# Modelado no lineal de aportes al sistema eléctrico

Informe final del proyecto FSE 2013-1-10764

30 de noviembre de 2016

# Tabla de contenidos

# 1 Introducción

El presente informe resume el trabajo realizado durante los 24 meses de duración del proyecto FSE-1-2013-1-10764, motivado por un análisis crítico de la herramienta SimSEE que actualmente se utiliza en UTE para determinar la operación óptima del sistema eléctrico uruguayo. De acuerdo a los objetivos planteados, a) se realizó un análisis en profundidad del SimSEE y sus potenciales falencias, b) se obtuvo un conjunto de modelos estadísticos no lineales alternativos a los utilizados hoy en día en SimSEE (específicamente, para el modelado de aportes hídricos), los cuales fueron incorporados como módulos al SimSEE y c) se desarrolló un marco de estadístico de evaluacion de los modelos apropiado al problema.

Actualmente la UTE utiliza el Simulador de Sistemas de Energía Eléctrica (SimSEE) [4, 1] para la toma de decisiones relacionada con el manejo de sus recursos energéticos que incluyen, entre otros, los aportes de las represas hidroeléctricas del país.

El SimSEE se compone esencialmente de dos bloques. El primero es un *simulador*, encargado de modelar posibles evoluciones del estado del sistema, incluyendo el estado de máquinas, represas, y posibles aportes como ser los hídricos o los eólicos. Para estos últimos, el simulador emplea actualmente un modelo llamado *Correlaciones en Espacio Gaussiano con Histograma* (CEGH), que es el foco de este trabajo. La segunda es un *optimizador* que calcula, mediante un algoritmo de programación dinámica estocástica [12], la decisión óptima instantánea para cada posible estado del sistema.

El presente proyecto nace de un análisis crítico realizado al simulador CEGH, en donde se caracterizan las propiedades del modelo estadístico que lo subyace, y se identifican un conjunto de limitaciones. Como resultado de dicho análisis, y de las limitaciones encontradas, este trabajo plantea una familia de modelos alternativos no lineales que no presentan las limitaciones del CEGH.

Para comparar los nuevos modelos con el CEGH se implementaron los modelos alternativos mencionados en el SimSEE y se evaluó su desempeño con datos históricos reales, evitando algunos de los problemas metodológicos utilizados previamente a la hora de evaluar modelos en SimSEE, como será mencionado más adelante.

Sin embargo, el evaluar el sistema sólo en SimSEE introduce un sesgo indeseable. Las limitaciones propias del SimSEE pueden afectar de manera distinta a cada modelo independientemente de su capacidad de reproducir fielmente las propiedades estadísticas de las series a simular, incluso aquellas que deberían ser relevantes en lo que refiere a la generación eléctrica. Por esto, también se consideró esencial disponer de un marco comparativo de los modelos a nivel estadístico, es decir, de una medida intrínseca de la capacidad de los distintos modelos de capturar la dinámica real de las series de aportes hídricos. Desafortunadamente, las medidas utilizadas hasta la fecha de comienzo de este proyecto, (derivadas de mecánica de los fluidos) resultan altamente insatisfactorios para caracterizar estadísticamente el desempeño de modelos como el CEGH o los propuestos. Es por esto que buena parte del proyecto se dedicó a desarrollar medidas apropiadas, y ésto en sí constituye un producto importante del proyecto.

## 1.1. Estructura del documento

El resto del documento se compone de la siguiente manera. En la sección 2 se describe de manera muy simplificada al SimSEE, el problema que busca resolver, y la forma en que se utiliza para resolverlo. En particular, se presentan algunas características conocidas de las series de aportes hídricas. La sección 4 describe el modelo CEGH y trata de caracterizar de manera teórica sus propiedades y limitaciones. En particular, se detallan algunas modificaciones no menores que el modelo original CEGH requiere para ser aplicado dentro del SimSEE. La sección 3 provee un análisis estadístico detallado de las series a modelar. Los resultados empíricos detallados en esta sección fueron el principal insumo para diseñar los modelos no lineales que luego se presentan en las secciones 5.1 y 5.7. La sección 6 provee un análisis crítico de las medidas estadísticas típicamente utilizadas para medir el desempeño de simuladores de series temporales, los inconvenientes que éstas presentan para el caso puntual tratado en este proyecto, y luego el llamado "marco comparativo intrínseco" desarrollado en este proyecto. Este último es de interés de por sí como herramienta genérica para medir la calidad de simulaciones de procesos. También se detalla el llamado "marco comparativo extrínseco", que no es otra cosa que el protocolo de prueba utilizado para evaluar los modelos desarrollados dentro del SimSEE. La sección 7 presenta los resultados obtenidos en términos de ambos marcos comparativos. Las conclusiones obtenidas del proyecto se resumen en la sección 8, junto con posibles lineas de trabajo futuro.

# 2 Descripción del problema

## 2.1. Notación

De aquí en adelante utilizaremos mayúsculas para referirnos a variables aleatorias, posiblemente multidimensionales, por ejemplo $X$, y minúsculas para realizaciones de ellas, $x$. El espacio donde toma valores $X$ será denotado por la letra caligráfica correspondiente, en el ejemplo $\mathcal{X}$. Si $X$ es multidimensional, utilizaremos $X[i]$ para referirnos a su $i$-ésimo elemento. Al trabajar con procesos estocásticos, utilizaremos subíndices para referirnos a índices dentro de las secuencias de variables aleatorias, $\{X_1, \ldots, X_n\}$; lo mismo con secuencias de realizaciones $\{x_1, \ldots, x_n\}$. Utilizaremos la notación $X_{i:j}$ para referirnos a una subsecuencia de variables, $X_{i:j} = \{X_i, \ldots, X_j\}$, $X_{:j} = X_{1:j}$ indica la subsecuencia desde el comienzo hasta $t$ (inclusive) y $X_{i:}$ o $X_{i:\infty}$ la subsecuencia desde $i$ en adelante.

## 2.2. Operación óptima del sistema eléctrico

El SimSEE es un conjunto de herramientas diseñado para optimización de la operación del sistema eléctrico de Uruguay [1]. Las dos más relevantes en lo que respecta a este proyecto son el *análisis serial* y el *optimizador*. El primero se utiliza para ajustar modelos de series temporales a partir de datos históricos de ellas; el modelo utilizado hoy en día es el CEGH (Correlación en Espacio Gaussiano de Histogramas) [2], del cual hablaremos más adelante. El otro componente, el optimizador, recibe una configuración de operación del sistema (llamada *sala*), donde se incluyen las distintas fuentes aleatorias y determinísticas que inciden en la generación eléctrica, y se busca obtener una *política de operación óptima* del sistema.

El algoritmo de optimización que determina la toma de decisiones óptima se basa en la minimización (aproximada) de la *función de Bellman* [12]. Sea $x_{1:n} = \{x_1, x_2, \ldots, x_n\}, x_t \in \mathcal{X}$ una secuencia de estados del sistema, con $\mathcal{X}$ el espacio de estados; el estado incluye por ejemplo el nivel de los embalses en las represas hidroeléctricas. La variable aleatoria $X \in \mathcal{X}$ representa los posibles valores que pueden tomar las $x_t$. Luego $v_t \in \mathcal{V}$ representa las variables de entrada al sistema en el tiempo $t$, y $w_t$ es el conjunto de decisiones tomadas por quien opera al sistema, por ejemplo, cuánto abrir una válvula; El sistema evolucionará hacia un nuevo estado $x_{t+1}$ en función de $x_t, v_t$ y $w_t$. Dicha evolución se modela mediantela *función de transición de estado* $f(\cdot)$ como,

$$x_{t+1} = f(x_t, v_t, w_t, t). \tag{2.1}$$

Las decisiones tomadas en el instante $t$, $w_t$ son función del estado $x_t$ y la entrada $x_t$ actuales. Llamamos a esta función *política de operación*,

$$w_t = o(x_t, v_t). \tag{2.2}$$

Finalmente, se define la función de *costo instantáneo* $\ell(\cdot)$ de operación como

$$c_t = \ell(x_t, v_t, w_t, t), \tag{2.3}$$

y a la función de *costo futuro* $L(\cdot)$ como

$$C_t = L(x_t, v_{t:\infty}, w_{t:\infty}) = \sum_{j=t}^{\infty} \ell(x_j, v_j, w_j, j). \tag{2.4}$$

Notar que esta función depende del estado actual $x_t$ y de *toda* la secuencia futura de entradas $v_{t:\infty}$, y las operaciones futuras $w_{t:\infty}$. Es inmediato ver que, de acuerdo a las definiciones dadas, el costo futuro puede calcularse recursivamente,

$$L(x_t, v_{t:\infty}, w_{t:\infty}) = \ell(x_t, v_t, w_t, t) + L(x_{t+1}, v_{t+1:\infty}, w_{t+1:\infty}) \tag{2.5}$$

o, lo que es lo mismo,

$$C_t = c_t + C_{t+1}.$$

Notar además que, si se dispone de la función $g(\cdot)$, $w_t$ es función de $x_t$ y $v_t$ por lo que puede eliminarse $w_t$ en la recursión anterior

$$L(x_t, v_{t:\infty}) = \ell(x_t, v_t, g(x_t, v_t), t) + L(x_{t+1}, v_{t+1:\infty}). \tag{2.6}$$

El objetivo final del SimSEE es definir la función $g^*(\cdot)$ de modo que el costo futuro $L$ en (2.6) sea minimizado. Debido a que la trayectoria futura es conocida sólo en probabilidad, la optimalidad de $g(\cdot)$ puede expresarse sólo en términos probabilísticos. La forma en que esto se realiza tradicionalmente es minimizando la función de Bellman en términos de la operación $w$:

$$w_t^*(x_t, v_t) = g^*(x_t, v_t) = \arg\min_{\omega}\{\ \ell(x_t, v_t, \omega) + E[\, L(x_{t+1}, v_{t+1})\,]\ \}. \tag{2.7}$$

La función de Bellman (2.7) implica el cálculo del valor esperado del costo futuro dado el estado y la entrada actual, con respecto al vector de entradas futuras con un horizonte infinito. Debido al tamaño del espacio de integración (comprendido por todas las posibles realizaciones de $v_{t+1:\infty}$), en la práctica se realiza una serie de aproximaciones mencionadas a continuación.

Para obtener la política de operación óptima $g^*(\cdot,\cdot)$, el optimizador debe resolver (2.7) para todos los pares posibles $(x, v)$. El SimSEE aproxima la minimización de la función de Bellman mediante una técnica conocida como *programación dinámica estocástica hacia atrás* (ABDP) [18]. Esto implica resolver el problema (2.7) empezando desde un horizonte finito prefijado $t = n$ (donde el problema es determinístico), y yendo hacia atrás en el tiempo de a un paso a la vez hasta llegar al tiempo presente $t = j$.

La mayor limitante del algoritmo ABDP es el tamaño potencialmente muy grande del espacio $\mathcal{X}\times\mathcal{V}$. Para empezar, tanto $x$ como $v$ deben ser conjuntos finitos, o sea que toda variable de estado o entrada debe ser *cuantificada* de alguna manera, con la consiguiente pérdida de información. Actualmente se realiza lo más sencillo, que es cuantificar de manera uniforme (una grilla) el espacio $\mathcal{X} \times \mathcal{V}$. Como veremos adelante, la cantidad de niveles de cuantificación representables en la práctica es tal que invalida muchas hipótesis importantes acerca de los modelos utilizados en SimSEE.

# 3 Propiedades estadísticas de las series de aportes

Todo modelo estadístico se beneficia de cualquier información a priori (fiable) que podamos tener sobre los datos que se desea modelar. El propósito de esta sección es mostrar y/o corroborar las propiedades estadísticas de las principales series de aportes a simular en este proyecto, que son las series de aportes hídricos a los embalses de Bonete, Palmar y Salto.

El desafío de cualquier modelo que se utilice para estas series es lograr capturar aquellas propiedades estadísticas que son relevantes para la generación de energía hidráulica con un número de puntos $|\mathcal{X} \times \mathcal{V}|$ manejable.

La información que tenemos a priori sobre las series puede resumirse de la siguiente manera:

1. Son no estacionarias.
2. Tienen cierta periodicidad (son cosas que en última instancia dependen del clima, que claramente tiene ciclos anuales).
3. Están correlacionadas temporalmente ya que los caudales de los ríos tienen mucha inercia.
4. Están correlacionadas espacialmente, ya que los caudales entre ríos distintos no son siempre independientes.
5. Están fuertemente condicionadas por el fenómeno 'El Niño'/'La Niña'; más específicamente, es muy bien conocida la dependencia de los aportes con el valor del índice N3.4, un promedo espacial sobre una cierta región de la costa de Ecuador, de la temperatura de la superficie del mar [17, 16].

En lo que sigue se muestran resultados experimentales sobre las series de aportes históricos disponibles a la fecha, en donde se busca explotar y/o corroborar las propiedades asumidas a priori anteriormente.

## 3.1. Distribuciones marginales empíricas

Uno de los primeros aspectos a tener en cuenta es el rango de valores que toman las series, y con qué frecuencia lo hacen. Esto podemos verlo mediante las distribuciones marginales empíricas. La figura 3.1 muestra los resultados de estimar dichas distribuciones para las tres series disponibles; puede verse que todas se ajustan muy bien a la distribución Gamma o la Weibull (ambas son muy similares).

## 3.2. Estacionareidad

Los aportes hídricos derivan de fenómenos climáticos en la región comprendida por las cuencas de los rios correspondientes. Naturalmente esto hace que la distribución de los aportes varíe con la estación del año. Esto, que ya es explotado en modelos como el CEGH [2], puede verificarse en la figura 3.2, en donde se muestran las distribuciones empíricas de las series para distintas

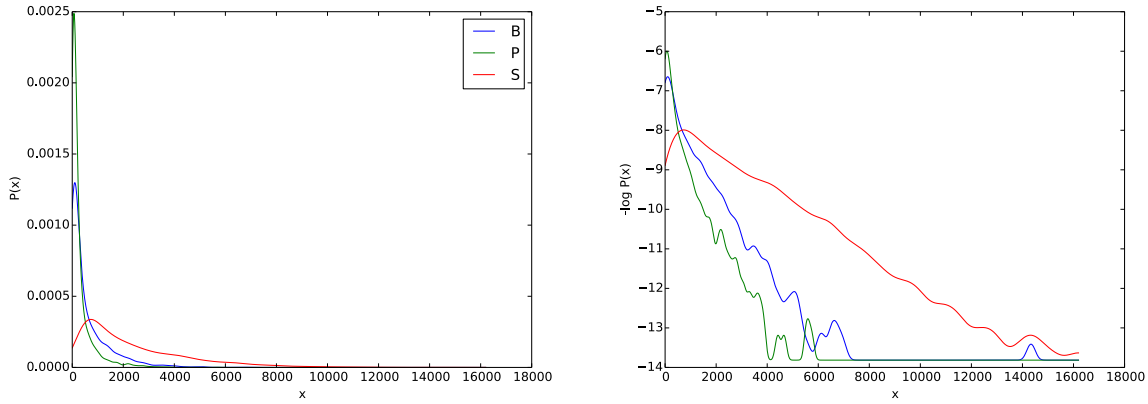


Figura 3.1: Distribuciones marginales de las series Bonete, Palmar y Salto entre 1909 y 2009 inclusive.

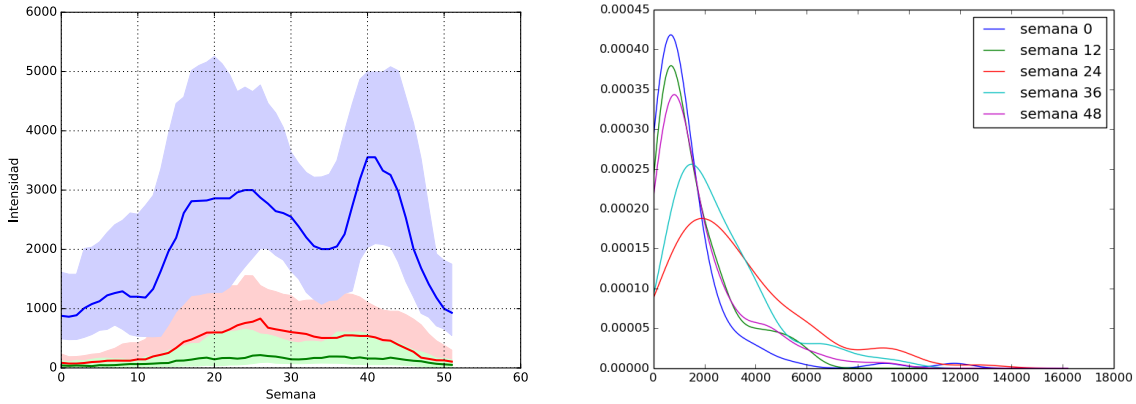


Figura 3.2: Izq.: percentiles 25-50-75 de Bonete, Palmar y Salto en función de la semana del año. Der.: distribución empírica de aportes a Salto para distintas semanas del año. Puede observarse la clara dependencia de todas estas estadísticas en función de la estación.

épocas del año, identificadas por una semana de referencia. Al igual que el CEGH, los modelos a desarrollar más adelante hacen fuerte uso de esta característica.

## 3.3. Correlación temporal

Como prácticamente toda señal física, es esperable que las series de aportes tengan algún tipo de continuidad o correlación entre muestras adyacentes. Esto puede verse en la figura 3.3, en donde se muestran las distribuciones de los error de predicción de orden 0 para las series de Bonete y Palmar para épocas más o menos coincidentes con las cuatro estaciones del año.

## 3.4. Dependencia con el índice N3.4

El índice N3.4 es la principal medida asociada con el fenómeno 'El Niño' [17, 16]. Su valor corresponde a la temperatura superficial del agua en un punto específico de la costa de Perú. Esencialmente, valores bajos del N3.4 se correlacionan con períodos de sequía en la región,

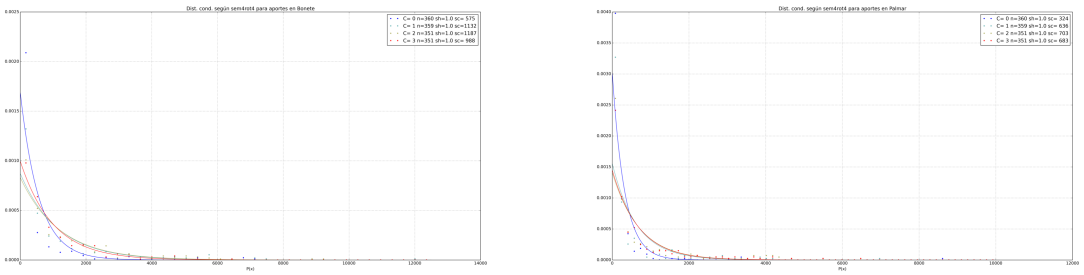

Figura 3.3: Distribución de errores de predicción de las series Bonete (izq.) y Palmar (der.) para distintas épocas del año. El hecho de que sean distribuciones altamente concentradas en 0 muestra la alta correlación que existe entre muestras adyacentes.

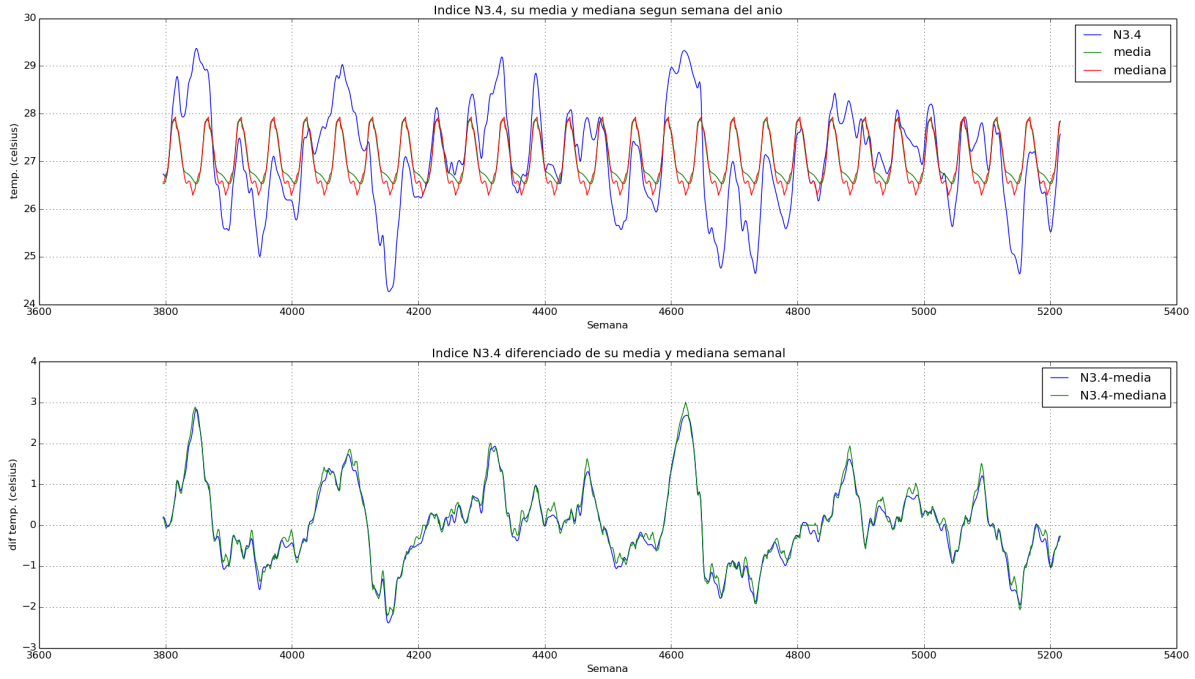


Figura 3.4: Curvas media y mediana anuales, y desviación de N3.4.

mientras que valores altos están asociados a períodos de mayores precipitaciones. Debido a la inercia de los fenómenos atmosféricos y la distancia entre el punto de medición y nuestra región, es de esperarse que haya un importante retardo entre las *desviaciones* del N3.4 y la manifestación de su efecto en nuestra región; por desviaciones nos referimos a la diferencia de la serie N3.4 respecto a su curva mediana anual, como puede verse en la figure 3.4. En trabajos previos como [15], dicho retardo se consideró variable según el mes del año. En nuestro caso, considerando que tal nivel de detalle es difícilmente capturable con la cantidad de datos disponibles, preferimos considerar dicho retardo constante a lo largo del año. El utilizar este índice fue una de las principales mejoras del CEGH respecto a su versión original [3]. En las figuras 3.5, 3.6 3.7 se muestra claramente la influencia que dicha variable tiene en los aportes a Bonete, Palmar y Salto. Naturalmente, esta información será incorporada en los modelos que nosotros desarrollemos en adelante. En particular, la figura 3.7 muestra que discriminar este valor en unos pocos rangos ya es suficiente para capturar su efecto en las series de aportes. Esto es consistente con los resultados reportados en [5], donde se trabaja con los terciles del N3.4 como factor condicionante.

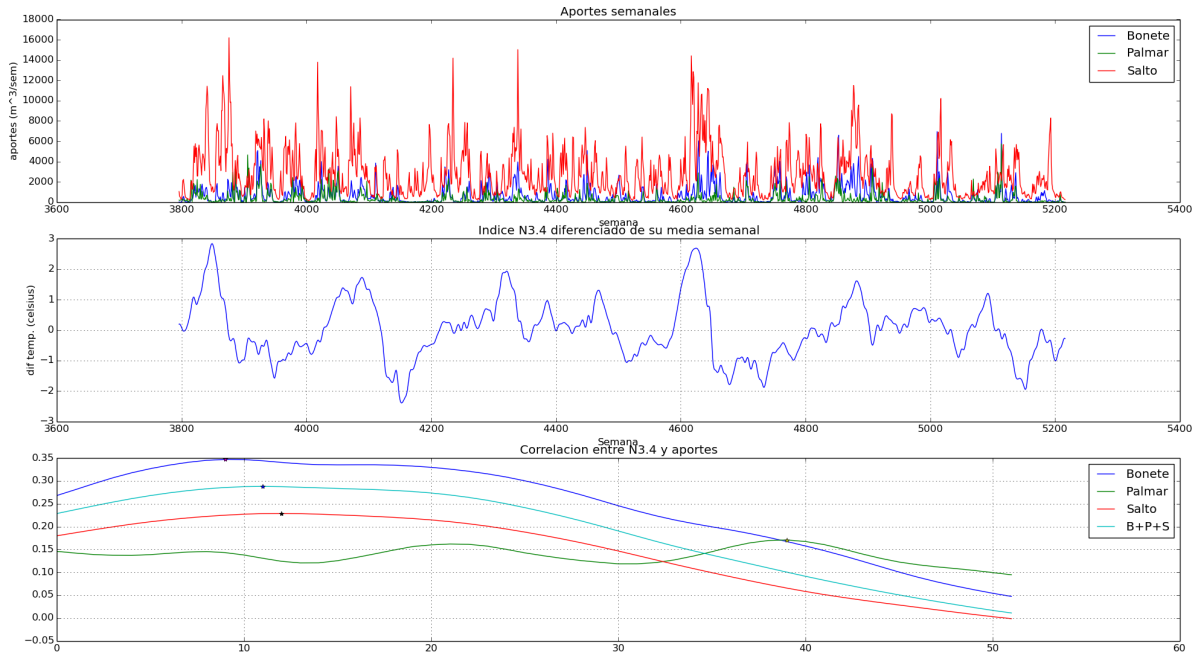


Figura 3.5: arriba) Series históricas de Bonete, Palmar y Salto; medio) desviación del N3.4; abajo) correlación entre cada una de las series de aportes y N3.4. Puede verse que el efecto de N3.4 sobre las series tiene un retraso entre 10 y 12 semanas.

## 3.5. Modelado de la serie N3.4

Más allá de la influencia del N3.4 sobre las series de aportes, interesa también modelar y predecir la evolución del propio índice N3.4. Para empezar, es claro que, siendo una temperatura medida en un punto fijo de la Tierra, el N3.4 tiene una fuerte componente estacional. Tal es así, que junto con el valor del N3.4, el índice publicado por la NOAA [1] viene acompañado de la *anomalía*, es decir, la desviación mensual del N3.4 respecto de su media histórica mensual. En adelante, en este proyecto, cuando hablemos del N3.4, hablaramos siempre de la anomalía y no del valor absoluto del N3.4

La anomalía del N3.4 es bastante suave, lo que sugiere que puede modelarse mediante un modelo autoregresivo. La figura 3.8 muestra que alcanza con un modelo de orden 2 para capturar razonablemente la dinámica de esta serie.

## 3.6. Dependencias entre series

Ya vimos cómo el comportamiento de las series de aportes depende de la época del año y el índice N3.4. Lo que resta explorar es la correlación que existe entre las series. En el caso de Bonete y Palmar esto es una consecuencia directa de que una (Palmar) está aguas abajo de la otra (Bonete). La figure 3.9 muestra esta correlación en escala logarítmica.

[1] http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices

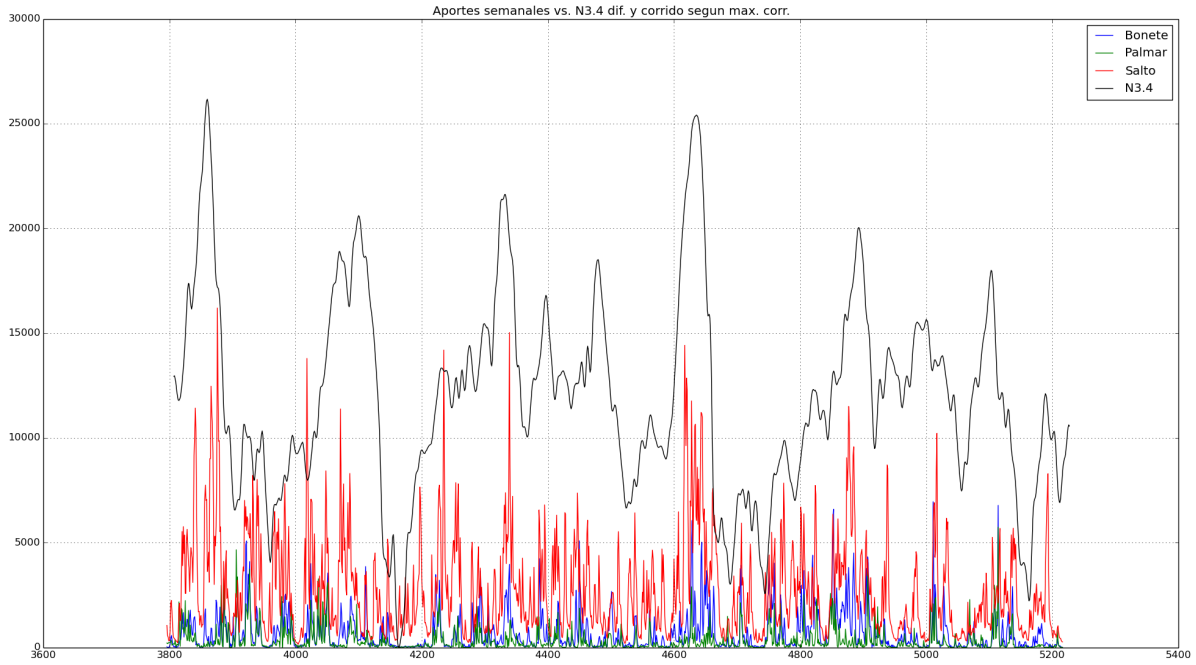


Figura 3.6: Series históricas de Bonete, Palmar y Salto y desviación N3.4 corregidas por desfasaje medio (11 semanas).

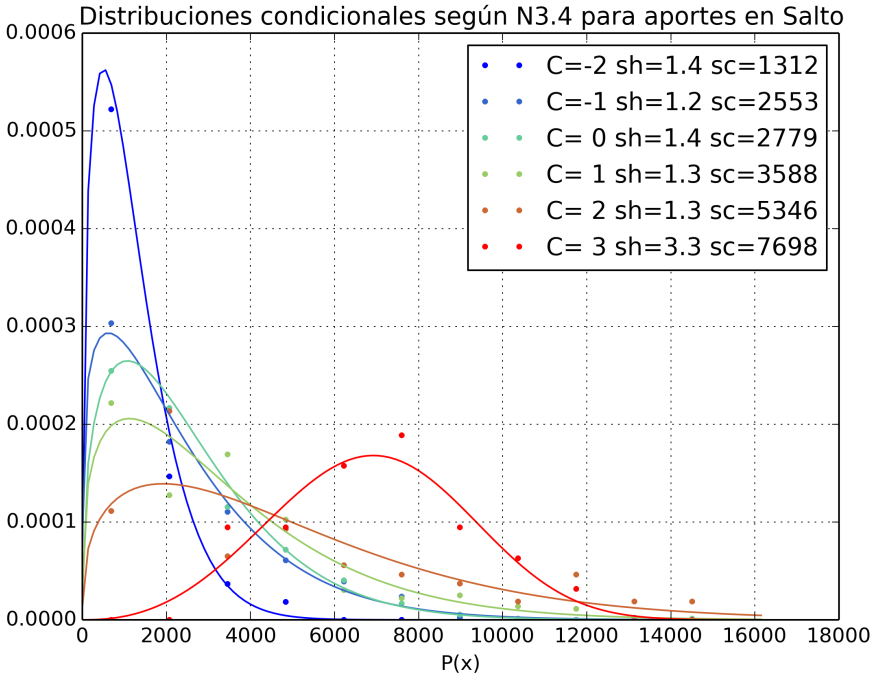


Figura 3.7: Distribución empírica de Salto condicionada a seis distintos rangos de valores de N3.4. Lo que puede verse, además de la fuerte dependencia de la distribución de Salto con el N3.4, es que las mayores diferencias se dan en los extremos, siendo menor la diferencia para los cuatro rangos contiguos centrales.


Figura 3.8: Memoria del N3.4. Lo que se muestra arriba es el error de predicción según el orden de modelo AR utilizado para predecirlo; de dicha gráfica se observa que tan sólo dos muestras pasadas son suficiente para capturar buena parte de la correlación de esta serie.

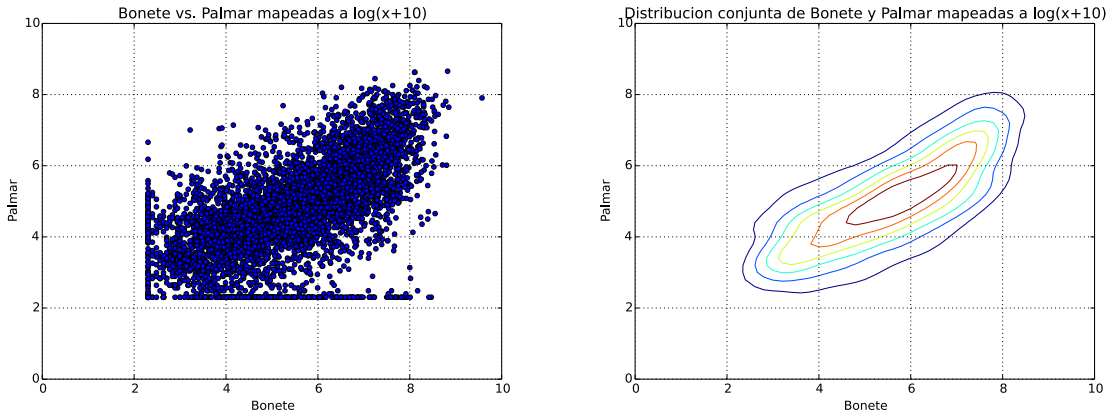


Figura 3.9: Distribución conjunta de series Bonete y Palmar en escala logarítmica. Los ejes principales son curiosamente $(1, 1)$ y $(1, -1)$, lo cual va de acuerdo con la fuerte correlación existente entre ambas series.

# 4 El modelo CEGH

## 4.1. Fundamentos

El objetivo del CEGH (Correlación en Espacio Gaussiano de Histogramas), a los efectos de su uso en el SimSEE, es proveer el modelo necesario para generar trayectorias del estado del sistema lo más realistas posibles, de modo de poder tomar decisiones a futuro en base a ellas.

El modelo propuesto en [4] busca capturar las siguientes características de la señal:

1. Correlación espacial (entre series)
2. Correlación temporal (intra series)
3. Distribución marginal de las series (figura 3.1)

La idea básica es transformar las series de aportes (no estacionaria, no Gaussiana) en procesos Gaussianos estacionarios para luego capturar las correlaciones en espacio Gaussiano mediante modelos autoregresivos.

## 4.2. Transformación a procesos Gaussianos

Sea $F(V)$ la distribución cumulativa asociada a una variable aleatoria $V$. Para llevar $V$ a una variable $Y$ Gaussiana basta con aplicar la transformación

$$Y = \phi(V) = G^{-1}(F(V))$$

donde $G^{-1}$ es la inversa de la distribución cumulativa Gaussiana de media 0 y varianza 1,

$$G(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2}\pi\sigma} e^{-\epsilon^2/2} d\epsilon.$$

En SimSEE, $F(V)$ es estimada mediante la distribución cumulativa empírica (ECDF). El CEGH incorpora la estacionareidad considerando que la distribución anterior depende de la semana del año $F(V|w), w = 0, \ldots, \tau - 1$, con $\tau = 52$ la cantidad de semanas por año. De esta manera se tienen entonces $\tau$ transformaciones

$$\{\phi_w(V) : w = 0, \ldots, \tau - 1\} \tag{4.1}$$

de la forma $\phi_w(v) = G^{-1}(\hat{F}(v|w))$ con

$$\hat{F}(v|w) = \frac{|\{V_{w+r\tau} \leq v : 0 \leq w + r\tau \leq n\}|}{|\{0 \leq w + r\tau \leq n\}|}, \tag{4.2}$$

donde $|\cdot|$ denota tamaño de un conjunto. En la ecuación (4.2), $\hat{F}_w(v)$ no es otra cosa que la proporción de veces que el valor observado en la $w$-ésima semana del año estuvo por debajo de $v$ en la serie de datos disponible, que abarca $n$ semanas. A las 52 transformaciones anteriores se les denomina "lentes".

El CEGH construye un grupo de 52 lentes para cada una de las series a analizar por separado. En nuestro caso, los aportes semanales de Bonete, Palmar, Salto, y (la anomalía de) el índice N3.4, a las que denominaremos de aquí en más como $V^b$, $V^p$, $V^s$ y $V^a$ respectivamente.

## 4.3. Redundancia temporal

Una vez transformadas las series, se captura la correlación temporal de cada una mediante un modelo lineal autoregresivo (AR) [11, Cap. 2]; esto es suficiente para caracterizar completamente a un proceso Gaussiano correlacionado. Para simular dicho proceso, se generan muestras pseudoaleatorias $\hat{\epsilon}_t \sim \mathcal{N}(0,1)$ y se las pasa por un filtro cuyos coeficientes $(a_1, a_2, \ldots, a_p)$ son los del modelo AR de orden $p$ estimado, produciendo una salida simulada $(\hat{v}_t)$ cuyo espectro de potencia corresponde con el de la serie original,

$$\hat{y}_t = \sum_{k=1}^{p} a_k \hat{y}_{t-k} + \sigma_e \hat{\epsilon}_t. \tag{4.3}$$

Los coeficientes $a$ que logran esto se obtienen resolviendo las ecuaciones de Yule-Walker,

$$\hat{a} = \mathbf{M}^{-1} r, \quad \sigma_e^2 = R(0) - \sum_{k=1}^{p} a_k R(k) \tag{4.4}$$

donde $R(k)$ es la autocorrelación empírica de la serie transformada $y_{1:n}$,

$$R(k) = \frac{1}{n-k} \sum_{j=k+1}^{n} y_j y_{j-k}, \tag{4.5}$$

la matriz Toeplitz $M = \{M_{ij} = \{R(i-j)\}\}$ y, $r = (R(1), R(2), \ldots, R(p))$.[1]

## 4.4. Redundancia espacial

Hasta ahora, todas las series del modelo CEGH son tratadas de forma independiente, cada una con su conjunto de lentes y coeficientes AR. Para incorporar la correlación existente entre series (lo que llamamos aquí "espacial"), se calculan las series de residuos $\epsilon_{1:n}^{b,p,s,a}$ de la aproximación AR en cada caso. Para bonete por ejemplo,

$$\epsilon_t^b = y_t^b - \sum_{k=1}^{p} a_k^b \hat{y}_{t-k}^b.$$

Una vez obtenidas las series de residuos $\epsilon_{1:\infty}$, $\epsilon_t = (\epsilon_t^a, \epsilon_t^b, \epsilon_t^p, \epsilon_t^s)$, se captura la correlación entre ellas mediante su matriz de autocovarianza empírica,

$$\hat{E} = \frac{1}{n-1} \sum_{t=1}^{n} \epsilon_t \epsilon_t^{\intercal}. \tag{4.6}$$

De la descomposición en valores singulares (SVD) de $\hat{E}$ se obtiene la matriz de *whitening* $\mathbf{D}$,

$$D = V \mathbf{\Sigma}^{1/2}, \; V \mathbf{\Sigma} V^T = \hat{E}. \tag{4.7}$$

La matriz $D$ permite generar muestras pseudoaleatorias de ruido $\hat{\epsilon} = (\hat{\epsilon}^a, \hat{\epsilon}^b, \hat{\epsilon}^p, \hat{\epsilon}^s)$ con la misma correlación empírica que las muestras reales $(\epsilon^a, \epsilon^b, \epsilon^p, \epsilon^s)$ a partir de un vector de muestras pseudoaleatorias $\mathcal{N}(0,1)$ independientes $\hat{\zeta} = (\hat{\zeta}_1, \hat{\zeta}_2, \hat{\zeta}_3, \hat{\zeta}_4)$ mediante $\hat{\epsilon} = D\hat{\zeta}$.

[1] Siendo $\mathbf{M}$ una matriz Toeplitz (las diagonales tienen valor constante), este problema lineal se puede resolver en forma eficiente utilizando, por ejemplo, el algoritmo de Levinson-Durbin [9, Cap 4].

## 4.5. Simulación en base al modelo

Para resumir, los parámetros aprendidos son

- Una distribución empírica cumulativa por semana, $\hat{F}(\cdot|w), w = 0, \ldots, \tau - 1$
- Un Modelo AR $(a^?, \sigma^?)$ por cada serie $? = a, b, p, s$
- Una matriz de correlación de ruido, $\mathbf{D} \in \mathbb{R}4\times4$

Una vez aprendidos estos parámetros en base a datos de entrenamiento, se simulan realizaciones del proceso mediante la generación pseudoaleatoria de secuencias de ruido blanco Gaussiano de media nula y varianza 1, las cuales son transformadas por $\mathbf{D}$ y luego inyectadas al filtro AR. Esto genera una secuencia simulada de variables aleatorias Gaussianas (correlacionadas) $\hat{y}_{1:n}$. La secuencia simulada final $\hat{v}_{1:n}$ se obtiene mediante $\hat{v}_t = \phi^{-1}_{t\,\%\tau}(y_i)$. Con esto último se logra reproducir los histogramas observados en la secuencia de entrenamiento.

## 4.6. Limitaciones del CEGH

El esquema de modelado y simulación descrito anteriormente es capaz de capturar correlaciones temporales y espaciales simultáneamente, y es muy fácil de entrenar. Sin embargo, se detecta en él una serie de problemas que es necesario tener en cuenta.

**Complejidad paramétrica** Uno de los primeros problemas del CEGH es su altísima cantidad de parámetros en comparación con la cantidad de datos disponibles para aprendizaje. La construcción de las "lentes" (4.1) implican estimar una distribución empírica cumulativa por cada una de las 52 semanas del año.

Es inmediatamente obvio que cualquier distribución empírica observada puede ser mapeada a una Gaussiana (o a cualquier otra) mediante la transformación ("lente") propuesta, pero, tiene sentido ajustarse tanto a los (escasos) datos empíricos?

En su versión original, y tal como fuera definido anteriormente en (4.2), el sobreajuste es extremo: sólo pueden generarse exactamente los valores que se dieron en el pasado (incluso tomándolos como enteros, los aportes toman valores entre 0 y 16000), con exactamente la misma frecuencia que ocurrieron en las semanas correspondientes. Teniéndose 100 años de historia, esto es 100 valores posibles por semana, habiendo ocurrido 1 vez cada uno en la mayoría de los casos (con excepción de los valores como 0, que ocurren muchas veces).

Este problema puntual, y sus efectos nocivos, fueron notados inmediatamente durante el desarrollo del CEGH. Actualmente se mitiga mediante una modificación en la estimación de las ECDF (4.2) donde se toman también valores ocurridos en semanas contiguas. Formalmente esto corresponde a un suavizado temporal de las ECDFs estimadas. De todos modos, la cantidad de parámetros sigue siendo enorme, y es el principal causante de sobreajuste del CEGH en su conjunto. El resto de los parámetros no aportan significativamente a esto.

**Limitaciones de la linealidad** Otro problema importante es que el CEGH está muy atado al concepto de llevar el problema a uno lineal, para así tratarlo mediante herramientas de modelado lineal de series, como los modelos autoregresivos. Esto fuerza un número de decisiones, y limita la expresividad del modelo. La consecuencia más importante de esto es en el caso multivariado: la única forma de capturar dependencia entre las distintas variables del sistema está dada por la transformación lineal del ruido inyectado a los modelos AR, $\epsilon = T(\zeta) := \zeta$, lo cual no permite expresar efectivamente dependencias no lineales entre ellas, como por ejemplo que el soporte de una esté acotada según la otra, algo que tiene sentido al modelar niveles de ríos cuyos afluentes dependen unos de otros, en particular si uno está aguas abajo del otro.

**Cuantificacion y linealidad** Como se mencionara en la sección 2.2, el SimSEE está fuertemente limitado por la cantidad posible de valores que el estado del sistema y las variables de entrada $(x, v)$ puedan tomar. En la práctica, esto impllica cuantificar los valores simulados por el CEGH, y sus estados, en una cantidad muy pequeña de valores. Para tener una idea, si se utiliza un modelo AR de orden 2, y se cuantifican los aportes y el N3.4 en 4 niveles cada uno, el espacio de estados resultante tiene $4^8 = 65536$ estados posibles!

Debido a esto, la versión actual del CEGH colapsa las tres series $b, p, s$ en una sóla variable mediante una proyección lineal, $u_t = p^\intercal(v_t^b, v_t^p, v_t^s)$. El vector $p$ utilizado actualmente es una estimación empírica del gradiente de la función de costo del problema de Bellman en el espacio original, luego de haber calculado la función de costo en una grilla muy fina de valores (ver detalles en [2]). Esto permite reducir el problema a una variable unidimiensional.

Sin embargo, incluso para una dimensión, el paso de cuantificación (llamémoslo $\Delta$) necesario para llevar una variable de aporte, que en la práctica puede considerarse continua, a una variable discreta de unos pocos valores, *es necesariamente muy grande*. En particular, $\Delta$ es muy grande en comparación con el nivel de ruido inyectado en la señal (la varianza empírica observada en los residuos de los modelos AR). Cabe recordar que para poder modelar un ruido de cuantificación como aditivo y aleatorio, es necesario que el paso de cuantificación sea muy inferior a la variación de la señal en cuestión.

La consecuencia final de todo esto es que el modelo resultante resulta altamente no lineal, y todas las transformaciones realizadas, así como la idea de capturar la correlación de la serie mediante un modelo autoregresivo, *dejan de ser válidas*.

Como respuesta a las anteriores observaciones, el presente proyecto tuvo como objetivo el desarrollo de modelos estadísticos de aportes que no dependieran de hipótesis de linealidad, y que tuvieran una cantidad de parámetros ajustables muy reducida, acorde a la cantidad de datos disponibles para su ajuste. Estos modelos serán detallados en la sección que viene.

# 5 Modelos propuestos

## 5.1. Modelos discretos – generalidades

La primera decisión importante sobre los modelos desarrollados en este proyecto deriva de las observaciones realizadas en la sección 4.6. Dado que los posibles niveles de aportes representables en la práctica son muy reducidos, tiene sentido trabajar con modelos estadísticos para variables aleatorias discretas.

La idea es hacer lo anterior preservando en lo posible la información condicional que aporta la serie N3.4, y la estación (semana del año) en la que se está, un factor que se nota claramente que modula fuertemente la distribución de las tres series de aportes (y el N3.4).

Debido a que la distribución empírica marginal de las tres series es altamente asimétrica (parece una Gamma), y para evitar mapeos, se opta aquí por realizar una cuantificación no uniforme de los datos. Para esto se parte a la serie en franjas según un conjunto prefijado de percentiles, siendo cada valor cuantificado el índice del percentil al que el valor original pertenece, y luego se reconstruye con el valor del percentil intermedio del rango correspondiente. Para fijar ejemplos, supongamos que las franjas son 0-25 (franja 0), 25-75 (1) y 75-100 (franja 2). un dato cae entre el percentil 25 y 75, se le asigna el índice 1. Al reconstruirlo, se utiliza el valor del percentil $(75 - 25)/2 = 50$.

Las series así discretizadas se modelan como procesos de Markov en cascada. Todos los modelos presentados en este capítulo mantienen esta idea general, introduciendo variantes en cómo se interrelacionan las series entre sí. A continuación presentamos los elementos comunes, es decir, el proceso de cuantificación y decuantificación, y el modelo de la serie N3.4, que se hace como primera etapa común a todo el resto.

Para simplificar ideas y análisis, los modelos presentados utilizan la misma cantidad de niveles de cuantificación $q$ para todas las series (salvo N3.4, que se trata de manera separada). Esto seguramente no sea lo óptimo (ver discusión en la sección 7. Claramente, es muy simple tanto en la programación como conceptualmente extender los modelos anteriores para que se utilicen distintos cantidades de niveles de cuantificación para las distintas series.

Un aspecto a resaltar de los modelos discretos es que son computacionalmente menos costosos que el CEGH, lo cual permite ejecutar más simulaciones por unidad de tiempo (y por ende mejorar las estimaciones de Monte Carlo).

Finalmente, los modelos propuestos tienen un conjunto muy reducido de parámetros entrenables, requiriendo muy poco almacenamiento y muy pocas muestras pasadas para ser ajustados con buena precisión.

### 5.1.1. Cuantificación

Recordemos que los índices temporales se especifican como subíndices, y que en el caso de variables multivariadas se especifica a uno de sus elementos con paréntesis rectos. Por ejemplo $u_3[5]$ sería el quinto elemento del valor de la variable $u$ en el tiempo $t = 3$.

La cuantificación de un dato cualquiera depende de un vector $u \in \mathbb{R}^q$ de umbrales de cuantificación, donde sólo se asume $u[i] < u[i+1]$ y otro vector $r \in \mathbb{R}^{q+1}$, con $u[i-1] < r[i] < u[i] <$

$r[i+1]$ de valores de reproducción. La cuantificación es un mapeo

$$\hat{v} = Q_u(v) : \mathcal{V} \to \{0, q-1\}$$

dado por

$$Q_u(v) = \arg\max_i\{v > u[i]\}.$$

La reconstrucción de un dato cuantificado está dada por

$$\bar{v} = Q_r^{-1}(\hat{v}) = r_{Q(\hat{v})}.$$

Sea $\eta_p(v), p = 0, 1, \ldots, 100$ el percentil $p$ de una serie $v_{1:n}$, que por definición es el valor del elemento $((p/100)\times n)$-ésimo en la serie $v_{1:n}$ ordenada de menor a mayor. Se tiene por ejemplo $\eta_0 = \text{mín}\{v_{1:n}\}$, $\eta_{100} = \text{máx}\{v_{1:n}\}$ y $\eta_{50} = \text{med}\{v_{1:n}\}$.

Dado un $q$ fijo, los umbrales de cuantificación $u$ y reproducción $r$ para $v_{1:n}$ están dados por

$$u[i] = \eta_{100\times 2i/2q}(v_{1:n}), \quad r[i] = \eta_{100\times(2i+1)/2q}(v_{1:n}),\ i = 0, 1, \ldots, q-1.$$

Por ejemplo para $q = 4$ tenemos

$$\begin{aligned} u &= \{0, \eta_{25}, \eta_{50}, \eta_{75}\} \\ r &= \{\eta_{12{,}5}, \eta_{37{,}5}, \eta_{62{,}5}, \eta_{87{,}5}\} \end{aligned}$$

La elección de los percentiles tiene varias ventajas. La más importante es que hace todo el modelado posterior *invariante a cualquier transformación monótona creciente* de las series (por ejemplo logaritmo, o restarle su media, o multiplicarla por algún valor positivo). Al ser rangos de percentiles del mismo ancho, la distribución de los valores cuantificados resulta aproximadamente uniforme. Esto a su vez ayuda a que los distintos estados definidos en base a los datos cuantificados tengan una buena representatividad.

**Periodicidad** Este aspecto se incorpora variando la cuantificación utilizada en cada semana $w = 0, \ldots, \tau - 1$, de modo que ahora el vector $u$ utilizado en la semana $w$ es $u_w$, y el de reconstrucción es $r_w$.

Dado que, incluso para valores pequeños de $q$, los datos históricos no son muchos, sigue siendo importante incorporar algún tipo de suavizado en las estimaciones de los umbrales; en la implementación de los modelos discretos se recurre a dos técnicas combinadas. Una es la misma que se utiliza en CEGH, es decir, para estimar los umbrales de la semana $w$ se recurre a datos no sólo de esa semana en el pasado, sino de todo el rango $[w - d, w + d]$ para un valor $d \geq 0$; actualmente $d = 2$. Luego, para suavizar las estimaciones, se utiliza *bootstrap* [10, Cap. 7] sobre el conjunto de datos resultante; la cantidad de muestreos bootstrap realizados actualmente es 10.

El procedimiento anterior es general y permite parametrizar la cuantificación en términos de una cantidad deseada de niveles $q$. Sin embargo, para la serie auxiliar N3.4, y en base a las observaciones realizadas en la sección 3.4, se fija de antemano $q = 3$ con los umbrales prefijados en:

$$\begin{aligned} u_w^a &= (0, \eta_{25}(v^a_{1:\tau:n}), \eta_{75}(v^a_{1:\tau:n}))^\intercal \\ r_w^a &= (\eta_{12{,}5}\,(v^a_{1:\tau:n}), \eta_{50}(v^a_{1:\tau:n}), \eta_{87{,}5}(v^a_{1:\tau:n}))^\intercal \end{aligned}$$

donde la notación $v^a_{1:\tau:n}$ indica la subsecuencia de $v^a_{1:n}$ tomada de a $\tau$, $(v^a_1, v^a_{1+\tau}, v^a_{1+2\tau}, \ldots)$.

La figura 5.1 muestra un ejemplo del procedimiento de cuantificación aplicado a las series históricas disponibles.

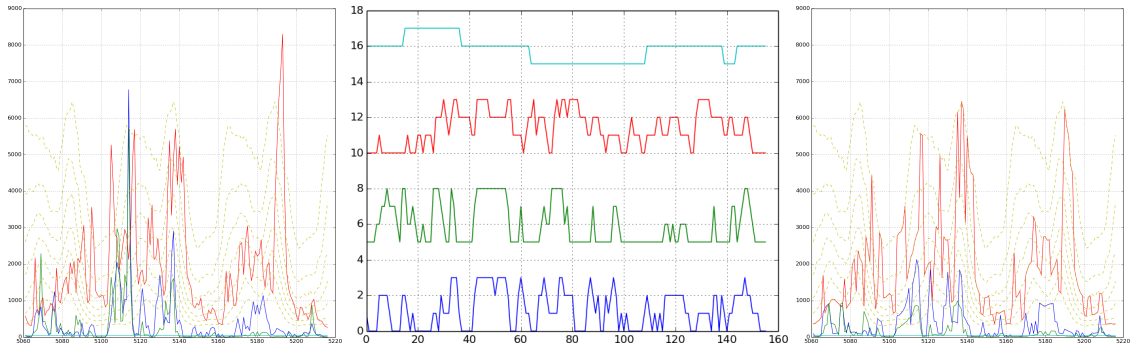

Figura 5.1: Proceso de cuantificación y reconstrucción para $q = 4$. Izquierda: serie original, las lineas amarillas marcan los percentiles según semana del año. Centro: series cuantificadas: N3.4 (celeste), Salto (rojo), Palmar (verde), Bonete (azul). Derecha: series reconstruidas.

### 5.1.2. Corrección de desfasaje del N3.4

Se sabe que el índice N3.4 tiene una influencia fuerte sobre las series Bonete, Palmar y Salto, pero ésta influencia tiene un retardo. De modo que para poder tomar en cuenta el valor del N3.4 en el condicionamiento de las series de aportes, tenemos primero que corregir ese retardo, que aquí llamaremos $\Delta$. (Como se mencionara en el capítulo 3, trabajos previos como [15] consideran que este retardo es una función periódica en el año. Aquí seguimos un camino simplificado).

La forma en que se hace aquí es buscando la correlación máxima entre cada serie y el N3.4. En congruencia con la justificación dada previamente para el uso de percentiles, se considera más robusto hacer esta correlación sobre las series cuantificadas en lugar de hacerlo sobre las series originales. Según los experimentos realizados sobre estas series, el retardo que mejor se ajusta a las tres series en su conjunto es $\Delta = 12$.

Notar que este parámetro afecta sólamente a la estimación del modelo, o bien a la simulación en caso de disponerse de la serie real N3.4. Si se simulan las cuatro series, se considera directamente que la simulación de $a$ está sincronizada con $b$, $p$ y $s$. Si la simulación es de menos de 10 semanas, se puede contar con los datos reales de N3.4 como condicionantes, y ahí aplicar el retardo correspondiente a cada serie.

### 5.1.3. Modelado de la serie N3.4

En adelante, denotaremos como $\tilde{v}$ al valor cuantificado de un dato originalmente continuo $v$. De acuerdo a lo descrito anteriormente tenemos que $\tilde{v^a} \in \{0, 1, 2\}$. En este modelo tomamos las dos muestras pasadas $\tilde{v}^a_{t-1}$ y $\tilde{v}^a_{t-2}$ como condicionantes de la distribución de la muestra $\tilde{v}^a_t$,

$$\tilde{V}^a_t \sim P(\tilde{V}^a_t = \tilde{v}^a_t | \tilde{V}^a_{t-1} = \tilde{v}^a_{t-1}, \tilde{V}^a_{t-2} = \tilde{v}^a_{t-2}),$$

o mas económicamente, $P(\tilde{v}^a_t | \tilde{v}^a_{t-1}, \tilde{v}^a_{t-2})$. El estado de este modelo está dado por $(\tilde{v}^a_{t-1}, \tilde{v}^a_{t-2})$. Representamos el estado como un índice mediante una función biyectiva que mapea el par de valores anteriores a un índice $x^a_t = 3\tilde{v}^a_{t-1} + \tilde{v}^a_{t-2}$. Ahora podemos escribir $P(\tilde{v}^a_t | \tilde{v}^a_{t-1}, \tilde{v}^a_{t-2}) = P(\tilde{v}^a_t | x^a_t)$.

Definimos ahora la matriz de transición de estados asociada al N3.4 como $\Pi^a \in [0, 1)^{9\times 3}$ con $\pi_{i,x} = P(\tilde{X}^a_t = i | x^a_t = x)$. Para completar el modelo de N3.4 ahora basta con estimar los elementos de $\Pi$. En este caso lo hacamos mediante el estimador de Krichevsky-Trofimov, que es como una versión suavizada del estimador de frecuencia básico,

$$\hat{\pi}^a_{i,x}(\tilde{v}^a_{1:n}) = \frac{n_{i|x} + 1/3}{n_x + 1} = \frac{1/3 + \sum_{j=0}^{n-1} 1(v^a_j = i)1(x^a_j = x)}{1 + \sum_{j=0}^{n-1} 1(x^a_j = x)} \tag{5.1}$$

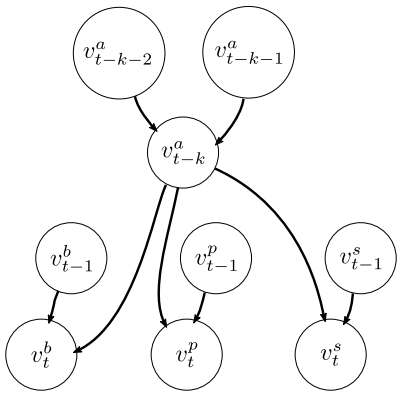


Figura 5.2: Esquema de dependencia estadística (tipo red bayesiana, donde las flechas indican dependencia estadística) para el modelo dicreto 1. La anomalía del N3.4 ($v^a_k$), retardada $k$ semanas, tiene un rol central en modular la distribución de las series de aportes.

donde $n_x$ es la cantidad de veces que el estado $x$ ocurrió en la serie $\tilde{v}^a_{1:n}$, y $n_{i|x}$ es la cantidad de veces que $\tilde{v}^a_{1:n}$ tomó el valor $i$ estando en el estado $x$.

## 5.2. Modelo discreto 1

Para las tres series $v^b$,$v^p$,$v^s$ (usaremos $v$ para referirnos a cualquiera de ellas), el factor condicionante en este modelo, es decir, el estado $x_t$, está dado por el valor retardado $k = 12$ muestras del N3.4, $\tilde{v}^a_{t-k}$, y el valor anterior de la serie cuantificada, $\tilde{v}_{t-1}$, $x_t = (\tilde{a}_{t-k}, \tilde{v}_{t-1})$. Dicho esto, y teniendo en cuenta que la anomalía N3.4 cuantificada $\tilde{v}^a$ toma 3 valores posibles, y que la serie cuantificada toma $q$ valores distintos, tenemos ahora $|\mathcal{X}| = 3q$ y $\Pi \in \mathbb{R}^{(3q)\times q}$. Salvo la conformación del estado $x$, la estimación es análoga al caso del N3.4. La estructura de dependencia estadística entre las variables anteriormente descrita se muestra en forma de red bayesiana en la figura 5.2.

### 5.2.1. Ajuste a datos

**Condiciones iniciales** Cantidad de niveles de cuantificación para Bonete, Palmar y Salto, $q$, parámetro de suavizado $d$, series $v^a_{1:n}$, $v^b_{1:n}$, $v^p_{1:n}$, $v^s_{1:n}$ de entrenamiento.

1. **Calcular umbrales** de cuantificación y reconstrucción. Para $\square = b, p, s$,

$$\left.\begin{array}{rcl} U^{\square}_w[i] & \leftarrow & \eta_{100\times 2i/2q}(v^{\square}_{1:n}) \\ R^{\square}_w[i] & \leftarrow & \eta_{100\times (2i+1)/2q}(v^{\square}_{1:n}) \end{array}\right\},\ i = 0, \ldots, q-1,\ w = 0, \ldots, \tau - 1$$

2. **Cuantifiar series.** Para $\square = b, p, s$,

$$\tilde{v}^{\square}_t = Q_{U^{\square}_w}(v^{\square}_t),\ w \leftarrow t\,\%\tau,\ t = 0, \ldots, n$$

3. **Calcular desfasaje óptimo con N3.4.**

$$k = \text{med}\{k^{\square} : \square = b, p, s\},\ k^{\square} = \arg\max_k\{\, c[k]\,\}, \quad c[k] = \sum_{t=-\infty}^{\infty} \tilde{v}^{\square}_t \tilde{v}^a_{t-k}$$

   (elementos fuera de rango son asumidos 0)

4. **Modelo N3.4.** estimar $\Pi^a$ usando (5.1) y el estado definido por

$$x^a_t \leftarrow 3\tilde{v}^a_{t-1} + \tilde{v}^a_{t-2}$$

5. **Modelo para los aportes.** para $\square = b, p, s$ estimar $\Pi^{\square}$ usando (5.1) y el estado definido por

$$x_t^{\square} \leftarrow q\tilde{v}_{j-k}^{a} + \tilde{v}_{j-1}^{\square}$$

### 5.2.2. Simulación

**Condiciones iniciales:** parámetros del modelo $\{\Pi^{\square}, U_w^{\square}, R_w^{\square} : \square \in \{a, b, p, s\}, w = 0, \ldots, \tau - 1\}$ y estado inicial dado por $(\tilde{v}_{-2-k}^{a}, \tilde{v}_{-1-k}^{a}, \tilde{v}_{-1}^{b}, \tilde{v}_{-1}^{p}, \tilde{v}_{-1}^{s})$

1. $t \leftarrow 0$, $\hat{v}_{-2-k}^{a} \leftarrow \tilde{v}_{-2-k}^{a}$, $\hat{v}_{-1-k}^{a} \leftarrow \tilde{v}_{-1-k}^{a}$, $\hat{v}_{-1}^{\square} \leftarrow \tilde{v}_{-1}^{\square}$,
2. $\hat{x}_{t-k}^{a} = 3\hat{v}_{t-k-2}^{a} + \hat{v}a_{t-k-1}$
3. Sortear $\hat{v}_{t-k}^{a}$ según la $\hat{x}_{t-k}^{a}$-ésima fila de $\Pi^{a}$, $\Pi[x_{t-k}^{a}, :]$
4. para $\square = b, p, s$:
   *a)* calcular estado $\hat{x}_t^{\square} = 3\hat{v}_{t-1}^{\square} + \hat{v}_{t-k}^{a}$
   *b)* sortear $\hat{v}_t^{\square}$ según $\Pi^{\square}[x_t^{\square}, :]$
   *c)* $\bar{v}_t^{\square} = (Q^{\square})^{-1}(\hat{v}_t^{\square})$
5. $t \leftarrow t + 1$ y volver a 2

### 5.2.3. Complejidad paramétrica

- Vectores semanales de cuantificación y reproducción para cada una de las series, $U_w^{a,b,p,s}$, $R_w^{a,b,p,s}, w = 0, \ldots, \tau - 1$. La primera aporta $52{\times}3$ parámetros, y las otras tres $52{\times}q$ parámetros.
- Matrices de transición de estados $\Pi^{a,b,p,s}$. La primera aporta 6 parámetros (la última columna es redundante) y la segunda $3{\times}q(q-1)$ parámetros.

**Resultados** La figura 5.3 muestra un ejemplo de simulación realizada utilizando la primera versión del modelo anteriormente descrita para $q = 4$. Se puede observar que la dinámica temporal se captura bastante bien, sobre todo en los períodos de sequía; no tan así en los picos. Los niveles de reconstrucción no parecen ser suficientes para reproducir la proporción de aportes entre picos y valles.

## 5.3. Modelo discreto 3

Esta versión tiene dos diferencias respecto a la anterior:

1. Para capturar la fuerte correlación entre Palmar y Bonete, siendo que Bonete está aguas arriba de Palmar, se redefine el estado de Palmar en el tiempo $t$ como

$$x_t^{p} = q \times v_{t-1}^{p} + v_{t-1}^{b}.$$

2. Asumiendo que toda la estacionareidad del N3.4 ya es capturada por su curva media, se asume que la anomalía $v^{a}$ es estacionaria, por lo cual sus niveles de cuantificación pasan a ser los mismos para todas las semanas del año.

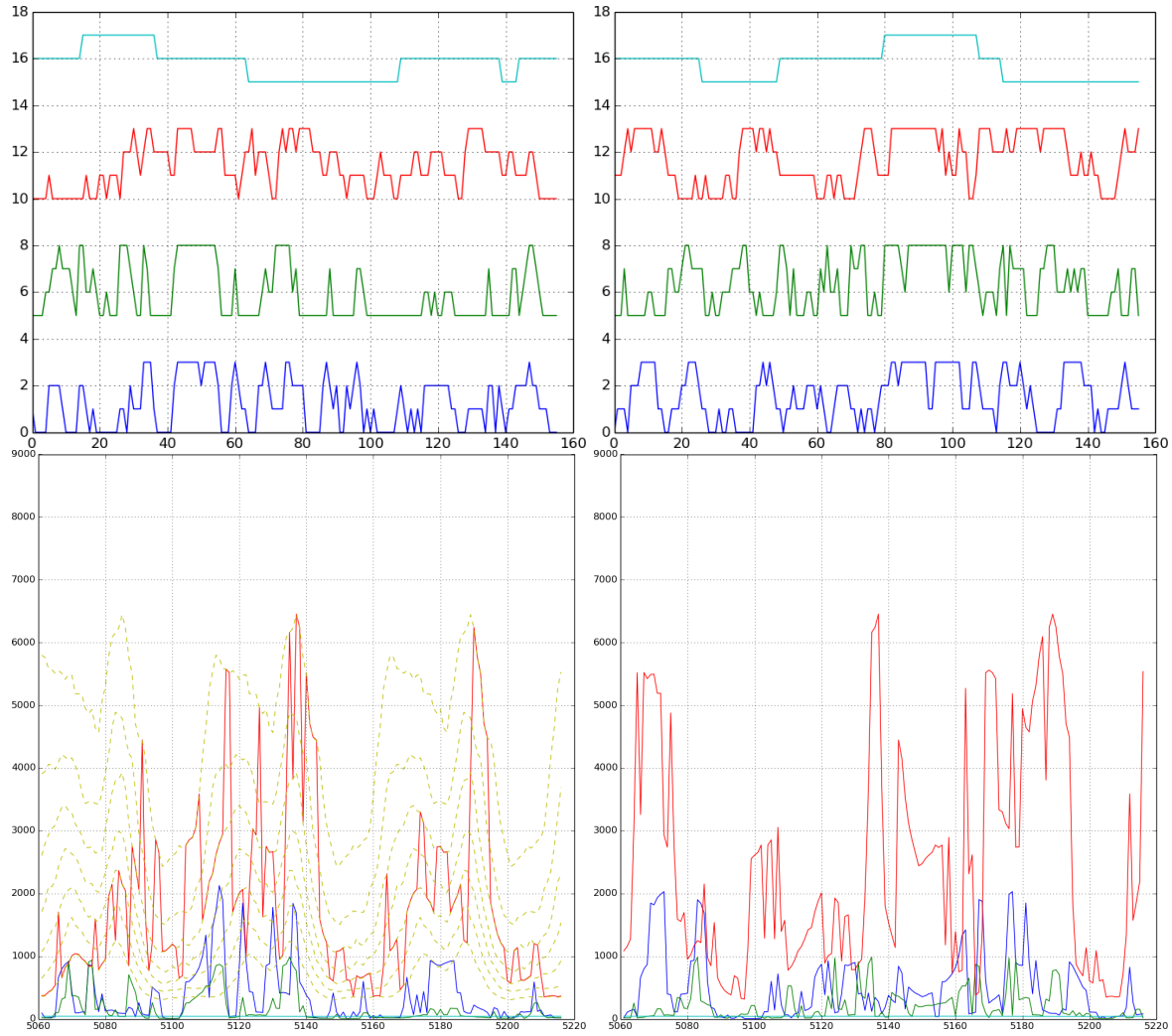

Figura 5.3: Simulación. Izq.: series reales. Der: series simulados. Arr.: series cuantificadas. Abj.: series reconstruidas.

## 5.4. Modelo discreto 4

Esta versión se basa en la figura 3.9. En ella se observa que, en escala logarítmica, las series Bonete y Palmar se descomponen en componentes principales que coinciden casi de manera exacta con las diagonales $Y = Z$ e $Y = -Z$ respectivamente, por lo que es mucho más fácil transformarlas mediante suma y resta. La transformación en cuestión es

$$V^y = \log_{10}(V^b + 10) - \mu_b + \log_{10}(V^p + 10) - \mu_p,$$
$$V^z = \log_{10}(V^b + 10) - \mu_b - (\log_{10}(V^p + 10) - \mu_p),$$

donde

$$\mu_b = (1/n)\sum_{t=1}^{n} \log_{10}(v_t^b + 10), \quad \mu_p = (1/n)\sum_{t=1}^{n} \log_{10}(v_t^p + 10)$$

son los valores medios de ambas series luego del logaritmo.

Hecho este cambio, la series $Y$ e $Z$ se tratan de manera idéntica a como se hacía con Bonete y Palmar, dependiendo de sí mismas y de la anomalía de N3.4 retardada $k$ semanas:

$$x_t^y = qv_{t-k}^a + v_{t-1}^y, \quad x_t^z = qv_{t-k}^a + v_{t-1}^z$$

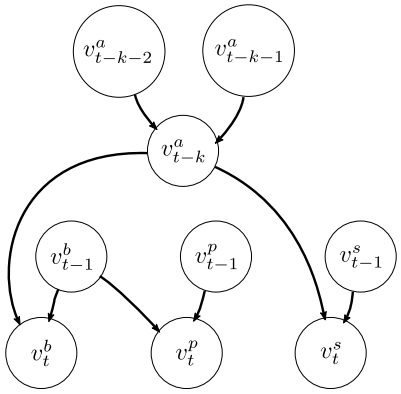


Figura 5.4: Esquema de dependencia estadística para el modelo discreto 3. La única diferencia respecto al modelo 1 (figura 5.2) es la sustitución del valor de Bonete anterior $v^b_{t-1}$ como determinante del valor actual de Palmar, $v^p_t$ en lugar del N3.4.

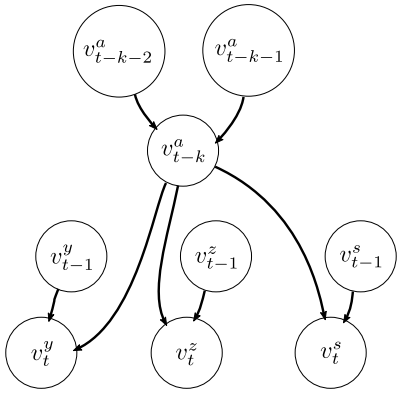


Figura 5.5: Esquema de dependencia estadística para el modelo discreto 4. La estructura de dependencias es análoga al modelo 1, pero aplicada sobre el cambio de variables $(v^b, v^p) \to (v^y, v^z)$.

.

El ajuste a datos y la simulación son análogos a los del modelo 1, haciendo y deshaciendo los cambios de coordenadas al principio y al final respectivamente.

### 5.4.1. Ajuste a datos

**Condiciones iniciales:** Cantidad de niveles de cuantificación para Bonete, Palmar y Salto, $q$, parámetro de suavizado $d$, series $v^a_{1:n}$, $v^b_{1:n}$, $v^p_{1:n}$, $v^s_{1:n}$ de entrenamiento.

1. Cambio de variables $(B, P) \to (Y, Z)$,

$$\begin{aligned}\mu_b &= (1/n)\sum_{t=1}^{n} \log_{10}(v^b_t + 10)\\ \mu_p &= (1/n)\sum_{t=1}^{n} \log_{10}(v^p_t + 10)\\ v^y_t &= \log_{10}(v^b_t + 10) - \mu_b + \log_{10}(v^p_t + 10) - \mu_p,\\ v^z_t &= \log_{10}(v^b_t + 10) - \mu_b - (\log_{10}(v^p_t + 10) - \mu_p),\end{aligned}$$

2. **Calcular Umbrales de cuantificación y reconstrucción.** Para $\square = b, y, z$,

$$\left.\begin{array}{rcl} U^{\square}_w[i] & \leftarrow & \eta_{100\times 2i/2q}(v^{\square}_{1:n}) \\ R^{\square}_w[i] & \leftarrow & \eta_{100\times (2i+1)/2q}(v^{\square}_{1:n}) \end{array}\right\},\ i = 0, \ldots, q-1,\ w = 0, \ldots, \tau - 1$$

3. **Cuantificación de las series.** Para $\square = a, b, p, s$,

$$\tilde{v}^{\square}_t = Q_{U^{\square}_w}(v^{\square}_t),\ w \leftarrow t\,\%\tau,\ t = 0, \ldots, n$$

4. **Calcular desfasaje óptimo con N3.4**.

$$k = \text{med}\{k^s, k^y, k^z\},\ k^{\square} = \arg\max_k\{\, c[k]\,\}, \quad c[k] = \sum_{t=-\infty}^{\infty} \tilde{v}^{\square}_t \tilde{v}^a_{t-k}$$

(elementos fuera de rango son asumidos 0).

5. **Modelo N3.4** estimar $\Pi^a$ usando (5.1) y el estado definido por

$$x_t \leftarrow 3\tilde{v}^a_{t-1} + \tilde{v}^a_{t-2}$$

6. **Modelo de aportes**. Para $\square = y, z, s$ estimar $\Pi^{\square}$ usando (5.1) y el estado definido por

$$x^{\square}_t \leftarrow q\tilde{v}^a_{j-k} + \tilde{v}^{\square}_{j-1}.$$

### 5.4.2. Simulación

**Condiciones iniciales** parámetros del modelo $\{\Pi^{\square}, U^{\square}_w, R^{\square}_w : \square \in \{a, \mathbf{y}, \mathbf{z}, s\}, w = 0, \ldots, \tau - 1\}$ y estado inicial dado por $(\tilde{v}^a_{-2-k}, \tilde{v}^a_{-1-k}, \tilde{v}^y_{-1}, \tilde{v}^z_{-1}, \tilde{v}^s_{-1})$

1. **Actualizar estado.** $t \leftarrow 0$, $\hat{v}^a_{-2-k} \leftarrow \tilde{v}^a_{-2-k}$, $\hat{v}^a_{-1-k} \leftarrow \tilde{v}^a_{-1-k}$, $\hat{v}^{\square}_{-1} \leftarrow \tilde{v}^{\square}_{-1}$,
2. $\hat{s}^a_{t-k} = 3\hat{v}^a_{t-k-2} + \hat{v}a_{t-k-1}$
3. **Sortear N3.4** $\hat{v}^a_{t-k}$ según la $\hat{x}^a_{t-k}$-ésima fila de $\Pi^a$, $\Pi[x^a_{t-k}, :]$

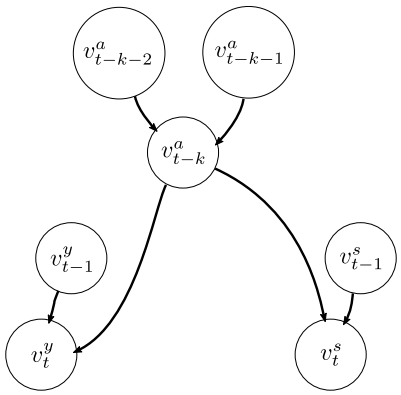


Figura 5.6: Esquema de dependencia estadística para el modelo discreto 4s1. La diferencia con el modelo 4 (5.5) es la eliminación total de la variable $v^z$; tanto Bonete como Palmar toman el mismo valor derivado de $v^y$.

4. **Simular aportes**. Para $\square = b, p, s$:

   *a*) calcular estado $\hat{x}_t^{\square} = 3\hat{v}_{t-1}^{\square} + \hat{v}_{t-k}^{a}$

   *b*) sortear $\hat{v}_t^{\square}$ según $\Pi^{\square}[x_{t-1}^{\square}, :]$

   *c*) $\bar{v}_t^{\square} = (Q^{\square})^{-1}(\hat{v}_t^{\square})$

5. **Deshacer cambio de variables.**

$$\begin{aligned} \hat{v}_t^b &\leftarrow 10^{\mu_b + (\hat{v}_t^y + \hat{v}_t^z)/2} - 10 \\ \hat{v}_t^p &\leftarrow 10^{\mu_b + (\hat{v}_t^y - \hat{v}_t^z)/2} - 10 \end{aligned}$$

6. $t \leftarrow t + 1$ y volver a 2

## 5.5. Modelo discreto 4s1

Esta variante busca reducir más aún el espacio de estados, eliminando por completo la variable Z del estado. En simulación, Bonete y Palmar es ambas toman el mismo valor. Esto reduce el espacio de estados por un factor $3q$.

## 5.6. Modelo discreto 4s2

Otra simplificación del Modelo 4 en donde la variable transformada Z sólo depende del N3.4 y no de su propio pasado, es decir

$$s_t^z = v_{t-k}^a.$$

Esto reduce el espacio de estados por un factor de $q$.

## 5.7. Modelo en espacio de variables de estado – GSSM

Una familia de modelos propuesta inicialmente en el proyecto fue la de los modelos de espacio en variables de estado (SSM - Space State Models) [8]. La representación general de un modelo no lineal - no gaussiano puede escribirse como

$$Y_t \sim p(Y_t|X_t) \tag{5.2}$$

$$X_{t+1} \sim p(X_{t+1}|X_t) \tag{5.3}$$

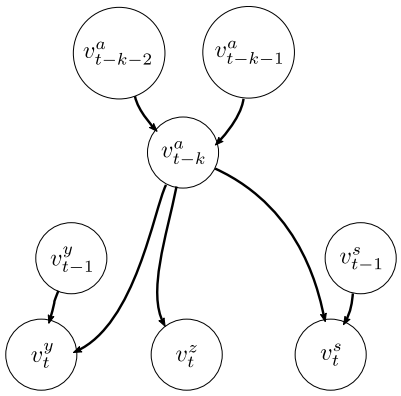


Figura 5.7: Esquema de dependencia estadística para el modelo discreto 4s2. En este caso la variable $v^z$ sólo depende del N3.4, $v^a_{t-k}$; ya no de su propio pasado.

donde $Y_t$ representa la variable observable que queremos modelar (en este caso los aportes a los embalses) y $X_t$ representa el estado del sistema (aquí estado del sistema hace referencia al subsistema que modela el fenómeno de aportes y no al sistema eléctrico completo (ver sección 2.2) del cual los aportes forman parte. La primera ecuación hace explícita la dependencia de la variable observable $Y_t$ con el estado $X_t$ a través de la distribución condicional $p(Y_t|X_t)$. La segunda modela la evolución del estado del sistema en el siguiente instante de tiempo $X_{t+1}$ dado el estado actual $X_t$ a través de la distribución condicional $p(X_{t+1}|X_t)$. La no linealidad - no gaussianidad del modelo deriva de la elección de la distribución en cada ecuación.

### 5.7.1. Motivación

La motivación de su uso tuvo dos razones principales. Primero, el modelo permite identificar variables explicativas que modelen fenómenos conocidos o que se sabe tienen efectos sobre los aportes. Por ejemplo, si el estado $X_t$ fuese una variable multidimensional, una de sus componentes puede modelar el efecto del Niño. Además, la introducción de otros efectos puede realizarse de forma natural en la representación del estado. También, la no linealidad - no gaussianidad puede introducirse de manera selectiva en parte de las variables de estado o en la variable observable, lo cual permite modelar de forma lineal - gaussiana los efectos para los que sí es aceptable hacerlo. Segundo, en la sección 2.2 se vio que el estado del sistema eléctrico de Uruguay se modela a través de un espacio $\mathcal{X}$, el cual incluye por ejemplo el nivel de los embalses en las represas hidroeléctricas, en particular, dado el estado actual del sistema en un instante de tiempo $x_t$, su evolución es determinada por la *función de transición de estado* $f(\cdot)$ (2.1). El modelo de aportes propuesto (5.2) y (5.3) hace explicito el estado del subsistema que modela los aportes y puede considerarse directamente un subespacio de $\mathcal{X}$, en este caso la evolución del estado de los aportes es determinada por (5.3).

Además de las razones anteriores, los modelos de espacio en variables de estados permiten representar familias de modelos simples, como los modelos autoregresivos y los modelos de espacio de estado discretos (ej. modelos Markov discretos), hasta modelos no lineales - no gaussianos (discretos y continuos), lo cual brinda una flexibilidad importante.

### 5.7.2. Modelo

De acuerdo a lo expuesto en la sección 3 para las distribuciones marginales de los aportes (fig. 3.1) es razonable asumir que siguen una distribución de la familia de las exponenciales (por ejemplo la distribución Gamma). La distribución Gamma es una distribución de probabilidad continua dependiente de dos parámetros, el parámetro de forma $k$ y de escala $\Theta$ (existen otras parametrizaciones como el factor de forma $\alpha$ y la tasa o cadencia $\beta$, pero no las consideraremos

para este modelo). Denotaremos a una variable aleatoria $X$ que sigue una distribución Gamma como

$$X \sim \Gamma(k, \Theta) = Gamma(k, \theta)$$

La función de densidad de probabilidad $p(x; k, \Theta)$ para la parametrización utilizada es

$$p(x; k, \Theta) = \frac{x^{k-1} e^{-\frac{x}{\Theta}}}{\Theta^k \Gamma(k)} \quad x \geq 0 \;\; \text{y} \;\; k, \Theta > 0$$

donde $\Gamma(k) = \int_0^\infty x^{z-1} e^{-x} dx$ es la función gamma evaluada en $k$.

En base esta observación se propuso utilizar la distribución Gamma para la ecuación de observación (5.2), para el estado asumiremos un modelo sencillo (Lineal y Gaussiano). El modelo inicial propuesto puede escribirse como,

$$Y_t \sim \Gamma(Y_t | k_t, \Theta_t), \qquad \Theta_t = Z_t \alpha_t \tag{5.4}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \qquad \eta_t \sim N(0, Q_t) \tag{5.5}$$

donde $\Theta_t$ representa el estado del sistema, (5.5) modela su evolución en forma lineal y gaussiana, $\Gamma(k, \Theta)$ es la distribución Gamma de parámetros $k$ y $\Theta$ (factor de forma y escala respectivamente) y $\eta_t$ serialmente independiente (i.e. $\eta_t$ y $\eta_s$ independientes $\forall\, t \neq s$) (un análisis detallado de esta familia de modelos puede verse en la sección 9.2 de [8]). De aquí en más llamaremos a este modelo GSSM (Gamma - State Space Model)

A continuación daremos los detalles de la heurística seguida para determinar completamente el modelo propuesto (i.e. $k_t$ y las matrices $Z_t$, $T_t$, $R_t$ y $Q_t$ $\forall\, t$), veremos también que el modelo puede mejorarse analizando algunas variantes propuestas, sin embargo, el modelo final utilizado corresponde a una variante simple de las que se mencionan. La razón principal de esta elección se debe a que el modelo presentado, al igual que el CEGH, es un modelo continuo, en particular las variables de estado utilizadas ($\Theta_t$) varían en un rango continuo, y tal como se mencionó en la sección 4.6 los niveles de cuantificación en la optimización de SimSEE son una limitante importante lo cual tiene como consecuencia directa que las mejoras que pueden conseguirse en base a mejorar el modelo se ven atenuadas por los niveles de cuantificación requeridos.

### 5.7.3. Estructura y ajuste a datos

La primer observación sobre la estructura del modelo es que $Y_t$ representa los aportes a los embalses de Palmar, Salto Grande y Bonete, por lo que $Y_t = (Y_t[0], Y_t[1], Y_t[2])$, en una primera versión del modelo GSSM se propuso utilizar una variable de estado por cada variable observable (por cada aporte) entonces $\Theta_t = (\Theta_t[0], \Theta_t[1], \Theta_t[2])$ y cada variable de estado $\Theta_t[i]$ se estima directamente de su respectiva serie de aportes.

Aquí puede introducirse una variante del modelo (no realizada en este trabajo), la serie del Niño puede modelarse por ejemplo con un modelo autoregresivo y determinar $\Theta_t$ a partir de ella. Otra variante podría utilizar directamente algún índice que represente al Niño como variable de estado (por ej. el índice N3.4) y nuevamente modelar $\Theta_t$ a partir de ella.

En la sección 3.4 de [8] se muestra que la familia de modelos autoregresivos puede modelarse con una estructura lineal como en (5.5) con $T_t = T$, $R_t = R$ y $Q_t = Q$ matrices invariantes en el tiempo, en base a esta observación es que inicialmente el modelo GSSM utiliza $T_t = T$, $R_t = R$ y $Q_t = Q$ matrices invariantes en el tiempo.

Por simplicidad $Z_t = Z = I$, la matriz identidad, y el factor de forma de cada distribución Gamma se toma invariante en el tiempo, $k_t = k = (k[1], k[2], k[3])$ y se usan sus estimadores MLE (directamente de cada serie de aporte) para determinarlos.

Hasta aquí el modelo GSSM puede reescribirse como

$$Y_t \sim \Gamma(Y_t | k, \Theta_t) \tag{5.6}$$

$$\Theta_{t+1} = T\Theta_t + R\eta_t, \qquad \eta_t \sim N(0, Q) \tag{5.7}$$

Para definir las matrices $T$, $R$ y $Q$ se sigue el método AVME (*Approximate via mode estimation* [ver sec. 10.6 de [8]]). El primer paso del método consiste en definir una secuencia inicial de realizaciones de $\Theta_t$ ($\theta^{(0)}_{1:n}$) que expliquen los datos observados (los aportes). Luego el algoritmo itera hasta encontrar una secuencia óptima $\theta^{*}_{1:n}$ . La implementación del modelo GSSM toma como condición inicial ($\theta^{(0)}_{1:n}[i]$, $i : 1, 2, 3.$) el logartimo de la media móvil (con una ventana de 13 semanas) de cada serie de aporte.

Cada una de estas tres series se modela con un proceso autoregresivo ARMA de parámetros p y q, la forma general de un proceso ARMA(p,q) se define como

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \psi_i \epsilon_{t-i}$$

con $\epsilon_i$ iid $\sim\ N(0, \sigma^2_\epsilon)$

Los procesos ARMA(p,q) determinan la forma de $T$, $R$ y $Q$ [ver sec. 3.4 de [8]]. En este caso se utilizaron procesos autoregresivos ARMA(2,2) para cada $\theta^{0}_{1:n}[i]$, la estimación por máxima verosimilitud de los parámetros de los procesos ARMA(2,2)) ($c$, $\phi_i$, $\psi_i$ y $\sigma_\epsilon$) determina completamente las entradas de $T$, $R$ y $Q$ distintas de cero. Llamemos $T^{(0)}$, $R^{(0)}$ y $Q^{(0)}$ a esta primer estimación.

El siguiente paso del método consiste en obtener $\theta^{(1)}_{1:n}[i]$ (i.e. obtener una segunda estimación del estado) a partir del modelo definido por (5.6) y (5.7) con $T = T^{(0)}$, $R = R^{(0)}$ y $Q = Q^{(0)}$. Para ello se maximiza la verosimilitud del estado dada la secuencia de aportes observados $y[i]_{1:n}$, esto es

$$\underset{\theta_{1:n}[i]}{\text{máx}} \quad p(\theta_{1:n}[i]|y[i]_{1:n}) \tag{5.8}$$

A partir de esta nueva $\theta^{1}_{1:n}[i]$ se vuelve a obtener una estimación de los parámetros de los procesos ARMA(2,2) y por tanto se obtiene $T^1$, $R^1$ y $Q^1$. El algoritmo converge en pocos pasos (ver sec. 10.6 de [8]) lo cual da como resultado $\theta^{*}_{1:n}[i]$ y sus correspondientes $T^*$, $R^*$ y $Q^*$ óptimas. Los detalles de la optimización de la verosimilitud (5.8) pueden verse en [sec. 10.6.2 de [8]]

Aquí es posible introducir una segunda variante al modelo, la estimación de $T^i$, $R^i$ y $Q^i$ a partir de la secuencia de estado $\theta^{1}_{1:n}[i]$ se puede realizar utilizando el algoritmo EM (Expectation Maximization) donde $T$, $R$ y $Q$ son las variables de la optimización (y no algunas de sus entradas con el caso anterior). Esto puede generar una mejora en la estructura de las matrices, por ejemplo introducir una estructura de dependencia entre las variables de estado $\Theta_t[1]$, $\Theta_t[2]$ y $\Theta_t[3]$ (el algoritmo presentado en este trabajo brinda esta posibilidad pero no se utiliza en la evaluación del modelo).

Para terminar esta sección se resume el algoritmo de ajuste del modelo GSSM.

### 5.7.4. Resumen del modelo y su ajuste a datos

El modelo final utilizado es

$$Y_t \sim \Gamma(Y_t|k, \Theta_t) \tag{5.9}$$

$$\Theta_{t+1} = T\Theta_t + R\eta_t, \qquad \eta_t \sim N(0, Q) \tag{5.10}$$

**Parámetros a estimar:**

1. Factor de forma $k[1]$, $k[2]$, $k[3]$ (3 parámetros)
2. $T$, $R$ y $Q \in\ M^{3\times3}$

Cada modelo ARMA(2,2) tiene 6 parámetros, por lo que $T$, $R$ y $Q$ tienen $6 \times 3 = 18$ entradas no nulas, en caso de que se utilicen las tres matrices como variables de estimación, en total contribuyen con $3 \times 3 \times 3 = 27$ parámetros.

Si utilizamos el criterio $\frac{\#\text{datos}}{\#\text{parámetros}} > 10$ se debe ajustar el modelo (en el peor caso) con al menos de 300 datos lo que equivale aproximadamente a dos años de cada serie de aportes.

**Ajuste a datos:**

1. A partir de las series de aportes de entrenamiento se obtiene su $\theta^{(0)}_{1:n}[i]$ correspondiente, $i : 1, 2, 3.$, aplicando una media móvil con una ventana de tamaño 13 semanas (el tipo de ventana puede elegirse, en este trabajo utilizamos la ventana *hamming*) al logaritmo de cada serie de aporte.
2. A partir de las series de aportes de entrenamiento se estima por MLE $\hat{k}[1]$, $\hat{k}[2]$ y $\hat{k}[3]$.
3. A partir de $\theta^{(i)}_{1:n}$ se estima por MLE $T^{(i)}$, $R^{(i)}$ y $Q^{(i)}$.
4. A partir de $\theta^{(i)}_{1:n}$ se obtiene $\theta^{i+1}_{1:n}$ a través del método AVME.
5. Itero hasta obtener $\theta^{*}_{1:n}[i]$ y $T^{*}$, $R^{*}$ y $Q^{*}$

### 5.7.5. Simulación

La simulación de series de este modelo es directa a partir de las ec. 5.6 y 5.7, para simular una secuencia de largo n se debe:

1. Generar una realización iid $\eta_{1:n}$ con $\eta_t \sim N(0, \sigma_\epsilon)$
2. A partir de un estado inicial $\theta_1$ se obtiene $\theta_2$ a través de (5.7)
3. A partir de $\theta_i$ se obtiene $\theta_{i+1}$ a través de (5.7). Repetir este paso hasta obtener $\theta_{1:n}$
4. A partir de $\theta_{1:n}$ y $k$ se generan muestras de $Y_t$ aplicando (5.6) y se obtiene las muestras simuladas $y_{1:n}$

### 5.7.6. Comentarios finales

Para finalizar esta sección se enfatiza que el modelo anterior permite introducir variantes, por ejemplo la serie del Niño puede introducirse como una variable de estado, la cual intuitivamente tiene una alta correlación con el estado $\theta$ del modelo anterior ya que el mismo es una medida del aporte instantáneo a cada embalse.

Otras variables explicativas del aporte pueden aún introducirse en el estado de forma natural y aplicar la misma metodología para ajustar el modelo.

En la sección 4.6 se vio que los niveles de cuantificación en la optimización del SimSEE son una limitante importante lo cual tiene como consecuencia de que las mejoras que pueden conseguirse en base a mejorar estos modelos continuos se ven atenuadas por los niveles de cuantificación requeridos.

La potencialidad descriptiva de estos modelos continuos debe estar garantizada al efectuar algún tipo de cuantificación en ellos con el fin de poder utilizarlos en el SimSEE, no obstante ello, veremos que ya esta primer versión del modelo GSSM presenta un desempeño superior al CEGH en varios de los índices del marco de evaluación propuesto en este trabajo, lo cual siguiere que es posible mejorar el CEGH. Su utilización en el SimSEE está justificada ya que el mismo tipo de cuantificación realizado en el CEGH puede realizarse con el GSSM. Finalmente, para este ultimo paso, es preciso también estudiar en profundidad como la descriptividad de los modelos continuos se ve afectada por la cuantificación realizada en ellos.

# 6 Marco de evaluación de modelos

En última instancia, la utilidad de cualquier modelo propuesto debe medirse en términos de cuánto ahorro puede representarle a la UTE utilizarlo al aplicarlo en la optimización de su política de operación (PO). La metodología con que realizamos esta evaluación es denominada *Marco extrínseco* de comparación, y será descrita al final de este capítulo.

Sin embargo, es necesario también disponer de un mecanismo de evaluación más directo. Algunas de las razones para ello son:

- Visualización inmediata de resultados durante el desarrollo y ajuste de los modelos
- Métrica trazable para, por ejemplo, ajustar parámetros de los modelos de manera automática mediante técnicas de optimización numérica.
- Perspectiva complementaria a la evaluación extrínseca, naturalmente sesgada por la implementación particular del método de optimización de la PO utilizado en SimSEE. (el propio SimSEE puede favorecer ciertos tipos de modelos frente a otros).

En este capítulo se motiva y luego expone el diseño de una serie de medidas de desempeño estadístico de los modelos propuestos.

## 6.1. Sobre las medidas utilizados anteriormente

Como antecedente de nuestro trabajo se realizó un proyecto FSE en 2009 [6] cuyo objetivo fue intentar mejorar el CEGH modificando algunos de sus componentes. Para evaluar el desempeño de los modelos propuestos, se propuso utilizar una serie de indicadores [14] de desempeño, a saber:

- Índice de Hurst
- Autocorrelación empírica
- Relación intensidad-duración-frecuencia (IDF) de eventos

De los tres, los dos primeros son indicadores estándar para caracterizar las propiedades estadísticas de modelos estocásticos. El primero, el de Hurst, es un índice escalar que indica si una serie de datos dada tiene *memoria larga* o no; no es realmente un índice de desempeño, sino un test estadístico realizado previamente en el proyecto para justificar cierto tipo de familias de modelos (puntualmente, FARIMA).

En todo caso, la metodología de validación de los modelos propuestos se basa en la comparación de las estadísticas (o índices) obtenidos para series simuladas, contra la calculada para la serie histórica real.

## 6.2. Índice de Hurst

Existen numerosas formas de estimar el índice de Hurst. La siguiente, llamada "R/S", por motivos que quedarán claros en breve, fue la utilizada en el proyecto FSE 2009 mencionado.

Consideremos la secuencia a analizar $x_{1:n}$. El índice se calcula de la siguiente manera:

- Se calcula la versión centrada de $v_{1:n}$, $\bar{v}_i = v_i - \hat{\mu}$ donde $\hat{\mu} = (1/n)\sum_i v_i$ es el valor medio de $v_{1:n}$.
- Se calcula la desviación acumulada de $v_{1:n}$, $z_i = \sum_{j \le i} \bar{v}_i$
- Se definen $R(t) = \text{máx}\{z_{1:t}\} - \text{mín}\{z_{1:t}\}$ el rango y $S(t) = \sqrt{(1/t)\sum_{i=1}^{t} \bar{v}_i^2}$ la varianza empírica calculada con la serie parcial $v_{1:t}$.
- El índice de Hurst se calcula como el exponente $H$ que, junto con una constante $C$, mejor se ajusta a

$$\frac{R(t)}{S(t)} = Ct^H$$

## 6.3. Autocorrelación empírica

La función de autocorrelación de un proceso estacionario en sentido amplio (WSS) $V_{1:n}$ se define como la función escalar $r(k) = E\left[(V_t - \mu)(V_{t-k} - \mu)\right]$, que por ser WSS no depende de $t$. La autocorrelación empírica $\hat{r}(k)$ se calcula simplemente como la autocovarianza empírica de la serie, separadamente para cada valor de $k$.

En el informe del proyecto FSE 2009 [6] se presenta gráficamente la autocorrelación empírica de 100 realizaciones de simulaciones de series temporales, contra la de la serie histórica *completa*. Notar que aquí hay cierto *sobreajuste* a los datos, ya que las simulaciones fueron calibradas en base a esa misma serie.

Teniendo en cuenta la periodicidad anual en los datos (que se refleja inmediatamente en la autocorrelación empírica), la evaluación de este índice se hace de dos maneras: una con la serie *cruda*, sin remover la periodicidad anual, y otra contra la serie luego de pasar por las *lentes Gaussianas* del CEGH, para que la serie sea "más IID".

## 6.4. Intensidad-Duración-Frecuencia (IDF)

Tomado del reporte del FSE 2009 [6],

> Surge muy claramente de los operadores del sistema la importancia de que los modelos representen adecuadamente la frecuencia y profundidad de los períodos de bajos aportes.

Sea $v_{1:n}$ una serie que queremos modelar; cada muestra $v_i$ se corresponde con una muestra semanal. Sea $\bar{v}_{1:n}(d)$ el resultado de filtrar a $v_{1:n}$ con una media móvil de duración $d$ semanas (en principio con ventana cuadrada pero podría ser algo más refinado). Decimos que $y_{1:n}$ es una *simulación* de $v_{1:n}$ si la primera preserva o aproxima un cierto conjunto de estadísticas de $v_{1:n}$ definidas a priori.

Sea $\hat{F}_{v_{1:n},d}(p)$ la distribución cumulativa empírica de la serie promediada temporalmente $\bar{v}_{1:n}(d)$. Para un largo $d$ de la ventana de promedidado temporal y un percentil $0 < p < 50$, el índice IDF se defiie para una serie $v_{1:n}$ como,

$$\text{IDF}(d, p; v_{1:n}) = \frac{\hat{F}^{-1}_{v_{1:n},d}(p/100)}{\hat{F}^{-1}_{y_{1:n},d}(1/2)} = \frac{\hat{F}^{-1}_{\bar{v}_{1:n},d}(p/100)}{\text{med}(\bar{v}_{1:n}(d))}. \tag{6.1}$$

El denominador es un término de normalización. Nos referimos pues a la IDF *no normalizada* a la medida correspondiente sin dicho término.

Como primera observación es importante resaltar que no queda claro, ni se fundamenta de ninguna manera en trabajos anteriores, por qué una medida como la IDF es apropiada para capturar las estadísticas deseadas a los efectos de simular aportes.

## 6.5. Nuevos índices propuestos

Más allá de la idoneidad de la IDF para capturar las estadísticas relevantes al problema, es claro de su definición que la IDF no es una medida de diferencias, sino una forma de observar las estadísticas de una serie particular. Debido a ésto, en trabajos anteriores sobre el tema, el desempeño entre series y simulaciones a nivel estadístico se realizó mediante la comparación visual entre las curvas IDF de las series reales $v_{1:n}$ y sus simulaciones $y_{1:n}$.

Siendo que el objetivo de este proyecto en cuanto a medidas de desempeño es obtener medidas estadísticas objetivas de calidad de simulaciones que se correlacionen positivamente con el desempeño obtenido en SimSEE en términos de ahorro esperado del sistema eléctrico, lo que en adelante se propone es una serie de medidas sensibles al tipo de características relevantes para el problema.

**IDP (Intensidad-Duración-Probabilidad)** Esta medida resulta natural en un problema como el de modelar aportes, dado que permite comparar directamente qué tan frecuente es la ocurrencia de distintos niveles de aporte en ambas series. En este caso, denotaremos por $P_{v_{1:n},d}$ a la distribución empírica de la serie promediada $\bar{v}_{1:n}(d)$. La formulación es la siguiente,

$$\mathrm{IDP}_d(v_{1:n}\,;\,y_{1:n}) = H\left(\hat{P}_{v_{1:n},d}\,;\,\hat{P}_{y_{1:n},d}\right) w(v)\,dv \tag{6.2}$$

donde la divergencia $H$ tiene la forma

$$H(P;Q) = \int_{\xi} h\left(P_{v_{1:n},d}(\xi)\,;\,Q_{y_{1:n},d}(\xi)\right) w(\xi)\,d\xi \tag{6.3}$$

Un ejemplos de $H(\cdot;\cdot)$ es la divergencia de Kullback-Leibler [7], que en definitiva se tomó como base para los resultados finales de este proyecto (medidas como la norma $\ell_1$ o la norma $\ell_2$ ponderadas están implementadas y fueron probadas, arrojando resultados similares),

$$H(F;G) = \int_{\xi} \hat{P}_{v_{1:n},d}(\xi) \log \frac{\hat{P}_{v_{1:n},d}(\xi)}{\hat{Q}_{y_{1:n},d}(\xi)}\,d\xi \tag{6.4}$$

**IDC (Intensidad-Duración-Condicional)** Durante la optimización de la operación, cada paso del algoritmo Backward Dynamic Stochastic Approximation utilizado por SimSEE utiliza al modelo estadístico de series subyacente para modelar no una serie completa sino tan sólo *una* muestra futura de la serie dado su estado actual.

Por lo anterior, en términos del sistema y por ende de los objetivos de este proyecto, la medida ideal de desempeño de un modelo de simulación de series debería depender qué tan bien dicho modelo simula las distribuciones *condicionales* de la serie, y no las distribuciones marginales aunque sea a distintas escalas.

Una primera medida basada en esta idea podría formularse simplemente en términos de las distribuciones empíricas condicionales de cada serie $v_{1:n}$ y $y_{1:n}$, donde el condicionamiento viene dado por que los valores de las últimas $d$ muestras de cada serie correspondan a un vector, $z_{1_d}$ al que denominamos *contexto*,

$$\mathrm{IDC}_d(v_{1:n};y_{1:n}) = \int_{z\in\mathcal{V}^d} H\left(P_{v_{1:n}}(\cdot|z_{1:d})\,;\,P_{y_{1:n}}(\cdot|z_{1:d})\right)\mu(z_{1:d})$$

Notar que el rol de la duración $d$ aquí es bastante distinto al utilizado anteriormente, si bien mantiene un vínculo con él a través del sentido de la "escala" que representa.

El mayor problema con la IDC es que *no es aplicable en la práctica*. Dada la cantidad posible de valores de los aportes (de 0 a decenas de miles) es muy poco probable que un contexto

cualquiera $z_{1:d}$ se repita en un histórico de tamaño razonable. Esto hace imposible tener estimaciones útiles de las distribuciones condicionales involucradas a menos que se realice algún tipo de agrupación de los contextos.

Esto último es realizable. Consideremos por ejemplo una función $g(z)$ que mapea cada posible contexto $z$ a un conjunto pequeño de valores enteros. La medida resultante podría ser entonces calculada como,

$$\mathrm{IDC}'_d(v_{1:n}, y_{1:n}) = \sum_c H\left(\hat{P}_{v_{1:n},d}(\cdot|g(z)=c)\,;\,\hat{P}_{y_{1:n},d}(\cdot|g(z)=c)\right). \tag{6.5}$$

El problema que persiste sin embargo es que *cualquier agrupación $g(z_{1:d})$ de contextos introduce un sesgo muy importante en la medida.* Precisamente, dicha agrupación podría ser utilizada como una representación del estado del sistema, y en base a ella definirse un nuevo modelo, el cual sería el mejor modelo posible para la medida IDC resultante!

Para evitar tal sesgo y a la vez mantener la importancia del condicionamiento al evaluar el desempeño es que se terminó por definir la siguiente medida.

**ISC: Índice de Simulación Condicional** Como se mencionó anteriormente, lo ideal sería, de acuerdo a cómo funciona el optimizador del SimSEE, tener algo que capture las distribuciones *condicionales* de las series, pero aplicar tal idea directamente da lugar a problemas de dilución estadística que no pueden ser resueltos sin incorporar sesgos importantes en la medida.

La idea detrás de este índice es muy simple: para cada posición temporal $t$ en la serie original $v_{1:n}$, se compara la distribución empírica de las siguientes $d$ muestras originales $v_{t:t+d}$ contra las distribuciones empíricas de series $y_{t:t+d}$ generadas por un simulador *$g$ que sólo tiene acceso a las muestras anteriores a $t$, $v_{1:t-1}$, incluyendo éstas el estado inicial de la simulación en $t$.* De esta manera se incorpora el condicionamiento con el pasado, al ser el propio simulador una función de las muestras pasadas. Para cada simulación $y_{t:t+d}$ utilizaremos el índice IDC como medida de divergencia entre las distribuciones condicionales de ella y la serie original. Luego se aproxima el valor esperado según las simulaciones de dicha divergencia mediante un promedio tomado sobre un número $r$ de simulaciones,

$$\mathrm{ISC}_d(v_{1:n}, g) = \frac{1}{n-d}\sum_{j=1}^{n-d}\frac{1}{r}\sum_{k=1}^{r}\delta(\,x_{j:j+d}\,;\,y_{1:d}(j,k)\,) \tag{6.6}$$

Claramente, podrían utilizarse otras formas de resumir el desempeño de las simulaciones que no fueran el promedio. x

## 6.6. Modelo de generación simplificado

Con infinitos datos podría pensarse en un modelo estadístico que capture perfectamente, en el límite, todas las propiedades estadísticas de las series a simular. Desafortunadamente, en el problema real que se plantea en este caso, la cantidad de datos no es tan grande. Es imposible entonces pedirle a un simulador que logre capturar *todas* las propiedades estadísticas de las series a simular.

El problema fundamental que surge entonces es *cómo ponderar a la propia medida* en términos de lo que finalmente realmente interesa. En nuestro caso, lo que importa es cómo se termina correlacionando un buen (o mal) ajuste con un buen (o mal) desempeño en términos de la potencia generada.

La observación clave aquí es que el propio modelo de generación de potencia, dado por los modelos físicos de los embalses, el modelo de generación eléctrica y la política de operación aplicada a ellos, actúa como filtro no lineal de las señales de aporte. Siendo no lineal, y siendo

un modelo con memoria, este filtro determina entonces qué diferencias estadísticas entre series reales y simuladas serán visibles a la salida de potencias. Además, claramente, medir diferencias en términos de potencias generadas se acerca mucho más al objetivo que se desea maximizar en la optimización.

Todo esto lleva a que lo ideal sería medir el desempeño de modelos en términos de las potencias generadas entre las simulaciones basadas en ellos, y las potencias generadas en base a series reales.

La solución que se plantea de aquí en más a ésto es la siguiente: disponer de un *modelo de generación simplificado*, que capture las características esenciales de él, sin depender de un modelo de condicionamiento, o un algoritmo de optimización complejo. Como características esenciales incluimos la *no linealidad* de la generación (tope de potencia máxima instantánea generada, tope de nivel de embalse, nivel mínimo de embalse), y la *memoria* del sistema, dada por los propios embalses.

En adelante presentamos algunos de los modelos de generación simplificados propuestos. En todo caso, la política de operación se determina de manera de optimizar la generación total en un período de tiempo de entrenamiento. Y en la siguiente sección se detalla el modelo de generación final utilizado en el marco de evaluación para sintetizar series de potencia a partir de las series de aportes.

Es importante recalcar que la simplicidad no es sólo una cuestión de comodidad; en nuestro caso, es importante tener un modelo sencillo para que las diferencias observadas en potencia sean razonablemente trazables hacia el modelo de aportes, de modo de servir de guía para el afinamiento de éstos.

**Geometría de los embalses** El modelo de embalse utilizado vincula de manera lineal el área del embalse con la altura $h$ del embalse, de manera que el volumen útil del embalse se relaciona de manera cuadrática con su altura, $v(h) = ah^2 + bh + c$. La altura del embalse varía entre un $h_{\text{mín}}$ para el cual $v(h_{\text{mín}}) = 0$, y un $h_{\text{máx}}$ arriba del cual el embalse es rebasado y el agua en exceso es drenada sin generar potencia.

$$P[kW] = \eta \times \rho[kg/m^3] \times g[m/s^2] \times \Delta[m] \times q[m^3/s]$$

donde $q$ es el caudal turbinado en un cierto instante por la represa.

**Modelo de interconexión** El sistema considerado en este proyecto cuenta con los aportes a las represas de Bonete (B), Palmar(S) y Salto(S). (La represa de Baygorria se modela junto con Palmar como una sola). La represa de Salto es independiente de las otras, mientras que Palmar se encuentra aguas abajo de Bonete, por lo que recibe el *caudal erogado* por la represa de Bonete en la *semana anterior*. El caudal erogado $e$ es la suma del caudal turbinado $q$ y el caudal vertido $z$, $e = q + z$.

**Política de operación** Para evitar el uso de un modelo de condicionamiento y, a la vez, tener una política simple pero razonable, la idea es obtener una política que maximice la generación de potencias en un período de la historia conocida dado su historial de lluvias, demandas, y en principio los distintos costos de generación.

Asumiremos que la energía hidráulica es *siempre la más barata*, de modo que optimizar la operación para una demanda dada durante un cierto período se reduce a minimizar la diferencia positiva entre la demanda y la potencia generada en cada instante; no se valora la sobregeneración en este modelo.

$$(x, u) = o(z, v)$$

donde $x$ es el caudal turbinado, $u$ el vertido, $v$ el aporte previsto (por ej. el aporte previsto en las siguientes $m$ semanas) y $z$ el nivel del mismo. La evaluaremos observando la serie de potencia generada al utilizarla como política de operación.

La evaluación de distintos modelos de aportes podrá realizarse a través de la comparación de la serie de potencia hidráulica obtenida con la política de operación $o(\cdot)$ al utilizar la serie histórica y el modelo. Esta comparación también podrá realizarse con la misma optimización anterior pero con la simulación de aportes del modelo en un período histórico y la serie óptima antes obtenida en el mismo período.

## 6.7. Modelo de central hidroeléctrica

Dada la escala semanal de tiempo, el modelo de central hidroeléctrica está basado en dos aspectos de su dinámica. El primero es el coeficiente de reserva del embalse

$$\rho_r \triangleq \textit{período de tiempo en semanas que puede operar la central a potencia nominal partiendo con su embalse en el nivel máximo y hasta que llega a su nivel mínimo, suponiendo que no hay aportes.}$$

La energía entregada en ese período es

$$E_{max} = \rho_r P_n T_s \tag{6.7}$$

donde $P_n$ es la potencia nominal de la central y $T_s$ un período de tiempo de una semana. El segundo es la suposición de un aspecto de diseño y es que la central puede operar a potencia nominal para todos los niveles admisibles de su embalse, como consecuencia el caudal turbinado será mayor en niveles bajos del embalse y menor para valores altos, de forma de suministrar siempre la potencia nominal. La ecuación que modela la potencia generada por la central (considerando que no hay aportes ni vertido) será

$$P_n = -2a_h V \frac{\delta V}{\delta t} \tag{6.8}$$

donde $V$ es el volumen del embalse y $a_h$ una constante determinada por el primer aspecto mencionado. Si integramos la ecuación anterior en un período $\Delta t = T_s \rho_r$ en el cual no hay aportes ni vertido, y si inicialmente el embalse está en su nivel máximo, al final del período el embalse se vaciará y la energía generada será $E_{max}$

$$E_{max} = \int_{\Delta t} P_n \delta t = \rho_r P_n T_s = \int_{\Delta t} -2a_h V \frac{\delta V}{\delta t} \delta t = \int_{V=V_{max}}^{V=0} -2a_h V \delta V = a_h V_{max}^2$$

y entonces

$$a_h = E_{max}/V_{max}^2 = \rho_r P_n T_s / V_{max}^2 \tag{6.9}$$

### 6.7.1. Modelo de generación hidráulica I

Consideraremos dos variantes de modelo anterior, en ambas supondremos que la energía demandada en una semana a la central hidroeléctrica se genera operando la misma a potencia nominal durante el intervalo de tiempo necesario, pero en la primera (y más simple), supondremos que tanto el aporte como el vertido semanal del embalse se realiza al final de la semana. No se fija una cota para el vertido y el mismo puede realizarse aún si el embalse no está en su nivel máximo. Consecuencia de ello la potencia generada por la central tiene como cota $E_{max}$

si $\rho_r \leq 1$ o $E_{max}/\rho_r = P_n T_s$ si $\rho_r > 1$. Si en una semana la energía demandada a la central es $E = P_n T$, entonces el volumen final de agua en el embalse queda determinado por

$$P_n T = E = a_h(V_o^2 - V_f^2) \tag{6.10}$$

donde $V_o$ y $V_f$ son el volumen al inicio y fin de la semana.

La curva de la figura 6.1 ilustra como varia el volumen del embalse desde su nivel máximo al mínimo en función de la energía generada. Los puntos $V_o$ y $V_f$ de la figura 6.1 muestran una

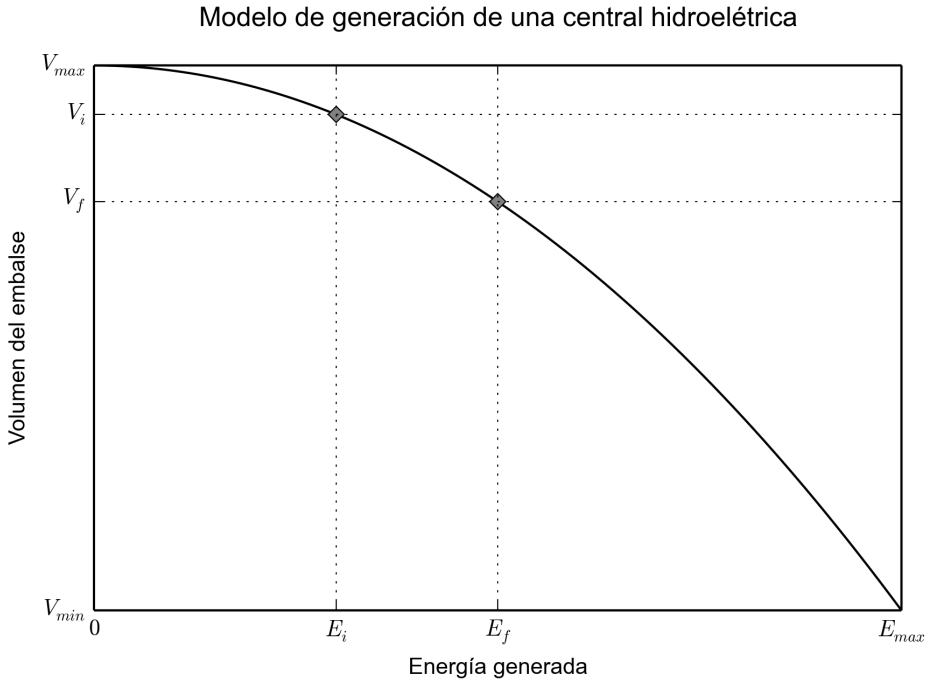


Figura 6.1: Variación del volumen del embalse desde su nivel máximo al mínimo en función de la energía generada. Los puntos $V_o$ y $V_f$ muestran una posible evolución del embalse en una semana, inicialmente el volumen es $V_o$, al final de la semana es $V_f$ y la energía generada es $E = E_f - E_i$

posible evolución del embalse en una semana, inicialmente el volumen es $V_o$, al final de la semana es $V_f$ y la energía generada es $E = E_f - E_i$.

### 6.7.2. Modelo de generación hidráulica II

La segunda variante es igual a la anterior pero supondremos que tanto el aporte como el vertido semanal se realiza en forma uniforme a lo largo del período $T$. En este caso tenemos que

$$\frac{\delta V}{\delta t} = \bar{v} - \bar{u} - \frac{P_n}{2a_h V} \tag{6.11}$$

donde $\bar{v}$ y $\bar{u}$ son el aporte y vertido por unidad de tiempo respectivamente. Nuevamente si la central debe generar $E = P_n T$, al integrar en un período $T$ obtenemos la ecuación que determina el volumen final en el embalse

$$T = \int_T \delta t = \int_{V_o}^{V_f} \frac{\delta V}{\bar{v} - \bar{u} - \frac{P_n}{2a_h V}}$$

y operando tenemos que

$$T = \begin{cases} \frac{V_f - V_o}{\bar{v} - \bar{u}} + \frac{P_n}{2a_h(\bar{v}-\bar{u})^2} \log\left(\frac{2a_h(\bar{v}-\bar{u})V_f - P_n}{2a_h(\bar{v}-\bar{u})V_o - P_n}\right) & si \quad \bar{v} - \bar{u} \neq 0 \\ a_h(V_o^2 - V_f^2)/P_n & si \quad \bar{v} - \bar{u} = 0 \end{cases} \tag{6.12}$$

## 6.8. Optimización de la política de operación

Definamos la notación a utilizar en el planteo del problema de optimización, la serie de aportes en un período dado la representamos como $v_{1:n} = (v_1, v_2, \ldots, v_n)^T$, donde $t = 1, \ldots, n$ es el índice semanal, $u_t = (u_t^b, u_t^p, u_t^s)$ corresponde al aporte en la semana $t$ a cada represa ($u_t^b$ aporte a Bonete, etc.) y $v \in \mathbb{R}^{3n}$. La demanda será el vector $d = (d_1, d_2, \ldots, d_n)^T \in \mathbb{R}^n$. Los vectores de aporte y demanda son datos conocidos y no serán variables independientes en la optimización. En forma análoga a $v$ tenemos $y^h \in \mathbb{R}^{3n}$, $x \in \mathbb{R}^{3n}$, $z \in \mathbb{R}^{3n}$ y $u \in \mathbb{R}^{3n}$ que representan respectivamente, la potencia generada por las centrales hidroeléctricas, el agua turbinada necesaria para generarla, el volumen de agua en el embalse al final de la semana y el agua vertida. Por último, $y^c = (y_1^c, y_2^c, \ldots, y_n^c)^T \in \mathbb{R}^n$ representa la energía térmica generada en cada semana.

### 6.8.1. Con el modelo de generación hidráulica I

La optimización de la política de operación en un período de tiempo maximizará la energía hidráulica generada basada en la suposición de que es la energía más barata. La operación óptima será entonces la solución del siguiente problema

$$\begin{aligned}
\underset{(x,y_h,z,u)}{\text{máx}} \quad & 1_{3n}^T \cdot y^h && && (6.13)\\
\text{sujeto a} \quad & 1_3^T \cdot y_t^h \leq d_t && \forall t = 1, \ldots, n. && (6.14)\\
& x_t^r + z_t^r + u_t^r = v_t^r + z_{t-1}^r && \forall t = 1, \ldots, n.\ \forall r = b, s. && (6.15)\\
& x_t^p + z_t^p + u_t^p = v_t^p + z_{t-1}^p + x_{t-1}^b + u_{t-1}^b && \forall t = 1, \ldots, n. && (6.16)\\
& y_t^{h,r} = a^{h,r}[{z_{t-1}^r}^2 - (z_{t-1}^r - x_t^r)^2] && \forall t = 1, \ldots, n.\ \forall r = b, p, s. && (6.17)\\
& z_t \leq z_{max} && \forall t = 1, \ldots, n. && (6.18)\\
& y_t^h \leq y_{hmax} && \forall t = 1, \ldots, n. && (6.19)\\
& x, y^h, z, u \geq 0 && && (6.20)
\end{aligned}$$

donde $1_n^T$ denota el vector columna de unos de tamaño n, (6.14) establece que en cada semana la suma de la potencia generada por cada represa sea menor o igual a la demanda. (6.15) y (6.16) representan el balance en cada represa y establecen para toda semana que el volumen final en el embalse más el agua turbinada y vertida es igual al volumen inicial de agua en el embalse más el aporte recibido, (6.16) corresponde al balance en Palmar que como está aguas abajo de Bonete recibe como aporte extra lo turbinado y vertido por Bonete en la semana anterior. (6.17) corresponde al primer modelo de generación hidráulica de la sección anterior (6.10) (con $a_h = (a_h^b, a_h^p, a_h^s)^T$) ya que

$$y_t^{h,r} = P_n^r T_t, \quad V_o = z_{t-1}^r \quad y \quad V_f = z_{t-1}^r - x_t^r,$$

el volumen final del embalse incorpora la suposición de que tanto el vertido como el aporte se realiza al final de la semana. Finalmente (6.18) y (6.19) son cotas al volumen de agua en los embalses y a la potencia generada por las centrales hidroeléctricas.

### 6.8.2. Con el modelo de generación hidráulica II

En este caso utilizamos en lugar de la restricción (6.17), la definida por el modelo de generación (6.12),

$$1 = \begin{cases} \frac{z_t^r - z_{t-1}^r}{v_t^r - u_t^r} + \frac{y_t,^{h,r}}{2a_h^r(v_t^r - u_t^r)^2} \log\left(\frac{2a_h^r(v_t^r - u_t^r)z_t^r - y_t^{h,r}}{2a_h^r(v_t^r - u_t^r)z_{t-1}^r - y_t^{h,r}}\right) & si \quad v_t^r - u_t^r \neq 0 \\ a_h^r({z_{t-1}^r}^2 - {z_t^r}^2)/y_t^{h,r} & si \quad v_t^r - u_t^r = 0 \end{cases} \tag{6.21}$$

ya que $\bar{v}_t^r = v_t^r/T_t$, $baru_t^r = u_t^r/T_t$, $P_{n,r}T_t = y_t^{h,r}$, $V_o^r = z_{t-1}^r$ y $V_f^r = z_t^r$.

### 6.8.3. Implementación

Con ambos modelos de generación hidráulica el problema de optimización puede plantearse en la forma

$$\begin{aligned} \max_w \quad & c'w \\ \text{s.a.} \quad & Dw \leq d \\ & Bw = u \\ & g(w) = 0 \\ & 0 \leq w \leq w^U \end{aligned} \tag{6.22}$$

con $w = (x^T, {y^h}^T, z^T, v^T)^T$. La solución se obtiene mediante el método SLSQP (Sequential Least Squares Programming) [13] (paquete en `Python` disponible en [1]).

## 6.9. Análisis y resultados

Para analizar cómo es la operación óptima en el período histórico se toman los siguientes períodos de tiempo

$$p_k = [ks_o + 1, ks_o + 2, \ldots, ks_o + s_o + s_d], \qquad k = 0, 1, 2, \ldots$$

la optimización se realiza para cada uno de los períodos pero los últimos $s_d$ datos (o semanas) se descartan, con el fin de atenuar el efecto de borde, sólo los primeros $s_o$ datos son válidos como solución.

A partir de la serie histórica de aportes $v = (v_1, v_2, \ldots, v_N)'$ (correspondiente a los aportes semanales desde 1909 a 2008) y de la serie de demanda $d = (d_1, d_2, \ldots, d_N)'$ se obtiene

$$v_k = [v_{ks_o+1}, v_{ks_o+2}, \ldots, v_{ks_o+s_o+s_d}]^T \text{ y } d_k = [d_{ks_o+1}, d_{ks_o+2}, \ldots, d_{ks_o+s_o+s_d}], \quad k = 0, 1, 2, \ldots$$

con los cuales se realiza la optimización. La condición inicial de los embalses para el período $k$ se toma igual al nivel de los embalses en $ks_o$

Con ambos modelos de generación hidráulica el problema a optimizar no es convexo pero podemos obtener la condición inicial en la cual en cada semana se turbina el máximo posible (respetando la demanda en esa semana). La llamaremos operación *greedy*, para obtenerla en cada semana primero se turbina el mínimo entre el máximo disponible con Salto y lo necesario para satisfacer la demanda, sino se alcanza la demanda se turbina en forma análoga con Palmar y lo mismo con Bonete. La solución óptima será para nosotros el óptimo local obtenido con la condición inicial anterior.

[1] http://www.pyopt.org/reference/optimizers.slsqp.html

### 6.9.1. Resultado con el modelo de generación I

Para hacer un análisis cualitativo de la optimización utilizaremos como medida la potencia anual generada en cada represa. El parámetro $s_o$ debe determinarse experimentalmente, en la elección del mismo hay un compromiso entre la dimensión del problema de optimización y la información a futuro disponible (i.e. aportes y demanda). Para elegirlo se tomó primeramente un valor alto del mismo y se lo fue bajando observando como variaba la potencia anual generada en cada represa, la potencia anual generada en la optimización no aumentaba sensiblemente para valores mayores a $s_o = 26$ semanas.

La figura 6.2 muestra la potencia total generada anualmente (i.e. la suma de lo generado por las tres centrales) al optimizar la operación y al utilizar la operación *greedy*.

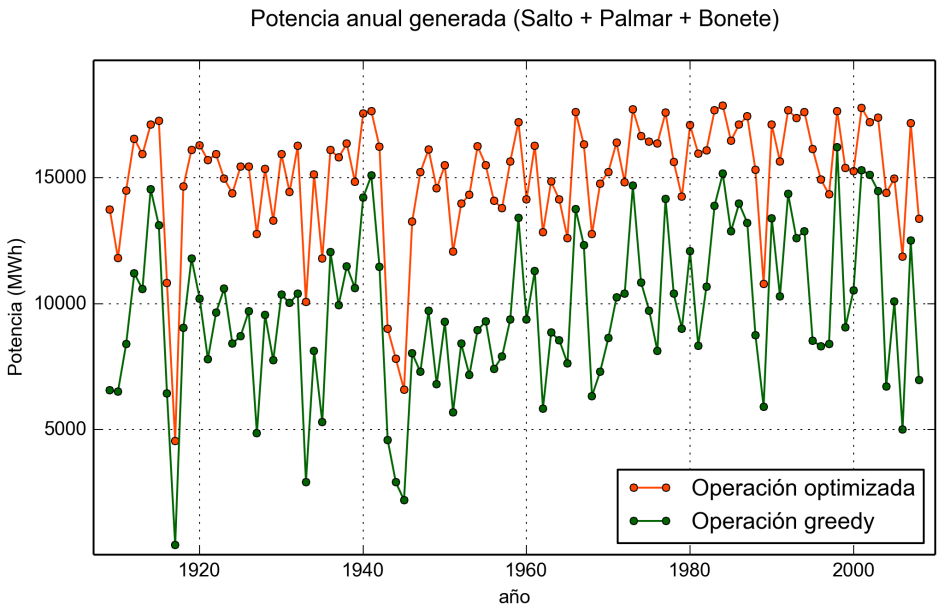


Figura 6.2: Potencia total generada anualmente (i.e. la suma de lo generado por las tres centrales) al optimizar la operación y al utilizar la operación *greedy*

#### Salto Grande

La figura 6.3 muestra el volumen semanal óptimo de agua turbinada y vertida en Salto Grande en función del aporte semanal recibido, si $x^{s,*} = (x_1^{s,*}, \ldots, x_N^{s,*})^T$, $u^{s,*} = (u_1^{s,*}, \ldots, u_N^{s,*})^T$ y $v^s = (v_1^s, \ldots, v_N^s)^T$ las figuras muestran los conjuntos de puntos

$$\{(v_t^s, x_t^{s,*})\}_{t=1,\ldots,N} \text{ y } \{(v_t^s, u_t^{s,*})\}_{t=1,\ldots,N}$$

En el caso de Salto Grande el resultado de la optimización es bastante simple, el volumen del embalse mayormente se mantiene en su nivel máximo, por ello el caudal turbinado en una semana es en general igual al aporte recibido en esa semana. La política de operación en Salto podemos definirla entonces como (la llamaremos operación *predictiva*)

$$x_t^s = u_t^s \qquad \text{(Política de operación } \textit{predictiva}\text{, Salto Grande)} \tag{6.23}$$

con la observación importante de que la política anterior debe realizarse una vez que el nivel del embalse este en su máximo. La cota del caudal turbinado que aparece en la figura 6.3 es consecuencia de la capacidad máxima de generación y el mismo efecto se producirá al utilizar

la política anterior ya que también debe respetarse esta capacidad máxima de generación. El vertido en salto es siempre cero pero cuando se llega al máximo caudal turbinado cambia a ser lineal con el aporte (esto con el fin de que se mantenga el balance de agua en el embalse).

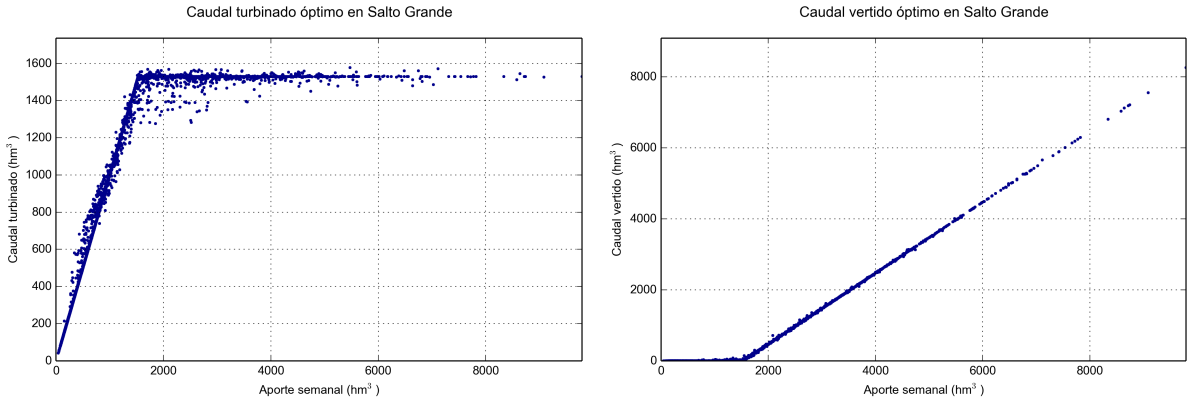


Figura 6.3: Volumen semanal óptimo de agua turbinada y vertida en Salto Grande en función del aporte semanal recibido.

Finalmente la figura 6.4 muestra comparativamente la potencia anual generada en Salto Grande utilizando la política *predictiva* y la operación optimizada. Podemos concluir que para Salto, la política de operación *predictiva* (ec. 6.23) bajo el modelo supuesto del sistema eléctrico de Uruguay obtiene un resultado similar al óptimo.

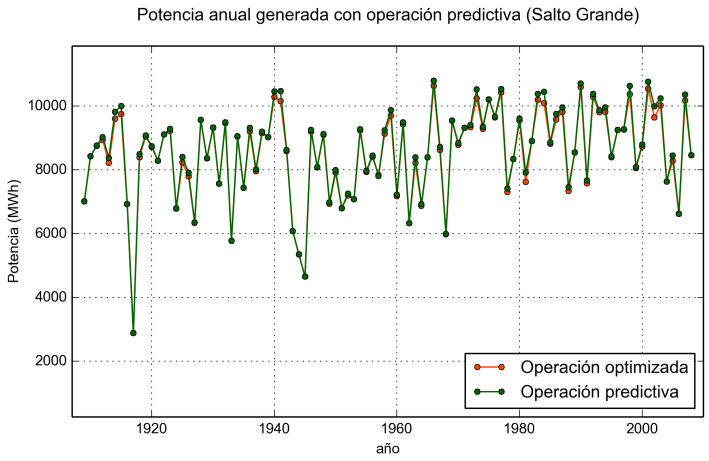


Figura 6.4: Potencia anual generada en Salto Grande utilizando la política de operación *predictiva* y la operación optimizada

### Palmar y Bonete

En el caso de Palmar y Bonete la figura 6.5 muestra que no es tan fácil definir un modelo simple de operación de la forma

$$(x_t, u_t) = o(z_{t-1}, v_t)$$

es decir, función del volumen inicial de agua en el embalse y del aporte recibido. La figura 6.5 muestra los conjuntos de puntos

$$\{(v_t^p, z_{t-1}^{p,*}, x_t^{p,*})\}_{t=1,\ldots,N} \text{ y } \{(v_t^b, z_{t-1}^{b,*}, x_t^{b,*})\}_{t=1,\ldots,N}$$

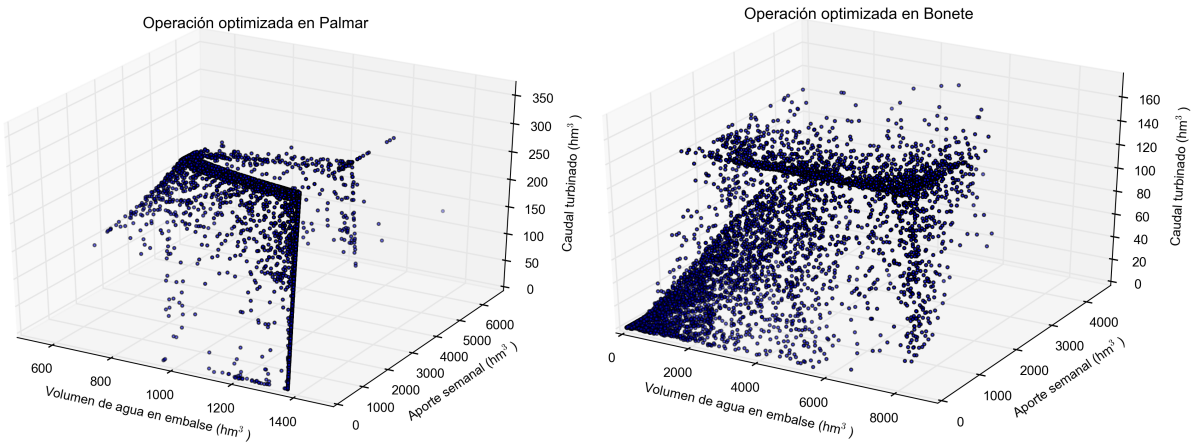


Figura 6.5: Operación optimizada en Palmar y Bonete.

Tampoco es sencillo obtener un modelo de la forma

$$(x_t, u_t) = o(z_{t-1}, v_{t,i+m})$$

con $v_{t,t+m} = (v_t, v_{t+1}, \ldots, v_{t+m})^T$, por lo que simplificaremos el modelo de optimización planteado en la sección 6.8.1 con el fin de obtener una política de operación. En este sentido se plantean las siguientes dos variantes:

- Producto de la política de operación encontrada para Salto Grande, el modelo optimiza la generación hidráulica solo considerando a Salto Grande y luego, con la demanda que no satisface Salto, se optimiza la operación de Palmar y Bonete.

- La otra variante es similar a la anterior, primero se optimiza solo considerando a Salto Grande, luego con la demanda que no satisface Salto se optimiza Bonete (sin considerar a Palmar en el modelo) y por último, con la demanda que no satisface Bonete, se optimiza Palmar.

La figura 6.6 muestra que la potencia anual generada con la primer variante es similar al modelo de optimización de la sección 6.8.1, con la segunda variante el desempeño es un poco peor pero sigue siendo superior que la política de operación *greedy*.

Al considerar la segunda variante del modelo de optimización (i.e. optimizando las tres centrales por separado) Bonete muestra una operación bastante simple, la figura 6.7 muestra la potencia generada en Bonete en función del nivel del embalse. La figura sugiere definir una política de operación de tipo *umbral*, es decir, turbino a potencia máxima pero manteniendo el nivel del embalse por encima de cierto umbral.

En el caso de Palmar la operación optimizada sigue siendo más compleja, en particular si bien hay cierto comportamiento de tipo *umbral* (figura 6.7), también lo hay del tipo *predictivo* (figura 6.8), tal como Salto Grande, además de otro tipo de operación. Utilizaremos de todos modos una política de tipo *umbral* para Palmar con el fin de tener una primer política simple.

Finalmente la figura 6.9 muestra como es el desempeño de las variantes definidas, i.e. los tres modelos de optimización, y utilizando las políticas definidas para cada central (en Salto de tipo *predictiva* y en Palmar y Bonete de tipo *umbral*). Todas están referidas a la generación con la política *greedy*, por lo que la gráfica muestra cuanto mejora la generación respecto a la política *greedy*.

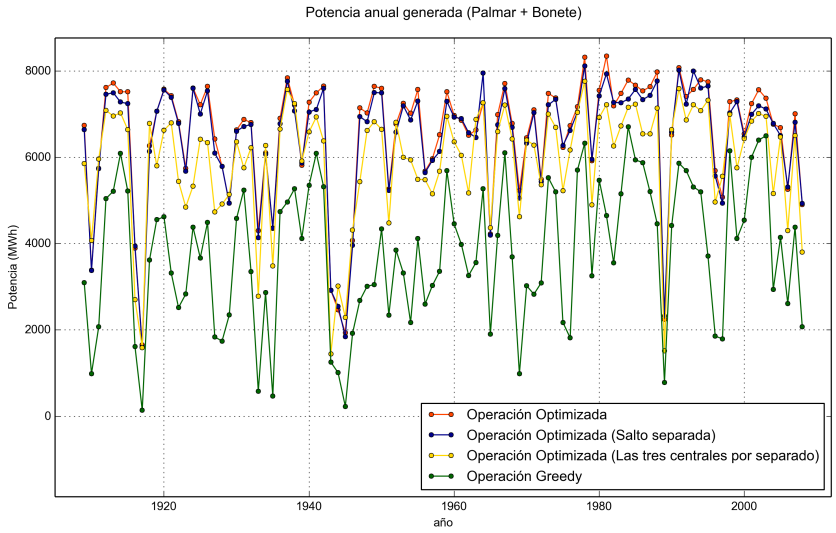


Figura 6.6: Potencia anual generada (Palmar+Bonete) con las distintas variantes del modelo de optimización.

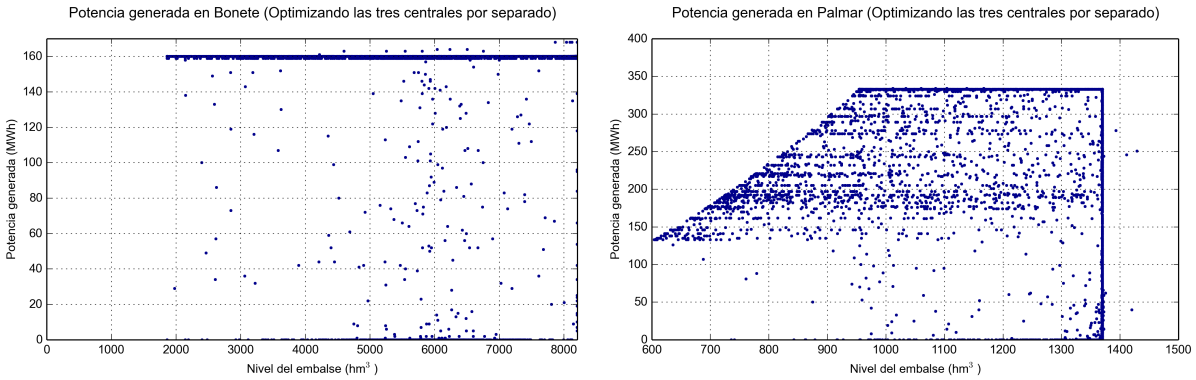


Figura 6.7: Potencia generada en Palmar y Bonete optimizando las tres centrales por separado

## 6.10. Variante utilizada en el marco de evaluación

Para finalizar indicamos que el modelo de generación utilizado el en marco de evaluación corresponde al que optimiza la generación de las tres centrales juntas.

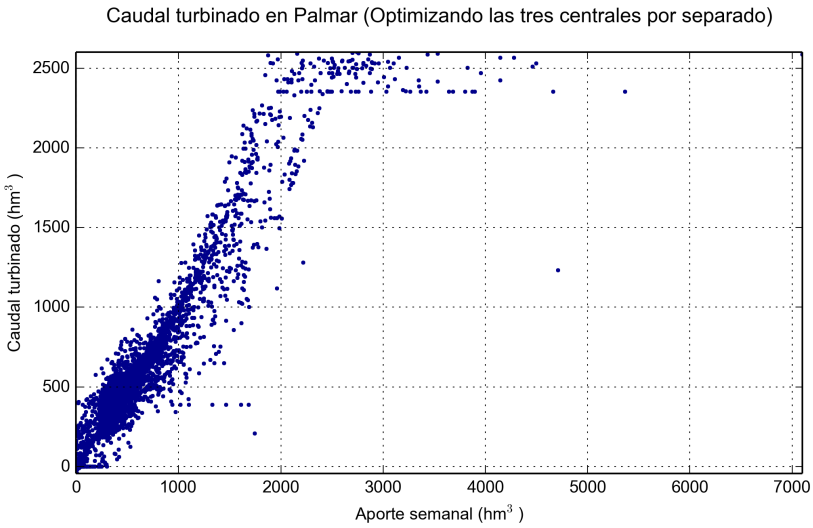


Figura 6.8: Caudal turbinado en Palmar optimizando las tres centrales por separado

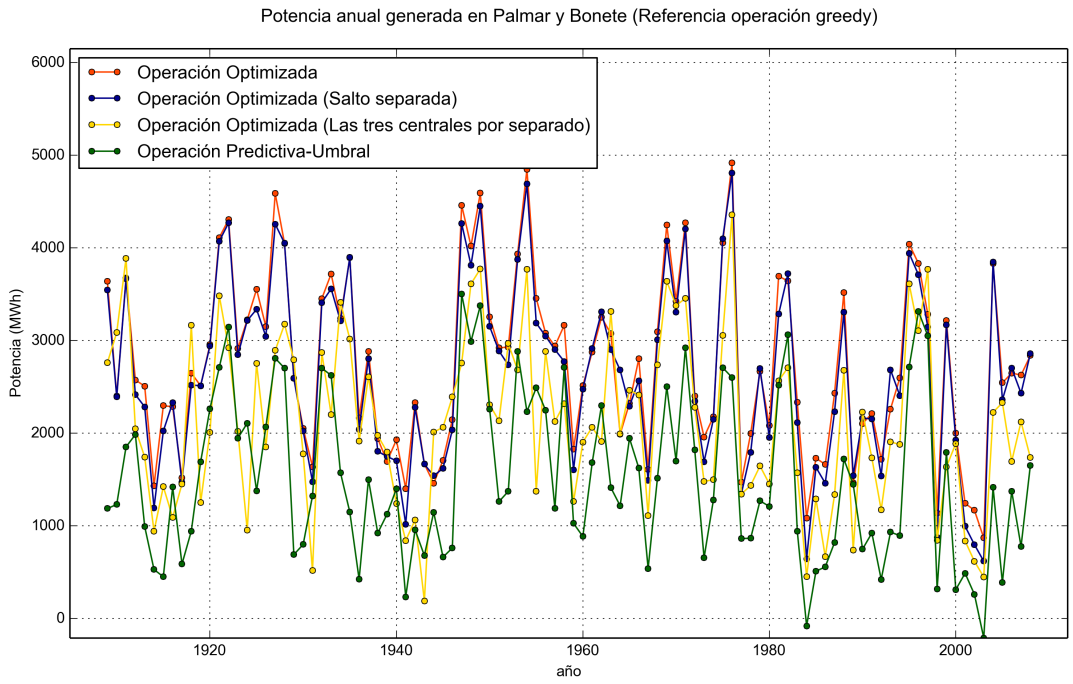


Figura 6.9: Mejora en la generación de potencia de las variantes del modelo de la sección 6.8.1 y del uso de la política de operación *predictiva* en Salto y de tipo *umbral* en Bonete y Palmar respecto a la política *greedy*.

# 7 Resultados

## 7.1. Datos utilizados

Los datos de aportes y demanda eléctrica utilizados en este proyecto fueron obtenidos del Despacho Nacional de Cargas. Concretamente, se trabajó con los datos de aportes registrados en las represas de Bonete, Palmar y Salto entre enero de 1909 y abril de 2009. Los datos sobre el índice N3.4 se obtuvieron de la NOAA (National Oceanic and Atmospheric Administration) [1], comenzando en enero del año 1982. Finalmente, para optimizar y evaluar operaciones se trabajó con datos de precio de crudo desde 1987 hasta 2015, y la demanda eléctrica semanal registrada hasta la fecha.

## 7.2. Marco de evaluación intrínseco

Como se mencionara anteriormente, denominamos como *evaluación intrínseca* de los modelos a aquella que se obtiene mediante medidas estadísticas y/o aproximadas de desempeño, previo a su aplicación directa en el SimSEE.

El índice finalmente utilizado, como fuera mencionado anteriormente, es el ISC con un par de pequeñas simplificaciónes. En primer lugar, en lugar de calcular y comparar un conjunto de simulaciones para cada tiempo $j = 1, \ldots, n$, se evalúa de a pasos de largo $\Delta j$; los resultados al final de este documento son obtenidos con $\Delta j = 7$ semanas. La cantidad de simulaciones corridas para cada tiempo $j$ fue fijada en 1000; cada simulación tiene una duración de 3 años, es decir, 156 semanas. En segundo lugar, en lugar de utilizar todo el pasado disponible para entrenar al simulador en cada tiempo $j$, se utilizan sólo los 9 años pasados; esto último se realizó en particular para facilitar la implementación de medidas con los mismos períodos en el SimSEE.

La medida secundaria utilizada por el ISC en (6.6) para comparar distribuciones es la IDP (6.2) con $d = 1, \ldots, 100$. A su vez, la función de divergencia utilizada en la IDPes la Kullback-Leibler (6.4).

Para calcular la potencia generada se utiliza el modelo de planta simplificado fijo en el cual Salto opera en modo "predictivo", turbinando un caudal igual al aporte previsto para ese mismo período de tiempo, y Bonete y Palmar turbinan en modo "umbral", turbinando todo lo que llegua sólo si se supera un umbral fijo que es $1950 hm^3$ para Bonete y $950 hm^3$ para Palmar.

Los años sobre los que se evaluó el sistema comienzan en enero de 1982, desde donde se dispone de información del índice N3.4, y terminan en 2009. Esto da un total de 1404 semanas. Dado $\Delta j = 7$, la evaluación completa del ISC implica calcular la IDPentre 1000 simulaciones para cada uno de los $1404/7 \approx 200$ instantes muestreados.

### 7.2.1. Resultados

La figura 7.1 muestra los resultados de aplicar las medidas de calidad intrínsecas a las tres series Bonete, Palmar y Salto, y finalmente a la serie de potencias generadas por el modelo de generación simplificado descrito anteriormente, a 5 modelos estadísticos distintos ,a saber: CEGH (no cuantificado), Discreto 3, Discreto 4, Discreto 4s1, GSSM (no cuantificado).

[1] http://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices

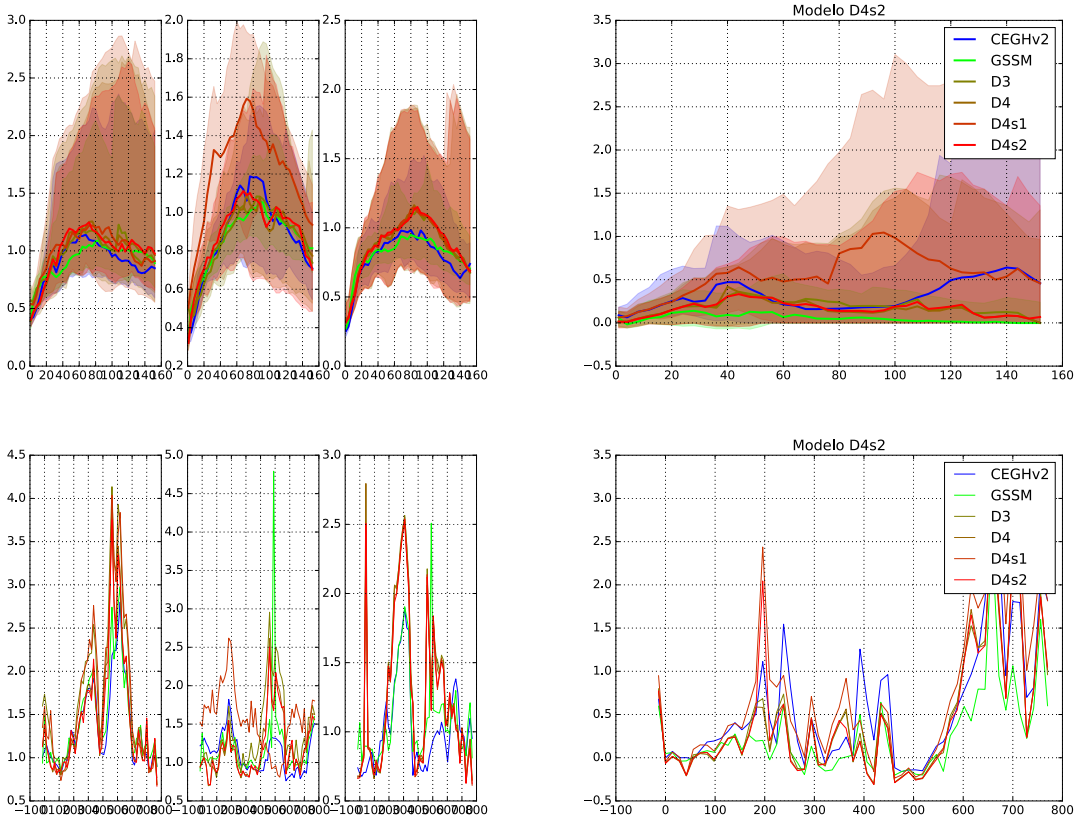


Figura 7.1: Resultados finales de comparación intrínseca de modelos evaluados. Los valores escalares asociados se encuentran en la tabla 7.1. Las dos gráficas superiores corresponden al índice ISC parametrizado según escala (duración de la ventana de promediado, en semanas), para cada una de las series Bonete, Palmar y Salto, y luego en potencia. Las bandas translúcidas corresponden a los percentiles entre el 10 % y 90 % de cada una de las curvas, mientras que las curvas corresponden a la mediana, en ambos casos de 1000 simulaciones tomadas cada 14 semanas entre 1982 y 2009. Las dos gráficas inferiores corresponden a la mediana del desempeño de los modelos en cada una de las semanas anteriormente mencionadas, donde el desempeño (ahora un escalar) es la suma de las curvas ISC con ponderación $w(d) = 1/d$.

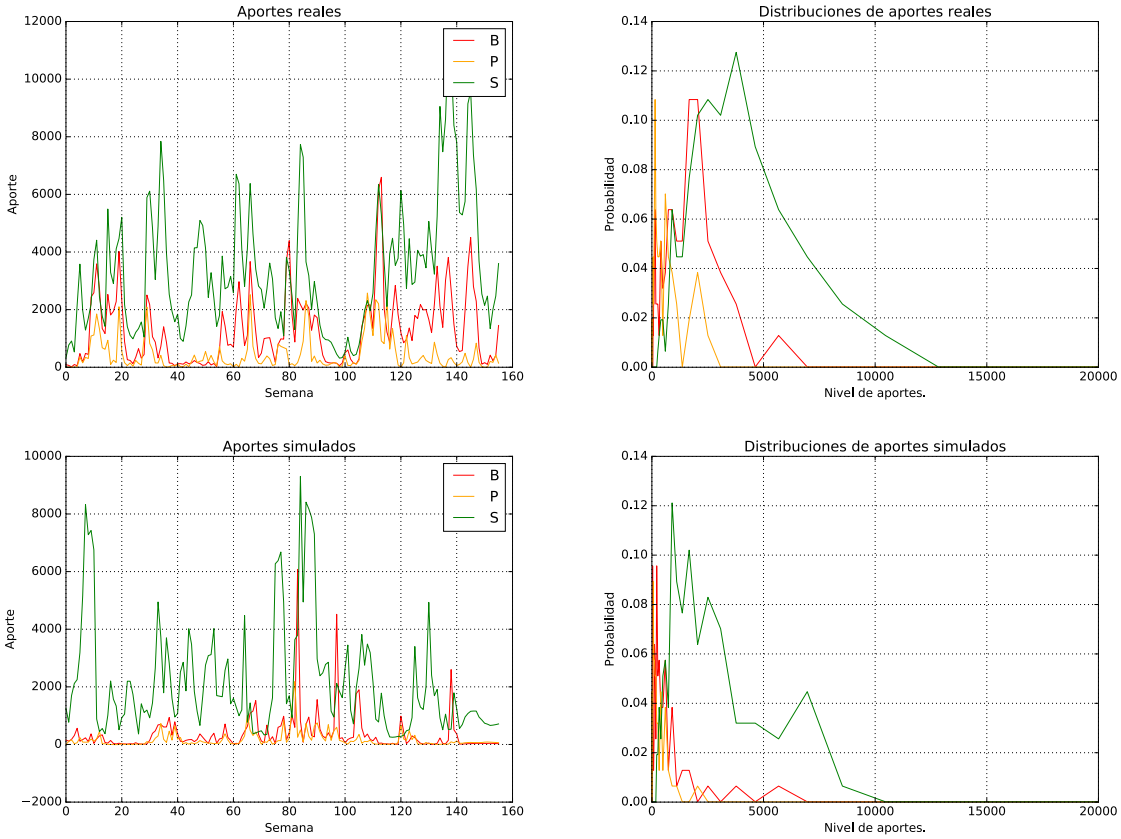


Figura 7.2: Caso particularmente bueno de cercanía entre distribución real y simulada de aportes.

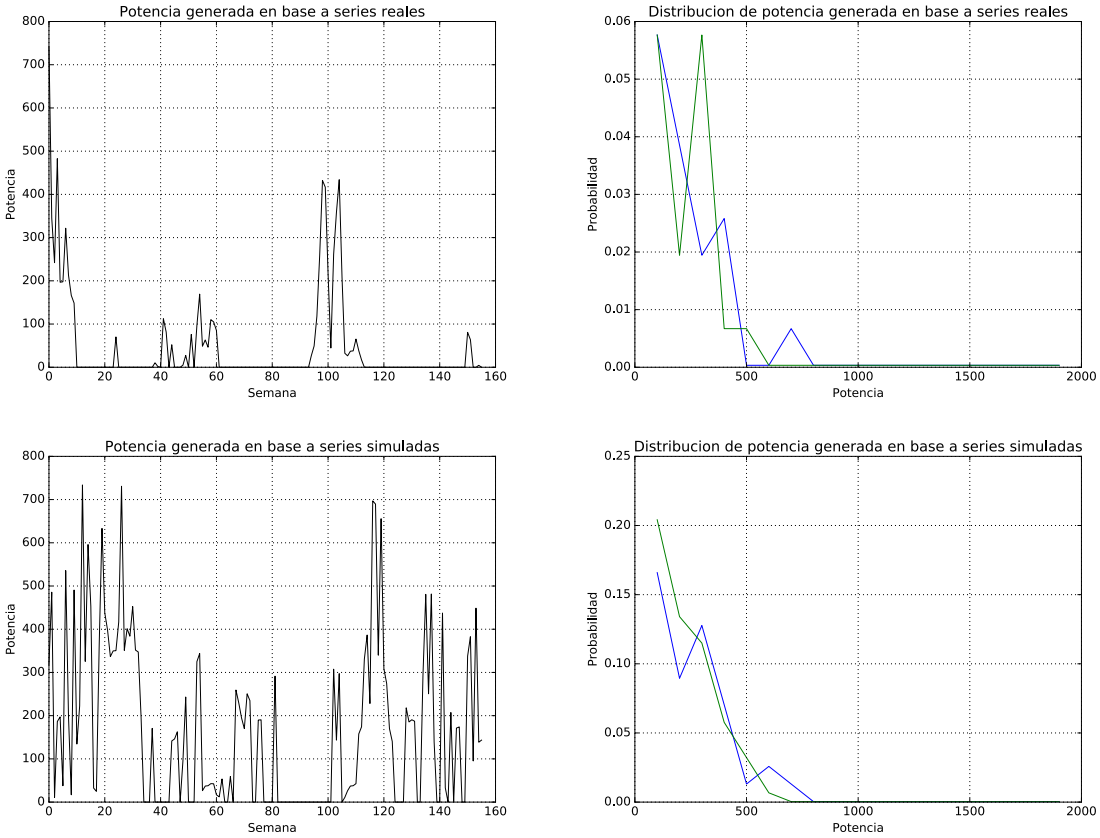


Figura 7.3: Caso particularmente bueno de cercanía entre distribución real y simulada de potencia.

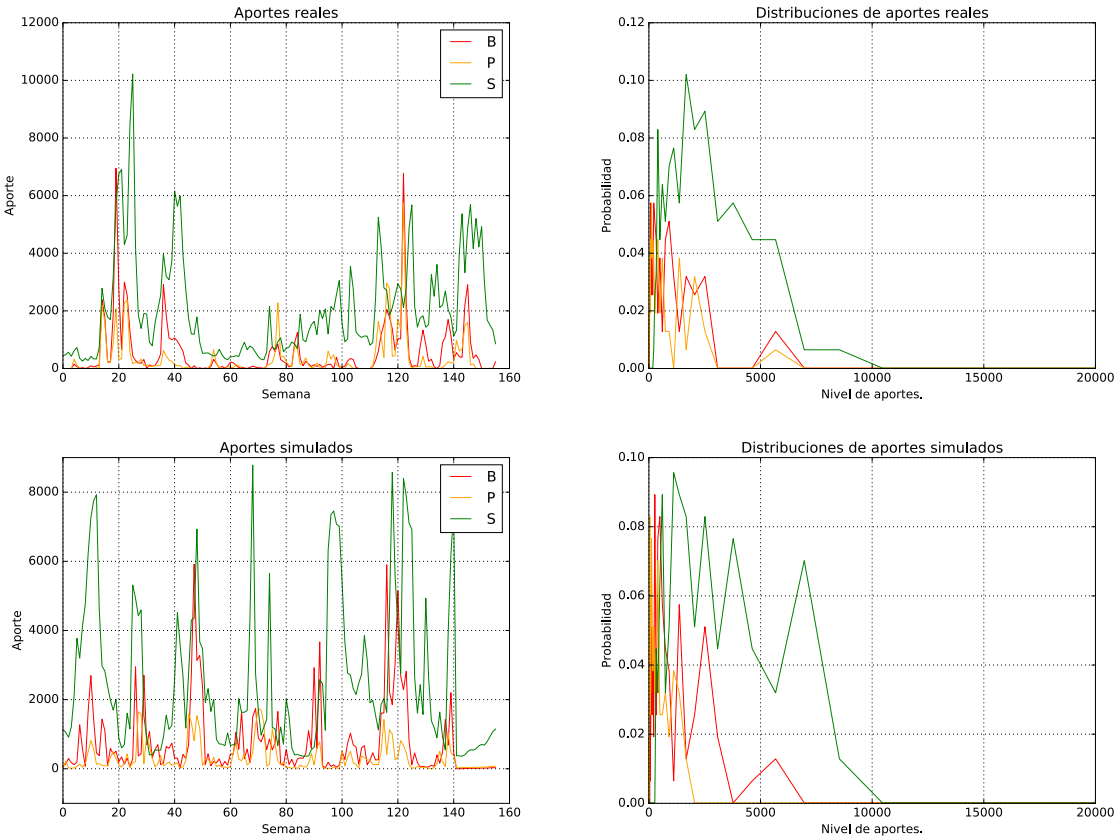


Figura 7.4: Caso particularmente *malo* de cercanía entre distribución real y simulada de aportes.

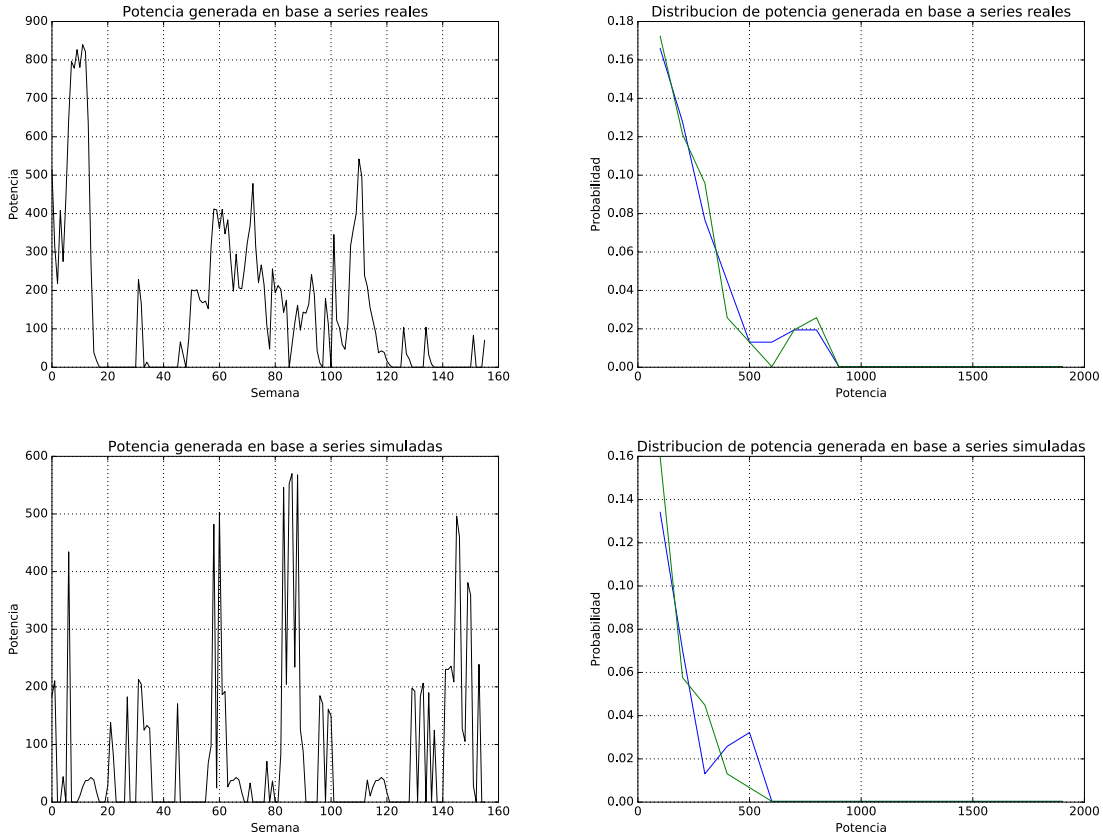


Figura 7.5: Caso particularmente *malo* de cercanía entre distribución real y simulada de potencia.

| | Bonete | Palmar | Salto | Potencia |
|---|---|---|---|---|
| CEGH* | **1.179** | 1.100 | **0.904** | 0.314 |
| GSSM* | 1.224 | 1.084 | 0.963 | **0.109** |
| D3 | 1.436 | 1.217 | 1.055 | 0.155 |
| D4 | 1.200 | **1.054** | 1.065 | 0.141 |
| D4s1 | 1.426 | 1.548 | 1.045 | 0.417 |
| D4s2 | 1.223 | 1.083 | 1.030 | 0.141 |

Tabla 7.1: Resultados finales correspondientes a sumar las curvas de la figura 7.1 con ponderación $w(d) = 1/d$. En negrita se muestran los mejores resultados. Los modelos marcados con $*$ corresponden a modelos continuos sin cuantificar.

Antes que nada, es muy importante tener claro que tanto CEGH y GSSM no tienen ningún tipo de cuantificación, mientras que los otros modelos están fuertemente cuantificados a sólo $q = 7$ niveles distintos. Es natural y esperable que, sin cuantificación alguna, ambos CEGH y GSSM produzcan mejores resultados. Desafortunadamente la implementación del modelo CEGH cuantificado, tal como se lo utiliza en SimSEE, no pudo culminarse a tiempo al final de este proyecto, por lo que no podemos comparar el desempeño de dicho modelo como debería hacerse. Lo que es seguro es que el aplicar la cuantificación al CEGH sólo empeorará y muy significativamente el desempeño observado en la figura anterior. En cuanto al GSSM, se intentó una cuantificación pero tratándose de un modelo aditivo continuo, como era de esperarse, los resultados fueron muy malos. La razón de incluir los modelos CEGH y GSSM en la figura 7.1 es que el problema de la cuantificación es inherente al método de optimización del SimSEE. Con otro mecanismo de optimización es previsible que puedan utilizarse dichos modelos con una cuantificación mucho más fina, por lo que es una buena idea tener como referencia su desempeño relativo en ese régimen.

**Comparación entre modelos discretos** Yendo a los modelos discretos, se observa un desempeño muy similar entre D3, D4 y D4s2, levemente a favor de D4s2. En esta configuración al menos (para esta cantidad de niveles de cuantificación), el D4s1 presenta un desempeño claramente inferior.

**Sobre la métrica de comparación** Una pregunta relevante es qué tan determinante es la métrica utilizada para comparar distribuciones a la hora de decidir entre un modelo y otro. El experimento de la figura 7.6 muestra las diferentes ISC que se obtienen al utilizar las tres medidas implementadas hoy en día: a) divergencia de Kullback-Leibler, b) promedio de norma $\ell_1$ (es decir, norma $\ell_1$ ponderada por la probabilidad de la distribución de referencia), y c) promedio de norma $\ell_2$. (Las curvas en este caso son versiones de menor calidad que las reportadas como "oficiales" al principio de este capítulo). Puede observarse que la medida Kullback-Leibler, que resulta intuitivamente más natural, arroja resultados muy similares a la distancia $\ell_1$. La norma $\ell_2$ produce algunos cambios en las comparaciones, en especial en lo que refiere al modelo D4s1 (que a esta altura debería descartarse) y al CEGH. La tabla 7.2 muestra los resultados de ponderar dichas curvas por $w(d) = 1/d$ como se hizo anteriormente para los casos $\ell_1$ y $\ell_2$.

## 7.3. Marco de evaluación extrínseco

Los siguientes resultados muestran el desempeño obtenido al implementaciones de los modelos X, Y, Z en SimSEE, junto con el CEGH que ya es parte de él. En este caso, se utilizaron 5 años para entrenar, desde 2001 a 2005 inclusive, y se evaluó la política de operación generada con los históricos de aportes y demandas registrados para los años 2006 a 2012 de manera indepen-

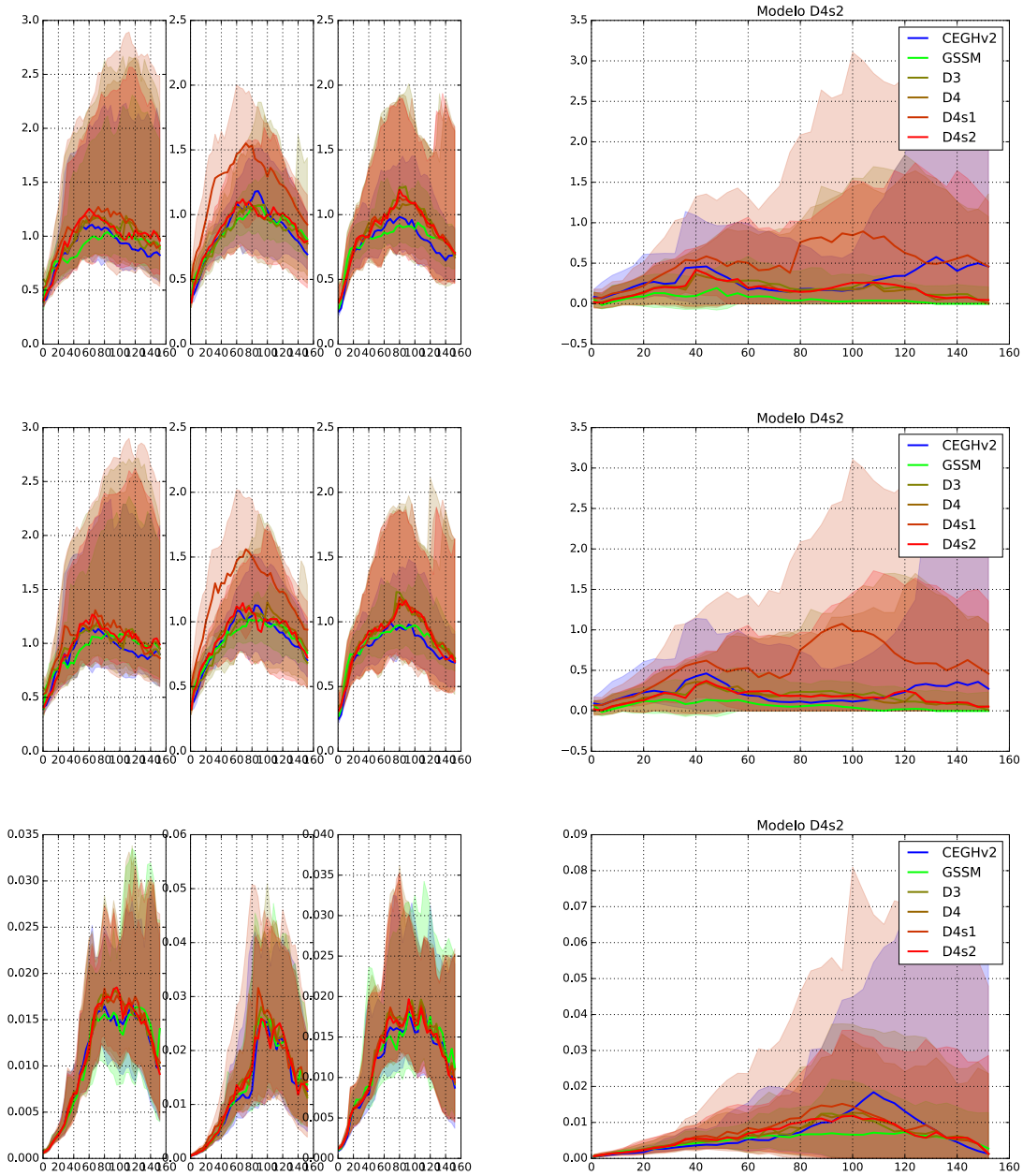


Figura 7.6: **Fila de arriba**: ISC en Bonete, Palmar, Salto y potencia generada utilizando la divergencia de Kullback-Leibler para comparar distribuciones. **Fila del medio**: mismas gráficas utilizando como medida la norma $\ell_1$ ponderada por la distribución de los datos reales. **Fila de abajo**: mismas gráficas utilizando la norma $\ell_2$ ponderada. Notar que los resultados obtenidos en los dos primeros casos son muy similares. La mayor diferencia ocurre con la norma $\ell_2$, que tiende a ser más extrema.

| | norma $\ell_1$ | | | | norma $\ell_2$ ponderada | | | |
|---|---|---|---|---|---|---|---|---|
| | Bonete | Palmar | Salto | Potencia | Bonete | Palmar | Salto | Potencia |
| CEGH* | 0.0737 | **0.0675** | **0.0869** | **0.0352** | 0.00650 | **0.00651** | **0.00813** | 0.00403 |
| GSSM* | **0.0717** | 0.0698 | 0.0891 | 0.0358 | **0.00632** | 0.00699 | 0.00850 | **0.00376** |
| D4 | 0.0731 | 0.0701 | 0.0933 | 0.0416 | 0.00654 | 0.00694 | 0.00898 | 0.00452 |
| D4S1 | 0.0764 | 0.0775 | 0.0932 | 0.0418 | 0.00682 | 0.00786 | 0.00893 | 0.00492 |
| D4s2 | 0.0726 | 0.0691 | 0.0923 | 0.0417 | 0.00660 | 0.00699 | 0.00877 | 0.00461 |

Tabla 7.2: Comparación de medidas de desempeño escalares según métrica utilizada para comparar distribuciones. La mayor diferencia se da, curiosamente, en potencia, donde nuestra implementación del CEGH sin cuantificar pasa de ser el peor a ser el segundo.

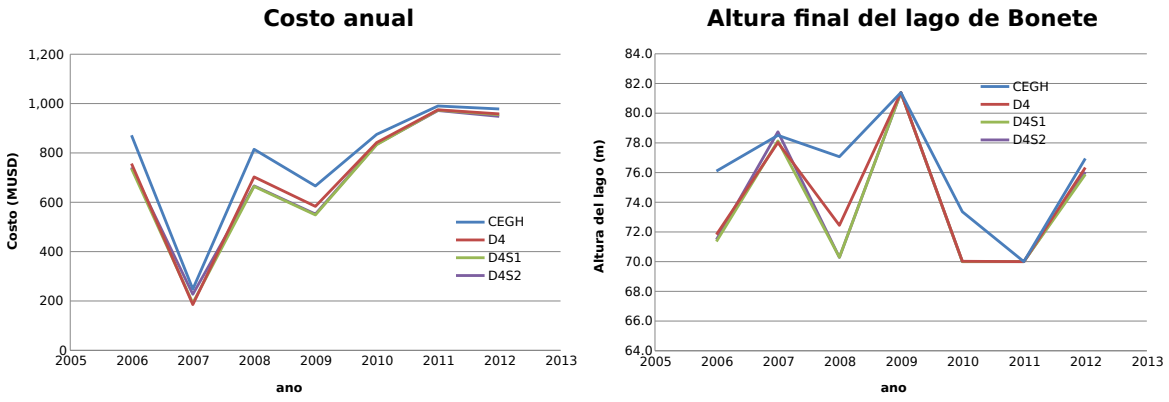


Figura 7.7: Resultados finales de evaluación en SimSEE. Izq.: costo de operación de cada año evaluado en dólares; Der.: altura final del embalse de Bonete en cada año. Las tablas 7.3 y 7.4 muestran estos resultados numéricamente.

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|---|
| CEGH | 872 | 248 | 815 | 666 | 875 | 990 | 978 |
| D4 | 757 | **185** | 703 | 583 | 842 | 975 | 958 |
| D4S1 | **737** | 192 | **664** | **549** | **833** | 974 | 955 |
| D4S2 | 741 | 227 | 666 | 552 | 835 | **972** | **948** |

Tabla 7.3: Costo total de operación en millones de dólares en cada uno de los años evaluados. A ésto se le suma un costo final proporcional a la baja en el nivel del embalse del Bonete (ver tabla 7.4). Contrariamente a lo esperado según los resultados intrínsecos, el modelo D4s1 produjo muy buenos resultados en todos los casos, siendo el mejor en general. Con esa salvedad, se mantiene que D4s2 es levemente mejor que D4, y que estos dos a su vez son mejores que CEGH (sin tomar en cuenta el nivel del embalse de Bonete al final de cada período).

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|---|
| H ini | 78.2 | 76.8 | 79.0 | 77.6 | 81.7 | 78.4 | 79.2 |
| CEGH | **76.1** | 78.5 | **77.1** | 81.4 | **73.4** | 70.0 | **77.0** |
| D4 | 71.8 | 78.0 | 72.4 | 81.4 | 70.0 | 70.0 | 76.3 |
| D4S1 | 71.4 | 78.1 | 70.3 | 81.4 | 70.0 | 70.0 | 75.9 |
| D4S2 | 71.5 | **78.7** | 70.3 | 81.4 | 70.0 | 70.0 | 76.0 |

Tabla 7.4: Nivel final del embalse de Bonete al final de cada período evaluado. Las diferencias a favor de CEGH en este caso son muy importantes en 2006 y 2008, y en menor grado en 2010 y 2012.

diente, de modo que el costo total anual es el reportado por la operación de ese año, mostrado en tabla 7.3, más la variación en el nivel del embalse del Bonete a fin de año, aproximadamente U$S 50 : por metro. Teniendo en cuenta estos dos números, el desempeño final de los modelos propuestos es mixto: para los años pares (2006, 2008, 2010 y 2012), el modelo CEGH supera en desempeño a las variantes ensayadas de los modelos discretos; la diferencia en costo futuro es particularmente alta en 2006 y 2008. Por otro lado, los modelos discretos arrojan ahorros muy importantes en los años 2007 y 2009, y en menor grado en 2011.

# 8 Conclusiones y trabajo futuro

Desde el punto de vista de los objetivos del proyecto, se logró desarrollar y evaluar de manera estadística y en la práctica un conjunto de modelos muy distintos al actualmente utilizado en SimSEE. Como subproducto se obtuvo además un marco de evaluación estadístico de modelos de aportes, incluyendo un modelo simplificado pero razonablemente realista de la operación de una planta hidroeléctrica.

Los modelos desarrollados, tanto sus prototipos en Python así como su implementación en SimSEE, se encuentran y se encontrarán siempre a disposición de la comunidad en el sitio web designado para ello (http://iie.fing.edu.uy/~nacho/simsee/), bajo la licencia GNU Public Licence v.3, lo que garantiza su libre utilización.

Considerando que los modelos obtenidos son prototipos cuyos diversos ajustes y variantes posibles no pudieron ser explorados en esta etapa, los resultados deben ser considerados como muy buenos; en algunos casos se llega incluso a superar el desempeño del sistema actual en años como 2007, 2009 y 2011.

Por otra parte, un estudio más profundo, imposible en los tiempos de este proyecto, será necesario para comprender ciertos resultados inesperados de las implementaciones en SimSEE, en particular el notable gasto en exceso de la reserva del embalse de Bonete, como puede observarse en la tabla 7.4 para los años 2006, 2008 y 2010; en esos casos, la pérdida de dicho nivel resulta en un costo total futuro superior al obtenido con CEGH.

Otro resultado no esperado es la discrepancia en desempeño intrínseco y extrínseco del modelo D4S1; esto no sucede con los otros modelos ensayados (incluso en algunos modelos intermedios no representativos que no fueron incluidos en este reporte).

Como trabajo futuro obvio queda entonces el refinamiento de los modelos desarrollados, así como la comprensión cabal de los fenómenos anteriormente mencionados. También queda a futuro el desarrollo de una versión cuantificada del modelo en espacio de estados GSSM, que sólo pudo evaluarse estadísticamente.

# Bibliografía


[1] G. Casaravilla, R. Chaer, and P. Alfaro. SimSEE - memoria final de ejecución proyecto pdt 47/12. Technical report, Universidad de la República. Facultad de Ingeniería. Instituto de Ingeniería Eléctrica, 2008.

[2] R. Chaer. Fundamentos del modelo CEGH de procesos estocásticos multivariables. Technical report, Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, Jun. 2013.

[3] R. Chaer, R. Terra, A. Díaz, and J. Zorrilla de San Martín. Considering the information of the niño 3.4 index in the operation of the electrical system of uruguay. In *International Association for Energy Economics International Conference, 33rd, (IAEE 2010). Rio de Janeiro, Brasil, 6-9 June 2010*, pages 1–14, 2010.

[4] Ruben Chaer. Simulación de sistemas de energía eléctrica. Master's thesis, Instituto de Ingeniería Eléctrica. Facultad de Ingeniería. Universidad de la República, 2008.

[5] Ruben Chaer, Rafael Terra, Alvaro Díaz, and Alvaro Brandino. Aproximación al modelado de los aportes hidráulicos a las represas del uruguay teniendo en cuenta el índice Niño 3.4. In *Encuentro de Potencia, Instrumentación y Medidas, EPIM 08. Montevideo, Uruguay.*, pages 52–57, 2008.

[6] E. Coppes, F. Barreto, C.Tutté, F. Maciel, M. Forets, E. Cornalino, M. Gurín Añasco, M. C. Álvarez, F. Palacios, D. Cohn, and R. Chaer. Memoria de proyecto FSE 2009. Technical report, Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, 2009.

[7] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, Inc., 2 edition, 2006.

[8] S. Durbin, J.; Koopman. *Time series analysis by state space methods.* Oxford, UK: Oxford University Press, 2001.

[9] G. Golub and C. van Loan. *Matrix Computations.* JHU Press, 3rd edition, 1996.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, 2nd edition, Feb. 2009.

[11] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, 3rd edition, 1995.

[12] P. Kall and S. W. Wallace. *Stochastic Programming.* John Wiley and Sons, second. edition, 1994.

[13] D. Kraft. A software package for sequential quadratic programming. Technical report, Aerospace Center — Institute for Flight Mechanics., Koln, Germany., 1988.

[14] F. Maciel, R. Terra, and A. Díaz. Incorporación de información climática en la simulación de aportes a represas en un modelo del sistema eléctrico. In *Congreso Latinoamericano de Hidráulica. (25º. : 9-12 Set. 2012 : San José, Costa Rica)*, 2013.

[15] F. Maciel, R. Terra, and A. Díaz. Mejoras en la simulación de aportes a represas hidroeléctricas para su incorporación a modelos de planificación energética. Technical report, Universidad de la República, 2009.

[16] Aldo Montecinos, Alvaro Díaz, and Patricio Aceituno. Seasonal diagnostic and predictability of rainfall in subtropical south america based on tropical pacific sst. *Journal of Climate*, 13(4):746–758, 2000.

[17] Gabriel Pisciottano, Alvaro Díaz, Gabriel Cazess, and Carlos R Mechoso. El niño-southern oscillation impact on rainfall in uruguay. *Journal of Climate*, 7(8):1286–1302, 1994.

[18] W. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality, 2nd Edition (Wiley Series in Probability and Statistics)*. John Wiley & Sons, 2nd. edition, Sept. 2011.