Redes Neuronales: Aplicación en la predicción de datos medioambientales.

WENCESLAO GONZÁLEZ MANTEIGA

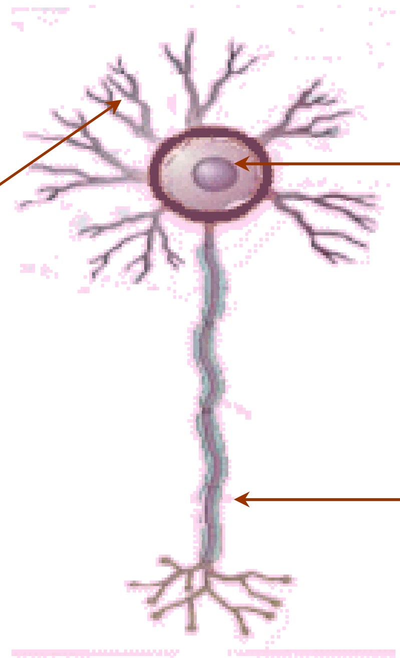
Departamento de Estadística e Investigación Operativa
Universidad de Santiago de Compostela

Las Redes Neuronales

EL ORIGEN

- A ***mediados de siglo*** surge un movimiento científico INTERDISCIPLINAR con el objetivo de crear ***inteligencia***
- **¿CÓMO?** Imitando al **cerebro humano**

LA NEURONA



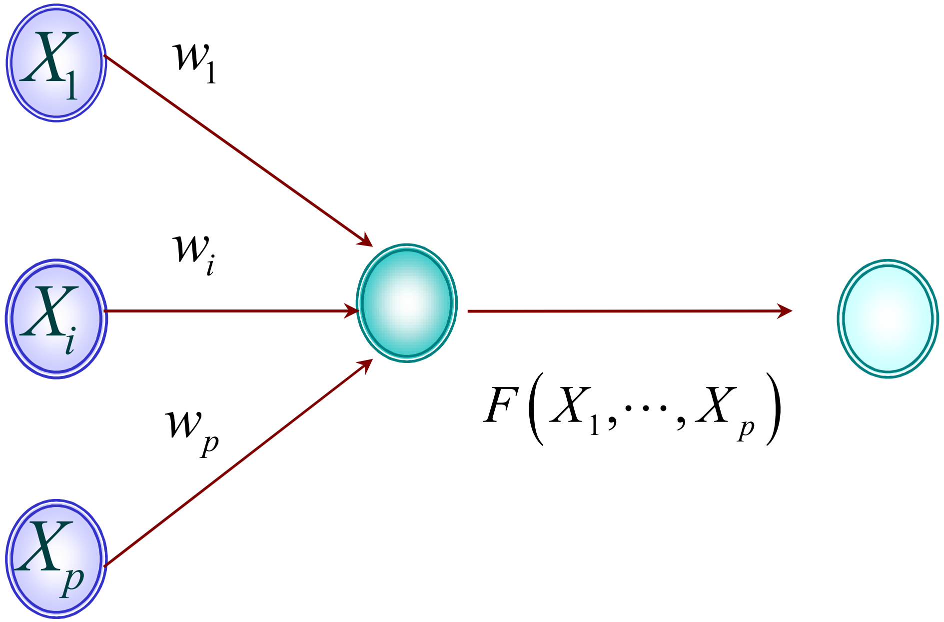
DENDRITAS
(RECOGE)

SOMA
(PROCESA)

AXÓN
(TRANSMITE)

Las Redes Neuronales

NODO



Distintos modos de procesar dan lugar a diferentes tipos de redes neuronales

Clasificación

↑ ***CONEXIONES***

- ***CONEXIONES HACIA DELANTE (FEEDFORWARD)***
- ***CONEXIONES HACIA DELANTE Y HACIA ATRÁS (FEEDFORWARD-FEEDBACK)***

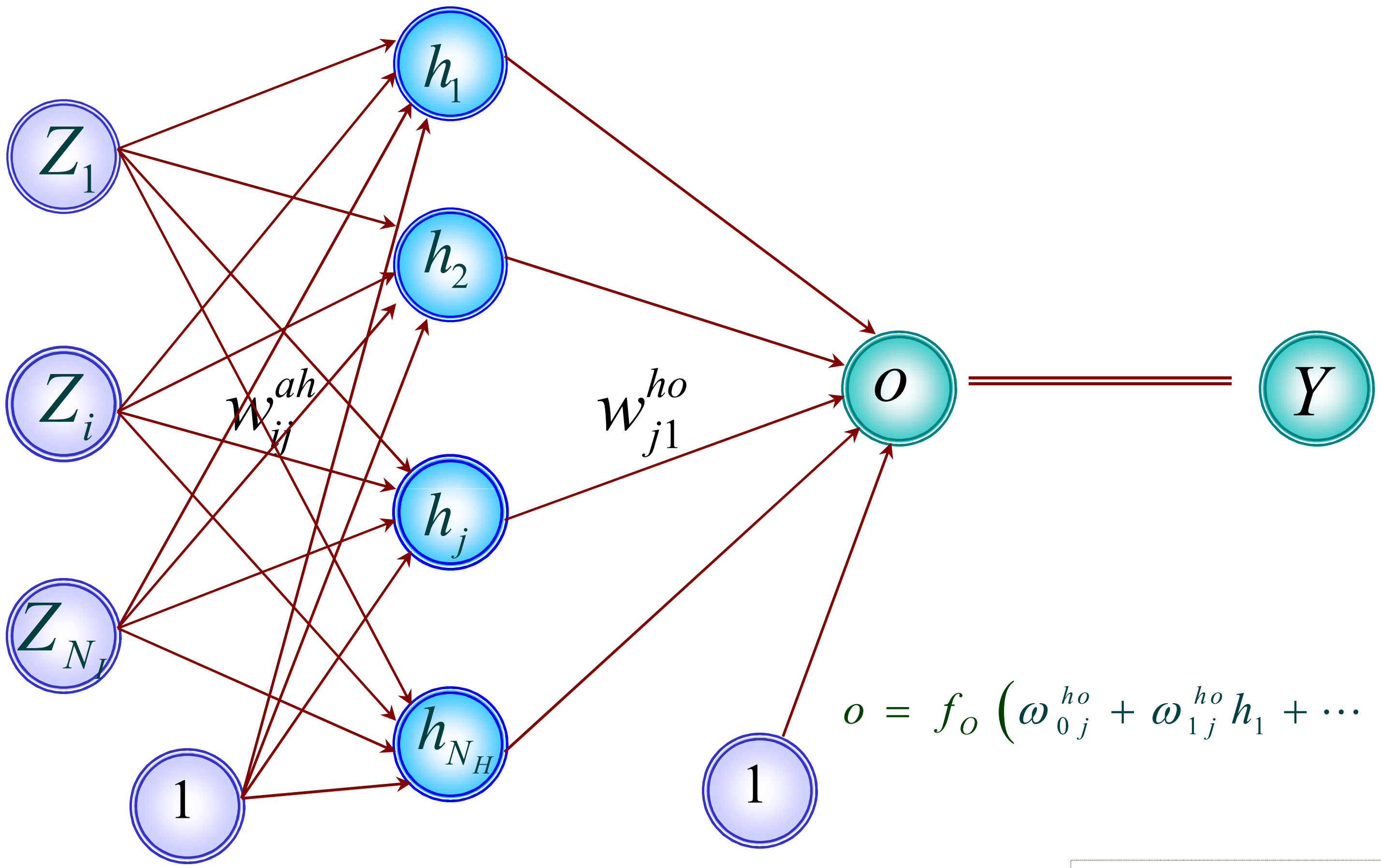
↑ ***TOPOLOGÍA***

- ***EL PERCEPTRÓN***
- ***RED NEURONAL DE BASE RADIAL***

↑ ***APRENDIZAJE***

- ***REDES ON-LINE FRENTE A REDES OFF-LINE***
- ***APRENDIZAJE SUPERVISADO FRENTE A NO SUPERVISADO***

Perceptrón con una Capa Oculta

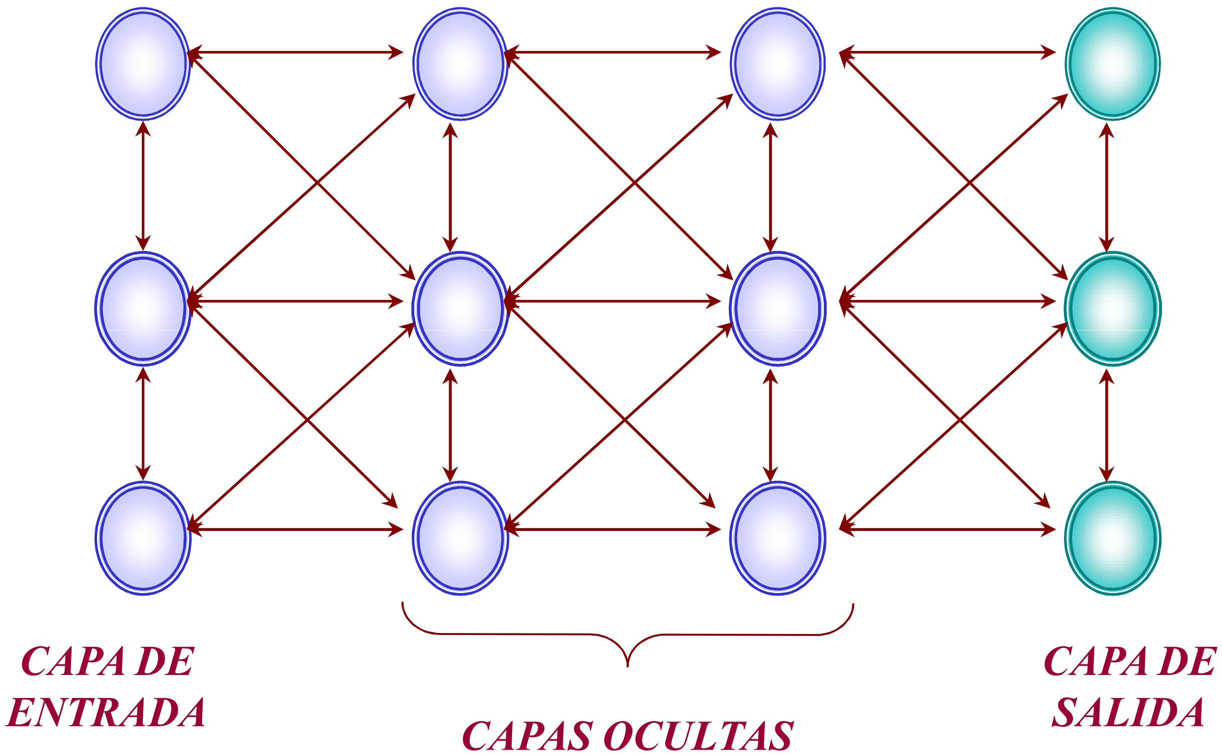


$$o = f_O\left(\omega_{0j}^{ho} + \omega_{1j}^{ho} h_1 + \cdots + \omega_{N_H j}^{ho} h_{N_H}\right)$$

Salida de la Red

$$h_j = f_h\left(\omega_{0j}^{ah} + \omega_{1j}^{ah} Z_1 + \cdots + \omega_{N_I j}^{ah} Z_{N_I}\right)$$

Red Neuronal



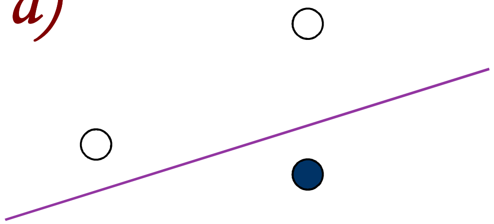
Dimensión VC. Ejemplos

3 Sea $U = \Re^2$ y sea C el conjunto de semiespacios abiertos

$$C = \{ax + by < c, \text{con } a, b, c \in \Re\}$$

C puede separar cualquier conjunto de **3** elementos no colineales, pero no puede separar ningún conjunto de **4** elementos, luego $\boxed{VCD(C) = 3}$

a)



b)

Dimensión VC en Redes Neuronales

PERCEPTRÓN SIMPLE

El perceptrón simple se comporta como un discriminante lineal en $\Re^{N_I}$

Sea $\mathcal{F} = \mathcal{P}_{\mathcal{N}_I}$ la familia de funciones asociadas al perceptrón

$$f : U = \Re^{\mathcal{N}_I} \to \Re, \text{ tales que}$$

$$f(\mathbf{Z}) = f\left(Z_1, \cdots, Z_{\mathcal{N}_I}\right) = w_{01}^{ao} + w_{11}^{ao} Z_1 + \cdots + w_{\mathcal{N}_I 1}^{ao} Z_{\mathcal{N}_I}$$

Esta familia está parametrizada por los vectores

$\left(w_{01}^{ao}, w_{11}^{ao}, \cdots, w_{N_I 1}^{ao}\right) \in \Re^{N_I}$ Luego $\boxed{\dim \mathcal{P}_{\mathcal{N}_I} = VCD\left(\mathcal{P}_{\mathcal{N}_I}\right) = \mathcal{N}_I + 1}$

La Estadística y las Redes Neuronales

El origen de las redes neuronales, totalmente ajeno a la estadística, hizo que esta relación no fuese evidente para aquellos científicos involucrados en el estudio de las redes neuronales. Crearon con el tiempo una terminología propia que en muchos casos podemos vincular con términos estadísticos

Este paralelismo no se detiene sólo en la terminología. Existe una estrecha vinculación entre las redes neuronales y la estadística. Muchos ***métodos estadísticos,*** unos clásicos y otros más recientes, han sido ***reescritos***, no siempre de forma consciente, como ***redes neuronales***.

Veamos con más detalle el paralelismo existente entre la terminología estadística y la terminología asociada a las redes neuronales

Redes Neuronales	Estadística
CARACTERÍSTICAS	VARIABLES
INPUTS	VARIABLES INDEPENDIENTES
OUTPUTS	PREDICCIONES
VALORES OBJETIVO O DE APRENDIZAJE	VARIABLES DEPENDIENTES
ERRORES	RESIDUOS
APRENDIZAJE, ENTRENAMIENTO, AUTOORGANIZACIÓN	ESTIMACIÓN
FUNCIÓN DE ERROR, DE COSTE O DE LIAPUNOV	CRITERIO DE ESTIMACIÓN

Redes Neuronales	Estadística
PATTERNS	OBSERVACIONES
PESOS	PARÁMETROS ESTIMADOS
APRENDIZAJE SUPERVISADO	ANÁLISIS DISCRIMINANTE Y REGRESIÓN
APRENDIZAJE SIN SUPERVISIÓN	REDUCCIÓN DE LA DIMENSIÓN
APRENDIZAJE COMPETITIVO	ANÁLISIS CLUSTER
GENERALIZACIÓN	INTERPOLACIÓN, ESTRAPOLACIÓN
VALIDATION SET	MUESTRA
	POBLACIÓN

La Estadística y las Redes Neuronales

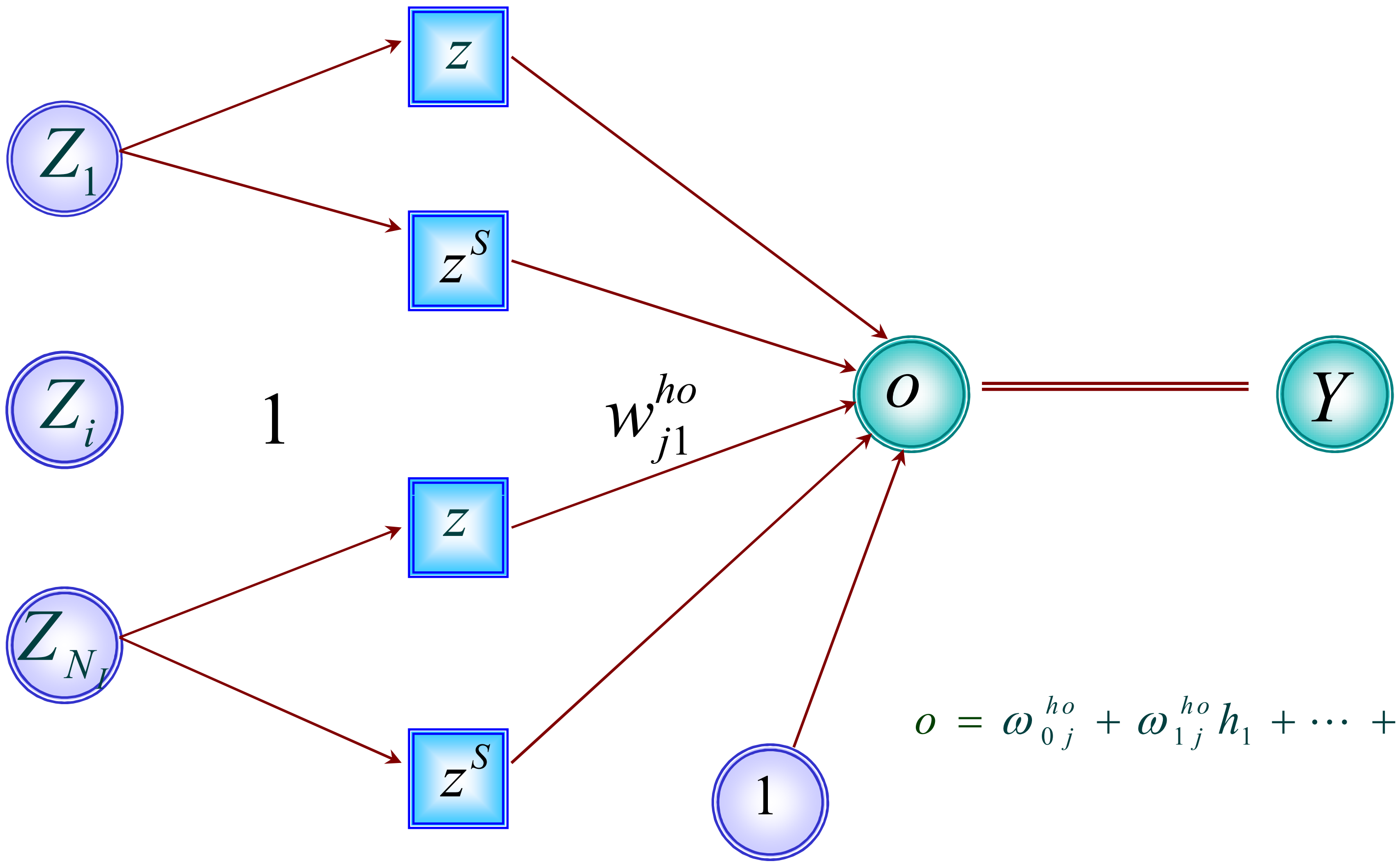
Hemos señalado que muchos métodos estadísticos pueden ser reescritos a través de alguna red neuronal.

Esto se extiende a métodos paramétricos y no paramétricos de regresión, de estimación de la función de densidad, técnicas de reducción de la dimensionalidad,...

Algunos de estos métodos son:

- ***Regresión Lineal***
- ***Regresión Polinómica***
- ***Regresión Logística***
- ***Modelos G.L.M.***
- ***Modelos G.A.M.***
- ***Regresión Projection Pursuit***
- ***Modelos G.G.A.M.***
- ***Single Index Model***
- ***Análisis Factorial***
- ***Análisis de Componentes Principales***

Regresión Polinómica



$$o = \omega_{0j}^{ho} + \omega_{1j}^{ho} h_1 + \cdots + \omega_{N_H j}^{ho} h_{N_H}$$

Salida de la Red

$$h_j = \left(z_i\right)^k, \; con\; 1 \le j \le N_I * S = N_H$$

$$j = (i-1) * S + k, \; con\; 1 \le i \le N_I, \; con\; 1 \le k \le S$$

Regresión G.L.M.

LA IDEA.

La ***regresión polinómica*** que tratamos previamente era una generalización inmediata de la ***regresión lineal***, en tanto en cuanto ésta es una regresión polinómica de grado 1.

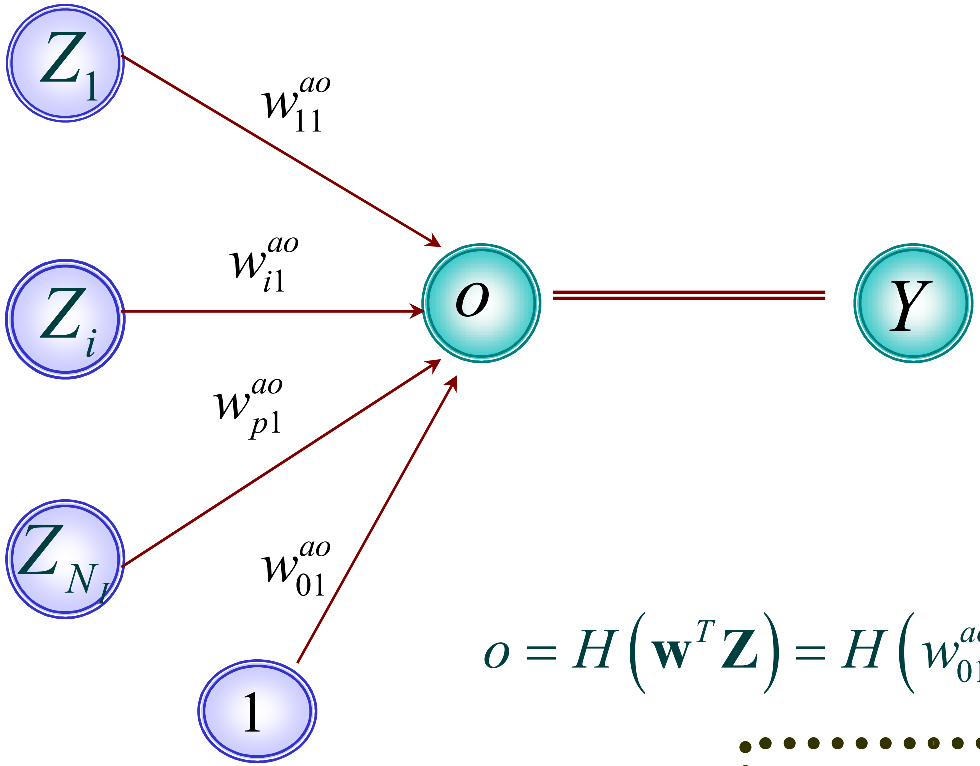
Otro camino por el que se ha tratado de generalizar la regresión lineal consiste en ***aplicarle a la combinación lineal una función determinada***. De este modo estimamos la relación entre un conjunto de variables regresoras $Z_1, Z_2, \cdots, Z_{N_I}$ y una variable respuesta Y a través de una función de la forma

$$\mathrm{E}\left[Y/Z_1, \ldots, Z_{N_I} \right] = H\left(a + b_1 Z_1 + b_2 Z_2 + \cdots + b_{N_I} Z_{N_I} \right)$$

Siendo H una *función conocida.*

Estos modelos reciben el nombre de **Modelos Lineales Generalizados**.

Modelos Lineales Generalizados



$$o = H\left(\mathbf{w}^T\mathbf{Z}\right) = H\left(w_{01}^{ao} + w_{11}^{ao}Z_1 + \cdots + w_{p1}^{ao}Z_p\right)$$

Salida de la Red

Regresión G.G.A.M.

LA IDEA.

Consideramos la misma idea que en los G.A.M. Pero en este nuevo contexto serán desconocidas tanto la función H como las $f_1, f_2, \cdots, f_{N_I}$ desconocidas.

$$\mathrm{E}\left[Y/Z_1,\ldots,Z_{N_I}\right] = H\left(f_1\left(Z_1\right) + f_2\left(Z_2\right) + \cdots + f_{N_I}\left(Z_{N_I}\right)\right)$$

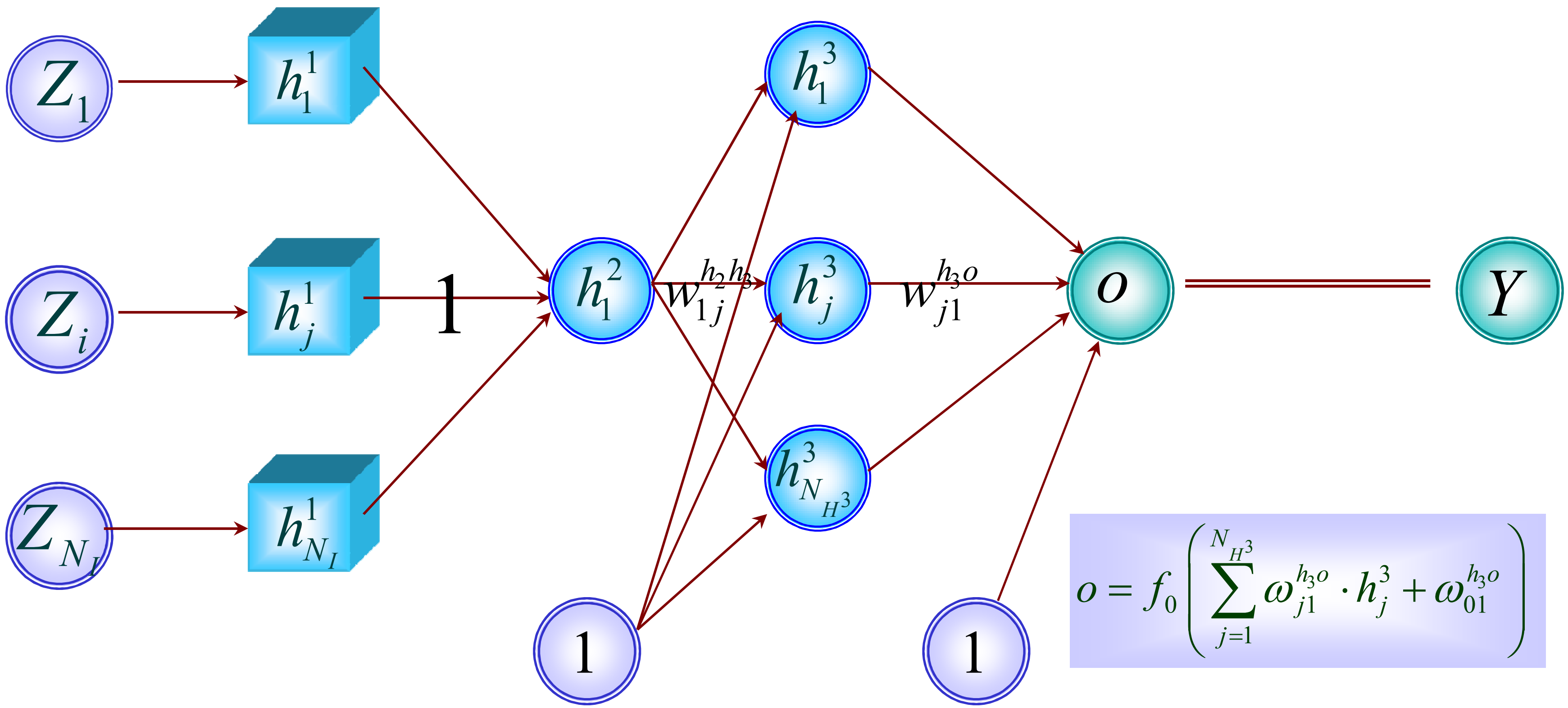
Estos modelos reciben el nombre de ***Modelos Aditivos Generalizados Generales.***

Hemos de estimar una función, H, y sus entradas son a su vez funciones estimadas, lo que aumenta el número de capas ocultas de la red. También ahora existe un método clásico, iterativo, de estimación de la H.

Según qué red, utilicemos para las estimaciones de las funciones $f_1, f_2, \cdots, f_{N_I}$

y de la H tendremos distintos modelos de redes que se adaptan a los G.G.A.M.

Modelos G.G.A.M.



$$o = f_0\left(\sum_{j=1}^{N_{H^3}} \omega_{j1}^{h_3 o} \cdot h_j^3 + \omega_{01}^{h_3 o}\right)$$

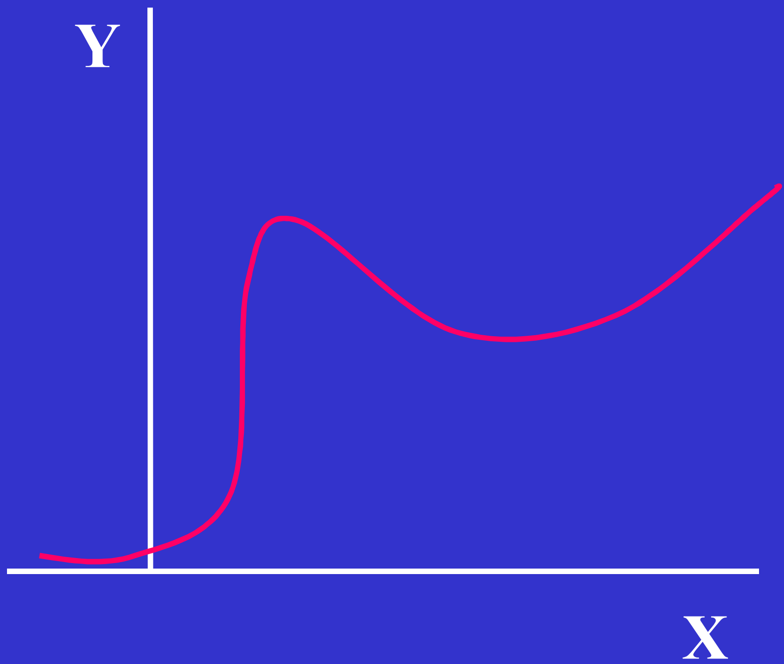
h_i^1 Corresponde a la previsión de $f_i\left(Z_i\right)$

Es una subred

$$h_1^2 = \sum_{i=1}^{N_I} h_i$$

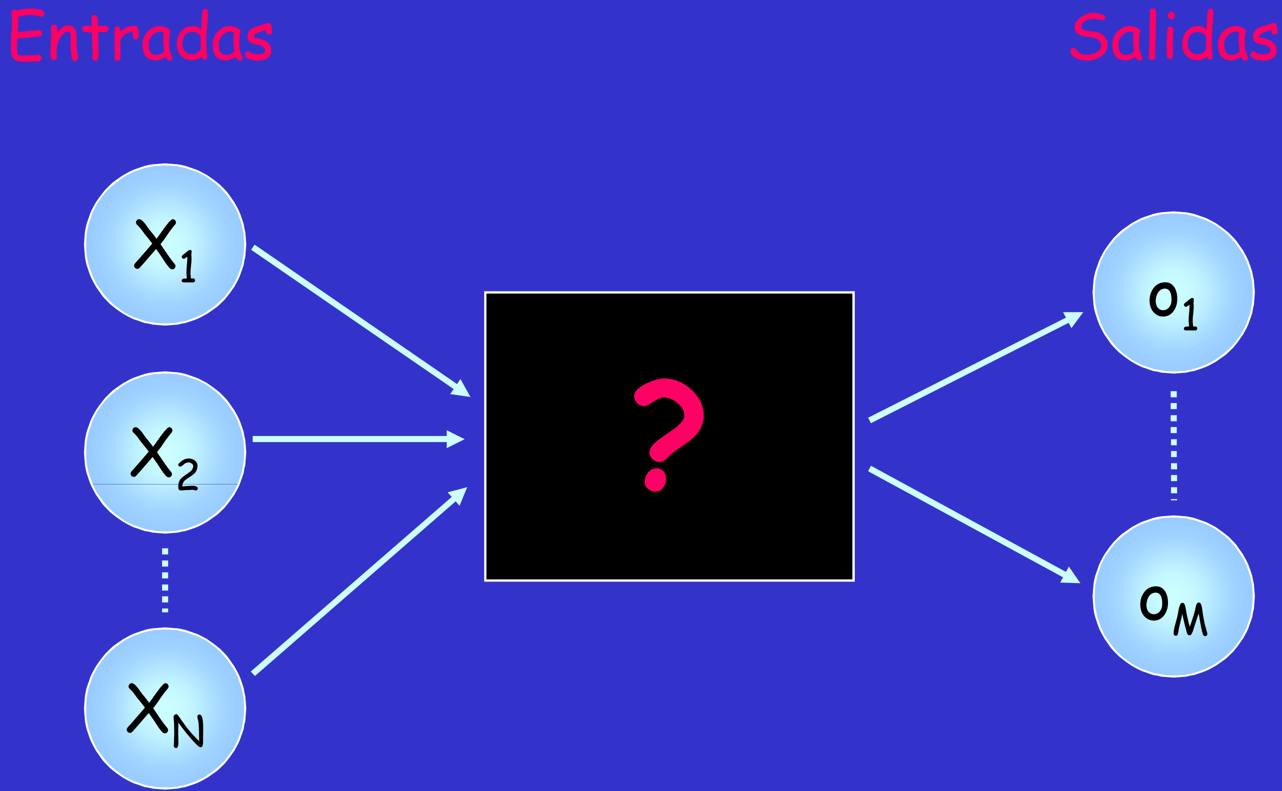
$$h_j^3 = f_{h^3}\left(\omega_{1j}^{h_2 h_3} \cdot h_1^2 + \omega_{0j}^{h_2 h_3}\right)$$

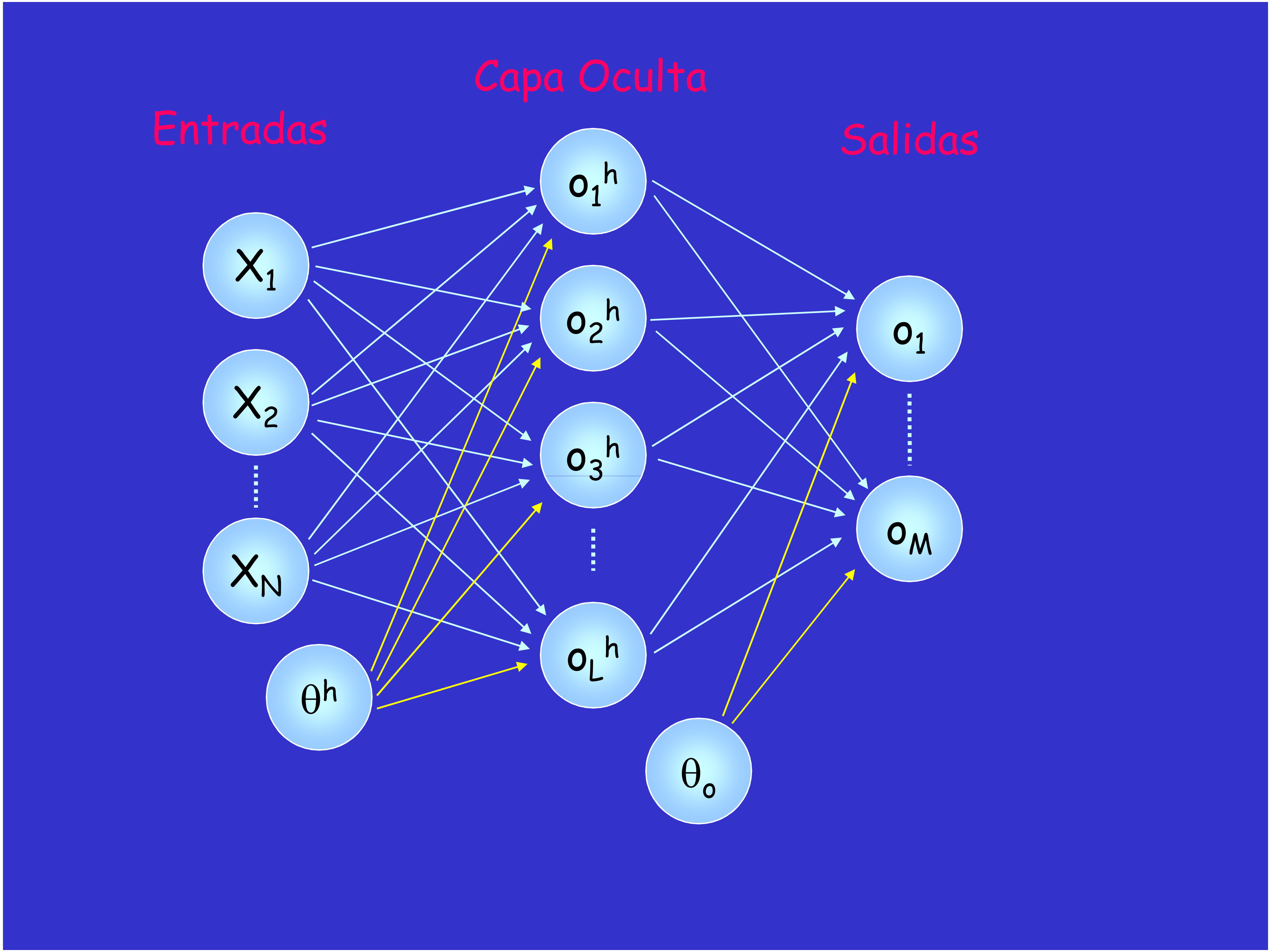
$$para \quad j = 1, \cdots, N_{H^3}$$



$$Y = F(X)$$

Utilizaremos las redes neuronales para estimar la función F que explica Y con X


Entradas
Salidas
X1
X2
XN
?
o1
oM



Capa Oculta
Entradas
Salidas
X_1
X_2
X_N
θ^h
o_1^h
o_2^h
o_3^h
o_L^h
o_1
o_M
θ_o

Topología de una RED NEURONAL

- N: dimensión del vector de entrada
- M: dimensión del vector de salida
- Fijar el nº de capas ocultas
- Fijar L: nº de nodos cada capa oculta
- Fijar las funciones de activación f_j

Salidas de la red neuronal

Dado $\mathbf{X} = (X_1, \ldots, X_N)$

$$o_k^o = \theta_k^o + \sum_{j=1}^{L} \omega_{jk}^o f_j\left(\theta_j^h + \sum_{i=1}^{N} \omega_{ji}^h X_i \right)$$

$$k = 1, \ldots M$$

Las funciones de activación habituales: f_j

- Identidad $f(x) = x$
- Logística $f(x) = \dfrac{1}{1 - e^{-x}}$
- Función umbral (clasificación)

Faltan por determinar los pesos

$$\omega_{11}^{h},\ \ldots,\ \omega_{LN}^{h},\ \omega_{11}^{o},\ \ldots,\ \omega_{ML}^{o}$$

y las tendencias

$$\boldsymbol{\theta}^{h} = (\theta_{1}^{h},\ \ldots,\ \theta_{L}^{h}) \qquad \boldsymbol{\theta}^{o} = (\theta_{1}^{o},\ \ldots,\ \theta_{M}^{o})$$

Los buscamos de forma que

$$o_{k} \approx y_{k}$$

$$\forall\, k = 1, \ldots, M$$

Entrenamiento

Conjunto de entrenamiento

$$(\mathbf{X}_1, \mathbf{Y}_1),\ \ldots,\ (\mathbf{X}_P, \mathbf{Y}_P)$$

Criterio de error

$$\frac{1}{2}\sum_{k=1}^{M}\left(y_k - o_k\right)^2$$

Buscamos lo pesos que minimizan el error

Algoritmo de entrenamiento: BACKPROPAGATION

(basado en algoritmo Steepest Descent)

Este algoritmo busca, en cada iteración del entrenamiento, la dirección de máximo descenso de la función de error.

$$\left(-\frac{\partial EC}{\partial \omega_{kj}^{o}}\right)$$

$$\left(-\frac{\partial EC}{\partial \omega_{ji}^{h}}\right)$$

Y modifica los pesos de la red en esa dirección.

$$\omega_{kj}^{o}(p+1)=\omega_{kj}^{o}(p)-\eta\frac{\partial EC}{\partial \omega_{kj}^{o}}\omega_{kj}^{o}(p)$$

$$\omega_{ji}^{h}(p+1)=\omega_{ji}^{h}(p)-\eta\frac{\partial EC}{\partial \omega_{ji}^{h}}\omega_{ji}^{h}(p)$$

Pasos del Algoritmo

1.- Aplicar $\mathbf{X_p} = (X_{p1}, \ldots X_{pN})$ a los nodos de entrada.

2.- Calcular las salidas de la capa de salida:

$$o_1, \ldots o_k, \ldots o_M$$

3.- Calcular los errores de la capa de salida:

$$e_k = (Y_k - o_k)^2 \quad \forall k = 1, \ldots M$$

Si el error es admisible ir al paso 6.

En otro caso continuar

4.- Se calculan las direcciones de máximo descenso.

5.- Se actualizan los pesos, según las ecuaciones vistas.

6.- Hacer p=p+1. Volver al paso 1.

Algunas observaciones:

- Será necesario recorrer varias veces el conjunto de entrenamiento.
- El orden en que se recorre el conjunto de entrenamiento debe ser aleatorio, y distinto cada vez.
- El criterio de parada:

  Error admisible.

  Nº máximo de iteraciones.
- η: paso o parámetro de velocidad de aprendizaje. Su valor determina la velocidad con la que la red aprende.

El conjunto de entrenamiento

¡¡ Es esencial escoger bien el conjunto de entrenamiento !!

- Sólo APRENDE lo que ocurre en el conjunto de entrenamiento

- No extrapola bien

Conjunto de entrenamiento erróneo → Mal aprendizaje → Mal funcionamiento de la red

El número de nodos de la capa oculta: L

Supongamos una red como la de la figura: red

El nº L determina el nº de parámetros de la red:

$$n^o\ parámetros = (N+1) * L + M * L$$

Por tanto: $$L \in \left[1, \frac{P}{N+1+M}\right]$$

Ya que la muestra de entrenamiento es de tamaño P.

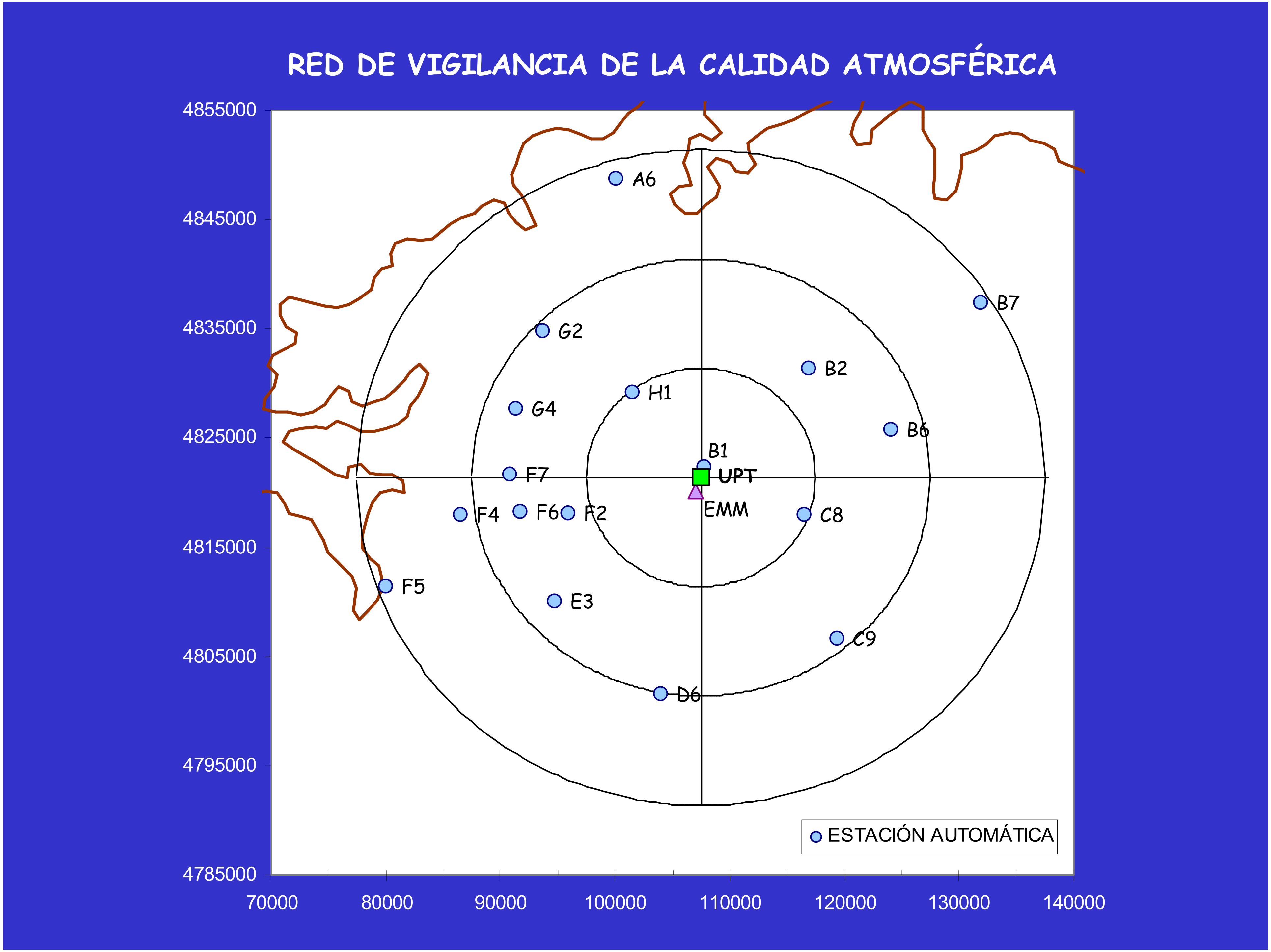
El efecto de L en los resultados de la red es similar al de la ventana en la estimación no paramétrica:

- Valores bajos de L: la red promedia los datos del conjunto de entrenamiento (sobresuaviza)

- Valores altos de L: la red reproduce los datos del conjunto de entrenamiento (infrasuaviza)

NUESTRO PROBLEMA

Predecir el nivel de inmisión de SO_2 alrededor de la U. P. T. de Endesa en As Pontes


RED DE VIGILANCIA DE LA CALIDAD ATMOSFÉRICA
4855000
4845000
4835000
4825000
4815000
4805000
4795000
4785000
70000
80000
90000
100000
110000
120000
130000
140000
A6
B7
G2
B2
H1
G4
B6
B1
UPT
F7
EMM
F4
F6
F2
C8
F5
E3
C9
D6
ESTACIÓN AUTOMÁTICA

- Recibimos datos cada 5 minutos
- La legislación controla la media horaria de la inmisión

Por tanto, nuestra variable de interés será:

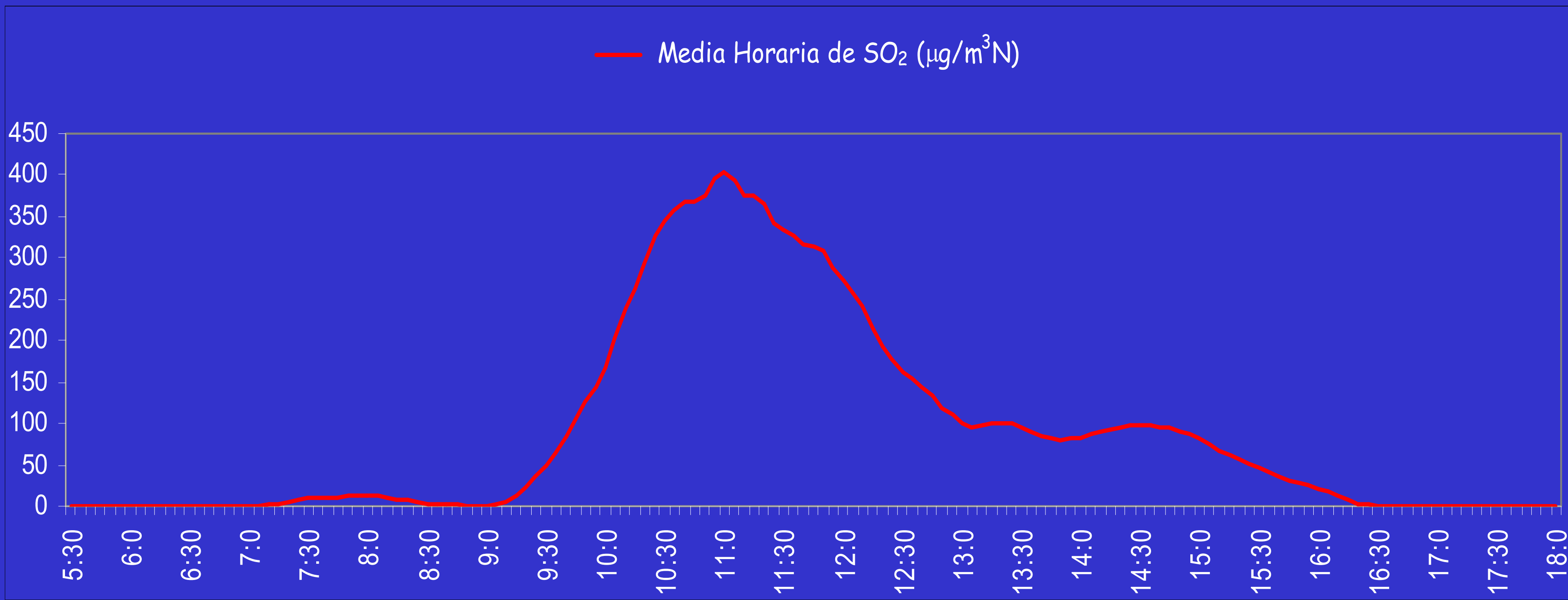
X_t Media horaria arrastrada de la inmisión de SO_2 en el instante t

Tenemos una serie temporal:

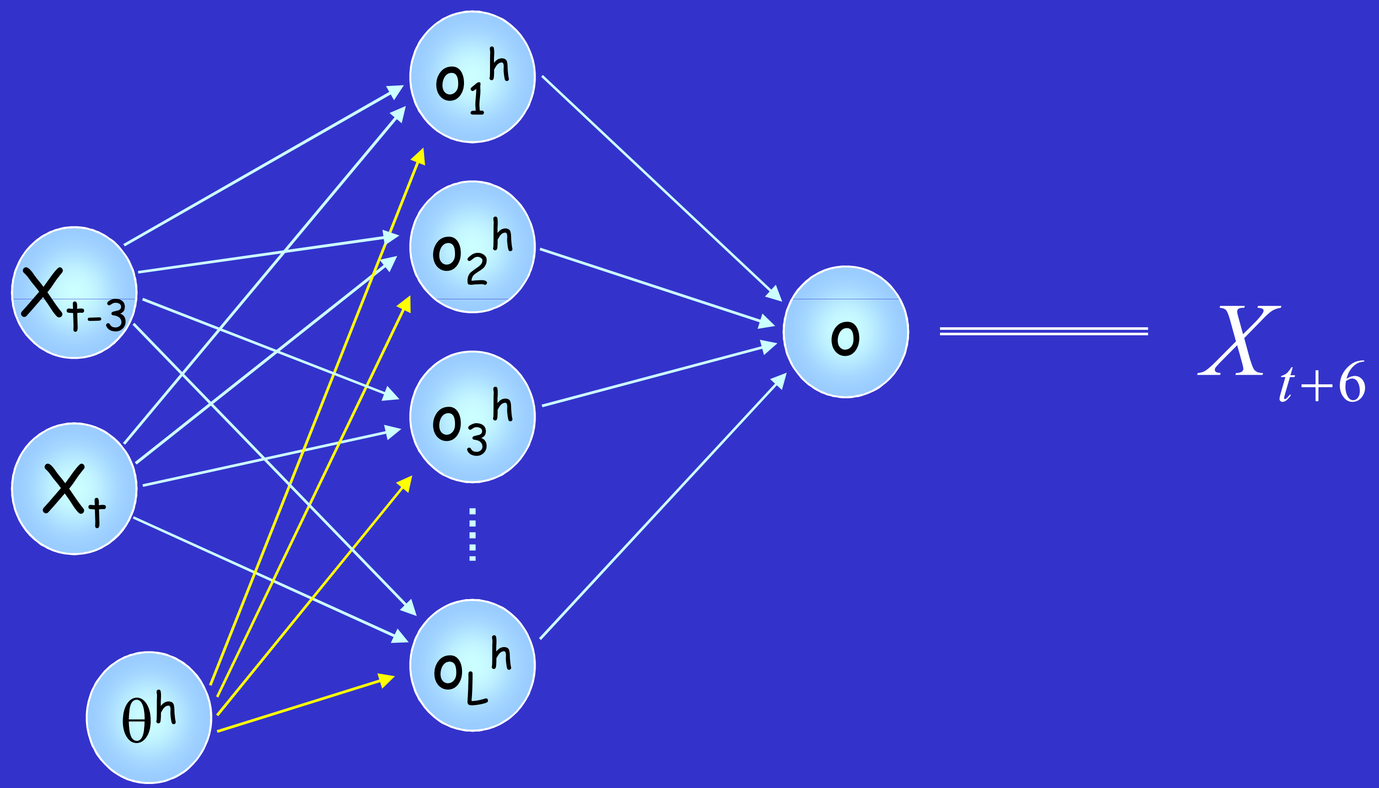
$$\ldots, X_{t-1}, X_t, X_{t+1}, \ldots, X_{t+h}, \ldots$$

- Habitualmente la serie $\{X_t;\, t = \ldots -1,\ 0,\ +1 \ldots\}$ se compone de valores cercanos a 0.

- En condiciones meteorológicas desfavorables, puede ocurrir un episodio de inmisión:



Red para predicción de niveles de inmisión con 1/2 hora de antelación

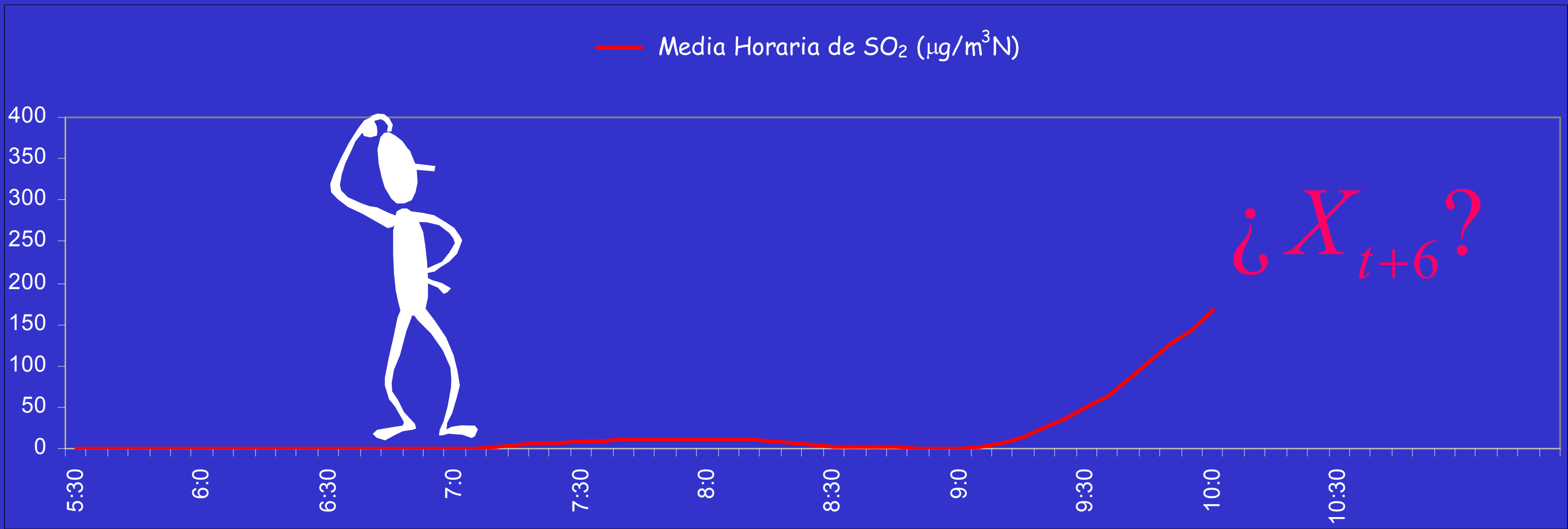


Nos interesa

➢ Predecir los episodios con antelación para que se puedan tomar medidas preventivas en la U. P. T.

Concretamente:

haremos predicciones a ½ hora de la media horaria



Diseño del conjunto de entrenamiento

- ~~Serie temporal~~
- Matriz Histórica:

2000 registros

Divididos en estratos

Utilizamos datos de 1999

$(X^1_{t-3}, X^1_t, X^1_{t+6})$

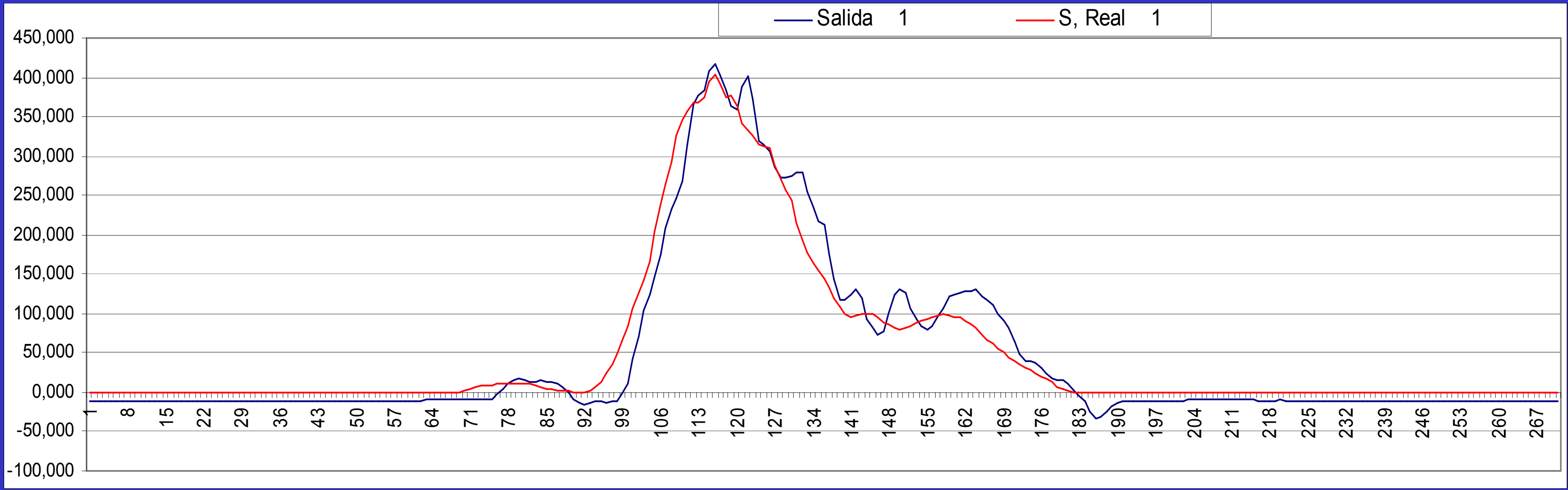
$(X^i_{t-3}, X^i_t, X^i_{t+6})$

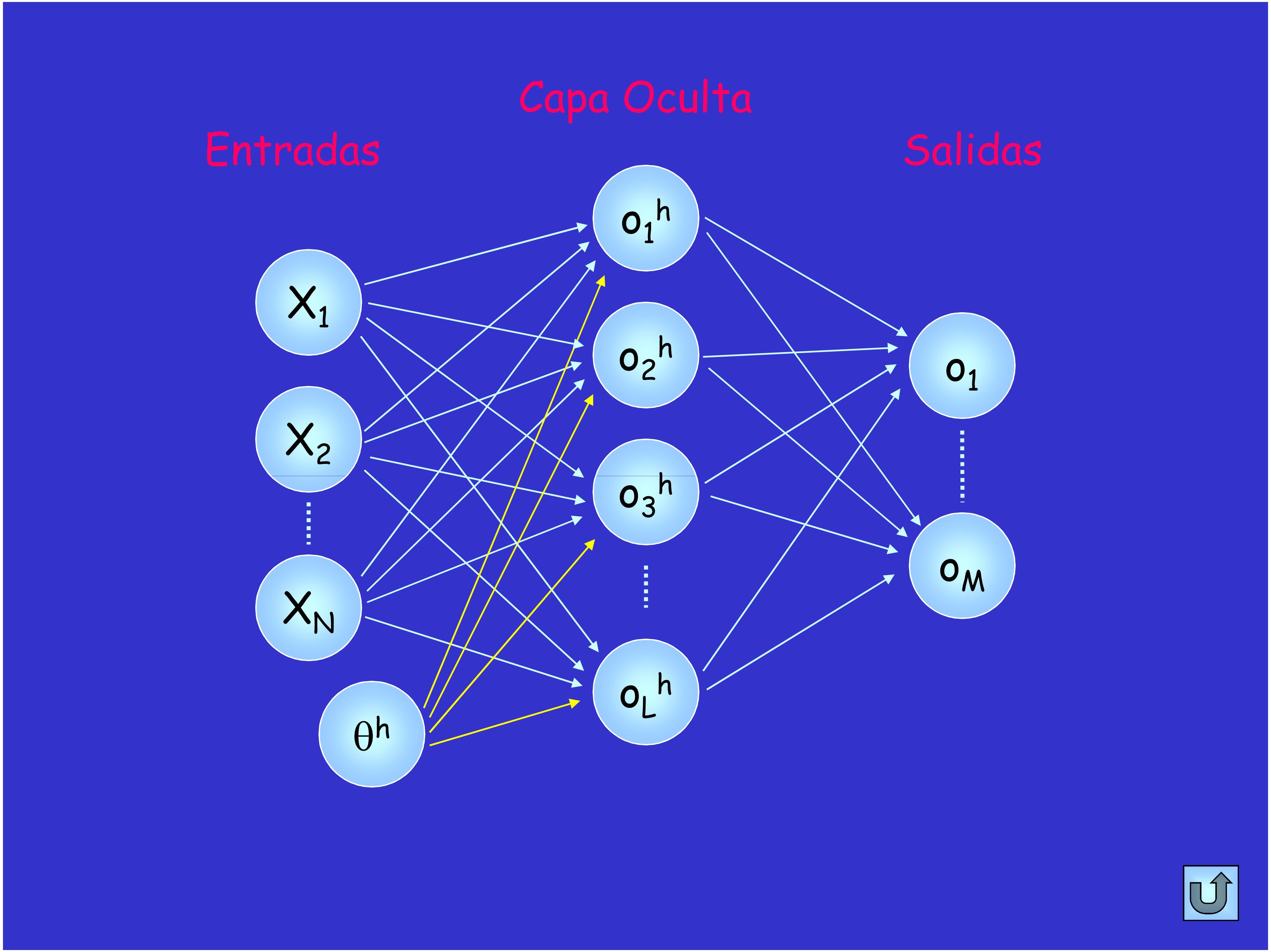
$\vdots$

$(X^P_{t-3}, X^P_t, X^P_{t+6})$

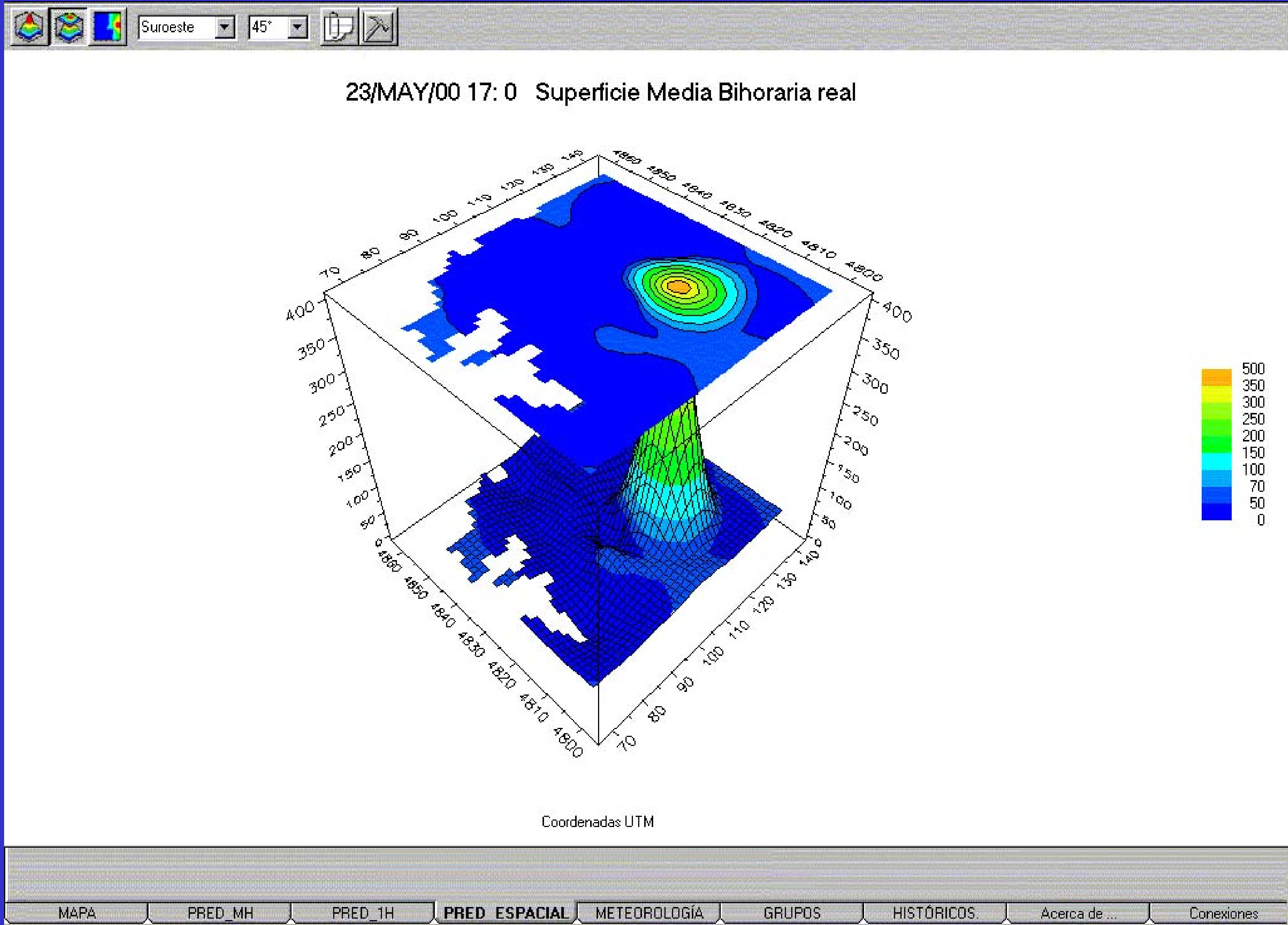
Resultados

20/05/2000 F4





Capa Oculta
Entradas
Salidas
X_1
X_2
X_N
θ^h
o_1^h
o_2^h
o_3^h
o_L^h
o_1
o_M


Suroeste
45°
23/MAY/00 17: 0 Superficie Media Bihoraria real
500
350
300
250
200
150
100
70
50
0
Coordenadas UTM
MAPA
PRED_MH
PRED_1H
PRED_ESPACIAL
METEOROLOGÍA
GRUPOS
HISTÓRICOS.
Acerca de ...
Conexiones

Predicción con Datos Funcionales

Consideramos variables aleatorias con valores en:

$$H = L^2\left(\left[a,b\right]\right)$$

Observando un proceso estocástico en tiempo continuo: $x(t), t \leq T$

Pretendemos predecir valores futuros: $x(t), t \geq T$

Consideramos el modelo estadístico:

$$X_n = \rho\left(X_{n-1}\right) + \varepsilon_n$$

ε_n es un ruido blanco fuerte de Hilbert

$\rho : H \to H$ es el operador funcional a estimar

Técnicas de Estimación

Modelo Autorregresivo de Hilbert (ARH)

$$\hat{X}_n = \hat{\rho}\left(X_{n-1}\right)$$

El estimador $\hat{\rho}$ se obtiene a partir de la expresión $D = \rho C$

con C y D matrices de covarianzas y covarianzas cruzadas.

Núcleo funcional (FK)

$$\hat{\rho}_{h_n}(x) = \frac{\sum_{i=1}^{n} X_{i+1} \cdot K\left(\frac{\left\|X_i - x\right\|}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\left\|X_i - x\right\|}{h_n}\right)}$$

con K núcleo gausiano.

Aplicación

Consideramos la serie de datos como:

observaciones de un proceso estocástico en tiempo continuo
que modeliza los niveles medios horarios de SO_2

Consideramos porciones de dicho proceso estocástico que representan 1/2 hora.
Por tanto consideraremos variables aleatorias que toman valores en:

$$H = L^2\left(\left[0,6\right]\right)$$

de la siguiente forma: $X_n(t) = x(6n+t)$

Matriz Histórica

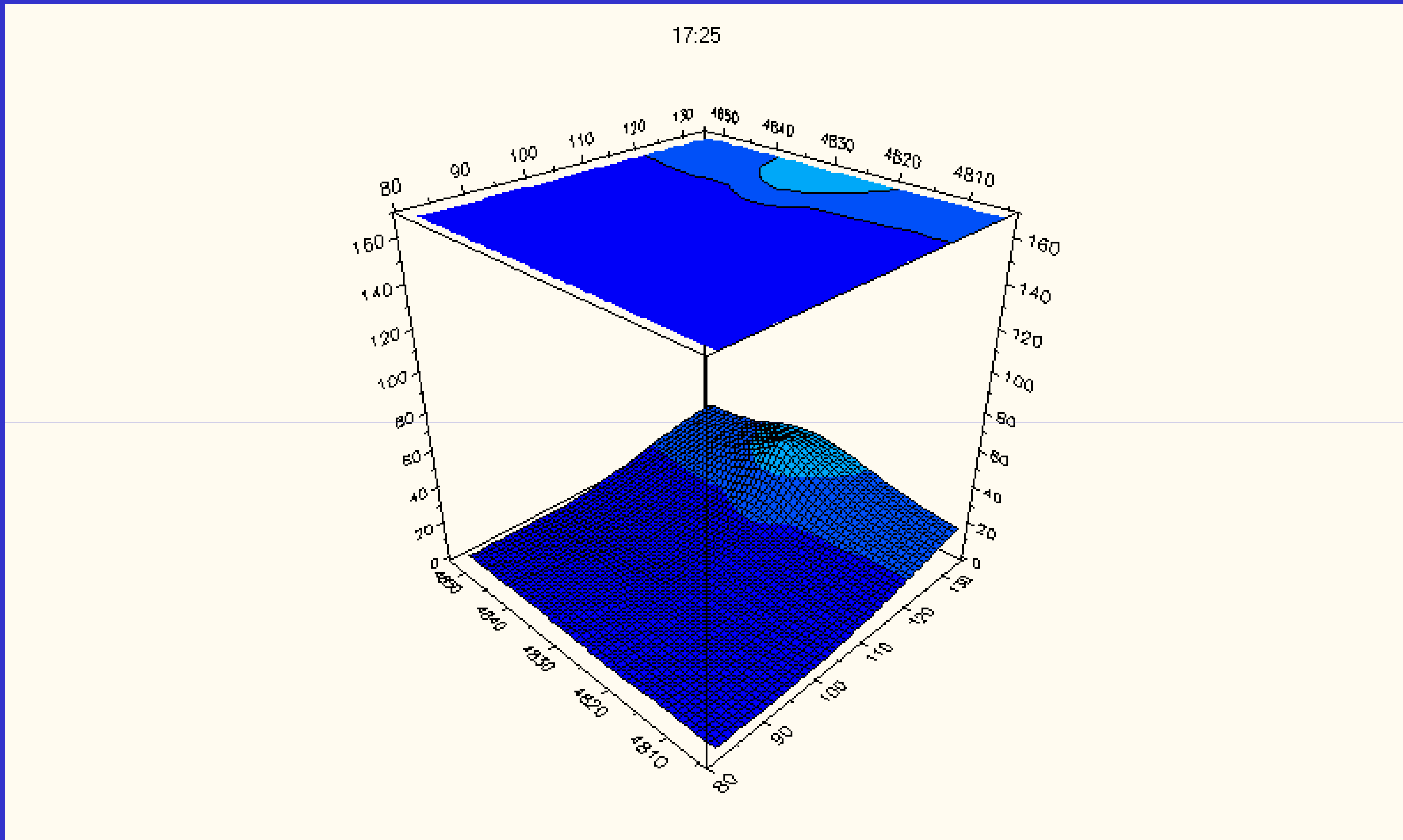
Necesitamos una matriz histórica adecuada a datos funcionales.
Hemos considerado una matriz formada por vectores de la forma:

$$\left(X_t, X_{t+1}\right)$$

Los registros de la matriz están divididos en estratos. Cada vector se incluye en un estrato según una cierta regla de clasificación.

Por niveles:
cada estrato tiene asignado un rango de valores de inmisión, y cada vector se clasifica según el rango al que pertenezca la última observación de: X_{t+1}

Por formas:
cada estrato tiene asignada una forma funcional, y cada vector se clasifica según la forma de: X_{t+1}



17:25
80
90
100
110
120
130
4850
4840
4830
4820
4810
160
140
120
100
80
60
40
20
0

Aplicación de Técnicas Estadísticas al Estudio Hidrológico de la Cuenca del Río Xallas

Wenceslao González Manteiga

María Castellano Méndez

JORNADAS LA MATEMÁTICA DEL AGUA

SANTIAGO DE COMPOSTELA , ABRIL 2002

El problema

La explotación hidroeléctrica de un río requiere un ***conocimiento de los recursos***, en este caso el agua, de los que se dispondrá en un futuro, con el fin de planificar la cantidad de energía que se va a generar y el momento adecuado para hacerlo.

Nuestro objetivo es proporcionar previsiones que sirvan de ayuda en la toma de decisiones en dos situaciones diferentes caracterizadas por el horizonte temporal con respecto al que deseamos realizar las predicciones.

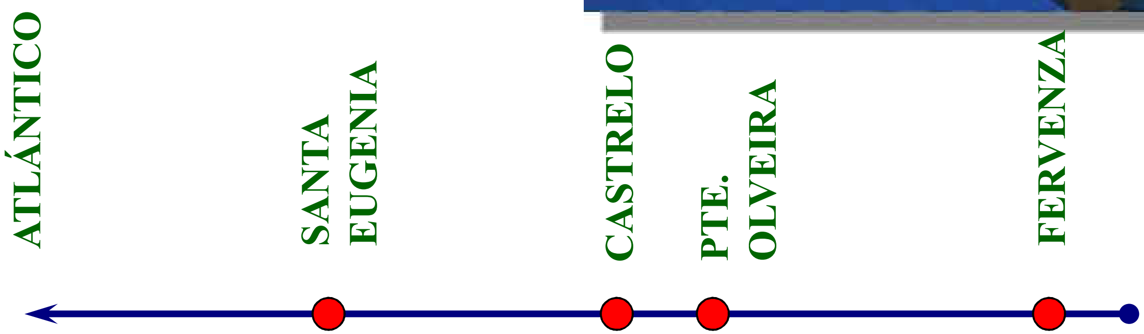
Abordaremos dos problemas diferentes:

- Predicción a Largo Plazo
- Predicción a Corto Plazo

Estos problemas presentarán características diferenciadas que nos llevarán a abordarlos desde diferentes perspectivas y por tanto a emplear ***diferentes técnicas estadísticas*** para su resolución.

El Río Xallas.

A lo largo del curso del río se encuentran situadas dos presas, la de ***Fervenza,*** al final de la cabecera del río, y la ***Santa Eugenia,*** situada justo antes de la desembocadura del río en plena *Costa da Morte*.



Río Xallas

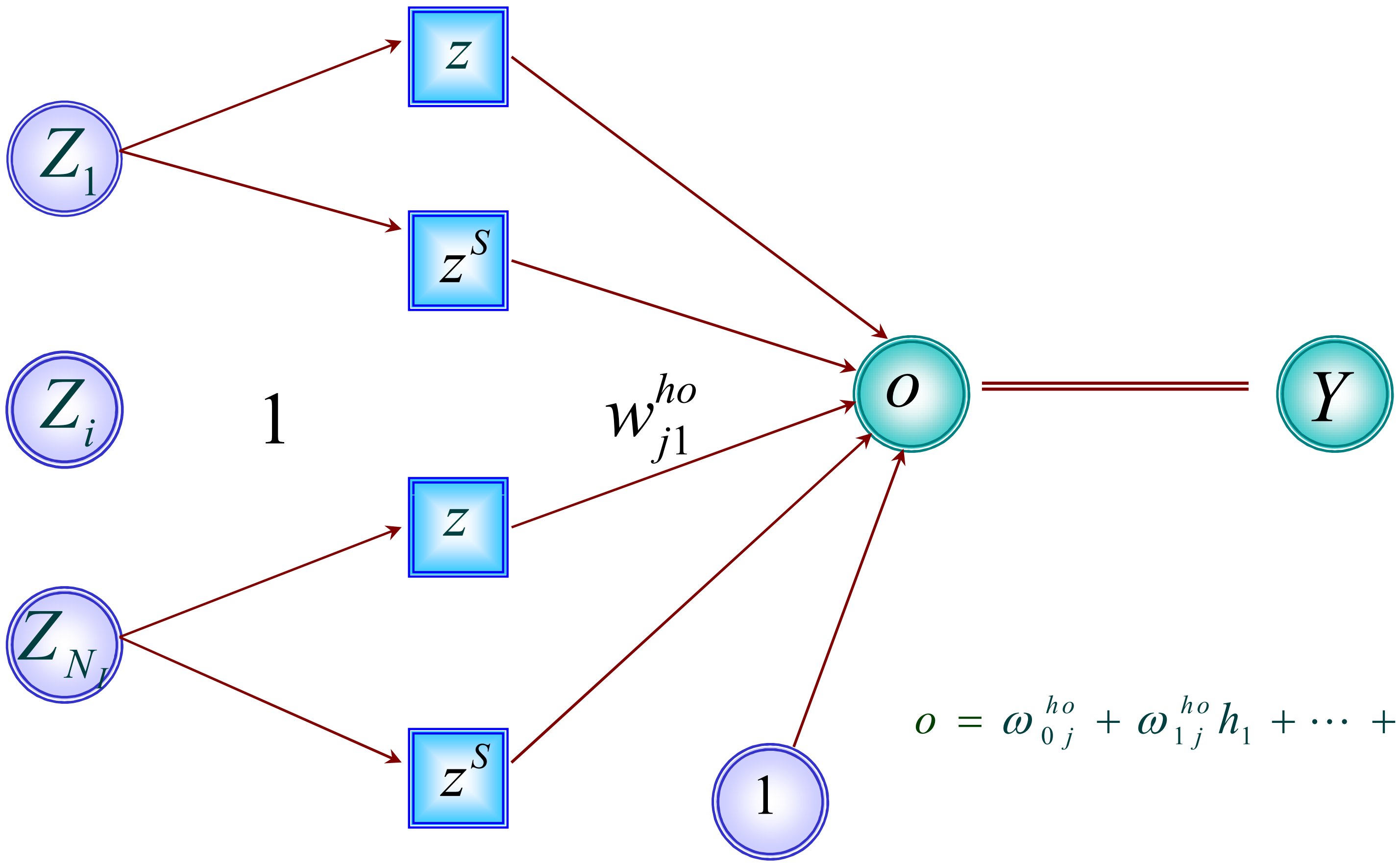
Fervenza

- Entradas:
  - *Lluvia acumulada en Fervenza en los últimos 15 días.*
  - *Lluvia en Fervenza del día previo*
  - *Lluvia en Fervenza del día objetivo* (previsión)
  - *Aportaciones en Fervenza del día previo*

- Conjunto de Entrenamiento:
  - *Datos diarios del período (3454 datos)*

- Conjunto de Validación:
  - *Datos diarios del período 16 de noviembre de 2000, 31 de enero de 2001.(77 datos)*

La Red Seleccionada



$$o = \omega^{ho}_{0j} + \omega^{ho}_{1j} h_1 + \cdots + \omega^{ho}_{N_H j} h_{N_H}$$

$$h_j = \left(z_i\right)^k \text{, } con\ 1 \le j \le N_I * S = N_H$$

$$j = (i-1) * S + k,\ con\ 1 \le i \le N_I,\ con\ 1 \le k \le S$$

Salida de la Red

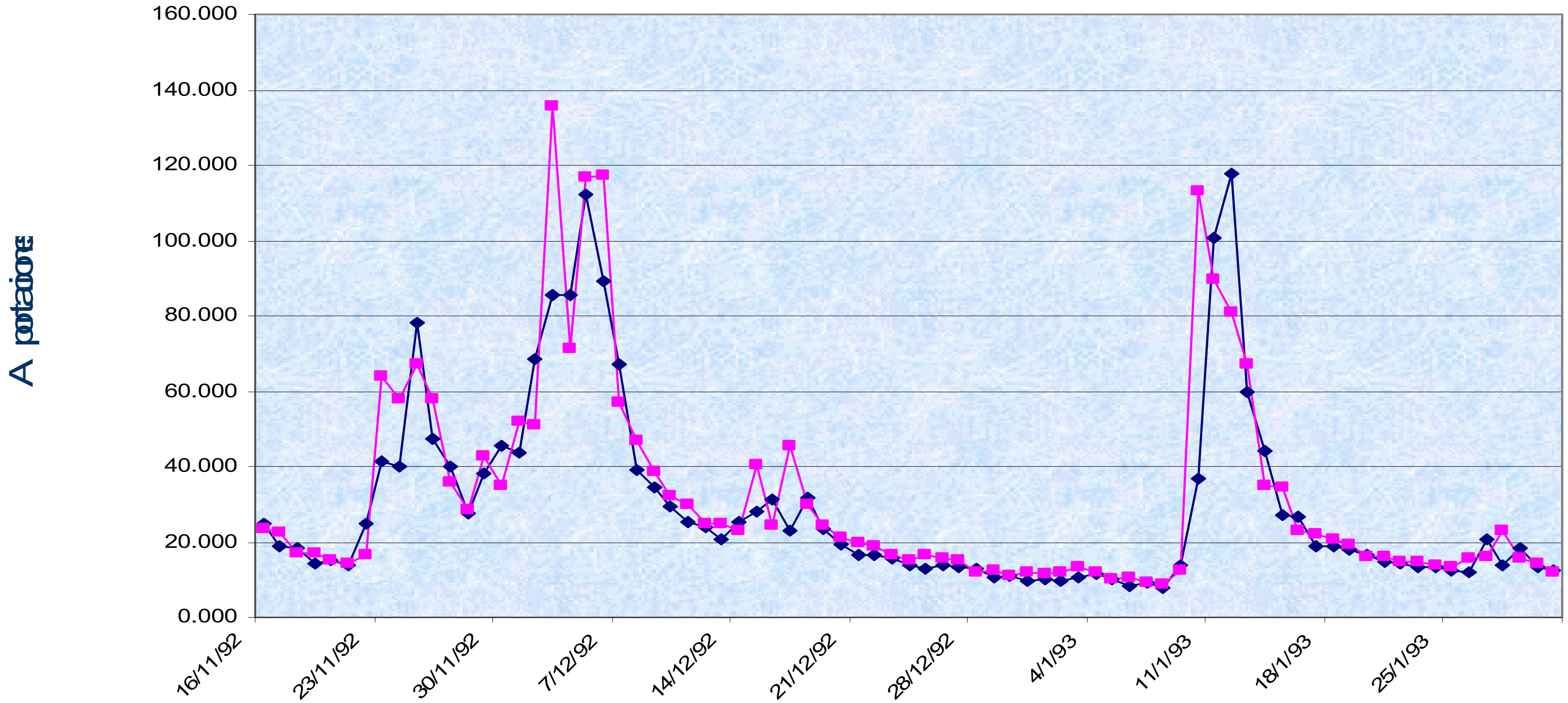
Perceptrón con una Capa Funcional Polinómica

- Características de la Red Polinómica:
  - *Grado del polinomio: 2*
  - *Número de pesos:9*
  - *Variables: ln(1+ vble)*
- Características del Entrenamiento:
  - *Paso:4x10-6*
  - *Número de iteraciones:5000*
  - *Pesos iniciales: U(0,0.5)*

- Resultados del Entrenamiento:
  - *Gráficas de las predicciones frente a la real.*

- Resultados de la Validación:
  - *Error relativo absoluto medio:**0.178***
  - *Gráficas de las predicciones frente a la real*

Sección del Conjunto de Entrenamiento (16/11/1992-31/01/1993)

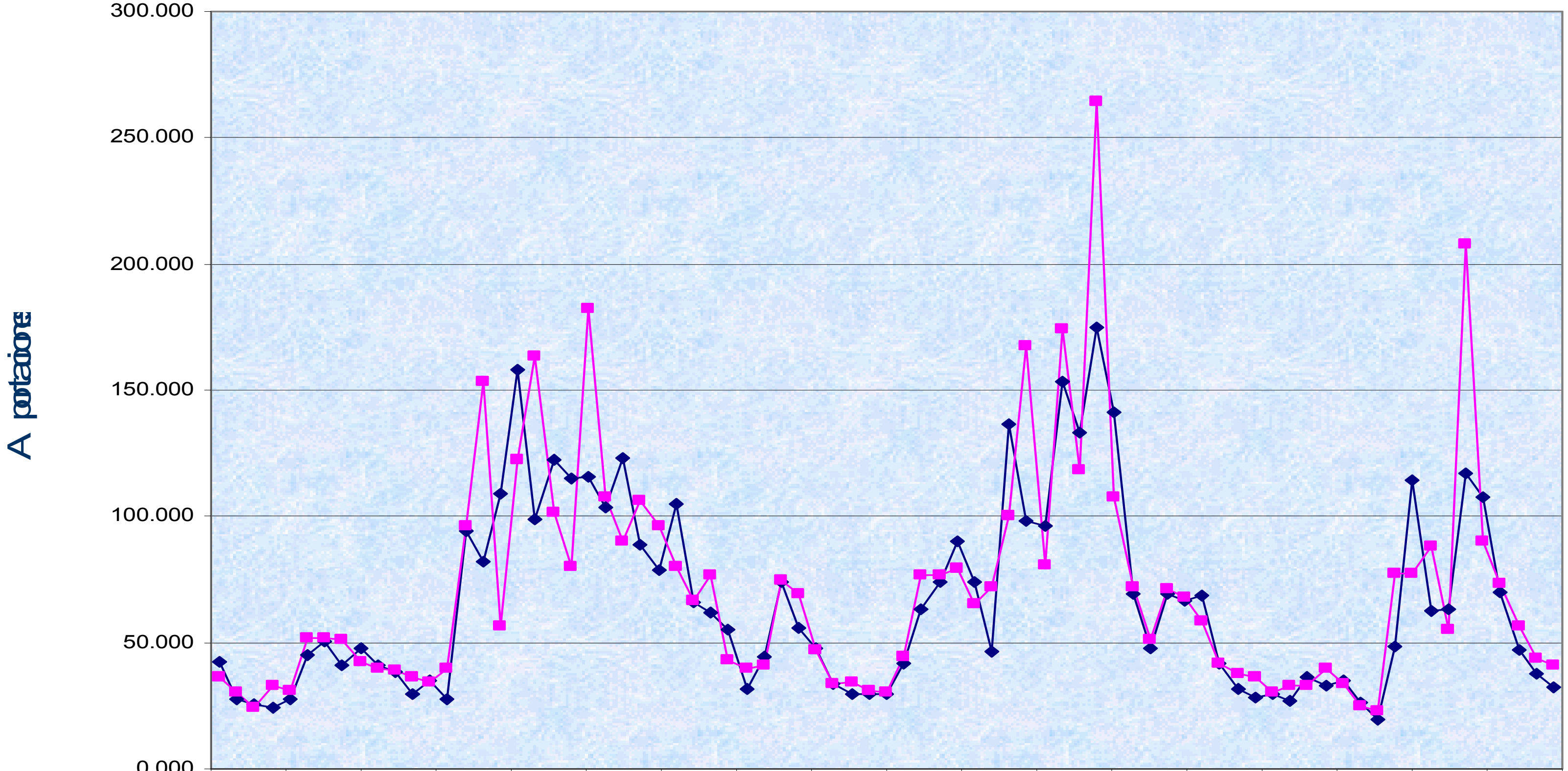


	16-nov-92	21-nov-92	25-nov-92	30-nov-92	4-dic-92	9-dic-92	13-dic-92	18-dic-92	22-dic-92	27-dic-92	31-dic-92	5-ene-93	9-ene-93	14-ene-93	18-ene-93	23-ene-93	27-ene-93
defprev	24.739	13.600	78.202	45.433	85.488	34.552	20.628	31.852	16.322	13.368	9.827	10.179	13.895	44.241	18.787	13.225	20.563
defreal	23.361	14.472	67.033	34.730	71.168	38.845	24.790	30.031	19.045	15.184	11.858	9.903	12.585	35.053	20.911	14.705	15.878

Datos

defprev defreal

Conjunto de Validación (16/11/2000-31/01/2001)



	N16	N20	N25	N29	D3	D7	D12	D16	D20	D25	D29	E2	E6	E11	E15	E19	E23	E28
prevdef	42.250	27.417	41.309	27.789	158.01	115.86	104.63	31.786	47.764	41.905	73.664	95.931	141.30	68.616	29.417	34.695	114.58	69.881
real	36.525	31.201	39.854	39.977	122.34	182.09	80.045	39.895	47.231	44.105	65.553	80.859	107.41	58.740	30.249	33.813	77.101	73.069

D ía

prevdef real