Introducción al Reconocimiento de Patrones

Predicción a un año de la situación hidrológica basado en Redes Neuronales.

Trabajo Final

Ing. Enrique Briglia
Proyecto GNL
UTE
Montevideo - Uruguay
ebriglia@ute.com.uy

Ing. Elias Carnelli
Sub Gerencia de Ingeniería de Protecciones
UTE
Montevideo - Uruguay
ecarnelli@ute.com.uy

Ing. Fernando Ron
Proyecto GNL - Sistemas de Gestión de Explotación
UTE
Montevideo - Uruguay
fron@ute.com.uy

***Abstract*—La posible instalación de una planta de regasificación en el Uruguay para ser utilizada en gran medida para suministrar combustible a generadores térmicos conectados a un sistema con alta dependencia del recurso hídrico, plantea como desafío el lograr realizar las mejores compras posibles de GNL con un año de anticipación. Para ello es de extrema importancia poder predecir la situación hidrológica del país para cada uno de los cuatro trimestres del siguiente año, a efectos de no realizar compras excesivas ni caer en una situación de déficit de combustible. El presente trabajo aplica diferentes técnicas para clasificar cada uno de los trimestres en Muy Seco , Seco, Medio, Húmedo y Muy Húmedo. Aplicaremos redes neuronales a la información histórica de lluvias incorporando posteriormente predicciones climáticas de NOA.**

IndexTerms-reconocimiento; patrones; predicción; clima; hidraulicidad; GNL;

Contenido

INTRODUCCIÓN

MOTIVACIÓN Y OBJETIVO

Uruguayse plantea la necesidad de realizar inversiones en generación eléctrica que acompañencrecimiento futuro previsto de la demanda. Para ello se maneja la alternativade incorporar, al actual Sistema Interconectado Nacional (SIN), generadores que exploten recursos renovables no tradicionales(fundamentalmente eólicos) que se complementarán con generadores térmicos a base de Gas Natural (GN) como combustible.A efectos de asegurar el suministro de combustible para estos generadores es que se proyecta lainstalación de una Planta de Regasificación de Gas Natural Licuado (PRGNL).

La instalación de la mencionada planta, requiere la realización de estudios técnicos para llevar adelante el proyecto siendo este el contextoen el que se plantea el presente trabajo.

Para la realización de los mencionados estudios se modela enSimSee (Facultad de Ingeniería s.f.)el SIN integrando las diferentes características (o restricciones) que una PRGNL presenta. Una de las restricciones más importantes introducidas por la PRGNL radica en la modalidad de compra de GNL, bajo el formato "TakeorPay" (ToP), que establece la necesidad de fijar, con un año de anticipación, un Plan Anual de Embarques (PAE) en el que Uruguay se compromete a comprar una determinada cantidad de cargamento de GNL a recibir en una fecha que se establece en el propio PAE.

Las características propias del SIN, hacen que la necesidades futuras de GNL sean dependientes de la disponibilidad futura del recurso hídrico. A modo de ejemplo, un año húmedo (es decir, que existirá mucho recurso hídrico) requerirá poco GNL, mientras que un año seco, requerirá mucho GNL. El problema radica entonces en estimar como se presentará el próximo año (como serán los siguientes cuatro trimestres? secos?, húmedos?) para poder acordar un PAE que no comprometa compras de combustible que se aparten demasiado de lo que ocurra en realidad, es decir, que no resulte en escases futura ni sobrantes de GNL. Si podemos lograr esto, habremos reducido significativamente el Costo de Abastecimiento de la Demanda (CAD).

El objetivo, entonces, de este trabajo, es lograr categorizar cada uno de los trimestres del siguiente año en 5 categorías, a saber: MUY SECO, SECO, MEDIO, HUMEDO, MUY HUMEDO. No es un objetivo de este trabajo predecir exactamente las lluvias con un año de anticipación ni determinar un PAE. Lo que se pretende obtener es una herramienta que en el futuro pueda ser utilizada para decidir el mejor PAE posible.

PLANTEO DEL PROBLEMA

El problema consistirá en resolver las etapas de Recolección de Datos, División de Datos, Pre procesamiento, Identificación de modelos y Evaluación con el objetivo de predecir la situación hidrológica de cada uno de los siguientes cuatro trimestres.Si bien existen diferentes modelos para predicción, el trabajo siguiente se basará en la utilización de Redes Neuronales Artificiales (RNA) como herramienta de predicción a partir de la información histórica disponible.

DESCRIPCIÓN DEL SISTEMA INTEGRADO NACIONAL

El sistema eléctrico uruguayo está constituido por un sistema de generación autónoma distribuida en el país de dimensión marginal y por el denominado Sistema Interconectado Nacional (SIN) fundamentalmente hidrotérmico.

En la Figura 1 se muestra, cuáles fueron las fuentes de generación eléctrica utilizadas en el año 2010 para suministrar la demanda y como esta fue abastecida en un 84% por el recurso hidráulico.La generación hidráulica,por su parte, se reparte en 4 centrales hidroeléctricas, por un lado Salto Grande (con gran capacidad instalada pero muy poca capacidad de reserva) y por otro lado 3 centrales en la cuenca del Río Negro de las cuales la Central Gabriel Terra posee la mayor capacidad de reserva de energía (equivalente a aprox. 4 meses de generación).

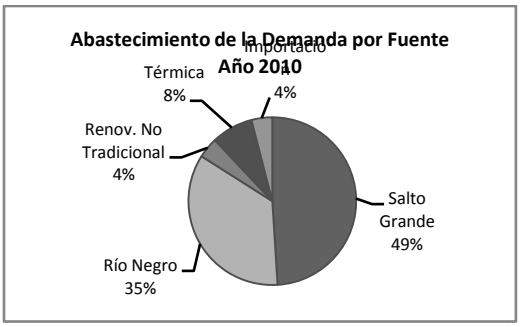


Figura 1 - Abastecimiento de la Demanda por Fuente para el año 2010

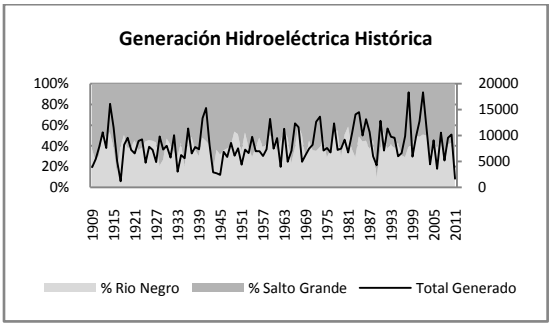


Figura 2 - Generación Hidroeléctrica Histórica del Uruguay

Un punto importante a considerar es la relación existente entre la energía generada en Salto Grande con respecto a la generada en Rio Negro (potencia vs. almacenamiento). Si observamos la información histórica que se muestra en la Figura 2, se puede apreciar, en el eje principal de la gráfica, que las centrales hidroeléctricas del Río Negro aportan en promedio aproximadamente el 40% del total de generación hidráulica.

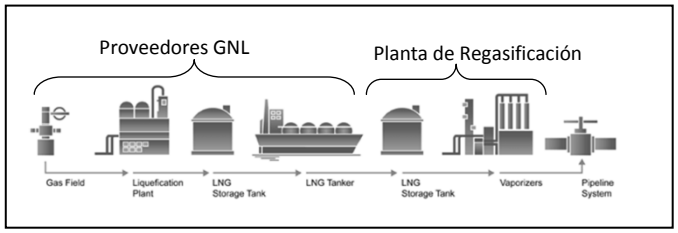


Figura 3- Cadena del GNL (6)

También en la Figura 2, pero en el eje secundario,se puede observar el total de generación hidroeléctrica y la variabilidad que esta presenta en función de la aleatoriedad hidráulica que viene dada por los aportes a los diferentes embalses de las precipitaciones sobre las cuencas del Río Uruguay y Río Negro.

PRINCIPALES RESTRICCIONES DE UNA PLANTA DE REGASIFICACIÓN.

La instalación de una PRGNL introduce la cadena del GNL que puede ser dividida, a efectos de este trabajo, en dos partes:

(a) Proveedores de GNL. Estos imponen restricciones relacionadas con la forma en que se debe comprar el GNL debiendo ser en modalidad TakeOrPay (ToP). Estos contratos, requieren que se estipule, con un año de anticipación, la agenda de entrega de embarques para el siguiente año definiendo el volumen de cada cargamento y la fecha en que este debe ser entregado. Permiten cierta flexibilidad para comprar cargamentos adicionales siempre que se soliciten con 90 días de anticipación.

(b) Planta de Regasificación. Introduce, como restricción fundamental, el concepto de tanque, con una capacidad finita de almacenamiento de GNL. Estos tanques tendrán una relación maxima 2:1 (solo es posible almacenar hasta dos cargamentos en el tanque. Hay que acotar que si al llegar un cargamento, el tanque no tiene espacio suficiente, este se pierde y por lo tanto no se tendrá otro hasta el siguiente cargamento estipulado en el PAE.

Los modelos de optimización y simulación mediante programación dinámica estocástica, implementan estas restricciones y simulan el comportamiento del sistema para un período de 15 años tomando 100 crónicas hidrológicas dadas por las lluvias históricas registradas. Para estas simulaciones se debe ingresar una agenda para todo el período de la simulación.

Actualmente, para poder determinar las agendas de la simulación,primeramente se determinan (sin considerar las restricciones impuestas por la PRGNL) las necesidades de combustible para cada crónica y para los 15 años y se calcula el valor esperado de los consumos de GN. Con este consumo medio se realiza una agenda en la que programamos un cargamento cada vez que se consume un volumen equivalente a la capacidad de cada cargamento.

En la Tabla 1 se puede apreciar una agenda donde figuran la cantidad de cargamentos a solicitar para cada mes de cada uno de los años. Se puede apreciar que la cantidad de cargamentos requeridos son relativamente pocos(perocon mucha energía).

Con la agenda determinada de esta manera, se procede a alimentar el modeloincluyendo todas las restricciones para evaluar así el comportamiento del sistema.

A modo de ejemplo, tomemos los resultados de la simulación para el año 2018 (en el cual se fijó una agenda 7 cargamentos) y procedemos a evaluar cómo se suministra la demanda para cada una de las crónicas hidrológicas. Para ello ordenaremos las crónicas hidrológicas (de MuyHúmedas a Muy Secas)tal como se muestra en la Figura 4.

Cargamentos	**Mes**											
Año	1	2	3	4	5	6	7	8	9	10	11	12
2014			1			1						
2015		1		1		1		1				1
2016		1		1		1		1				1
2017	1		1	1		1		1			1	
2018	1	1		1	1		1		1			1
2019	1	1	1		1	1		1			1	1
2020	1		1	1		1	1		1		1	
2021	2	1	1		1	1	1		1		1	1
2022	1	1	1	1		1	1	1		1		1
2023	2	1	1	1	1		1	1	1		1	1
2024	2	1	1	1	1	1	1		1	1	1	1
2025	1	1	2		2	1	1	1		1	1	1
2026	2	1	2	1	1	1	1	1	1		1	2
2027	1	2	1	1	1	2	1	1	1		2	1
2028	2	1	2	1	1	2	1	1	1	1	1	2
2029	1	2	1	2	1	1	2	1	1	1	1	2

Tabla 1 - Agenda de cargamentos simulada a 15 años

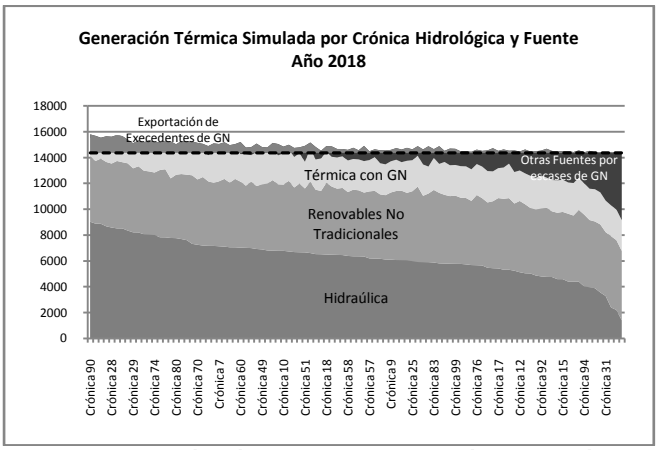


Figura 4 - Generación Térmica Simulada por Crónica Hidrológica y por Fuente para el año 2018. (Crónicas ordenadas de Muy Húmedas a Muy Seca).

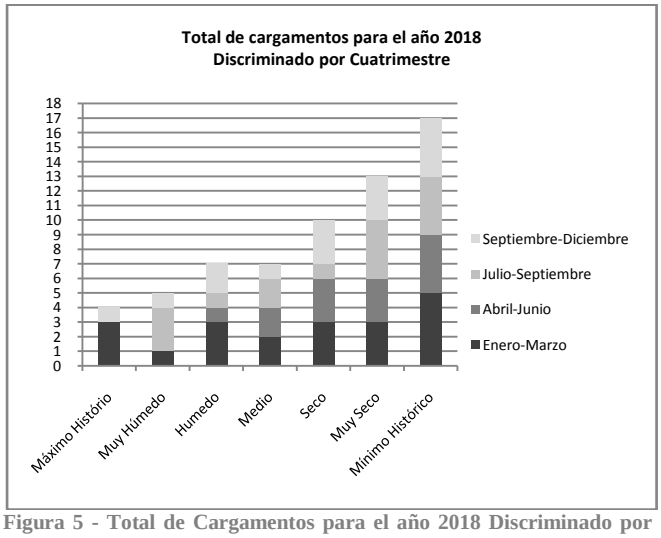


Figura 5 - Total de Cargamentos para el año 2018 Discriminado por Cuatrimestre. Estos datos son el resultado de la simulación con las precipitaciones históricas.

La agenda se determinó sin hacer pronósticos sobre la situación hidrológica futura, introduciendo el mismo volumen de GNL en cada crónica, provocando así que en los años húmedos se produzcan excedentes, mientras que en los secos el sistema debe recurrir a otros combustibles más costosos para suministrar la demanda eléctrica. Si esto se traduce a costos, el armar una agenda de esta manera dejándola librada luego a la suerte de los procesos estocásticos, podemos llegar a duplicar el CAD por tener faltantes de combustible (comparando Secos con Húmedos)

Para analizar la cantidad de cargamentos que debieron solicitarse según la crónica hidrológica que se presente seleccionamos las crónicas correspondientes a los percentiles 15, 30, 50, 70, 85 conjuntamente con los máximos y mínimos históricos de precipitaciones.

La Figura 5 muestra que si nos centramos en el 15 % de las crónicas más húmedas, vemos que serían necesarios 5 cargamentos para el año y 4 cargamentos para un año equivalente al más húmedo de la historia del Uruguay.Si por el contrario, nos situamos en el 15% de las crónicas más secas, la cantidad de cargamentos necesarios para el año sería de 13, siendo el 17 el máximo dado por un año equivalente al más seco de la historia del Uruguay.

Es decir, si se planea una agenda para un año medio, y se presenta un año muy húmedo, tendríamos un excedente de 2 cargamentos, pero si el año que se presenta es seco, necesitaríamos 7 cargamentos más. Esta última es la situación más problemática para el SIN.

Además de la cantidad de cargamentos, es necesario determinar en cuando estos cargamentos serán necesarios, lo que estará determinado por la estacionalidad conocida de la demanda eléctrica y fundamentalmente por la hidraulicidad.

Tenemos entonces que sería extremadamente conveniente, disponer de más información que permita cada año armar una agenda, más adecuada a las previsiones climáticas de ese mismo año, que incurra en los menores sobrecostos del abastecimiento de la demanda. No solo esto, que la clasificación no debe ser exacta como para predecir las lluvias, sino que con predecir el nivel de hidraulicidad (de Muy Seco a Muy Húmedo) sería suficiente, siendo en definitiva esto lo que lo que motiva este trabajo.

ANTECEDENTES

A partir del primer modelo matemático hidrológico desarrollado por Mulvany en 1850 (On the use of self registering rain and flow gauges, (1850)), se han desarrollado una serie de modelos que, dependiendo de cómo describen internamente los procesos hidrológicos, se pueden clasificar en modelos basados en el conocimiento o basados en los datos.

Los procesos basados en el conocimiento, tanto los mecánicos como los conceptuales, requieren que se describa el comportamiento y las iteraciones de los procesos que suceden tanto en la superficie de los ríos como en el subsuelo, debiendo modelar principios como conservación de masa, momento y energía, en conjunto con flujos, precipitaciones, filtración, evaporación, escurrimiento, drenaje, etc. Algunos ejemplos de estos modelos son el SystemaHydrologiqueEuropeen (SHE) (Abbot et al. 1986ª, b), Institute of HydrologyDistributedModel (IHDM) (Beven, et al., 1987, Calver and Wood, 1995), (Wagener et al. 2004)

Por lo contrario, los procesos basados en los datos,capturan el mapeo existente entre las entradas (ej. precipitaciones, evaporación, temperatura, etc.) y las salidas (aportes) basándose solamente en la información recolectada y sin considerar directamente los procesos físicos que rigen el sistema. Estos modelos pueden ser divididos en dos tipos (Jain and Srinivasulu, 2004):

(i) Modelos con estructura predefinida pero parámetros desconocidos. Se basan en asumir que las series temporales son estocásticas, normales y procesos invariantes en el tiempo.

(ii) Modelos con estructura y parámetros desconocidos. Estos toman ventaja de las series temporales para derivar la estructura del modelo y sus correspondientes parámetros.

Algunos ejemplos del primer tipo de modelos son lo modelos de regresión lineal y no lineal (box and Jenkins, 1976; Salas et al. , 1985), sistemas lineales restringidos (Natale and Todini, 1976), modelos de perturbaciones lineales (Nash and Barsi, 1983), modelos de funciones de trasferencia (Kachroo, 1992ª; Kachroo, 1992; Liang et al, 1992)

El segundo tipo incluye métodos no-paramétricos como k-nearest-neighbor (K-NN) (Karlsson y Yakowist, 1997ª, Solomatine et al 2008), métodos basados en sistemas dinámicos (DBMS) (Jayawardena and Lai, 1994), sistemas de inferencia difusa (Liong et al. 2000; Jacquin and Shamseldin, 2006), regresión con vectores de soporte (SVR) (Sivapragasam et al., 2001; Yu et al., 2006) y programación genética (Koza 1992, Babovic and Keijer, 2000), redes neuronales artificiales (RNAs) (Hsu et al., 1995; Solomatine and Dulal, 2003; Khan and Coulibaly, 2006),

Los modelos basados en datos surgen como fáciles y rápidos dado que la calibración de los parámetros está relacionada exclusivamente con la estructura del modelo y de las entradas sin considerar los procesos hidrológicos que ocurren, siendo adecuados para identificar un mapeo directo entre entradas y salidas.

El presente trabajo utiliza los modelos basados en datos, particularmente en las redes neuronales artificiales como modelo de predicción.Para ello, deberemos, a partir de los datos disponibles, identificar cuáles serán las entradas y las salidas más adecuadas para el problemade predecir la situación hidrológica (muy seco, seco, medio, etc.) a un año discriminando por semestre/trimestre.

Se tomó como trabajo de referencia fundamental el realizado por (C. C. Wu 2010) en que se propone un modelo modular para la predicción hidrológica basado en Redes Neuronales y utilizando diferentes técnicas para la selección de entradas y pre-procesamiento de los datos.

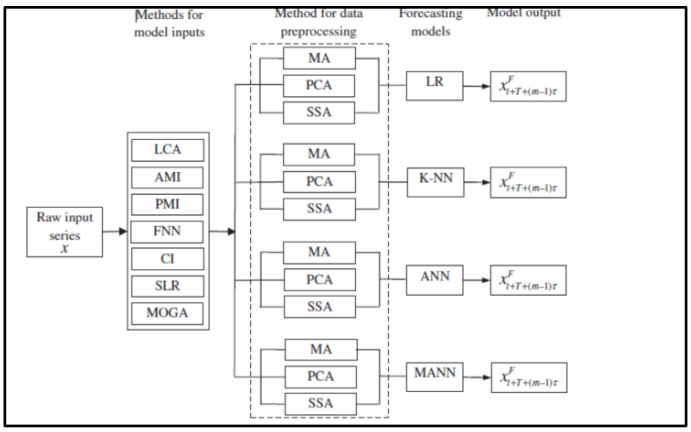


Figura 6 - Modelo de C.L.WU y K.W.Chau

I. CONCEPTOS BÁSICOS DE TÉCNICAS APLICADAS

CLUSTERING

Las técnicas de agrupamiento, buscan descubrir una estructura dentro de un conjunto de datos dividiéndolo en subconjuntos que tengan cierta *coherencia*. , entendiendo como tal al hecho de que muestras dentro del mismo cluster sean lo más *parecidas* posibles para lo cual es necesario definir una noción de similitud entre muestras.

El algoritmo Fuzzy C-means (FCM) le asigna a cada muestra x_j un grado de "pertenencia" a cada cluster dado por una probabilidad:

$$\hat{P}(\omega_i|\mathbf{x_j})$$

Con $\sum_{i=1}^{c} \hat{P}(\omega_j|\mathbf{x_j}) = 1; \forall j = 1...n$

La función a minimizar es :

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n}\left(\hat{P}(\omega_i|\mathbf{x_j})\right)^b \left\|\mathbf{x_j} - \boldsymbol{\mu_i}\right\|^2$$

Se plantea entonces

$$\frac{\partial J}{\partial \mu_i} = 0 \text{ y } \frac{\partial J}{\partial \hat{P}_j} = 0$$

Que tiene como solución:

$$\boldsymbol{\mu_i} = \frac{\sum_{j=1}^{n}\left(\hat{P}(\omega_i|\mathbf{x_j})\right)^b \mathbf{x_j}}{\left(\hat{P}(\omega_i|\mathbf{x_j})\right)^b}$$

$$\hat{P}(\omega_i|\mathbf{x_j}) = \frac{1}{\sum_{r=1}^{c}\left(\frac{\left\|\mathbf{x_j} - \boldsymbol{\mu_i}\right\|^2}{\left\|\mathbf{x_j} - \boldsymbol{\mu}_r\right\|^2}\right)^{\frac{1}{b-1}}}$$

El parámetro "*b*" determina el grado de mezcla entre los distintos clusters, si *b=0* no se aceptan mezclas y por lo tanto es k-means. Los centros de los clusters $\boldsymbol{\mu_i}$ se encuentran cerca de los puntos que tienen alta probabilidad estimada de estar en el cluster. Como difícilmente se encuentren soluciones analíticas, la estimación de los parámetros se realiza de acuerdo al siguiente algoritmo.

- Se inicializan $\hat{P}(\omega_i|\mathbf{x_j})$ teniendo en cuenta que $\sum_{i=1}^{c} \hat{P}(\omega_i|\mathbf{x_j}) = 1; \forall j = 1...n$
- Calculo de $\boldsymbol{\mu_i}$ a partir de

$$\boldsymbol{\mu_i} = \frac{\sum_{j=1}^{n}\left(\hat{P}(\omega_i|\mathbf{x_j})\right)^b \mathbf{x_j}}{\left(\hat{P}(\omega_i|\mathbf{x_j})\right)^b}$$

- Actualización de $\hat{P}(\omega_j|\mathbf{x_j})$ a partir de

$$\hat{P}(\omega_i|\mathbf{x_j}) = \left(\sum_{r=1}^{c}\left(\frac{\left\|\mathbf{x_j} - \boldsymbol{\mu_i}\right\|^2}{\left\|\mathbf{x_j} - \boldsymbol{\mu}_r\right\|^2}\right)^{\frac{1}{b-1}}\right)^{-1}$$

- Si los $\hat{P}(\omega_i|\mathbf{x_j})$ y los $\boldsymbol{\mu_i}$ varían menos de un cierto umbral con respecto a la iteración anterior se finaliza el proceso; de lo contrario volver al paso 2

Debemos hacer notar que los resultados pueden ser sensibles a la cantidad de clusters a buscar, a las condiciones iniciales y al parámetro "*b*" correspondiente al exponente de la función a minimizar.

VALIDACIÓN DE CLUSTERS

La validación de los clusters refiere al problema de que tan bien una partición se adapta a los datos. El algoritmo de clustering siempre intenta encontrar la partición que ajuste mejor dado una cantidad de clusters a buscar y los diferentes parámetros del algoritmo. Sin embargo esto no significa que sea la mejor solución pues tanto la cantidad de clusters a buscar como los parámetros del algoritmo pueden no corresponder a los mejores grupos que se pueden obtener de los datos.

Hay dos aproximaciones para determinar el número apropiado de clusters:

- Comenzando con un número suficientemente grande de clusters y sucesivamente reducir este número uniendo clusters similares con algún criterio predeterminado. (Compatible ClusterMergin)
- Generar diferentes cantidades de clusters y utilizar métricas que permitan medir la bondad de los mismos. (aproximación que tomaremos es este trabajo)

La literatura propone una variedad de índices scalares que permiten medir la bonad de los diferentes clusters, pero ninguno de ellos es perfecto por sí mismo, por lo que se hace necesario evaluar un conjunto de ellos. En este trabajo evaluaremos las siguientes métricas:

- ***PartitionCoeficient (PC)***: mide la cantidad de "solapamientos" entre los clusters (Bezdek 1981) yse define como:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m$$

Siendo μ_{ij} la pertenencia de la muestra *j* al cluster *i*. y *m* el parámetro exponente de la función a minimizar.

A mayores valores se tendrán las mejores particiónes.

- ***ClasificationEntropy (CE):*** mide que tan difusa es la partición (similar al PartitionCoeficient) y se define como:

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log(\mu_{ij})$$

Siendo μ_{ij} la pertenencia de la muestra *j* al cluster *i*.

A mayores valores se tendrán las mejores particiónes.

- ***PartitionIndex (SC)***: es la razón de la suma de la compactación y la separación de los clusters y se define como

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^m \left\|x_j - v_i\right\|^2}{N_i \sum_{k=1}^{C}\|v_k - v_i\|^2}$$

Este índice es útil cuando se comparan diferentes particiones con igual número de clusters.

A menores valores de SC se tendrán mejores partición.

- ***SeparationIndex (S)***: por el contrario del PartitionIndex (SC), este índice utiliza la mínima distancia de separación para validar la partición

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m \left\|x_j - v_i\right\|^2}{N\, min_{i,k}\|v_k - v_i\|^2}$$

A menores valores de SC se tendrán mejores partición.

Para evaluar estas métricas, se realizarán un conjunto de iteraciones y en cada una de ellas se evaluará el valor esperado de cada una de estas métricas y su desviación estándar como medida de error.

Estas métricas se complementarán con el porcentaje de muestras que cambian de cluster en el conjunto de iteraciones.

En términos generales se procederá de la siguiente forma:

- *Iterar para cada valor de m*
  - *Iterar para cada valor de c*
    - *Realizar n iteraciones ejecutando el algoritmo de FuzzyKmeans evaluando las métricas.*
    - *Determinar que muestras cambiaron de cluster en las n iteraciones y que porcentaje de muestras sobre el total son las que cambian de cluster.*
    - *Calcular el valor esperado y la desviación estándar de cada métrica entre las n iteraciones.*

CORRELACIÓN

Dada una serie temporal $(Y_t, Y_{t-1},...,Y_{t-N})$ se tienen las siguientes definiciones:

Media: $\mu_Y = E[Y_t]$

Varianza: $\sigma_Y = E[(Y_t - \mu_Y)^2]$

Autocovarianza de orden *"k"*:

$$\gamma_k = E[(Y_t - \mu_Y)(Y_{t-k} - \mu_Y)] \quad k = 1,2,...$$

Observación: $\gamma_0 = \sigma_Y$

Auto correlación simple de orden *"k"*:

$$\rho_k = \frac{E[(Y_t - \mu_Y)(Y_{t-k} - \mu_Y)]}{(E[(Y_t - \mu_Y)^2])^{1/2}(E[(Y_{t-k} - \mu_Y)^2])^{1/2}} = \frac{\gamma_k}{\sigma_Y} = \frac{\gamma_k}{\gamma_0} \quad k = 1,2,...$$

El primer mínimo de la función de auto correlación puede ser un buena guía para determinar cuales retardos (lags) son relevantes para ser identificados como entradas a una la red neuronal. Si se elige demasiado pequeño, los elementos serán muy parecidos y no aportarán nueva información al proceso de reconstrucción, mientras que si se elige un retardo demasiado grande no se podrá establecer la relación existente entre las variables retardadas, puesto que no estarán correlacionadas. Típicamente se elige $\rho_k < 0.2$

Una alternativa analizar los mínimos de la función AMI: "Averaged Mutual Information"con la siguiente definición:

$$I(k) = \sum_{t,t+\tau}^{t-k} P(x_t, x_{t-k}) \log\left[\frac{P(x_t, x_{t-k})}{P(x_t)P(x_{t-k})}\right]$$

Como en el caso anterior se busca que los valores de x_t, x_{t-k} sean suficientemente independientes, pero no tan independientes para que no exista ninguna correlación entre ellos

CONCEPTOS BÁSICOS DE REDES NEURONALES ARTIFICIALES (RNA)

Las RNA pueden pensarse como una alternativa para levantar las limitaciones de las funciones discriminantes lineales que provienen de las propiedades impuestas a las funciones que determinan la frontera de separación. Con ellas se pueden elegir funciones no lineales más adecuadas que permiten resolver problemas mucho más complejos.

El problema central es elegir cuales son las funciones no lineales adecuadas.

Redes Neuronales Multicapa

En la figura se representa una Red Neuronal Multicapa, de tres capas (entrada, oculta y salida). En la entrada de cada unidad *j* de la capa oculta se tiene la suma de las entradas:

$$net_j = \sum_1^d x_i \omega_{ji} + \omega_{j0} = \sum_0^d x_i \omega_{ji} = \mathbf{w_j^t x}$$

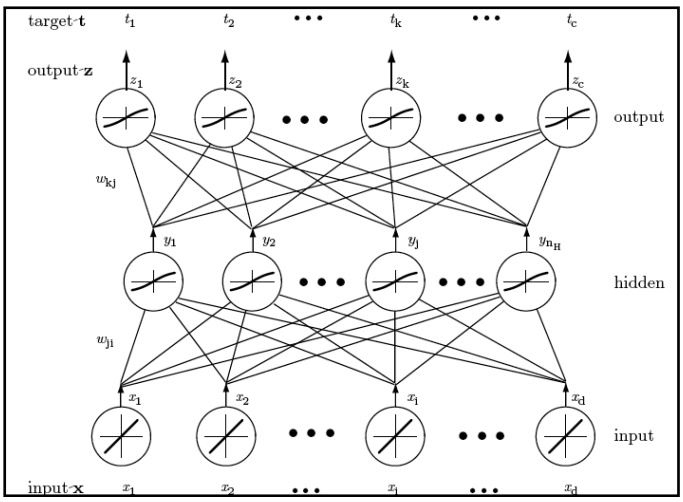
Como las funciones $f(.)$ de transferencia de la capa oculta son no lineales, la salida será no lineal:

$$y_j = f(net_j)$$

Para la capa de salida se tiene un razonamiento análogo:

$$net_k = \sum_1^{n_H} y_j \omega_{kj} + \omega_{k0} = \sum_0^{n_H} y_j \omega_{k0} = \mathbf{w_k^t y}$$

$$z_k = f(net_k)$$



De lo anterior surge que la función discriminante se pude expresar de la siguiente forma

$$g_k(\mathbf{x}) = z_k = f\left(\sum_1^{n_H} \omega_{kj} f\left(\sum_1^d \omega_{ji} x_i + \omega_{j0}\right) + \omega_{k0}\right)$$

En principio y de acuerdo con Kolmogorov, toda decisión pude ser implementada por una red de tres capas con siempre esta cuente con el número correcto de unidades de la capa oculta. Él establece que con un número de unidades de la capa oculta igual a *2n+1* y eligiendo adecuadamente Ξ_j , ψ_{ij}, toda función *g(x)* en el dominio I^n con I^n; I = [0; 1], se puedeaproximar de la forma:

$$g(\mathbf{x}) = \sum_{j=1}^{2n+1} \Xi_j \left(\sum_1^d \psi_{ji}(x_i)\right)$$

Si bien Kolmogorov, no asegura "suavidad" (y por lo tanto es poco aplicable en los casos prácticos) esta expresión se puede aproximar a partir de una red de tres capas.

Entrenamiento

El método de entrenamiento conocido como "Backpropagation" es el más generalizado y simple método de aprendizaje supervisado y es una extensión del LMS utilizado para sistemas lineales.

$$J(\mathbf{w}) = 1/2 \sum_{k=1}^{c} (t_k - z_k)^2 = 1/2 (\mathbf{t} - \mathbf{z})^2$$

Donde $\mathbf{t}$ es el vector objetivo y $\mathbf{z}$ es el vector de salida de la red ,*c* es la dimensión de dichos vectores y $\mathbf{w}$ representa los pesos.

La regla de aprendizaje se basa en gradiente descendente, los pesos se inicializan con valores aleatorios y van cambiando de manera de reducir el error

$$\Delta \mathbf{w} = -\eta \frac{\partial \mathbf{J}}{\partial \mathbf{w}}$$

Si se aplica esta expresión a la red de tres capas de la figura 1, a partir de las ecuaciones anteriormente vistas se tiene que

$$\Delta w_{k,j} = \eta (t_k - z_k) f'(net_k) \frac{\partial net_k}{\partial w_{kj}}$$

$$y_j = \frac{\partial net_k}{\partial w_{kj}}$$

$$\Delta w_{j,i} = -\eta x_i f'(net_j) \sum_{k=1}^{n} w_{k,j} (t_k - z_k) f'(net_k)$$

Recomendaciones de Diseño

En (Ajoy K. Palit 2005)se establecen un conjunto de recomendaciones, métodos y herramientas prácticas para atacar este tipo de problemas

Preparación de los Datos

En cuanto a la normalización de los datos, en se proponen dos criterios

Normalización simple

$$x_{ni} = \frac{x_i}{x_{\max}}$$

Normalización Lineal, este criterio es el más utilizado

$$x_{ni} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

A partir de la totalidad de los datos se extraen tres conjuntos destinados a los procesos de Entrenamiento, Validación y Testeo

En cuanto a la definición del conjunto de Entrenamiento en [(2)] se establece una regla empírica:

$$N = \frac{W}{\varepsilon}$$

Dónde:

N : Número total de patrones utilizados para el entrenamiento

ε :Erro aceptado

W :Es el número de pesos en la red

Esta regla al vincular el número total de pesos red, con el error y numero de patrones de patrones de entrenamiento, la misma impone una restricción hacia la topología que finalmente tendrá la red y por lo tanto se puede interpretar como una primera recomendación Arquitectura de la Red

Arquitectura de la Red

La tarea de definir la Arquitectura de la Red con el objetivo de atacar un problema de predicción es una tarea compleja dado que existen varias posibles soluciones, por lo tanto requiere de mucha habilidad y experiencia práctica

La tarea implica:

- Determinación de los nodos de entrada

En general, la regla viene dada por la cantidad de variables independientes presentes en el conjunto de datos, para el caso de problemas de predicción está determinada por el número de lagged que se utilizaran para realizar el pronóstico y este número tiene que ser el menor posible que asegure una correcta predicción. La estructura de autocorrelación de los datos puede servir de herramienta para determinar el número de entradas .

- Determinación de los nodos de salida

El número de salidas depende del horizonte de predicción, en la practica la predicción de un único paso es lo más frecuentes; para el caso de que sea necesario predecir escalones múltiples, se presentan dos opciones : tantas salidas como escalones sea necesario predecir o una única salida pero con un proceso iterativo

- Selección del número de capas ocultas

Como ya fue expuesto, el teorema de Superposición de Kolmogorov indica que una única capa oculta seria alcanzaría en todos los casos, pero como también fue ya expuesto dicho teorema presenta limitaciones en la práctica que habilitan la posibilidad de que para algunos casos particulares sean necesarias más de una capa oculta, en estos casos se corre el riesgo de caer durante el entrenamiento en mínimos locales.

- Selección del número de unidades en las capas ocultas

Si bien no hay una forma directa de determinar el número de neuronas óptimo de la capa oculta, existen sin embargo algunas reglas practicas como por ejemplo:

1. El 75% del número de entradas a la red
2. $N_h = \alpha\sqrt{N_i \times N_0}$

   N_i : Es el número de neuronas de capa de entrada

   N_0 : Es el número de neuronas de capa de salida

   α : Es una constante 0.5<<2
3. Baum and Haussler (1989)

$$N_h = \frac{N_{tr} \times E_{tol}}{N_{dp} + N_0}$$

N_{tr} : Es el número de ejemplos de entrenamiento

N_{dp} : Es el número de datos por ejemplo de entrenamiento

N_0 : Es el número de neuronas de capa de salida

E_{tol} : Es el error tolerado

4. Akike

$$AIC = Nk\ln(\sigma^2) + 2K \text{ con}$$

$$\sigma^2 = 1/N\sum_{t=1}^{N}(z_t - z*_t)^2$$

N : Es el número de ejemplos de entrenamiento

K : Es el número de paramentos en el modelo

k : Es el número de neuronas de capa de salida

z^*_t : Es la salida de la red

z_t : Es verdadero valor

- Determinación del patrón de interconexión entre nodos

Una vez determinado el número de neuronas de cada una de las capas es necesario determinar el patrón de interconexión entre neuronas del cual surge el número de pesos que presenta la red. Como ya se vio anteriormente, este número está relacionado con el numero de patrones destinado al entrenamiento y error admitido

Para el inicio en se recomienda la interconexión completa en donde todas las neuronas de una capa están interconectadas con todas las neuronas de la siguiente capa, para este caso el numero de pesos que presenta la red es:

$$W = N_h \times (N_i + N)$$

- Selección de la función de activación de las neuronas

En problemas de predicción las utilizadas son

- Sigmoide

$$y = \frac{1}{1+e^{-x}}$$

- Tangente hiperbólica

$$y = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

- Escalón o Rampa (capa de salida)

Entrenamiento de la Red

La inicialización consiste en asignar l valor inicial que se tendrán a los pesos para comenzar el proceso de entrenamiento.

Se establece que una decisión difícil dado que la velocidad de entrenamiento está fuertemente influenciada por este valor. Con el objetivo de prevenir la posible saturación de las neuronas, se sugiere elegir un una distribución aleatoria de valores pequeños .

Para que las aprendan a la misma velocidad se propone la inicialización con : $-\hat{w} < w_{kj} < \hat{w}$ y para evitar la saturación y se establece que :

- Para los nodos $\hat{w} = \frac{1}{\sqrt{d}}$ con d la cantidad de entradas

- Para los nodos de la capa de salida $\hat{w} < \frac{1}{\sqrt{n_H}}$

*Método de Entrenamiento: **"Backpropagation"***

Para la aplicación las decisiones a tomar son acerca de

- "paso" o tasa de aprendizaje :Magnitud con que se actualizan los pesos

- Momento : Requerido para escapar de mínimos locales

Una elección apropiada del "paso" es importante debido a que método de gradiente descendente sufre problemas de convergencia lenta y baja robustez; acelerar la convergencia tomando "pasos" altos trae consigo el peligro de tener un comportamiento oscilatorio de la red alrededor del mínimo [(2)] .

Una alternativa posible para mitigar este problema y poder tomar un "paso" alto se recomienda introducir el "*momento*" en la regla de aprendizaje [(2)]

$$\mathbf{w}(m+1) = \underbrace{\mathbf{w}(m) + \Delta\mathbf{w}(m)}_{\text{gradient descent}} + \underbrace{\alpha\Delta\mathbf{w}(m-1)}_{\text{momentum}}$$

Donde α generalmente e se toma un valor constante perteneciente al intervalo (0.5; 0.9)

II. METODOLOGÍA

DATOS DISPONIBLES SEMANALES

La información disponible consiste de los datos históricos registrados para *n=102* años, a partir del año 1909, correspondientes a los promedios semanales (se toman 52 semanas por año) de aportes a los embalses de las centrales *Dr. Gabriel Terra*, *Constitución* y *Salto Grande*, constituyendo así una Serie Temporal Multivariada (STM) $X_{semanal} = \{x_1, x_2, .., x_{52*n}\}$ siendo cada una de las muestras $x_t = \{a_t^{Terra}, a_t^{Constituci\acute{o}n}, a_t^{Salto\ Grande}\}$ con a_t^i =aporte medio al embalse *i* durante la semana *t* con $t = 1..52 * n$.

Si bien el SIN se compone con 4 centrales hidrológicas, solo se dispone de la información de aportes relacionada con tres de ellas. Esto se debe (tal como se puede apreciar en la Figura 7) a que la Central "Rincón de Baygorria", para la cual no se dispone de la información histórica, posee un embalse y una cuenca relativamente pequeña con respecto a las restantes centrales por lo que se denomina "central de pasada" pues esta podrá estar generando casi exclusivamente si la central aguas arriba ("Dr. Gabriel Terra") se encuentra vertiendo y/o generando

En la Figura 8 se pueden apreciar las 3 series temporales, mientras que en la Tabla 2 se tienen algunos datos estadísticos de cada una de ellas. Por otra parte se puede ver como un análisis de correlación lineal indica que existe alta correlación entre la serie de Terra y las restantes.

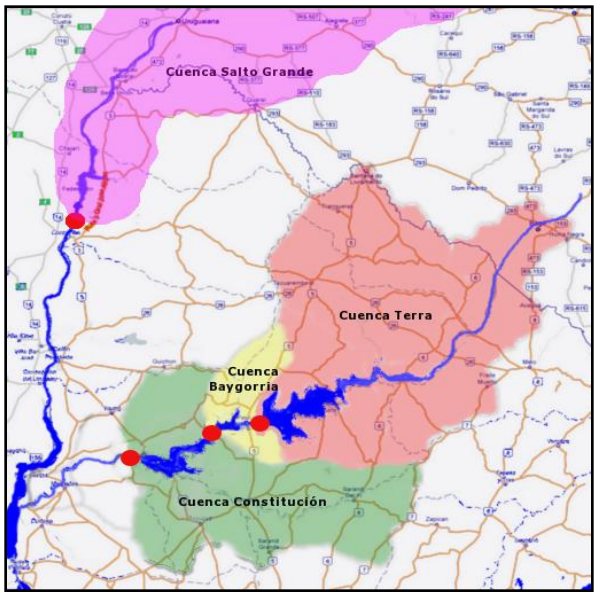


Figura 7 - Centrales hidráulicas, embalses y cuencas

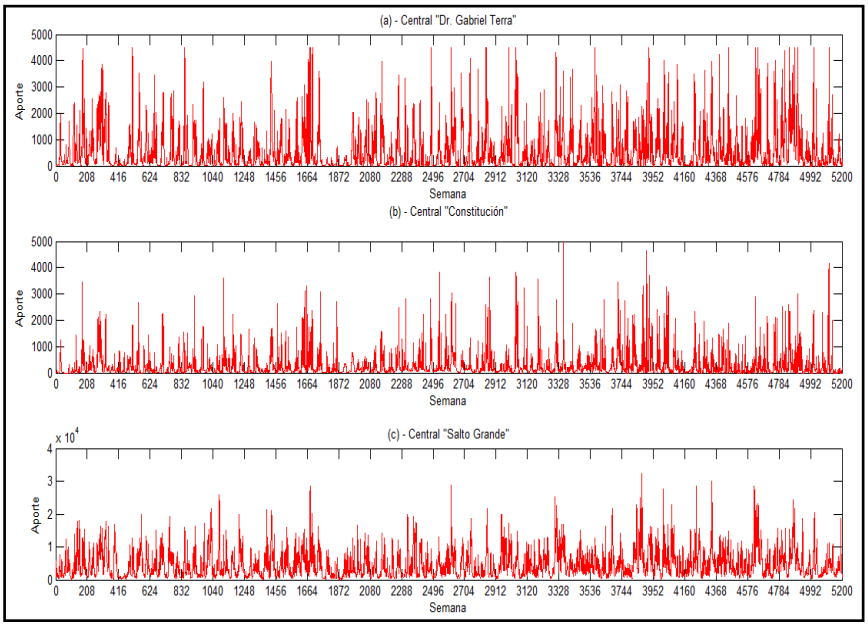


Figura 8 - Series temporales de aportes medios semanales por embalse

	Terra	*Constitución*	*Salto Grande*
Media	590,47	294,72	2344,94
Error típico	11,73	6,86	29,37
Mediana	238,00	115,00	1664,50
Desviación estándar	849,86	497,05	2128,43
Varianza de la muestra	722257,83	247056,95	4530203,04
Mínimo	0,00	0,00	71,00
Máximo	14335,00	5738,00	16197,00

Tabla 2 - Valores estadísticos de cada serie de aportes.

	Terra	*Constitución*	*Salto Grande*
Terra	1	0.5898	0.5299
Palmar	0.5898	1	0.2737
Salto	0.5299	0.2737	1

Tabla 3 - Coeficientes de correlación lineal

DATOS DISPONIBLES TRIMESTRALES

Se generaron los datos agregados trimestrales. Para ello se fueron agrupando de a trece semana consecutivas a partir de la primer semana de cada año y se realizó el promedio de aportes en cada una de las cuencas. De esta manera se construyó una serie temporal de *n=408* trimestres correspondientes a los aportes promedios trimestrales a los embalses.

Se tiene entontes una nueva STM $X_{trimestral} = \{x_1, x_2, .., x_{4*n}\}$ donde cada una de las muestras $x_t = \{a_t^{Terra}, a_t^{Constituci\acute{o}n}, a_t^{Salto\ Grande}\}$ con a_t^i =aporte medio al embalse *i* durante el trimestre *t*con $t = 1..4 * n$.

En la Figura 9 se tiene una representación de la nueva serie temporal, mientras que en Tabla 4 y Tabla 5 se tienen los estadísticos y las correlaciones respectivamente.

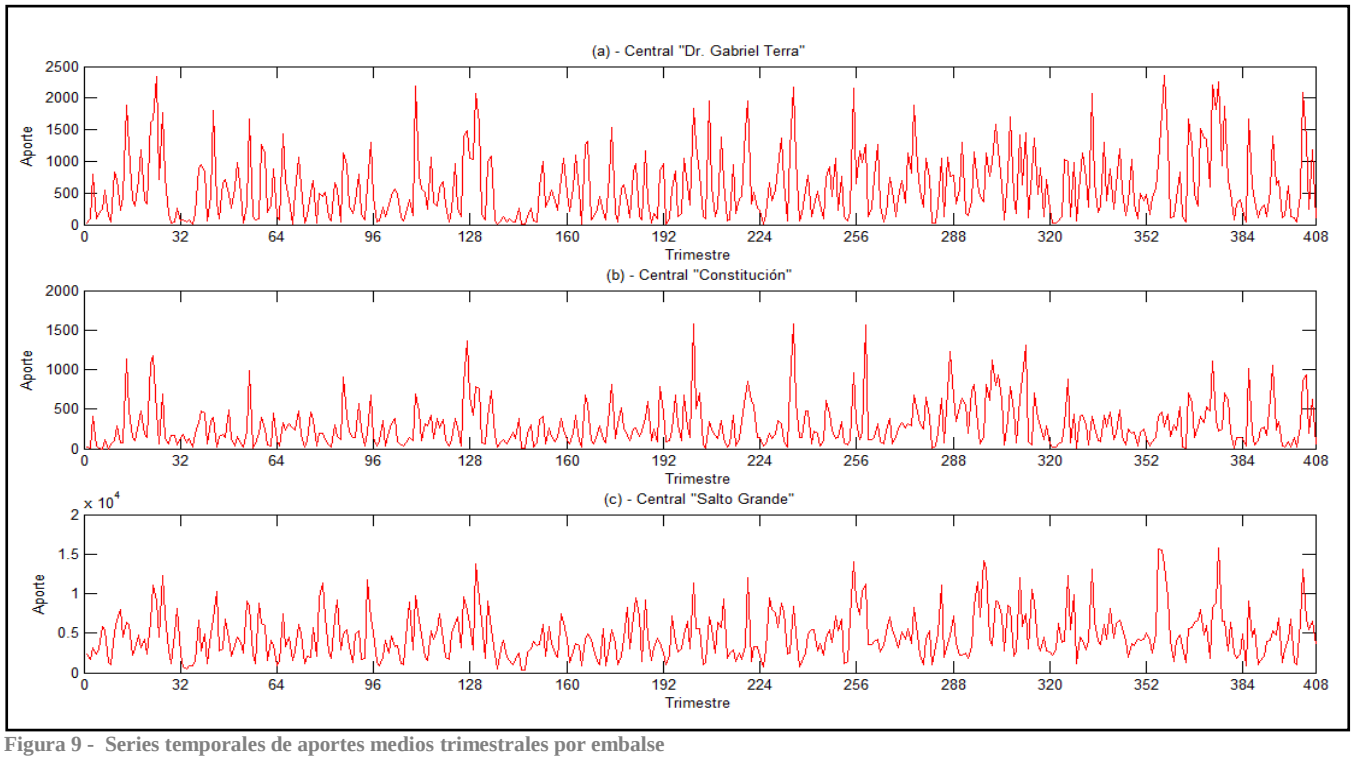


Figura 9 - Series temporales de aportes medios trimestrales por embalse

	Terra	*Constitución*	*Salto Gr nde*
Media	592,19	298,13	4733,11
Error típico	26,45	14,13	150,01
Mediana	442,80	209,68	4186,04
Desviación estándar	534,31	285,47	3030,09
Varianza de la muestra	285485,08	81491,32	9181429,62
Mínimo	3,98	0,00	382,03
Máximo	2364,37	1578,63	15812,77

Tabla 4 -Valores estadísticos de cada serie de aportes trimestrales

	Terra	*Constitución*	*Salto Grande*
Terra	1	0,7417	0,6725
Constitución	0,7417	1	0,4506
Salto Grande	0,6725	0,4506	1

Tabla 5 - Coeficientes de correlación lineal en aportes trimestrales

MODELO ADOPTADO

El proceso general involucra las etapas básicas referentes a la Recolección de Datos, la División de Datos (Entrenamiento, Test y Validación), el Pre procesamiento de los datos, la selección de las entradas, la identificación de la RNA y la Evaluación del desempeño. El punto clave radica en la selección e identificación del modelo, para lo cual se requiere un proceso de entrenamiento (Figura 10) que permita aproximarnos minimizando las diferencias entre los valores objetivo Y y las predicciones realizadas Y'. Este proceso posee una gran cantidad de factores que afectan la exactitud del modelo incluyendo el filtrado de los datos, el modelado de las entradas, el escalado de los datos, y la configuración del modelo.

La exactitud del modelo estará altamente influenciada por la cantidad de datos disponibles. Un conjunto óptimo de datos debe tener un tamaño grande y ser representativo de los probables valores que se puedan presentar.

La práctica común consiste en dividir el conjunto de datos en entrenamiento/test y validación. Es necesario que estos sean representativos de la misma población pues una red neuronal no puede extrapolar más allá del rango de valores que se utilicen para entrenar (Minnsan Hall 1996). Generalmente, si el conjunto de validación tiene valores que exceden el rango representado por los de entrenamiento, entonces la predicción será muy pobre. Se nos presentan entonces diferentes modelos posibles para la resolución de problema planteado en función del tipo de predicción que realice la RNA, como manejar la multidimensionalidad de los datos y como predecir más de un paso de tiempo en el futuro. A modo de ejemplo, podríamos tener una RNA que realice predicciones de hidraulicidad (MUY SECO, SECO, etc) directamente o puede predecir aportes, lo cual requiere ´de una clasificación posterior en un nivel de hidraulicidad. Asimismo, la RNA puede realizar predicciones en cada dimensión por separado o de todas las dimensiones en conjunto, la predicción del futuro pueden ser por múltiples salidas en la RNA, iterando en las predicciones de una RNA que predice un solo paso en el futuro, implementando una RNA que se retroalimente directamente con la predicción, etc.

En el presente trabajo nos enfocaremos en una RNA que predice 4 pasos de tiempo (4 salidas de la red), para una serie univariada (cada dimensión se trabaja por separado) y en la cual se predicen valores de aportes (lo que implica la necesidad de clasificar las predicciones). En la Figura 11 se presenta un esquema básico del modelo adoptado en este trabajo el cuan se detalla en la Figura 12. Se buscó dejar un modelo que admita fácilmente cambiar los criterios (por ejemplo para determinar los lag's, o el tipo de clasificador, etc.) para en un futuro poder evaluar diferentes alternativas.

En lo que respecta estrictamente a la construcción de la RNA, en la Figura 13 se pude apreciar más detalladamente cómo se procede.

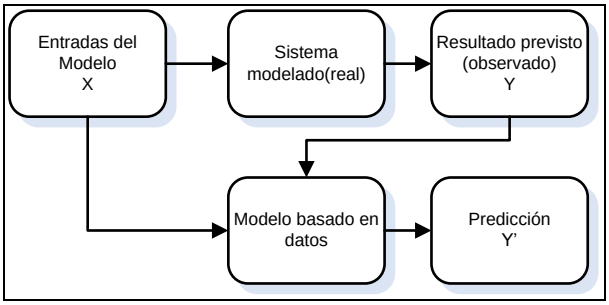


Figura 10 - Entrenamiento de modelos basados en datos adoptado por Mitchell, 1997

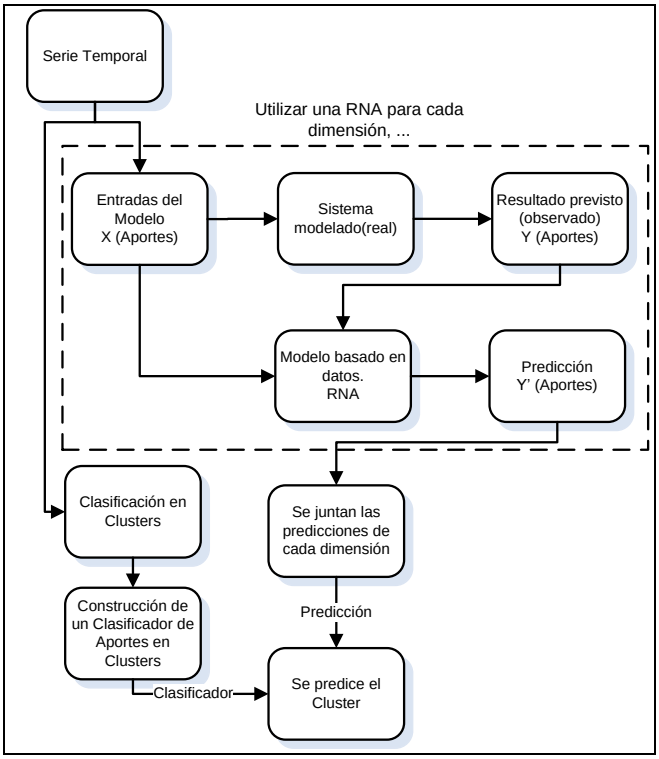


Figura 11 - Esquema básico del modelo adoptado

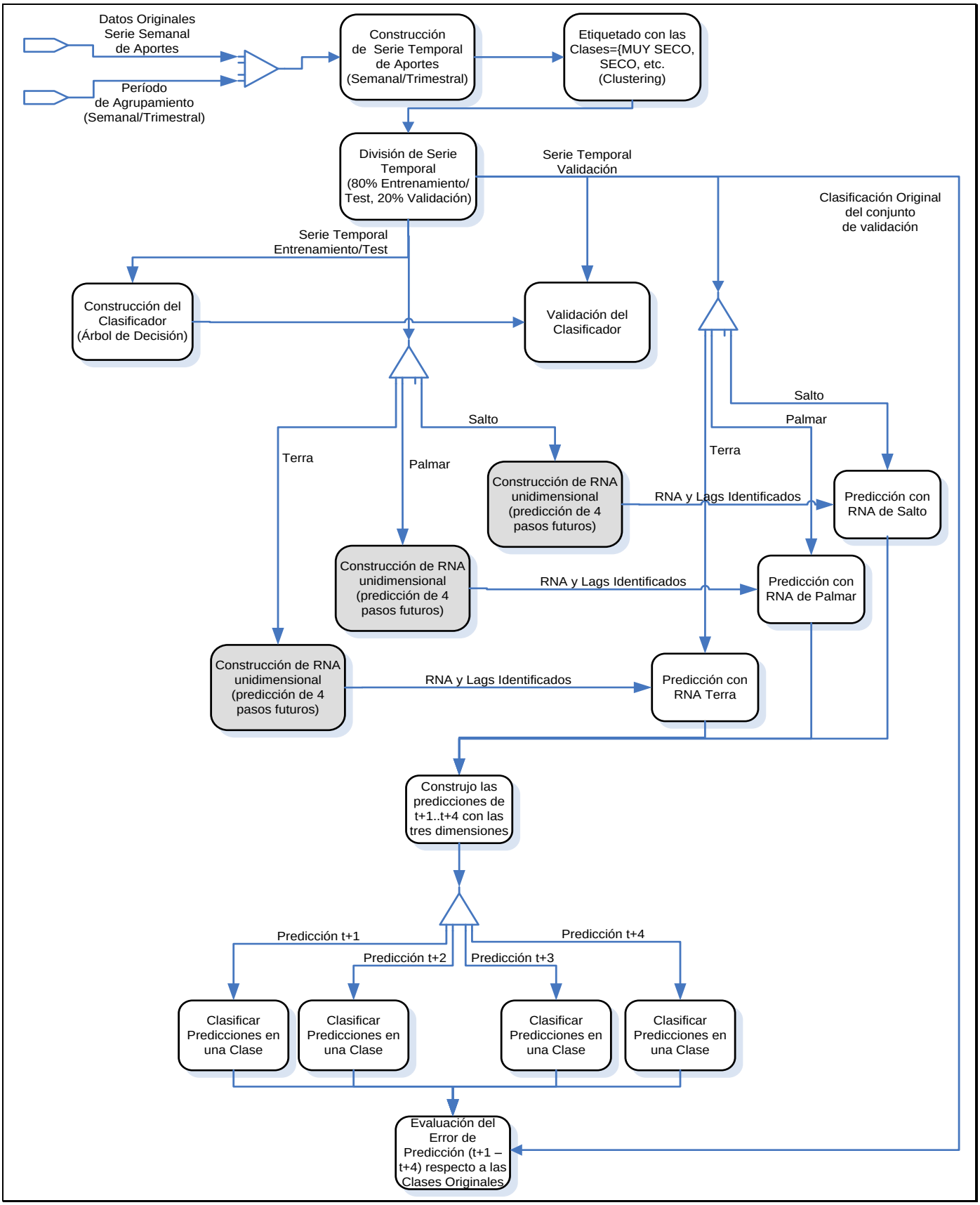


Figura 12 - Detalle del esquema del modelo adoptado

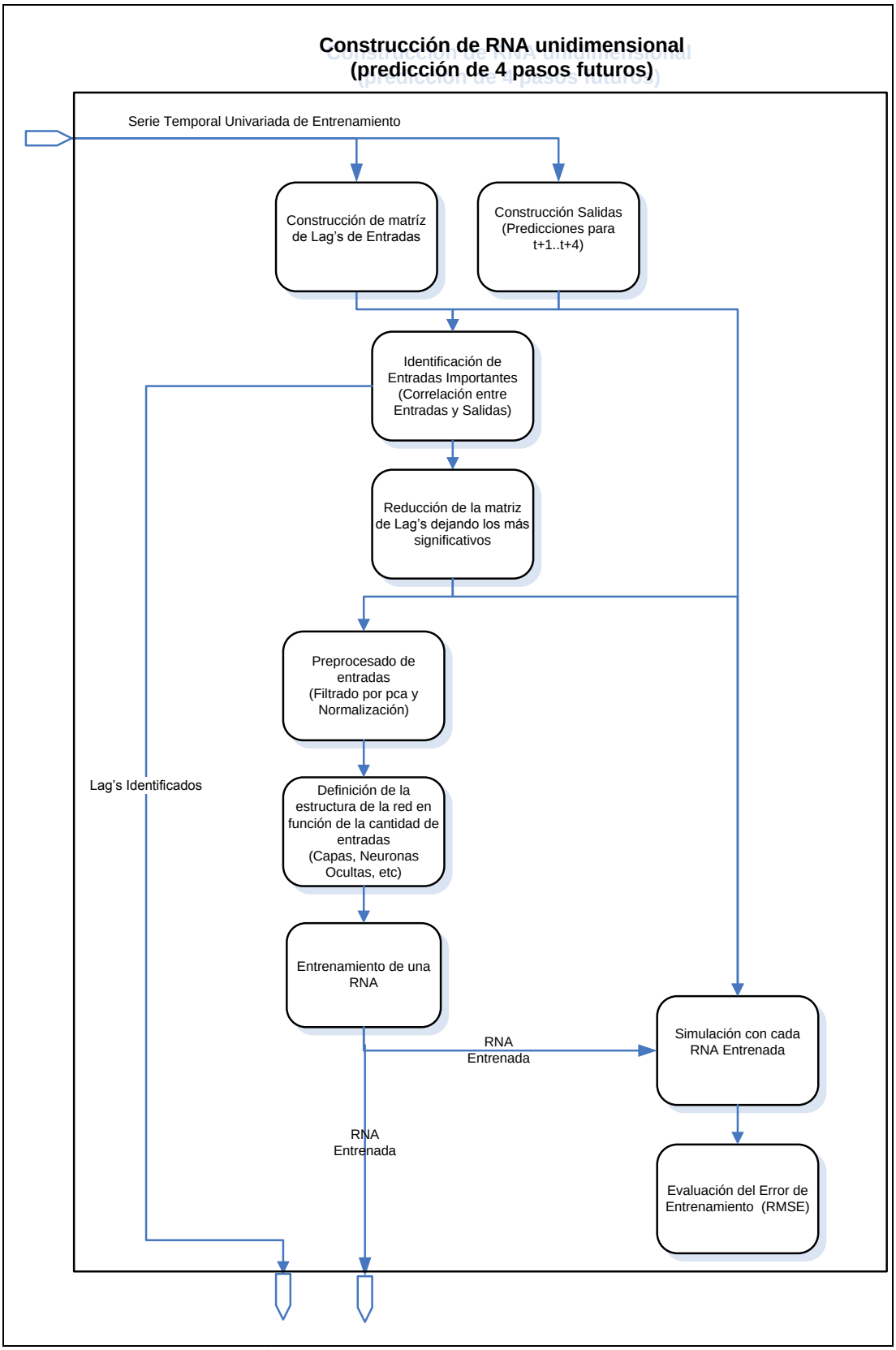


Figura 13 - Detalle de la construcción de la RNA

ETIQUETADO DE LAS MUESTRAS (CLUSTERING)

Dado que el objetivo del presente trabajo consiste es discriminar cuando estamos en presencia de un período MUY SECO, SECO, MEDIO, HUMEDO y MUY HUMEDO, debemos ser capaces de clasificar la información histórica en 5 grupos a efectos de posteriormente poder efectuar la predicción. Para ello utilizaremos una técnica de agrupamiento no supervisada (clustering) basado en el algoritmo de Fuzzy C-Means (FCM) y validaremos que la clasificación en 5 clusters sea adecuada, para lo cual aplicaremos la técnica buscando diferentes cantidades de clusters con diferentes parámetros.

El algoritmo utiliza distancia Euclidea, lo cual es adecuado para nuestro caso pues la información disponible corresponde a aportes (0 indica nula hidraulicidad, valores mayores indican hidraulicidad mayor).

Si bien el algoritmo de FCM, genera un agrupamiento para las muestras, dado que éste es sensible a la inicialización, entonces tendremos que en corridas independientes, no solo los centroides se pueden mover, sino que además el identificador del cluster puede cambiar para un mismo centroide. A esto se suma el hecho de las muestras pueden cambiar de cluster. En resumen, en corridas diferentes se puede clasificar una muestra en el grupo *i* y en el *j* pudiendo estos grupos ser el mismo o no. Esto implica la necesidad de determinar entonces si en realidad se trata del mismo cluster o no y por lo tanto deberemos normalizar el etiquetado de los clusters.

Para validar las particiones, se realizarán *n* iteraciones y se evaluarán las métricas mencionadas anteriormente. Luego de cada iteración, se ordenan los centroides en función de su distancia al origen y se actualizará el cluster asignado a cada muestra en función de cómo cambio el orden de los centroides. Luego de ordenar los centroides, si uno de ello paso de la posición i a la j, entonces renombramos el cluster i por j. El cluster 1 corresponderá así siempre a la mínima hidraulicidad y el último cluster agrupará las muestras que corresponden a la mayor hidraulicidad.

Si bien la cantidad de cluster a buscar está dada por el objetivo del problema (5), validaremos que la información histórica se pueda clasificar en esa cantidad de cluster o nó. Por lo tanto:

- Se validara la cantidad de clusters buscados tomando la totalidad de las muestras.
- Se buscará el parámetro m del algoritmo FCM tomando la totalidad de las muestras.
- Se estudiará la sensibilidad de las particiones a la inicialización.
- Se estudiará la sensibilidad de las particiones al reducir la candidata de muestras.

En términos generales se procedió de la siguiente forma con el total de las muestras:

- *Iterar para cada valor de m=[2 3 4 5 6 7 8]*
  - *Iterar para cada valor de c=[2 3 4 5 6 7]*
    - *Realizar 100 iteraciones ejecutando el algoritmo de FCM, normalizando el etiquetado de los clusters y evaluando las métricas.*
    - *Determinar que muestras cambiaron de cluster en las 100 iteraciones y que porcentaje de muestras sobre el total son las que cambian de cluster.*
    - *Calcular el valor esperado y la desviación estándar de cada métrica entre las 100 iteraciones.*

Luego de validar la cantidad de clusters a buscar y de determinado el mejor exponente del algoritmo para esa cantidad de clusters, procederemos a ejecutar nuevamente el algoritmo de FCM, con los parámetros definidos, fijando así el cluster de cada muestra y sus centroides.

Posteriormente particionaremos las muestras en 10 folder y le aplicaremos FCM comparando los centroides obtenidos con el sub-muestreo con respecto a los obtenidos utilizando la totalidad de las muestras. Para cada nuevo centroide calcularemos la distancia con respecto al original y el porcentaje respecto a su distancia al origen. Esto se repetirá tanto para las muestras semanales como para las trimestrales.

CLASIFICADOR DE APORTES EN CLUSTERS

A partir del etiquetado de todas las muestras, construiremos un clasificador que nos permita, ante la presencia de una nueva muestra de aportes, poder determinar a que categoría corresponde esta nueva muestra. Este será utilizado para clasificar las predicciones de la RNA en alguna de las categorías de MUY SECO a MUY HÚMEDO.

Para la construcción del clasificador, partiremos de la clasificación resultante de aplicar el algoritmo de FCM, es decir, del etiquetado dado por el cluster y de los centroides obtenidos. El cluster asignado a cada muestra corresponderá a la etiqueta de la misma. Una vez divididos los datos de Entrenamiento/Test por un lado y de Validación por otro, entrenamos y testeamos el clasificador y posteriormente lo validaremos.

Si bien se pueden entrenar diferentes tipos de clasificadores, en primera instancia simplemente utilizaremos un árbol de decisión y evaluamos su performance de entrenamiento con Cross Validation 10 fold y posteriormente lo validaremos aplicando el clasificador al conjunto de validación.

El algoritmo de construcción del árbol utilizará la medida de impureza de Gini como criterio para ramificar los nodos y efectuará un post-podado por mínimo costo-complejidad a efectos de evitar el overfitting.

ESTRUCTURA DE LA RNA

En principio y de acuerdo con Kolmogorov toda decisión pude ser implementada por una red neuronal de tres capas siempre que esta cuente con el número correcto de unidades de la capa oculta. Asimismo, en problemas de modelado de procesos hidrológicos, se utilizan comúnmente redes neuronales con una única capa oculta pues se considera que provee de la suficiente complejidad para simular la dinámica y las propiedades no lineales de los procesos hidrológicos (C. C. Wu 2010).

Definido que se tendrá una única capa oculta, el diseño de la arquitectura de la red consistirá en la especificación de las entradas (definiendo el tamaño de la capa de entrada), la definición de las salidas de la red (dada por la cantidad de predicciones futuras a realizar) y en la determinación de la cantidad de nodos de la capa oculta.

En este trabajo, se tendrán tres redes donde cada una de ellas intentará predecir el aporte en t+1, t+2, t+3 y t+4 en cada embalse por separado.

La cantidad de nodos en la capa de salida de cada una de estas redes será de 4, la cantidad de entradas podrá variar en función de los lag's que se determinen y la cantidad de nodos de la capa oculta se calculará como el 75% de las entradas.

IDENTIFICACIÓN DE LAS ENTRADAS IMPORTANTES PARA LA RNA

Dada la serie temporal de aportes $\{x_1, x_2, .., x_N\}$, esta puede ser reorganizada en una serie de vectores con retraso $X_t = \{x_t, x_{t-\tau}, x_{t-2\tau}, .., x_{t-(m-1)\tau}\}$, donde $X_t \in R^m$, τ es el tiempo de retardo expresado como un múltiplo de período de muestreo (semanas/trimestres en nuestro caso) y m es la dimensión embebida.

Supongamos que las salidas observadas en T-pasos de tiempo en el futuro son y_{t+T}, entonces la información histórica puede ser expresada como el par $\{X_t, y_{t+T}\}: t = 1, .., n$

La relación existente entre el vector de entradas X_t en el instante ty la predicción de la salida y_{t+T}^F en los momentos $t+T$ puede expresarse como:

$$y_{t+T}^F = f(\{X_t, y_{t+T}\}) + e_t$$

Siendo e_t es el ruido típico y $f(.)$ la función que realiza el mapeo (predicción) que en este trabajo es aproximada por un modelo basado en Redes Neuronales.De esta forma, la serie temporal se transforma en un conjunto de muestras independientes

Ahora bien, los valores que se definan de la dimensión m y τ, determinarán el conjunto de entradas con retraso que se consideren.Con un $\tau = 1$ se tendrá un vector de entradas de dimensión m, mientras que con valores de $\tau > 1$ se descartarán datos (toma 1 cada τ) bajando así la dimensión del vector. Estos valores provocarán que el modelo pierda o gane información que puede o no ser significativa. Si el modelo gana demasiada información se puede caer en un sobreajuste, mientras que si pierde información relevante se tendrá una mala representación del sistema (falta de información vs sobreajuste a los datos).

Es así, entonces, que la selección de las entradas del modelo sea de suma importancia. El objetivo es determinar cuáles son los retrasos (lag's) más significativos que debemos utilizar para la predicción de la salida del sistema. De acuerdo con (C. C. Wu 2010), en el estudio de problemas hidrológicos se han utilizado una variedad de métodos que se pueden agrupar en las siguientes categorías: (a) Conocimiento a priori del sistema, (b) Análisis de componentes principales (PCA), (c) Análisis de correlación lineal como ser correlación cruzada, autocorrelación o correlación parcial) , (d) Análisis de correlación Nolineal como ser Average Mutual Information, Partial mutual information, (e) Estudios Eurísticos (Prueba y Error), (f) Extracción de conocimiento (Análisis de Sensibilidad), (g) Métodos Compuestos, combinando por ejemplo PCA con técnicas analíticas de correlación, (h) Selección automática por ejemplo utilizando métodos de optimización global basados en programación genética.

Todos estos métodos buscan identificar las relaciones más significativas entre las entradas $\{x_t, x_{t-\tau}, x_{t-2\tau}, .., x_{t-(m-1)\tau}\}$y la salida y_{t+T}del modelo para valores de m y τ, de forma tal que permitan identificar que valores mantener en las entradas eliminado las restantes.

Así, para un m y τ dados se busca el subconjunto $L \subseteq \{1,2,..,(m-1)\}$ de los valores de τ que correspondan a los retrasos más significativos dentro de la ventana *m.* Se tendrá así un vector de entradas $X_t = \{x_t, x_{t-i\tau}\}$ de dimensión menor a *m* considerando solo los lag's más significativos.

En lo que respecta al trabajo, aplicaremos solo la técnica de Auto Correlación Lineal para determinar cuáles son las entradas (lag's) más significativas. Para ello se estudiará la autocorrelación lineal entre la salida y los lag's candidatos y se realizará un Test de Hipótesis de No Correlación. En la medida que una columna retorne en el test valores menores (inferiores a un cierto umbral) esta

será mas significativa y por lo tanto candidata a ser considerada.

El resultado de estos métodos determinarán las entradas a considerar en la Red Neuronal condicionando así su estructura.

PREPROSESAMIENTO DE LAS ENTRADAS A LA RNA

Se aplica PCA. Para elo se buscan las componentes principales que explican el 90% de la varianza, se reduce la dimensionalidad aplicando la matriz de transformación y posteriormente se vuelve a aplicar la transformación inversa. De esta forma se logra eliminar ruido de la señal original

III. RESULTADOS CON LA SERIE SEMANAL

RESULTADO DE LA CLUSTERIZACIÓN DE LAS MUESTRAS SEMANALES DE ENTRENAMIENTO

En la Tabla 1Tabla 6se presentan los porcentajes de muestras que cambian de cluster durante todas las 100 iteraciones y se puede apreciar que en la medida que la cantidad de clusters a buscar sea superior a 6, aumenta la movilidad de las muestras entre los cluster. En particular, para el caso de 5 clusters, si bien se presenta alguna movilidad de las muestras(al utilizar los valores 2 y 3 como parámetro del exponente de FCM) esta es muy baja por lo que la sensibilidad a la inicialización del algoritmo es muy baja.

Si acudimos a los indicadores escalares de cada iteración podremos ver que a lo largo de las 100 iteraciones estos tienen una varianza muy baja (Tabla 8) por lo cual podemos tomar el valor esperado entre todas las iteraciones como valor válido para cada escalar, siendo estos los valores que se muestran en la Tabla 7, donde se presentan en función de los parámetros considerados. Esta tabla, muestra los valores más convenientes con fondo blanco y lo va oscureciendo en la medida que el valor pasa a ser menos conveniente.

Por otra parte, en la Figura 14, se pueden apreciar los centroides que se obtienen si se realiza un sub-muestreo con una partición de 10 Fold. Esta gráfica incluye los centroides que se obtuvieron con la totalidad de las muestras.

.

	Exponente						
Clusters	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	0,04%	0,04%	0,00%	0,00%	0,00%	0,00%	0,00%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	0,51%	0,51%	0,51%	0,51%	0,00%	0,51%	0,51%

Tabla 6 - Porcentaje de muestras que cambian de cluster en las 100 iteraciones

		Exponente						
	Clusters	2	3	4	5	6	7	8
PartitionCoeficient (PE)	**2**	0,863	0,577	0,337	0,185	0,098	0,051	0,026
	3	0,797	0,415	0,175	0,066	0,024	0,008	0,003
	4	0,752	0,328	0,108	0,031	0,008	0,002	0,001
	5	0,718	0,266	0,073	0,017	0,004	0,001	0,000
	6	0,682	0,230	0,054	0,011	0,002	0,000	0,000
	7	0,653	0,200	0,040	0,007	0,001	0,000	0,000
ClassificationEntropy (CE)	**2**	0,231732	0,442656	0,549210	0,602794	0,632092	0,649482	0,660519
	3	0,369525	0,728621	0,897961	0,978137	1,019886	1,043702	1,058336
	4	0,471393	0,938228	1,153994	1,250910	1,299101	1,326216	1,342642
	5	0,553039	1,116801	1,358231	1,464454	1,517392	1,546556	1,563772
	6	0,638478	1,249737	1,523051	1,640579	1,696738	1,727241	1,745784
	7	0,712229	1,375839	1,670506	1,791216	1,848841	1,879985	1,898548
PartitionIndex (SC)	**2**	8,030E-05	6,763E-05	5,239E-05	3,943E-05	2,919E-05	2,138E-05	1,554E-05
	3	5,525E-05	4,430E-05	2,970E-05	1,897E-05	1,185E-05	7,249E-06	4,366E-06
	4	4,120E-05	3,199E-05	1,958E-05	1,115E-05	5,988E-06	3,154E-06	1,649E-06
	5	3,352E-05	2,630E-05	1,465E-05	7,410E-06	3,644E-06	1,760E-06	8,187E-07
	6	2,995E-05	2,021E-05	1,165E-05	5,581E-06	2,496E-06	1,100E-06	4,904E-07
	7	2,938E-05	1,843E-05	9,777E-06	4,289E-06	1,805E-06	7,219E-07	2,898E-07
SeparationIndex (S)	**2**	3,234E-10	7,324E-10	7,146E-10	4,402E-10	1,923E-10	1,140E-10	8,697E-11
	3	2,099E-09	2,467E-09	2,160E-09	3,402E-10	1,245E-10	9,545E-11	2,901E-10
	4	7,212E-08	6,058E-08	2,958E-08	1,821E-09	8,752E-10	6,173E-10	5,573E-10
	5	1,600E-05	1,797E-06	3,179E-08	1,295E-08	2,656E-08	6,568E-10	8,656E-10
	6	2,109E-06	8,939E-05	5,710E-07	3,114E-07	3,023E-08	4,630E-09	1,802E-09
	7	4,467E-05	1,078E-05	9,181E-06	5,045E-07	4,522E-09	7,534E-09	3,145E-09

Tabla 7 - Valores de los indicadores escalares sobre la bondad del clústering en función de los parámetros

Indicador	Máxima Varianza
PartitionCoeficient (PE)	7,99E-09
ClassificationEntropy (CE)	4,14E-08
PartitionIndex (SC)	1,31E-15
SeparationIndex (S)	7,99E-09

Tabla 8 - Máxima Varianza entre todas las iteraciones

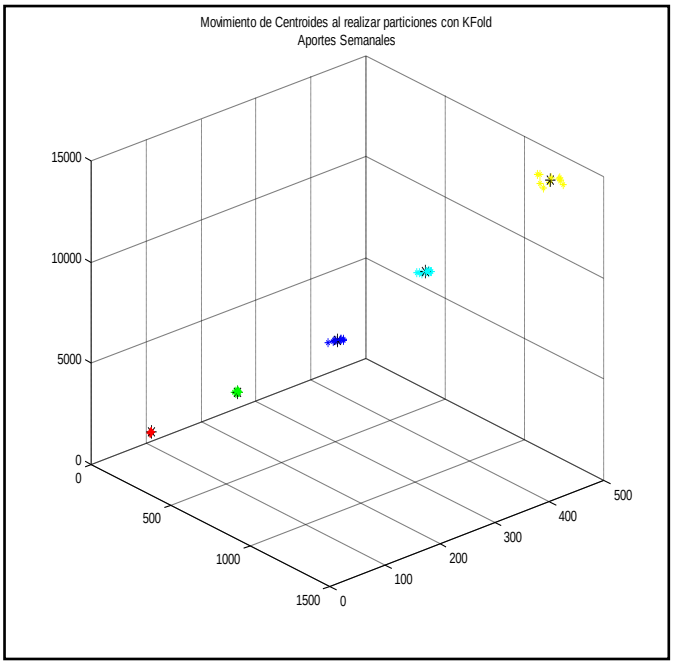


Figura 14 - Ubicación de los centroides al reducir la cantidad de muestras.

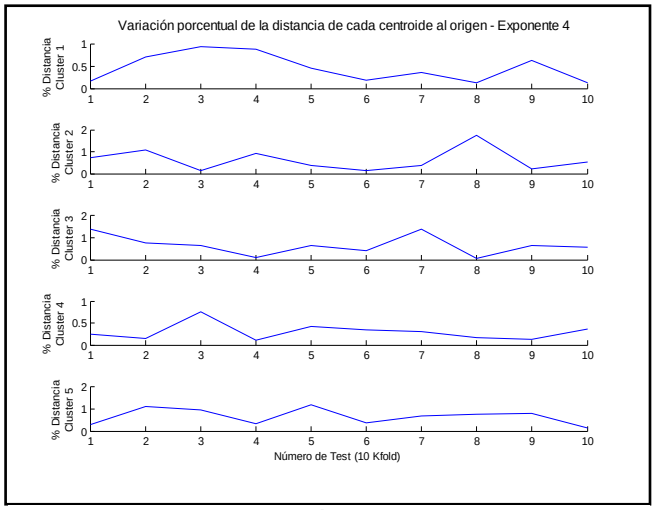


Figura 15 - Porcentaje de variación de la distancia respecto al origen para Exponente=4 al variar la cantidad de muestras

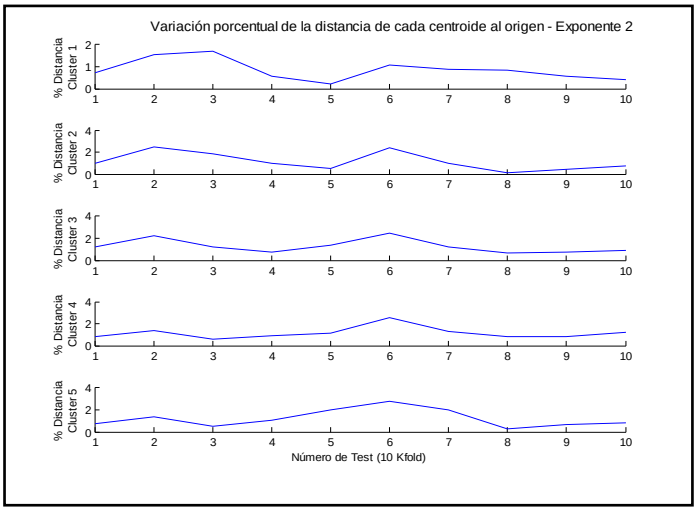


Figura 16 - Porcentaje de variación de la distancia respecto al origen para Exponente=2 al variar la cantidad de muestras

Conclusión: Podemos concluir que utilizar los valores de c=5 y exponente = 4 parece ser una decisión bastante adecuada pues:

- Si tomamos el porcentaje de muestras que cambian de cluster, este es muy bajo para c=5 (<0.04%)
- Posee valores de PE cercanos al máximo posible que se presenta para el caso de c=2 clusters.
- El CE presenta valores próximo al máximo posible que se da en el caso de c=7
- Por otra parte, si nos concentramos en c=5 y observamos los indicadores CE, SC y S, podemos ver que en la medida que aumentamos el exponente mejora la clasificación, lo cual se contradice que el indicador PE, existiendo una compromiso entre los valores de exponente 3 y 4
- Con exponente 4 parece tener mayor estabilidad que con exponente 2 al reducir la cantidad de muestras.

También podemos concluir que la clasificación de estas muestras es bastante estable y poco sensible a la inicialización.

Por otra parte, si reducimos la cantidad de muestras que se utilizan para calcular los centroides podemos observar que estos son bastante estables. Si bien se nota mayor inestabilidad en la medida que nos alejamos del origen, porcentualmente nunca sobrepasa el 2% de la distancia.

Por otra parte, para el problema es más crítico identificar la sequia que la extrema humedad.

Clasificador para Aportes Semanales:

Se construyó el árbol de decisión considerando las muestras de entrenamiento resultando en el que se muestra en la Figura 17.

Las dimensiones correspondientes a los aportes de Terra Palmar y Salto se muestran en la imagen como T, P y S respectivamente. Lo primero que se puede observar es que a la dimensión correspondiente a los aportes de Palmar no es necesaria para clasificar las muestras.

Entrenamiento y Test : La matriz de confusión resultante (Tabla 9) junto con el porcentaje de muestras mal clasificadas (0,47%) nos indica que el clasificador posee un performance muy buena con los datos de entrenamiento (aplicando crossvalidation).

Validación: Al validar el clasificador con el conjunto de validación se tiene una performance similar. La matriz de confusión resultante (Tabla 10Tabla 9) junto con el porcentaje de muestras mal clasificadas (0,94%) nos indica que el clasificador posee un performance muy buena con los datos de validación y por lo tanto no hay señales de overfitting en el entrenamiento.

	Muy Seco	Seco	Medio	Húmedo	Muy Húmedo
Muy Seco	801	2	0	3	0
Seco	4	1083	3	0	0
Medio	0	0	1510	0	0
Húmedo	4	0	0	544	4
Muy Humedo	0	0	0	0	300

Tabla 9- Matriz de Confusión para el clasificador (Conjunto de Entrenamiento de muestras semanales)

	Muy Seco	Seco	Medio	Húmedo	Muy Húmedo
Muy Seco	81	1	0	0	0
Seco	2	162	3	0	0
Medio	0	0	237	0	0
Húmedo	0	0	1	291	1
Muy Húmedo	0	0	0	0	284

Tabla 10 - Matriz de Confusión aplicando el clasificador al conjunto de validación.

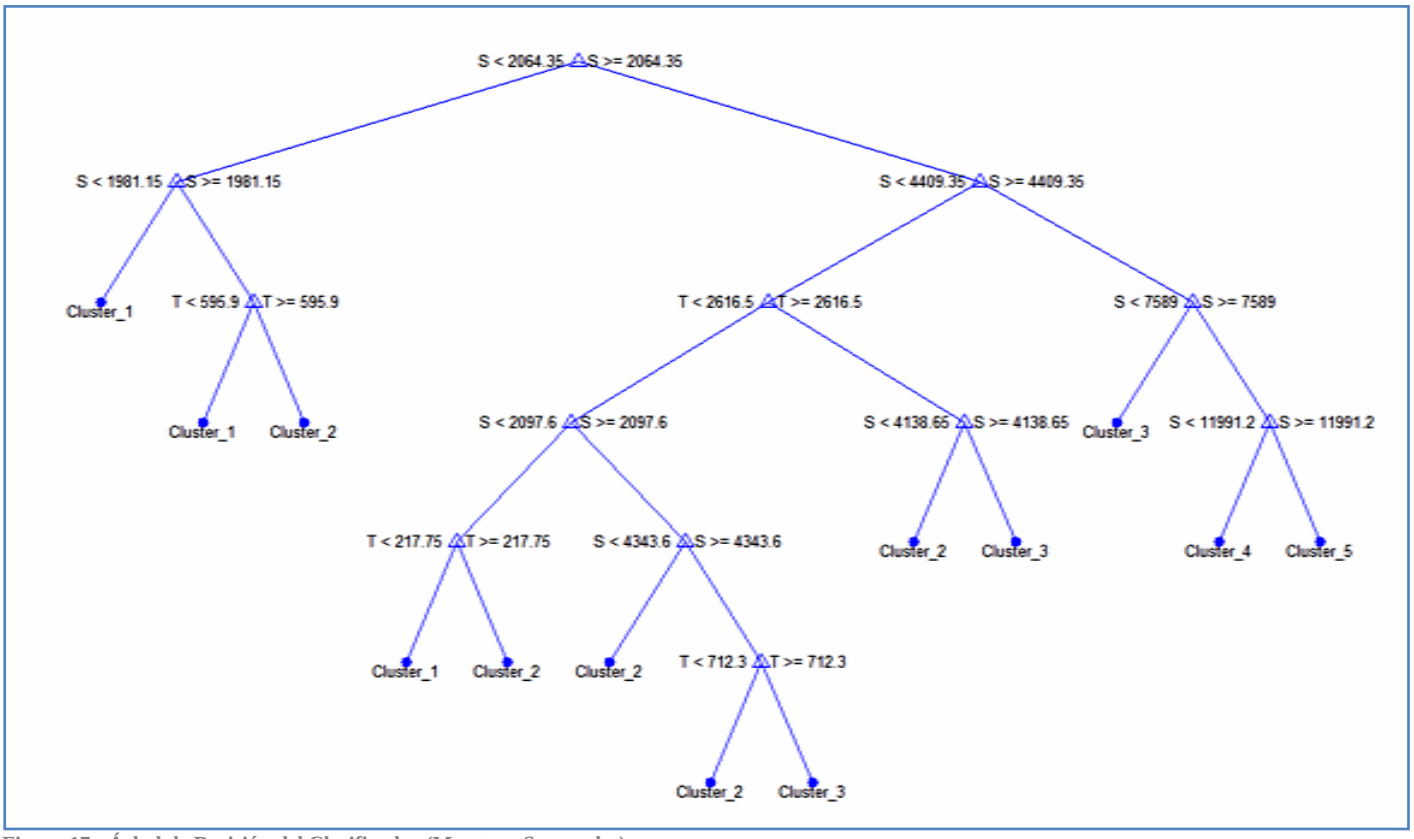


Figura 17 - Árbol de Decisión del Clasificador (Muestras Semanales)

IDENTIFICACIÓN DE LAS ENTRADAS IMPORTANTES PARA LA RNA

Se decidió manejar una historia máxima de 208 semanas (equivalente a 4 años) y al analizar la correlación entre las predicciones de t+1..t+4 se utilizó un umbral de 0.01 y tomar las 50 columnas con menor valor en de test de hipótesis de no correlación que sea inferior a dicho umbral.

Para cada uno de las dimensiones se encontraron los lags con mayor correlación con t+1, t+2, t+3 y t+4, uniéndolos (eliminando repetidos) para crear un conjunto de entradas común para los 4 pasos de predicción. Se obtuvo así los siguientes lag's por cada dimensión:

- **Terra**: 1 al 59 menos el 15 al 18 y 35 al 39
- **Palmar**:1 al 59 menos el 16 al 19 y 31 al 39
- **Salto**:1 al 59 menos el 22 al 23 y 31 al 38 66-63

Nota: Si observamos las correlaciones se puede apreciar cierta estacionalidad anual (recordar que son 52 semanas)

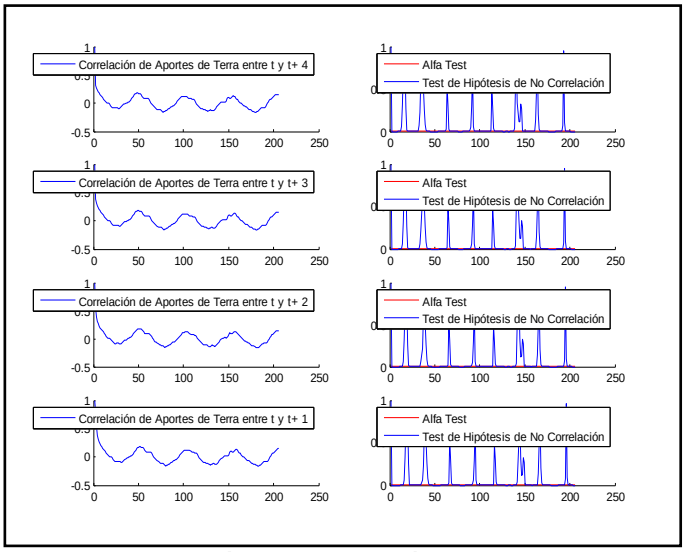


Figura 18 - Correlación y Test de Hipótesis para las muestras semanales de Terra

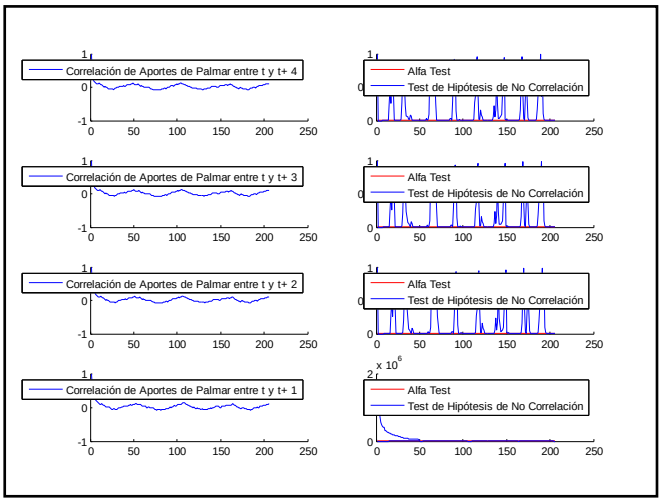


Figura 19 - Correlación y Test de Hipótesis para las muestras semanales de Palmar

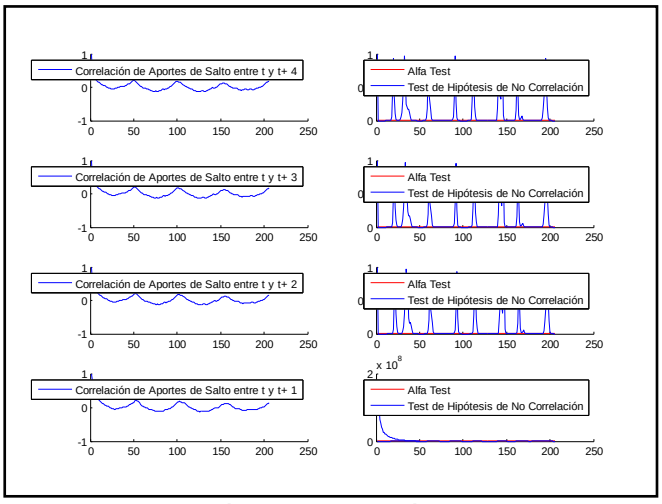


Figura 20 - Correlación y Test de Hipótesis para las muestras semanales de Salto

CONSTRUCCIÓN DE LAS RNA'S POR DIMENSIÓN

Se entrenó una RNA diferente para la predicción del aportes de Terra, Palmar y Salto. Cada RNA tiene 4 salidas (predicción de t+1..t+4). Las figuras muestran las performance de la red con los datos de entrenamiento y el error evaluado. Se puede apreciar que la red que presenta mayor error es la que predice los aportes de Terra. En el caso de Salto, la predicción de punto t+4 parece ser deficiente (de hecho presenta mayor error que las predicciones de t+1..t+2)

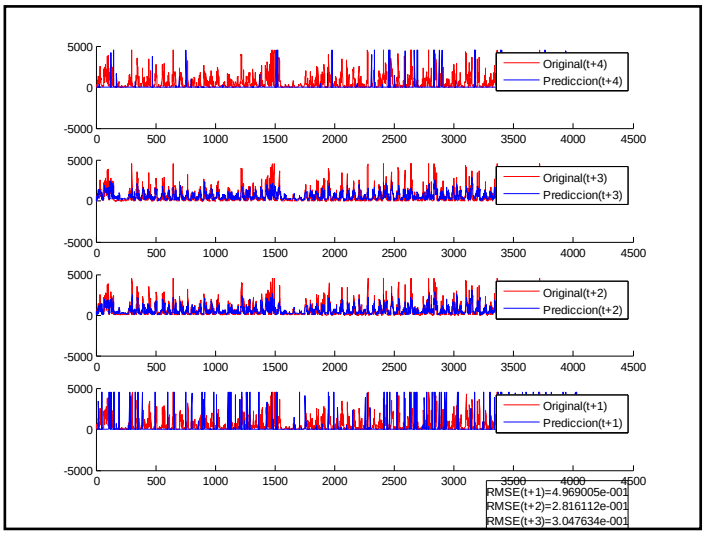


Figura 21 - Entrenamiento RNA - Terra (Semanal)

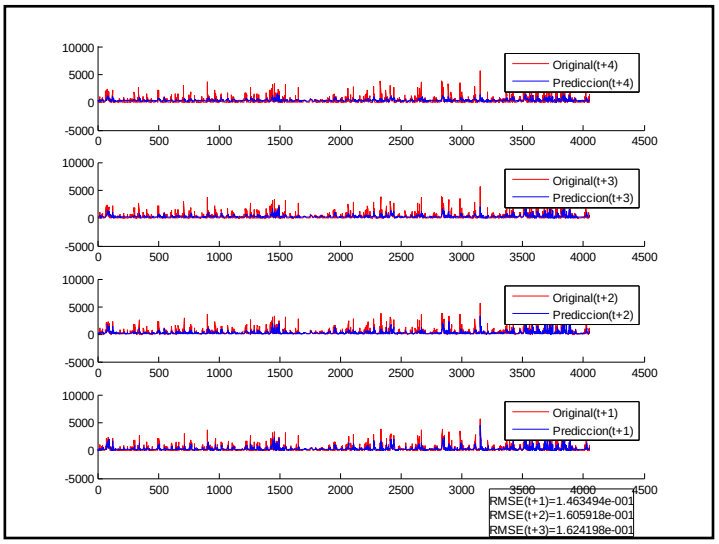


Figura 22 - Entrenamiento RNA - Palmar (Semanal)

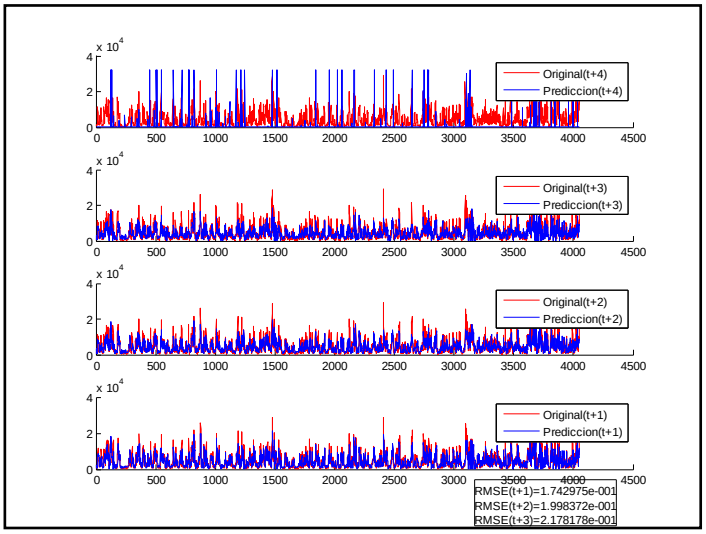


Figura 23 - Entrenamiento RNA Salto – Semanal

VALIDACIÓN DEL MODELOSEMANAL

Una vez entrenada la red neuronal y el clasificador definido, se tomó el conjunto de validación, se construyeron las entradas a cada RNA tal como lo indican los lag's identificados durante la etapa de entrenamiento y se procedió a hacer las predicciones de los aportes en Terra, Palmar y Salto. Posteriormente estas predicciones individuales se juntaron hasta formar un conjunto de muestras que etiquetada con el clasificador entrenado. Se comparó posteriormente esta clasificación con la originalmente dada a las muestras a partir de la aplicación del FCM.

Como resultado se presentan las matrices de confusión y los porcentajes de muestras mal clasificadas.

Se puede observar que no solo en porcentaje de muestras mal clasificadas es alto, sino que también hay categorías en los que no pudo clasificar ni una sola muestra bien. Por ejemplo, para t+4 solo acertó en las categorías de MUY HUMEDO.

	Matriz de confusión						*% Muestras mal clasificadas*
		Muy Seco	Seco	Medio	Húmedo	Muy Húmedo	
t+1	Muy Seco	16	1	0	0	64	73%
	Seco	20	0	0	0	137	
	Medio	18	0	0	1	203	
	Húmedo	7	0	0	0	268	
	Muy Húmedo	3	0	0	0	251	
		Muy Seco	Seco	Medio	Muy Húmedo	Muy Húmedo	
t+2	Muy Seco	23	16	52	46	20	70%
	Seco	20	8	22	20	10	
	Medio	23	12	58	88	42	
	Húmedo	30	9	70	114	52	
	Muy Húmedo	19	9	46	86	94	
		Muy Seco	Seco	Medio	Húmedo	Muy Húmedo	
t+3	Muy Seco	9	13	31	22	5	71%
	Seco	14	18	51	53	20	
	Medio	18	27	55	87	37	
	Húmedo	6	19	80	126	44	
	Muy Húmedo	6	20	50	95	83	
		Muy Seco	Seco	Medio	Húmedo	Muy Húmedo	
t+4	Muy Seco	0	3	0	0	153	75%
	Seco	0	1	0	0	78	
	Medio	0	4	0	0	220	
	Húmedo	0	7	0	0	269	
	Muy Húmedo	0	2	0	0	252	

I. RESULTADOS CON LA SERIE TRIMESTAL

RESULTADO DE LA CLUSTERIZACIÓN DE LAS MUESTRAS TRIMESTRALES DE ENTRENAMIENTO

Se presentan los datos análogos a los realizados para el caso de las muestras semanales. El resultado es básicamente similar.

Algunas pequeñas diferencias se pueden aprecias como ser el hecho de que con un valor de exponente = 4 parece ser ligeramente más inestable al reducir la cantidad de muestras.

Conclusión: utilizar los valores de c=5 y exponente = 2 parece ser una decisión bastante adecuada.

	Exponente						
Clusters	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	39,22%	0,00%	0,00%	0,00%	39,22%	39,22%	39,22%

Tabla 11 - Porcentaje de muestras que cambian de cluster en las 100 iteraciones (muestras Trimestrales)

		Exponente						
	Clusters	2	3	4	5	6	7	8
PartitionCoeficient (PE)	**2**	0,833	0,520	0,287	0,151	0,078	0,040	0,020
	3	0,769	0,366	0,144	0,052	0,018	0,006	0,002
	4	0,727	0,292	0,085	0,023	0,006	0,002	0,000
	5	0,698	0,221	0,057	0,012	0,003	0,001	0,000
	6	0,668	0,204	0,043	0,007	0,001	0,000	0,000
	7	0,632	0,174	0,033	0,007	0,001	0,001	0,001
ClassificationEntropy (CE)	**2**	0,279537	0,491204	0,583722	0,626445	0,648757	0,661690	0,669778
	3	0,420035	0,785287	0,935578	1,002634	1,036573	1,055666	1,067317
	4	0,519411	0,989212	1,197128	1,279217	1,318411	1,339813	1,352594
	5	0,596855	1,190755	1,399096	1,495877	1,537998	1,560153	1,573030
	6	0,674254	1,302596	1,556397	1,673449	1,718205	1,742487	1,755892
	7	0,764858	1,438817	1,703755	1,808132	1,865496	1,893204	1,905966
PartitionIndex (SC)	**2**	1,318E-04	1,146E-04	9,169E-05	7,053E-05	5,286E-05	3,899E-05	2,847E-05
	3	7,996E-05	6,859E-05	4,696E-05	2,964E-05	1,814E-05	1,094E-05	6,551E-06
	4	5,572E-05	4,251E-05	3,288E-05	1,917E-05	1,037E-05	5,476E-06	2,847E-06
	5	4,836E-05	3,622E-05	1,946E-05	1,331E-05	6,569E-06	3,092E-06	1,428E-06
	6	4,213E-05	2,810E-05	1,475E-05	9,909E-06	4,655E-06	2,090E-06	8,873E-07
	7	4,354E-05	2,752E-05	1,354E-05	5,859E-06	3,001E-06	1,412E-06	5,573E-07
SeparationIndex (S)	**2**	8,329E-01	5,201E-01	2,873E-01	1,513E-01	7,805E-02	3,977E-02	2,013E-02
	3	7,692E-01	3,662E-01	1,438E-01	5,186E-02	1,791E-02	6,044E-03	2,015E-03
	4	7,273E-01	2,921E-01	8,544E-02	2,294E-02	5,960E-03	1,533E-03	3,989E-04
	5	6,980E-01	2,213E-01	5,659E-02	1,210E-02	2,501E-03	6,008E-04	2,753E-04
	6	6,683E-01	2,042E-01	4,322E-02	7,155E-03	1,386E-03	2,039E-04	4,153E-05
	7	6,322E-01	1,741E-01	3,348E-02	6,984E-03	1,465E-03	5,613E-04	5,568E-04

Tabla 12 - Valores de los indicadores escalares sobre la bondad del clústering en función de los parámetros (muestras Trimestrales)

Indicador	Máxima Varianza
PartitionCoeficient (PE)	4,86E-06
ClassificationEntropy (CE)	2,24E-05
PartitionIndex (SC)	9,67E-13
SeparationIndex (S)	4,86E-06

Tabla 13 - Máxima Varianza entre todas las iteraciones (muestras Trimestrales)

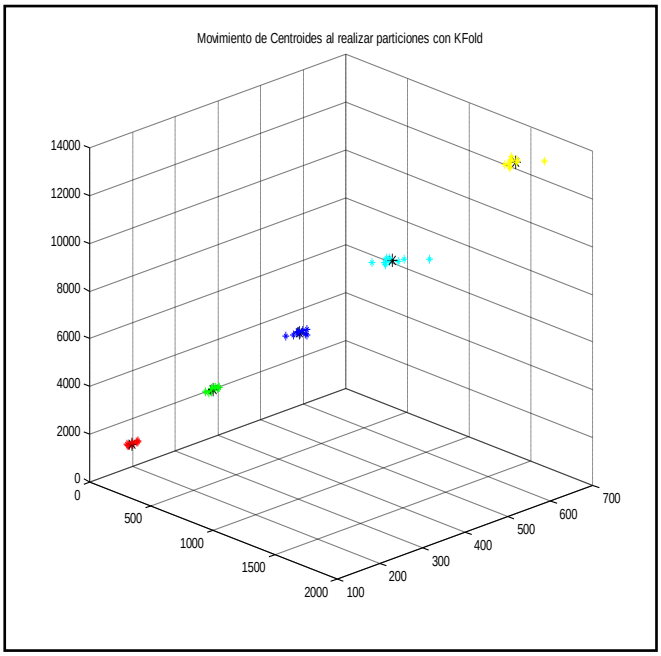


Figura 24 - Ubicación de los centroides al reducir la cantidad de muestras trimestrales

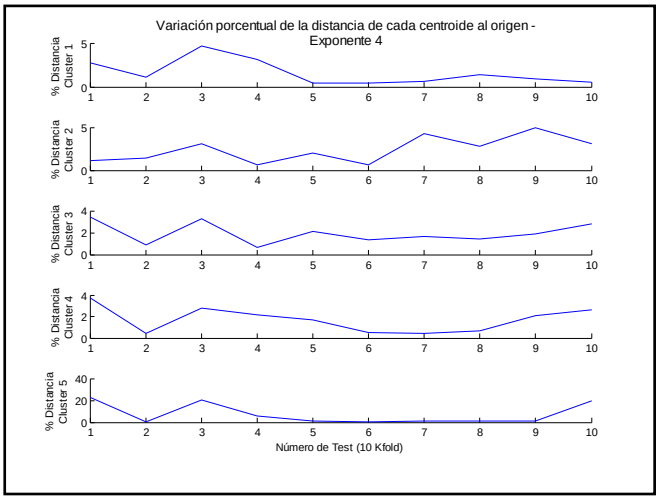


Figura 25 - Porcentaje de variación de la distancia respecto al origen para Exponente=4 al variar la cantidad de muestras trimestrales

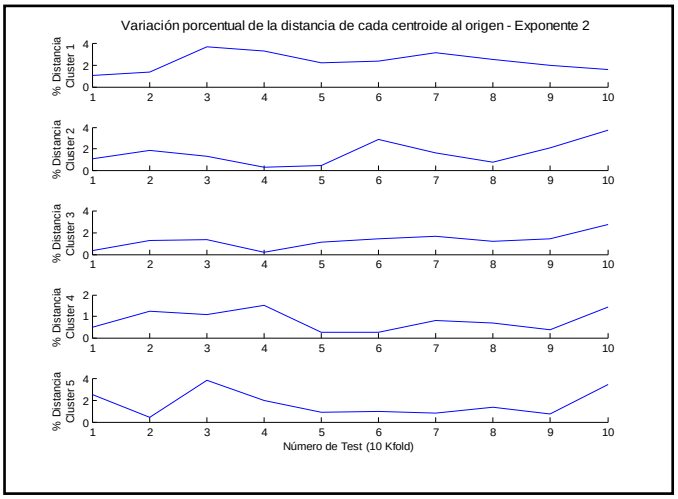


Figura 26 - Porcentaje de variación de la distancia respecto al origen para Exponente=2 al variar la cantidad de muestras trimestrales

Clasificador para Aportes Trimestrales:

Se construyó el árbol de decisión considerando las muestras de entrenamiento resultando en el que se muestra en la Figura 27

Las dimensiones correspondientes a los aportes de Terra, Palmar Salto se muestran en la imagen como T, P y S respectivamente. Lo primero que se puede observar es que en este caso, la dimensión correspondiente a los aportes de Terra y Palmar no es necesaria para clasificar las muestras.

Entrenamiento y Test : La matriz de confusión resultante (Tabla 14) junto con el porcentaje de muestras mal clasificadas (0,15%) nos indica que el clasificador posee un performance muy buena con los datos de entrenamiento (aplicando crossvalidation).

Validación: Al validar el clasificador con el conjunto de validación se tiene una performance similar. La matriz de confusión resultante (Tabla 15Tabla 10Tabla 9) junto con el porcentaje de muestras mal clasificadas (0,122%) nos indica que el clasificador posee un performance muy buena con los datos de validación y por lo tanto no hay señales de overfitting en el entrenamiento.

	Muy Seco	Seco	Medio	Muy Humedo	Humedo
Muy Seco	102	2	0	3	0
Seco	0	101	1	0	0
Medio	0	1	68	1	0
Muy Humedo	0	0	0	35	0
Humedo	0	0	0	0	15

Tabla 14 - Matriz de Confusión para el clasificador (Conjunto de Entrenamiento de muestras trimestrales

	Muy Seco	Seco	Medio	Muy Humedo	Humedo
Muy Seco	22	0	0	0	0
Seco	0	7	3	0	0
Medio	0	0	16	0	0
Muy Humedo	1	0	0	30	0
Humedo	0	0	00	0	6

Tabla 15 - Matriz de Confusión aplicando el clasificador al conjunto de validación.

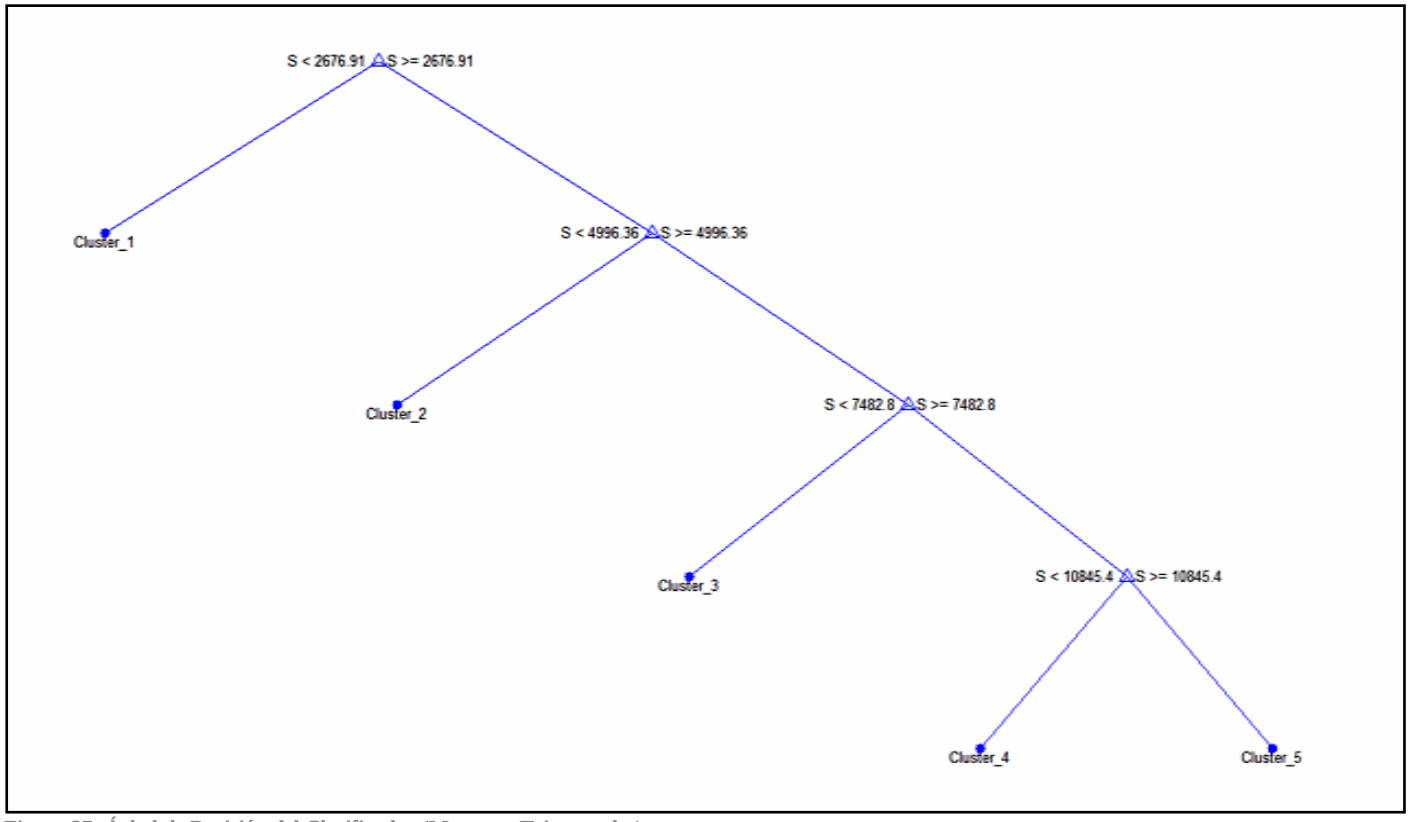


Figura 27 -Árbol de Decisión del Clasificador (Muestras Trimestrales)

IDENTIFICACIÓN DE LAS ENTRADAS IMPORTANTES PARA LA RNA

Se decidió manejar una historia máxima de 52 trimestres (equivalente a 13 años) y al analizar la correlación entre las predicciones de t+1..t+4 se utilizó un umbral de 0.05 y tomar las 30 columnas con menor valor en de test de hipótesis de no correlación que sea inferior a dicho umbral.

Para cada uno de las dimensiones se encontraron los lags con mayor correlación con t+1, t+2, t+3 y t+4, uniéndolos (eliminando repetidos) para crear un conjunto de entradas común para los 4 pasos de predicción. Se obtuvo así los siguientes lag's por cada dimensión:

- **Terra**: 1 al 30
- **Palmar**:1 al 30
- **Salto**:1 al 32 menos 11 y 17

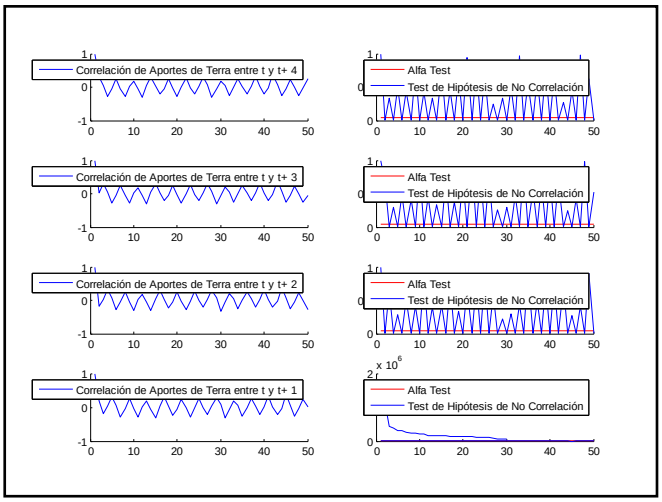


Figura 28 - Correlación y Test de Hipótesis para las muestras Trimestrales de Terra

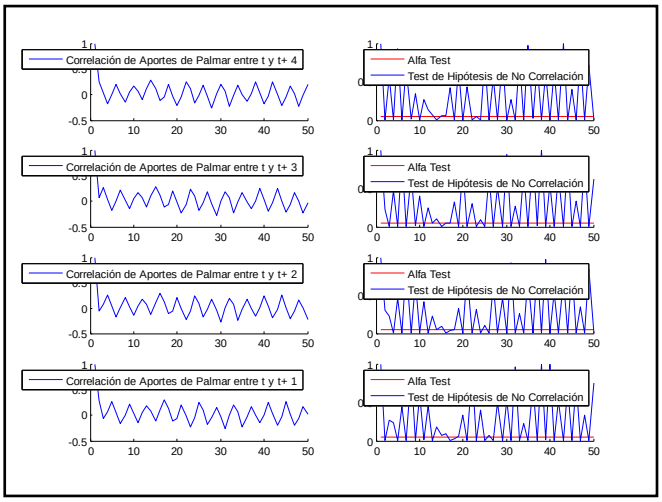


Figura 29 - Correlación y Test de Hipótesis para las muestras Trimestrales de Palmar

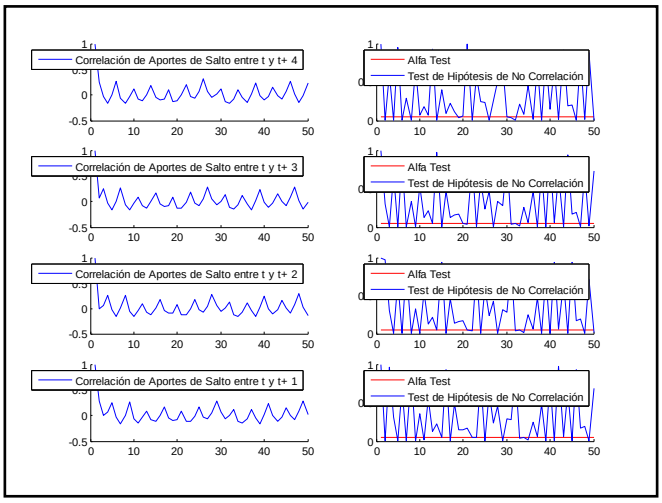


Figura 30 - Correlación y Test de Hipótesis para las muestras Trimestrales de Salto

Nota: Si observamos las correlaciones se puede apreciar la estacionalidad anual (recordar que un año son 4 muestras).

CONSTRUCCIÓN DE LAS RNA'S POR DIMENSIÓN

Análogamente al caso semanal, se entrenó una RNA diferente para la predicción del aportes de Terra, Palmar y Salto. Cada RNA tiene 4 salidas (predicción de t+1..t+4). Las figuras muestran las performance de la red con los datos de entrenamiento y el error evaluado. Se puede apreciar que la red que presenta mayor error es la que predice los aportes de Terra. En el caso de Salto, la predicción de t+1..t+4 parece ser mejor que las restantes.

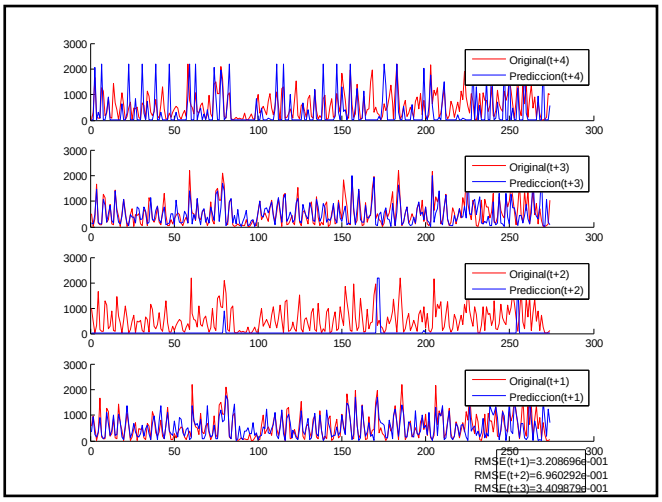


Figura 31 - Entrenamiento RNA - Terra (Trimestral)

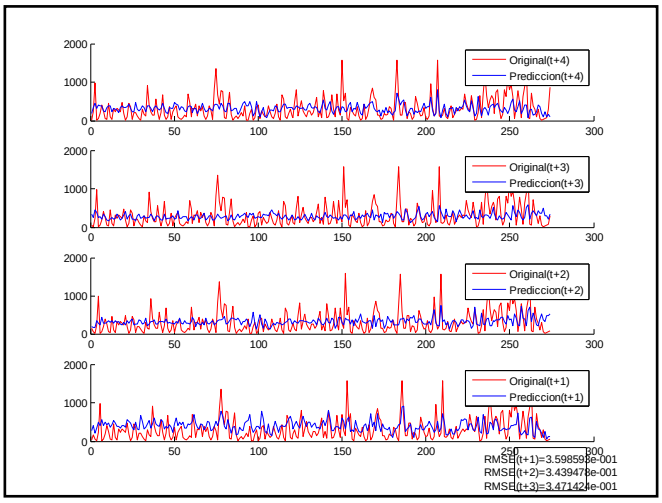


Figura 32 - Entrenamiento RNA - Palmar (Trimestral)

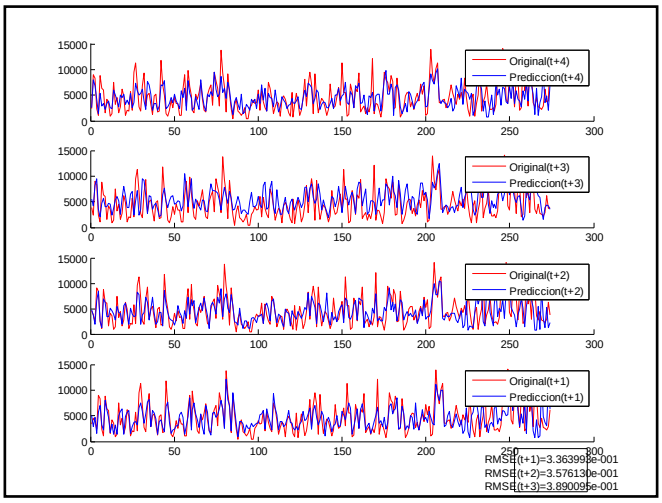


Figura 33 - Entrenamiento RNA - Salto (Trimestral)

VALIDACIÓN DEL MODELOTRIMERSTRAL

Una vez entrenada la red neuronal y el clasificador definido, se tomó el conjunto de validación, se construyeron las entradas a cada RNA tal como lo indican los lag's identificados durante la etapa de entrenamiento y se procedió a hacer las predicciones de los aportes en Terra, Palmar y Salto. Posteriormente estas predicciones individuales se juntaron hasta formar un conjunto de muestras que etiquetada con el clasificador entrenado. Se comparó posteriormente esta clasificación con la originalmente dada a las muestras a partir de la aplicación del FCM.

Como resultado se presentan las matrices de confusión y los porcentajes de muestras mal clasificadas.

Se puede observar que no solo en porcentaje de muestras mal clasificadas es alto, sino que también hay categorías en los que no pudo clasificar ni una sola muestra bien. Por ejemplo, para t+3 solo acertó en las categorías de MUY SECO.

	Matriz de confusión						*% Muestras mal clasificadas*
		Muy Seco	Seco	Medio	Húmedo	Muy Húmedo	
t+1	Muy Seco	0	10	0	1	0	69%
	Seco	1	11	0	0	0	
	Medio	0	10	2	1	1	
	Húmedo	0	4	0	0	0	
	Muy Húmedo	0	2	0	0	0	
		Muy Seco	Seco	Medio	Muy Húmedo	Muy Húmedo	
t+2	Muy Seco	1	3	3	2	1	74%
	Seco	2	7	0	3	0	
	Medio	3	6	2	6	1	
	Húmedo	0	2	1	1	0	
	Muy Húmedo	0	0	1	1	0	
		Muy Seco	Seco	Medio	Húmedo	Muy Húmedo	
t+3	Muy Seco	12	0	0	0	0	72%
	Seco	15	0	0	0	1	
	Medio	4	0	0	0	0	
	Húmedo	9	0	0	0	0	
	Muy Húmedo	2	0	0	0	0	
		Muy Seco	Seco	Medio	Húmedo	Muy Húmedo	
t+4	Muy Seco	4	4	5	3	0	62%
	Seco	2	1	0	1	0	
	Medio	4	2	3	1	0	
	Húmedo	2	0	2	7	0	
	Muy Húmedo	1	0	0	0	1	

II. CONCLUSIONES

Este trabajo se presenta como una primera aproximación para la predicción a largo plazo de la situación hidrológica, a continuación se destacan los principales puntos analizados:

- **Clasificación no supervisada de información histórica:** La utilización del FCM se presenta un clasificador estable (baja sensibilidad condiciones iniciales y a cantidad de muestras). Por otra parte la cantidad de cinco clusters (cantidad compatible con el problema) se muestra cercana a la cantidad optima de clusters de acuerdo a los distintos indicadores.

- **Clasificación de aportes en situación hidrológica:** El clasificador construido a partir de las etiquetas del FCM (árbol de decisión binario) no solo muestra una buena performance sino que también revela que los aportes de Palmar no son significativos para la clasificación (esa dimensión no es utilizada en ninguno de los nodos del árbol de decisión), esto concuerda con los expertos que indican que los aportes de Terra son suficientes para indicar el estado hidrológico del sistema.

- **Identificación de entradas relevantes para la RNA:** Utilizar el método de autocorrelación lineal seleccionando mediante un test de hipótesis de no correlación no se muestra muy adecuado. Los lag's determinados por este mecanismo tienden a ser consecutivos y no reflejar periodicidad .

- **El modelo global adoptado como predictor:**No dio buenos resultados por lo que es necesario continuar con los trabajos.

-

III. POSIBLES TRABAJOS FUTUROS

Este primer trabajose concentró, principalmente, en la clasificación de las crónicas hidrológicas, también se pretendió mostrar una primera entrega de la predicción de los niveles de aportes para el futuro año.

Esta predicción se realizó considerando un único método de modelado de entradas (auto correlación lineal), un único método de pre procesamiento utilizado fue PCA y se consideró un único modelo de predicción (redes neuronales).

Los trabajos futuros se pueden sintetizar en las siguientes líneas de acción que se pueden combinar entre sí:

- ***Enriquecer el modelo con mas datos de entrada***: Detección e incorporación de predicciones de otras variables climáticas provenientes de NOA.

- ***Cambiar el tipo de predicción***: Cambiar el modelo para que realice la predicción directa de la hidraulicidad en lugar de los aportes, evitando así la necesidad de un clasificador extra.

- ***Cambiar la forma de predecir múltiples pasos en el futuro:*** por ejemplo, utilizando una RNA que se retroalimente con su propia predicción, o implementando una red que realice la predicción de un paso de tiempo y luego itere para obtener los siguientes, etc.

- ***Pasar de múltiples redes predictoras univariadas a una red multivariada*:** por ejemplo, utilizando una RNA que se retroalimente con su propia predicción, o implementando una red que realice la predicción de un paso de tiempo y luego itere para obtener los siguientes, etc.

- ***Estudio de Diferentes Técnicas que mejoren la predicción***: En base al trabajo de C.L.WU, K.W. Chau, C Fan, tal cual lo muestra la Figura 34**Error! Reference source not found.**, se podrán tomar diferentes criterios para los distintos pasos del modelo:

    - Distintos métodos de determinación de entradas: AverageMutalInformation, False NearestNeighbors, Correlation Integral, Partial Mutual Information, Multi-objetive GeneticAlgorithm, Stepwise Linear Regression.

    - Distintos modelos de preprocesamiento: Media Movil, PCA y SSA.

    - Distintos modelos de predicción:Regresión Lineal, k-NN y Redes Neuronales Artificiales Modulare.s

IV. REFERENCIAS

"Curso Introduccion al Reconocimiento de Patrones": Apuntes de Clase. Facultad de Ingeniería, Universidad de la Republica, 2011.

ADME. http://www.adme.com.uy/sin/sistemaElectrico.php.

Administración Nacional del Mercado Eléctrico Uruguayo . *Informe Mensual del Mercado Eléctrico.* ADME. 2010. http://www.adme.com.uy/mmee/infmensual.php (último acceso: 25 de 11 de 2011).

Ajoy K. Palit, Dobrivoje Popovic. *Computational Intelligence in Time Series Forecasting - Theory and Engineering Applitcations.* Springer-Verlag London, 2005.

Balazs Balasko, Janos Abonyi and Balazs Feil. *Fuzzy Clustering and Data Analysis Toolbox For use With Matlab.*

Bezdek, J. C. . *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, 1981.

CMS Energy/Canadian Center for Energy Information. *Center For Energy.* http://www.centreforenergy.com/AboutEnergy/ONG/LiquifiedNaturalGas/Overview.asp?page=1 (último acceso: 25 de 11 de 2011).

Duda, y Hart. «Pattern Clasification.» Cap. 6. 2000.

Facultad de Ingeniería. *Sitio oficial del SimSee.* http://iie.fing.edu.uy/simsee/ (último acceso: 25 de 11 de 2011).

Mulvany, T.J. *On the use of self registering rain and flow gauges, (1850).* Proc. Institute Civ. Eng. 4(2) 1-8, 1850.

Wu, C.L., Chau, K.W., Fan,C. «Prediction of rainfall time series using modular artificial neural networks.» *Journal of Hydrology 389*, 2010: 146–167.

Wu, C.L., Chau, K.W., Li, Y.S.,. «Methods to improve neural network performance in daily flows prediction.» *Journal of Hydrology 372*, 2009: 80-93.