag{3.9}$$

Ahora, para calcular los dos otros parámetros:

$$\begin{aligned} r_{yj} &= \frac{1}{s_{yj-1}s_{yj}} Ln\left[r_j\sqrt{e^{s_{yj-1}^2}-1}\sqrt{e^{s_{yj}^2}-1}+1\right] \\ b_{yj} &= \frac{r_{yj}s_{yj}}{s_{yj-1}} \end{aligned} \tag{3.10}$$

Se usan los 4 parametros para generar los datos sintéticos y(i,j). Luego para obtener los caudales sintéticos que siguen una distribución lognormal se toma el exponencial de y:

$$\bar{Q}_j = e^{y_j} \tag{3.11}$$

### 3.1.2. Generación sintética de flujos

El primer trabajo para generar caudales sintéticos es el de escoger una buena distribución. Se calcula los estimadores (la media, la varianza, etc.) con el método: *maximum likelihood estimation.*

Con n observaciones independientes {x1,. . . , xn} de una variable aleatoria, la función de densidad de probabilidad es:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n\,|\theta) = f_X(x_1\,|\theta)\,\ldots f_X(x_n\,|\theta) \tag{3.12}$$

Donde $\theta$ es el vector de los parámetros ($\mu$ y $\sigma$).

Se tiene sólo que maximizar la función f. Por ejemplo para la función log normal:

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}[\ln(x) - \mu]^2) \tag{3.13}$$

Entonces:

$$\begin{aligned} \overset{\wedge}{\mu} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ln(x_i) \\ \overset{\wedge}{\sigma^2} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} (\ln(x_i - \overset{\wedge}{\mu})^2 \end{aligned} \tag{3.14}$$

Se verifica que la ley log normal da resultados correctos, con el test de Kolmogorov-Smirnov. La comparación entre la distribución de probabilidad y la distribución empírica está definida como: $Prob(Xi < x) = i/n$

Ahora para verificar que la ley normal funciona, se calcula la desviación máxima entre las dos curvas:

$$D = \text{máx}_{1 \leq i \leq n}(F(Y_i) - \frac{i-1}{n}, \frac{i}{n} - F(Y_i)) \tag{3.15}$$

Después de calcular la desviación es fácil ver si la distribución da una buena representación de la realidad.

Después es fácil generar muchas distribuciones con las mismas características de los datos de entrada. En efecto con un algoritmo simple se puede generar datos sintéticos.

## 3.2. Modelo Estocástico Periódico basado en Redes Neuronales de Campos

### 3.2.1. Descripción

El comportamiento caótico y la no-linearidad de los datos a fomentado recientes investigaciones en la generación de series temporales con Redes Neuronales (Kantz y Schreiber, 2004; Campos, 2010) los modelos tradicionales que hacen uso de aproximaciones lineales se han convertido en modelos poco eficientes y de aplicabilidad limita-

da, y los modelos no-lineales, necesitan un conocimiento profundo del dominio para su construcción (Campos, 2010; Han y Wang, 2009) Una de las características que hacen ventajoso el uso de Redes Neuronales es la no necesidad de asumir un tipo de distribución a priori, aprenden la distribución a través de ejemplos y manejan datos de diversas fuentes con diferentes niveles de precisión y ruido. (Vieira y cols., s.f.; Prudencio, 2002)

El uso de redes neuronales hace que el proceso estocástico neuronal sea un modelo no-lineal capaz de capturar las características de la serie temporal, sin la necesidad de hacer suposiciones a priori sobre el comportamiento de la serie o efectuar algún tipo de descomposición en la misma. Para ello es preciso que las entradas de las redes neuronales del modelo de proceso estocástico neuronal tengan una memoria de corto plazo, la cual debe contener los términos pasados de la serie temporal a ser simulada. Los parámetros del modelo de proceso estocástico neuronal corresponden a los pesos sinápticos de las redes neuronales y para simular las realizaciones estocásticas es necesario adicionar un valor aleatorio a las salidas de las redes neuronales. Estos valores aleatorios son obtenidos a través de las distribuciones de probabilidad de los residuos de las redes neuronales del proceso estocástico neuronal.

Para poder trabajar con las series temporales periódicas, los parámetros del modelo de proceso estocástico neuronal se deben ajustar no sólo al intervalo del tiempo de la serie sino también al periodo. El proceso estocástico neuronal es modelado con una componente estocástica para cada periodo de la serie. Por ejemplo, en el caso del periodo mensual en proceso estocástico neuronal esta compuesto por 12 componentes estocásticas (una para cada mes), y si el periodo seria semestral, la cantidad de componentes estocásticos seria de 2 y en el caso de la serie no periódica apenas se usa un compuesto estocástico. Cada componente estocástico del proceso neuronal estocástico esta formada por una red neuronal y por una distribución de probabilidad para generar valores aleatorios en la generación de escenarios como se ilustra en la Figura 3.1.

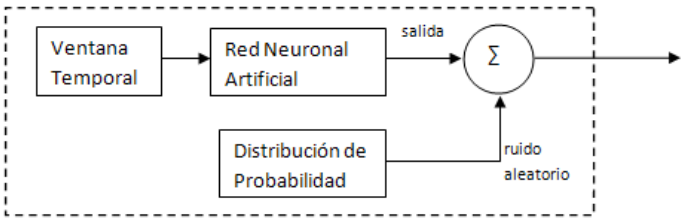



Figura 3.1: Componente estocástico del proceso estocástico neuronal. (Campos, 2010)

Cuando el proceso estocástico neuronal está formado por mas de un componente estocástico ocurre un encadenamiento entre ellos, donde el valor de la serie dado por el componente estocástico de un periodo forma parte de la ventana temporal de entradas de la red neuronal del componente estocástico del siguiente periodo.

El proceso estocástico neuronal es clasificado como un modelo estocástico periódico no-linear autoregresivo genérico.

### 3.2.2. Proceso Estocástico Neuronal

Sea $Z(t)$ una serie temporal con un periodo estacionario $s$ y con $n$ observaciones simultáneas en todos los periodos. El índice de tiempo $t$ es descrito por la Ecuación 3.16

$$t = (r-1) \cdot s + m \tag{3.16}$$

donde:

$r = 1 \ldots n$ es el número de observaciones de cada periodo de la serie.

$m = 1 \ldots s$ corresponde a un periodo de la serie.

$s \epsilon N$ y es el total de periodos de la serie.

$n \cdot s$ es el tamaño de la serie observada.

Para que las redes neuronales *feedforward* se comporten como un modelo de procesamiento temporal, es necesario que ellas presenten habilidades de memoria de corto plazo, la cual es realizada a través de técnicas de "ventana" (Gutierrez, 2003). Esta técnica consiste en introducir memoria en las neuronas de la primera capa escondida, otorgando de esta forma a las neuronas valores pasados de la serie temporal. Por eso el proceso estocástico neuronal es clasificado como un modelo autoregresivo.

La red neuronal de cada componente estocástico del proceso estocástico neuronal posee un número determinado de términos pasados de la serie, llamados orden de la red neuronal. El orden de la red neuronal del componente estocástico del periodo $m$ es representado por $p_m$. Para obtener un valor de la serie en un instante de tiempo $t$, el proceso estocástico neuronal accede al componente estocástico $m$ correspondiente y su red neuronal recibe los $p_m$. La Figura 3.2 muestra la estructura de la red neuronal de orden $p_m$.

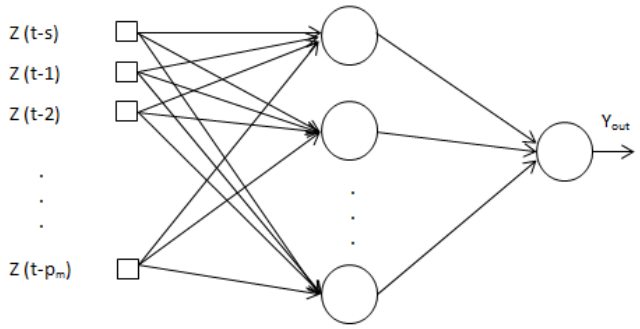



Figura 3.2: Red neuronal del proceso estocástico neuronal de orden $p_m$.

La Figura 3.3 representa en detalle a la neurona perteneciente a la capa oculta de la red neuronal de orden $p_m$, cuya salida esta dada por la Ecuación 3.17

$$y_i = \varphi(\omega_{i,0} \cdot Z(t-s) + (\sum_{j=1}^{p_m} \omega_{i,j} \cdot Z(t-j)) + \theta_i) \tag{3.17}$$

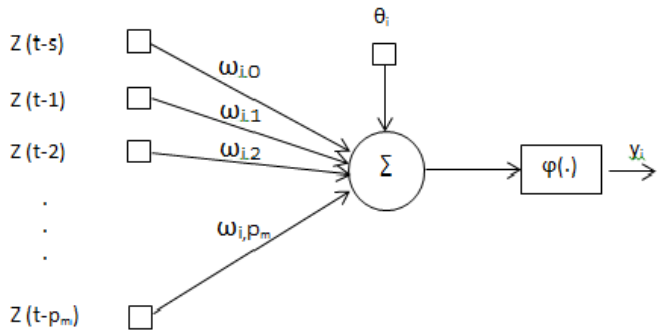



Figura 3.3: Neurona de la capa oculta de red neuronal del proceso estocástico neuronal de orden $p_m$.

donde $\varphi$ es la función de activación de la neurona $i$, $\omega_{i,j}$ es el peso sináptico de la conexión entre la entrada $j$ y la neurona $i$ y $\theta_i$ es el bias de esta neurona.

Considerando que la red neuronal de orden $p_m$ contiene $l_m$ neuronas en la capa oculta, esta puede ser representada como se muestra en la Figura 3.4, donde esta salida es calculada por la Ecuación 3.18:

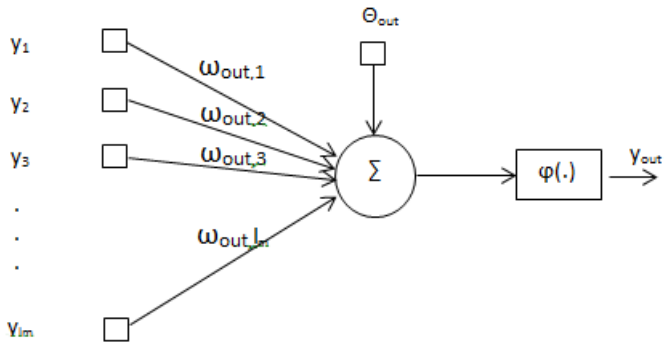



Figura 3.4: Neurona de salida de una red neuronal del proceso estocástico neuronal con $l_m$ neuronas en la capa oculta.

$$y_{out} = \varphi_{out}(\sum_{i=1}^{l_{mm}} \omega_{out,i} \cdot y_i + \theta_{out}) \tag{3.18}$$

donde $\varphi_{out}$ es la función de activación de la neurona de la capa de salida representado por $out$, $\omega_{out,i}$ es el peso sináptico de la conexión entre la entrada $i$ y la neurona $out$ y $\theta_{out}$ es el bias de la neurona.

Como se ve en la Figura 3.1, la salida de un componente estocástico corresponde a la sumatoria de la salida de las redes neuronales con un valor aleatorio proveniente de la distribución de probabilidad de residuos de la red neuronal. La serie temporal $Z(t)$ que posee como índice de tiempo $t$ descrito por la Ecuación 3.16 es simulada a través de la siguiente Ecuación:

$$Z(t) = y_{out} + \alpha(t) \tag{3.19}$$

donde $\alpha(t)$ es el valor aleatorio proveniente de la distribución de probabilidad de los residuos de la red neuronal de los componentes estocásticos del periodo $m$. Uniendo las Ecuaciones 3.17 3.18 4.3 obtenemos la descripción matemática de la componente estocástica del periodo $m$ del proceso estocástico neuronal.

$$Z(t) = y_{out}(\Sigma_{i=1}^{l_m}\omega_{out,i} \cdot \varphi_i[\omega_{i,0}Z(t-s) + (\Sigma_{j=1}^{p_m}\omega_{i,j}Z(t-j)) + \theta_i]\theta_{out}) + \alpha(t) \tag{3.20}$$

Los términos de las series son simulados por el proceso estocástico. La Figura 3.5 muestra de forma genérica el encadenamiento de los componentes estocásticos del proceso estocástico neuronal en un determinado tiempo $t$.

Se debe ajustar el modelo del proceso estocástico neuronal a la serie temporal histórica a ser simulada, este ajuste debe seguir los siguientes pasos:

- Determinar la estructura del modelo,
- Estimar los primeros parámetros y
- Validar los residuos.

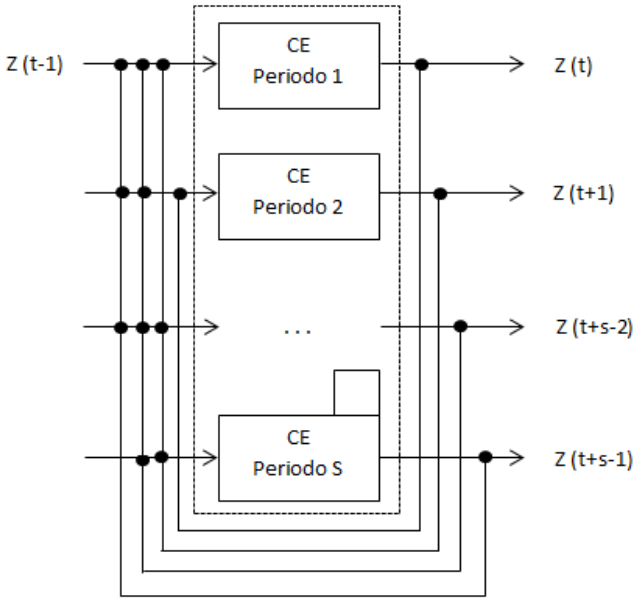



Figura 3.5: Encadenamiento entre las entradas/salidas de las componentes estocásticas del proceso estocástico neuronal.

### 3.2.3. Determinación de la Estructura de los Procesos Estocásticos Neuronales

La arquitectura de la red neuronal consiste en determinar el número de entradas, salidas, capas ocultas, neuronas por capa, padrón de conexión entre las neuronas y la función de activación. Para la determinación de la arquitectura general de la red neuronal se usa una sola capa oculta(según Haykin (Haykin, 2001)) con funciones sigmoideas para la activación de las neuronas. Posee una sola neurona en la capa de salida y el número de neuronas de la capa oculta es determinado en forma empírica (probando las diferentes arquitecturas de redes neuronales y variando el número de neuronas en la capa oculta). El modelo del proceso estocástico neuronal referenciado por $PEN(p, l)$. La cantidad de parámetros del modelo es la suma del número de parámetros (número de

pesos sinápticos, incluyendo el bias de la red neuronal) de cada componente estocástico del proceso estocástico neuronal.

$$\sum_{m=1}^{s} p_m^{l_m} + 2 \cdot l_m + 1 \tag{3.21}$$

La definición del modelo $PEN(p, l)$ consiste en la identificación de los términos $p$ y $l$, los cuales pueden ser determinados a partir de estudios preliminares sobre la serie o por tentativa de error.

En el modelo de proceso estocástico neuronal, los pesos de la red son ajustados por un algoritmo de entrenamiento supervisado, donde los parámetros utilizados son formados por el conjunto de entradas y el conjunto de salidas deseadas. Este algoritmo de entrenamiento es ejecutado por un número dado de épocas donde en cada época los pesos sinápticos son ajustados de forma independiente.

Para cada red neuronal es creado un conjunto de padrones de entrenamiento con salidas deseadas, y datos de entrada normalizados dentro los límites establecidos por la función de activación.

Como el entrenamiento es supervisado, la respuesta de la neurona de salida es comparada con la respuesta deseada que se encuentra en el padrón de los datos. La diferencia de estos valores corresponden al error usado en el ajuste de pesos sinápticos por el algoritmo de entrenamiento, y el calculo del desempeño del entrenamiento.

La métrica para medir el desempeño de los modelos de series temporales es el error medio porcentual absoluto (MAPE) (Tang, 1991). El MAPE es calculo a través de la Ecuación 3.22

$$MAPE = \frac{1}{N} \cdot \sum_{k=1}^{N} \left| \frac{Z(k) - Y(K)}{Z(K)} \right| \cdot 100 \tag{3.22}$$

donde $N$ corresponde al total de padrones y $Z(k)$ es el valor de la $k$-ésima salida deseada del padrón de entrenamiento del periodo $m$. $Y(k)$ es la salida desnormalizada

de la red neuronal del periodo $m$ para el $k$-ésimo padrón de entrada.

El objetivo de la etapa de evaluación es generar un escenario de $x \cdot s$ elementos como se ilustra en la Figura 3.6, envolviendo de esta forma todas las redes neuronales del proceso estocástico neuronal. La construcción del escenario es realizada de forma secuencial a través del encadenamiento entre las redes, donde la red neuronal $m = 1$ muestra la ventana temporal y genera la salida, el cual es el primer elemento del escenario de evaluación que es usado en la ventana temporal de la red $m + 1$.

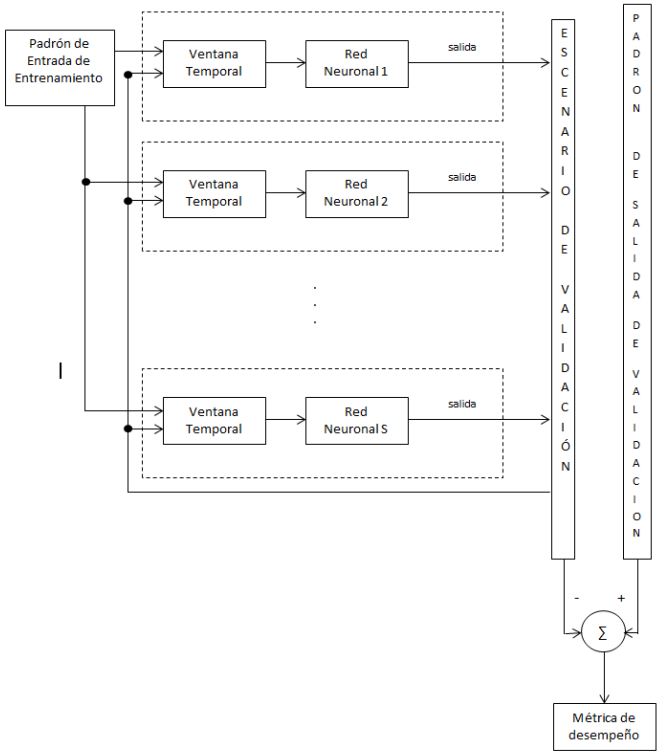



Figura 3.6: Evaluación de las redes neuronales del proceso estocástico neuronal.

Se calcula una métrica de desempeño, similar a la usada en el entrenamiento usando los datos de evaluación. Para calcular la métrica de desempeño de evaluación, se compara los datos del escenario con los datos que se encuentran en el padrón de salida deseada. De esta forma se tiene dos tipos de métrica de evaluación:

1. Por escenario: se calcula la métrica recorriendo de manera secuencial todo el escenario.

2. Se comparan los valores de $x$ del periodo $m$ presentes en el escenario con los patrones $x$ de salida del conjunto de evaluación de periodo $m$.

Existen dos formas de calcular el MAPE de evaluación obteniendo dos tipos de función de costo para evaluar la interrupción de entrenamiento:

1. En conjunto: Se interrumpe el entrenamiento de todas las redes neuronales del proceso estocástico neuronal como se muestra en la Figura 3.7.

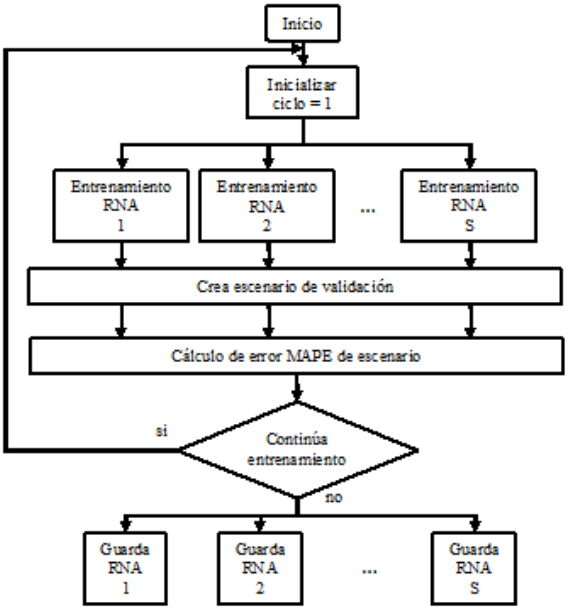



Figura 3.7: Evaluación de las redes neuronales del proceso estocástico neuronal.

2. Separado por red neuronal: Cuando el MAPE comienza a subir la red neuronal interrumpe su entrenamiento en distintas épocas, como se ilustra en la Figura 3.8.

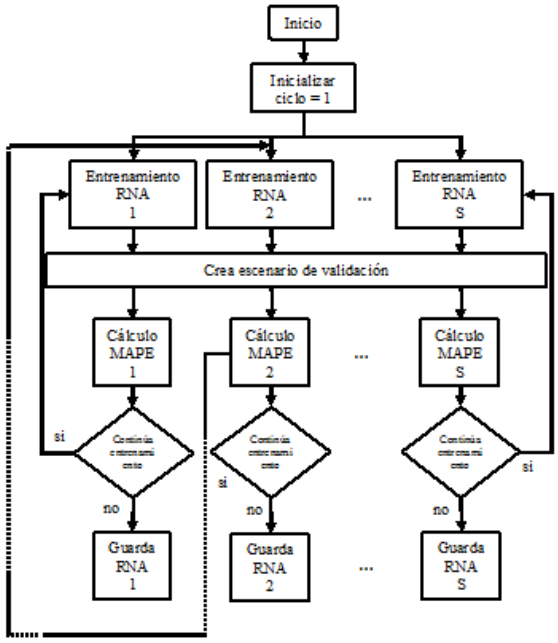



Figura 3.8: Evaluación de las redes neuronales del proceso estocástico neuronal.

### 3.2.4. Evaluación de los Residuos Generados

Durante la fase de entrenamiento de la red neuronal artificial es calculado un conjunto de diferencias entre la salida dada por la red neuronal y la salida deseada del padrón de entrenamiento. Al termino del entrenamiento, el conjunto de diferencias obtenido por la red neuronal del periodo $m$ corresponde a las serie de residuos del estimador.

En esta etapa se busca ajustar una distribución de probabilidad teórica que tenga una buena adherencia con la serie de residuos de la red neuronal entrenada para el periodo $m$, luego a través de la distribución de probabilidad teórica se obtiene una descripción aproximada de las características de los residuos. Para verificar la adherencia de la distribución se usa la prueba de Kolmogorov-Smirnov el cual ayuda a conseguir el menor error de ajuste, en este caso corresponde a la distribución de probabilidad

del componente estocástico del periodo $m$ del proceso estocástico neuronal como se muestra la Figura 3.1.

## 3.3. Otros Trabajos Relacionados

### 3.3.1. Razonamiento Basado en Casos en el Descubrimiento de Conocimiento y Minería de Datos

Maleky Kalawaty presenta el 2010 la tesis de PHD con algunas contribuciones en tres áreas de investigación: razonamiento basado en casos, descubrimiento de conocimientos y representación del conocimiento. Se introduce un lenguaje para representar variaciones entre casos. Primero se muestra como este lenguaje puede ser utilizado para representar la adaptación del conocimiento y modelar la fase de adaptación en el razonamiento basado en casos. Este lenguaje es luego aplicado a la tarea de aprendizaje del conocimiento de adaptación. El proceso de descubrimiento del conocimiento, llamado CabamakA, aprende el conocimiento adaptado por generalización a partir de una representación de variaciones entre los casos. La discusión continúa sobre cómo hacer este proceso de descubrimiento del conocimiento operacional en una adquisición de conocimiento. La discusión conduce a la proposición de un nuevo enfoque para la adquisición de conocimiento de adaptación, en el cual el proceso de descubrimiento de conocimiento es lanzado como una manera oportunista en el tiempo de resolución del problema. Los conceptos introducidos en esta tesis son ilustrados en el dominio de tema a través de su aplicación en el sistema TAAABLE, de razonamiento basado en casos, que constituye el dominio de la aplicación del estudio (Malek y Kanawati, 2009).

### 3.3.2. Razonamiento Basado en Casos en aplicaciones con series de tiempo

Basado en Ning Xiong (Funk y Xiong, 2006). Este trabajo discute sobre el rol e integración del descubrimiento del conocimiento (DC) en sistemas de razonamiento basado en casos (RBC). La opinión general es que DC es complementaria a la tarea de conocimiento de retención y puede ser tratado como un proceso separado fuera del tradicional ciclo RBC. A diferencia de la retención de conocimiento que esta relacionado a experiencias de casos específicos, los objetivos del DC en la elicitación del nuevo conocimiento son más generales y valiosas para mejorar las diferentes tareas del RBC. El trabajo se ejemplificó por un escenario de aplicación real en la medicina en el que series de tiempo de patrones son analizados y clasificados. Como un único patrón no puede transmitir la información suficiente en la aplicación, las secuencias de patrones son más adecuadas. Por lo tanto, es más ventajoso si las secuencias de patrones y su co-ocurrencia con las categorías pueden ser descubiertas. La evaluación de los casos que contienen series clasificadas en un número de categorías e inyectadas con secuencias de indicadores muestra que el enfoque es capaz de identificar secuencias ocultas. En una aplicación clínica con una biblioteca de casos representativa del mundo real, estas secuencias clave mejoraran la habilidad de clasificación y puede generar investigación clínica para explicar la co-ocurrencia entre ciertas secuencias y clases.

### 3.3.3. Aplicación del Razonamiento Basado en Casos para series de tiempo de datos de Pronóstico Financiero

Sobre la capacidad de pronóstico del RBC se tiene el trabajo de Pei-Chann Chang *Application of a Case Based Reasoning for Financial Time Series Data Forecasting* (Chang y cols., 2009).

Este trabajo establece un modelo de predicción de series de tiempo financieros, por

clustering y la evolución del *Support Vector Machine* para las acciones de S & P 500 en los E.E.U.U. Este modelo de predicción integra una técnica de clustering de datos con RBC ponderado, clustering con un *Support Vector Machine (SVM)* para construir un sistema de toma de decisiones basado en datos históricos y técnicas de indexación. El precio futuro de las acciones es predicho por el modelo propuesto y la precisión de modelo de predicción se mejora al dividir la data histórica en diferentes clusters. En general, los resultados apoyan el nuevo modelo para predecir el precio de acciones al mostrar que puede reaccionar precisamente a la tendencia actual del movimiento del precio de las acciones a a partir de estos casos más pequeños. La tasa de éxito del modelo RBC-SVM es 93,85 %, el más alto rendimiento, a la fecha.

## 3.4. Consideraciones finales

En este capítulo se ha presentado los modelos usados en la literatura para la generación de series temporales asociadas a variables climatológicas, el modelo lineal de Thomas Fiering, luego un modelo basado en redes neuronales (no-lineal, propuesto recientemente)y otros especializado a fenómenos con características periódicas (Campos, 2010; El-Shafie y El-Manadely, 2011; Ochoa-Rivera, 2008; Bao y Cao, 2011); de las propuestas se destaca la contribución de Luciana Conceicao, que trabajan si información a priori y que no requieren de una formulación compleja, se evidencian algunas limitaciones sobre la aplicabilidad de las propuestas para caracterizar información oculta. Luego se presentan algunos trabajos, donde se muestra la capacidad del Razonamiento Basado en Casos para descubrir información oculta, se tiene los trabajos de Maria Malek en su tesis doctoral *Case-based Reasoning in Knowledge Discovery and Data Mining* (Malek y Kanawati, 2009) de Ning Xiong (Funk y Xiong, 2006) que trabaja sobre series temporales, sobre la capacidad de pronóstico del RBC se tiene el trabajo de Pei-Chann Chang (Chang y cols., 2009).

En el siguiente Capítulo se describirá, a un nivel de detalle significativo, el Razonamiento Basado en Casos, se apreciará sus ventajas y desventajas, su capacidad para trabajar con información oculta, finalmente se discutirá sobre su aplicabilidad en la generación de series temporales estocásticas.

# Capítulo 4

# Propuesta: Modelo Estocástico a partir de Razonamiento Basado en Casos para la Generación de Series Temporales

---

En este capítulo se presenta el nuevo modelo de Proceso Estocástico a partir del Razonamiento Basado en Casos; el objetivo es generar series temporales que muestran información oculta. Para ello en la etapa de representación se indexa las series temporales de los registros históricos en una estructura de datos de acceso secuencial, Para ello se propone, en la etapa de representación un modelo con memoria a corto plazo, multidimensional. se sugiere la indexación en una estructura de acceso secuencial; luego en la etapa de recuperación, la búsqueda y generación de un componente determinístico a partir de la extensión de los modelos con memoria auto-regresiva de 3 términos, donde se cambia los parámetros promedio, varianza, coeficiente de correlación y pendi-

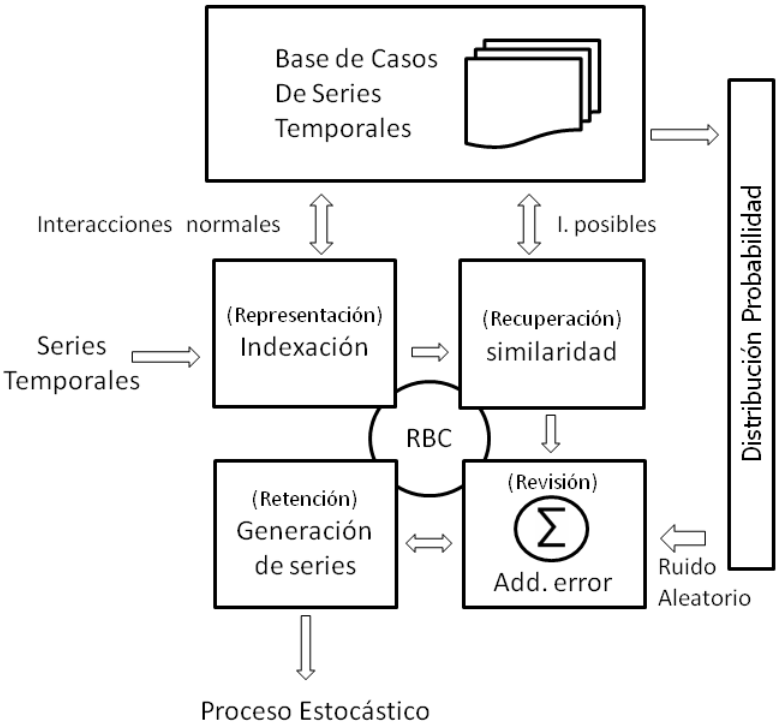



Figura 4.1: Etapas del Proceso Estocástico a partir del Razonamiento Basado en Casos.

ente de la recta de regresión, por una función de similitud. La búsqueda por similitud usará la distancia euclidiana basada en la ubicación de objetos en el espacio euclidiano representado por un vector $(n+1) - dimensional$ donde $n$ es una entrada ponderada por el coeficiente de correlación de las variables relativas al caso de búsqueda; en la etapa de reutilización se genera una realización estocástica, agregando un error aleatorio, proveniente de una distribución de probabilidad asociada a la ventana de similitud buscada; la etapa de Retención almacena las series temporales generadas que cumplan las consideraciones físicas; vea la Figura 4.1 Etapas del Proceso Estocástico a partir del Razonamiento Basado en Casos. A continuación vea el detalle de la propuesta.

## 4.1. Componente estocástico

El Proceso Estocástico a partir de Razonamiento Basado en Casos es modelado con un componente estocástico para cada periodo de la serie; para un periodo mensual, el nuevo proceso está compuesto por 12 componentes estocásticos (uno para cada mes), y si el periodo es semestral, la cantidad de componentes estocásticos seria 2, para una diaria se tendría 360 y para el caso de una serie no periódica solo un componente estocástico. Cada componente esta formado por la Base de Casos, un razonador basado en casos con su algoritmo de recuperación, una distribución de probabilidad para generar valores aleatorios, los elementos del componente estocástico son ilustrados en la Figura 4.2.

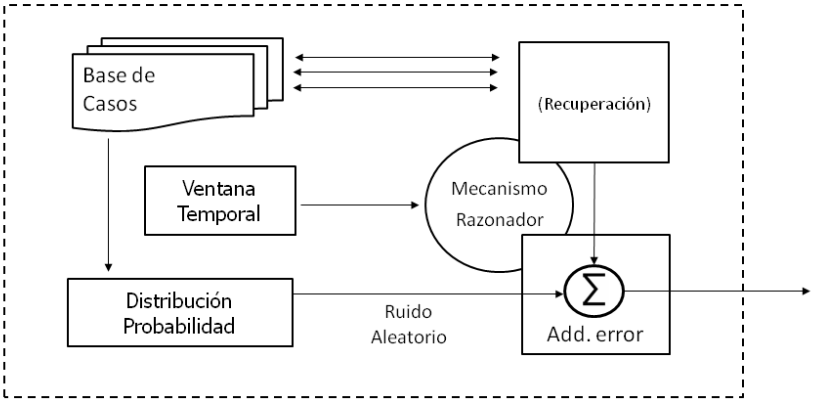



Figura 4.2: Componente estocástico del proceso estocástico a partir de Razonamiento Basado en Casos.

Cuando el proceso estocástico a partir de Razonamiento Basado en Casos está for-

mado por mas de un componente estocástico ocurre un encadenamiento entre ellos, donde el valor de la serie, dado por el componente estocástico de un periodo, forma parte de la ventana temporal de entradas del componente estocástico del siguiente periodo; el proceso estocástico a partir de Razonamiento Basado en Casos es clasificado como un modelo estocástico periódico auto-regresivo genérico.

## 4.2. Representación e Indexación de casos

Como se menciono en la sección 2.4.3, la base de un sistema RBC es la memoria de casos, a diferencia de otros métodos que usan abstracciones o modelos basados en dominio (redes neuronales, inferenciales, clasificadores en general); se representan a partir de registros históricos de series temporales, organizados por el espacio temporal y/o geográfico.

### 4.2.1. Representación de Casos

La entidad caso, para series temporales debe relacionar variables con características comunes.

Se presenta a continuación el diseño del esquema para una Base de Casos de registros temporales, una representación gráfica la tiene en la Figura 4.3.

$$e = (x, y_1, y_2, z_1, z_2, ..., z_n) \tag{4.1}$$

donde $x$ es un indice. $y_1$ es un atributo que describe la temporalidad del registro histórico. $y_2$ es un atributo que describe la ubicación del registro histórico. $z_1, z_2, ..., z_n$ son atributos que describen las $n$ dimensión relativas al registro histórico.

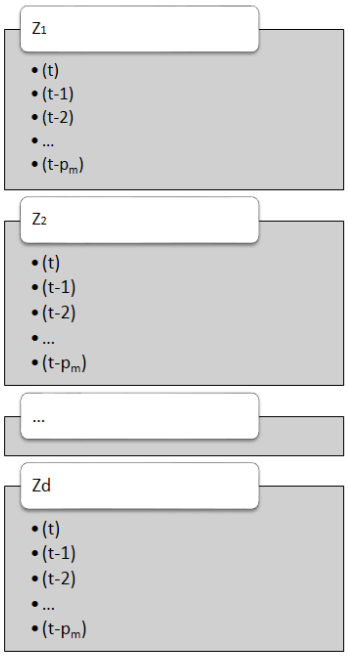



Figura 4.3: Registro de Caso Serie Temporal Genérico

### 4.2.2. Indexación de casos para series temporales

Puesto que el RBC trabaja con la memoria de toda la serie histórica, la indexación es importante para obtener los casos similares en un tiempo rápido. Se sugiere que los índices sean abstractos para permitir la recuperación en varias circunstancias (Bonzano, Cunningham, y Smyth, 1997). De acuerdo a la sección 2.4.3: Indexación de casos, se indexa todos los valores de atributos numéricos que influyan en la generación de un dato para la serie temporal. para saber el grado de importancia, se pondera de acuerdo al coeficiente de correlación de los atributos. La clave primaria es asignada al registro a buscar en las consultas, y la clave secundaria a los atributos asociados.

### 4.2.3. Indexación sobre una estructura de acceso métrico

Se sugiere usar una estructura que soporte rangos y búsquedas multidimensionales ponderadas, con un método de acceso métrico, se recomienda utilizar el $Omni - secuencial$ con un memoria estructurada en $flatmemory$, para mas detalle sobre estos métodos, vea las secciones 2.4.3, 2.5.4.

## 4.3. Recuperación de casos para series temporales

Para que el proceso de recuperación en un RBC, representado en la Figura 2.7, se comporte como un modelo de procesamiento temporal, es necesario que presente habilidades de memoria de corto plazo, para ello en la formulación de un caso se debe incluir retrasos temporales, con una ponderación basada en el coeficiente de correlación y una técnica de "ventana" (Gutierrez, 2003). Esta técnica introduce memoria en el razonador, a través de las series de tiempo pasadas; por eso el proceso estocástico es clasificado como un modelo auto-regresivo.

El razonador de cada componente estocástico del proceso posee un numero determinado de términos pasados de la serie, se llamará orden o grado del razonador. El orden del razonador, del componente estocástico del periodo $m$, es representado por $p_m$. Para obtener un valor de la serie en un instante de tiempo $t$, el proceso accede al componente estocástico $m$ correspondiente y su razonador recibe los $p_m$; asociado al orden se tiene dimensiones $d$, la primera dimensión corresponde a los datos históricos de las serie trabajada ($d = 1$), las dimensiones adicionales son series de temporales asociadas por el coeficiente de correlación $w$ a la primera dimensión, el razonador trabaja con todas $d$ dimensiones; a mas dimensiones, mejores resultados. La Figura 4.4 muestra la estructura de un razonador de orden $p_m$ y dimensión $d$.

Se tiene una nueva formulación de las variables: Sea $Z_1(t)$ una serie temporal con un periodo estacionario $s$ y con $n$ observaciones simultáneas en todos los periodos,

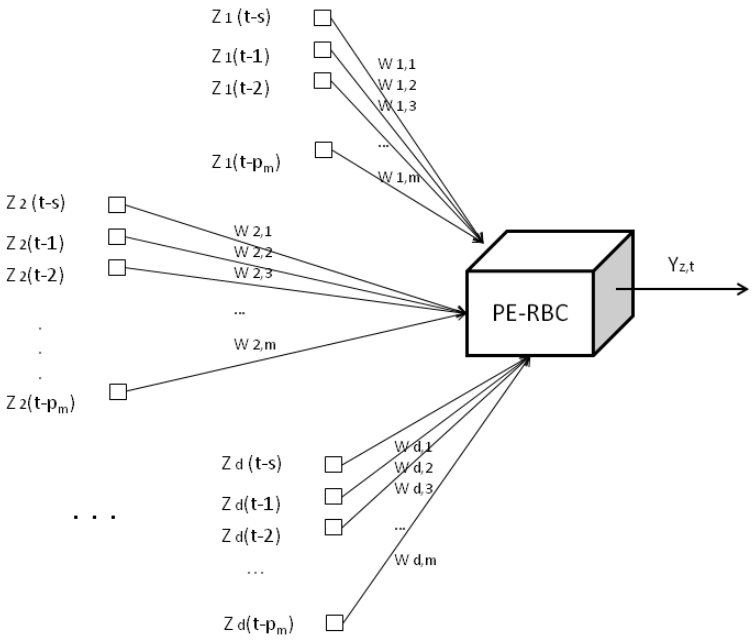



Figura 4.4: Proceso Estocástico Genérico a partir de Razonamiento Basado en Casos de orden $p_m$ y $d$ dimensiones.

correlacionada a series asociadas $Z_2(t)$ ... $Z_d(t)$.

Se describe un índice de tiempo $t$, vea la Ecuación 4.2

$$t_d = (r-1) \cdot s + m \tag{4.2}$$

donde:

$r = 1 \ldots n$ es el número de observaciones de cada periodo de la serie.

$m = 1 \ldots s$ corresponde a un periodo de la serie.

$s$ es el total de periodos de la serie $s \epsilon N$.

$d$ son las dimensiones de la series.

$\beta w_d$ es la ponderación extraído del coeficiente de correlación de la serie $d$ con la serie generada.

En la Figura 4.2 se aprecia que la salida de un componente estocástico, corresponde a la recuperación de h series temporales con un mecanismo razonador y un valor aleatorio proveniente de la distribución de probabilidad, asociado a un error del mecanismo razonador. La serie temporal $Z(t)$ que posee como indice de tiempo $t$ es simulada a través de la siguiente ecuación:

$$Z(t) = y_t + \alpha(t) \tag{4.3}$$

donde $\alpha(t)$ es el valor aleatorio proveniente de la distribución de probabilidad asociado a los errores de los componentes estocásticos del periodo $m$.

$Y_t$ es la salida del mecanismo razonador, el mecanismo razonador se basa en una medida de similitud. La expresión que expresa la nueva forma de modelar el proceso estocástico, teniendo en cuenta la medida de similitud, es:

$$Z_{j+1} = Sim_j(Z_j, BC), +\alpha(j) \tag{4.4}$$

donde: $Z_j$ es el componente estocástico en el instante de tiempo $j$.
$Sim_j(Z_j, BC)$ es la función de similitud para el mes j en base a los datos históricos registrados en las series temporales de $BC$.
$\alpha(j)$ es un error aleatorio que proviene de una distribución de probabilidad para el instante de tiempo $j$ generado.

### 4.3.1. Concepto de similitud

En el contexto de Generación sintética de series temporales, se asume que las series presentan un fenómeno de persistencia observable, el cual se encontrará por una medida de similitud de la persistencia sobre los datos históricos, se define como caso un subconjunto de una serie histórica observada. El trabajar con este concepto es posible por los enfoques siguientes:

- Se basa en el cálculo de la distancia, entre los casos en donde se determina el caso más similar por una medida (es decir métrica) de evaluación de similitud.
- El segundo enfoque está relacionado con las estructuras representación/indexación de los casos, el cual recorre en busca de un caso similar, aquí se enfatiza la utilidad de los métodos de acceso métrico.

### 4.3.2. Distancia Euclidiana Ponderada

Es forma mas directa para medir una distancia, esta basado en la ubicación de los objetos en el espacio Euclideano (es decir un conjunto ordenado de números reales). Formalmente la distancia Euclidiana entre los casos se expresará de la siguiente manera:

$$BC = \{e_1, e_2, ...e_N\} \tag{4.5}$$

donde $BaseCasos$ es la librería de $N$ casos correspondiente a las series históricas almacenadas, y $e_i$ representa una medida en el instante $i$.

Ademas se tiene la colección de atributos correspondientes a las dimensiones asociadas $\{F_j(j = 1, 2, ..., n)\}$ para indexar los registros; luego:

$$e_i = (x_{i1}, x_{i2}, ..., x_{in}, \theta_i) \tag{4.6}$$

donde: $e_i$ es el i-ésimo caso en la librería , se representado por un vector $(n+1)-$ *dimensional* $x_{ij}$ corresponde al valor de la dimensión $F_j(1 \leq j \leq n)$ $\theta_i$ corresponde a los valores de ubicación no indexados $V(i = 1, 2, ..., N)$.

Para cada valor de la serie representada en el caso $\{F_j(j = 1, 2, ..., n)\}$, se asigna un peso $w_j(w_j \in [0, 1])$ asignado a la j-ésima dimensión para indicar la influencia de dicha observación en nuestro valor buscado, este se obtiene a partir del coeficiente de correlación entre los atributos, previamente calculado.

Entonces, para la ventana temporal $e_p$ y la salida buscada $e_q$ en la librería de registros históricos, la distancia métrica ponderada se define como:

$$d_{pq}^{(w)} = d^{(w)}\left(e_p, e_q\right) \tag{4.7}$$

$$d_{pq}^{(w)} = \left[\sum_{j=1}^{n} w_j^2 (x_{pj} - x_{qj})^2\right]^{1/2} \tag{4.8}$$

$$d_{pq}^{(w)} = \left(\sum_{j=1}^{n} w_j^2 x_j^2\right)^{1/2} \tag{4.9}$$

donde $x_j^2 = (x_{pj} - x_{qj})^2$. Cuando todos los pesos son iguales a 1, la distancia métrica ponderada definida anteriormente degenera a la medida Euclidiana $d_{pq}^1$, esto quiere decir que es denotado por $d_{pq}$.

La medida de similitud entre dos datos; $SM_{pq}^{(w)}$, se define como:

$$SM_{pq}^{(w)} = \frac{1}{1 + \alpha d_{pq}^{(w)}} \tag{4.10}$$

Donde $\alpha$ es una constante, cuanto más alto sea el valor de $d_{pq}^{(2)}$ , la similitud entre $e_p$ y $e_q$ es mas bajo. Cuando todos los pesos toman valor de 1, la medida de similitud es denotado por $SM_{pq}^{(1)}$, $\in [0, 1]$.

Para cada característica una medida de distancia ha sido definida. La medida de

distancia para el j-ésimo atributo esta denotado por $\rho_j$ ; que es , $\rho_j$ es un mapeo de $F_j x F_j a[0, \infty]$ (donde $F_j$ es denotado como el dominio del j-ésimo atributo) con las siguientes propiedades:

$$\rho_j(a, b) = 0 \leftrightarrow a = b \tag{4.11}$$

$$\rho_j(a, b) = \rho_j(b, a) \tag{4.12}$$

$$\rho_j(a, b) \leq \rho_j(a, c) + \rho_j(c, b) \tag{4.13}$$

Se pueden definir otros atributos como la transición diferencial, y otros numéricos generados a partir de los históricos se tiene:

$$\rho_j(a, b) =\mid a - b \mid, a, b \in R. \tag{4.14}$$

donde

$$\rho_j(A, B) = max_{a \in A, b \in B} \mid a - b \mid siAyBsonintervalos. \tag{4.15}$$

Para estos atributos, la distancia entre dos casos $e_p$ y $e_q$ se calcula por:

$$d_{pq}^w = \sqrt{\sum_{j=1}^{n} w_j^2 \rho_j^2(e_{pj}, e_{qj})} \tag{4.16}$$

### 4.3.3. Ponderación vía coeficientes de correlación

Según la sección 2.4.4 se ponderan las variables intervinientes en el mecanismo razonador asignándole un peso en función del impacto o influencia de estos en el resultado, para ello se puede usar un experimento o técnicas de agrupación a un coeficiente

de correlación. El mecanismo razonador usa la distancia euclidiana ponderada de la salida del componente estocástico buscado contra los $(n+d)-1$ dimensiones y ordenes de las series asociadas, el peso de la ponderación es representada por $w$, el cual es generado por el coeficiente de correlación de $Z$ con las dimensiones y ordenes asociadas. En procesos periódicos se puede definir valores que describen la estructura de correlación lineal de un periodo con los periodos anteriores, puede ser de orden 1 con el inmediato anterior, o una correlación de orden 2 que describe la dependencia del periodo $m$ con respecto a los periodos $m-2$, o generalizando, una correlación de orden $k$ que representa la dependencia del periodo $k$ con respecto al periodo $m-k$.

**Cálculo del peso** Los valores que puede tomar el coeficiente de correlación $r$ son: $-1 < r < 1$; si se realiza una ponderación los valores negativos, generación valores inconsistentes, por lo que se usa una escala relativa de fuerza de $[0 \quad a \quad 1]$

El signo indica la dirección de la correlación, positiva o directamente proporcional (a mayor A mayor B o a menor B menor A) y negativa o inversamente proporcional (a menor A mayor B o viceversa).

La cifra indica la *fuerza de la correlación*. Una correlación perfecta tendría una cifra cercana al 1 o -1, mientras que una ausencia de correlación tendría una cifra cercana al 0.

El coeficiente se calcula aplicando la siguiente fórmula:

$$r = \frac{\frac{1}{n} * \sum((X_i - X_m) * (Y_i - Y_m))}{\sqrt{(\frac{1}{n} * \sum(X_i - X_m)^2) * (\frac{1}{n} * \sum(Y_i - Y_m)^2)}} \tag{4.17}$$

donde el numerador se denomina covarianza y se calcula de la siguiente manera:

$$\widehat{\gamma}^{m_{(k)}}=\frac{1}{N}\sum_{i=1}^{N}\left(z_{(i-1)p+m}-\widehat{\mu}_m\right)\left(z_{(i-1)p+m-k}-\widehat{\mu}_m\right) \tag{4.18}$$

$$\widehat{\rho}^{m_{(k)}}=\frac{\gamma^{m_{(k)}}}{\widehat{\sigma}_m\widehat{\sigma}_{m-k}} \tag{4.19}$$

donde $m = 1, \ldots, p$ y $p$ = numero de periodos, en cada par de valores $(x, y)$ se multiplica el valor de $x$ menos su media, multiplicado por el valor de $y$ menos su media. Se suma el resultado obtenido de todos los pares de valores y este resultado se divide por el tamaño de la muestra.
El denominador se calcula el producto de las varianzas de $x$ y de $y$, y a este producto se le calcula la raíz cuadrada.

### 4.3.4. Formulación del nuevo proceso estocástico

Concatenando las ecuaciones 4.10, 4.5, 4.8 y usando álgebra relacional para la proyección y selección de los casos sobre la base de casos $BC$ indexada sobre la estructura de acceso métrico; se tiene la descripción matemática de la componente estocástica (CE) para el periodo $j$ del modelo de Proceso Estocástico Basado en Razonamiento Basado en Casos. Es la contribución mas importante de esta tesis.

$$Z_{j+1} = \{(\Pi_Z(\sigma_{SM_{pz}^{(w)})\approx 1}(BC)))\} + \alpha(j) \tag{4.20}$$

donde:

- $Z_j$ Es el componente estocástico en el periodo $j$.
- $(\Pi_Z A)$ es la proyección de la salida del mecanismo razonador sobre $(\sigma_{SM_{pz}^{(w)})\approx 1}(BC)$.

- $(\sigma_B)$ es la selección de los casos que cumplan el criterio $SM_{pz}^{(w)}) \approx 1$
- $SM_{pq}^{(w)}) \approx 1$ es la función de similitud del caso buscado $pq$, vea la ecuación 4.10.
- $\alpha(j)$ es un error aleatorio para el instante que proviene de la distribución de probabilidad de la ventana de similitud.
- $BC$ es la base de casos de trabajo, vea la ecuación 4.5.

Extendiendo la expresión se tiene

$$Q_{j+1} = \{(\Pi_Q(\sigma_a(BC)))\} + \alpha(j) \tag{4.21}$$

$$a = \Big(1/\Big(1 + \alpha\Big( \Big[\sum_{j=1}^{n} w_j^2 (x_{pj} - \ x_{qj}\ )^2\Big]^{1/2}\Big)\Big) \approx 1\Big) \tag{4.22}$$

la obtención del $\alpha(j)$ se explica en la sección Adaptación de casos.

## 4.4. Reutilización y adaptación de casos

La adaptación, en el contexto del RBC, se usa para corregir el error de la solución; en la propuesta se estudiara inversamente, es decir para generar un error, esto confiere al proceso la característica estocástica deseada.

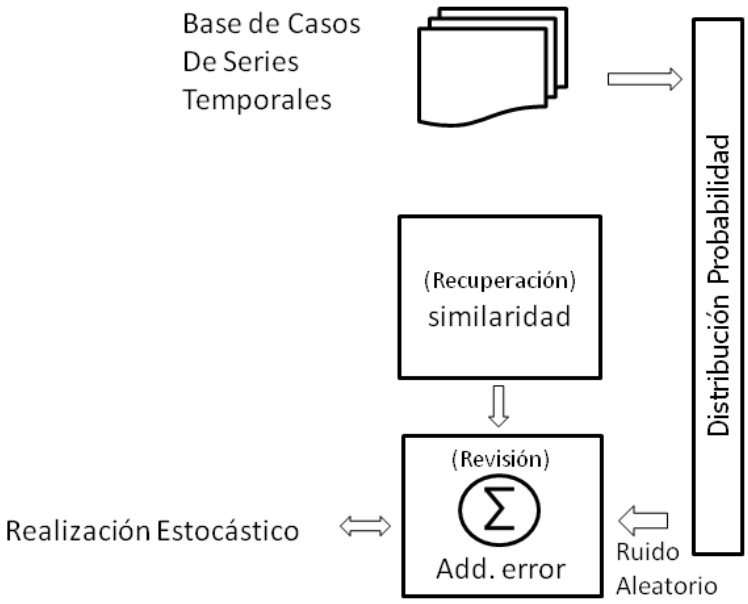



Figura 4.5: Adaptación de casos con error aleatorio

### 4.4.1. Componente aleatorio

La adaptación de casos transforma la salida del razonador en un componente estocástico, basado en la propuesta de (Awchi y cols., 2009) se propone la reutilización del componente aleatorio heredado del modelo de Thomas Fiering; Basado en el trabajo de (Campos, 2010), se sugiere también agregar un error aleatorio, que proviene de una distribución de probabilidad , asociada a las distancias del valor determinístico contra los registros históricos, todo ello bajo un umbral de búsqueda, ambas propuestas son aceptables, se debe estudiar su comportamiento para evaluar su aplicación, registros densos pueden sugerir usar un componente a partir de las distancias.

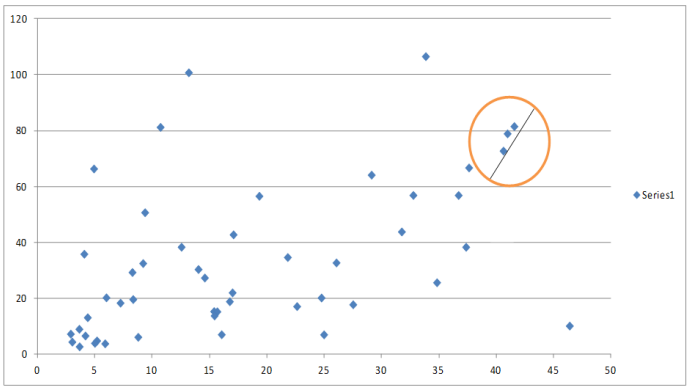



Figura 4.6: Umbral de 10 % para la generación de la distribución de probabilidad

**Umbrales**

El modelo usa un umbral para la generación de la distribución de probabilidad, después que el componente determinístico propone un valor por similitud, se analizan los cercanos bajo el umbral de búsqueda para producir el componente aleatorio, vea en la Figura 4.6, si se usa un umbral de 100 el modelo se comportara similar a Thomas Fiering y la Figura 4.7, el umbral es determinado por la fuerza de la similitud, valores muy similares generarán un umbral pequeño; si la similaridad es cercana a 0 el umbral es el rango.

## 4.5. Retención

### 4.5.1. Encadenamiento de Componentes Estocásticas

Puesto que se propone un Modelo de Proceso Estocástico genérico, es necesario encadenar los términos de las serie que son simulados. La Figura 4.8 muestra de forma genérica el encadenamiento de los componentes estocásticos del proceso estocástico a

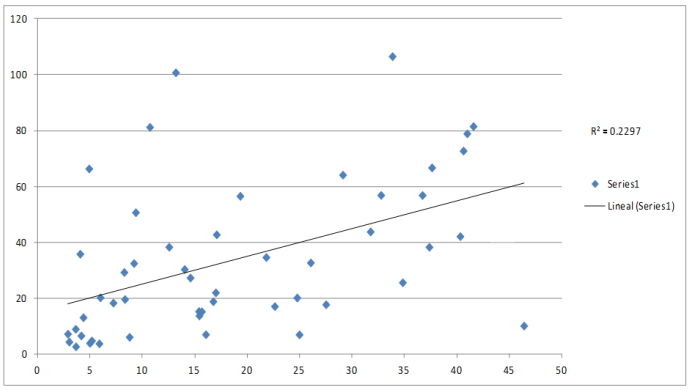



Figura 4.7: Umbral de 100 % para la generación de la distribución de probabilidad

partir de RBC en un determinado tiempo $t$, si el encadenamiento es exitoso se procede a la retención de los valores y la generación del proceso estocástico para todos los periodos.

### 4.5.2. Generación de escenarios

Finalmente, basado en la propuesta de (Campos, 2010) para la generación de escenarios, se concatenan las salidas de los componentes estocásticos de cada periodo, la union de todos estos componentes se le llama «Realización estocástica» o serie temporal generada, vea la Figura 4.9.

## 4.6. Consideraciones Finales

Luego de evaluar los modelos auto regresivos periódicos y ensayar una extensión con RBC, se espera evaluar todas las estrategias abordadas por la técnica para la generación de series temporales en un Proceso Estocástico y recuperar los componentes ocultos, se continuará con el análisis residual para incorporar la componente estocástica formal, y

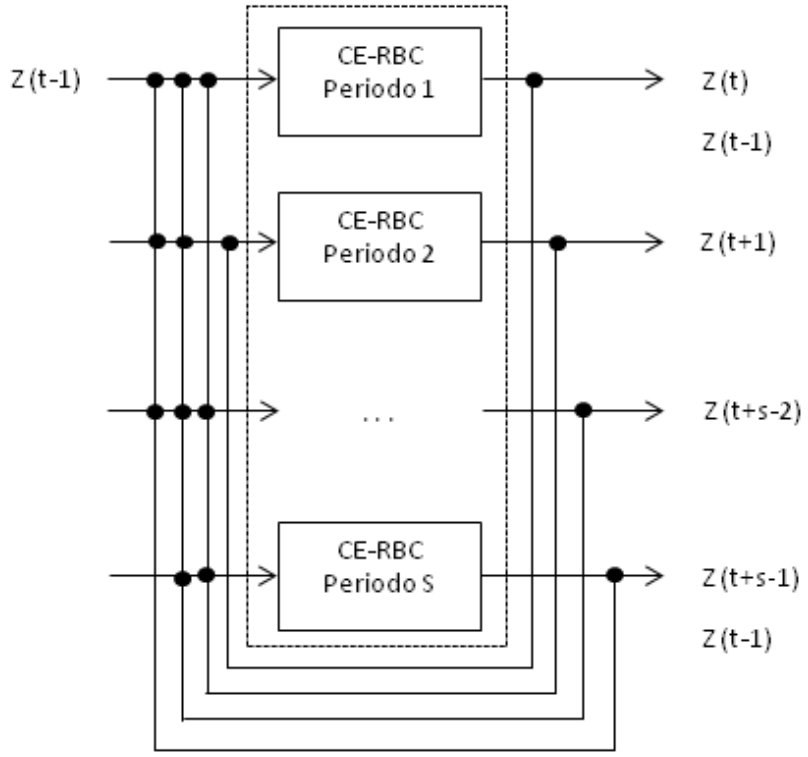



Figura 4.8: Encadenamiento entre las entradas/salidas de las Componentes Estocásticas del Proceso Estocástico a partir de Razonamiento Basado en Casos

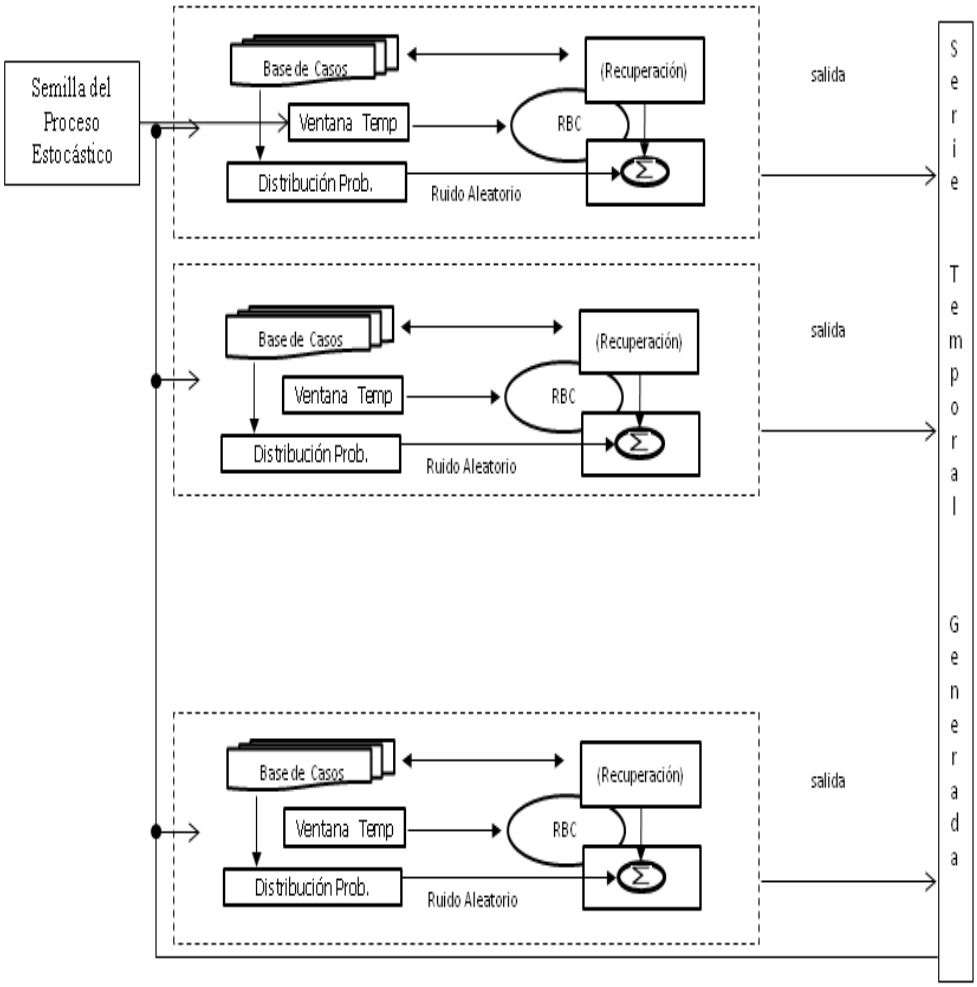



Figura 4.9: Generación de un escenario del Procesos estocástico, a partir de los componentes estocásticos.

la evaluación de los métodos de acceso métrico; a continuación se hará una exploración de otras técnicas para ajustar el modelo propuesto, y finalmente se probará en la generación de caudales sintéticos del caso de estudio.

# Capítulo 5

# Estudio de Caso

---

La evaluación del modelo aplica la generación de variables hidrometeorológicas (Caudales, Evaporación, Precipitación) en la cuenca del Chili, se escogieron tres estaciones de medición: el Pañe, Aguada blanca y el Frayle, se estudiaron periodos mensuales. Los modelos de comparación utilizados son el Modelo de Thomas Fiering y el Modelo Estocástico Neuronal de Luciana. los parámetros utilizados para evaluar a nivel mensual son la media, desviación estándar, el coeficiente de asimetría, máximos y mínimos. A continuación se presenta la caracterización de la cuenca, el contexto de aplicación, los experimentos y finalmente la discusión de los resultados.

## 5.1. Caracterización del área de estudio

La cuenca del río Chili se encuentra ubicada al sur del Perú, y su ámbito está comprendido entre las coordenadas geográficas siguientes:

- $15°37'$ y $16°47'$ de Latitud Sur.
- $70°49'$ y $72°26'$ de Longitud Oeste.

Políticamente, se encuentra en la región de Arequipa, abarcando las provincias de Arequipa, Caylloma y Camaná, y algunos pequeños sectores ubicados en las regiones de Puno, Cusco y Moquegua.
El área de la cuenca, hasta su desembocadura en el Océano Pacífico y sin incluir la sub cuenca del Río Siguas, es de 12,542 $km^2$ . Sus altitudes varían de los 0 a 6,056 $msnm$.

A continuación, se describe la climatología de las zonas geográficas donde se ubican las estaciones de medición tomadas en consideración para realizar las pruebas en esta investigación y las características de éstas (Ver Figura 5.1).

### 5.1.1. Estaciones de medición

#### El Pañe

Ubicada en la sub-cuenca El Pañe, que está localizada en el extremo norte de la cuenca del río Chili, está sobre los 4 585 m.s.n.m. presenta un clima húmedo (tropical). Tiene una extensión de 198 $Km^2$, una precipitación media diaria de 2.21 $mm/d$, la evapotranspiración promedio es de 4 $mm/d$ y el caudal medio diario es de 2.66 $m^3/s$.

La estación El Pañe, cuenta con una estación climatológica y limnimétrica. Realizando mediciones desde 1950, hasta 1964 las descargas naturales de las lagunas de El Pañe. A partir de 1965, hasta la fecha, en que la presa El Pañe entró en funcionamiento, la estación mide las descargas reguladas, con cortos periodos de interrupción a mediados de la década de los 70.

Actualmente, la estación llamada también Oscollo, que es operada por AUTODEMA, está ubicada en el inicio del canal de derivación Pañe-Bamputañe, aproximadamente a unos 100 $m$ de la presa. La sección del canal en este lugar es rectangular, con

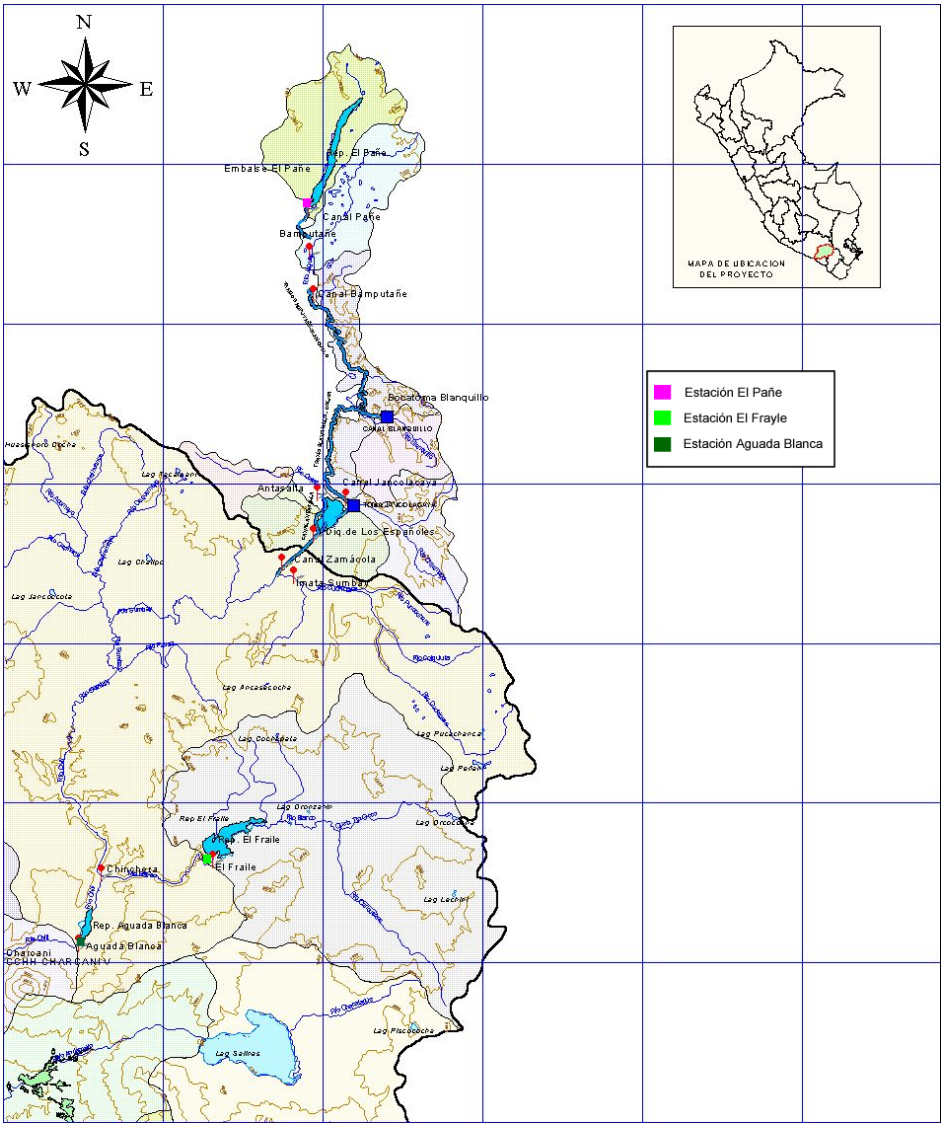



Figura 5.1: Localización de las estaciones de medición consideradas para la investigación.

paredes de concreto de 2.00 $m$ de alto y piso de concreto; su ancho es de 2.70 $m$ y tiene una mira de 2.00 $m$ de alto, ubicada en su margen izquierda (Oviedo T., Umeres R., Franco R., Vílchez, y Butrón, 2001) (Oviedo Tejada, 2004).

**Estación El Frayle**

Ubicada en la sub-cuenca El Fraile, que abarca desde el nacimiento de los ríos Yamanayo, Collpamayo, Paltimayo, Cancusane, Pasto Grande (entre otros ríos menores); hasta el río Blanco (que nace de la confluencia de los ríos ya mencionados) presentando un área de drenaje de 1041 Km2 y finaliza en el embalse El Fraile ubicado sobre el río Blanco a una altitud media de 4000 m.s.n.m., regulando los recursos hídricos. Teniendo una precipitación media anual de 386 $mm$, un caudal medio anual de 3.32 $m^3/s$

La estación El Frayle, cuenta con una estación climatológica y limnimétrica. Realizó mediciones durante desde 1953 hasta 1957 de las descargas naturales de El Frayle, luego, dejó de operar, y desde 1964 hasta la fecha, mide las descargas reguladas del reservorio El Frayle, cuya construcción finalizó en 1959 y entró en funcionamiento en 1964. Esta estación de aforos, mide las descargas reguladas por el embalse El Frayle y se encuentra ubicada en el cauce del río Blanco, aproximadamente a unos 50.00 $m$ aguas abajo, del lugar en que ingresan, las filtraciones se ocurren en la represa lateral conocida como Dique de Bloques (Oviedo T. y cols., 2001) (Oviedo Tejada, 2004).

**Estación Aguada Blanca**

Ubicada en la subcuenca mismo nombre, que presenta una climatología semiárida. La estación Aguada Blanca, hasta antes de 1989 medía las descargas reguladas y no reguladas del embalse Aguada Blanca. Desde 1989, las descargas reguladas del embalse se miden en la Central Hidroeléctrica de Charcani V. Desde 1989, la estación mide la

suma de derrames que se producen en el aliviadero Morning Glory y las descargas que se efectúan por la compuerta de regulación.

Consecuentemente, desde 1989, las salidas totales del embalse Aguada Blanca, son la suma de lo que mide la estación Aguada Blanca (ó mas precisamente, la estimación que se hace de las salidas por la compuerta de regulación, y los caudales que se obtienen del limnígrafo ubicado en la cresta del vertedero) mas el caudal turbinado por la Central Hidroeléctrica. Cuenta con una estación climatológica y limnimétrica (Oviedo T. y cols., 2001) (Oviedo Tejada, 2004).

## 5.2. Contexto del caso de estudio

La generación de series temporales se da en el contexto de una arquitectura para la planificación de Recursos Hídricos, vinculada a un Sistema de Soporte de Decisiones, las salidas del Nuevo Proceso Estocástico son probadas en el Generador de escenarios.

### 5.2.1. Generador de escenarios

Dentro del caso de estudio se enmarca el generador de escenarios de series temporales (GST), este permite encontrar posibles series de datos (precipitación y evaporación) para simulaciones a futuro, que permitirán proyectar posibles escenarios de condiciones climáticas y de demanda de agua; son usadas técnicas matemáticas (estadísticas, estocásticas), inteligentes (redes neuronales) y complementariamente la propuesta para la generación de estos registros.
Se puede ver el flujo de datos en la Figura 5.2, Allí se toman los registros históricos de la cuenca, luego son almacenados en la base de datos y mediante el uso de modelos matemáticos (estadísticos o estocásticos), inteligentes (redes neuronales), y la propuesta; se generan datos sintéticos, que luego son almacenados en una base de datos:

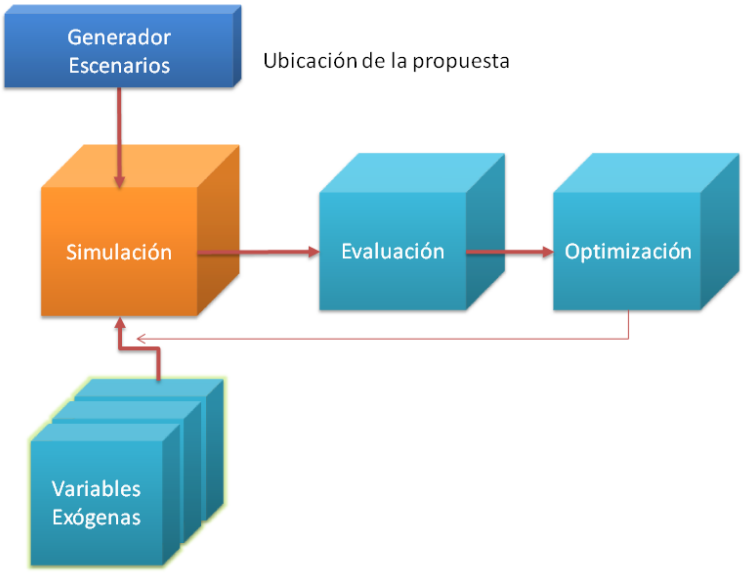



Figura 5.2: Arquitectura del sistema de planificación que incluye la generación estocástica de escenarios

"*BD Series Generadas*", que pueden ser usados para la generación de los diferentes escenarios climatológicos.

## 5.3. Formulación del RBC

Para el casos de estudio se debe formular la entidad **caso**, para ello se relaciona los atributos precipitación, evaporación y caudales de una estación, vea la Figura 5.3: Registro de Caso Serie Temporal.

Se propone un diseño de esquema para la Base de Casos:

$$e = \{T, XY, E, E_1, E_2, Q, Q_1, Q_2, P, P_1, P_2\} \tag{5.1}$$

donde:

$e$ : es el esquema de los casos

$T$ : es la referencia temporal para mes=$(modulo(RT, 12))$ y año=$(RT)$

$XY$ : es la geo-referencia del dato registrado

$E$ : Evaporación

$E_1$ : Evaporación con un retraso

$E_2$ : Evaporación con dos retrasos

$Q$ : Caudal

$Q_1$ : Caudal con un retraso

$Q_2$ : Caudal con dos retrasos

$P$ : Precipitación

$P_1$ : Precipitación con un retraso

$P_2$ : Precipitación con dos retrasos

El orden del razonador es 2. las dimensiones son 5, se debe resaltar que si las dimensiones tienden a infinito, el umbral de búsqueda sera cercano a 1 y el modelo se convertirá en determinístico, pudiendo ser usado en tareas de pronóstico.

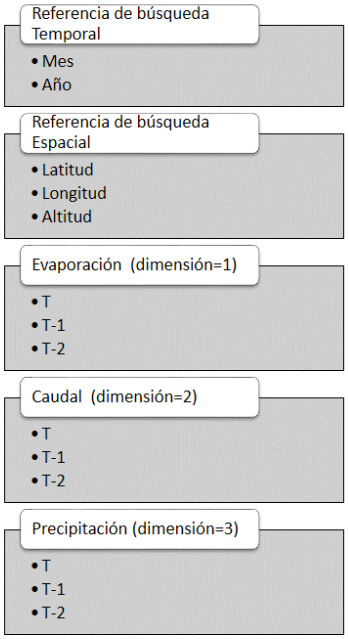



Figura 5.3: Registro de Caso Serie Temporal

## 5.4. Experimentos

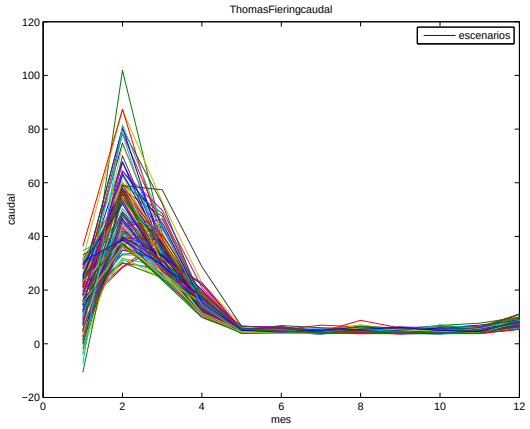
### 5.4.1. Procesos Estocástico de Thomas-Fiering

Los valores generados por el modelo estocástico Neuronal de TF 3.1 corresponden a las variables hidrometeorológicas: Caudales, Evaporación y Precipitación, el área de estudio es la cuenca del rio Chili, las estaciones de medición son El Pañe, Aguada blanca y el Frayle, se generan 100 realizaciones en periodos mensuales, el año de pronostico es el el año 2000, finalmente los registros históricos corresponden al periodo de 1970 a 1999.

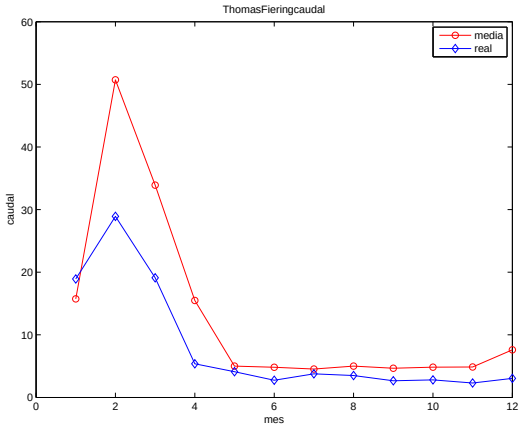
En la figura 5.4 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca de Aguada Blanca.

En la figura 5.5 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca del Frayle.

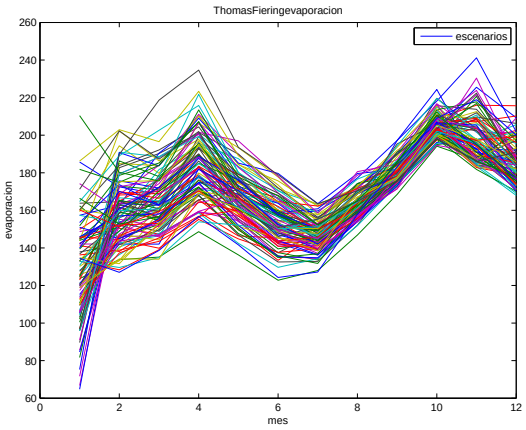
En la figura 5.6 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca del Pañe.



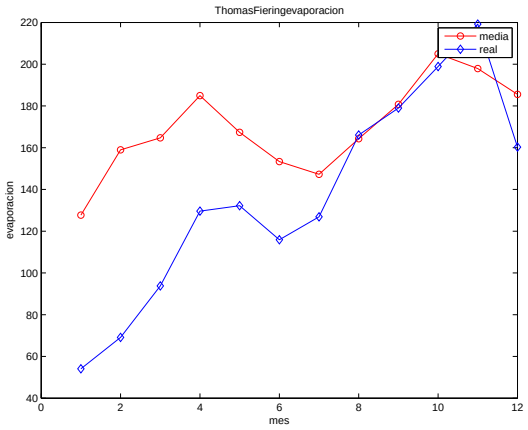

(a) Series temporales de caudal



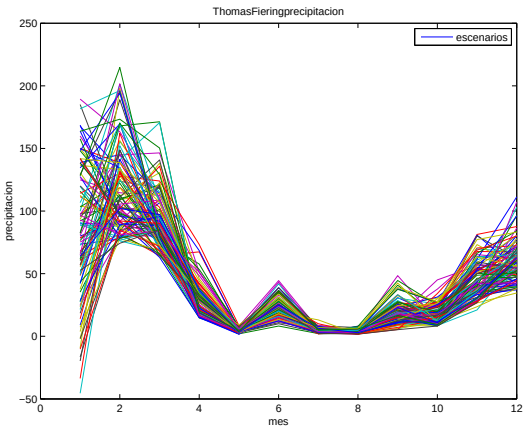

(b) Series temporales de caudal media



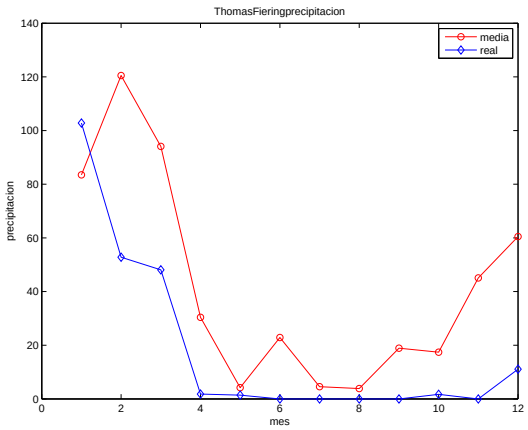

(c) Series temporales de evaporación




(d) Series temporales de evaporación media

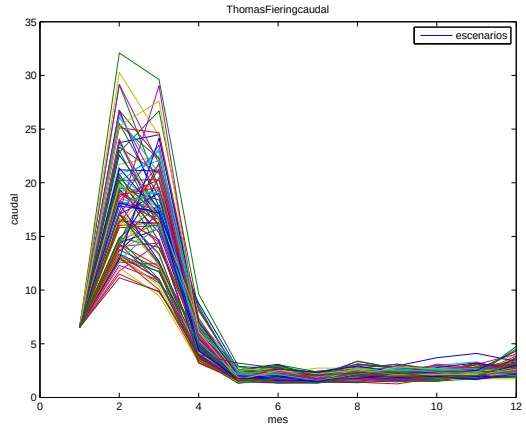



(e) Series temporales de precipitación

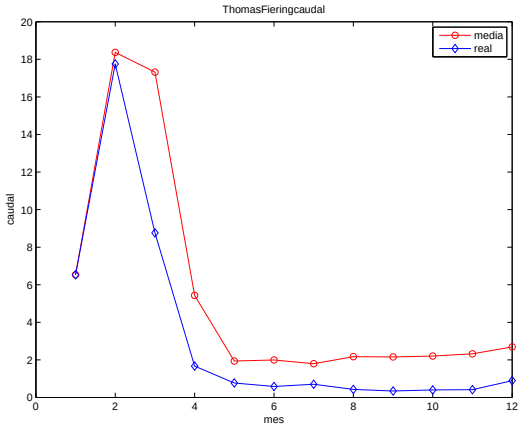



(f) Series temporales de precipitación media

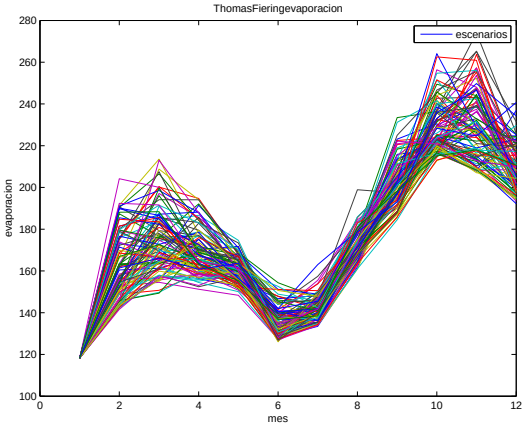
Figura 5.4: Series generadas por el modelo Thomas Fiering, data histórica de Aguada Blanca : años 1970-1999, data sintetizada: 2000.



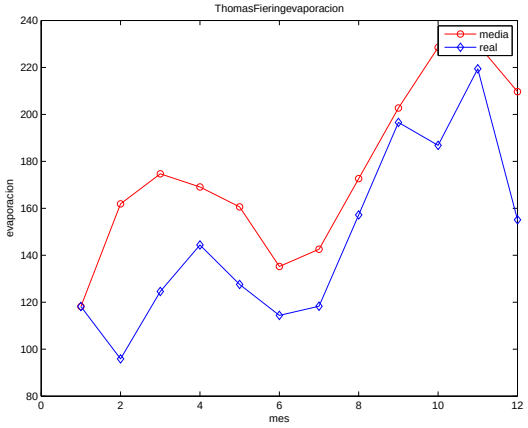

(a) Series temporales de caudal



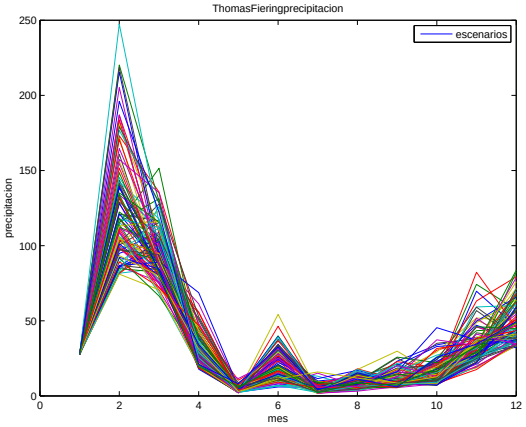

(b) Series temporales de caudal media



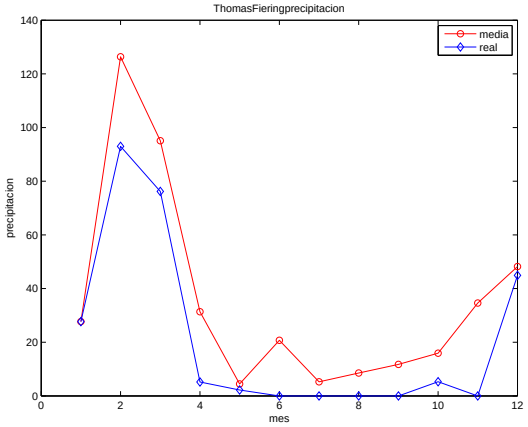

(c) Series temporales de evaporación




(d) Series temporales de evaporación media

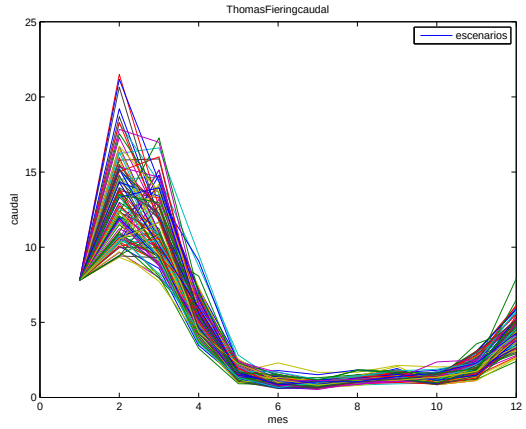



(e) Series temporales de precipitación

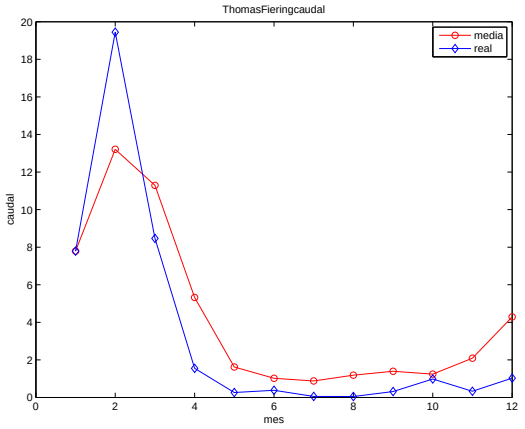



(f) Series temporales de precipitación media

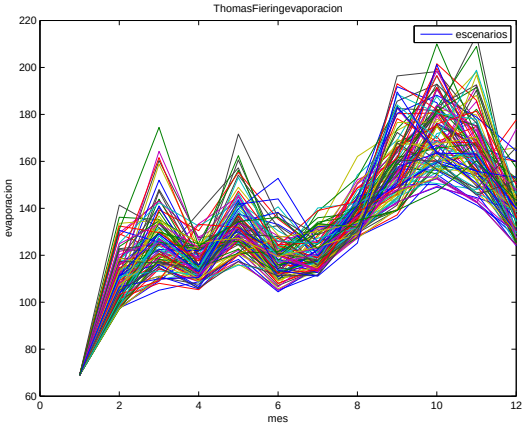
Figura 5.5: Series generadas por el modelo Thomas Fiering, data histórica del Frayle : años 1970-1999, data sintetizada: 2000.



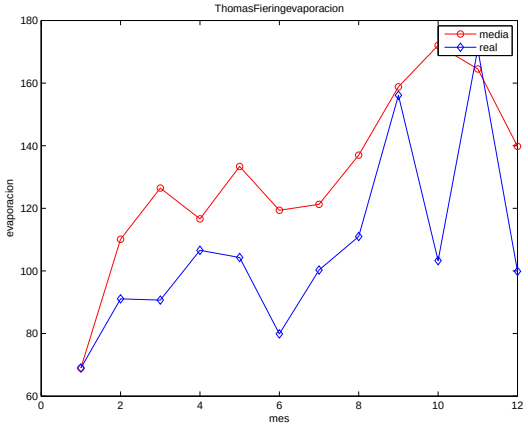

(a) Series temporales de caudal



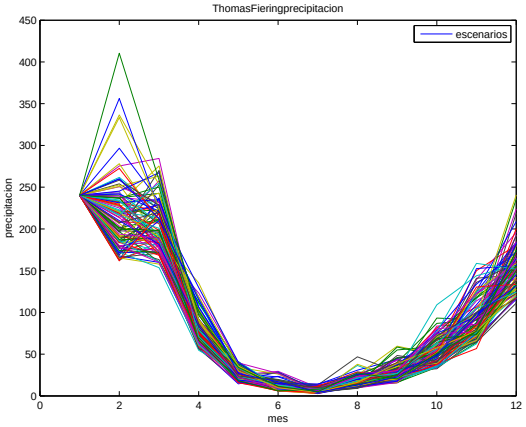

(b) Series temporales de caudal media



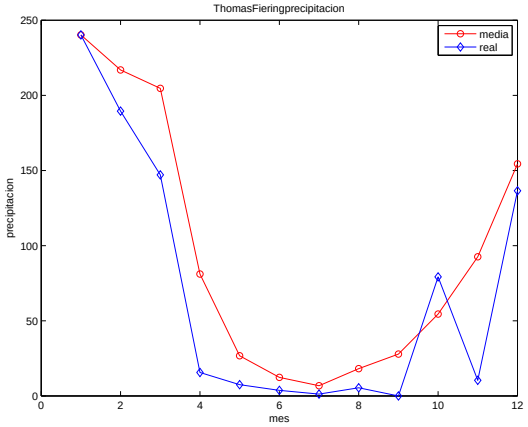

(c) Series temporales de evaporación




(d) Series temporales de evaporación media




(e) Series temporales de precipitación




(f) Series temporales de precipitación media

Figura 5.6: Series generadas por el modelo Thomas Fiering, data histórica del Pañe : años 1970-1999, data sintetizada: 2000.

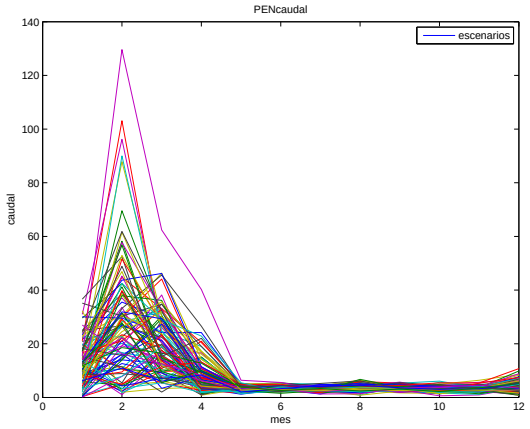
### 5.4.2. Proceso Estocástico Neuronal (PEN)

Los valores generados por el modelo estocástico Neuronal (PEN) de Luciana (Campos, 2010) corresponden a las variables hidrometeorológicas: Caudales, Evaporación y Precipitación, el área de estudio es la cuenca del rio Chili, las estaciones de medición son El Pañe, Aguada blanca y el Frayle, se generan 100 realizaciones en periodos mensuales, el año de pronostico es el el año 2000, finalmente los registros históricos corresponden al periodo de 1970 a 1999.

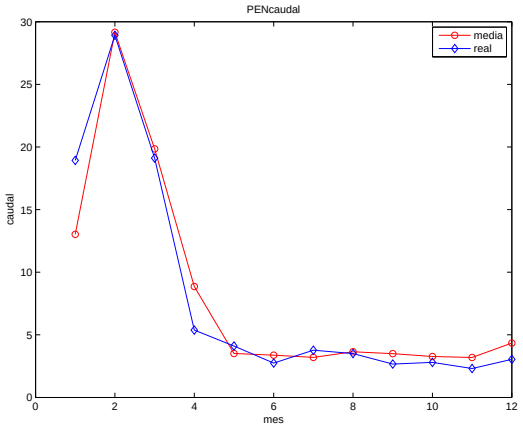
En la figura 5.7 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca de Aguada Blanca.

En la figura 5.8 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca del Frayle.

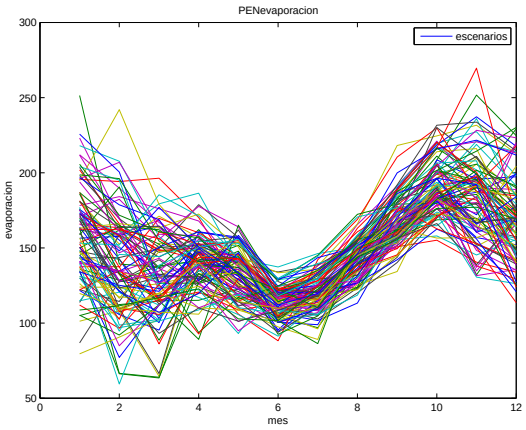
En la figura 5.9 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca del Pañe.



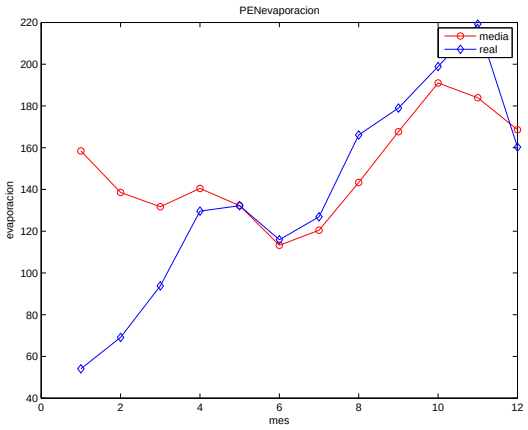

(a) Series temporales de caudal



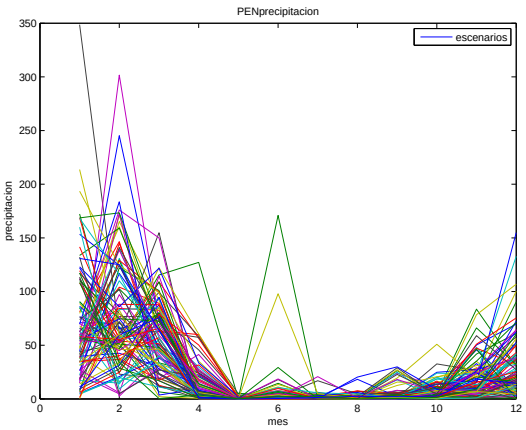

(b) Series temporales de caudal media



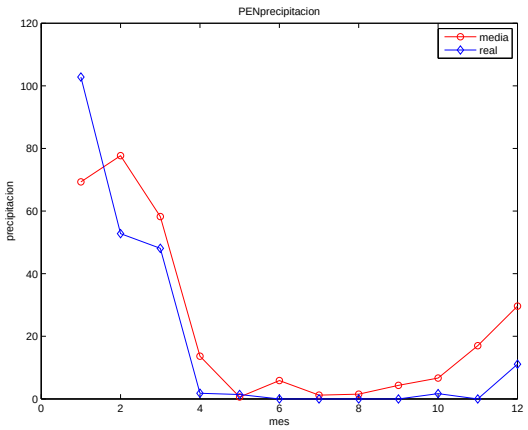

(c) Series temporales de evaporación




(d) Series temporales de evaporación media

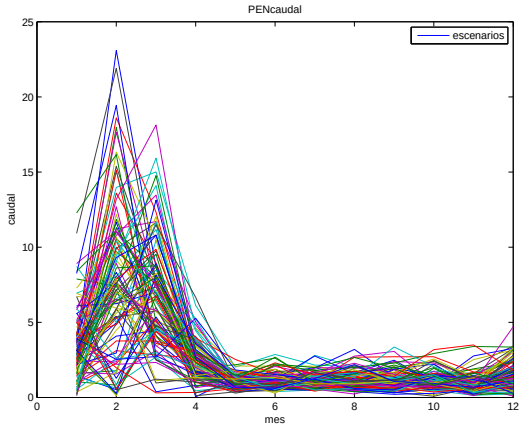



(e) Series temporales de precipitación

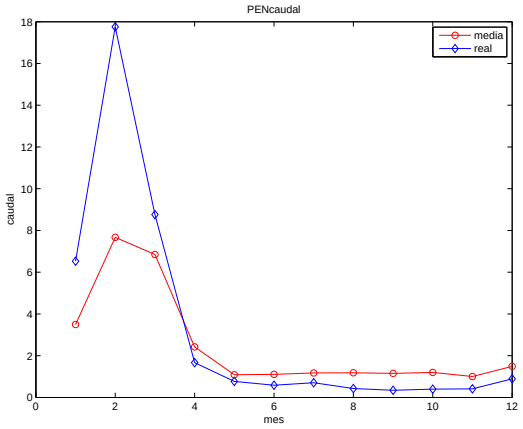



(f) Series temporales de precipitación media

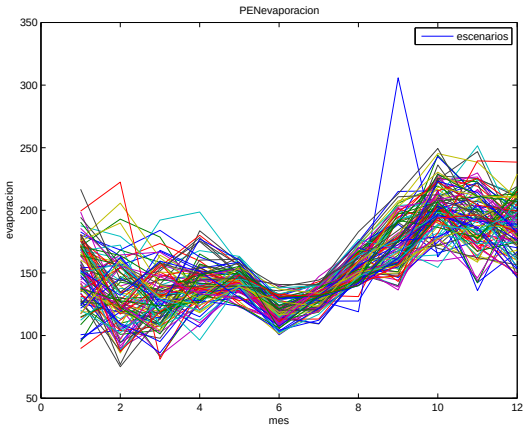
Figura 5.7: Series generadas por el modelo PEN, data histórica de Aguada Blanca: años 1970-1999, data sintetizada: 2000.



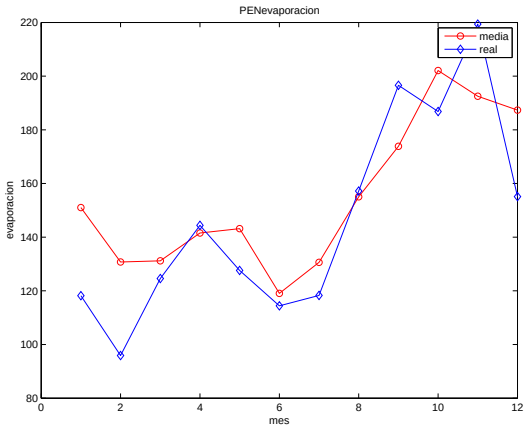

(a) Series temporales de caudal



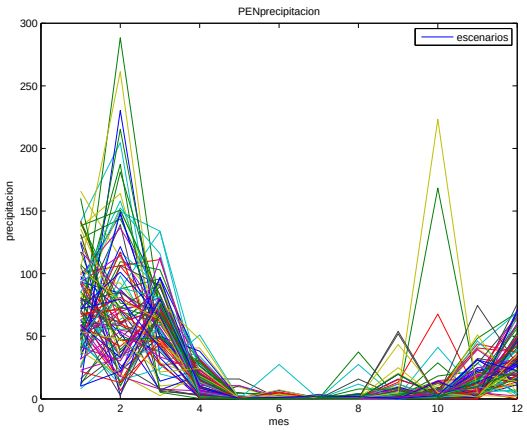

(b) Series temporales de caudal media



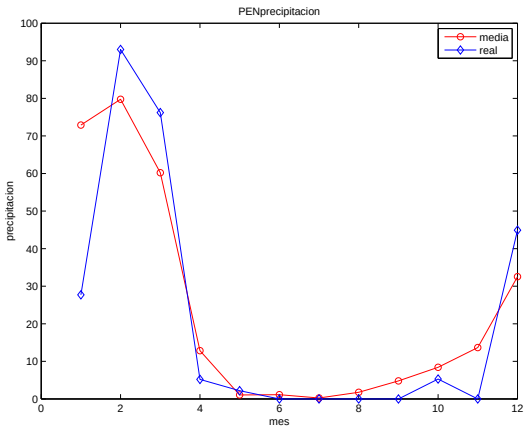

(c) Series temporales de evaporación




(d) Series temporales de evaporación media

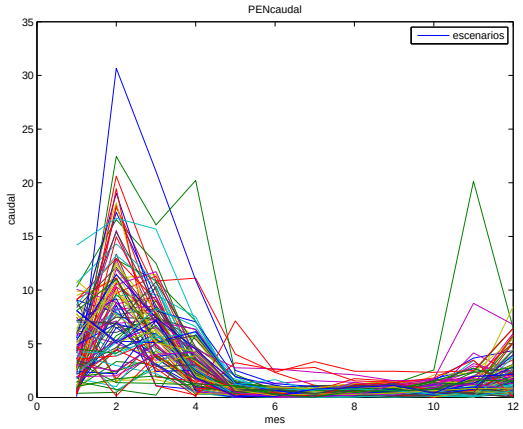



(e) Series temporales de precipitación

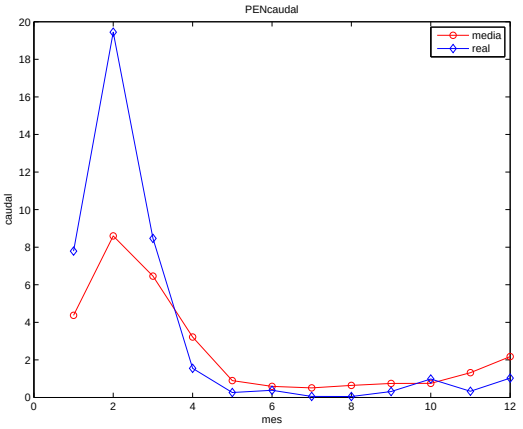



(f) Series temporales de precipitación media

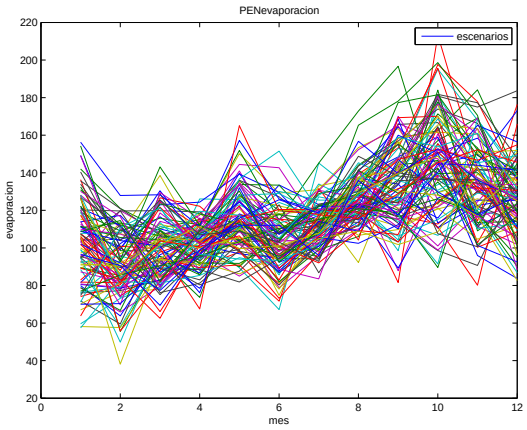
Figura 5.8: Series generadas por el modelo PEN, data histórica del Frayle: años 1970-1999, data sintetizada: 2000.



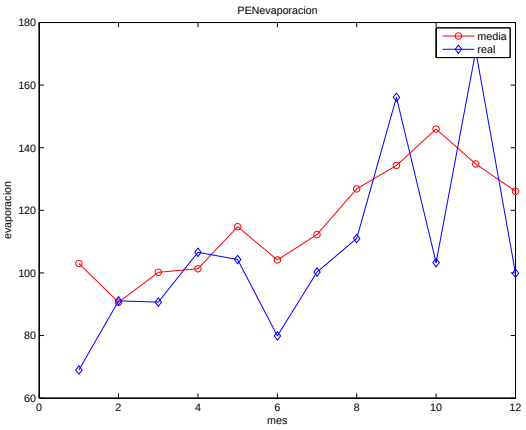

(a) Series temporales de caudal



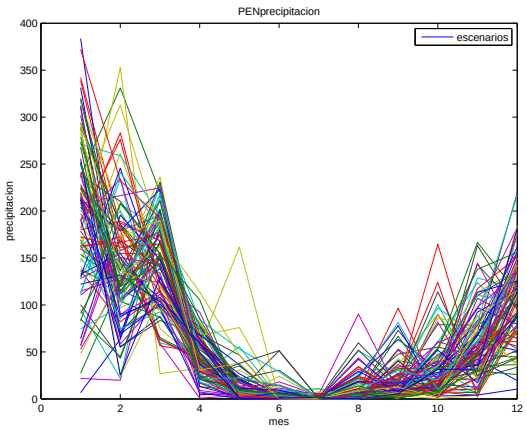

(b) Series temporales de caudal media



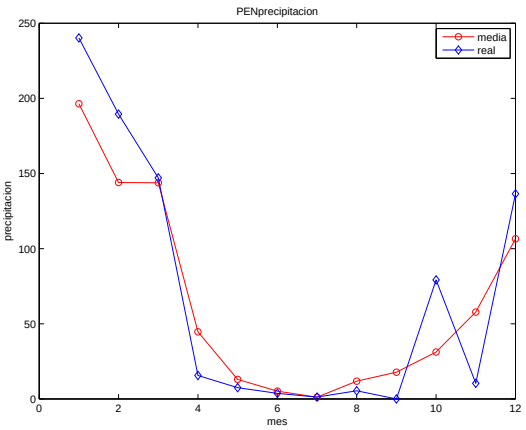

(c) Series temporales de evaporación




(d) Series temporales de evaporación media




(e) Series temporales de precipitación




(f) Series temporales de precipitación media

Figura 5.9: Series generadas por el modelo PEN, data histórica del Pañe: años 1970-1999, data sintetizada: 2000.

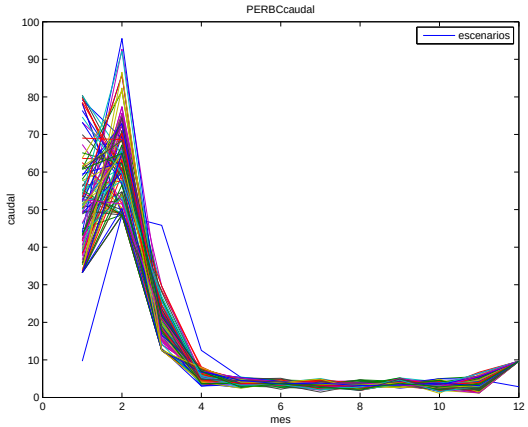
### 5.4.3. Proceso Estocástico a partir de Razonamiento Basado en Casos

Los valores generados por la propuesta, el modelo estocástico a partir de Razonamiento Basado en Casos, corresponden a las variables hidrometeorológicas: Caudales, Evaporación y Precipitación, el área de estudio es la cuenca del rio Chili, las estaciones de medición son El Pañe, Aguada blanca y el Frayle, se generan 100 realizaciones en periodos mensuales, el año de pronostico es el el año 2000, finalmente los registros históricos corresponden al periodo de 1970 a 1999.

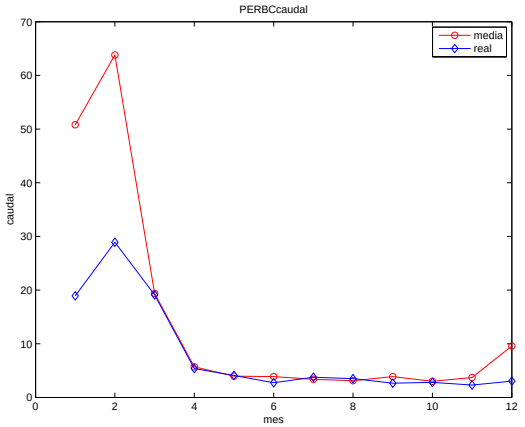
En la figura 5.10 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca de Aguada Blanca.

En la figura 5.11 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca del Frayle.

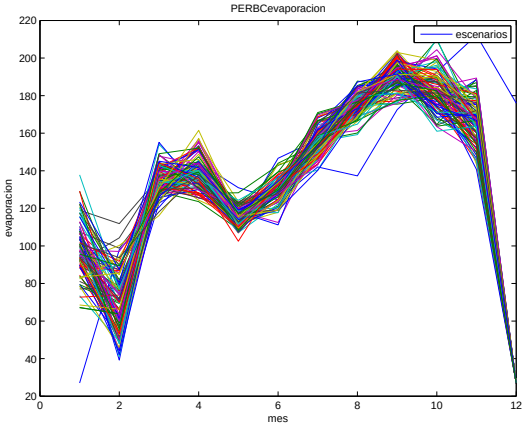
En la figura 5.12 muestra los caudales, precipitaciones, evaporaciones y la comparación de las medias de los datos observados de la subcuenca del Pañe.



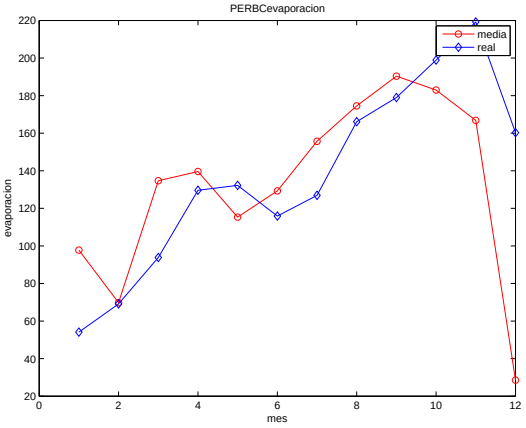

(a) Series temporales de caudal



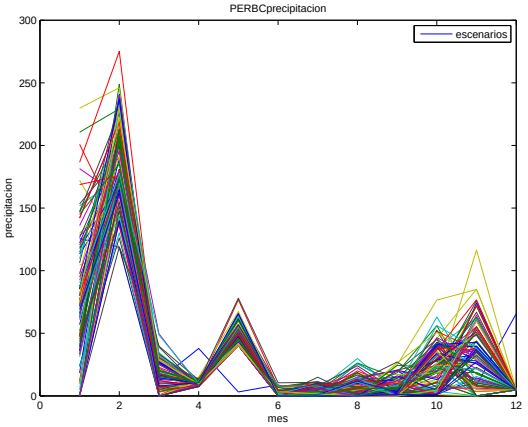

(b) Series temporales de caudal media



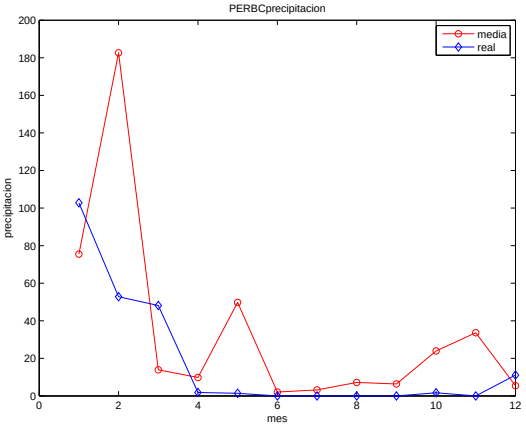

(c) Series temporales de evaporación




(d) Series temporales de evaporación media

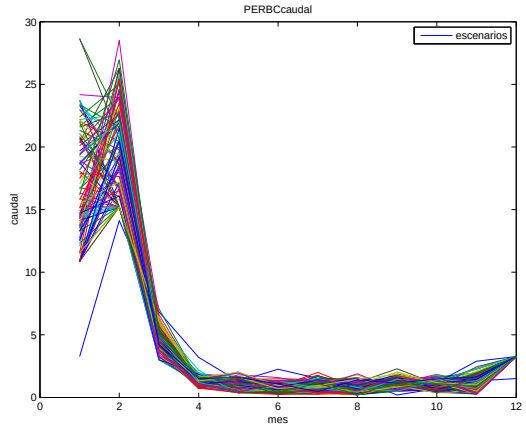



(e) Series temporales de precipitación

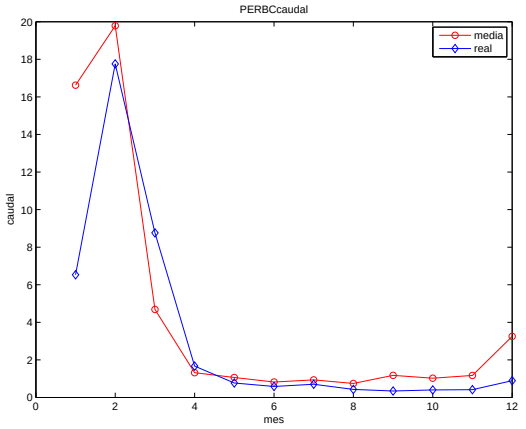



(f) Series temporales de precipitación media

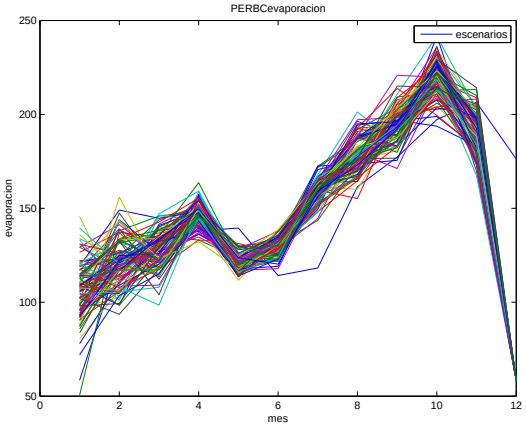
Figura 5.10: Series generadas por el modelo PERBC, data histórica de Aguada Blanca : años 1970-1999, data sintetizada: 2000.



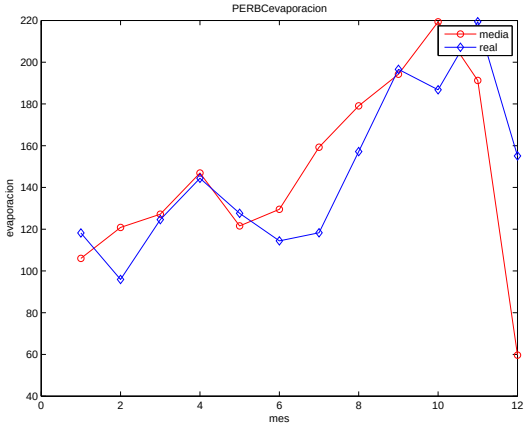

(a) Series temporales de caudal



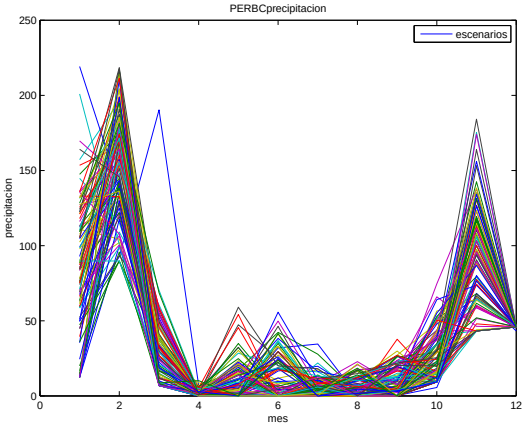

(b) Series temporales de caudal media



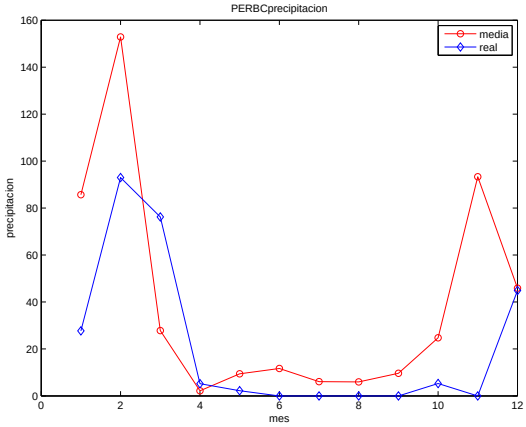

(c) Series temporales de evaporación




(d) Series temporales de evaporación media

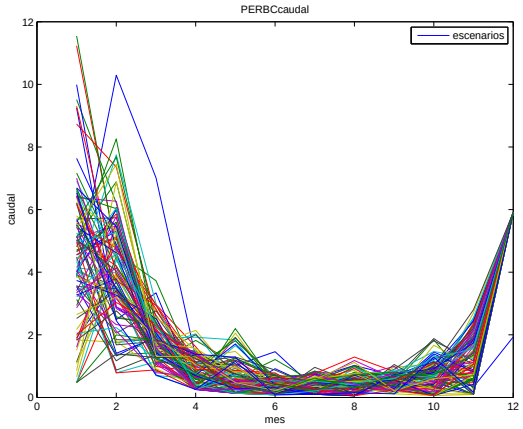



(e) Series temporales de precipitación

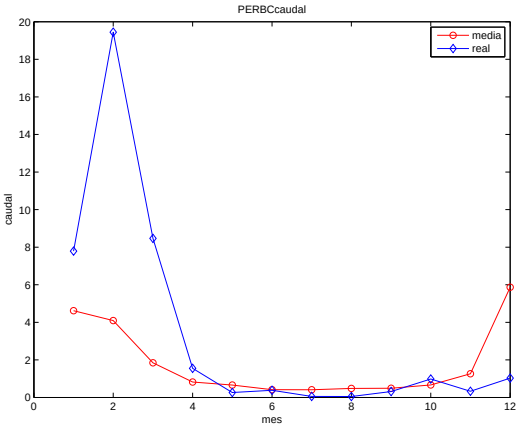



(f) Series temporales de precipitación media

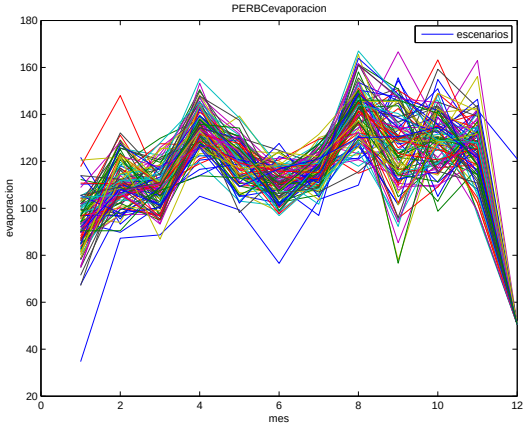
Figura 5.11: Series generadas por el modelo PERBC, data histórica del Frayle : años 1970-1999, data sintetizada: 2000.



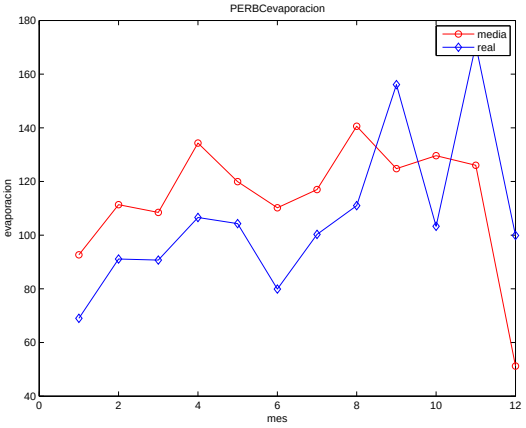

(a) Series temporales de caudal



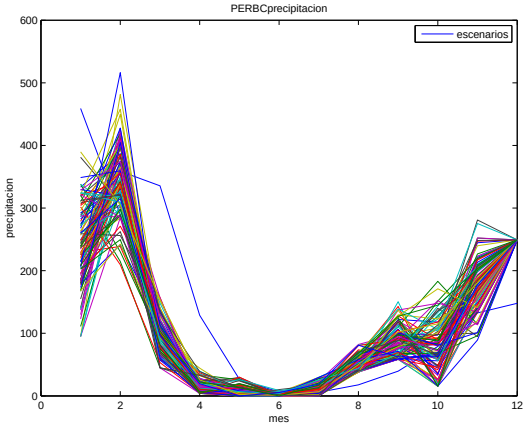

(b) Series temporales de caudal media



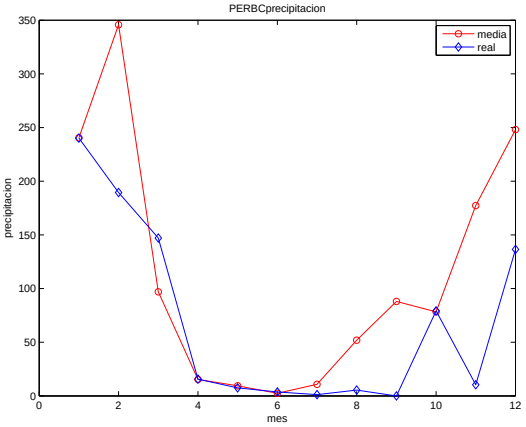

(c) Series temporales de evaporación




(d) Series temporales de evaporación media




(e) Series temporales de precipitación




(f) Series temporales de precipitación media

Figura 5.12: Series generadas por el modelo PERBC, data histórica del Pañe : años 1970-1999, data sintetizada: 2000.

## 5.5. Analisis de resultados

### 5.5.1. Estimadores de primer orden

Un análisis detallado de la media: Cuadro 5.1, desviación estándar: Cuadro 5.2 y la asimetría: Cuadro 5.3 para todos los experimentos de los modelos TF, PEN y PERBC muestran que se conservan satisfactoriamente las características de la serie histórica, sin embargo se ven generaciones leptocúrticas para el modelo PERBC (el propuesto) respecto a sus similares, incluso la serie histórica, esto se debe por las multidimensionalidad de la propuesta, el modelo ajusta los pronósticos y reduce la incertidumbre, una propiedad del RBC (Pal y Shiu, 2004; Loucks y cols., 2005).

| | | **Media** | | | |
|---|---|---|---|---|---|
| | | **Hist** | **TF** | **PEN** | **PERBC** |
| ***Estacion*** | ***Variable*** | | | | |
| *Pañe* | *Caudal* | 2.6698 | 4.2776 | 2.5228 | 1.8015 |
| | *Evaporación* | 115.5414 | 130.6950 | 116.2029 | 113.8477 |
| | *Precipitación* | 62.9222 | 94.6741 | 64.4482 | 113.7261 |
| *Frayle* | *Caudal* | 2.9951 | 5.4106 | 2.4860 | 4.3828 |
| | *Evaporación* | 161.0736 | 175.4696 | 154.8506 | 146.2414 |
| | *Precipitación* | 25.1347 | 35.8219 | 24.1087 | 39.6004 |
| *Aguada Blanca* | *Caudal* | 7.7259 | 12.9597 | 8.2448 | 14.5204 |
| | *Evaporación* | 144.8400 | 172.4811 | 149.1352 | 132.1097 |
| | *Precipitación* | 23.4403 | 42.1531 | 23.8099 | 34.4451 |

Cuadro 5.1: Comparación anualizada de **Medias** para el Caudal, Evaporación, Precipitación de la serie Histórica (Hist), el modelo de Thomas Fiering (TF) el modelo Estocástico Neuronal (PEN) y la propuesta (PERBC)

| | | Desviación Estándar | | | |
|---|---|---|---|---|---|
| | | **Hist** | **TF** | **PEN** | **PERBC** |
| ***Estacion*** | ***Variable*** | | | | |
| *Pañe* | *Caudal* | 1.2905 | 0.7628 | 1.5330 | 0.3990 |
| | *Evaporación* | 5.6919 | 3.8721 | 5.9567 | 2.1740 |
| | *Precipitación* | 19.7444 | 8.4634 | 15.7573 | 11.8989 |
| *Frayle* | *Caudal* | 2.2584 | 1.5651 | 1.2745 | 1.1454 |
| | *Evaporación* | 4.9789 | 3.8990 | 6.3512 | 2.5545 |
| | *Precipitación* | 10.8996 | 9.1974 | 13.3485 | 8.8962 |
| *Aguada Blanca* | *Caudal* | 6.7306 | 3.2589 | 6.1120 | 3.1942 |
| | *Evaporación* | 6.4658 | 5.5242 | 7.8475 | 2.5269 |
| | *Precipitación* | 14.4746 | 9.6855 | 15.3894 | 10.2891 |

Cuadro 5.2: Comparación anualizada de la **Desviación Estándar** para el Caudal, Evaporación, Precipitación de la serie Histórica (Hist), el modelo de Thomas Fiering (TF) el modelo Estocástico Neuronal (PEN) y la propuesta (PERBC)

| | | Asimetría | | | |
|---|---|---|---|---|---|
| | | **Hist** | **TF** | **PEN** | **PERBC** |
| ***Estacion*** | ***Variable*** | | | | |
| *Pañe* | *Caudal* | 0.1036 | 0.3775 | 0.2683 | 0.1818 |
| | *Evaporación* | 0.3662 | -0.1889 | -0.1015 | 0.116 |
| | *Precipitación* | 1.8423 | 0.1232 | 0.1510 | 0.5295 |
| *Frayle* | *Caudal* | -0.7527 | 0.1723 | -0.0386 | -0.0838 |
| | *Evaporación* | 0.2947 | -0.0942 | 0.2497 | -0.0108 |
| | *Precipitación* | 0.0029 | 0.2603 | 0.5279 | 0.3729 |
| *Aguada Blanca* | *Caudal* | -0.3793 | 0.0018 | 0.0449 | 0.8721 |
| | *Evaporación* | -1.1077 | -1.2389 | 0.0209 | -0.4597 |
| | *Precipitación* | 0.1656 | 0.5924 | 0.3319 | -0.3132 |

Cuadro 5.3: Comparación anualizada de la **Asimetría** para el Caudal, Evaporación, Precipitación de la serie Histórica (Hist), el modelo de Thomas Fiering (TF) el modelo Estocástico Neuronal (PEN) y la propuesta (PERBC)

### 5.5.2. Máximos y mínimos

Los eventos máximos y mínimos fueron reproducidos satisfactoriamente, PERBC genera mínimos extremos; el Cuadro 5.5.2 muestra el comportamiento de los mínimos sobre las precipitaciones del Pañe, Frayle y Aguada Blanca observándose que el modelo PERBC consigue generar mínimos 0, lo cual representan a la serie histórica (TF y PEN tienen valores aproximados), esto permite inferir un buen desempeño para generar series que contemplen sequías del modelo PERBC sobre el TF y el PEN.

### 5.5.3. MSE y RMSE

El Error Medio Cuadrático (MSE) y la Raíz del Error Medio Cuadrático (RMSE), permiten una comparación sobre el error medio de las las generaciones sobre el valor observado. Luego de analizar el MSE y el RMSE para (TF), el Proceso Estocástico Neuronal (PEN) y el Proceso Estocástico Basado en Casos (PERBC) se puede observar en los Cuadros 5.5, 5.6 que todos los modelos son malos predictores, esto se debe directamente al componente aleatorio agregado; sin embargo, en varios casos, el PERBC presenta una ligera ventaja sobre los otros (vea caudal y precipitación en el Pañe, y todas en Aguada Blanca), esto se debe a su naturaleza multidimensional que finalmente genera series temporales leptocúrticas; el PEN también tiene ventajas sobre TF.

| | | | Maximos | | | | Minimos | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Hist** | **TF** | **PEN** | **PERBC** | **Hist** | **TF** | **PEN** | **PERBC** |
| ***Estacion*** | ***Variable*** | | | | | | | | |
| *Pañe* | *Caudal* | 20.0480 | 21.5023 | 30.6604 | 11.5438 | 0.0030 | 0.5233 | 0.0016 | 0.0550 |
| | *Evaporación* | 195.0000 | 213.7827 | 213.3103 | 166.9922 | 55.0000 | 69.0000 | 38.1141 | 34.7000 |
| | *Precipitación* | *331.3000* | *410.4703* | *383.6119* | *516.2470* | 0 | 2.7337 | 0.0010 | 0 |
| *Frayle* | *Caudal* | 32.2200 | 32.0925 | 23.0862 | 28.6761 | 0.0390 | 1.2670 | 0.0381 | 0.1950 |
| | *Evaporación* | 246.0000 | 274.0724 | 305.8154 | 241.7700 | 87.5000 | 118.2000 | 75.1152 | 50.5215 |
| | *Precipitación* | 210.4000 | 247.4411 | 288.4087 | 219.2043 | 0 | 1.6128 | 0.0010 | 0 |
| *Aguada Blanca* | *Caudal* | 105.1480 | 91.2014 | 129.6086 | 95.5330 | 1.3410 | -6.8586 | 0.1044 | 1.0900 |
| | *Evaporación* | 240.0000 | 244.1905 | 269.6172 | 212.2295 | 72.0000 | 60.8776 | 59.4493 | 27.0500 |
| | *Precipitación* | 240.3000 | 253.4967 | 348.6200 | 275.0995 | 0 | -39.806 | 0.0013 | 0 |

Cuadro 5.4: Comparación anualizada de los **Máximos y mínimos** para el Caudal, Evaporación, Precipitación de la serie Histórica (Hist), el modelo de Thomas Fiering (TF) el modelo Estocástico Neuronal (PEN) y la propuesta (PERBC)

| | | TF | PEN | PERBC |
|---|---|---|---|---|
| Estacion | Variable | | | |
| Pañe | Caudal | 6.6861 | 11.6363 | 6.2 |
| | Evaporación | 968.6778 | 554.3066 | 869.1 |
| | Precipitación | 1453.9 | 889.9396 | 643.1 |
| Frayle | Caudal | 9.1 | 9.8539 | 10.9 |
| | Evaporación | 1224.1 | 439.8503 | 1180.3 |
| | Precipitación | 345.6 | 242.5547 | 561.7 |
| Aguada Blanca | Caudal | 62.2868 | 4.3 | 19.2 |
| | Evaporación | 2412.4 | 1611.3 | 2127.7 |
| | Precipitación | 1209 | 224.9 | 811.9 |

Cuadro 5.5: Error Medio Cuadrático

| | | TF | PEN | PERBC |
|---|---|---|---|---|
| Estacion | Variable | | | |
| Pañe | Caudal | 2.5857 | 3.4112 | 2.4900 |
| | Evaporación | 31.1236 | 23.5437 | 29.4805 |
| | Precipitación | 38.1295 | 29.8319 | 25.3594 |
| Frayle | Caudal | 3.0118 | 3.1391 | 3.3015 |
| | Evaporación | 34.9867 | 20.9726 | 34.3555 |
| | Precipitación | 18.5898 | 15.5742 | 23.7002 |
| Aguada Blanca | Caudal | 7.8922 | 2.0821 | 4.3818 |
| | Evaporación | 49.1167 | 40.1414 | 46.1270 |
| | Precipitación | 34.7712 | 14.9969 | 28.4939 |

Cuadro 5.6: Raíz del Error Medio Cuadrático

# Capítulo 6

# Conclusiones y trabajo futuro

---

## 6.1. General

El uso del Razonamiento Basado en Casos para la formulación de un nuevo modelo de Proceso Estocástico para la generación de series temporales, genera razonablemente realizaciones que muestran información que TF y PEN aproximan, particularmente para el caso de valores mínimos extremos, Luego el uso de casos multidimensionales y de grados superiores genera series leptocúrticas, lo que en ciertos casos no reproduce las características de la serie histórica, pero que reduce la incertidumbre. Computacionalmente una estructura de datos de acceso secuencial permite la indexación en memoria de todos los casos facilitando las tareas de búsqueda de datos y relaciones ocultas; finalmente, gracias a la aplicación del Álgebra Relacional y sus operadores de Proyección y Selección, junto a la medida de similaridad como restricción de búsqueda, permite proponer un modelo, genérico, que puede ser implementado en una amplia variedad de Lenguajes de Programación y Bases de Datos con soporte a búsqueda multidimensional; que finalmente, puede ser aplicado en una amplia gama de fenómenos de persistencia observable, de comportamiento estocástico no lineal.

## 6.2. Específicas

1. Se ha descrito teóricamente los procesos estocásticos, conceptos de variable aleatoria, modelos lineales ARMA, PARMA, se ha visto la importancia del ruido blanco como un bloque que describe un Proceso Estocástico básico; Luego la definición de series temporales y algunos estimadores usados para describirlos.

2. Se ha presentado los modelos usados en la literatura para la generación de series temporales asociadas a variables climatológicas, el modelo lineal de Thomas Fiering, luego un modelo basado en redes neuronales propuesto por Luciana Conceicao Campos, que trabaja sin información a priori y que no requieren una formulación compleja, se evidenciaron las limitaciones sobre la aplicabilidad de las propuestas para caracterizar información oculta. Luego se presentaron los trabajos Maria Malek, Ning Xiong, Pei-Chann Chang, donde se muestra la capacidad del Razonamiento Basado en Casos para descubrir información oculta, sobre series temporales y tareas de pronóstico.

3. Se ha detallado y descrito, significativamente, el Razonamiento Basado en Casos, mostrando su capacidad para trabajar con múltiples dimensiones y grados de información, registrando de manera formal información y relaciones ocultas, finalmente se discutió su aplicabilidad en la generación de series temporales estocásticas.

4. Se ha logrado formular un nuevo modelo llamándose «Modelo Estocástico a partir de Razonamiento Basado en Casos para la Generación de Series Temporales» (PERBC), siendo un modelo genérico que puede ser implementado en una amplia gama de fenómenos no lineales de comportamiento estocástico; con la capacidad de manejar todos los casos incorporados a la memoria; Auto-regresivo, en series temporales que presenten un fenómeno de persistencia observable.

5. Se Aplicó el modelo propuesto (PERBC) en la generación de series temporales para la generación de escenarios en la Cuenca del Rio Chili, en las estaciones de El Pañe, Aguada Blanca, El Frayle, para las variables hidrometeorológicas: Caudal, Evaporación y Precipitación. los resultados muestran que el modelo , en algunos casos tiene una baja capacidad para reproducir las características generales de la serie observada, lo cual es generado aceptablemente por el modelo TF y el PEN, sin embargo en la mayoría de los casos logra mostrar eventos extremos, lo que evidencia su habilidad para mostrar detalles ocultos que los modelos TF y PEN no logran.

6. Un análisis detallado de la media, Cuadro 5.1; desviación estándar, Cuadro 5.2 y la asimetría, Cuadro 5.3 para todos los experimentos de los modelos TF, PEN muestran que conservan satisfactoriamente las características de la serie histórica, Sin embargo se observan generaciones leptocúrticas para el modelo PERBC (el propuesto) respecto a sus similares (vea la desviación estándar), no siendo tan descriptivo como los otros, ahora bien se puede concluir que el modelo ajusta los pronósticos y reduce significativamente la incertidumbre, una propiedad del RBC por su manejo multidimensional (Pal y Shiu, 2004; Loucks y cols., 2005).

   Los eventos máximos y mínimos fueron reproducidos satisfactoriamente, PERBC genera mínimos extremos, en el Cuadro 5.5.2 el comportamiento de los mínimos sobre las precipitaciones del Pañe, Frayle y Aguada Blanca representan a la serie histórica (el modelo PERBC consigue generar mínimos 0); TF y PEN tienen valores aproximados, para los máximos TF y PEN son mas generosos que PERBC; sin embargo en lineas generales se puede inferir un mejor desempeño para generar series que contemplen valores extremos (sequías) del modelo PERBC sobre el TF y el PEN.

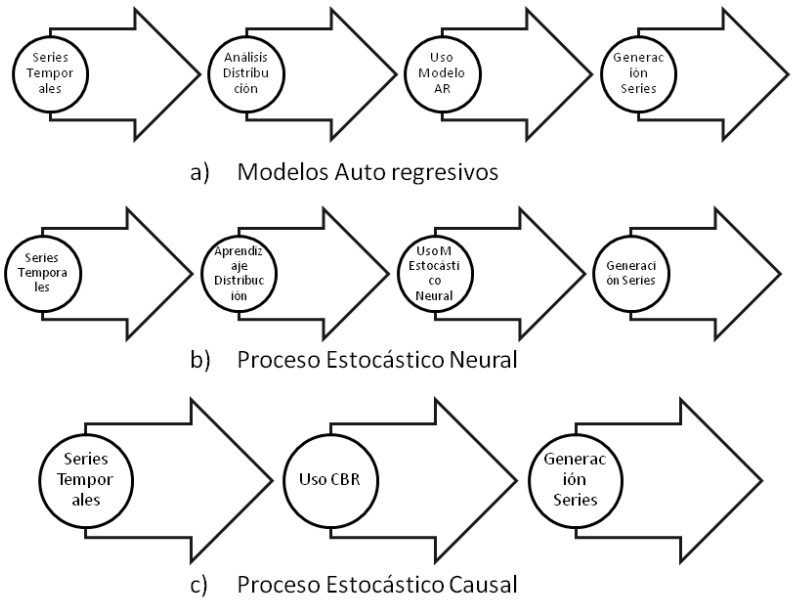



Figura 6.1: a) Modelos Autoregresivos VS b) Proceso Estocástico Neural VS c) Proceso Estocástico RBC (Propuesta).

## 6.3. Ventajas del modelo

- Como se muestra en el análisis de resultados el modelo PERBC tiene la habilidad de descubrir características ocultas y reproducirlas en la generación de series temporales, particularmente los mínimos extremos, y algunos máximos; los modelos TF y PEN reproducen aproximaciones.

- El formularlo de manera genérica permite incluir mas dimensiones y grados, considere por ejemplo incluir una dimensión espacial, datos georeferenciados de imágenes satelitales, fenómenos paralelos en otras ubicaciones geográficas pero de similares características, el modelado de fenómenos de otros áreas distintas a las presentadas en la Tesis.

- Adicionalmente, es un modelo que no requiere una formulación a priori, ni tareas de aprendizaje, el uso del ciclo de vida del RBC lo hacen relativamente automático, vea la Figura 6.1.

- El Álgebra Relacional mejora la expresividad matemática de la propuesta, computacionalmente es un beneficio relativo ya que, siendo una expresión matemática, es factible de ser implementado en diferentes lenguajes informáticos, con diferentes estructuras de indexación multidimensional

- Es una contribución complementaria en el área de representación planificación, desarrollo, administración, de muchos sistemas reales; vinculados a fenómenos hidrometeorológicos, financieros, biológicos y físicos.

## 6.4. Desventajas del modelo

- Tiene generaciones leptocurticas, en algunos casos no representan a la serie histórica.

- El uso de la memoria de todos los casos para la generación de las Series Temporales genera una dependencia a los métodos de acceso métrico; sino se usa, su desempeño es bajo para grandes volúmenes de información, considere el caso de incluir series temporales de imágenes.

- Existen modelos que tratan información extrema, véase los modelos de Régimen Extremo, se debe notar que un dato oculto no necesariamente siempre es extremo, luego el modelo no siempre encuentra datos extremos máximos.

## 6.5. Contribuciones

- Se puede usar el nuevo modelo PERBC como complemento en las tareas de análisis de escenarios junto a los modelos tradicionales, el modelo se destaca por la habilidad de incluir características ocultas (ejemplo: datos extremos) en las realizaciones, lo que permite evaluar eventos extremos (sequías, heladas, lluvias torrenciales) esto permitirá a un tomador de decisión desarrollar acciones técnicas de previsión, que finalmente puedan evitar pérdidas económicas y sociales (Construcción de defensas rivereñas para evitar inundaciones, implantación de políticas de consumo de agua para mejorar la disponibilidad del recurso hídrico, ajustando el impacto del evento sobre el área vulnerable correspondiente)

- La propuesta se clasifica como un modelo estocástico periódico auto-regresivo genérico.

## 6.6. Trabajo futuro

1. Es conocido que los estadísticos de primer orden (media, varianza, desviación típica) no contienen información suficiente para capturar detalles ocultos sobre los datos; por lo que se recomienda extender el modelo para trabajar con estadísticos

de orden superior, considerando la existencia de investigaciones recientes en esta área (de la Rosa, Agüera-Pérez, Palomares-Salas, Sierra-Fernández, y Moreno-Muñoz, 2012).

2. La propuesta fue implementada sobre el lenguaje M, un lenguaje interpretado; para justificar plenamente el uso de la memoria plana sobre los registros almacenados es recomendable la implementación sobre un lenguaje compilado, este trabajo futuro permitirá la evaluación de diferentes estructuras de acceso métrico.

3. Dada las características de estimación por similaridad, el componente determinístico del modelo se puede extender para completación de datos, análisis de consistencia de datos, análisis de doble masa, y ciertas tareas de pronóstico.

4. Se debe considerar la estimación del componente aleatorio a partir de un análisis de las distancias de similaridad, basado en la propuesta de campos sobre la creación del componente aleatorio a partir de los residuos (Campos, 2010); se cree que mejoraría las generaciones.

## 6.7. Reflexiones finales

Se han generado 2700 series temporales, 32400 datos; en todas ellas la incertidumbre esta presente; se sabe que en los sistemas de recursos hídricos, esta incertidumbre se debe a factores que afectan el desempeño del sistema y que no son conocidos. El éxito y desempeño de cada componente frecuentemente depende de condiciones futuras en aspectos meteorológicos, demográficos, económicos, sociales, técnicos y políticos; todos los cuales pueden influir en los beneficios futuros, costos, impacto ambiental, aceptación social. La incertidumbre también se debe a la naturaleza estocástica de los procesos meteorológicos, como la precipitación, evaporación, temperaturas, así como la

población futura, consumo de agua por persona, patrones de irrigación, prioridades en el uso de agua; todo lo cual afecta la demanda y nunca se conoce con certeza. (Loucks y cols., 2005)

Como se analizó, los modelos lineales tratan la incertidumbre, manejando estadísticos de primer orden, lo cual es aceptable si la incertidumbre es razonablemente pequeña y no afecta el desempeño; en estos casos el planificador puede evaluar la importancia de la incertidumbre mediante un análisis de sensibilidad. Ahora bien, usar modelos tradicionales, en un modelo complejo, puede generara una pobre representación del desempeño. Un análisis completo requiere de la evaluación tanto de los resultados esperados del proyecto, el riesgo y posible magnitud de las fallas del sistema en un contexto físico, social, económico y ecológico; se puede ver que modelos como los de Luciana (Campos, 2010), Taymoor (Awchi y cols., 2009) y otros incluyen nuevos análisis para la generación de series temporales, sin embargo su formulación es compleja, luego los modelos basados en aprendizaje (redes neuronales) aveces no reproducen características ocultas debido a su habilidad para la generalización; finalmente, se puede sentenciar que es complejo lidiar con la incertidumbre, el modelo propuesto es un intento más por administrarla, si bien es cierto la habilidad de manejar información de múltiples variables reduce la incertidumbre, lo cierto es que humanamente aun es imposible administrarla y todo se convierte en aproximaciones de una realidad subjetiva, se requiere de una inteligencia sobresaliente con naturaleza divina, aun no disponible, para gobernar y gerencias todos los fenómenos que rodean nuestra futura y escasa existencia.

## 6.8. Publicaciones generadas

Se presenta las diferentes publicaciones logradas en el transcurso de esta investigación.

1. Modelo Estocástico a partir de Razonamiento Basado en Casos para la Generación de Series Temporales, José Herrera Quispe, Yessenia Yari, Luis Alfaro, Yván Túpac. Jornadas Peruanas de Computación; Chiclayo PERU 2013.

2. Red Neuronal aplicada a la generación de caudales mensuales estocásticos, José Herrera Quispe, Yessenia Yari, Yvan Túpac. Jornadas Peruanas de Computación; Chiclayo PERU 2013.

3. Stochastic Processes Using Case-based Reasoning for Generation of Time Series. A. José Herrera Quispe, B. Luis A. Alfaro Casas, C. Yessenia Yari 1, and Yvan Tupac. 12th Grace Hopper Celebration of Women in Computing, BALTIMORE USA Octubre 2012.

4. A Novel Stochastic processes using slope of correlation limited by thresholds and similarity for generation of time series flows A. José Herrera Quispe, B. Luis A. Alfaro Casas, C. Yessenia Yari 1, and Yvan Tupac. FCS'12 - The 2012 International Conference on Foundations of Computer Science, NEVADA USA Julio 2012.

   OTRAS RELATIVAS:

5. Optimización Inteligente de Reglas de Operación a partir de Series Temporales de Caudales, Jornadas Chilenas de Computación Santiago de Chile, 2012.

6. Razonamiento Basado en Casos en el reconocimiento de dígitos manuscritos del MNIST, José Herrera Quispe, Luis Alfaro, Cesar Beltran Castañon. Jornadas Peruanas de Computación; Puno PERU 2012.

7. Case Based Reasoning in recognition of MNIST - The 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition; A. José Herrera Quispe, B. Luis A. Alfaro Casas, Cesar Beltran Castañon; Nevada USA 2011

8. Optimal Calibration of Parameter of a Conceptual Rainfall-Runoff Model Using Genetic Algorithm, A. José Herrera Quispe, B. Luis A. Alfaro Casas, C.Jorge Luis Suaña, WORLDCOMP'11 ; Las Vegas USA 2011

9. Modelo Gr4j usando Algoritmos Genéticos. Caso: Cuenca Del Rio Chili INTERCOM - IEEE, PERU 2010.

# Referencias


Awchi, T. A., Srivastava, D., y cols. (2009). Analysis of drought and storage for mula project using artificial neural network and stochastic generation models. *Hydrology Research*, *40*(1), 79–91.

Baeza-Yates, R. A., Cunto, W., Manber, U., y Wu, S. (1994). Proximity matching using fixed-queries trees. En *Cpm* (p. 198-212).

Bao, H., y Cao, J. (2011, January). Delay-distribution-dependent state estimation for discrete-time stochastic neural networks with random delay. *Journal of Neural Networks & Computer Science*, *24*, 19–28. doi: http://dx.doi.org/10.1016/j.neunet.2010.09.010

Bareiss, R. (1989). Exemplar-based knowledge acquisition. *Perspectives in artificial intelligence*, *2*, 1–169.

Beard, L. R., y Kubîk, H. (1967). Monthly streamflow simulation. *Computer Program*, 1–6.

Bonzano, A., Cunningham, P., y Smyth, B. (1997). Using introspective learning to improve retrieval in cbr: A case study in air traffic control. *Case-Based Reasoning Research and Development*, 291–302.

Bozkaya, T., y Özsoyoglu, Z. M. (1997). Distance-based indexing for high-dimensional metric spaces. En *Sigmod conference* (p. 357-368).

Brillinger, D. (2001). *Time series: data analysis and theory*. Society for Industrial and Applied Mathematics.

Brin, S. (1995). Near neighbor search in large metric spaces. En *21th international conference on very large data bases (vldb 1995)* (p. 574-584).

Brittan, M. R. (1961). *Probability analysis applied to the development of synthetic hydrology for the colorado river*. Bureau of Economic Research, University of Colorado.

Brockwell, P., y Davis, R. (2009). *Time series: Theory and methods*. Springer.

Cadavid, J., y Salazar, J. (2008). Generación de series sinteticas de caudales usando un modelo matalas con medias condicionadas. *Avances en Recursos Hidráulicos*, 17–24.

Campos, L. C. D. (2010). *Modelo estocastico periodico baseado em redes neurais*. Tesis Doctoral no publicada, Pontificia Universidade Catolica do rio de Janeiro, Rio de

Janeiro - Brasil.

Chang, P.-C., Tsai, C.-Y., Huang, C.-H., y Fan, C.-Y. (2009). *Application of a case base reasoning based support vector machine for financial time series data forecasting* (Vol. 5755; D.-S. Huang, K.-H. Jo, H.-H. Lee, H.-J. Kang, y V. Bevilacqua, Eds.). Springer Berlin, Heidelberg.

Chávez, E., Navarro, G., Baeza-Yates, R., y Marroquín, J. L. (2001, septiembre). Searching in metric spaces. *ACM Comput. Surv.*, *33*(3), 273–321.

Cheng, A., y Bear, J. (2008). *Modeling time series of groundwater flow and contaminant transport*. Springer.

Ciaccia, P., Patella, M., y Zezula, P. (1997). M-tree: An efficient access method for similarity search in metric spaces. En *Proceedings of the 23rd international conference on very large data bases* (pp. 426–435). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Colston, N., y Wiggert, J. (1970). A technique of generating a synthetic flow record to estimate the variability of dependable flows for a fixed reservoir capacity. *Water Resources Research*, *6*(1), 310–315.

Craw, S., Jarmulak, J., y Rowe, R. (2001). Maintaining retrieval knowledge in a case-based reasoning system. *Computational Intelligence*, *17*(2), 346–363.

de la Rosa, J. J. G., Agüera-Pérez, A., Palomares-Salas, J. C., Sierra-Fernández, J. M., y Moreno-Muñoz, A. (2012). A novel virtual instrument for power quality surveillance based in higher-order statistics and case-based reasoning. *Measurement*, *45*(7), 1824 - 1835. doi: http://dx.doi.org/10.1016/j.measurement.2012.03.036

De Mantaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., ... others (2005). Retrieval, reuse, revision and retention in case-based reasoning. *Knowledge Engineering Review*, *20*(3), 215.

Dohnal, V., Gennaro, C., Savino, P., y Zezula, P. (2003). D-Index: Distance Searching Index for Metric Data Sets. *Multimedia Tools Appl.*, *21*(1), 9–33.

Elmasri, R., y Navathe, S. (2010). *Database systems: Models, languages, design, and application programming*. Pearson.

Elmasri, R., y Navathe, S. (2011). *Fundamentals of database systems*. Addison Wesley Publishing Company Incorporated.

El-Shafie, A., y El-Manadely, M. (2011). An integrated neural network stochastic dynamic programming model for optimizing the operation policy of aswan high dam. *Hydrology research*, *42*(1), 50–67.

Fiering, M. B. (1967). Streamflow synthesis. *Cambridge, Harvard University Press, 1967. 139 P.*

Filho, R. F. S., Traina, A. J. M., Jr., C. T., y Faloutsos, C. (2001). Similarity search without tears: The omni family of all-purpose access methods. En *Icde* (p. 623-630).

Funk, P., y Xiong, N. (2006). Case-based reasoning and knowledge discovery in medical applications with time series. *Computational Intelligence*, *22*(3-4), 238–253.

Gangyan, Z., Goel, N., y Bhatt, V. (2002). Stochastic modelling of the sediment load of the upper yangtze river (China). *Hydrological sciences journal*, *47*(S1), 93–105.

Gutierrez, J. (2003). *Monitoramento da instrumentaco da barragem de corumbai por redes neurais e modelos de box and jenkins*. Dissertacao de mestrado pontifica universidade catolica do rio de janeiro, Departamento de Engenharia Civil.

Hajdinjak, M., y Bierman, G. (2011). Extending the relational algebra with similarities. *Poslano v Mathematical Structures in Computer Science*.

Hammond, K. (1989). *Case-based planning: viewing planning as a memory task*. Academic Press Professional, Inc.

Han, M., y Wang, Y. (2009). Analysis and modeling of multivariate chaotic time series based on neural network. *Expert Systems with Applications*, *36*(2, Part 1), 1280 - 1290. doi: DOI:10.1016/j.eswa.2007.11.057

Haykin, S. (2001). *Redes neurais: Princípios e prática.* (Bookman, Ed.). Porto Alegre, RS.

He, W., Xu, L. D., Means, T., y Wang, P. (2009). Integrating web 2.0 with the case-based reasoning cycle: A systems approach. *Systems Research and Behavioral Science*, *26*(6), 717–728. doi: 10.1002/sres.976

Hinrichs, T. (1992). *Problem solving in open worlds: A case study in design*. Lawrence Erlbaum.

Hjaltason, G. R., y Samet, H. (2003). Index-driven similarity search in metric spaces. *ACM Trans. Database Syst.*, *28*(4), 517-580.

Hochreiter, R., y Pflug, G. (2007). Financial scenario generation for stochastic multi-stage decision processes as facility location problems. *Annals of Operations Research*, *152*(1), 257–272.

Jaeger, H. (2000). Observable operator models for discrete stochastic time series. *Neural Computation*, *12*(6), 1371–1398.

Jr., C. T., Traina, A. J. M., Seeger, B., y Faloutsos, C. (2000). Slim-trees: High performance metric trees minimizing overlap between nodes. En *Edbt* (p. 51-65).

Julian, P. R. (1961). *A study of the statistical predictability of stream-runoff in the upper colorado river basin*.

Kantz, H., y Schreiber, T. (2004). *Nonlinear time series analysis*. Cambridge University Press.

Kjeldsen, T. R., y Rosbjerg, D. (2004). Choice of reliability, resilience and vulnerability estimators for risk assessments of water resources systems/choix destimateurs de fiabilite, de resilience et de vulnerabilite pour les analyses de risque de systemes de ressources en eau. *Hydrological sciences journal*, *49*(5).

Kolodner, J. (1983a). Maintaining organization in a dynamic long-term memory*. *Cognitive science*, *7*(4), 243–280.

Kolodner, J. (1983b). Reconstructive memory: A computer model*. *Cognitive Science*, *7*(4), 281–328.

Lajmi, S., Ghedira, C., y Benslimane, D. (2006). Wesco cbr: Web services via case based reasoning. En *Icebe 06. ieee international conference* (pp. 618–622).

Lee, C., Cheng, K., y Liu, A. (2008). A case-based planning approach for agent-based service-oriented systems. En *Systems, man and cybernetics, 2008. smc 2008. ieee international conference on* (pp. 625–630). (Dept. of Computer Science & Inf. Eng., Nanhua Univ., Chiayi)

Lee, C., Liu, A., y Huang, H. (2010). Using planning and case-based reasoning for web service composition. *Journal ref: Journal of Advanced Computational Intelligence and Intelligent Informatics*, *14*(5), 540–548.

Loor, P. D., Bénard, R., y Chevaillier, P. (2011). Real-time retrieval for case-based reasoning in interactive multiagent-based simulations. *Expert Systems with Applications*, *38*(5), 5145 - 5153. doi: DOI:10.1016/j.eswa.2010.10.048

Loucks, D., Van Beek, E., Stedinger, J., Dijkman, J., y Villars, M. (2005). *Water resources systems planning and management: an introduction to methods, models and applications*. Paris: UNESCO.

Malek, M., y Kanawati, R. (2009). Case-based reasoning in knowledge discovery and data mining (Tesis Doctoral, Wiely). *Recherche*.

Meng, T., Somani, S., y Dhar, P. (2004). Modeling and simulation of biological systems with stochasticity. *Silico Biol*, *4*(3), 293–309.

Navarro, G. (2002, agosto). Searching in metric spaces by spatial approximation. *The VLDB Journal*, *11*(1), 28–46.

Ochoa-Rivera, J. C. (2008). Prospecting droughts with stochastic artificial neural networks. *Journal of Hydrology*, *352*(1-2), 174 - 180. doi: DOI:10.1016/j.jhydrol .2008.01.006

Oviedo T., J., Umeres R., H., Franco R., R., Vílchez, G., y Butrón, D. (2001). *Diagnóstico de gestión de la oferta de agua de la cuenca quilca - chili* (Inf. Téc.). INADE - AUTODEMA.

Oviedo Tejada, J. M. (2004). *Propuesta de asignaciones de agua en bloque (volúmenes anuales y mensuales) para la formalización de los derechos de uso de agua en los valles chili regulado y chili no regulado del programa de formalización de derechos de uso de agua - profodua* (Inf. Téc.). Ministerio de Agricultura - Instituto Nacional de Recursos Naturales - Intendencia de Recursos Hídricos - Administración Técnica del Distrito de Riego Chili.

Pal, S., y Shiu, S. (2004). *Foundations of soft case-based reasoning*. John Wiley & Sons.

Peng, C.-s., y Buras, N. (2000). Dynamic operation of a surface water resources system. *Water Resources Research*, *36*(9), 2701–2709.

Prudencio, R. (2002). *Projeto híbrido de redes neurais*. Tesis de Master no publicada, Mestrado em ciencias da computacao - Universidade Federal de Pernambuco.

Raman, H., y Sunilkumar, N. (1995). Multivariate modelling of water resources time series using artificial neural networks. *Hydrological Sciences Journal*, *40*(2), 145–

163.

Ramirez, F. O. P. (2007). *Introducción a las series de tiempo. métodos paramétricos.* Editora Correo Restrepo.

Romero, O., Marcel, P., Abelló, A., Peralta, V., y Bellatreche, L. (2011). Describing analytical sessions using a multidimensional algebra. *Data Warehousing and Knowledge Discovery*, 224–239.

Ruiz, E. V. (1986, julio). An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recogn. Lett.*, *4*(3), 145–157.

Salas, J. D., Tabios III, G. Q., y Bartolini, P. (1985). Approaches to multivariate modeling of water resources time series1. *JAWRA Journal of the American Water Resources Association*, *21*(4), 683–708.

Schank, R. (1982). Dynamic memory: A theory of reminding and learning in computers and people. *New York*.

Schank, R., Abelson, R., y cols. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures* (Vol. 2). Lawrence Erlbaum Associates Nueva Jersey.

Sebag, M., y Schoenauer, M. (1994). A rule-based similarity measure. *Topics in case-based reasoning*, 119–131.

Simoudis, E. (1992). Using case-based retrieval for customer technical support. *IEEE Expert*, *7*(5), 7–12.

Simpson, R. (1985). A computer model of case-based reasoning in problem solving: an investigation in the domain of dispute mediation.

Singh, V., y Yadava, R. (2003). *Water resources system operation: proceedings of the international conference on water and environment (we-2003), december 15-18, 2003, bhopal, india* (n.º v. 1). Allied Publishers.

Smyth, B., y Champin, P. (2009). The experience web: A case-based reasoning perspective. En *Grand challenges for reasoning from experiences, workshop at ijcai* (Vol. 9).

Srikanthan. (2002). *Stochastic generation of monthly rainfall data.* CRC for Catchment Hydrology.

Sumathi, S., y Esakkirajan, S. (2007). *Fundamentals of relational database management systems.* Springer.

Sycara, K. (1988). Using case-based reasoning for plan adaptation and repair. En *Proceedings of the darpa case-based reasoning workshop* (Vol. 425, p. 434).

Tang, C. F. P. A., Z.; Almeida. (1991). Time series forecasting using neural networks vs box-jenkins methodology. *SIMULATION*, *57*, 303-310.

Taylor, S. (2008). *Modelling financial time series.* World Scientific Pub Co Inc.

Thomas, H., y Fiering, M. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. *Design of water resource systems*, 459–493.

Tokdemir, O., y Arditi, D. (1999). Comparison of case-based reasoning and artificial neural networks. *Journal of computing in civil engineering*, *13*, 162.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Uhlmann, J. K. (1991). Satisfying general proximity/similarity queries with metric trees. *Inf. Process. Lett.*, *40*(4), 175-179.

Ünal, N., Aksoy, H., y Akar, T. (2004). Annual and monthly rainfall data generation schemes. *Stochastic Environmental Research and Risk Assessment*, *18*(4), 245–257.

Vieira, C., de Carvalho Júnior, W., y Solos, E. (s.f.). Utilização de redes neurais artificiais para predição de classes de solo em uma bacia hidrográfica no domínio de mar de morros césar da silva chagas elpídio inácio fernandes filho 2.

Weber, G. (1995). Examples and remindings in a case-based help system. *Advances in Case-Based Reasoning*, 165–177.

Wei, W. W.-S. (1994). *Time series analysis*. Addison-Wesley Redwood City, California.

Wilkinson, D. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, *10*(2), 122–133.

Zadeh, L. (2003). *Foreword of foundations of soft case-based reasoning*. Berkely, CA.

Zezula, P., Amato, G., Dohnal, V., y Batko, M. (2006). *Similarity search: The metric space approach* (Vol. 32). Springer.