Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXVII, págs. 3607-3625 (artículo completo)
Alberto Cardona, Mario Storti, Carlos Zuppa. (Eds.)
San Luis, Argentina, 10-13 Noviembre 2008

APLICACIÓN DE REDES NEURONALES ARTIFICIALES PARA LA PREDICCIÓN DE CALIDAD DE AIRE

Lucila L. Chiarvetto Peralta[a,b,c] **, Fernando A. Rey Saravia**[a,c]**, y Nélida B. Brignole**[a;d]

[a]*Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Departamento de Ciencias e Ingeniería de la Computación Universidad Nacional del Sur (UNS), Av. Alem 1253, Bahía Blanca (B8000CPB), Argentina*

[b]*Thinknet, http://www.thinknetgroup.com.ar*

[c]*Comité Técnico Ejecutivo (CTE,) Av. San Martín 3400, Municipalidad de Bahía Blanca (B8103CEV), Argentina*

[d]*Planta Piloto de Ingeniería Química (UNS - CONICET) Camino La Carrindanga Km.7 Bahía Blanca (CC717 (8000)) - Argentina*

Palabras Clave: PM_{10}, redes neuronales, medio ambiente

Resumen. Se llevó a cabo la implementación de un predictor para el promedio diario de material particulado (PM) y el diseño y desarrollo de este software es descrito en este trabajo. El daño producido por el PM en la salud humana, está relacionado con el pequeño tamaño de las partículas. Las redes neuronales artificiales (RNs) han mostrado ser un método eficiente y universal en la aproximación de funciones para cualquier tipo de dato. Una RN fue escogida porque se ha demostrado que son eficaces cuando son aplicadas a predicciones de la calidad de aire. En comparación con otros trabajos similares, sólo una red fue realizada, pero varios prototipos fueron desarrollados y evaluados para evitar la arbitrariedad en las decisiones de diseño. Se experimentaron tres aspectos en particular del diseño de una RN: la normalización de los datos, la selección de la arquitectura y la selección de la función de activación. En base a nueve variables de entrada: dos estacionales, y siete meteorológicas; se determinó que la mejor candidata es una RN compuesta por: una capa de entrada lineal de nueve neuronas artificiales (NA), una capa oculta de catorce NA y una capa de salida de una NA; ambas con una función de activación tangente hiperbólica. Durante el desarrollo de un sistema de Data Warehousing (DW) para el monitoreo y control de polución en la ciudad de Bahía Blanca (Pcia. de Buenos Aires, Argentina), el conjunto de requerimientos incluía la necesidad de contar con herramientas que permitan la predicción de las concentraciones de varios contaminantes. En el futuro, esta herramienta terminada podrá ser embebida en dicho DW. Este trabajo es el comienzo del desarrollo de un entorno de predicción más complejo que abarcará diversos contaminantes en aire.

1 INTRODUCCIÓN

La contaminación de aire es la presencia en la atmósfera de sustancias resultantes de la actividad humana o de procesos naturales en concentraciones suficientes durante un determinado tiempo y en circunstancias tales como para afectar el medio ambiente. Esta contaminación supone un problema de salud ambiental grave que afecta a países desarrollados y en vías de desarrollo de todo el mundo. En una escala global, se emiten a la atmósfera grandes cantidades de gases y partículas potencialmente nocivas, lo cual afecta el ambiente y, por ende, la salud humana. Asimismo, estas emisiones dañan los recursos necesarios para el desarrollo sustentable del planeta a largo plazo.

1.1 El PM_{10}

La Agencia de Protección Ambiental (EPA, http://www.epa.gov/oar/particlepollution/basic.html) de los Estados Unidos define el material particulado (PM) como un conjunto de partículas sólidas y líquidas presentes en la atmósfera. Esta definición no realiza ninguna distinción sobre el tamaño o composición de dichas partículas. El material particulado con una dimensión no mayor a 10 micrómetros, es especificado como PM_{10}. La medición de PM_{10} en la atmósfera se expresa en términos de concentración de masa, expresado usualmente en microgramos por metro cúbico ($\mu g/m^3$).
La norma de calidad de aire del Decreto 3395/96 reglamentario de la Ley Provincial 5965 de la provincia de Buenos Aires (publicado en el boletín oficial del día 27/9/96) establece una concentración de 150 $\mu g/m^3$ para un período de exposición diario y de 50 $\mu g/m^3$ para un año de exposición. La medida de la exposición diaria corresponde al período entre la hora cero del día 1 y la hora cero del día 2, según lo establece la normativa correspondiente. El límite de concentración que establece la norma para un período de exposición diario, fue sobrepasado en seis oportunidades en el año 2004, en once oportunidades en 2005, y en once oportunidades en 2006.
El daño producido por el PM en la salud humana, está directamente relacionado con el tamaño de las partículas (EPA, http://www.epa.gov/oar/particlepollution/health.html). El PM_{10} produce los mayores problemas, debido a que, por su tan pequeño tamaño, puede ingresar muy profundamente en los pulmones, e incluso hasta el torrente sanguíneo.

1.2 El Comité Técnico Ejecutivo (CTE)

La ciudad de Bahía Blanca está localizada en el sudeste de la República Argentina, con una población de alrededor de 300.000 habitantes. Esta localidad está rodeada por un gran polo petroquímico y un activo puerto de aguas profundas, donde los buques realizan cargas y descargas de gran cantidad de cereales.

Mediante la Ley 12.530/00 de la provincia de Buenos Aires en el artículo 10, se crea el CTE, conformado por profesionales de la ingeniería y de la química. Este ente tiene a su cargo la ejecución de aquellas tareas que llevan a lograr los objetivos enunciados en la Ley. El CTE es un organismo gubernamental, dependiente de la Municipalidad de Bahía Blanca, que tiene como principal objetivo asegurar a la comunidad una mejor calidad de vida en armonía con el progreso industrial y tecnológico, esto implica en particular efectuar un eficaz monitoreo de la calidad de aire.

Este trabajo surge como resultado de la búsqueda de requerimientos necesarios para la implementación de un sistema de Data Warehousing (DW) para el control y monitoreo on-

line de la polución ambiental, durante el período agosto de 2006 a diciembre de 2007, en el CTE como parte del diseño del montaje de un Observatorio Ambiental para la ciudad de Bahía Blanca (Rey Saravia et al., 2006).

Brevemente un DW, es una colección integrada de bases de datos personalizadas, destinadas a apoyar la función de un sistema de soporte de decisión (DSS), donde cada unidad de datos es relevante para algún momento en el tiempo. Un DSS, es un sistema que es utilizado para apoyar las decisiones de gestión. (http://inmoncif.com/library/glossary)

La extracción de los requerimientos fue realizada por la primera autora de este trabajo. Mientras el sistema de DW permitía reportar los datos obtenidos del presente y pasado, contar con información a futuro resultaba una necesidad por parte del ente gubernamental. La predicción de PM_{10} es el primer desafío a resolver. Para cumplir esta necesidad en su totalidad se requerirán predictores para monóxido de carbono (CO), ozono (O_3), dióxido de azufre (SO_2), amoníaco (NH_3) y óxidos de nitrógeno (NO_x, NO, NO_2).

2 METODOLOGÍA DE DESARROLLO

Para poder cubrir esta necesidad resulta conveniente desarrollar un modelo predictivo para un determinado contaminante por vez. Se ha puesto particular atención a la forma en la cual fue construido el predictor, porque la predicción de PM_{10} sólo constituye la primera etapa de un proyecto de gran escala. Una vez desarrollado y validado el modelo, se revisará todo el proceso por el cual fue construido, intentando extraer los conocimientos que resultaron beneficiosos, con el fin de aplicarlos a la predicción de otros contaminantes. Existen estándares de calidad que deben ser satisfechos para lograr un predictor totalmente operacional. Lograr este importante propósito requiere de amplia experimentación en la elaboración de modelos. Se eligió modelar estos fenómenos mediante redes neuronales (RNs), en principio porque permiten ser acopladas al DW brindándole valores numéricos on-line. Cabe destacar que el desarrollo de cada RN exige un entrenamiento previo en base a datos históricos disponibles. Por esta razón, se eligió una metodología modular para el desarrollo.

Las RNs han mostrado ser un método eficiente y universal en la aproximación de funciones para cualquier tipo de dato (Lek y Guégan, 1999). La función está determinada por un conjunto de variables meteorológicas, estacionales y de concentraciones de otros contaminantes al momento de la predicción que determinan el aumento o disminución de un contaminante en particular. Las RNs son de especial utilidad cuando dicha función es desconocida, son capaces de resolver complejos patrones entre la fuente de emisión y la concentración (Gardner y Dorling, 1999). Por otro lado, han mostrado ser superiores en predicción de calidad de aire en comparación a métodos tradicionalmente estadísticos (Grivas y Chaloukou, 2006).

3 UN MÉTODO BIOINSPIRADO: LAS REDES NEURONALES ARTIFICIALES

Biológicamente, las redes neuronales están compuestas por un grupo de neuronas conectadas de forma química o físicamente. La conexión química se realiza mediante el mecanismo de neurotransmisores, mientras que la conexión física, llamada sinapsis, es realizada mediante las dendritas de una neurona ligadas electrónicamente a un número extenso de axones de otras neuronas. La inteligencia artificial ha tratado de modelar simplificadamente el funcionamiento del cerebro humano; las redes neuronales artificiales, o simplemente RNs, constituyen un modelo matemático aproximado del funcionamiento cerebral.

Una RN esta compuesta por un número de nodos, unidades o neuronas artificiales (NA) vinculadas a través de enlaces. Algunas neuronas artificiales están conectadas al ambiente externo, siendo designadas como unidades de entrada de información desde el exterior o en contrapartida, unidades de salida. Las NAs además están compuestas por un conjunto de enlaces de entrada provenientes de otras NAs, y por otro conjunto de enlaces que probablemente se dirigen hacia otras NAs. Estos enlaces (análogos a la sinapsis biológica) son considerados la principal fuente de memoria a largo plazo, y tienen un peso asociado a ellos. Usualmente los algoritmos de aprendizaje modifican estos pesos aproximando las relaciones entre los datos de entrada y salida propuestos como ejemplos con el fin de realizar el entrenamiento.

Cada NA tiene asociada una función de activación que es aplicada a la suma ponderada de los valores de los enlaces (v_i) por sus pesos (p_i), permitiendo propagar el resultado de la computación (v_k) al próximo nivel. Existe una gran variedad de funciones de activación. En la Figura 1 se puede apreciar la forma en la que se lleva acabo la computación dentro de la NA.

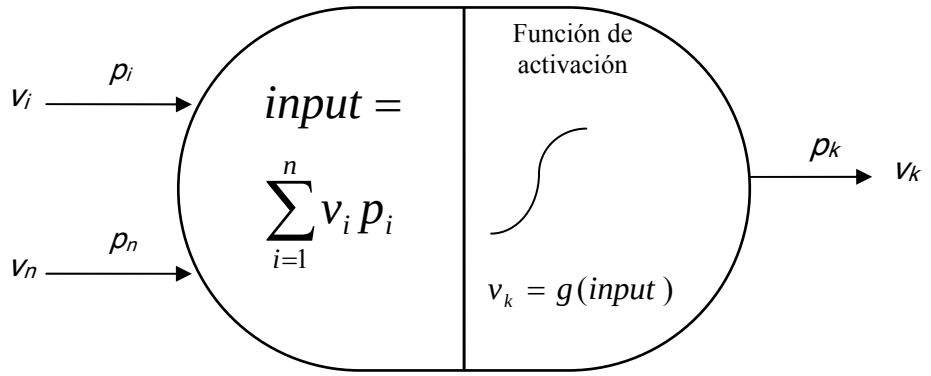


Figura 1: Nodo, Unidad o Neurona artificial.

La idea es que cada unidad realice una computación local basada en las computaciones realizadas por sus antecesoras, sin la necesidad de algún control global sobre el conjunto de unidades.

Existe una gran variedad de RNs, cada una de las cuales resulta en una variedad de resultados computacionales. La distinción principal se realiza entre las redes que son recursivas y las que no tienen esta propiedad, más conocidas por su definición en inglés feed-forward.

El diseño de una RN involucra determinar básicamente los siguientes aspectos: la selección de una topología adecuada para el problema a resolver, recursiva o no; la elección del tipo de función de activación más adecuada; el algoritmo de aprendizaje; la correcta definición de los ejemplos con los cuales se va a entrenar la RN.

Una categoría dentro de las RNs no recursivas son las RN multicapa. Estas RNs están compuestas por un conjunto de NA interconectadas formando capas, conteniendo al menos dos de ellas. Cada capa C_i esta compuesta por NA que reciben sus valores de entrada de la capa C_{i-1} y los valores producidos son enviados a la capa C_{i+1}. Usualmente en la capa C_0 o de entrada, no se realiza computo y es utilizada para ingresar los datos a la RN, análogamente la última capa C_n es utilizada para comunicar los resultados computados. La capa C_k con $0<k<n$, es llamada capa oculta.

3.1 Las RNs en la predicción de calidad de aire

En particular las RNs han sido aplicadas frecuentemente en el campo de las Ciencias Atmosféricas (Gardner y Dorling, 1998), particularmente en predicción. Han sido utilizadas para modelar y predecir concentraciones horarias de óxidos de nitrógeno (NOx), principalmente emitidos por el transito vehicular, en el área urbana de la ciudad de Londres, concluyendo que las RNs son capaces de modelar la relación local entre NO_2/NO_X y las variables meteorológicas (Gardner y Dorling., 1999). En la India, para la capital Nueva Delhi, se utilizó el mismo enfoque para modelar las emisiones de NO_2 provenientes del tráfico vehicular (Shiva Nagendra y Khare, 2005).

En otras ciudades del mundo se han realizado enfoques similares. En Helsinki, fueron utilizadas para predicción de PM_{10} y NO_2 (Kukkonen et al., 2003). En gran Atenas, durante los Juegos Olímpicos del año 2004, modelos basados en RNs fueron utilizados para predecir concentraciones horarias de PM_{10}, mostrando ser más eficaces que los modelos de regresión simple, recomendando este modelo como candidato para su utilización de forma operacional (Grivas y Chaloukou, 2005). En la ciudad de Milán, Italia, donde el O_3 y PM_{10} constituyen una de las principales preocupaciones referentes a la calidad de aire asociados con la mortalidad y las hospitalizaciones por causas cardiorrespiratorias, las RNs fueron utilizadas para un análisis de la relevancia de las distintas variables (Corani, 2005). Corani llega a la conclusión que cuando la radiación solar es alta, esta variable se transforma en el principal predictor, y cuando la misma disminuye, la temperatura y la lluvia se transforman en variables claves.

El O_3 puede tener un impacto negativo sobre la salud pública, cuando se presenta en los niveles más bajos de la atmósfera y en cantidades suficientes. Por lo que en la ciudad de Kuwait, se desarrollaron RNs para la predicción diaria de estas concentraciones (Abdul-Wahab y Al-Alawi., 2002). Lograron como conclusión principal saber que la contribución de la meteorología en la variación del ozono se encuentra en el rango del 33% y 40%, aproximadamente.

Las ciudades de Valencia y Bilbao, en España, no son la excepción. En Valencia utilizan las RNs para el estudio de relevancia de las variables utilizadas para la predicción de O3 troposférico (Gomez-Sanchis et al., 2006), mientras que en Bilbao son utilizadas para la predicción horaria de distintos contaminantes (SO_2, CO, NO_2, NO y O_3) (Ibarra-Berastegui et al., 2008).

4 PREDICCIÓN DE PM_{10}

Las RNs, como cualquier otro sistema de software, son construidas mediante un proceso. Son conocidos los beneficios de la formalización del modelo de proceso que permite la construcción de un sistema de software. La definición del MDP es uno de los principios fundamentales de la Ingeniería de Software, cuyo principal objetivo es la obtención de software de calidad.

Una ventaja adicional de la formalización del modelo del proceso de desarrollo (MPD), es que el mismo puede ser repetido, y como consecuencia natural, mejorado. Cabe destacar que la RN (producto del proceso) no puede ser utilizada en forma confiable para predicciones en distintas condiciones de tiempo y espacio, para la cual, fue entrenada. Lo que sí puede ser reutilizado es el MPD que permitió su creación, en nuevas condiciones de tiempo (otro

intervalo) y espacio (distinto sitio geográfico). Por eso resulta tan importante la formalización del MPD, de esta forma también se abaratarían costos de desarrollo (otro objetivo, propuesto por la Ingeniería de Software).

El modelo de desarrollo propuesto está enfocado en el desarrollo de RNs multicapas con una capa oculta en particular. Si bien existen otras posibles arquitecturas, la arquitectura multicapa ha mostrado ser útil en la predicción de calidad de aire (Gardner y Dorling., 1998, 1999; Abdul-Wahab y Al-Alawi., 2002; Kukkonen et al., 2003; Shiva Nagendra y Khare., 2005; Grivas y Chaloukou, 2005; Corani, 2005; Gomez-Sanchis et al., 2006; Ibarra-Berastegui et al., 2008). El motivo antes mencionado es por el cual nuestros esfuerzos fueron concentrados en este tipo de arquitectura en particular, sin embargo no descartamos en futuros desarrollos intentar el mismo enfoque con otras arquitecturas.

El proceso de desarrollo definido para este sistema en particular consiste de las siguientes etapas:

1. Selección de las variables de entrada
2. Selección de los métodos de evaluación
3. Normalización
4. Selección de la arquitectura: cantidad de NA para capa oculta
5. Selección de la función de activación
6. Selección del algoritmo de aprendizaje
7. Entrenamiento y validación de RN
8. Evaluación operacional del modelo
9. Evaluación científica: conclusiones
10. Articulación al sistema de DW

Como resultado de una revisión bibliográfica previa, se obtuvo un conjunto de las distintas heurísticas y técnicas que les han resultado productivas a otros autores en la predicción de calidad de aire (mediante RNs). Para cada una de las heurísticas correspondientes a cada etapa del desarrollo, se construirá un prototipo para evaluar cual de ellas resulta en nuestro caso más efectiva para nuestro contexto. De esta forma, se evita la arbitrariedad en el diseño y también se pretende asegurar la calidad en el producto.

La selección de las variables de entrada está limitada a la disponibilidad de las mediciones realizadas. En nuestro caso no se cuenta con mediciones tomadas de radiación solar, la cual sería una variable de utilidad, por lo tanto no puede ser incluida.

En las etapas comprendidas entre 3 y 6, las posibles decisiones de diseño son evaluadas. La evaluación se realizará mediante prototipos. Si bien estos prototipos son RNs operacionales, no deben ser considerados como RNs funcionales. El propósito de su construcción es evaluar y comparar que tan beneficiosas resultan en si mismas las distintas decisiones de diseño. Por este motivo no se realiza ninguna técnica de validación de los prototipos. La construcción del prototipo en la etapa $I+1$ se hará sobre los resultados obtenidos en la etapa I.

En la etapa 7, las decisiones de diseño se encuentran justificadas, y es posible comenzar la construcción de la RN funcional, como técnica de validación se utilizara validación cruzada (en inglés, cross validation).

4.1 Datos y materiales

La Estación de Monitoreo Continuo de aire de Bahía Blanca (EMCABB) es una cabina móvil equipada con analizadores de los contaminantes de aire, en particular PM_{10}, y una estación meteorológica. Mediante este equipo, se puede realizar un monitoreo continuo de la calidad de aire, para determinar los niveles base de la zona en la cual se encuentra emplazada. (http://www.bahiablanca.gov.ar/cte/emcabb.html).

Los datos utilizados en este trabajo, fueron los obtenidos por la EMCABB, durante el período del año 2004 al año 2006 inclusive. Los promedios diarios fueron obtenidos en base a los promedios horarios. Para el calculo de los promedios diarios no fueron tenidos en cuenta aquellos días que tuvieran faltantes en sus promedios horarios. Como resultado se obtuvo un conjunto de datos de 554 elementos. Del conjunto antes mencionado se dividió de forma aleatoria, en tres subconjuntos para distintas utilidades a lo largo del desarrollo del predictor. El primero de estos subconjuntos con una cardinalidad del 70 % (388 patrones) respecto al conjunto total, esta destinado al entrenamiento de la red. El segundo y tercero de estos subconjuntos, con 83 patrones cada uno de ellos (esto es una cardinalidad del 15% respecto al conjunto total), están destinados a la validación y evaluación de la RN respectivamente. En el tratamiento de los datos se utilizó como motor de base de datos a Microsoft SQL Server 2005 Express Edition®.

Los prototipos fueron realizados mediante un editor gráfico provisto por un framework para la creación de RNs, llamado JOONE (Java Object Oriented Neural Engine, www.jooneworld.com) basado en el lenguaje de programación Java© provisto por Sun Microsystems® (www.java.sun.com). Sin embargo, para la construcción de la RN functional, se utilizará la librería provista por JOONE, en el entorno de programación NetBeans (www.netbeans.org).

5 PROCESO DE DESARROLLO

5.1 Definición del predictor

El objetivo de la RN, es la predicción de promedios diarios de PM_{10} para la ubicación de la EMCABB en el barrio de Villa Delfina en la ciudad de Bahía Blanca. La utilización de la RN entrenada con datos para un sitio e intervalo de tiempo en particular, sólo pueden ser utilizadas de forma confiable para predicciones en dichas condiciones. Cualquier predicción en otro intervalo de tiempo o lugar, no serían confiables (Fox, 1981; Gardner y Dorling, 1998).

5.2 Selección de las variables de entrada

Las variables de entrada se dividen en dos conjuntos: meteorológicas y estacionales. Las distintas variables meteorológicas son: dirección predominante del viento (DV), velocidad promedio del viento predominante (VV), velocidad promedio (V, teniendo en cuenta todas las direcciones), velocidad máxima promedio (V_{max}, teniendo en cuenta toda dirección posible), temperatura ambiente promedio (T), presión promedio (P), humedad promedio (H). Estos promedios fueron calculados sobre los promedios horarios. Las variables estacionales son: mes del año (M) y día de la semana (DS); estas variables intentan capturar la influencia de los cambios estacionales y el tránsito vehicular sobre el PM_{10}. Estas variables están representadas de forma escalar.

Grivas y Chaloukou consideran a DS, dentro de las variables más útiles en el caso de la predicción horaria de PM_{10} para Gran Atenas. Si bien en el presente trabajo las predicciones no son para intervalos horarios, el resultado obtenido por Grivas y Chaloukou será considerado como una heurística en el caso de la selección de variables. En la Figura 2 se muestra la relación local entre los promedios diarios de PM_{10} y los días de la semana, durante el fin de semana se puede notar un descenso de dichos promedios. En la Figura 3 se muestra como varía el PM_{10} respecto cada mes del año.

En la Figura 4 hasta la Figura 10, se muestra la relación entre el PM_{10} y las distintas variables meteorológicas, en estos gráficos se puede notar cómo la variación de estas variables tienen un impacto en la variación del PM_{10}, estas relaciones justifican su inclusión en el desarrollo del predictor.

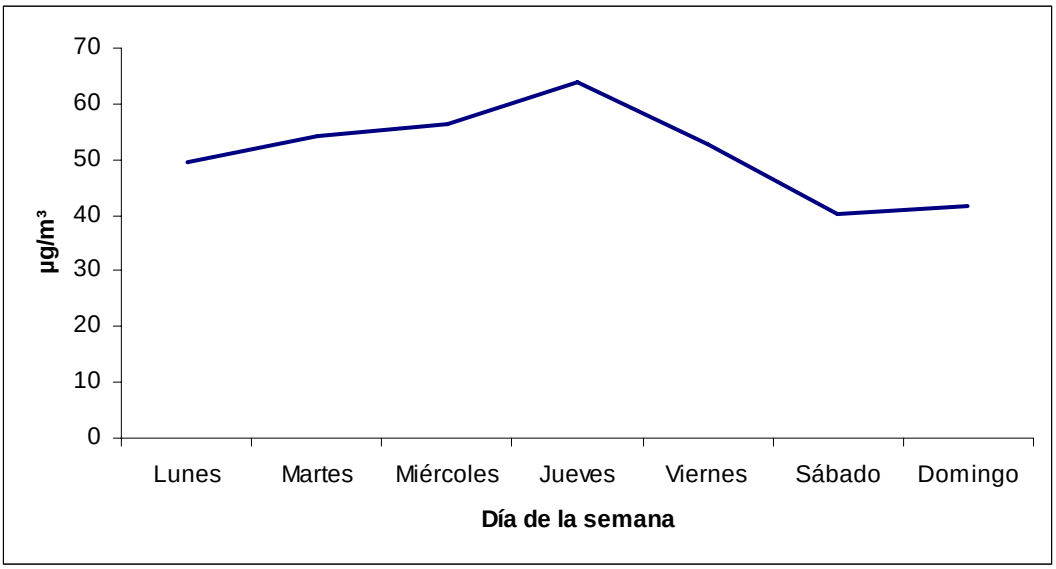


Figura 2: Relación entre los promedios diarios de PM_{10} y DS. Este gráfico muestra intuitivamente cómo las variaciones entre el fin de semana (sábado y domingo) y los días laborales (lunes a viernes).

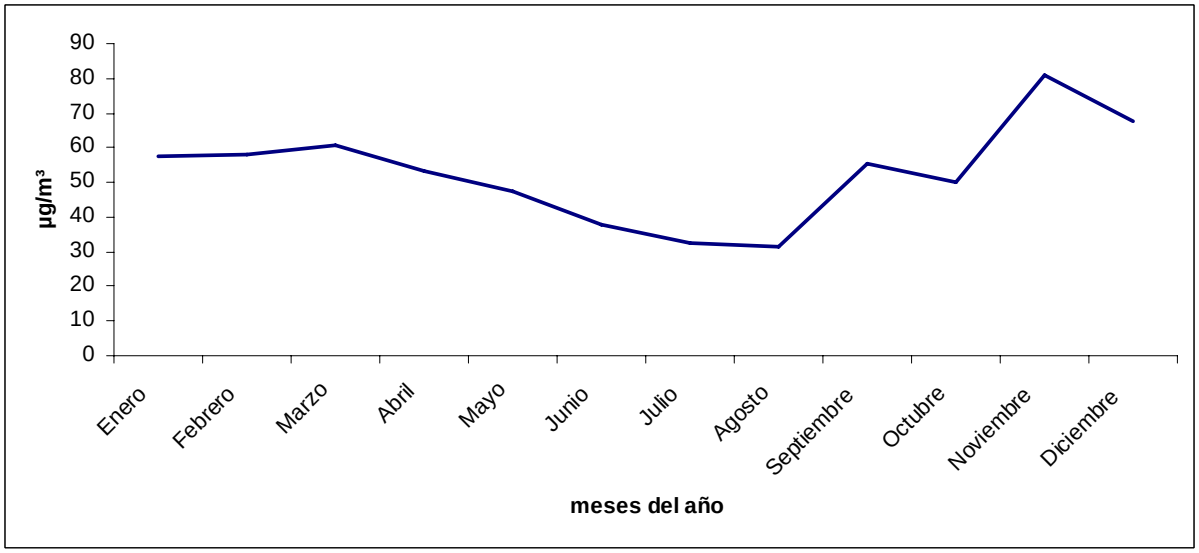


Figura 3: Relación entre los promedios diarios de PM_{10} y M. Este gráfico muestra implícitamente como las

variaciones estacionales.

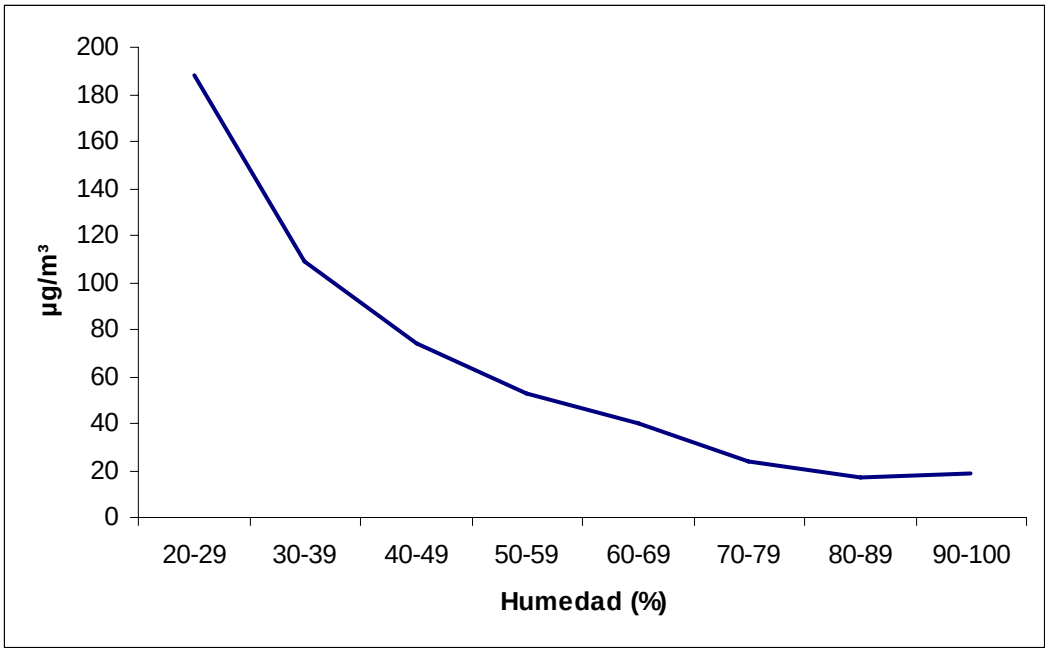


Figura 4: Relación entre los promedios diarios de PM_{10} y H agrupados por intervalos de 10 unidades.

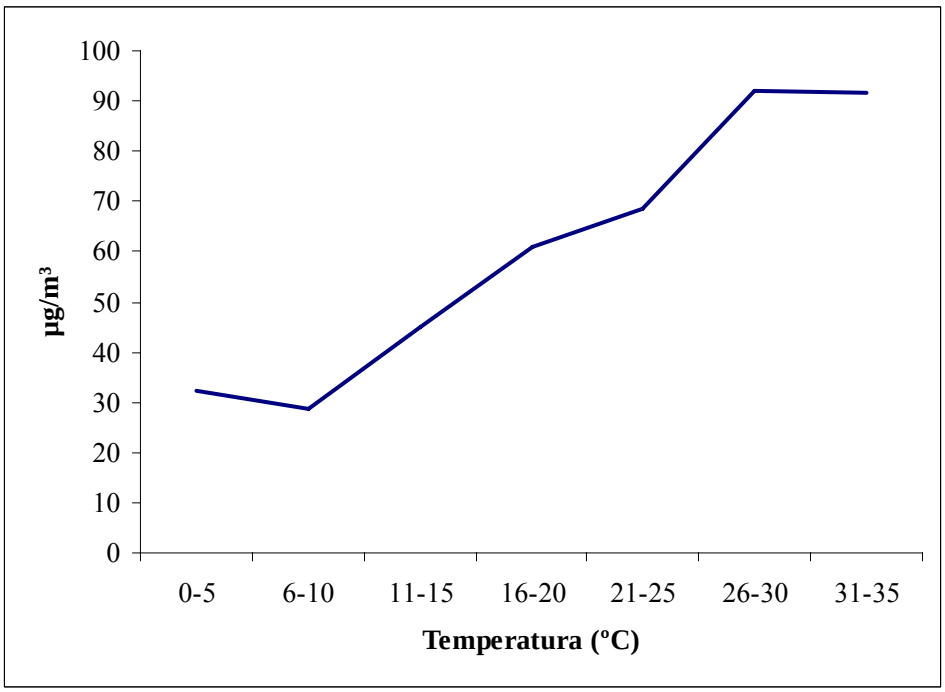


Figura 5: Relación entre los promedios diarios de PM_{10} y T agrupados por intervalos de 5 grados centígrados.

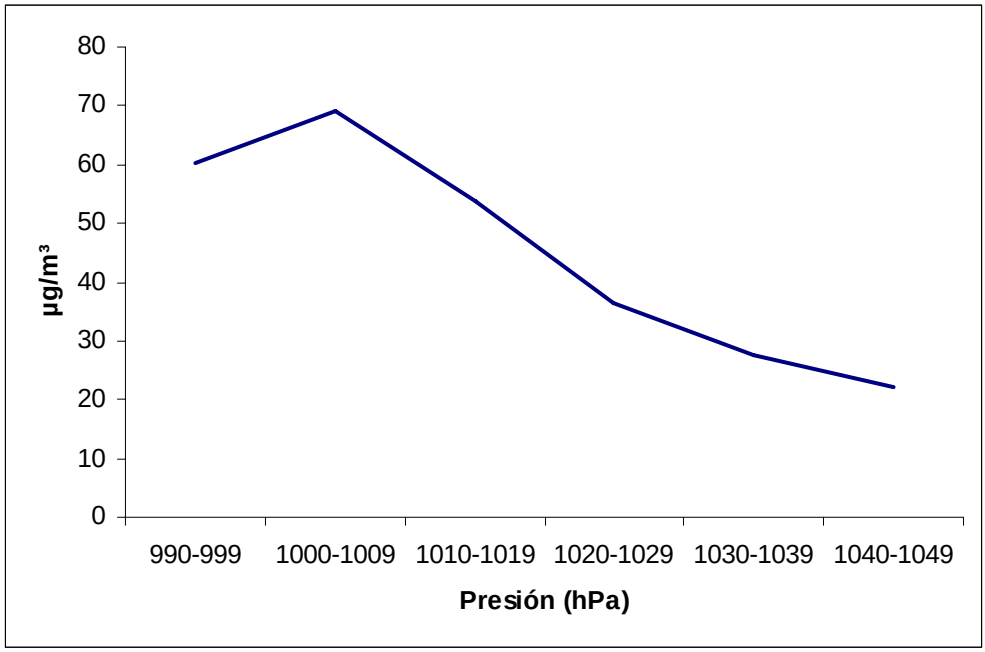


Figura 6: Relación entre los promedios diarios de PM_{10} y T agrupados por intervalos de 10 hectopascales (hPa).

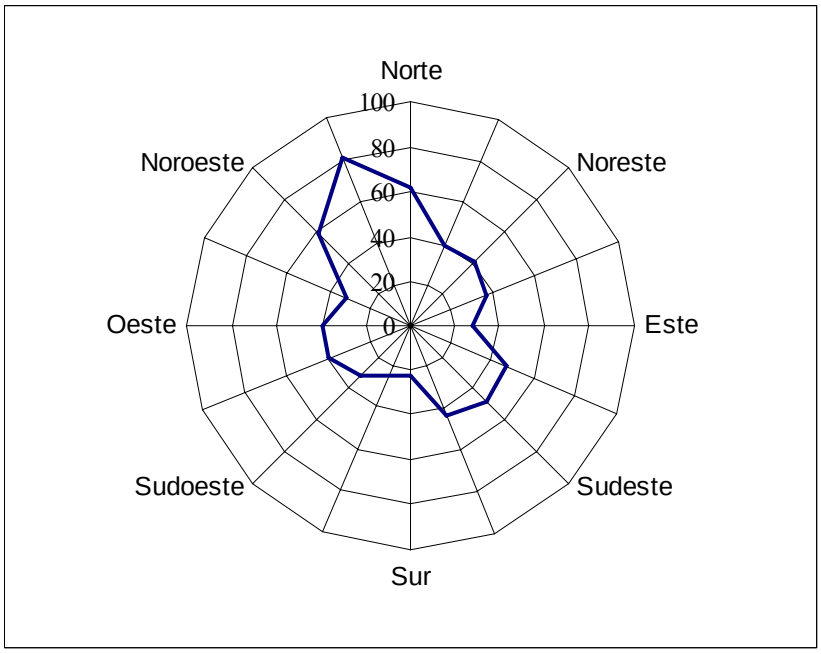


Figura 7: Rosa de los vientos. Relación entre los promedios diarios de PM_{10} y DV.

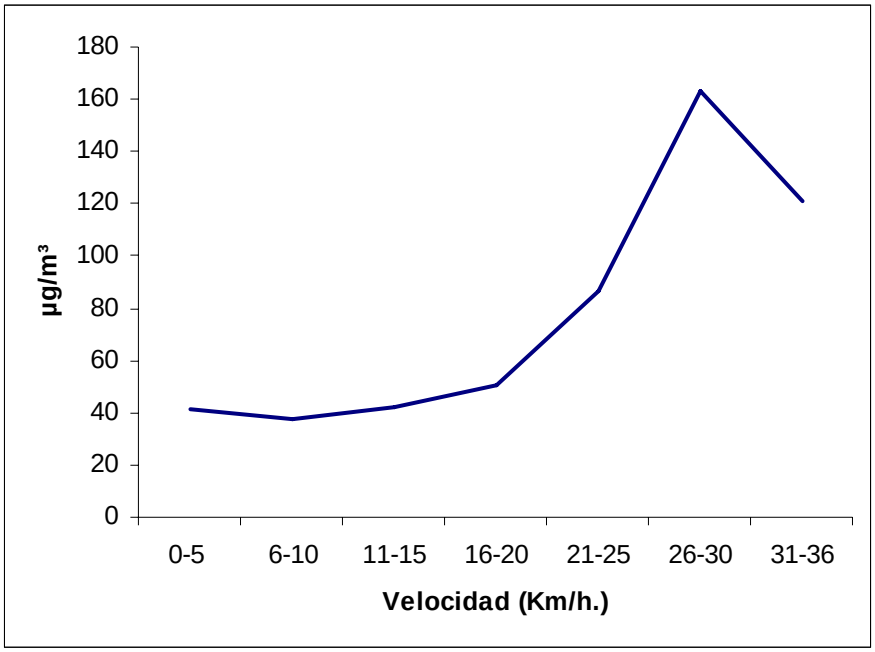


Figura 8: Relación entre los promedios diarios de PM_{10} y V agrupados por intervalos de 5 km. por hora.

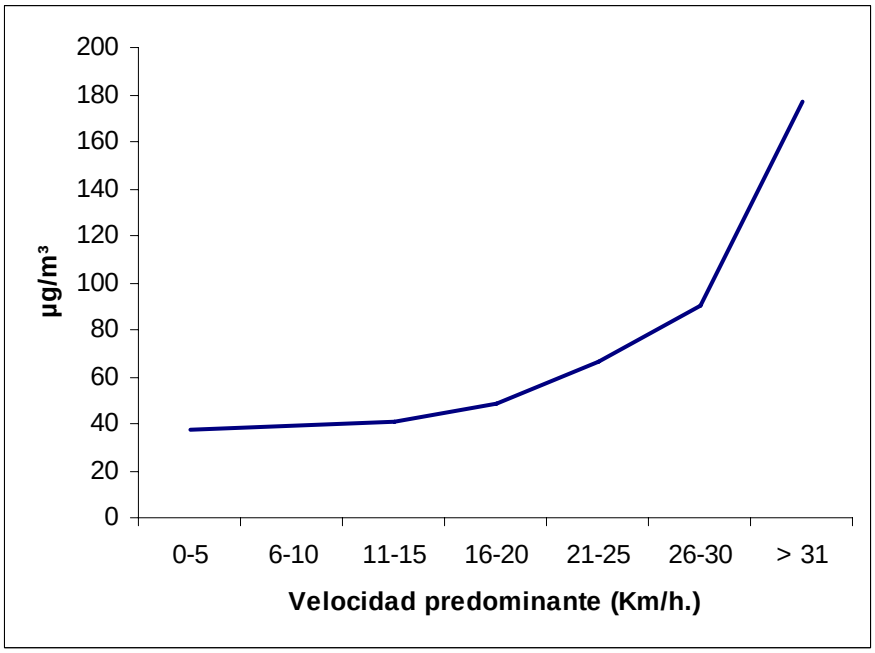


Figura 9: Relación entre los promedios diarios de PM_{10} y VV agrupados por intervalos de 10 km. por hora.

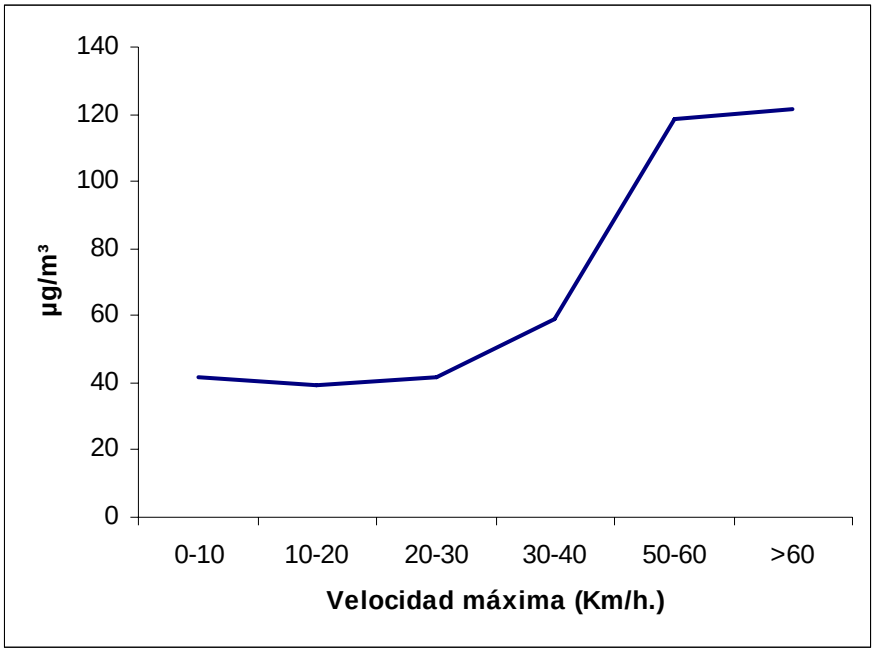


Figura 10: Relación entre los promedios diarios de PM_{10} y V_{max} agrupados por intervalos de 10 km. por hora.

5.3 Métodos de evaluación de los modelos

Como sugiere Fox (1981), se debe distinguir entre la evaluación operacional y la evaluación científica, de la performance del modelo de calidad de aire. La evaluación científica implica algún entendimiento sobre la relación causa-efecto que subyace en el modelo, mientras que la evaluación operacional implica un análisis respecto al conjunto de datos en un contexto de aplicación particular. En este caso, dicho contexto implica la utilización de métodos de evaluación acordes a las RNs. Por otro lado, se debe diferenciar entre la evaluación operacional de los prototipos y la RN funcional.

- Evaluación operacional de prototipos: para que la evaluación de los prototipos sea confiable, éstos deben ser desarrollados bajo la mayor igualdad de condiciones posible: mismos pesos iniciales, misma cantidad de épocas (epochs, en inglés), mismo algoritmo de entrenamiento (en el caso de no esté siendo evaluado), mismo conjunto de datos, etc. Además, se debe contar con un método de evaluación que permita comparar entre distintas implementaciones de RNs. Una medida de evaluación descriptiva, es el índice de acuerdo (en inglés, index of agreement) o *d*, esta medida ha sido definida por Willmott (1982), y es utilizada por otros autores (Corani, 2005; Abdul-Wahab y Al-Alawi, 2002; Shiva Nagendra y Khare, 2006; Grivas y Chaloukou, 2006).
  El índice de acuerdo, se define como:

$$d = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (|P_i - \overline{O}| + |O_i - \overline{O}|)^2}, \qquad 0 \le d \le 1 \quad (1)$$

  Donde:

- N es la cantidad de datos observados
- P_i es el i-ésimo valor predicho
- O_i es el i-ésimo valor observado
- $\overline{O}$ es el promedio de todos los valores observados

El índice de acuerdo, puede tomar valores en el intervalo [0,1]. Donde d=1 implica una perfecta correspondencia entre el valor observado y el valor predicho, contrariamente d=0 representa un desacuerdo total.

Aunque RMSE (root mean square error) y MAE (mean absolut error) se encuentran entre las mejores medidas de performance (Willmott, 1982). Sin embargo en el caso de ser necesarias para la comparación entre modelos no siempre pueden ser aplicables dado que están sujetas a una unidad de medida.

- Evaluación operacional de la RN: esta evaluación es la que se realizará a la RN una vez que la misma esté construida. Esta evaluación utilizará las medidas *d*, RMSE y MAE, pero también incluirá otras medidas de análisis más específicas.

5.4 Normalización de los datos

Diferentes formas de normalización son utilizadas en la implementación de RNs. Se evaluaron tres fórmulas distintas de normalización, con la intención de encontrar cual de ellas resulta más efectiva. Esta efectividad debe ser leída en términos de habilidad de predicción.

Las fórmulas evaluadas fueron:

$$N_1(valor) = \frac{valor - \overline{X}}{S} \quad (2)$$

$$N_2(valor) = \frac{valor - \overline{X}}{\overline{X}} \quad (3)$$

$$N_3(valor) = 2 * \left(\frac{(valor - x_{\min})}{(x_{\max} - x_{\min})} \right) - 1.0 \quad (4)$$

Donde:

- $\overline{X}$ es la media muestral
- S es el desvío muestral
- x_{min} es el menor valor de la variable
- x_{max} es el mayor valor de la variable

Las ecuaciones (2) y (3) son las formas más naturales de pensar en la normalización, por su uso estadístico. La ecuación (4) fue utilizada por Gardner y Dorling (1999). La representación de la dirección del viento merece una mención especial. Dos opciones fueron consideradas:

- la dirección del viento como ángulo medido en radianes, a la cual posteriormente se le aplico la función trigonométrica coseno. Esta representación fue utilizada por Grivas y Chaloukou; también por Shiva Nagendra y Khare. (2006).

- La dirección del viento como una variable escalar, que puede tomar los valores: norte, nornordeste, estenordeste, este, etc. Siendo un total de 17: cada dirección de la rosa de los vientos y la calma.

Como producto de las tres fórmulas de normalización propuestas y las dos formas de representar el viento, seis opciones de normalización fueron evaluadas. Todas ellas fueron calculadas sobre la misma RN, la misma arquitectura, idénticos pesos iniciales y el mismo algoritmo de entrenamiento. Esto prototipo puede ser caracterizado por:

- Una capa de entrada con una función de activación lineal.
- Una capa oculta y una de salida, con la tangente hiperbólica como función de activación. Esta función tiene un conjunto imagen [-1,1]
- El algoritmo de entrenamiento utilizado fue Back Propagation, con una tasa de aprendizaje de 0.3 y un momento (en inglés, momentum) de 0.4, durante 1500 épocas (epochs, en inglés).

Como resultado, no se observó una diferencia significativa entre la representación escalar del viento y la representación mediante el coseno del ángulo, esto puede observarse en la Figura 11. Sin embargo, se aprecia una notable diferencia en el desempeño de las funciones de normalización. La ecuación (4) resulto la más efectiva, la evaluación concluyó d=0,85844077 para la representación escalar del viento, y la representación angular d=0,82919493. En la figura 12 se muestran las distintas medidas de evaluación utilizadas, es deseable que RMSE y MAE tiendan a cero, mientras d debe tender a uno. La diferencia tan marcada entre las funciones de normalización, se debe principalmente al dominio de la función de activación, en el caso de la tangente hiperbólica su dominio que se encuentra entre [-1, 1], que coincide con la imagen de (4). Esta consideración será tenida en cuenta para futuros desarrollos.

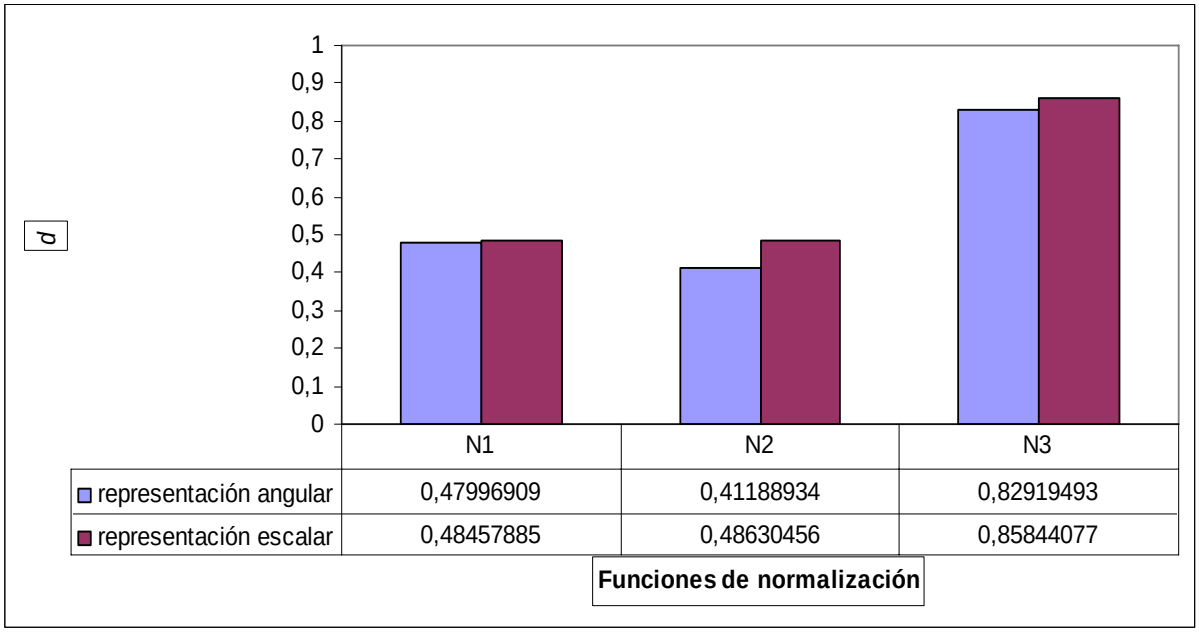
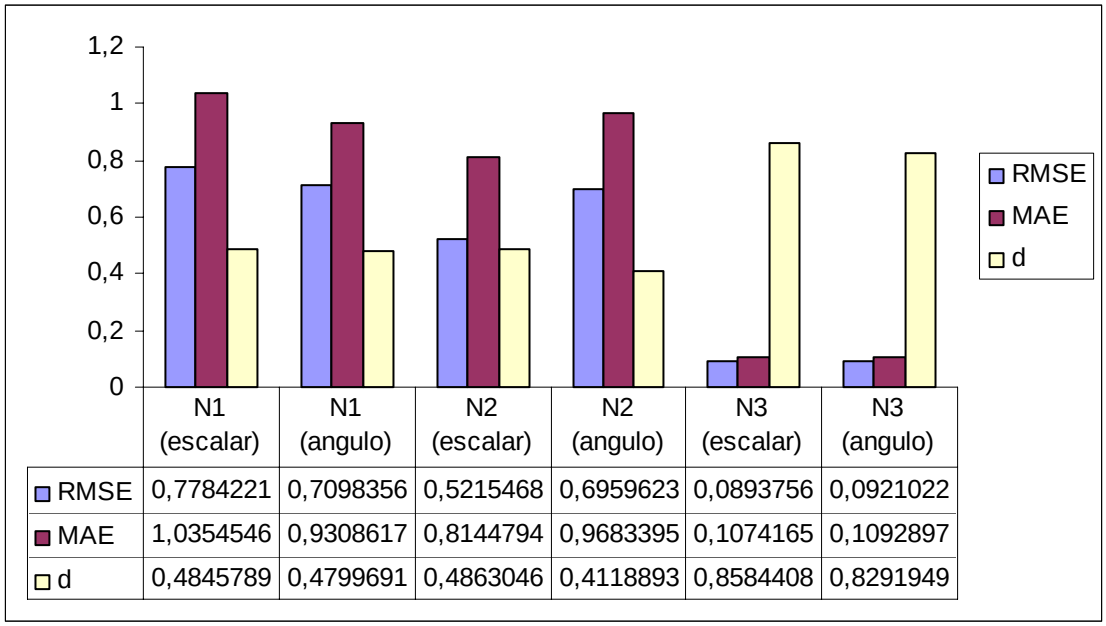


Figura 11: Comparación entre la representación angular y la representación escalar.



	N1 (escalar)	N1 (angulo)	N2 (escalar)	N2 (angulo)	N3 (escalar)	N3 (angulo)
RMSE	0,7784221	0,7098356	0,5215468	0,6959623	0,0893756	0,0921022
MAE	1,0354546	0,9308617	0,8144794	0,9683395	0,1074165	0,1092897
d	0,4845789	0,4799691	0,4863046	0,4118893	0,8584408	0,8291949

Figura 12: resultado de la evaluación de los distintos prototipos. Es deseable que RMSE y MAE tiendan a cero, mientras *d* debe tender a uno

5.5 Selección de la arquitectura

La selección de la cardinalidad de las capas ocultas (en caso de ser más de una) parece ser una cuestión poco clara al momento del diseño de una RN. Demasiadas unidades pueden llevar a poca capacidad de generalización. Por otro lado, pocas unidades pueden llevar a que la RN no posea la capacidad suficiente para resolver el problema en cuestión.
Nagendra y Khare proponen tres heurísticas distintas para la elección de la cantidad de NA en la capa oculta:

- H_1= número de neuronas de entrada + neuronas de salida
- H_2= el máximo número de neuronas en la capa oculta, es dos veces el número de neuronas en la capa de entrada.
- H_3= el número de los patrones de entrenamiento dividido por cinco veces el número de neuronas de entradas y el número de neuronas de salida.

El número de variables de entrada corresponde a la cantidad de neuronas en la capa de entrada, análogamente, la cantidad de neuronas en la capa de salida corresponde a la cantidad de variables que se quieren ser aproximadas. Aplicando estas heurísticas a nuestro problema se obtuvo:

- H_1= 9 unidades + 1 unidad = 10 unidades
- H_2= 9 unidades * 2 =18 unidades
- H_3= $\dfrac{388 \text{ de los patrones de entrenamiento}}{5 * (9 \text{ unidades} + 1 \text{ unidad})} = 7.76 \approx 8$

Estas heurísticas nos marcan un mínimo de 8 unidades y un máximo de 18 unidades. Por lo que se construyeron prototipos variando la cantidad de NA en su capa oculta, esta variación corresponde al intervalo [8, 18]. Estos prototipos pueden ser caracterizados por:

- Una capa de entrada con una función de activación lineal.
- Una capa oculta y una de salida, con la tangente hiperbólica como función de activación. Esta función tiene un conjunto imagen [-1,1]

- El algoritmo de entrenamiento utilizado fue Back Propagation, con una tasa de aprendizaje de 0.3 y un momento de 0.4, durante 500 épocas.
- La función de normalización para los datos expresados por la ecuación (4) con la representación escalar del viento. Esta normalización resulto ser la más efectiva en la etapa anterior.

En la Tabla 1 se muestran los resultados obtenidos para cada prototipo, en la Figura 13 se muestra el índice de acuerdo para cada prototipo correspondiente. El criterio utilizado en la selección para los prototipos en esta etapa, consiste de la menor cantidad de neuronas que lleven al mayor índice de acuerdo de predicción. Aplicando el criterio anterior la mejor opción resulto ser el prototipo con 14 unidades en la capa oculta.

Cantidad de neuronas en la capa oculta	RMSE	MAE	*d*
8	0,091167204	0,109301952	0,830914507
9	0,101671739	0,118752893	0,811406344
10	0,094818407	0,112522168	0,812211078
11	0,092700519	0,11062	0,822816594
12	0,094497533	0,113076684	0,828657155
13	0,090111593	0,10717084	0,842615694
14	**0,089528757**	**0,107073494**	**0,844807878**
15	0,094011871	0,112402791	0,817463968
16	0,095056689	0,114384107	0,827291684
17	0,092382896	0,108489059	0,831875015
18	0,095171846	0,113319794	0,8097027

Tabla 1: Resultado de la evaluación de los distintos prototipos para la cantidad de NA en la capa oculta. Es deseable que RMSE y MAE tiendan a cero, mientras *d* debe tender a uno.

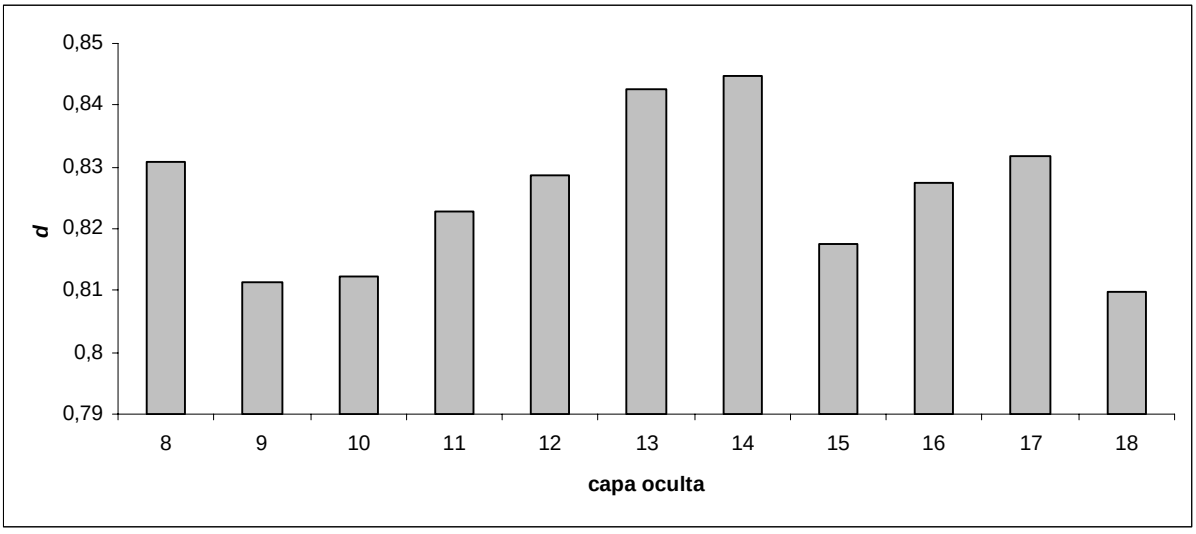


Figura 13: Resultado de la evaluación de los distintos prototipos para la cantidad de NA en la capa oculta, según índice de acuerdo.

5.6 Selección de la función de activación

La selección de función de activación en la bibliografía, muestran dos opciones reiteradas: la tangente hiperbólica y sigmoidea. La tangente hiperbólica fue utilizada por Shiva Nagendra

y Khare, Gardner y Dorling; mientras que Grivas y Chaloulakou utilizaron ambas. En la etapa de selección de normalización se concluyo que resultaba en nuestro caso más efectiva la función de normalización expresada por la ecuación (4) con la representación escalar del viento, esta representación mapea los valores al rango [-1, 1]. Dado que el dominio de la función sigmoidea se encuentra entre [0, 1] resulta claro que esta función de activación tendrá un desempeño notablemente menor a la tangente hiperbólica, teniendo en cuenta que el dominio de la tangente hiperbólica coincide con la normalización elegida. En base a este análisis concluimos en la utilización de la tangente hiperbólica como función de activación.

5.7 Selección del algoritmo de aprendizaje

En las etapas anteriores del desarrollo, se utilizó como algoritmo de aprendizaje Back-propagation, con el cual se entrenaron a los prototipos. En futuros desarrollos serán evaluados otros dos algoritmos provistos por JOONE:

- Back-propagation por lotes.
- Resilient back-propagation.

El objetivo de la comparación entre los distintos algoritmos, es determinar cual resulta más efectivo en términos de capacidad de predicción. Se puede encontrar detalles de sobre los mismos en la documentación provista por los desarrolladores en: http://ufpr.dl.sourceforge.net/sourceforge/joone/JooneCompleteGuide.pdf.

6 RESULTADOS Y DISCUSIÓN

Si hubiésemos construido una RN para cada combinación de todas las posibles decisiones de diseño, se hubiesen construido 132 RNs (seis opciones de normalización, once opciones en la cardinalidad de la capa oculta, dos funciones de normalización y tres algoritmos de aprendizaje). En lugar de esto, solo 22 prototipos serán construidos y una RN. De esta forma el MPD reduce los costos de desarrollo de SW.

A este momento de desarrollo, podemos justificar la forma en la que será construida la RN. La misma será:

- La función de normalización (4) con una representación escalar del viento.
- Una capa de entrada lineal de nueve NAs
- Una capa oculta de catorce NAs con la función de activación tangente hiperbólica.
- Una capa de salida de una NA con una función de activación tangente hiperbólica.

Queda pendiente la selección del algoritmo que será utilizado para el aprendizaje de la misma. El presente trabajo se define el MPD que es utilizado en la construcción del predictor. Si bien el proceso no fue concluido, se aprecian mejoras que pueden ser aplicadas al mismo. La primera observación refiere a las etapas tres y cinco. La selección de la función de normalización debe ser coherente con la función de activación, motivo por el cual estas etapas deben ser fusionadas en una. Un par ordenado (F_n, F_a) donde F_n: $D_i \rightarrow I_i$ es una función de normalización y F_a: $D_j \rightarrow I_j$, el requisito mínimo para ser considerado apto de evaluación es: $D_i \subseteq D_j$ y $I_i \subseteq I_j$.

Se plantean las siguientes dudas respecto a la construcción de los prototipos: ¿debería utilizarse alguna técnica de validación cuando son desarrollados? Esto aumentaría el tiempo de desarrollo, por tanto los costos del mismo. ¿Resulta beneficioso? Para contestar esta pregunta, debería repetirse todo el proceso, i.e. volver a construir el predictor, utilizando por ejemplo, validación cruzada como técnica de validación.

Sería deseable que el proceso permita la construcción de redes multicapa de dos capas ocultas, estas redes suelen ser menos propensas al sobre ajuste (overfitting, en inglés).

7 CONCLUSIONES

Se ha logrado diseñar prototipos eficientes para predecir el contaminante PM_{10} presente en la atmósfera de la localidad de Bahía Blanca. La capacidad de predicción proporcionada por esta herramienta resultaría de utilidad para la comunidad porque permitiría tomar medidas precautorias para la protección de la salud de la población. La metodología elegida en este trabajo servirá como base de desarrollo para otros predictores de distintos contaminantes de interés.

Se planteó una red neuronal con nueve variables de entrada: dos estacionales, y siete meteorológicas. Luego, se desarrollaron distintos prototipos con el objetivo de determinar la forma de representación más conveniente, la arquitectura más ventajosa y la función de activación más fructífera. Se observó que un buen diseño es fundamental para lograr un grado aceptable de certeza en las predicciones. Para evaluar el desempeño de la red, se emplearon medidas de evaluación de performance de modelos de calidad de aire. Los resultados concluidos de la normalización de los datos, la selección de la arquitectura y la selección de la función de activación, resultan promisorios. Queda pendiente la construcción de la RN. Su evaluación como modelo determinará si el predictor es confiable para su integración en el sistema DW, convirtiéndose de está forma en un predictor de utilidad práctica.

Una vez construidas las RNs, se plantean para ellas distintas utilidades diferentes. Por un lado, como predictores que publicarán sus resultados, de forma pública y privada. Como contraparte, como componente principal de un sistema experto off-line que permita el estudio local de los efectos meteorológicos y estaciónales sobre la dispersión de los contaminantes in situ.

8 REFERENCIAS

Fox D. G. Judging Air Quality Model Performance. *Bulletin American Meteorological Society*, 62:599–561, 1981.

Willmott C. J. Some comments on the Evaluation of Model Performance. *Bulletin American Meteorological Society*, 63:1309–1313, 1982.

Gardner M. W., Dorling S. R.. Artificial Neuronal Networks (The Multilayer Perceptron)—A Review Of Applications In the Atmospheric Sciences. *Atmospheric Enviroment*, 32:2627-2636, 1998.

Lek S., Guégan. J. F. Artificial neuronal networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120:65-73, 1999.

Gardner M. W., Dorling S. R.. Neuronal Network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Enviroment*, 33: 709-719, 1999.

Abdul-Wahab S. A., Al-Alawi S. M.. Assessment and prediction of tropospheric ozone concentration levels using artificial neuronal networks. *Enviromental Modelling & Software*, 17:219-228, 2002.

Kukkonen J., Partanen L., Karppinen A., Ruuskanen J., Junninen H., Kolehmainen M., Niska H., Dorling S., Chatterton T., Foxall R., Cawley G.. "Extensive evaluation of neuronal networks models for the prediction of NO2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki". *Atmospheric Enviroment*, 37:4539-4550, 2003.

Corani G.. Air quality prediction in Milan: feed-forward neuronal networks, pruned neuronal networks and lazy learning. *Ecological Modelling*, 185:513-529, 2005.

Shiva Nagendra S. M., Khare M.. Artificial neuronal network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. *Ecological Modelling*, 190:99-115, 2006.

Grivas G., Chaloulakou A.. Artificial neuronal network model for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric Enviroment*, 40;1216-1229, 2006.

Gómez-Sanchis J., Martín-guerrero J. D., Soria-Olivas E., Vila-Frances J., Carrasco J. L., del Valle-Tascón S.. Neuronal Networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration. *Atmospheric Enviroment*, 40:6173-6180, 2006.

Rey Saravia F.A., Puliafitto E., Brignole N.B.. "Mounting an Environmental Observatory in an Urban-Industrial Area", XXII-CIIQ-2006: XXII Congreso Interamericano de Ingeniería Química y V Congreso Argentino de Ingeniería Química, Buenos Aires, Argentina, 1 al 4/10/2006.

Ibarra-Berastegui G., Elias A., Barona A., Saenz J., Ezcurra A., Diaz de Argadoña J.. From diagnosis from prognosis for forecasting air pollution using neuronal networks: Air pollution monitoring in Bilbao. *Enviromental Modelling & Software*, 23:622-637, 2008.