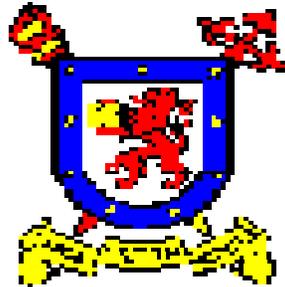


UNIVERSIDAD DE SANTIAGO DE CHILE. Facultad de Ingeniería.



**DESARROLLO DE UN PREDICTOR DE MP10
BASADO EN LA UTILIZACIÓN DE REDES NEURONALES**

Trabajo final de curso presentado por Jorge
Antonio Reyes Molina para la obtención del Diploma
del Curso de Postítulo en Consultoría Medioambiental.

Profesor-tutor: Carolyn Palma Tolosa

28/08/2000; Macul, Provincia de Santiago.

1. Introducción

2. Objetivos

3. Marco de Referencia Teórico

4. CARACTERIZACION DE LOS REGISTROS

Los registros corresponden a los valores de MP10, hora por hora, medidos en la Estación A de la red de monitoreo MACAM (Plaza Gotuzzo) Esta estación se ubica a unos 100 m al NW del Palacio de La Moneda y a 150 m de un semáforo donde se observa la mayor congestión de Santiago. El periodo abarcado corresponde a aquel donde los días críticos incrementan su intensidad y frecuencia: desde el 01/05 al 30/09 ("Periodo Invernal") Esto equivale a 3672 datos por serie (1994, 1995 y 1996)

La estadística de las series se muestra en el siguiente apartado.

4.1. Estadística Clásica de la Serie de MP10

Año 1994

$\langle x \rangle = 134.5$

Moda = 78.1

Valor mín = 1

Valor máx = 723

$\sigma = 85.1$

Por razones algorítmicas, se trabajó con los primeros 148 días de cada serie. Para el año 1994 se encontró lo siguiente:

NIVEL MAXIMO	NUMERO DE DIAS
0	110
1	21
2	16

3	1
---	---

Tabla 4.1. Días v/s Niveles máximos para el año 1994.

Año 1995

Esta serie sobresale de las otras debido a que sus parámetros estadísticos son los mínimos: menor promedio, menor máximo, menor moda y menor dispersión.

$\langle x \rangle = 116.6$

Moda = 54.7

Valor mín = 1

Valor máx = 466

$\sigma = 69.5$

NIVEL MAXIMO	NUMERO DE DIAS
0	126
1	18
2	4
3	0

Tabla 4.2. Días v/s Niveles máximos para el año 1995.

Año 1996

$\langle x \rangle = 118.4$

Moda = 109.4

Valor mín = 1

Valor máx = 613

$\sigma = 74.4$

NIVEL MAXIMO	NUMERO DE DIAS
0	124
1	20
2	4
3	0

Tabla 4.3. Días v/s Niveles máximos para el año 1996.

La estadística de las series indica que abunda el nivel cero (80%) y escasea el nivel 3 (< 1%) Por otro lado, es importante señalar que la dinámica de MP10 no corresponde a la de un sistema aislado, debido a que las autoridades reaccionan frente a la contaminación con medidas que tendrían que amortiguar los incrementos de esta, interfiriendo así con su evolución natural. Esto podría, eventualmente, provocar contradicciones en cualquier predictor que sea entrenado sin considerar como input la intervención humana. Por lo tanto, hay que hacer otra hipótesis para que sea lícito entrenar a la red con los registros íntegros sin tener que filtrar los periodos en que el sistema deja de ser aislado (sería nocivo filtrarlos, ya que justamente son esos los periodos que deseamos que la red sea capaz de predecir) La nueva hipótesis se deduce de la anterior exposición: tenemos que aceptar, al menos provisoriamente, que el sistema sigue siendo aislado a pesar de las medidas anticontaminantes dictadas por la autoridad. Afirmar lo contrario nos obligaría a filtrar los días críticos, sobreviviendo únicamente el nivel cero.

4.2. Búsqueda de Orden en el Desorden

Después de efectuar diversas pruebas asociadas a la Teoría del Caos, se llegó a resultados interesantes:

AÑO	S	LZ	EL
1994	0.421	0.693	0.466 ± 0.020
1995	0.490	0.516	0.431 ± 0.021
1996	0.596	0.616	0.481 ± 0.020

Tabla 4.4. Caracterización no lineal de las series.

- S: Entropía.

S es una medida del desorden propio de las series, cumpliéndose lo siguiente:

S < 0: Presencia de orden.

S > 0: Presencia de desorden.

La entropía de las series resultó ser mayor que cero, de acuerdo con lo esperado.

Obsérvese además, el siguiente detalle: la entropía informática de las series muestra una clara tendencia al alza. En caso de corresponder a una tendencia global, significaría que cada vez será más difícil predecir la contaminación. Esto puede estar relacionado con la entropía física que – en un sistema aislado – crece con el tiempo, aunque la conexión no es tan directa como nos gustaría que lo fuera.

- LZ: Complejidad relativa de Lempel Ziv.

LZ mide la complejidad algorítmica de la serie. Los extremos son:

0: Perfectamente predecible.

1: Totalmente azaroso.

LZ de las series resultó estar inclinada hacia lo azaroso (valores del orden de 0.6)

- EL: Mayor Exponente de Lyapunov en base e.
EL mide la sensibilidad de la serie respecto de las condiciones iniciales (grado de caos)

Las posibilidades son:

$EL \leq 0$: serie periódica

$EL > 0$: caos

EL infinito: datos random, totalmente descorrelacionados.

EL resultó ser mayor que cero para las tres series, lo que indica que la dinámica de MP10 depende fuertemente de las condiciones iniciales (usual en meteorología) Los valores resultaron ser del orden de 0.45, lo cual es abordable con las actuales herramientas matemáticas existentes.

Por último, es importante señalar la posibilidad de que el año 1995 haya sido un año con una dinámica anómala, debido a que LZ y EL experimentan en ese periodo una baja notoria. Más adelante se observará que la función de autocorrelación confirma nuestro supuesto.

- La Figura 4.1 muestra las primeras 367 horas para las series originales (MP10 hora por hora), mientras que la Figura 4.2 muestra lo mismo para las series suavizadas (C24 hora por hora) Se observa que C24 suaviza notoriamente a las series originales, lo cual mejorará la capacidad de predicción. En términos técnicos, lo que hace C24 es quitar el “ruido” de MP10 y así el predictor no invierte recursos en aprender desviaciones “anómalas” (como un peak de contaminación) He aquí una crítica al predictor por el hecho de utilizar C24 en lugar de los valores originales de MP10: El suavizado hace extremadamente improbable alcanzar el nivel 3, de modo que la validación cuantifica básicamente la capacidad de predecir el nivel cero, es decir, el nivel menos importante en términos de salud pública. Por lo visto, hace falta un tercer parámetro, aparte de K y PSP, para cuantificar la calidad de la predicción.

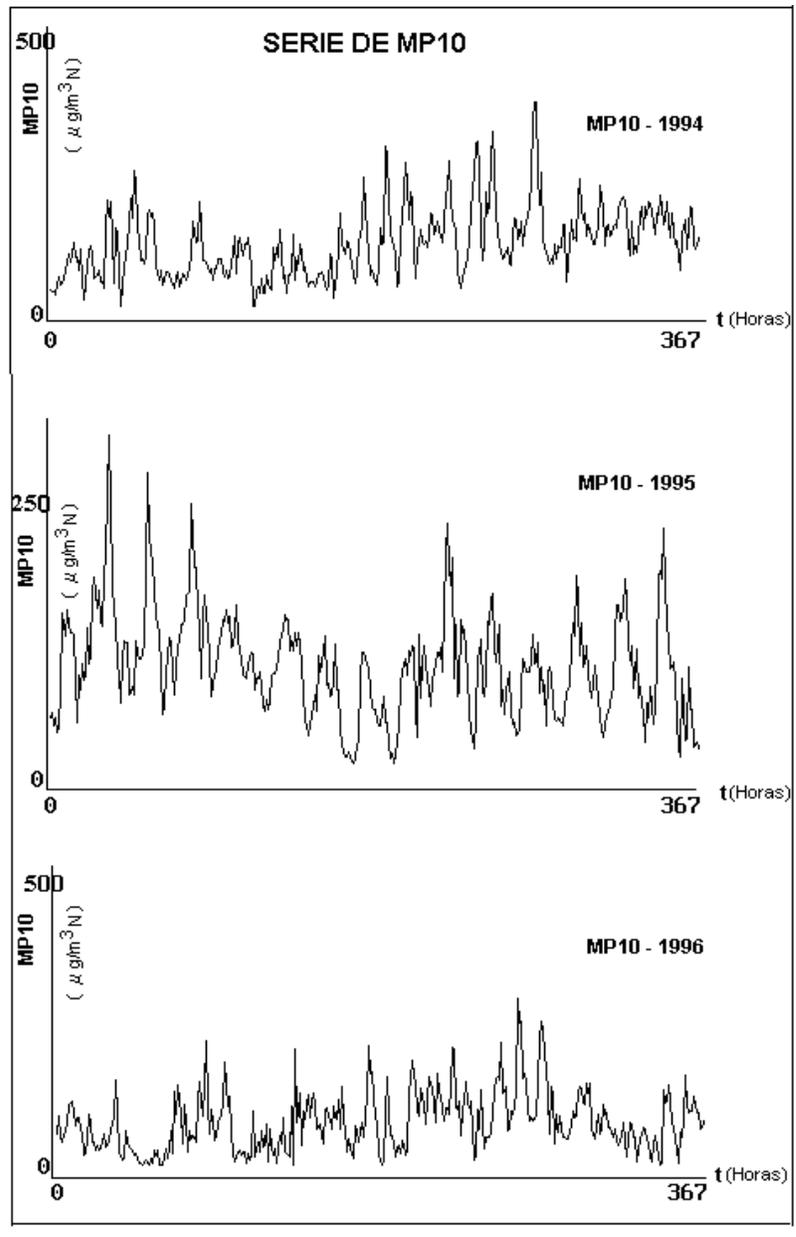


Figura 4.1.
MP10 v/s t.

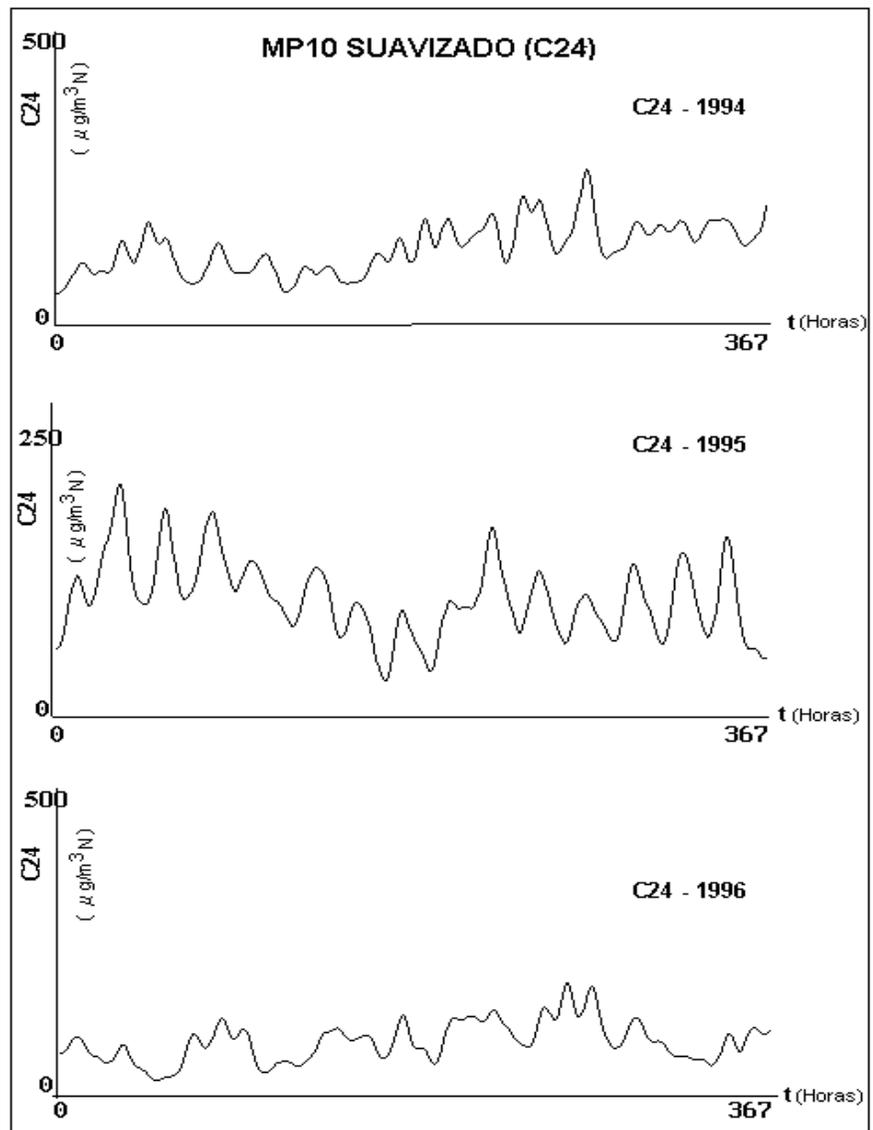


Figura 4.2.

C24 v/s t.

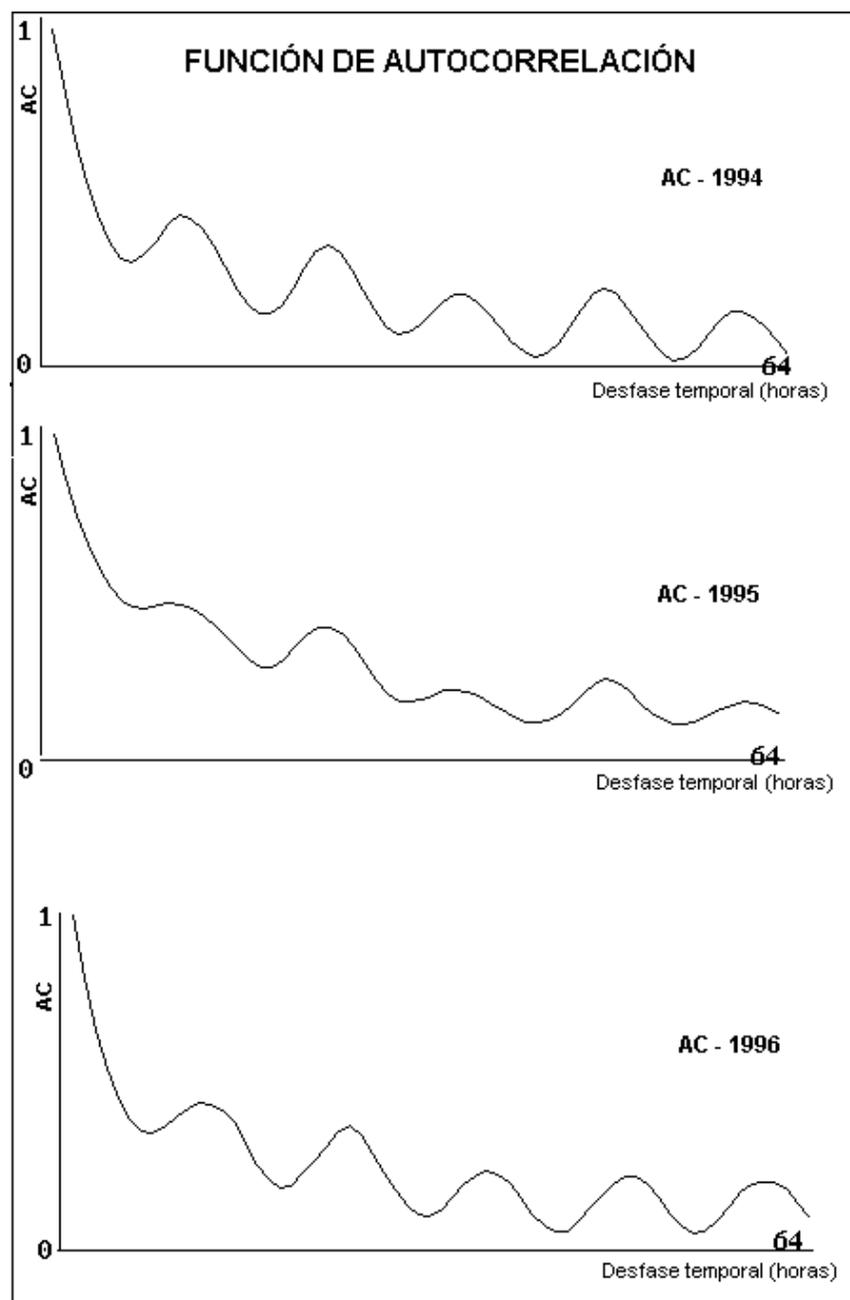


Figura 4.3.

AC v/s t.

- Las curvas de MP10 y C24 aparentan ser muy caóticas y seguramente en la dinámica de la contaminación deben prevalecer interacciones no lineales. Sin embargo, la función de autocorrelación – que indica el promedio de la cantidad de información binaria transmitida por dos puntos de la serie separados por un intervalo “t” – demuestra que bajo el desorden (*aparente*) hay orden (Fig. 4.3) Esta curva es suave, continua y cíclica (hecho realmente sorprendente y favorable) Se observan peaks sistemáticos en $t = 12, 24, 36, 48, \text{etc.}$, lo que indica que la dinámica de MP10 se repite cada 12 horas (al menos en términos cualitativos) Llama la atención la función de autocorrelación del año 1995, ya que el patrón tiende a perderse: la correlación en múltiplos impares de 12 es notoriamente menor que para los múltiplos pares, no ocurriendo lo mismo con los otros años. Este hecho confirma la opinión de que el año 1995 se rigió por una dinámica anómala.
- El porqué de la forma de la función de autocorrelación puede entenderse si se observa con detalle los registros de MP10. Al hacerlo, se constatará que los peaks diarios se dan a las 10:00 AM y a las 10:00 PM, aproximadamente. El peak de la mañana tiene que asociarse a la alta congestión vehicular que se da alrededor de las 8:00 AM, más el tiempo necesario para que el sistema tome constancia del hecho (“inercia”) Lo mismo puede decirse respecto del peak de congestión de las 8:00 PM.

5. DESARROLLO DEL PREDICTOR

Una de las premisas que ha orientado este trabajo consiste en asumir que una Red Neuronal Artificial es capaz de predecir los valores de C24 con un buen grado de confiabilidad y anticipación. Ahora tenemos que responder diversas preguntas para ir definiendo el perfil de nuestra red. Primeramente necesitamos caracterizar el set de entrenamiento y el de test, hecho que nos permitirá definir la arquitectura fundamental del predictor.

5.1. Vectores de Entrenamiento y de Test.

Según el Decreto 59 / 98, el predictor debe cumplir las siguientes condiciones:

- El periodo de entrenamiento (“periodo de generación de la información”) debe ser previo al de test (“periodo de validación”)
- El set de entrenamiento debe ser representativo de un periodo equivalente, al menos, al mismo conjunto de días del año para el que fue concebido el método de pronóstico. Como los registros corresponden al periodo “Invierno”, se concluye que el predictor ha sido concebido – al menos – para ser aplicado en invierno.

Las dos anteriores exigencias se cumplen entrenando con el periodo invierno del año 1994 y validando con el periodo invierno de los años 1995 y 1996.

- Las siguientes variables son válidas como input:
 - Emisiones de MP respirable y precursores.
 - Condiciones meteorológicas.
 - Ciclos de emisiones observados.
 - Procesos de acumulación y remoción de contaminantes.
 - Condiciones topográficas.
- La calidad de la predicción se cuantificará con el parámetro “Confiabilidad” y como mínimo tendrá que ser igual a 65%. Además se utilizará el parámetro “PSP” (Porcentaje de Semejanza Promedio)
- La anticipación implica ser capaz de predecir para el día siguiente o un periodo superior.

Estas dos últimas exigencias pueden cumplirse parcialmente prediciendo C24 máximo para el día $n+1$ a las 12:00 del día n . Obviamente la cota inferior de la confiabilidad dependerá del comportamiento de la red.

Resumiendo, las características fundamentales del predictor serán las siguientes:

- El input corresponderá a un ciclo completo de la dinámica de C24: desde las 13:00 del día $n-1$ a las 12:00 del día n . Cabe señalar que en mayo apareció una nueva reglamentación que permite hacer la predicción a las 19:00, por lo que la confiabilidad mejorará.
- El output corresponderá al valor máximo de C24 del día $n+1$ (desde la 1:00 a las 24:00)
- Vectores de entrenamiento: 25 valores de C24 del periodo "Invierno" (01/05 a 30/09) del año 1994, desglosados en 24 como input y uno como output.
- Vectores de test: análogo al anterior, con la salvedad de que corresponden a los años 1995 y 1996.
- Cardinalidad: por razones algorítmicas, se tendrá 148 vectores para cada año.

5.2. Predictores Posibles.

Como predictor se utilizará una Red Neuronal supervisada que aprenderá a predecir mediante Backpropagation. Las posibilidades a examinar serán las siguientes:

- Red tipo Perceptrón.

Arquitectura: 24 neuronas de entrada y una de salida.

Subtipos:

- PL: Perceptrón con función de transferencia lineal.
 - PNL: Perceptrón con función de transferencia no lineal ($\tanh(x)$)
- Red tipo Multicapas.

Arquitectura: 24 neuronas de entrada, 6 en la capa intermedia y una de salida. Se ha utilizado subjetivamente seis neuronas en la capa intermedia porque encuentro razonable comprimir la información de entrada en un factor cuatro.

Subtipos:

- ML: Multicapas con función de transferencia lineal.
- MNL: Multicapas con función de transferencia no lineal.
- Red tipo Multicapas Kolmogorov.

Arquitectura: 24 neuronas de entrada, 49 en la capa intermedia y una de salida. Según Kolmogorov (1957), dada una función n-dimensional “de buen comportamiento”, ésta podrá ser implementada por una red neuronal con la siguiente arquitectura: n neuronas en la capa de entrada (24), $2n+1$ en la capa intermedia (49) y una neurona en la capa de salida.

Subtipos:

- MKL: Multicapas Kolmogorov con función de transferencia lineal.
- MKNL: Multicapas Kolmogorov con función de transferencia no lineal.

Resumiendo, se desarrollará un total de seis redes neuronales y nuestro predictor corresponderá a la red de mayor confiabilidad observada con el set de validación.

El Decreto 59/98 exige especificar las siguientes características del predictor:

- Periodo de uso del pronóstico en el año calendario: desde el 01/05 al 30/09, al menos.
- Zona geográfica de aplicación: Estación A de la Red MACAM.
- Requerimientos para la operación del pronóstico: se debe medir MP10 hora por hora y con esta serie se debe calcular C24. La red se alimenta de los valores de C24 y no de MP10.
- Hora de comunicación del pronóstico: el pronóstico se comunicará al mediodía, lo que implica – en la práctica – que la última medición de MP10 (en $t = 12:00$) fue extrapolada, por ejemplo, a las 11:45 AM (no estoy considerando el retraso del instrumental de monitoreo) Cabe señalar que la red demora aproximadamente un minuto en hacer la predicción.
- Capacidad predictiva del pronóstico: la red predice a las 12:00 del día n el valor máximo de C24 para el día n+1.

- Estimación y caracterización del error en la metodología: el error de la “metodología de pronóstico” (la red neuronal) se definirá “a posteriori”, según la calidad de la predicción que se observe con el set de validación (años 95 y 96)

6. RESULTADOS

Los resultados se indican a continuación:

Perceptrón Lineal

Validación con el año 1995:

$$K = 83.1\%$$

$$PSP = 116.5 \pm 35.44\%$$

Validación con el año 1996:

$$K = 81.8\%$$

$$PSP = 114.2 \pm 33.4\%$$

Perceptrón No Lineal

Validación con el año 1995:

$$K = 83.8\%$$

$$PSP = 116.9 \pm 35.7\%$$

Validación con el año 1996:

$$K = 81.1\%$$

$$PSP = 114.7 \pm 33.6\%$$

Multicapas Lineal

Validación con el año 1995:

$$K = 85.8\%$$

$$PSP = 117.5 \pm 35.3\%$$

Validación con el año 1996:

$$K = 79.7\%$$

$$PSP = 115.1 \pm 33.2\%$$

Multicapas No Lineal

Validación con el año 1995:

$$K = 84.5$$

$$PSP = 118.1 \pm 35.3\%$$

Validación con el año 1996:

$K = 79.1\%$

$PSP = 116.0 \pm 33.7\%$

Multicapas Kolmogorov Lineal

Validación con el año 1995:

$K = 85.8\%$

$PSP = 117.3 \pm 35.2\%$

Validación con el año 1996:

$K = 79.7\%$

$PSP = 114.8 \pm 33.1\%$

Multicapas Kolmogorov No Lineal

Validación con el año 1995:

$K = 83.8\%$

$PSP = 117.6 \pm 35.5\%$

Validación con el año 1996:

$K = 79.1\%$

$PSP = 115.3 \pm 33.2\%$

7. COMENTARIOS Y CONCLUSIONES

El comportamiento de las redes resultó ser muy bueno. Lo ideal es que K y PSP sean cercanos al 100% y eso es lo que ocurrió, aunque no deja de llamar la atención la desviación estándar de PSP del orden del 30% (en todo caso es un valor aceptable) Lamentablemente los resultados son muy similares entre sí: K es del orden del 80% y PSP del 115%. Esto significa que estamos trabajando en el límite de la capacidad de las redes neuronales. Equivalentemente, se puede generalizar y afirmar que frente a este problema ninguna red es mejor que otra. A pesar de lo anterior, y si observamos los leves porcentajes de diferencia, es posible asegurar lo siguiente:

- El año 1995 es mejor representado por las redes ML y MKL, en términos de K, y por la PL, en términos de PSP.
- El año 1996 es mejor representado por la red PL en términos de K y de PSP.

Nuestro predictor está totalmente definido:

- Perceptrón lineal.
- 24 neuronas de entrada y una de salida.
- K observado (años 95 y 96): $83 \pm 1\%$.
- PSP observado: $115 \pm 35\%$

Además, nuestras hipótesis han sido provisoriamente confirmadas:

- Una red neuronal es capaz de predecir la contaminación con un buen nivel de confiabilidad (83%) y anticipación ("un" día) Cabe recordar que la confiabilidad del primer predictor de la CONAMA era del 59%, mientras que la del segundo (elaborado por el experto estadounidense Joseph Cassmasi) es del 78%.
- Una red neuronal es capaz de deducir el estado meteorológico aproximado a partir de la serie de C24. De no ser así, los valores de K y PSP habrían sido inaceptables, lo que nos estaría indicando que los parámetros de entrada eran insuficientes para modelar el fenómeno.
- Las medidas anticontaminantes influyen tan poco en el sistema que puede asumirse que el MP continúa su evolución natural a pesar de los esfuerzos humanos por detener su incremento. Muchas veces ha ocurrido que después de un par de días críticos seguidos, llegaba la lluvia y sólo entonces se conseguía un descenso al nivel 0.

Conclusión: Lo simple fue lo mejor. Podemos afirmar que mientras no se incluya explícitamente variables distintas al MP10, la dinámica quedará bien representada por un perceptrón lineal. El resultado no deja de sorprender: los efectos no lineales en la serie de MP10 son despreciables.

8. GLOSARIO

Complejidad Relativa de Lempel Ziv.

Medida de la complejidad algorítmica de la serie. Su valor se obtiene mediante el algoritmo de Kaspar y Schuster.

Entropía Informática.

Medida del grado de desorden de la serie. Equivale a la suma de los exponentes de Lyapunov positivos en base e y se obtiene mediante el algoritmo de Grassberger y Procaccia.

Función de Autocorrelación, "AC" (normalizada a uno)

Función que indica el grado de dependencia de los datos de la serie con sus vecinos. Se obtiene multiplicando $x(n)$ con $x(n-t)$ y sumando el resultado sobre todos los datos. Posteriormente se normaliza de modo que $AC(0) = 1$ (dependencia máxima)

Mayor Exponente de Lyapunov en Base e.

Estimación de la divergencia experimentada por trayectorias similares en el espacio de fase. Se obtiene mediante el algoritmo de Wolf.

Perceptrón.

Red neuronal de dos capas (entrada y salida) que modifica su set de sinapsis mediante aprendizaje supervisado (vectores de entrenamiento formado por estímulos y respuestas esperadas)

Red Neuronal.

Simulación informática del cerebro apoyada en tres pilares: neuronas o unidades de procesamiento, sinapsis o intensidad de las conexiones neuronales y un método de aprendizaje.

Retropropagación.

Método de aprendizaje supervisado donde la red neuronal modifica iterativamente su set de sinapsis con el fin de optimizar su capacidad de reproducir el set de entrenamiento. En este caso se busca minimizar la función Error (energía cinética relativa)

9. FUENTES DE INFORMACION

9.1 Bibliografía.

- Abarbanel, H.; Brown, R.; Sidorowich, J. Y L. Tsimring, 1993. *The analysis of observed chaotic data in physical systems*. Rev. Mod. Phys. 65: 1331-1392.
- Boznar, M.; Lesjak, M. Y P. Mlakar, 1993. *A neural network based method for short term predictions of ambient SO₂ concentrations in highly polluted industrial areas of complex terrain*. Atmospheric Environment 27 B: 221-230.
- Gardner, M. Y S. Dorling, 1998. *Artificial neural networks (the multilayer perceptron)- A review of applications in the atmospheric sciences*. Atmospheric Environment 31: 4103-4117.
- Kulkarni, D.; Parikh, J. Y R. Pratap, 1997. *Simulation of characteristics and artificial neural network modeling of electroencephalograph time series*. Physics Review E 55: 4508-4511.
- Pérez, P.; Trier, A. Y J. Reyes, 2000. *Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile*. Atmospheric Environment 34: 1189-1196.
- Prendez, M.; Egido, C.; Tomas, C.; Seco, J.; Calvo, A. Y H. Romero, 1995. *Correlation between solar radiation and total suspended particulate matter in Santiago, Chile – Preliminary results*. Atmospheric Environment 29: 1543-1551.
- Reyes, J. Y P. Perez, 1998. *Predictibilidad del material particulado PM_{2.5} en Santiago de Chile utilizando técnicas de modelación de sistemas dinámicos y redes neuronales*. Contribuciones Científicas y Tecnológicas, Area Ciencias Básicas (USACH) Año XXVI, n° 121: 107-108.

9.2. Otras Fuentes de Información.

- CONAMA. Página Web: www.conama.cl.
- Sanguinetti, Roxana. Entrevista del 10/03/2000 con Roxana Sanguinetti, Meteoróloga de la CONAMA.
- Molina, Roberto. Decreto 59/98, enviado vía E-mail por Roberto Molina, profesional de la CONAMA (diciembre de 1999)
- Trier, Alex. Base de datos de MP10 proporcionada por Alex Trier, profesor del Dpto. de Física de la Universidad de Santiago.