Aplicación de algoritmos de clustering desarrollados en el entorno FIR a la predicción de la concentración de ozono

Pilar Gómez** Angela Nebot** Francisco Mugica*

**Depto. Lenguajes y Sistemas Informáticos
Universidad Politécnica de Cataluña,
Módulu C6 – Campus Nord
Jordi Girona Salgado, 1-3, Barcelona 08034, España.
Teléfono: (343) 4015642: Fax: (343) 4017014
pgomez@lsi.upc.es Becaria IPN y SUPERA
angela@lsi.upc.es

*Centro de Inv. En Ciencia Aplicada y Tecnología Avanzada (CICATA)
Instituto Politécnico Nacional,
Legaria 694, Col. Irrigación C.P. 115000, México, D. F.
Teléfono: (525) 5577596; Fax: (525) 5575103
fmugica@mexico.com

Resumen: *El presente trabajo tiene como objetivo estudiar la aplicación de diferentes algoritmos de clustering desarrollados en el entorno de la metodología FIR al problema de la predicción a largo plazo de las concentraciones de ozono en la zona centro de la ciudad de México. La investigación realizada se centra en la identificación de modelos para la predicción del ozono desde dos perspectivas distintas: modelado "estacional" y modelado "mensual". El modelado "estacional" tiene como objetivo identificar modelos para una determinada estación del año (período no lluvioso, en este caso). El modelado "mensual" tiene como objetivo identificar modelos para cada mes del año (mes de enero, en este caso). Los algoritmos de clustering seleccionados para este estudio han sido dos jerárquicos, Complete Linkage y Ward Linkage y tres no jerárquicos, Equal Frequency Partition, K-means y Fuzzy C-means.*

Palabras claves: Modelos de contaminación ambiental; Razonamiento Inductivo Difuso; Clustering

1 INTRODUCCIÓN

Este trabajo se puede considerar la continuación de la investigación llevada a cabo en el reporte de investigación LSI-01-40-R titulado "Predicción a largo plazo de la concentración de ozono usando la metodología de Razonamiento Inductivo Difuso (FIR)" y publicado posteriormente en el congreso internacional European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (EUNITE'2001), celebrado en Tenerife en diciembre del 2001. En él se estudió el problema del modelado y predicción del contaminante ozono en la ciudad de México utilizando la metodología FIR. La investigación realizada se centró en la identificación de modelos desde dos perspectivas distintas. La primera pretendía obtener un modelo "estacional" con el objetivo de pronosticar el período no lluvioso del año 2000, para lo cual se utilizaron datos consecutivos de enero a mayo del 2000. La segunda pretendía obtener un modelo "mensual" con el objetivo de pronosticar el mes de enero del 2000, para lo cual se utilizaron los datos de los meses de enero de los años anteriores (1996-1999). Los resultados obtenidos fueron esperanzadores y permitieron concluir que el potencial de la metodología FIR para predecir esta aplicación es considerable. Sin embargo, los modelos generados no eran capaces de predecir los picos más altos del contaminante ozono, aspecto que es de especial relevancia si se tiene en cuenta que el objetivo final de esta línea de investigación es la predicción de posibles contingencias ambientales.

El presente trabajo tiene como objetivo estudiar la aplicación de diferentes algoritmos de clustering, desarrollados en el entorno de la metodología FIR [Pinyol, 2002], al mismo problema de contaminación ambiental tratado en la anterior investigación. Cabe remarcar que, en el trabajo anterior, la mayoría de los modelos reportados utilizaron como método de discretización el Equal Frequency Partition (EFP), que a partir de un número de clases predefinido (granularidad) reparte los elementos de la variable a ser discretizada entre las diferentes clases de manera que cada clase tenga el mismo número de elementos. El método EFP es extremadamente simple y resulta muy útil en aquellos casos en que los datos disponibles del sistema representen los posibles comportamientos de éste mediante un número similar de registros (elementos). Para el resto de los modelos se usó el algoritmo Fuzzy C-means.

En la siguiente sección se describirán brevemente los algoritmos de clustering que se han utilizado en este estudio. En la sección 3 se presenta el modelado y predicción del ozono en la ciudad de México utilizando dichos algoritmos en el proceso de fusificación de la metodología FIR. Se divide el estudio en las dos perspectivas de modelado comentadas anteriormente: "mensual" y "estacional". Finalmente, se darán las conclusiones del trabajo desarrollado. En este reporte no se describe en ningún momento la metodología del Razonamiento Inductivo Difuso (FIR), ni el problema de la contaminación ambiental en la ciudad de México, por considerar que ambos están ampliamente descritos en el trabajo precedente [Nebot *el al.,* 2001]. Para una descripción de FIR que permita al lector no conocedor de esta metodología entender el trabajo que aquí se reporta referirse a [Cellier *et al.,* 1996; Nebot *et al.,* 1998].

2 ALGORITMOS DE CLUSTERING

Una de las clasificaciones más comunes que se realizan de los algoritmos de clustering es la división entre algoritmos jerárquicos y no jerárquicos. Los primeros construyen una jerarquía de nodos y a partir de ella encuentran los conjuntos o clases. Los segundos, utilizan diferentes técnicas de agrupación sin ningún tipo de jerarquía. En principio los métodos jerárquico están aportando más información ya que generan un árbol de relaciones entre todos los elementos. Sin embargo, son bastante más costosos en espacio y tiempo de cómputo. En este estudio se utilizaran dos algoritmos *jerárquicos (Complete* y *Ward Linkage*) y tres no jerárquicos (*Equal Frequency Partition, K-means* y *Fuzzy C-means*). En esta sección se introducirán estos cinco algoritmos.

2.1 CLUSTERING JERÁRQUICO

El clustering jerárquico genera una partición de los datos de entrada en k grupos mediante un árbol construido a partir de una función de distancia (en este caso la Euclidiana). La dificultad de este método radica en la construcción de esta jerarquía que relacione todos los elementos a partir de su distancia. Hay dos posibles alternativas para la construcción del árbol que se pueden ver como un proceso de construcción *forward* o un proceso *backward*. En el primero de ellos, inicialmente se toma cada punto como un solo grupo (o clase) y se construye la jerarquía fusionando los grupos más cercanos hasta que se llega al grupo que engloba todos los puntos. En el segundo, inicialmente se toma un único grupo (clase) que contiene todos los puntos y se sigue el proceso inverso al anterior, generando nuevos grupos en función de la distancia a sus componentes. Los dos algoritmos que se utilizan en este trabajo son del primer tipo y se acostumbran a llamar aglomerativos. Al segundo grupo se les llama algoritmos particionados. Tanto en un caso como en el otro, en cada paso del algoritmo se toma una decisión (agrupar o dividir) que será definitiva para el resto del proceso. Es por este motivo que los algoritmos jerárquicos son muy sensibles a los outliers. Sin embargo, los outliers pueden ser detectados al graficar el dendrograma (forma gráfica de representación del árbol).

Los algoritmos aglomerativos se basan en el algoritmo *Linkage*, donde diferentes variaciones en la fórmula de cálculo de las distancias al fusionar dos grupos conlleva algoritmos distintos [Mucha and Sofyan, 2000]. El algoritmo de Linkage empieza calculando una matriz D de distancias de todos los puntos entre si. Posteriormente encuentra los dos grupos con menor distancia (basándose en D) y los agrupa en uno solo. Recalcula la nueva matriz D de distancias que tendrá ahora una dimensión menos y se repite el proceso de nuevo hasta llegar a que el número de grupos sea uno. El cálculo de las distancias es el que puede variar, generándose los diferentes algoritmos.

Sea P y Q los dos grupos con menor distancia que se ha decidido agrupar, la distancia entre el nuevo conjunto (que llamaremos $P \cup Q$) y el conjunto R se puede calcular mediante la siguiente fórmula:

$$D[R, P \cup Q] = \delta_1 D[R, P] + \delta_2 D[R, Q] + \delta_3 D[P, Q] + \delta_4 |D[R, P] - D[R, Q]|$$

donde las deltas se definen en la tabla 1 y dan origen a los diferentes algoritmos Linkage.

Tabla 1. Tabla de cálculo de las distancias de los algoritmos Linkage

	δ_1	δ_2	δ_3	δ_4
Single Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Simple Average Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Average Linkage	$\frac{np}{np+nq}$	$\frac{nq}{np+nq}$	0	0
Centroid	$\frac{np}{np+nq}$	$\frac{nq}{np+nq}$	$-\frac{np \cdot nq}{(np+nq)^2}$	0
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{nr+np}{nr+np+nq}$	$\frac{nr+nq}{nr+np+nq}$	$-\frac{nr}{nr+np+nq}$	0
Flexible Method	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	β	0

Donde:

np = nº de elementos del grupo P: $N(i=1..n \mid x_i \in P)$,
nq = nº de elementos del grupo Q: $N(i=1..n \mid x_i \in Q)$,
nr = nº de elementos del grupo R: $N(i=1..n \mid x_i \in R)$,
β = Parámetro de entrada dado por el usuario

En la tabla 1 se describen los diferentes algoritmos de *Linkage* implementados en el entorno de FIR a pesar de que en este estudio solo se han utilizado el *Complete* y el *Ward.*

Complete Linkage: Este algoritmo utiliza la distancia máxima entre los datos, por lo que los árboles resultan muy equilibrados. Como desventaja principal se tiene que son más sensibles a los outliers.

Ward Linkage: Este método realiza la agrupación lo más homogénea posible, utiliza la medida de perdida de información entre los grupos para dirigir las futuras agrupaciones.

2.2 CLUSTERING NO JERÁRQUICO

Son algoritmos que no utilizan el concepto de jerarquía para realizar la agrupación, aunque la mayoría de ellos también utiliza la medida de la distancia.

Equal Frecuency Interval

Este algoritmo no analiza los valores de los datos para realizar la agrupación, simplemente crea grupos con el mismo número de elementos. Este algoritmo empieza con una ordenación de los datos y posteriormente divide el vector en tantas partes iguales como venga indicado por el número de clases preestablecido.

K-means

La idea es asignar cada elemento a aquel cluster que tenga su centroide más cerca de él [Mucha and Sofyan, 2000]. Se entiende por centroide de un conjunto de datos aquel punto ficticio que hace que la distancia entre él y los elementos del conjunto sea mínima. Se puede encontrar simplemente calculando la media aritmética de cada dimensión de los datos del grupo. Sea un conjunto de datos en el espacio R^m, x_j el elemento j-ésimo del conjunto, x_{ji} la dimensión i_ésima del j-ésimo elemento, el valor de la dimensión i_ésima de su centroide se calculará mediante la ecuación 1.

$$c_i = \frac{\sum_{j=1}^{N} x_{ji}}{N} \qquad (1)$$

El algoritmo genera, inicialmente, una partición aleatoria de los datos en tantos clusters como se haya indicado. Se calculan los centroides de los clusters generados mediante la ecuación 1. En este punto se realiza la iteración, con la idea de asignar cada elemento al cluster que tenga el centroide más cercano a él. Si es diferente al que pertenece actualmente, se coloca el elemento en el nuevo cluster y se recalculan los centroides de los dos grupos involucrados. Esto se repite hasta que no haya movimiento de los elementos o hasta que el número de iteraciones llegue al límite. Este método genera grupos de datos bastante compactos y, generalmente, con buena distribución.

Fuzzy C-means

Este algoritmo utiliza la lógica difusa de manera que a cada elemento se le asigna una función de pertenencia a cada uno de los clusters [Bezdek et al., 1984]. Dado un elemento y una clase, esta función indica el grado de pertenencia de ese elemento a esa clase (cluster). Al igual que el anterior, este algoritmo genera, inicialmente, una partición. En este caso, la partición corresponderá a realizar una matriz de pertenencias con la posibilidad de asignar valores aleatorios. Seguidamente se calculan los centroides, mediante la fórmula descrita en la ecuación 2. El centrodie del cluster k, v_k se calculará de la manera siguiente:

$$v_k = \frac{\sum_{i=1}^{N} (u_{ik})^m x_i}{\sum_{i=1}^{N} (u_{ik})^m} \qquad (2)$$

donde N es el número de elementos, x_i el elemento i_ésimo, u_{ik} la pertenencia del elemento 'i' en el cluster 'k' y *m* el exponente de fusificación.

En este punto se inicia la iteración, donde se calculan las distancias de todos los nodos con los centroides de los clusters existentes. Posteriormente se calcula el grado de pertenencia de cada elemento. Sea D_{ik} la distancia entre el elemento i_ésimo del conjunto inicial i el centroide k_ésimo, el grado de pertenencia u_{ik} vendrá definido por la siguiente fórmula:

$$u_{ik} = \sum_{j=1}^{K} \left(\frac{D_{ik}}{D_{jk}} \right)^{\frac{2}{m-1}} \qquad (3)$$

Los elementos Djk = 0 no se incluirán en la fórmula y valdrán cero en el sumatório.

A continuación se calculará la función objetivo para la iteración actual, mediante la ecuación 4. Sea N el nº de elementos, K el nº de clusters, x_i el elemento i_ésimo, v_k el cluster k_ésimo, u_{ik} el grado de pertenencia del elemento 'i' en el cluster 'k' y d(a,b) la distancia euclidiana entre los puntos a i b, la función de coste es:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} (u_{ik})^m d(x_i, v_k)^2 \qquad (4)$$

Finalmente, se recalculan los centroides como se hacia anteriormente, utilizando la ecuación 2. El proceso de iteración termina cuando el número de iteraciones sea igual al definido por el usuario o

hasta que la diferencia de la función objetivo de dos iteraciones sucesivas sea menor que el parámetro introducido.

3 MODELADO Y PREDICCIÓN DEL OZONO

La investigación realizada se centra en la identificación de modelos para la predicción del ozono a largo plazo considerando solo un punto clave de la Zona Metropolitana de la Ciudad de México. La zona elegida para estos estudios fue la zona centro (merced), por ser una zona que cuenta con mayor actividad humana durante los días y las horas laborales, y por ende, está propensa a un mayor índice de contaminación. En el primer apartado se trabajó en la predicción del período no lluvioso, mayo del año 2000. En el segundo apartado se pronostica el segundo mes de invierno, enero del 2000.

Los datos horarios disponibles para este estudio son proporcionados por el sistema RAMA que mide los cinco contaminantes atmosféricos centinela (O3, NO, NO2, SO2, CO), junto con las condiciones ambientales y meteorológicas (humedad relativa, temperatura, velocidad del viento y dirección del viento) [Instituto Mexico & Los Alamos, 1994]. La consideración de estas variables es importante por ser las que pueden disipar o concentrar en algún momento los contaminantes. Estas mediciones se registran cada hora durante las 24 horas del día, los 365 días del año, por lo tanto la frecuencia de las mediciones es alta. Con esta forma de medición se generan tres variables temporales: la hora del día, el día de la semana y el año. Los datos horarios para la zona centro de la Ciudad de México, están disponibles desde el año de 1996 hasta el mes de mayo del 2000. La base de datos mencionada presenta como principal característica un alto porcentaje de datos perdidos, los cuales se presentan en todas las variables en distintos instantes de tiempo. FIR es capaz de tratar con este tipo de datos, por lo que no se requiere un tratamiento previo de la base de datos.

Los parámetros generales establecidos para la identificación de los modelos en los dos apartados desarrollados son: Las variables de entrada: velocidad del viento (vv), medida en metros por segundo (m/s); la dirección del viento (dv) en grados (0° a 365°); la temperatura (te) en grados centígrados (°C); la humedad relativa (hu) en porcentaje (%); las variables temporales, hora del día (ho) (0 a 24) y los días de la semana (ds) (representados del 1al 7). La variable de salida es el ozono (O3) que está medido en partes por millón (PPM).

La máscara candidata propuesta tiene una profundidad de 3, y cubre un período de tiempo de 2 horas. Esta propuesta se considera conveniente ya que el intervalo de muestreo es de una hora y con esta profundidad se consideran relaciones causales con respecto a la salida, con una anticipación de dos horas. La máscara candidata de la ecuación 5 se define con el máximo de posibles relaciones causales y temporales (valores –1) respecto a la salida en el instante *t*, por lo tanto no se deja de estudiar ninguna posibilidad de causalidad para la identificación del modelo cualitativo.

$$t/x \quad ho \quad ds \quad vv \quad dv \quad te \quad hu \quad O3 \tag{5}$$

t- 2δt	-1	-1	-1	-1	-1	-1	-1
t - δt	-1	-1	-1	-1	-1	-1	-1
t -	-1	-1	-1	-1	-1	-1	+1

Máscara candidata

Por otro lado, se ha decidido incluir en el estudio los mejores modelos obtenidos en el trabajo anterior [Nebot *et al.*, 2001], de manera que se puedan comparar los resultados que se obtengan en esta investigación con los que se reportaron en ese trabajo utilizando las mismas máscaras.

Para cuantificar la capacidad de predicción se utiliza el error cuadrático medio normalizado por la variancia, *MSE*, que se define en la ecuación 6.

$$\mathrm{MSE} = \frac{\mathrm{E}[(y_1(t) - y_2(t))^2]}{y_{var}} \tag{6}$$

en donde y_{var} es la variancia, el vector y_1 son los valores reales y los valores de predicción se encuentran almacenados en el vector y_2.

En esta investigación se parte de los mismos casos que se plantearon en la investigación anterior. La diferencia principal entre estos tres casos radica en el proceso de discretización de los datos; tanto en el aspecto de definir el número de clases, como en la forma de determinar las marcas. En la investigación precedente se utilizaron dos métodos de discretización, el *Equal Frequecy Partition* (EFP) y el *Fuzzy C-means* (FCM), ambos descritos en la sección anterior. El EFP fue usado en los casos **A** y **B,** mientras que el FCM se utilizó únicamente en el caso **C**. En el estudio actual, para los tres casos se han usado los algoritmos *Comlpete Linkage*, *Ward Linkage* y *K-means*. A continuación se describe las características de cada caso:

Caso A. Todas las variables son discretizadas en 3 clases.

Caso B. Las variables se discretizan de la siguiente manera: la hora del día y el día de la semana en 6 clases, la dirección del viento y la temperatura en 3 clases, la velocidad del viento y el ozono en 2 clases, la humedad relativa en 4 clases.

Caso C. Las variables se discretizan de la siguiente manera: hora del día, día de la semana y dirección del viento en 5 clases, el resto de las variables en cuatro clases.

Una vez terminada la discretización de las variables y el proceso de fusificación, el proceso de modelado cualitativo de FIR (referirse a [Cellier *et al.,* 1996; Nebot *et al.,* 1998]) se encarga de identificar la máscara óptima y las reglas basadas en patrones. La máscara seleccionada y la base de reglas patrón conforman el modelo FIR. Este modelo es posteriormente utilizado para efectuar la simulación cualitativa, que permite la validación del modelo mediante el pronóstico del conjunto de

datos de prueba. El error *MSE* se calcula respecto a los valores reales del ozono y los valores de la predicción obtenida con la metodología FIR.

3.1 MODELO "ESTACIONAL": PERÍODO NO LLUVIOSO

Para identificar este modelo se utilizan los datos de enero a abril del 2000 y para validarlo se usan los datos del mes de mayo [Instituto Mexico & Los Alamos, 1994]. El total de la base de datos cuenta con 3,648 registros de los cuales 929 registros (25.466%) son datos perdidos. Los datos perdidos para este período corresponden a las variables velocidad del viento, dirección del viento, temperatura y humedad relativa.

En la tabla 1 se presentan los resultados obtenidos en la predicción de las concentraciones de ozono del período no lluvioso del año 2000. La primera columna describe el caso desarrollado, la segunda señala si es máscara óptima o subóptima. La tercera columna describe las relaciones causales de la máscara, usando el formato de posición. Los números indican las posiciones donde se encuentran los elementos negativos (m-entradas) en la máscara y el último corresponde a la posición de la salida. Se enumera la máscara de arriba abajo y de izquierda a derecha. Las siguientes columnas contienen el error *MSE* obtenido cuando se predice el conjunto de prueba. En cada uno se indica el método de discretización utilizado. La columna sombreada corresponde a la discretización realizada en cada caso en la investigación precedente [Nebot *et al.,* 2001]. Como ya se ha mencionado anteriormente, en ese trabajo se usó el método EFP en los casos **A** y **B**, mientras que en el caso **C** se optó por usar el algoritmo de FCM.

*Tabla 1. Máscaras para la predicción de las concentraciones de ozono de los casos A, B y C. (**Modelo "Estacional"**).*

Caso	*Modelo*	*Relación Causal de la Máscara*	*MSE% EFP*	*MSE% Comp. Link.*	*MSE% Ward Link.*	*MSE% K-means*
A	Subóptimo	(14,15,17,21)	30.9834	24.0386	27.7362	24.9449
A	Subóptimo	(8,14,17,21)	33.0107	27.2309	25.8784	24.5323
A	Subóptimo	(10,14,15,21)	33.3741	23.0226	27.3248	27.6318
A	Óptimo	(1,14,21)	73.2878	42.4796	44.7391	42.3054
B	Subóptimo	(10,14,15,18,21)	25.1895	Sin salida	Sin salida	22.0184
B	Subóptimo	(10,14,15,21)	27.6692	25.1780	27.9893	24.8014
B	Subóptimo	(10,11,14,15,21)	30.4642	Sin salida	Sin salida	26.6509
B	Óptimo	(15,16,21)	109.8774	39.8243	53.7155	51.1712
			MSE% Fuzzy C-Means			
C	Subóptimo	(11,14,15,17,21)	26.1003	23.5581	Sin salida	24.7650
C	Subóptimo	(11,14,15,18,21)	36.0287	29.9534	40.9527	39.2880
C	Subóptimo	(9,14,15,18,21)	37.8871	31.0066	40.1889	34.2861
C	Óptimo	(14,17,21)	166.2589	190.5801	128.0047	146.0796

Como se puede observar en la tabla 1, en el caso **A** todos los algoritmos alternativos de clustering usados (*Complet Linkage*, *Ward Liankage* y *K-means*), mejoran la predicción respecto a la obtenida en la anterior investigación con el método de *Equal Frequency Partition*. Los errores disminuyen considerablemente en todos los casos y el error menor (23.0226%) se consigue con el modelo formado por la máscara (10, 14, 15 21), de complejidad 4, y usando el algoritmo *Complete Linkage*. Recordemos que en el caso **A** todas las variables se descretizaban en 3 clases. La máscara (10, 14, 15 21) se representa en forma matricial en la ecuación 7.

t/x	ho	ds	vv	dv	te	hu	o3	
$t - 2\delta t$	0	0	0	0	0	0	0	
$t - \delta t$	0	0	-1	0	0	0	-2	(7)
t	-3	0	0	0	0	0	+1	

Este modelo afirma que le valor actual del ozono, depende de los valores de la velocidad del viento y del ozono una hora anterior a la actual y de la hora del día en el instante actual. En la figura 1 se puede apreciar la predicción de los datos de test obtenida usando esta máscara (ecuación 7). La línea discontinua corresponde a la señal de predicción mientras que la línea continua representa los datos de test reales del ozono.

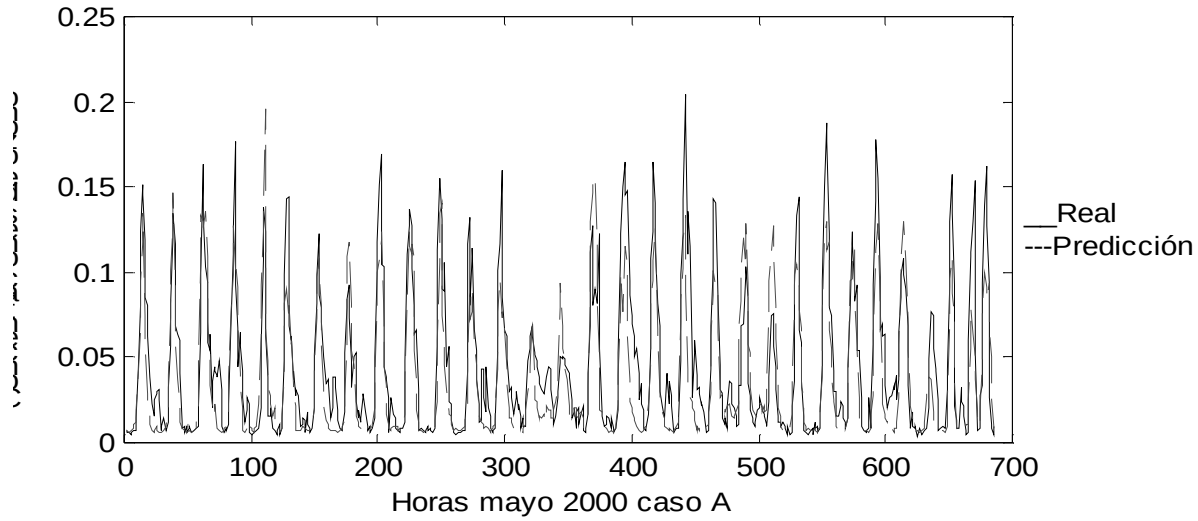

Figura 1. Predicción del conjunto de prueba con la máscara (10,14,15,21) y usando el algoritmo *Complete Linkage* (Modelo estacional: caso A).

En el caso **B** se considera un número diferente de clases para las distintas variables con la finalidad de estudiar sí una discretización más ajustada de las variables más relevantes o influyentes (las que aparecen en las máscaras con errores de predicción inferiores) permitirían mejorar el poder de predicción del modelo (máscara + base de reglas patrón). Los resultados de los modelos utilizando los métodos de clasificación *Complete Linkage, Ward Linkage* y *K-means* no presentan una mejora

significativa con respecto a los obtenidos con el método *Equal Frequecy Partition* (ver tabla 1). Cabe señalar que al aumentar considerablemente el número de clases de algunas de las variables, las máscaras de mayor complejidad (5) dejan de predecir en algunos casos ("sin salida" en la tabla 1). Esto es debido a que aumenta la expresividad del modelo pero, a su vez, disminuye su poder de predicción (predictivilidad) [Cellier *et al.,* 1996]. *K-Means* obtiene el menor error (22.0184%) utilizando el modelo (10,14,15,18,21) de complejidad cinco. La representación matricial de este modelo se muestra en la ecuación 8.

t/x	ho	ds	vv	dv	te	hu	o3	
$t - 2\delta t$	0	0	0	0	0	0	0	
$t - \delta t$	0	0	-1	0	0	0	-2	(8)
t	-3	0	0	-4	0	0	+1	

La salida del ozono en este modelo depende de las mismas variables del modelo del caso **A** representado en la ecuación 7, más de la variable dirección del viento (dv) en el instante actual. En la figura 2 se puede apreciar la predicción obtenida usando este modelo sobre el conjunto de datos de prueba.

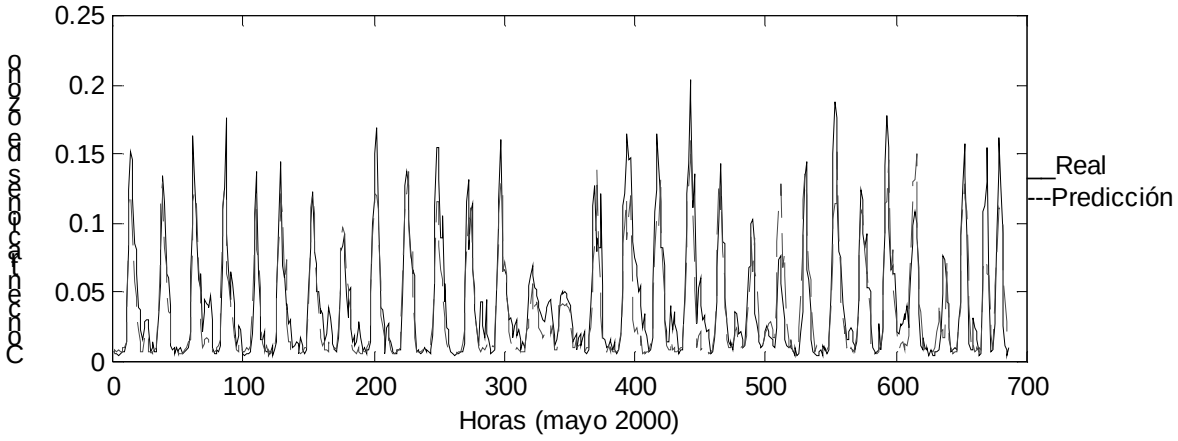

Figura 2. Predicción del conjunto de prueba con la máscara (10,14,15,18,21) y usando el algoritmo *K-means* (Modelo estacional: caso B).

Finalmente, analizando los errores presentados en la tabla 1 referentes al caso **C**, vemos que el algoritmo *Fuzzy C-means* usado en la anterior investigación obtiene mejores resultados que el *Ward Linkage* y resultados muy parecidos al algoritmo *K-means*. Para este caso, el algoritmo *Complete Linkage* es el que da errores menores, mejorando ligeramente los resultados obtenidos por el *Fuzzy C-means.*

3.2 MODELO "MENSUAL": MES DE ENERO

Para este estudio se utilizan los datos del mes de enero de los años 1996 a 1999 para identificar los modelos, y para probarlos o validarlos se predice la concentración del ozono del mes de enero del 2000 [Instituto Mexico & Los Alamos, 1994]. El total de registros disponible es de 3.720 de los cuales 443 (11.908%) son datos perdidos. Todas las variables (excepto ho y ds) presentan valores perdidos.

En la tabla 2 se presentan los resultados obtenidos en la predicción de las concentraciones de ozono (enero del 2000), usando los modelos "mensuales". Al igual que en la tabla 1, la primera columna indica el caso desarrollado, la segunda señala si es máscara óptima o subóptima. La tercera indica el método de clasificación. La cuarta columna describe las relaciones causales de la máscara. Las siguientes columnas contienen el error *MSE* obtenido cuando se predice el conjunto de prueba (test). En cada uno se indica el método de discretización utilizado. La columna sombreada corresponde a la discretización realizada en cada caso en la investigación precedente.

*Tabla 2. Máscaras para la predicción de las concentraciones de ozono de los casos A, B y C. (**Modelo "Mensual"**).*

Caso	*Modelo*	*Relación Causal de la Máscara*	*MSE% EFP*	*MSE% Comp. Link.*	*MSE% Ward Link.*	*MSE% K-means*
A	Subóptimo	(8,14,18,21)	25.3632	25.2917	23.6871	Sin salida
A	Subóptimo	(1,14,17,21)	25.4041	25.0474	28.4126	31.1698
A	Subóptimo	(10,14,15,18,21)	27.9998	Sin salida	Sin salida	Sin salida
A	Óptimo	(1,14,21)	89.0537	35.0448	36.8264	33.2961
B	Subóptimo	(10,14,15,21)	29.0347	33.2708	Sin salida	30.6307
B	Subóptimo	(14,15,17,21)	31.0473	30.6648	Sin salida	30.9179
B	Subóptimo	(8,14,17,21)	31.1448	31.0013	32.0539	Sin salida
B	Óptimo	(14,15,21)	67.6596	34.5689	Sin salida	50.2285
			MSE% Fuzzy C-Means			
C	Subóptimo	(14,15,18,21)	38.6147	Sin salida	Sin salida	37.6155
C	Subóptimo	(8,14,18,21)	41.9671	37.2999	Sin salida	Sin salida
C	Subóptimo	(4,14,15,21)	45.1167	Sin salida	Sin salida	Sin salida
C	Óptimo	(1,14,21)	94.0393	38.6350	34.8889	Sin salida

Los modelos que se identificaron en estos casos no presentan una disminución significativa del error con respecto a los errores obtenidos por la metodología FIR cuando utiliza para la discretización el método EFP (casos **A** y **B**) o el algoritmo FCM (caso **C**). El problema principal radica en que los modelos FIR identificados no son capaces de predecir completamente la señal de prueba, debido a que algunos de los conjuntos antecedentes que se deben predecir no se encuentran en la

base de reglas patrón. Al aumentar la complejidad de las máscaras y el número de clases en las que son discretizadas las variables, aumenta el número de estados legales del modelo. Como el número total de registros de datos se mantiene constante, la frecuencia de observación de cada estado disminuye rápidamente y como consecuencia también lo hace el poder de predicción del modelo. En estos casos el modelo se caracteriza por una alta expresividad pero presenta una reducida predictivilidad.

El mejor resultado de este estudio se obtiene con el modelo (8,14,18,21) de complejidad 4, que está representado en forma matricial en la ecuación 9, y usando el algoritmo *Ward Linkage.*

t/x	ho	ds	vv	dv	te	hu	o3	
$t - 2\delta t$	0	0	0	0	0	0	0	
$t - \delta t$	-1	0	0	0	0	0	-2	(9)
t	0	0	0	-3	0	0	+1	

El error obtenido en este caso es de 23.6871%, ligeramente inferior a los errores obtenidos con los demás algoritmos de clustering que son del orden del 25%. La figura 3 muestra la mejor predicción obtenida con un modelo "mensual".

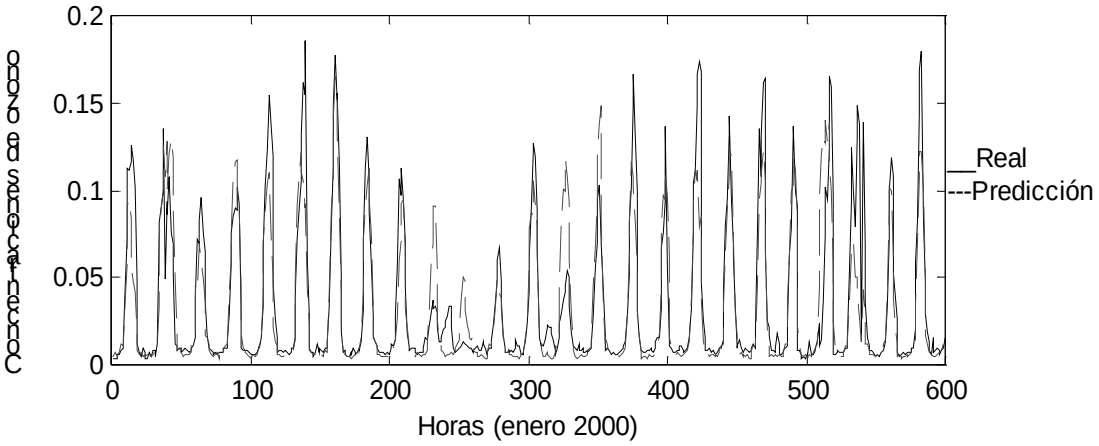

Figura 3. Predicción del conjunto de prueba con la máscara (8,14,18,21) y usando el algoritmo *Ward Linkage* (Modelo mensual: caso A).

3.3 ANALISIS DE RESULTADOS

Los algoritmos de clasificación *Complete Iinkage*, *Ward Linkage, Fuzzy C-means* y *K-means* obtienen una agrupación de los datos de las variables velocidad del viento, dirección del viento y ozono más acorde con la información inherente que la que se obtiene con el método *Equal Frenquency Partition*.

Los resultados obtenidos en el modelo "estacional" utilizando las nuevas marcas en la identificación del modelo óptimo para los casos **A** y **B** mejoraron apreciablemente con respecto a los resulta-

dos obtenidos cuando se usó el método EFP. Sin embargo, no existe diferencia significativa entre los 3 algoritmos de clustering introducidos en este estudio. Ninguno de ellos destaca sobre los demás atendiendo a su error de predicción. La comparación que se hizo en el caso **C** en donde el algoritmo inicial de agrupación fue el FCM no presentó resultados significativos con respecto a la predicción realizada con los otros algoritmos mencionados. Cabe señalar que el aumento de la granularidad en las variables más significativas no deriva en una mejora de la predicción. Por lo tanto, a igual poder de predicción es preferible seleccionar el modelo más simple que en este estudio se corresponde al mejor modelo del caso **A**.

Los resultados obtenidos en el modelo "mensual" no se pueden considerar significativos puesto que la mayoría de los modelos obtenidos con las nuevas particiones no son capaces de predecir por completo el conjunto de datos de prueba. Como ya se ha comentado anteriormente esto es debido a que aumenta la complejidad de la máscara y el número de clases (granularidad) de las variables manteniéndose constante el número de registros de la base de datos.

Por otro lado, en este trabajo se confirma que los mejores modelos identificados son los que incluyen la influencia de las variables: días de la semana, velocidad del viento, dirección del viento y ozono. Es importante comentar que a pesar de las mejoras obtenidas en las predicciones usando los mejores modelos identificados, la señal de predicción no consigue alcanzar los picos superiores de la señal de validación.

4 CONCLUSIONES

El objetivo principal de este trabajo de investigación ha sido estudiar que tanto influye el método de discretización de los datos utilizado por la metodología FIR en el pronóstico de las concentraciones de ozono a largo plazo, en la zona Centro de la Ciudad de México. Para lograrlo, se partió de un trabajo anterior [Nebot *et al.,* 2001], en el que se utiliza el método EFP en el proceso de discretización de los datos proporcionados por RAMA. Los modelos "estacionales" obtenidos (enero a mayo del 2000) por la metodología FIR utilizando marcas generadas con los algoritmos jerárquicos *Complete linkage* y *Ward Linkage* y con el algoritmo no jerárquico *K-means* fueron capaces de capturar y seguir el comportamiento del sistema con mayor precisión que los modelos cuya partición de datos se realizó con el método *Equal Frequency Partition*. Esta mejora en el seguimiento de la señal se ve reflejada en la disminución del error de predicción. Sin embargo no se logró que la señal de predicción alcanzara los picos de la señal real (los niveles peligrosos de ozono). Los resultados que se obtienen en el modelado "mensual" (enero de 1996 a enero del 2000) utilizando diferentes algoritmos de clasificación no presentan mejoras significativas.

FIR es una metodología inductiva que captura la base de reglas a partir de los datos observados y basa su predicción en el conocimiento derivado de estos. Un problema importante que se presentó, es que a medida que la complejidad de la máscara y la granularidad de las variables aumentan también se incrementa el número de estados legales. Como el número total de registros de los datos

observados permanece constante, la frecuencia de observación de cada estado disminuye rápidamente, y así la capacidad de predicción de los modelos. Este ha sido el caso en muchas de las predicciones.

5 REFERENCIAS BIBLIOGRÁFICAS

[Bezdek et al., 1984] Bezdek J., R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. Computers E Geosciences, 10(2-3): 191-203, 1984.

[Cellier *et al*., 1996] Cellier F.E., A. Nebot, F. Mugica and A. de Albornoz, Combined qualitative/quantitative simulation models of continuous-time processes using FIR tecniques. International Journal of General Systems 24 (1-2). Pp. 95-116. 1996.

[Gómez *et al*., 2001] Gómez P., A Nebot, F. Mugica, and F. Wotawa, Fuzzy Inductive Reasoning for the Prediction of Maximum Ozone Concentration. In Proceedings ESS'01: European Simulation Symposium, page 8pp., Marseille, France, 18-20 October 2001.

[Instituto Mexico & Los Alamos, 1994] Instituto Méxicano del Petróleo & Los Alamos National Laboratory, Estudio Global de la Calidad del Aire en la Ciudad de México. Final technical report – la-12699, México, 1994.

[Mucha and Sofyan, 2000] Hans-Joachim Mucha and Hizir Sofyan, Xplore Application Guide, Chapter 9:"Clustering Analysis". Electronic Book: http://www.quantlet.de/scripts/xag/htmlbook/

[Nebot *et al*., 1998] Nebot A., F.E. Cellier and M. Vallverdú, Mixed quantitative/qualitative modeling and simulation of the cardiovascular system, Computer Methods and Programs in Biomedicine 55, pp. 127-155, 1998.

[Nebot *et al*., 2001] Nebot A., F. Mugica and P. Gómez, Long term prediction of maximum ozone concentration using fuzzy inductive reasoning, EUNITE'01: European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, pp. 91-101, 2001.

[Pinyol, 2002] Pinyol I., Algorismes de Clustering per al tractament de dades, Projecte Final de Carrera de la Facultat d'Informàtica de Barcelona, UPC, 2002.

Monitoreo Atmosférico en el Valle de México.
http://www.sima.com.mx/sima/df/df.html
Instituto Nacional de Ecología
http://www.ine.gob.mx/ucamp/index.html
Laboratorio de Modelación Ambiental. http://uninet.mty.itesm.mx/lab_model/labmodam.html
http://uninet.mty.itesm.mx/lab_anali/labanali.html
Investigación y Desarrollo.
http://uninet.mty.itesm.mx/CCA/info/investades.html
Calidad del aire

http://www.calidad-del-aire-gob.mx/sima/ddf/contamin.htm1#PST
http://www.calidad-del-aire-gob.mx/mx/sima/ddf/res_imk.html
Investigación en la Universidad Autónoma de México
http://cueyatl.uam.mx/uam/publicaciones/boletines/tips/enero97/cincouno.html