BIOMATEMÁTICA ***XIII*** *(2003) 113–127*
Uma Publicação do Grupo de Biomatemática IMECC – UNICAMP

Modelamiento de la contaminación atmosférica por particulas: Comparación de cuatro procedimientos predictivos en Santiago, Chile

C. Silva[1], S. Alvarado , R. Montaño ,
Universidad de Chile, Facultad de Medicina, Escuela de Salud Publica.
P. Pérez ,
Universidad de Santiago de Chile, Facultad de Ciencia,
Departamento de Física.

Resumen. La contaminación atmosférica por particulas en suspensión es un problema mayor para la Salud Pública de muchas ciudades: Santiago de Chile es una de ellas. Modelar el nivel de esta contaminación se ha convertido en un problema de interés político-administrativo ya que la predicción de episodios críticos debe guiar la toma de decisiones muy importantes, especialmente en el período invernal. La necesidad de aplicar medidas paliativas frente a tales episodios requiere predecir adecuadamente (en precisión y anticipación) algunos indicadores como PM10 o PM2,5. En este trabajo se comparan resultados obtenidos utilizando análisis discriminante no-parametrico, redes neuronales, regresión lineal múltiple y modelos MARS "multivariate adaptive regression splines", esta última metodología compite exitosamente frente a las metodologías estadísticas antes mencionadas.

Palabras claves*: MARS, PM10, PM2.5, Predicción, Calidad del aire.*

1. Introducción

La ciudad de Santiago de Chile (33.5^o S, 70.8^o W) se encuentra localizada en el valle central de la Región Metropolitana, limitada por cordones montañosos transversales y la cordillera de los Andes. El centro de la ciudad tiene una elevación de 520 m. El área de la Región Metropolitana de Santiago excede los 15.000 km^2, con una población aproximada de 5,3 millones de habitantes (Silva et al., 2001). Por la combinación de factores meteorológicos y topográficos Santiago presenta una mala ventilación atmosférica en invierno, que determina la acumulación de partículas y gases; por otra parte en verano el incremento de la radiación solar favorece las reacciones fotoquimicas que dan origen al ozono (Sanhueza et al., 1998).

[1]csilva@machi.med.uchile.cl

La medición automatica y continua de material particulado comenzo en Santiago en el año 1987, cuando se instaló una Red Automatica denominada MACAM1-RM, compuesta por cinco estaciones y dedicada principalmente a evaluar la calidad del aire del centro de Santiago (4 estaciones) contando con una quinta estación de tipo móvil. Desde el año 1997 la red de monitoreo MACAM2-RM del Servicio de Salud del Ambiente de la Región Metropolitana (SESMA) está midiendo constantemente la concentración de los contaminantes más típicos, incluyendo al PM10. Los métodos estadísticos utilizados para modelar la contaminación ambiental en Santiago causada por partículas en suspensión en la atmósfera aprovechando información referente a variables asociadas como visibilidad, temperatura, humedad relativa y/o velocidad del viento, han incluído: series cronológicas con función de intervención (Trier e Firinguetti, 1994; Silva et al., 1994), análisis discriminante no-paramétrico (Silva y Trier, 1995), redes neuronales (Pérez et al., 1998) y modelos MARS (Silva et al., 1998, 2001; Alvarado et al., 2002, 2003). En este trabajo se reportan los resultados obtenidos utilizando análsis discriminante o-parametrico, redes neuronales, regresión lineal múltiple y modelos MARS "multivariate adaptive regression splines", esta ultima tecnica compite exitosamente frente a las metodologías estadísticas antes mencionadas.

2. Analisis Discriminante

En instalaciones de la red MACAM1-RM (Estación Gotuzzo) se ha estado midiendo regularmente desde hace algunos años el nivel de contaminación por particulas inhalables (diámetro inferior a 10μm); actualmente se ha incorporado instrumental que permite el análisis de partículas de diámetro inferior a 2.5μm. Para este estudio se han utilizado mediciones realizadas cada doce horas. El objetivo es intentar describir, con adecuada precisión y anticipación, el comportamiento de esta variable (episodios críticos en los niveles de contaminación) en función de su comportamiento previo y de un conjunto de variables meteorológicas: velocidad del viento, humedad relativa, temperatura y dirección del viento.

2.1. Clasificación parametrica

Se disponía de $n = 210$ observaciones compuestas de valores de $\mathrm{x}_1 =$ temperatura, $\mathrm{x}_2 =$ humedad relativa, $\mathrm{x}_3 =$ velocidad del viento y $\mathrm{x}_4 =$ nivel de concentración de material particulado de diámetro aerodinámico menor o igual a 2.5μm espaciadas cada 12 horas desde el 19 de Julio al 30 de Octubre de 1993.

Se definió una variable auxiliar "Y" que correspondéria a las condiciones de calidad de aire "normal", "alerta" y "emergencia".

$$\mathrm{Y} = \begin{cases} 0, & si \quad x_4 \leq 50 \\ 1, & si \quad 50 < x_4 \leq 150 \\ 2, & si \quad x_4 > 150 \end{cases}$$

Podemos aprovechar de procesar la información multivariante usando algún algoritmo de clasificación asociado al análisis discriminante clásico. Para tal efecto combinamos k de

estos vectores ($k = 1, 2, 3$ o 4) generando $n - k$ vectores $\boldsymbol{x}$ de dimensión 4k a cada uno de los cuales asociamos el valor de Y correspondiente a q períodos ($q = 1$, 2, 3) hacia adelante.

Por ejemplo, $k = 4$ significa información sobre 36 horas hacia el pasado y $q = 3$ implica clasificación (en términos de Y) 36 horas hacia el futuro.

Los resultados se pueden presentar en tablas de clasificación cruzada que resumen la clasificación real de cada caso y la asignada por el procedimiento en cuestión. Básicamente, una observación $\boldsymbol{x}$ es clasificada en al categoría h si $p(h|\boldsymbol{x})$ es el máximo de $p(j|\boldsymbol{x})$ para $j = 1, 2, 3$ donde

$$p(j|\boldsymbol{x}) = \frac{exp(-\frac{1}{2}d_j^2(\boldsymbol{x})\,)}{\sum_{i=1}^{3} exp(-\frac{1}{2}d_i^2(\boldsymbol{x})\,)};$$

para

$$d_i^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x_i})^{'}\boldsymbol{S_i^{-1}}(\boldsymbol{x} - \boldsymbol{x_i}) + ln|\boldsymbol{S_i}|$$

Es decir, $d_i^2(\boldsymbol{x})$ es la distancia generalizada de Mahalanobis de la observación $\boldsymbol{x}$ al centroide de la muestra i-ésima. (Bajo el supuesto de normalidad multivariante, $p(j|\boldsymbol{x})$ correspondería a la probabilidad a posteriori de que $\boldsymbol{x}$ perteneciera a la j-ésima población).

En base a dos períodos y prediciendo cuatro pasos adelante (48 horas) tenemos en la Tabla N^o 1:

Población	Clasificada como			
de origen	0	1	2	TOTAL
0	69 (74.2)	18 (19.4)	6 (6.4)	93 (100.0)
1	26 (28.6)	54 (59.3)	11 (12.1)	91 (100.0)
2	0 (00.0)	0 (00.0)	21 (100.0)	21 (100.0)
TOTAL	95 (46.3)	72 (35.1)	38 (18.5)	205 (100.0)

Tabla N^o 1. Predicción dos períodos y cuatro pasos para PM2.5. Con una tasa global de error de 22.16 % (25.8 % para la clase 0, 40.7 % para la clase 1 y 0.0 % para la clase 2). Nótese que ninguna emergencia "escapa" al proceso de clasificación mientras sólo 6 de 38 alarmas serían erradas.

En base a tres períodos y prediciendo cuatro pasos adelante (48 horas) tenemos en la Tabla N^o 2:

Población	Clasificada como			
de origen	0	1	2	TOTAL
0	73 (78.5)	19 (20.4)	1 (1.1)	93 (100.0)
1	22 (24.2)	63 (69.2)	6 (6.6)	91 (100.0)
2	0 (00.0)	0 (00.0)	20 (100.0)	20 (100.0)
TOTAL	95 (46.6)	82 (40.2)	27 (13.2)	204 (100.0)

Tabla N^o 2. Predicción tres períodos y cuatro pasos para PM2.5. Con una tasa global de error de 20.9 % (21.5 % para la clase 0, 30.8 % para la clase 1 y 0.0 % para la clase 2). Nótese que ninguna emergencia "escapa" al proceso de clasificación mientras sólo 1 de 27 alarmas serían erradas.

Ya que importantes decisiones administrativas relacionadas con el nivel de contaminación se toman usando la información correspondiente a material particulado PM10, resulta relevante re-estudiar la clasificación usando la tricotomía definida por la variable auxiliar Y_{10}

$$y_{10} = \begin{cases} 0 & si \quad x_4 \leq 150; \\ 1 & si \quad 150 < x_4 \leq 300 \text{ y} \\ 2 & si \quad x_4 > 300 \end{cases}$$

que correspondería a las condiciones "normal", "alerta" y "emergencia".

En base a tres períodos y prediciendo cuatro pasos adelante (48 horas) tenemos en la Tabla N^o 3:

Población	Clasificada como			
de origen	0	1	2	TOTAL
0	80 (81.6)	18 (18.4)	0 (0.0)	98 (100.0)
1	24 (24.2)	75 (79.8)	0 (0.0)	99 (100.0)
2	0 (00.0)	0 (00.0)	7 (100.0)	7 (100.0)
TOTAL	104 (51.0)	93 (45.6)	7 (3.4)	204 (100.0)

Tabla Nº 3. Predicción tres períodos y cuatro pasos para PM10. Con una tasa global de error de 14.2 % (18.4 % para la clase 0, 24.2 % para la clase 1 y 0.0 % para la clase 2).

2.2. Clasificacion No-Parametrica

Sin asumir un supuesto de Normalidad multivariante es posible replantear el problema de clasificación. Para ello se recurre a algún procedimiento de análisis discriminante basado en estimación de funciones de densidad de probabilidad usando funciones kernel (Hang, 1982; Scott, 1992).

La clasificación de una observación $\boldsymbol{x}$ se basa en las probabilidades a posteriori de pertenencia a cada grupo, calculadas a partir de las densidades específicas del grupo estimadas usando el conjunto preliminar.

El método kernel usa un radio fijo r y una función kernel K_t para estimar la densidad t-ésima en cada observación $\boldsymbol{x}$. Por ejemplo, la propuesta de Epanechnikov es:

$$K_t(\mathrm{z}) = \begin{cases} c(t)\left(1 - \frac{\boldsymbol{z}' \boldsymbol{V_t^{-1}} \boldsymbol{z}}{r^2}\right), & \text{si } \boldsymbol{z}' \boldsymbol{V_t^{-1}} \boldsymbol{z} \leq r^2 \\ 0, & \text{en otro caso.} \end{cases}$$

siendo $c(t) = (1 + \frac{p}{2})/v_t(r)$ en donde tenemos que:

$v_t(r) = r^p \left|V_t\right|^{\frac{1}{2}} v_0 = r^p \left|V_t\right|^{1/2} \frac{\pi^{1/2}}{\Gamma(\frac{p}{2}+1)}$ es el volumen del elipsoide p-dimensional $\left\{\boldsymbol{z} \,\middle|\, \boldsymbol{z}' \boldsymbol{V_t^{-1}} \boldsymbol{z} \leq r^2\right\}$ si $\boldsymbol{V}_t$ es la matriz de varianzas-covarianzas interna del $t-$ésimo grupo.

La densidad de probabilidad para una observación $\boldsymbol{x}$ en ese grupo es estimada por $f_t(\boldsymbol{x}) = \frac{1}{n_t} \sum_{y \in G_t} K_t(\boldsymbol{x} - \boldsymbol{y})$ y la probabilidad posteriori de pertenencia al grupo t estará dada por $p(t \,|\, \boldsymbol{x}) = \frac{q_t \, f_t(\boldsymbol{x})}{\sum_h \; q_h \, f_h(\boldsymbol{x})}$ siendo q_h la probabilidad a priori de pertenencia al grupo h.

En nuestra situación de estudio para material particulado de PM2.5 y en base a dos períodos y prediciendo cuatro pasos adelante (48 horas) se tiene en la Tabla Nº 4:

Población	Clasificada como			
de origen	**0**	**1**	**2**	TOTAL
0	93 (100.0)	0 (0.0)	0 (0.0)	93 (100.0)
1	0 (00.0)	91 (100.0)	0 (00.0)	91 (100.0)
2	0 (00.0)	0 (00.0)	20 (100.0)	20 (100.0)
TOTAL	93 (100.0)	91 (100.0)	20 (100.0)	204 (100.0)

Tabla Nº 4. Predicción dos periodos y cuatro pasos de PM2.5 con una clasificación óptima.

En las mismas condiciones y en referencia a material particulado PM10 obtuvimos aplicando el procedimiento no-paramétrico recién descrito, se tiene en la Tabla Nº 5:

Población	Clasificada como			
de origen	**0**	1	**2**	TOTAL
0	98 (100.0)	0 (00.0)	0 (00.0)	98 (100.0)
1	0 (00.0)	99 (100.0)	0 (00.0)	99 (100.0)
2	0 (00.0)	0 (00.0)	7 (100.0)	7 (100.0)
TOTAL	98 (100.0)	99 (100.0)	7 (100.0)	204 (100.0)

Tabla Nº 5. Predicción dos períodos y cuatro pasos para PM10. con una clasificación igualmente óptima.

3. Modelos MARS

Friedman (1991), propone una novedosa metodología llamada MARS, esta metodología intenta construir un modelo de Regresión no-lineal, que esté basado en un producto de funciones base spline. MARS es una generalización de la Recursive Partitioning Regression (PR), la que divide al espacio de las variables predictoras en diferentes subregiones, intentando una aproximación local en cada subregion (Lewis y Stevens, 1991; Friedman y Roosen, 1995).

MARS produce un modelo para la respuesta en estudio, que automáticamente selecciona las variables que aparecen en la ecuación final. También indica lo complejo de la relación entre la respuesta y cada variable predictora, a partir del numero de funciones bases usadas (Silva et al., 2001).

3.1. Modelo para un predictor

Para una variable respuesta y, la variable predictora x y el error aleatorio ϵ se asume un modelo del tipo:
y = f(x) + ϵ

Seleccionando K "nodos" $t_k, \quad k = 1, \cdots, K$ se definen $K+1$ "regiones" sobre el rango de x. Se asocia a cada nodo la función spline lineal, generando una familia de funciones bases:

$$B_k^{(q)}(x) = \begin{cases} x^j & j = 0, \cdots, q \\ (X - t_k)_+^q & k = 1, \cdots, K \end{cases}$$

Para la aproximación de orden q, se estima la función $jf_q(x) = \sum_{k=0}^{K+q} a_k B_k^{(q)}(x)$ generalmente el orden de la función spline que se tome debe ser menor o igual a tres, para que la función y sus q-1 derivadas sean continuas en jf_q, jf'_q y jf''_q.

Esta restricción y el uso de polinomios en cada subregion produce funciones suavizadas y ajustadas.Ya que cada polinomio tiene grado q, tenemos q+1 parámetros que deben ser ajustados usualmente por mínimos cuadrados.

Para evaluar este modelo, Friedman (1991) propone usar la estadística Generalized Cross Validation $GCV = A \times \sum_i \left(y_i - jf_q(x_i)\right)^2 \Big/ N$, en donde $A = \left(1 - C(M)/N\right)^{-2}$ y $C(M) = 1 + traza\left(\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\right)$, el numerador es la falta de ajuste sobre los datos de entrenamiento y el denominador es un termino penalizado que refleja la complejidad del modelo

3.2. Extensión a p predictores

Para $x = x_1, x_2, \ldots, x_p$ la función spline es definida análogamente que para el caso univariante. El espacio R^p es dividido en un conjunto de regiones disjuntas y dentro de cada región se ajusta un polinomio de p variables. Para $p > 2$ se toman regiones disjuntas que definen la aproximación spline como productos tensores de intervalos disjuntos en cada una de las variables delineadas por la ubicación del nodo. Así ubicando K_j nodos en cada variable produce un producto de $K_j + 1$ regiones, j = 1, ..., p.

Un conjunto de funciones bases que generan el espacio de las funciones spline sobre todo el conjunto de regiones, es el producto tensorial de las correspondientes básales spline uni-dimensionales asociadas con la ubicación de los nodos en cada variable dada por:

$$jf_q(\tilde{x}) = \sum_{k_p} \ldots \sum_{k_p} a_{k_1 \ldots k_p} \prod_{j=1}^{p} B_{k_j}^{(q)}(x_j)$$

$$= a_0 + \sum a_m \prod_{j=1}^{p} [s_{km}(x_{v(k,m)} - t_{km})]_+^q$$

$$= a_0 + \sum_{K_m=1} f_{i(x_i)} + \sum_{K_m=2} f_{ij(x_i,x_j)} + \sum_{K_m=3} f_{ijk(x_i,x_j,x_k)} + \ldots$$

3.3. Estación Gotuzzo Modelos MARS

Para este estudio se han utilizado mediciones realizadas cada doce horas. El objetivo es intentar describir, con adecuada precisión y anticipación, el comportamiento de esta variable en función de su comportamiento previo y de un conjunto de variables meteorológicas: velocidad del viento, humedad relativa, temperatura y dirección del viento.

Se disponía de n = 286 observaciones compuestas de valores de x_1 = temperatura, x_2 = humedad relativa, x_3 = velocidad del viento, x_4 = dirección del viento, x_5 = hora de la medición (0 o 12 hs) y PM10 = nivel de concentración de material particulado de diámetro aerodinámico menor o igual a 10μm, espaciadas cada 12 horas desde el 1 de Mayo al 30 de Septiembrede 1994.

Se definieron una variable "dummy" para x_5 y otras tres para incorporar al modelo cuatro direcciones principales del viento. Para aplicar la metodología MARS se estandarizó las variables x_1, x_2 y x_3 ($\mu = 5$, $\sigma = 1$) junto con transformar logarítmicamente la respuesta a modelar $y = \ln(\text{PM10})$.

Como se ve en la Tabla N^o 6, los tres modelos resumidos son satisfactorios en terminos de estimación (uno, dos o tres pasos hacia adelante) del nivel de PM10. Para mayor anticipación, 48 horas o más, la situación desmejora.

Modelo	GCV	r	mpab
1: p= 2, q= 1	0.937	0.59	0.14
2: p= 2, q= 2	1.104	0.50	0.15
3: p= 2, q= 3	1.091	0.44	0.16

Tabla N^o 6 . Comparación de los tres modelos MARS construidos para estación Gotuzzo.

En estos tres modelos las variables retenidas por MARS son: dos valores rezagados de PM10, día de la semana, un valor de humedad relativa (el menor orden de rezago), la temperatura está presente con dos valores en los modelos 1 y 2, pero sólo con uno en el modelo 3 y la velocidad del viento está presente sólo en los modelos 2 y 3.

3.4. Estación Gotuzzo: modelos MARS y Redes Neuronales

Las Redes Neuronales y MARS han sido aplicadas, en problemas de predicción para comparar su velocidad y exactitud (De Veaux et al., 1993; Steinberg, 2001).

La aplicación sobre la cual se analizaran estos métodos, corresponde al material particulado en el aire medido con un filtro de 10 micrones, estos datos fueron registrados en el período de Mayo a Septiembre de 1994, tomados diariamente, en promedio de una hora, dando un total de 3457 registros en la estación Gotuzzo ubicada en el centro de Santiago; además se tiene la información correspondiente a las variables meteorológicas de temperatura, humedad relativa, dirección y velocidad del viento, para el mismo período y cada una hora.

Para el análisis de estos registros, se consideraron dos subconjuntos de datos, uno de ellos involucra a los meses de julio y agosto correspondiente al periodo de invierno donde las condiciones de ventilación son desfavorables, registrándose en este periodo los mayores niveles de contaminación, el tamaño de muestra para estos meses es de N=1431, el otro conjunto involucra la totalidad de las observaciones.

Las variables involucradas son: x_1 = Temperatura (^{o}C), x_2 = Humedad relativa, x_3 = velocidad del viento (m/s), x_4 = Dirección del viento, x_5 = Mes del año (de Mayo a Septiembre), x_6 = Día de la semana (de Lunes a Domingo), x_7 = Hora de registro (de 0 a 23), x_8 = Nivel de concentración por material particulado de diámetro menor a 10 μm (PM10).

Tres variables dummy fueron definidas para el manejo de la dirección del viento (N - S, E - O, NE - SO y NO - SE), se utilizó una escala logarítmica para la variable respuesta PM10, y un operador de rezago de un período de 12 horas.

Con el objeto de comparar los resultados de ambas metodologias se consideraron las estadísticas: generalized cross validation GCV, la correlación lineal entre el valor observado y el predicho $r = \text{corr}(y_i, jf_{\text{q}}(x_i))$ y la proporción de error medio absoluto mpab = $\frac{1}{N}\sum_i \frac{\left|y_i - jf_{\text{q}}(x_i)\right|}{y_i}$.

Para la aplicación de las redes neuronales, se utilizo el software NeuroShell 2 (1995), con una arquitectura de multicapas feedforward, para cada red de entrenamiento se selec-

cionó un 15 % de los datos como conjunto de chequeo o prueba, en cada una de ellas se utilizó una capa de entrada formada entre 8 y 12 neuronas según corresponda y en cuyas conexiones con las capas ocultas se utilizaron tangentes hiperbólicas, funciones logísticas, lineales (-1 1), y normales como funciones de activación. Los criterios de entrenamiento de "Backpropagation" seleccionan los patrones aleatorios y utilizan un error medio inferior al 0.2 % para detener el entrenamiento de la red en conjunto con la cantidad de entrenamiento hecho a través de un monitoreo del error de predicción de la red. Una vez que la red ha sido entrenada se obtiene las contribuciones y los valores estimados por la red, estos resultados pueden adjuntarse al archivo de datos de manera tal de obtener gráficos y el cálculo del error cuadrático medio entre el verdadero valor y el predicho por la red neuronal.

Período Invierno (Julio - Agosto)

Los resultados del entrenamiento de las Redes Neuronales y MARS para el período de invierno Julio - Agosto se resumen en la Tabla N^o 7, estos valores corresponden a la información tomada con 12 horas de diferencia (rezago) respecto del nivel observado de contaminación por partículas.

	Redes		Mars	
Mes	r	mpab	r	mpab
Julio	0.54	0.18	0.69	0.13
Agosto	0.37	0.14	0.54	0.11
Total	0.48	0.16	0.64	0.12

Tabla N^o 7. Resultados de Redes Neuronales y Modelos MARS periodo invierno.

En la Tabla N^o 7, podemos observar que las correlaciones más altas entre el valor observado y la estimación son las obtenidas por el modelo MARS, y que en promedio el porcentaje de error es de un 12 %, en cambio en las redes neuronales este porcentaje aumenta a un 16 %.

Para predecir, se utilizan los patrones de entrenamiento de la red, con lo cual se obtuvieron los pesos de las conexiones de las neuronas en las distintas capas. A partir de estos pesos y con nuevos vectores de entrada se estiman los valores para las 12 horas siguientes. Este proceso se puede replicar hasta que el error de predicción no sea demasiado grande, si no es necesario reentrenar la red. En la Figura N^o 1 se muestra un gráfico de los valores predichos obtenidos por las redes y Mars y los valores observados.

Período Total (Mayo - Septiembre)

La resultados del entrenamiento de las Redes Neuronales y MARS para el total de las observaciones del periodo mayo - septiembre se resumen en la Tabla N^o 3, estos valores

corresponden a la información tomada con 12 horas de diferencia (rezago) respecto del nivel observado de contaminación por partículas.

	Redes		Mars	
Mes	r	mpab	r	mpab
Mayo	0.52	0.115	0.55	0.111
Junio	0.42	0.128	0.44	0.125
Julio	0.58	0.157	0.60	0.152
Agosto	0.53	0.118	0.58	0.115
Septiembre	0.55	0.117	0.60	0.112
Total	0.53	0.127	0.57	0.123

Tabla Nº 8. Modelos MARS y Redes Neuronales período total.

Podemos observar en la Tabla Nº 8 que los valores correspondientes a MARS son levemente mejores que los obtenidos por las Redes Neuronales Para este modelo con las Redes Neuronales se probaron distintos tipos de arquitecturas, aunque finalmente se utilizó el mismo tipo de red de multicapas feedforward que la del período de Invierno; sin embargo, se obtuvo una mayor precisión en las estimaciones, ya que hubo una mayor cantidad de patrones para su entrenamiento.

Las predicciones, utilizando los patrones de entrenamiento de la red neuronal y el modelo Mars descrito anteriormente para el período completo se muestra en la siguiente Figura:

3.5. Estación Pudahuel: modelos MARS y Regresión Lineal Múltiple

Para este estudio se utilizaron las bases de datos de la estación Pudahuel de la red de monitoreo MACAM2-RM, de los años 1998, 1999 y 2000. Se utilizó esta estación debido a que ella presenta los mayores índices de concentración de PM10; por tal motivo es la estación con mayor influencia en el momento de toma de decisiones respecto a decretar episodios críticos. En cada año se seleccionaron las mediciones realizadas en el período 01 de abril al 31 de agosto. El ajuste del modelo se ha validado con la muestra del año siguiente al año del modelo generado, de esta manera se busca garantizar la independencia de los datos usados para validar el modelo respecto a los usados en su construcción.

Se usaron 152 observaciones multidimensionales compuestas por una variable respuesta y 13 variables predictoras. Las variables predictoras se definieron como: promedio horario concentración PM10 a las 0:00 hrs del día N (pm0), promedio horario concentración PM10 a las 6:00 hrs del día N (pm6), promedio horario concentración PM10 a las 12:00 hrs del día N (pm12), promedio horario concentración PM10 a las 18:00 hrs del día N (pm18), máximo de concentración del promedio móvil 24 hrs de PM10 entre las 19 hrs día N-1 y 18 hrs día N (pm10h), máxima temperatura entre 19 hrs día N-1 y 18 hrs día N (mth), mínima humedad relativa entre las 19 hrs día N-1 y 18 hrs día N (mhrh), temperatura máxima menos mínima entre 19 hrs día N-1 y 18 hrs día N (dth), promedio velocidad del viento entre 19 hrs día N-1

y 18 hrs día N (vvh), máxima temperatura día N+1 (mtm), mínima humedad relativa día N+1 (mhrm), temperatura máxima menos mínima día N+1 (dtm) y promedio velocidad del viento día N+1 (vvm). La respuesta en estudio es el máximo de concentración del promedio móvil 24 hrs de PM10 del día N+1 (pm10m). Se construyeron once modelos MARS por año, trabajando con tres conjuntos de variables predictoras: el total de variables, sólo variables de ayer y hoy y finalmente el conjunto de variables pm0, pm6, pm18, dtm, dth y vvm.

Las regresiones lineales múltiples, se construyeron usando el mismo conjunto de variables seleccionadas por MARS con el fin de comparar ambas metodologías. Finalmente se trabajó con el último conjunto de variables predictoras.

Las autoridades de la Comisión Nacional del Medio Ambiente (CONAMA), han definido cuatro niveles de concentraciones de PM10, con el objeto de tomar decisiones administrativas al momento de generarse episodios críticos, los niveles son: bueno 0–193μg/m^3; alerta 194–239 μg/m^3; pre-emergencia 240–329 μg/m^3 y emergencia PM10 $>$ 340 μg/m^3 (CONAMA, 1998). Para nuestro estudio dicotomizamos la respuesta en dos clases; I : pm10m $<$ 240 μg/m^3 y II: pm10m $\geq$ 240 μg/m^3, es decir, "bueno o alerta" versus "pre-emergencia o emergencia".

Con el objeto de comparar los modelos se consideraron las siguientes estadísticas: (1) generalized cross validation GCV, (2) correlación lineal entre el valor observado y el predicho $\mathrm{r} = \mathrm{corr}(\mathrm{y}_i, jf_{\mathrm{q}}(x_i))$ y(3) proporción de error medio absoluto $\mathrm{mpab} = \frac{1}{N}\sum_i \frac{\left|y_i - jf_{\mathrm{q}}(\,x_i\,)\right|}{y_i}$. Además se consideró la proporción de aciertos en cada clase, adicionalmente se construyeron modelos variando el número de funciones bases para ver si se mejoraban las predicciones.

Los resultados de los modelos que se muestran a continuación corresponden a la validación, hecha aplicando cada modelo a los datos del año siguiente.

Modelos año 1998

Para los modelos del año 1998, el proceso de selección (forward/backward) de MARS se inicia con seis predictores: pm0, pm6, pm18, dtm, dth y vvm. En este caso MARS selecciono para los Modelos Nº 1 y 3 el total de los predictores, mientras que para el Modelo Nº 2 selecciono a pm0, pm6, pm18, dtm y vvm.

El modelo Nº 1 corresponde a un modelo aditivo sin interaccion entre predictores y con una entrada de 60 funciones base. El Modelo Nº 2 corresponde a un modelo multiplicativo con un nivel de interacción entre predictores con 40 funciones base de entrada, finalmente el Modelo Nº 3 es un modelo mutiplicativo con dos niveles de interacciones entre predictores y 40 funciones base de entrada.

Observando la Tabla Nº 9, se puede apreciar que los Modelos MARS superan a la Regresión Lineal Múltiple ampliamente en proporcion de aciertos a Clase II, la proporcion de error medio absoluto (mpab) se encuentran bastante parejos para los Modelos Nº 1 y 2, en cambio el mpab para el Modelo Nº 3 destaca MARS con un 16 % de error contra RLM con un 21,6 %.

Modelo 1	GCV	r	mpab	Clase I	Clase II
60 fb	1383,89	0,884	0,202	96 %	70 %
RLM	-	0,914	0,205	99 %	30 %
Modelo 2					
40 fb	1099,17	0,844	0,229	98 %	50 %
RLM	-	0,864	0,220	100 %	35 %
Modelo 3					
40 fb	1117,78	0,920	0,160	98 %	75 %
RLM	-	0,865	0,216	100 %	25 %

Tabla Nº 9. Resumen de modelos MARS y Regresión lineal múltiple para el año 1998.

Modelos año 1999

Para los Modelos del año 1999, el proceso de selección (forward/backward) de MARS se inicia con seis predictores: pm0, pm6, pm18, dtm, dth y vvm. En este caso MARS selecciono para los Modelos Nº 5 y 6 el total de los predictores, mientras que para el Modelo Nº 4 selecciono a pm0, pm6, pm18, dtm y vvm. El Modelo Nº 4 corresponde a un modelo aditivo sin interaccion entre predictores y con una entrada de 15 funciones base, el Modelo Nº 5 corresponde a un modelo multiplicativo con un nivel de interacción entre predictores con 20 funciones base de entrada, finalmente el Modelo Nº 6 es un modelo mutiplicativo con dos niveles de interacciones entre predictores y 40 funciones base de entrada.

Modelo 4	GCV	r	mpab	Clase I	Clase II
15 fb	1322,10	0,910	0,201	98 %	72 %
RLM	-	0,895	0,217	100 %	22 %
Modelo 5					
20 fb	1121,81	0,892	0,225	97 %	72 %
RLM	-	0,910	0,193	96 %	67 %
Modelo 6					
40 fb	1117,78	0,883	0,220	96 %	78 %
RLM	-	0,913	0,200	97 %	78 %

Tabla Nº 10. Resumen de modelos MARS y Regresión lineal múltiple para el año 1999.

En la Tabla Nº 10 se puede apreciar que los Modelos MARS superan a la Regresión Lineal Múltiple ampliamente en proporcion de aciertos a Clase II, salvo en el Modelo Nº 6 en donde ambas proporciones de acierto son idénticas con un 78 %. La figura N^o 1 muestra predicciones entre el mejor modelo MARS y RLM.

Figura N^o 1 Valores observados y predichos por MARS y RLM Modelo Nº 6 año 1999.

MARS entrega un modelo explícito para la respuesta en estudio, que automáticamente selecciona las variables que aparecen en la ecuación final. También indica lo complejo de la relación entre la respuesta y cada variable predictora, a partir del numero de funciones bases usadas .

En el Modelo Nº 6 del año 1999 se puede apreciar la selección de todos los predictores

ingresados al sistema y a su vez la ubicación de los nodos:

Pm10m = 45.328 + 0.466 * max(0, PM18 - 16.000)+ 0.380 * max(0, PM6 - 1.000)- 9.353 * max(0, VVM - 13.000) * max(0, DTM - 11.000)+ 0.987 * max(0, DTH - 1.000) * max(0, VVM - 13.000) * max(0, DTM - 11.000) + 0.083 * max(0, 108.000 - PM18) * max(0, 13.000 - VVM) * max(0, DTM - 11.000)- 0.102 * max(0, PM18 - 184.000) * max(0, 11.000 - DTM)+ .900773E-03 * max(0, PM0 - 340.000) * max(0, PM18 - 16.000)+ 1.207 * max(0, DTM - 7.000) * max(0, DTH - 5.000)- 0.174 * max(0, VVM - 6.000) * max(0, DTM - 7.000) * max(0, DTH - 5.000)

Figura N^o 2. Contribución de las interacciones a la respuesta del Modelo Nº 6

4. Conclusiones

La adaptación de algoritmos de análisis discriminante, paramétricos y no-paramétricos, a la clasificación predictiva de eventos de riesgo ambiental en la ciudad de Santiago ha resultado alentadora. Es particularmente interesante el desempeño observado con la metodología no-paramétrica: como ella no requiere supuestos distribucionales nos coloca en posición más realista y flexible.

Tanto las Redes Neuronales como el método estadístico para regresiones no lineales MARS, tienen que necesariamente fijar (poner a punto) parámetros que especifican la complejidad del modelo; en el caso de las RN su arquitectura, el número de nodos en las capas ocultas y la cantidad de capas juegan un rol fundamental, en cambio en MARS se necesita especificar el número máximo de funciones bases y el orden máximo de interacción entre las variables predictoras; sólo un conocimiento a priori del problema nos podría ayudar, y aún así resulta difícil para cada problema en particular.

Usando las redes neuronales uno debe tener cuidado de no sobre ajustar los datos, este problema es particularmente dañino en conjuntos de datos pequeños, ya que un sobreajuste implica una mala estimación cuando se aplica a nuevos datos, en cambio MARS tiene la capacidad de "podar el modelo" después de ajustar los datos, a través, de la penalización en el criterio GCV por el número de parámetros estimados por el modelo, consiguiendo de esta manera un modelo de máxima parsimonia.

Por otra parte MARS selecciona automáticamente las variables predictoras y detecta posibles interacciones entre ellas generando modelos más flexibles; ya que las interacciones están siempre restringidas a alguna subregión, estas interacciones quedan expresados algebraicamente a través de las funciones basales, logrando de esta forma establecer un modelo. En las Redes neuronales una vez seleccionada su topología, (el número de capas, neuronas, funciones de activación ,etc) y entrenada la red, podemos aplicar su aprendizaje a un nuevo conjunto de datos, pero sin embargo, no podemos describir un modelo en función de las variables de entrada.

Una vez seleccionado el modelo óptimo, MARS reajusta el modelo para cada variable, de modo de determinar el impacto en la calidad del modelo al eliminar dicha variable; así se asigna un ranking relativo desde la variable más importante a la menos importante. De esta manera, se definen variables reemplazantes o competidoras con lo cual la metodología de MARS permite tratar los valores missing o faltantes.

De esta manera, se observó que en los problemas donde se tiene un tamaño de mues-

tra menor, el modelo MARS funciona mejor, obteniendo aprendizajes y predicciones mas precisas, en un menor tiempo de estimación.

En el caso de la aplicación en la Estación Pudahuel se han originado cambios de concentraciones anuales y mensuales de material particulado entre 1997 y 2001, ellos se pueden deber a: medidas de descontaminación global, medidas extraordinarias los días de episodios críticos, uso de modelo de pronóstico y factores meteorológicos. Tales cambios influyen en el desempeño de los modelos y hace que éstos no sean tan complejos en su estructura. A su vez se ven diferencias de un año a otro, lo que podría estar influenciado por condiciones meteorológicas particulares de cada año, por ejemplo año seco, fenómeno del Niño u otros cambios climatológicos.

MARS se desempeña de manera similar al cambiar la cantidad de funciones bases para un determinado conjunto de variables predictoras y el año para el cual se construyó el modelo y aquel con que se valida. Esto podría explicarse ya que las series de tiempo de PM10 con el transcurrir del tiempo muestran una tendencia descendente y variaciones de concentración más suaves.

5. Discusión

Los procedimientos de análisis discriminante no-parametrico y modelos MARS aplicados en la Estación Gotuzzo han mostrado su eficiencia para modelar y predecir la contaminación atmosférica por material particulado superando a otras metodologías alternativas.

Series cronológicas con funciones de transferencia involucrando variables meteorológicas dió un 40 % de proporción media de error (Silva et al., 1994). Redes neuronales con suavizamiento previo dieron aproximadamente 30 % de error (Pérez et al., 1998), estos resultados son consistentes con los descritos por otros autores (De Veaux et al., 1993; Steinberg, 2001) que encuentran que MARS es en muchas aplicaciones , más exacto y más rápido que las redes neuronales.

Ruttland y Garreaud (1995), aplicaron análisis discriminante para desarrollar un Meteorological Air-Pollution-Potencial Index (MAPPI) que predice con 12 horas de anticipación y 73 % de exactitud, los episodios de alta contaminación potencial del aire. La discriminación no-parametrica es recomendable para propósitos de predicción cualitativa o clasificación con la ventaja adicional del fácil acceso a software típico (SAS o equivalente).

En el caso de las aplicaciones a la Estación Pudahuel las variables predictoras que mejor explican la respuesta serian las concentraciones puntuales de PM10: pm0, pm6 y pm18 y las variables meteorológicas: dtm, vvm y dth lo que es consistente con (Silva et al., 2001), en el sentido de que MARS selecciona adecuadamente variables relacionadas con persistencia de condiciones de ventilación, las que tienen relación con la meteorología.

Las metodologías de predicción aplicadas nos entregan modelos adecuados para estudiar la contaminación por material particulado, la regresión lineal múltiple es inferior en aciertos que MARS y coloca a esta última metodología, como una mejor herramienta de predicción. Este ultimo punto es consistente con resultados de otros autores que muestran que MARS es más eficiente que otras técnicas (De Veaux et al., 1993; Silva et al., 2001).

Agradecimientos

Los autores agradecen el apoyo financiero de los proyectos FONDECYT 1930096, 1970418, 1010085; DICYT-USACH 9533SZ y a la Organización Panamericana de la Salud (OPS).

Referencias

Alvarado, S., Silva, C., y Pérez, P. (2002). Predicción de calidad del aire para material particulado pm10 en la estación de monitoreo pudahuel de la red macam-2, comparación de dos modelos predictivos. Quinto Congreso Latinoamericano de Sociedades de Estadística (CLATSE V), 28 de Octubre a 1 ° de Noviembre, Caseros, Buenos Aires, Argentina.

Alvarado, S., Silva, C., y Pérez, P. (2003). Modelos mars; una aplicación a una muestra de material particulado en la región metropolitana. "I Coloquio Nacional de Estadística, 9 y 10 de Enero del 2003, Facultad de Medicina, Universidad de Chile".

CONAMA (1998). Comisión Nacional del Medio Ambiente. Decreto Ley N ° 59, Diario Oficial de la Republica de Chile. Lunes 25 de mayo de 1998".

De Veaux, R. D., Psichogios, D. C., y Ungar, L. H. (1993). A comparison of two nonparametric estimation schemes: Mars and neural networks. *Computers Chemical Engineering*, 17, N ° 8:819–837.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–194.

Friedman, J. H. y Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research*, 4:197–217.

Hang, D. J. (1982). *Kernel Discriminant Analysis*. Research Studies Press. John Wiley & Sons.

Lewis, P. A. W. y Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). Journal of the American Statistical Association Vol. 86, N ° 416.

MARS (2001). *User Guide*. Salford Systems.

NeuroShell 2 (1995). *User's Manual*. Ward Systems Group, Inc..

Pérez, P., Trier, A., Silva, C., y Montaño, R. M. (1998). Prediction of atmospheric pollution by particulate matter using a neural network. Proc. of the 1997 Conf. on Neural Inf. Proc., Dunedin, New Zealand, Springer-Verlag, Vol 2.

Ruttland, J. y Garreaud, R. (1995). Meteorological air pollution potential for Santiago, Chile: towards an objective episode forecasting. *Environmental Monitoring and Assessment*, 34:223–244.

Sanhueza, P., Vargas, C., y Jimenez, J. (1998). Mortalidad diara en Santiago y su relación con la contaminación del aire. *Revista Médica de Chile*, 127(2):235–242.

Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization.* John Wiley & Sons.

Silva, C., Firinguetti, L., y Trier, A. (1994). Contaminación ambiental por partículas en suspensión: Modelamiento estadístico. Actas XXI Jornadas Nacionales de Estadística, Concepción, Noviembre.

Silva, C., Pérez, P., y Trier, A. (2001). Statistical modelling and prediction of atmospheric pollution by particulate material: two nonparametric approaches. *Environmetrics*, 12:147–159.

Silva, C., Pérez, P., Trier, A., y Montaño, R. M. (1998). Modelamiento estadístico y predicción mars de la contaminación ambiental por partículas en suspensión. VII Congreso Latinoamericano de Probabilidad y Estadística Matemática, Córdoba, Octubre.

Silva, C. y Trier, A. (1995). Modelamiento estadístico y predicción de la contaminación ambiental por partículas en suspensión. Actas del VII Congreso Internacional de Biomatemática, Buenos Aires, Octubre.

Steinberg, D. (2001). An alternative to neural nets: Multivariate adaptive regression splines (mars). *PC AI's*, 15, N ° 1:28–41.

Trier, A. e Firinguetti, L. (1994). A time series investigation of visibility in an urban atmosphere – I. *Atmospheric Environment*, 28(5):991–996.