UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE CIENCIA

Departamento de Física

Pronóstico de concentraciones de material particulado PM2,5 en la ciudad de Temuco

Marcelo Eduardo Navarro Linares

Profesor Guía:

Patricio Pérez Jara

Tesis para optar al Título Profesional de Ingeniero Físico.

Santiago – Chile

2017

“Pronóstico de concentraciones de material particulado PM2,5 en la ciudad de Temuco”

Marcelo Eduardo Navarro Linares

Este trabajo de graduación fue preparado bajo la supervisión del profesor guía Dr. Patricio Pérez J. del departamento de física y ha sido aprobado por los miembros de la comisión calificadora del candidato,…

……..
Dr. Ernesto Gramsch

……..
Dr. Raúl Cordero

……..
Dr. Luis Díaz

……..
Dr. Enrique Cerda
Director

……..
Dr. Patricio Pérez
Profesor Guía

Resumen

La ciudad de Temuco en Chile, actualmente es una de las ciudades del país con mayor niveles de contaminación atmosférica, por lo tanto es importante implementar políticas públicas en favor de la descontaminación del aire. Uno de estos, es el plan de descontaminación implementado el año 2014,- Para que este plan de descontaminación funcione de manera preventiva y no reactiva es necesario un método de pronóstico del material partículado fino PM2,5.

Con estos antecedentes se construyó un modelo predictivo de material particulado fino PM2,5 para la ciudad de Temuco (incluyendo la comuna vecina de Padre las Casas) utilizando redes neuronales artificiales. Para esto, se determinó la estructura óptima de la red, identificando las variables de entrada que mayor efectividad otorgan al pronóstico considerando las características ambientales y sociales de la ciudad.

Además se determinó mediante el criterio de información de Akaike la cantidad óptima de neuronas en la capa oculta y se utilizó el algoritmo de "Bayesian regularization backpropagation" para el entrenamiento de la red. Todo esto para obtener como dato de salida el mayor promedio móvil de 24h de la concentración de material partículado PM2,5 del día siguiente.

Los resultados concluyeron en que el modelo más simple y óptimo es entrenar la red con datos de entrada de estación *"Padre las Casas"* de Temuco, en el cual se obtuvo un porcentaje de asertividad del 71,1% en la predicción del PM2,5 y errores en los episodios contingentes o críticos como máximo del 52,2% en las alertas y hasta un 0% en las emergencias.

Con estos resultados, se recomienda el uso de este modelo predictivo como una alternativa o complemento al modelo existente para la toma de decisiones políticas.

Palabras Claves: PM2,5, episodio crítico, red neuronal artificial, predicción, promedio móvil horario, predictibilidad.

Dedicatoria

“If history is to change, let it change! If the world is to be destroyed, so be it! If my fate is to be destroyed... I must simply laugh!!”

- Magus (From Chrono Trigger)

“Know thy self, know thy enemy. A thousand battles, a thousand victories.”

- Sun Tzu

Dedico este trabajo a mi familia, y sobre todo a mi madre *Patricia*. Sin el cariño, la preocupación, el apoyo y el esfuerzo que han realizado durante todos estos años, este trabajo no habría sido posible. Muchas gracias por todo.

Además, quiero agradecer a todas las personas que de alguna forma contribuyeron con un grano de arena a este trabajo y que me acompañaron y apoyaron en mi formación.

Agradecimientos

Quiero agradecer en primer lugar a mi profesor guía Dr. *Patricio Pérez*, por su guía, paciencia y sabiduría entregada para ayudarme a completar este trabajo.

Finalmente, quiero agradecer a mi mentor *Mg. Gerardo González*, gracias a su apoyo y guía en este proceso es que este trabajo pudo ser realizado.

Tabla de contenido

Índice de Tablas

Índice de Figuras

Introducción

“The wise man does at once what the fool does finally.”

- Niccolo Machiavelli

La ciudad de Temuco en la región de la Araucanía, con una población total de 245.347 habitantes de acuerdo al CENSO realizado en el año 2012, en la actualidad es una de las ciudades con mayores problemas de contaminación atmosférica del país, encontrándose concentraciones el doble o el triple que las vistas en Santiago, la cual cuenta con 25 veces mayor población que Temuco.

El 10 de enero del año 2013 la zona correspondiente a las comunas de Temuco y Padre las Casas fueron declaradas zonas saturadas por PM2,5 por el SEREMI del medio ambiente de la región de la Araucanía. A raíz de esto, el año 2014 se aprueba el ante proyecto de descontaminación atmosférica para esta zona y el año 2015 se aprueba el D.S 38 como Plan de descontaminación atmosférica por PM2,5 para las comunas de Temuco y Padre las Casas y actualización del plan de descontaminación por PM10 de las mismas comunas.

En las figuras 0.1 y 0.2 se pueden observar los días de episodios críticos distribuidos en cada uno de los rangos de Alerta, Preemergencia y Emergencia Ambiental en las estaciones Las Encinas y Padre las Casas de la ciudad de Temuco.

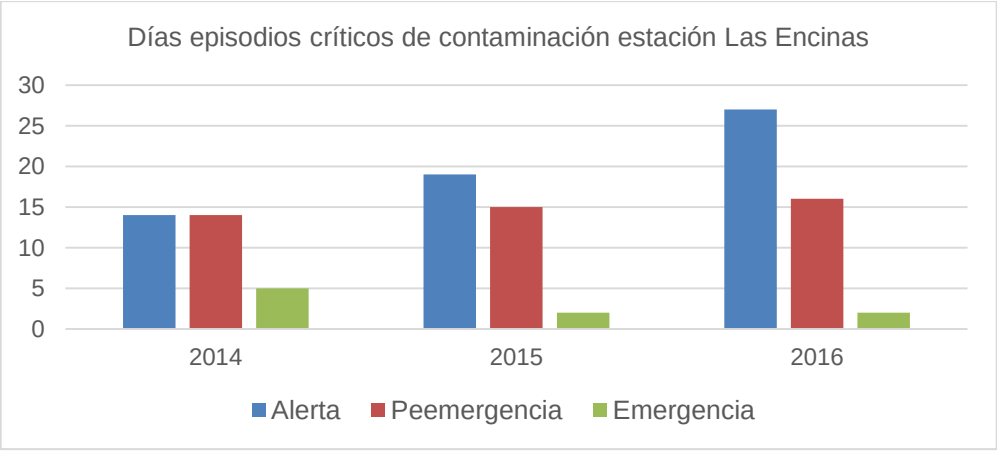

Figura 0.1 Días de episodios de contaminación según datos históricos de material particulado PM2,5 en estación Las Encinas (Fuente: Elaboración propia).

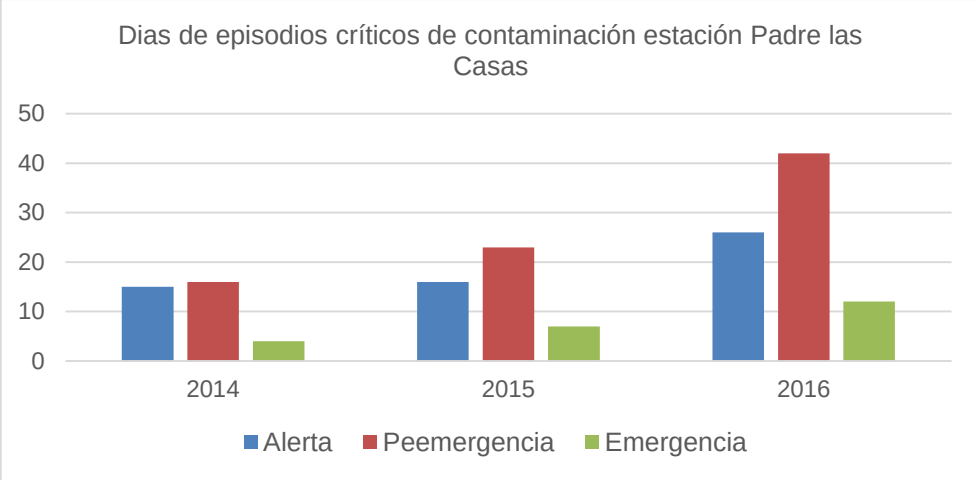

Figura 0.2: Días de episodios de contaminación según datos históricos de históricos de material particulado PM2,5 en estación Padre las Casas (Fuente: Elaboración propia).

Estos datos muestran la cantidad de días que esta ciudad se encuentra en niveles peligrosos de concentración de material particulado PM2,5, y es por esto la necesidad de crear un método de pronóstico de PM2,5.

Este trabajo está dividido en cuatro capítulos y una conclusión final, a continuación, se mostrará un resumen de cada capítulo:

- Capítulo 1: Es un resumen de los conceptos fundamentales de contaminación atmosférica necesarios para el entendimiento de este trabajo. Además, abarca los riesgos a la salud de las personas de los altos niveles de concentración de PM2,5 y la normativa actual vigente.
- Capítulo 2: Abarca el estudio de las series temporales no lineales, enfocado en métodos necesarios para el análisis de datos de este trabajo.
- Capítulo 3: Comienza con una breve explicación de las redes neuronales artificiales y su relación con la neurona biológica. El grueso de este capítulo es la explicación de la estructura de la red, su funcionamiento, el algoritmo de aprendizaje y la cantidad de neuronas en la capa oculta.
- Capítulo 4: Este es el capítulo final, aquí se observa el desempeño de los algoritmos predictivos que se construyeron y se analizan los resultados.

Alcances y Limitaciones

Este proyecto contempla la creación de un modelo predictivo para el material particulado PM2,5 para uso en la ciudad de Temuco (es decir, es un proyecto enfocado en una zona específica) utilizando redes neuronales artificiales. Además, contempla solo la utilización de variables ambientales y no variables antropogénicas.

Adicionalmente, no se consideró la nueva estación “Cerro Ñielol” para el estudio, esto debido a que la estación fue habilitada el presente año (2017) y no se dispone de la cantidad de datos necesarios para aplicar el modelo de red neuronal.

Objetivo

El presente trabajo tiene como objetivo diseñar un modelo que permita pronosticar la concentración de material particulado fino PM2,5 en la atmosfera de la ciudad de Temuco.

Hipótesis

Las concentraciones de material particulado pueden ser pronosticadas mediante análisis de series temporales no lineales, uno de estos métodos en concreto es a través de modelos de redes neuronales artificiales. Identificando de manera correcta los parámetros de entrada, que en este caso corresponden a parámetros ambientales y del clima característicos de la ciudad de Temuco se puede conseguir un pronóstico certero.

Metodología

La metodología utilizada en este trabajo se describe a continuación:

Se reúnen los datos horarios del material particulado PM2,5 y de datos meteorológicos necesarios para la implementación del modelo propuesto. Los datos se recopilan a través de la plataforma del “Sistema Nacional de Calidad del Aire” (SINCA), la cual es operada por el Ministerio del Medio Ambiente, de las estaciones “Las Encinas” y “Padre las Casas II” de la comuna de Temuco.

El periodo de datos elegidos es el período más frío del año, desde el 01 de abril hasta el 31 de octubre en los años de estudio, los cuales fueron 2014, 2015 (entrenamiento de la red artificial) y 2016 (prueba de la red neuronal). Debido a que los pronósticos de material particulado se realizan el día anterior, solo se disponen de datos como un máximo de hasta las 19 horas del día anterior, por lo tanto, la red neuronal artificial considerará el promedio móvil de 24 horas hasta las 19 horas.

Luego se analizan las series temporales de datos (PM2,5 y meteorológicos). Mediante métodos estadísticos, se calcula la correlación entre concentración de PM2,5 y demás variables de interés. Si se quiere generar un pronóstico a las 19 horas del día actual, de tal manera de predecir el máximo del promedio móvil de 24 horas del día siguiente (lo cual define la calidad del aire para ese día), se debe considerar datos disponibles sólo hasta esa hora. Es así como se analizó la correlación de calidad del aire con: PM2,5, temperatura ambiente, velocidad del viento, humedad relativa, presión atmosférica, precipitaciones, radiación solar y dirección del viento; además del dato pronosticado por la Dirección Meteorológica de Chile de la temperatura mínima del día a pronosticar.

Después de seleccionar las variables que parecen permitir un mejor pronóstico, se construirá un modelo de redes neuronales artificiales en Matlab, utilizando la extensión *"Neural Network Toolbox"*. Para estp se utilizará un modelo autorregresivo con entradas y salidas autorregresivas NARX. Para determinar el número de neuronas en las capas ocultas se utilizará el método AIC o criterio de información de Akaike, el cual utiliza la cantidad de parámetros, el tamaño de la muestra y la función de verosimilitud. El algoritmo de aprendizaje que entrega mejor rendimiento al problema es el Bayesian regularization backpropagation.

La red se entrenará y validará con los datos obtenidos de la estación *"Las Encinas"* y la estación *"Padre las Casas"* de forma independiente generando dos modelos, y un modelo extra con datos cruzados de ambas estaciones para obtener un modelo integrativo. El entrenamiento y la validación se realizarán con los datos del periodo 2014 y 2015 respectivamente. Luego de fijar los pesos de conexión de la red neuronal se procederá a probar su capacidad de generalización los datos de las estaciones "Las Encinas" y "Padre las Casas" del periodo 2016. Para analizar la calidad de los modelos se utilizarán 2 métodos: Correlación cruzada entre valor pronosticado y valor observado y la construcción de una "tabla de contingencia".

Además se probarán diferentes configuraciones de datos de entrada para entrenamiento de la red, obteniendo 4 pruebas para los 3 modelos a modo de comparar resultados y determinar la mejor construcción de la red.

Finalmente, se evaluará la respuesta de la red añadiendo la medición de concentración de los gases NO2 y CO, pero, debido a la disponibilidad de los datos en el sistema SINCA se evalúo con los datos de los años 2015 para entrenamiento-validación y 2016 para prueba de la red ya que los datos del año 2014 en periodos de tiempo prolongados no existen mediciones de los parámetros.

Capítulo 1: Contaminación atmosférica y material particulado

"We do not inherit the earth from our ancestors; we borrow it from our children."

- Native American Song

1.1 Fundamentos

La Contaminación atmosférica es uno de los mayores problemas que enfrentan las grandes ciudades del planeta, la cual puede ser originada por eventos naturales como los volcanes, incendios o por eventos antropogénicos debido a las actividades humanas (J. Fenger 1999).

Uno de los contaminantes perjudiciales para la salud es el material partículado de fracción respirable (diámetro aerodinámico menor a 10 μm). A menor diámetro aerodinámico de las partículas atmosférica mayor es el nivel de penetrabilidad en el sistema respiratorio, destacándose el material partículado fino (menor a 2,5 μm) que corresponde a partículas que son capaces de atravesar los alvéolos pulmonares. Numerosos estudios epidemiológicos realizados a partir de la década de 1990 (Dockery et al., 1993; Pope y Dockery, 1996) han demostrado la existencia de efectos adversos para la salud derivados de la exposición puntual o prolongada a niveles elevados de material particulado atmosférico (PM).

Las emisiones de contaminantes a la atmósfera provienen de dos tipos de fuentes: las naturales (que proceden de fenómenos naturales como erupción, incendios forestales, polvo en suspensión generado por vientos, entre otros) y las artificiales (o antropogénico) que se producen debido a las actividades humanas diarias, como por ejemplo industrias, chimeneas de los hogares, quemas agrícolas, tubo de escape de los vehículos, los vertederos, actividad minera, plantas térmicas, a carbón, petróleo o gas, etc.

De entre los elementos contaminantes de especial interés es el material particulado, a diferencia de los contaminantes gaseosos como óxidos de carbono, de nitrógeno, de azufre, etc. El material particulado típicamente contienes aerosoles, polvo natural y antropogénico, nitratos, amonio, sulfatos, carbono orgánico, carbono vegetal,. El material particulado se puede clasificar en términos de su tamaño o diámetro. En la figura 1.1 se muestra, a modo de comparación, el tamaño del material particulado de 2,5 μm y 10 μm respecto del diámetro de un cabello humano promedio.

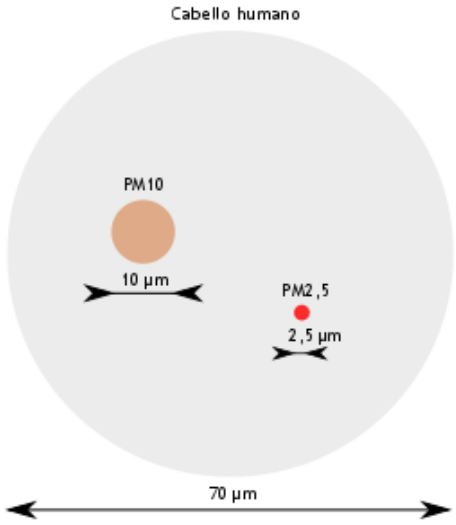

Figura 1.1: Relación entre material particulado y el grosor de un cabello humano.

1.2 Efectos del material particulado en la salud

El PM10 tiene un grado de penetración limitado en el sistema respiratorio humano pudiendo alcanzar el árbol traqueobronquial. Sin embargo, si no son demasiado pequeñas, allí quedan adheridas a la capa gel del sistema mucociliar transportador y son llevados corriente arriba por la actividad ciliar del epitelio bronquial, hasta alcanzar la faringe, donde son deglutidas y eliminadas por el tubo digestivo.

El PM2,5 en cambio pueden alcanzar zonas más profundas del pulmón, llegando a vías aéreas donde no hay células ciliadas como, por ejemplo, los bronquiolos terminales y bronquiolos respiratorios. Allí pueden ejercer su acción directa por un tiempo indefinido o bien pueden ser transportados en la sangre a otros órganos y sistemas.

El PM1 posee un grado de penetración en el sistema respiratorio mayor que el PM2,5 provocando una respuesta inflamatoria mayor al tener un área de contacto proporcionalmente más grande (Raúl Morales, 2006).

Los efectos del material particulado PM2,5 en la salud humana están descritos en la tabla 1.1

Tabla 1.1: Efectos adversos respiratorios del material particulado PM2,5.

Corto Plazo	*Largo Plazo*
- Aumento de la morbilidad y mortalidad respiratoria. - Disminución en la función pulmonar. - Inflamación mononuclear. - Interferencia en mecanismos de defensa pulmonar: fagocitosis y depuración mucociliar.	- Bronquitis crónica. - Genotoxicidad. - Factor de riesgo de cáncer pulmonar y EPOC.

1.3 Legislación Nacional para el PM 2,5 en Chile

De acuerdo al Decreto Supremo Nº 12 del ministerio del Medio Ambiente de Chile, publicado en el diario oficial el 9 de mayo de 2011, se establece la norma primaria para material particulado fino respirable PM2,5.

Para cuantificar la concentración de PM2,5 en términos simples se ha creado el Índice de Calidad del Aire referido a Partículas (ICAP) el cual es una función lineal discontinua a dos trazos de la concentración de PM2,5. La pendiente en el primer trazo (0 – 50 [µg/m^3]) es considerablemente mayor a la segunda (50 – 170 [µg/m^3]) lo que hace que una vez superado los 50 [µg/m^3] el ICAP ascienda con menor intensidad que el primer tramo.

Tabla 1.2: ICAP PM2,5

ICAP	*Concentración PM2,5 [µg/m^3]*
0	0
100	50
500	170

Según dicta la normativa, los límites de la concentración dictados por el Artículo 3º. La norma primaria anual de calidad del aire para material particulado fino es veinte microgramos por metro cúbico (20 µg/m3l, y cincuenta microgramos por metro cúbico (50 µg/m3), como concentración de 24 horas.

Así mismo, según el Artículo 4º. Se considerará sobrepasada la norma primaria de calidad del aire para material particulado fino respirable PM2,5, en los siguientes casos:

a) Cuando el percentil 98 de los promedios diarios registrados durante un año, sea mayor a 50(µg/m3), en cualquier estación monitora calificada como EMRP; o

b) Cuando el promedio tri-anual de las concentraciones anuales sea mayor a 20(µg/m3), en cualquier estación monitora calificada como EMRP.

Las concentraciones de 24 horas que definen episodios de contaminación atmosférica están indicadas en el Artículo 5°

Tabla 1.3: Definición de episodios críticos para PM2,5

Nivel	*Concentración 24 horas PM2,5 [µg/m^3]*
Bueno	0 - 50
Regular	51 - 79
Alerta	80 – 109
Preemergencia	110 – 169
Emergencia	>170

Las concentraciones serán obtenidas a partir de un pronóstico de calidad del aire, o bien, en caso que no se cuente con este pronóstico, de la constatación de las concentraciones de Material Particulado Respirable PM2,5 a partir de las mediciones provenientes de alguna de las estaciones de monitoreo de calidad del aire calificadas como EMRP.

Las metodologías de pronóstico serán definidas al momento de elaborar el respectivo Plan de Descontaminación o Prevención, debiendo para estos efectos emplearse los modelos de pronósticos más actualizados en la materia.

De acuerdo a la resolución 609 de la subsecretaría del Ministerio del Medio Ambiente, los pronósticos de material particulado deben ser emitidos por la DGAC (Dirección general de Aeronáutica Civil) en el rango horario de las 20:00 y 20:30 horas, por lo tanto solo se pueden contar con los datos actualizados hasta las 20:00 horas.

1.4 Clima y Contaminación Atmosférica en la Ciudad de Temuco

Temuco es una ciudad y comuna de Chile, capital de la provincia de Cautín y la región de La Araucanía. La comuna de Temuco cuenta, según el censo de 2002, con una población de 245.347 habitantes en una superficie total de 464 km². Situada aproximadamente a 122 metros sobre el nivel del mar, ubicada en la depresión intermedia del país en lo que se denomina las

terrazas fluviales del río cautín que se desarrollan en forma encajonada entre el cerro Ñielol (350 msnm) y el cerro Conun Huenu (360 msnm).

Según la clasificación climática, Temuco se localiza en la categoría que Koeppen define como Cfs, en los cuales a pesar de que las lluvias descienden en el verano, no puede calificarse dicho período como una estación seca. De esta forma, corresponde a la de un clima templado lluvioso o húmedo con precipitaciones concentradas en al menos 10 meses del año. En este clima, las temperaturas medias mensuales son inferiores a los 18° C, aunque el promedio de las máximas puede superar los 25° C, en tanto las mínimas pueden alcanzar los 2° C. Las precipitaciones, en tanto, son siempre superiores a los 1.250 mm anuales, registrándose lluvias durante el verano, las que, aunque no alcanzan montos importantes, si permiten señalar que se registran precipitaciones en época estival. Esta clasificación coincide con otras la cuales también definen el sector como clima mediterráneo frío.

De acuerdo al inventario de emisiones de la ciudad de Temuco del año 2015 que se observa en la figura 1.2, la principal fuente contaminante corresponde a la combustión de leña residencial, con un 97% de incidencia en la contaminación atmosférica.

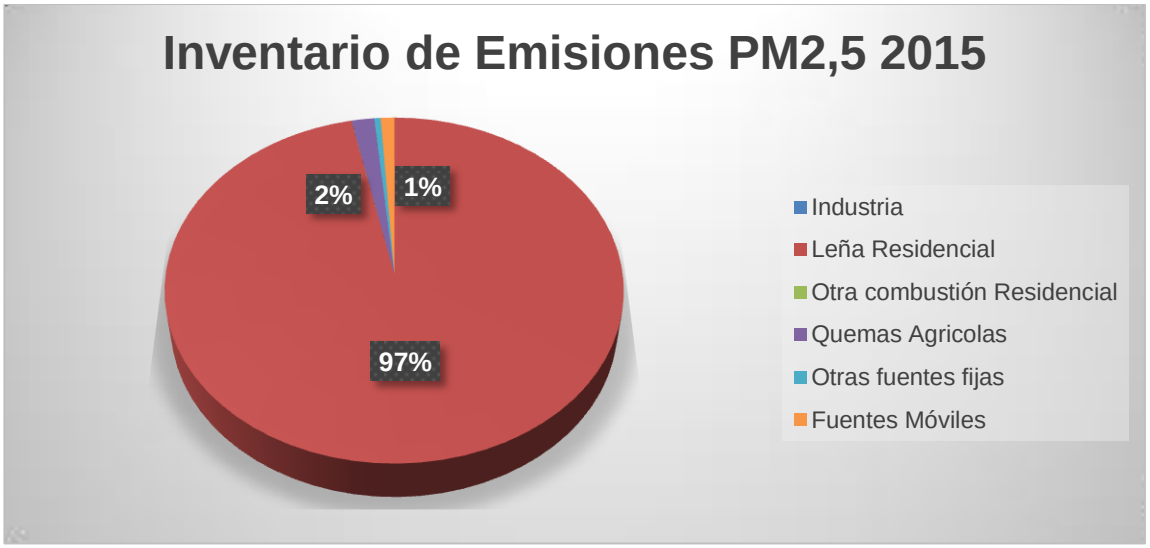

Figura 1.2: Inventario de emisiones PM2,5 del año 2015 indicando fuente contaminante.

Considerando el clima y las fuentes de emisión de material particulado PM2,5 no es de extrañar que los niveles de concentración aumenten el periodo invierno de gran manera en comparación con otras ciudades. Donde la sensación térmica para las personas es un factor fundamental en la contaminación del aire debido a la calefacción mediante combustión de leña residencial.

Por lo tanto, es de gran importancia estudiar los factores climáticos que afectan directamente a la sensación térmica como las temperaturas ambientales, además de los factores de ventilación de la ciudad, propias de las ciudades en la cuenca central de Chile.

El Plan de Descontaminación Atmosférica de Temuco y Padre Las Casas PDA MP10 y MP2,5 D.S.Nº8/2015 MMA establece una serie de medidas en favor de mitigar y mejorar las condiciones atmosféricas de la ciudad de Temuco. Dentro de las principales características se encuentran

- Monitoreo y reporte mensual de Leña Seca (SERNAC). ARTICULO Nº13 PDA.
- Recambio de artefactos. ARTICULOS Nº26 Y 27 PDA.
- Operación sistema registro de artefactos. ARTICULO Nº28 PDA.
- Exigencia de compensación de emisiones de un 120% para proyectos en el SEIA que generen más de 0,5 ton/año de MP. ARTICULO Nº58 PDA.
- Gestión de Episodios Críticos. ARTICULOS Nº63 al 68 PDA.
- Programa educativo difusión de medidas en RRSS, actividades deportivas masivas, trabajo con escuelas SNCAE, cuenta pública. ARTICULO Nº70 PDA.
- Jornadas con Líderes sociales con Seremi de Gobierno. ARTICULO Nº 74 PDA.

Donde, para este estudio lo de mayor relevancia es la gestión de episodios críticos GEC debido a la necesidad de realizar pronósticos de episodios críticos. Las principales características de la gestión de episodios críticos son:

- Operación y coordinación GEC: 1º de abril al 30 de Septiembre.
    - La declaratoria de un episodio Crítico se establece desde Intendencia (por resolución).
    - La Seremi del MA establecerá a través de una resolución zonas territoriales de gestión de episodios.
    - Las medidas las Fiscaliza la seremi de Salud.
- Pronósticos diarios.
- Seguimiento y reporte calidad de aire diario.
- Campaña difusión GEC (M$37.500) Radio, RRSS, etc. SEMAFOROS.

Las medidas de estas gestiones se pueden ver en la figura 1.3

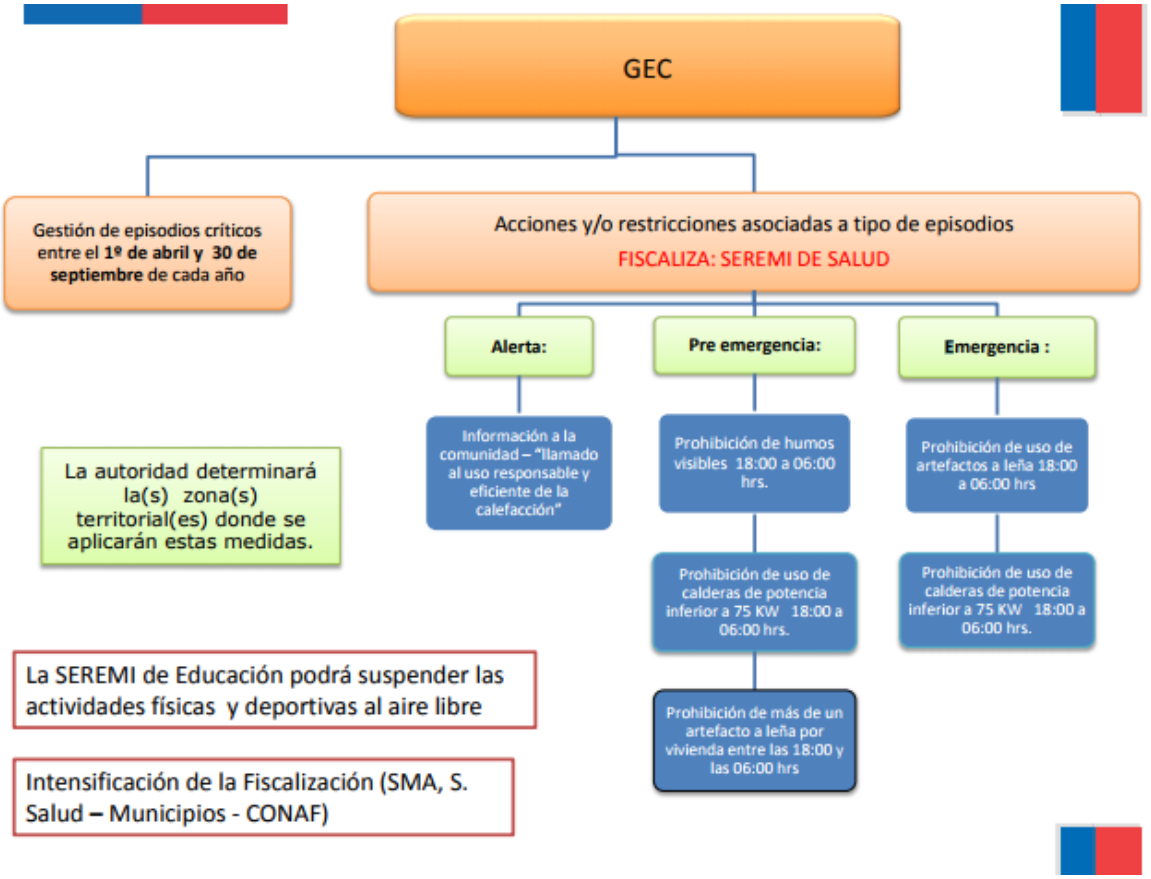

Figura 1.3. . Mapa de acción de gestión de episodios críticos GEC del D.S.Nº8/2015 MMA. (Fuente: D.S.Nº8/2015 MMA).

De acuerdo al plan de descontaminación, se aprobó el 3 de Mayo del 2016 mediante publicación en el diario oficial la metodología de pronóstico para calidad del aire.

"La Metodología de Pronóstico de Calidad de Aire para Material Particulado Respirable PM2,5, denominada "Modelo Predictivo de Calidad de Aire para Material Particulado Fino Respirable MP2,5 WRF-MMA", que tiene por objeto entregar un pronóstico de la máxima concentración promedio de Material Particulado MP2,5, para 24 horas, expresada en microgramos por metro cúbico (µg/m3), para las estaciones de monitoreo de material particulado MP2,5 con representatividad poblacional (EMRP), ubicadas en las zonas saturadas de Chillán y Chillán Viejo, Temuco y Padre Las Casas y Osorno.

El modelo de pronóstico WRF-MMA, se encuentra basado en el modelo Weather Research and Forecasting with Chemistry (WRF-Chem) de losautores Skamarock et al., 2008 y Grell et al. 2005, versión 3.6.1. Este modelo es de tipo Online, es decir, resuelve la meteorología y la química atmosférica en forma simultánea para generar la estimación de contaminantes atmosféricos, tanto en superficie como en altura. La configuración del modelo para su uso en

Chile se basa en el estudio de P. Saide et al., de 2016, denominado Air Quality Forecasting for Winter-Time PM2.5 Episodes Occurring in Multiple Cities in Central and Southern Chile, en el cual la modelación del Monóxido de Carbono (CO) es utilizada como trazador del MP2,5 debido a una alta correlación entre ellos, en especial durante la ocurrencia de episodios críticos de contaminación atmosférica. El modelo WRF-MMA permite predecir la calidad del aire para MP2,5 y la meteorología asociado con 3 días de anticipación. Para ello, utiliza datos meteorológicos de escala regional obtenidos de forma automática desde el Global Forecasting System (GFS) del National Center for Environmental Information y datos de uso de suelo de U.S Geological Survey (USGS), en conjunto con los datos del inventario de emisiones de trazador de CO, de la zona de aplicación. La zona de aplicación de dicho pronóstico, conforme a la configuración mencionada, se extiende desde la Región Metropolitana hasta la Región de Los Lagos (Osorno)." (Extracto de publicación en el diario oficial).

De acuerdo al estudio realizado por P. Saide et al., de 2016. El porcentaje de predictibilidad para la estaciones de Padre las Casas II es de un 73%[1] con un porcentaje de falsas alarmas de un 23%, y para estación las Encinas es de un 64% con un 36% de falsas alarmas. Estos valores son considerados de acuerdo a los episodios, definidos en el estudio cuando la concentración del PM2,5 es sobre los 80μg/m^3. Estos resultados corresponden a la evaluación de datos del período abril-septiembre del año 2014.

[1] Valor de la *persistencia* de la predictibilidad en el tiempo.

Capítulo 2: Estudio de Series Temporales no Lineales

"Measure what is measurable, and make measurable what is not so."

-Galileo Galilei

2.1 Fundamentos

Una serie temporal se puede definir como una secuencia de datos u observaciones de una variable física tomada de intervalos de tiempo, definidos como variables continuas o discretas. Matemáticamente hablando, una s. de t. se define como un conjunto de mediciones de cierto fenómeno o experimento, registrados secuencialmente en el tiempo. Normalmente estas series temporales están basadas en valores medidos que portan ruido. Los valores de estas series generalmente contienen una parte sistemática (determinista) y otra componente estocástica (aleatoria), que representa a la interferencia ruidosa, la cual a su vez provoca fluctuaciones estadísticas alrededor de los valores deterministas (D. Peña S. 2005).

No obstante, no se puede observar ninguna de estas partes en las series reales. Por lo que los métodos de pronóstico tratan de aislar la parte sistemática. Así, los pronósticos están basados en la parte sistemática de la serie. Sin embargo, la parte aleatoria determina la forma de la distribución. Se puede afirmar, entonces, que el propósito del análisis de una s. de t. es comprender la variabilidad de éstas; identificar las oscilaciones regulares e irregulares de estas series; describir las características de estas oscilaciones y estudiar los procesos.

Por lo general, el análisis de las series cronológicas se refiere a estimar los factores (o componentes) que producen el comportamiento general o patrón de las series; y el uso de estas estimaciones para predecir el comportamiento futuro de la serie.

La tendencia se entiende como el movimiento continuo de una variable durante un período de tiempo extendido. En otras palabras, es la conducta a largo plazo de la variable durante un período de longitud prolongado, y refleja la dirección general de la s. de t. la cual puede ser ascendente o descendente. Por su parte, la variación cíclica o estacionalidad son las variaciones periódicas, en el nivel general de la actividad, durante un período relativamente prolongado. Series de tiempo periódicas con componentes periódicas son periódicas. Se puede afirmar que una serie temporal es estacionaria si su media y varianza son constantes en el tiempo y la autocovarianza depende sólo del rezago temporal. El ruido aleatorio puede estar presente en toda o parte de la serie temporal.

Una forma de visualizar la tendencia, es mediante el suavizamiento de la serie:

$$Z(t) = F\big(x(t)\big) \qquad [2.1]$$

En que F se denomina Filtro lineal.

El filtro lineal más usado es el Promedio Móvil (moving average, MA). Con este filtro se busca eliminar las componentes estacionales y estocásticas (ruidosas)[2]. En la figura 2.1 se puede observar cómo actúa el suavizamiento de la tendencia de datos de la concentración de PM2,5 correspondiente al período 2014 de la estación las encinas de Temuco.

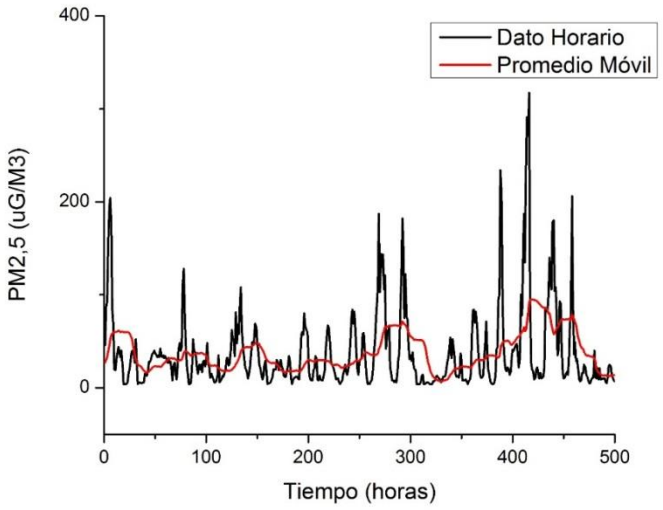

Figura 2.1: Extracto de datos de PM2,5[3] horarios y con filtro lineal de promedio móvil 24 horas.

2.2 Relación entre series temporales

Cuando se desea cuantificar la relación o asociación entre dos series del tiempo o del clima, o entre una de ellas y otra variable de naturaleza no climática, usualmente se recurre a métodos paramétricos como el coeficiente de correlación lineal, como el de Pearson o a no paramétricos de Spearman o el de Mann-Kendall. En muchas situaciones no hay indicios de correlación entre los eventos expresados por estos coeficientes, puesto que ellos solo expresan la asociación en tiempo simultáneo, pero cuando se les aplica a las series la técnica estadística de la correlación cruzada, es posible que resalten asociaciones muy claras entre las series.

[2] Se suaviza cuando existen muchos cambios bruscos o movimientos irregulares.
[3] Datos correspondientes al periodo 2014 de estación las encinas.

Si bien, al aplicar la correlación cruzada a datos no estacionarios puede llevar a que, aunque los coeficientes resultasen altos y significativos, no tendrían validez y las correlaciones podrían ser consideradas espurias, en este caso se utilizará esta correlación como un filtro preliminar de variables a utilizar para el modelo predictivo.

Así, el dato de interés es el promedio móvil 24-horas más alto del día a pronosticar (es decir el dato t+1), y como solo se dispone de los datos hasta las 20:00 horas, se utilizará para comparar los datos promedio horario de las 18:00 horas y los datos promedio móvil 24 horas de las 18:00 horas, de modo de disponer del pronóstico dentro de los horarios correspondientes, ya que la información de esta hora tiene un lag de actualización de 1h y media aproximadamente, es decir el dato de las 18:00 horas estará disponible entre las 19:00 y 20:00 horas.

Para el caso de las estaciones de monitoreo de Temuco, Las Encinas y Padre las Casas II, del cálculo de las correlaciones cruzadas (cuyo valor no es de relevancia para este estudio), se obtienen las siguientes variables de interés:

- PM2,5 promedio horario de las 18:00 horas para ambas estaciones,
- PM2,5 promedio móvil 24h de las 18:00 horas para ambas estaciones.
- Temperatura ambiente promedio horario de las 18:00 horas para ambas estaciones ya que permite conocer el comportamiento de la temperatura en las últimas horas antes del pronóstico.
- Temperatura ambiente promedio móvil 24h de las 18:00 horas para ambas estaciones ya que nos indica el perfil térmico del día.
- Temperatura ambiente mínima pronosticada para el día siguiente, este se complementa con los datos anteriores de temperatura para completar el perfil de comportamiento de la temperatura.
- Velocidad del viento promedio móvil 24h de las 18:00 horas para ambas estaciones.
- PM10 promedio horario de las 18:00 horas para ambas estaciones.
- PM10 promedio móvil 24h de las 18:00 horas para ambas estaciones.

Estas variables son las que serán utilizadas como input para el modelo de redes neuronales que se describirá en los siguientes capítulos, es de interés señalar que no se consideraron variables como la presión atmosférica, las precipitaciones y la radiación solar. Además señalar que por la naturaleza de las estaciones, no se disponen de los datos de NO_2 de las estaciones. En Las Encinas, este parámetro no se mide actualmente y en Padre las Casas, la medición de este parámetro comenzó el 2014 y existen muchos espacios sin información, por lo que podría llegar a ser contraproducente considerar este dato en las condiciones para el periodo considerado en el estudio.

Capítulo 3: Redes Neuronales Artificiales

"There are not more than five musical notes, yet the combinations of these five give rise to more melodies than can ever be heard. There are not more than five primary colors, yet in combination they produce more hues than can ever been seen. There are not more than five cardinal tastes, yet combinations of them yield more flavors than can ever be tasted."

- Sun Tzu

3.1 La neurona biológica

Las Redes Neuronales Artificiales (ANR) están inspiradas en las neuronas biológicas del cerebro humano (E.R. Kandel & J-H- Schwartz 1985) el que consiste de unas cien mil millones de células que conectadas entre sí son capaces de realizar tareas complejas como memoria, asociación, aprendizaje, clasificación, percepción, atención, comunicación, razonamiento, inspiración, conciencia, entre otros atributos del sistema; esto gracias a las aproximadamente 10^{15} interacciones interneuronales que se producen en total.

La unidad "básica" del cerebro humano es la neurona -célula nerviosa especializada- y más aún del Sistema Nervioso Central (SNC). Las neuronas se comunican entre sí a través de la sinapsis. Cada neurona puede tener hasta 60.000 puntos de contacto o sinapsis, que le permite comunicarse con otras neuronas.

La señal generada por la neurona y que es transportada a lo largo del axón es un impulso eléctrico (E.R. Kandel & J-H- Schwartz 1985). Ver figura 3.1. En cambio, la señal que se transmite entre los terminales axónicos de una neurona y las dendritas de la neurona siguiente es de origen químico; más específicamente, se efectúa mediante las moléculas de ciertas sustancias químicas transmisoras denominadas neurotransmisoras que fluyen a través de unos contactos que se denominan sinapsis. La mayoría de las sinapsis ocurren entre el terminal axónicos de una neurona (que lleva un impulso eléctrico) y las dendritas de la neurona siguiente (que recibe el estímulo).

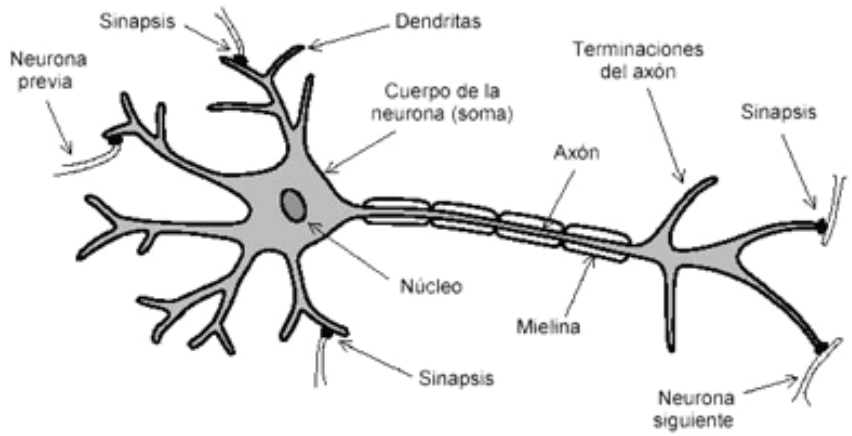

Figura 3.1: Esquema de una neurona.

La capacidad de aprendizaje de nuestro cerebro, así como su resistencia a las averías locales (robustez) se debe fundamentalmente al funcionamiento colectivo y simultaneo de las neuronas que lo componen (capacidad de procesamiento en paralelo), estando organizadas éstas en redes altamente interconectadas. Por ejemplo, al efectuar un recuerdo nuestro cerebro realiza un gigantesco cálculo que, si se compara con un ordenador, el cual haría la misma tarea de manera secuencial, le tomaría muchísimo más tiempo en realizar la misma tarea.

Pese a lo reducido de la velocidad de procesamiento de cada neurona, el paralelismo le permite capturar la imagen de un rostro al primer encuentro. Su capacidad de memoria es tan amplia que almacena las experiencias de una vida entera y es tan versátil que el recuerdo de una escena trae a colación asociaciones de imágenes visuales, sonidos, olores, sabores, sensaciones táctiles y emociones. La mayoría de las veces, los recuerdos se originan a partir de impresiones sensoriales.

3.2 Estructura y funcionamiento de una red neuronal artificial

Las Redes Neuronales Artificiales son algoritmos para tareas cognitivas, tales como el aprendizaje y la optimización, las cuales están, en algún aspecto, basado en conceptos derivados de la investigación sobre la naturaleza del cerebro.

La investigación ha demostrado (J.A. Freeman y D.M. Skapura 1993, Mc Cullogh y Pitts, 1943) cómo las redes que consisten de un gran número de elementos de proceso muy simples, las neuronas, pueden ser usados para efectuar tareas computacionales. En estas redes se unen

cientos de unidades individuales de procesamiento a través de conexiones, imitando las propiedades del cerebro.

La estructura de una red neuronal se muestra en la figura 3.2

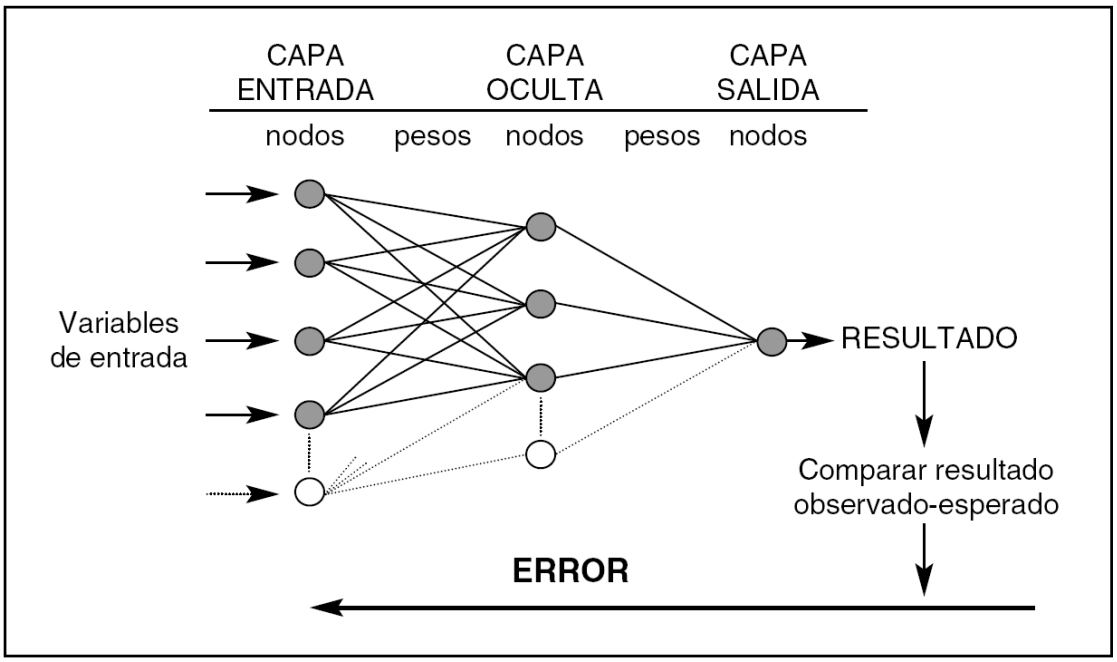

Figura 3.2: Estructura de una red neuronal artificial.

La misma está constituida por neuronas interconectadas y arregladas en tres capas. Los datos ingresan por medio de la “capa de entrada”, pasan a través de la “capa oculta” (es posible que la red tenga más de una capa oculta) y salen por la “capa de salida”. Comparando las figuras 3.1 y 3.2 se puede observar las similitudes entre una neurona biológica una neurona artificial.

La distribución de neuronas dentro de la red se realiza formando niveles o capas, con un La distribución de neuronas dentro de la red se realiza formando niveles o capas, con un número determinado de dichas neuronas en cada una de ellas. A partir de su situación dentro de la red, se pueden distinguir tres tipos de capas:

- De entrada: es la capa que recibe directamente la información proveniente de las fuentes externas de la red.

- Ocultas: son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de redes neuronales.

- De salidas: transfieren información de la red hacia el exterior.

En la Figura 3.1 se puede ver el ejemplo de la estructura de una posible red multicapa, en la que cada nodo o neurona únicamente está conectada con neuronas de un nivel superior. Notar que hay más conexiones que neuronas en sí; en este sentido, se dice que una red es totalmente conectada si todas las salidas desde un nivel llegan a todos y cada uno de los nodos del nivel siguiente.

Una red neuronal debe aprender a calcular la salida correcta para cada constelación (arreglo o vector) de entrada en el conjunto de ejemplos. Este proceso de aprendizaje se denomina: proceso de entrenamiento o acondicionamiento. El conjunto de datos (o conjunto de ejemplos) sobre el cual este proceso se basa es, por ende, llamado: conjunto de datos de entrenamiento. Si la topología de la red y las diferentes funciones de cada neurona (entrada, activación y salida) no pueden cambiar durante el aprendizaje, mientras que los pesos sobre cada una de las conexiones si pueden hacerlo; el aprendizaje de una red neuronal significa: adaptación de los pesos.

En otras palabras, el aprendizaje es el proceso por el cual una red neuronal modifica sus pesos en respuesta a una información de entrada. Los cambios que se producen durante el mismo se reducen a la destrucción, modificación y creación de conexiones entre las neuronas. En los sistemas biológicos existe una continua destrucción y creación de conexiones entre las neuronas.

En los modelos de redes neuronales artificiales, la creación de una nueva conexión implica que el peso de la misma pasa a tener un valor distinto de cero. De la misma manera, una conexión se destruye cuando su peso pasa a ser cero.

Después del proceso de entrenamiento los pesos de las conexiones en la red neuronal quedan fijos. Como paso siguiente se debe comprobar si la red neuronal puede resolver nuevos problemas, del tipo general, para los que ha sido entrenada. Por lo tanto, con el propósito de validar la red neuronal se requiere de otro conjunto de datos, denominado conjunto de validación o testeo. Cada ejemplo del conjunto de evaluación contiene los valores de las variables de entrada, con su correspondiente solución tomada; pero ahora esta solución no se le es otorgada a la red neuronal. Luego se compara la solución calculada para cada ejemplo de validación con la solución conocida. El conjunto de validación, a diferencia del conjunto de test, se puede usar para determinar el número óptimo de iteraciones de entrenamiento para evitar lo que se conoce como sobre-entrenamiento

3.3 Construcción de la red neuronal artificial utilizando Matlab

De acuerdo a las características del problema, la estructura de la red neuronal artificial para Matlab se construirá con la estructura autorregresivo. En modelamiento de series temporales, se utiliza el modelo no lineal autorregresivo. Un modelo autorregresivo corresponde a una representación de un tipo de proceso aleatorio el cual depende de forma estocástica (por lo tanto imperfectamente predecible) de sus valores anteriores.

Esto significa que el modelo considera valores de las series temporales de datos anteriores de las mismas series, y de datos actuales y anteriores de series exógenas, las cuales son series externas que influyen en las series de interés. La ecuación definitoria para el modelo es:

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-d), x(t-1), x(t-2), \dots, x(t-d)) \qquad [3.1]$$

Donde x e y son la entrada y salida respectivamente y d corresponde al parámetro de *delay*. Se observa que para la predicción, el modelo utiliza datos anteriores de la misma salida (que en este caso corresponde a los datos históricos de la concentración promedio móvil 24h máxima del día anterior).

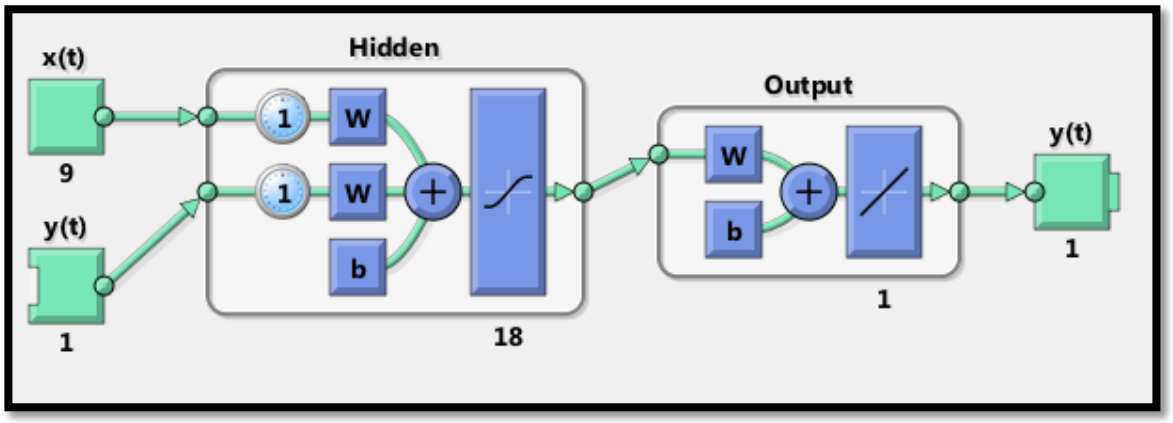

Figura 3.3: Esquema de una red neuronal artificial en Matlab.

La función de entrenamiento óptima para el problema de predicción ambiental es la función “Bayesian regularization backpropagation” en lugar del algoritmo “backpropagation” estándar. El algoritmo de regularización Bayesiano es una aproximación para prevenir el sobre-entrenamiento de la red. La complicación de la regularización Bayesiana es que puede entregar malos resultados si el modelo está mal especificado ya que para reducir el tiempo de cálculo elimina la necesidad de la validación cruzada, minimizando a una combinación lineal entre errores cuadrados y pesos utilizando el Jacobiano para realizar los cálculos (MacKay, 1992).

En sí, el algoritmo de entrenamiento en regularización Bayesiana difiere en los algoritmos estándar en la aproximación al realizar los cálculos de los pesos, aproximando los cálculos para reducir el tiempo de cálculo y el sobre-entrenamiento.

3.4 Determinación de número de neuronas en la capa oculta

Para determinar el número óptimo de neuronas en las *Hidden Layers* o capas ocultas de la red se utilizó el criterio de información de redes (NIC) utilizando el criterio de información de Akaike (AIC) (Noboru Murata, Shuji Yoshizawa, Shun-ichi Amari 1992). Para calcular el número de capas ocultas primero se debe calcular el AIC de la siguiente forma:

$$AIC = 2k - 2\ln(\hat{L}) + \frac{2k(k+1)}{n-k-1} \qquad [3.2]$$

Donde k es el número de parámetros, n el tamaño de la muestra y $\hat{L}$ es la función de verosimilitud. Pero el AIC puede ser calculado usando la suma de errores cuadrados RSS de la regresión

$$AIC = 2K + n\ln\left(\frac{RSS}{n}\right) + \frac{2K(K+1)}{n-K-1} \qquad [3.3]$$

Donde K es el número de funciones de peso disponibles y $RSS = \sum \epsilon_I$, es decir es la suma de los errores cuadrados corregidos de los datos de cada parámetro. K se calcula como:

$$K = k \cdot ℏl + ℏl \qquad [3.4]$$

Donde $ℏl$ son la cantidad de neuronas en la capa oculta, donde $ℏl \in IN$. Estos valores fueron calculados mediante software matemático Matlab.

Graficando los valores posibles de las neuronas en la capa oculta en función del valor calculado del AIC, se obtiene el número de neuronas óptimo para la capa oculta tal como se muestra en la figura 3.4, donde el menor valor posible del criterio AIC corresponde a la cantidad óptima de neuronas. Lo cual, en este caso corresponde a 18 neuronas en la capa oculta para optimizar el problema.

Finalmente se utilizaron los datos correspondientes al periodo 2014 para el proceso de aprendizaje y entrenamiento de la red neuronal artificial, los datos del periodo 2015 para la validación de la red.

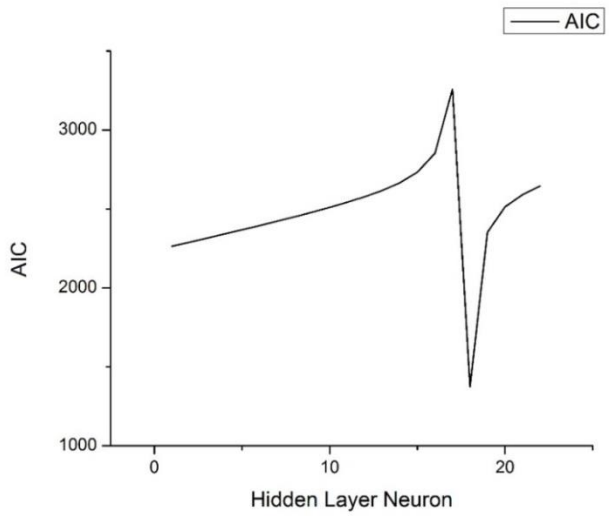

Figura 3.4: Número de neuronas en la capa oculta en función del criterio de información de Akaike AIC.

3.5 Respuesta de la red neuronal artificial y preparación de pruebas

Para encontrar los parámetros de entrada de la red neuronal artificial óptimos se procedió a probar diferentes combinaciones, en primera instancia variando la estación base para realizar el aprendizaje de la red, así se obtuvieron tres tipos de datos de entrada:

- Red entrenada en base a datos de estación “Las Encinas”.
- Red entrenada en base a datos de estación “Padre las Casas”.
- Red entrenada en base a datos combinados de ambas estaciones (PM2,5, PM10).

Por otro lado, se varió los datos de entrada para así analizar la predictibilidad de las redes en función de las variables de entradas, se utilizaron 4 pruebas utilizando los datos enlistados en la sección 2.2:

- Prueba 1: Se utilizaron los datos a excepción del PM10 y la presión atmosférica de ambas estaciones.
- Prueba 2: Se utilizaron los datos a excepción del PM10 de ambas estaciones.
- Prueba 3: Se utilizaron los datos a excepción de la presión atmosférica de ambas estaciones.
- Prueba 4: Se utilizaron todos los datos.

- Prueba 5: Se utilizaron los datos de la prueba 3, añadiendo además el dato concentración PM2,5 promedio horario y el promedio móvil 24h de las 17:00 horas, es decir una hora antes añadida a efecto de observar mejor la distribución del PM2,5 para ambas estaciones. [4]

Así, se obtuvieron 15 redes neuronales diferentes con diferentes combinaciones. En la tabla 3.1 se puede observar la respuesta preliminar del entrenamiento y la validación de datos de cada una de las redes construidas.

Se debe notar que a mayor error cuadrático medio (MSE por sus siglas en inglés), existe mayor dispersión entre los datos de salida y los datos reales, lo que indica que hay mayor probabilidad de que los datos tengan un mayor error asociado. Adicionalmente un número cercano a 1 en el ajuste R cuadrático índica una mayor linealidad entre los datos de salida y los datos reales, es decir, mientras mayor el valor, menor probabilidad de error asociado a la predicción.

Tabla 3.1: Respuesta del entrenamiento y validación de cada red neuronal.

N° de Red	*Pruebas*	*Tipo de Entrenamiento*	*Entrenamiento*		*Validación*	
			MSE	R	MSE	R
1	Prueba 1	Las Encinas	852,53	0,811	693,07	0,762
2		Padre las Casas	987,09	0,823	975,59	0,753
3		Mixto	889,18	0,836	640,41	0,886
4	Prueba 2	Las Encinas	960,82	0,783	363,67	0,895
5		Padre las Casas	983,10	0,829	725,07	0,832
6		Mixto	1062,26	0,796	753,09	0,881
7	Prueba 3	Las Encinas	941,89	0,785	546,70	0,837
8		Padre las Casas	885,17	0,813	1391,95	0,847
9		Mixto	1060,97	0.809	718,05	0,834
10	Prueba 4	Las Encinas	856,65	0.796	825,22	0,817
11		Padre las Casas	897,08	0,830	850,05	0,862
12		Mixto	963,16	0,820	1196,71	0,778
13	Prueba 5	Las Encinas	695,25	0,835	1522,93	0,671
14		Padre las Casas	964,88	0,824	1082,39	0,759
15		Mixto	887,45	0,846	1007,88	0,711

[4] La prueba 5 se realizó después de conocer los resultados de las pruebas anteriores, por lo que se dio énfasis en mejorar la red de mejor rendimiento.

Capítulo 4: Análisis y pronóstico de material particulado en la ciudad de Temuco

“Nothing can beat science!”

- Luca (From Chrono Trigger)

4.1 Análisis mediante coeficiente de correlación cruzada

Como se mencionó anteriormente, la red neuronal artificial fue testeada con los datos del PM2,5 en promedio 24h más alto del día a pronosticar, del periodo 2016 de ambas estaciones (para el caso de datos mixtos, se utilizó el dato mayor de ambas).

Para análisis de esta predicción se comparará los datos predichos por la red neuronal construida tal como se indicó en el capítulo anterior, una forma simple de analizar la relación entre series es utilizando el método de la correlación cruzada, para esto se utilizará la serie de datos temporales de la predicción del PM2,5 en promedio 24h y se compararon con los datos observados del mismo parámetro en el mismo día.

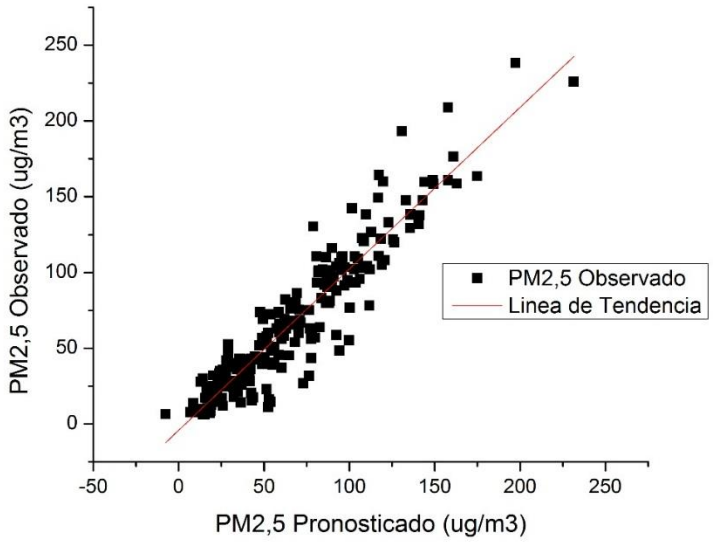

Figura 4.1: Ejemplo del método de correlación utilizado, corresponden a datos pronosticados y observados del mayor promedio horario 24h de PM2,5 para estación Las Encinas.

Para efectos de lectura, los datos están resumidos en la figura 4.2, donde Corrx es el coeficiente de correlación cruzada de Pearson, R es el coeficiente R^2 y Px corresponde a la prueba desde la 1 hasta la 5.

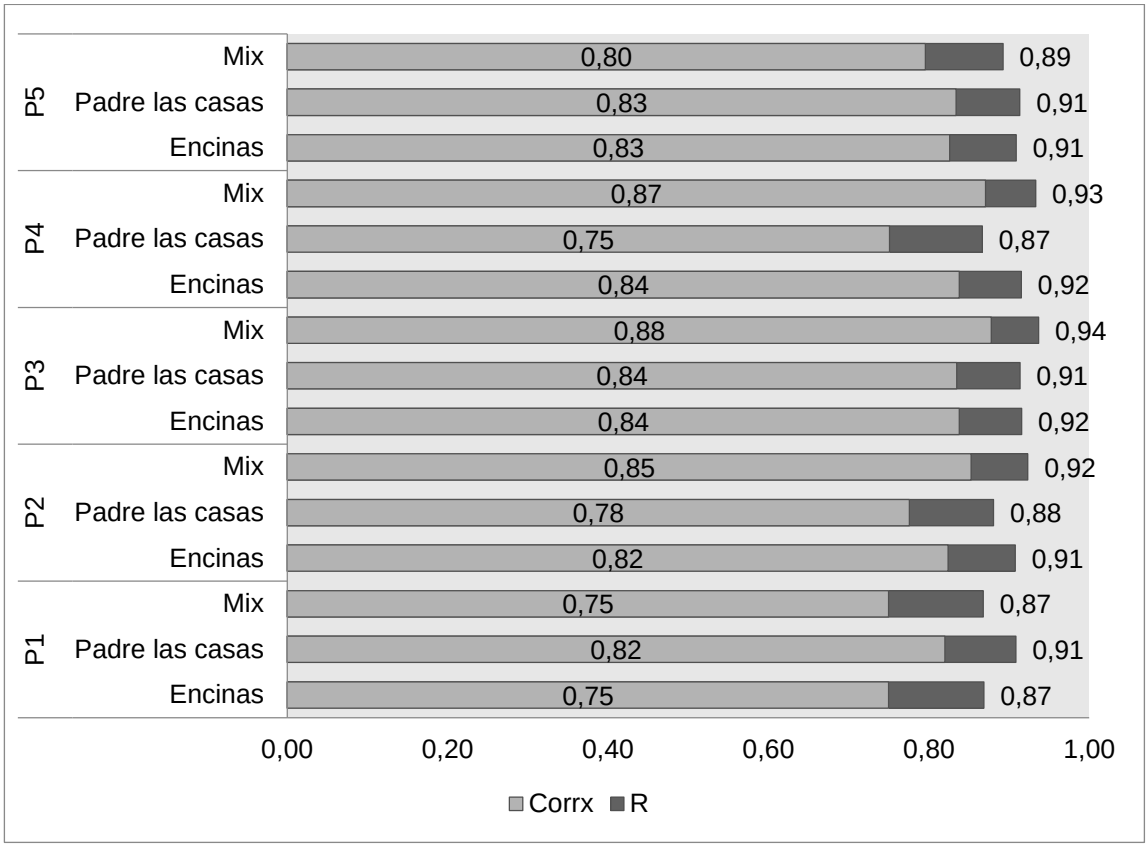

Figura 4.2: Correlación cruzada y coeficiente R de las diferentes redes neuronales artificiales.

Se puede observar de la figura 4.2 que la prueba con mejor resultado general es la prueba n° 3, como se evidenció en capítulos anteriores, la estación “Padre las Casas” es la de mayor interés ya que es la que contiene una mayor cantidad de episodios críticos y de contingencia. Por lo tanto es importante considerar esta estación como base del modelo predictivo. Así las pruebas con mayor éxito en padre las casas son: P1, P3 y P5. Por otro lado las pruebas con mejor desempeño en datos mixtos son las: P2, P3 y P4. Lo que se condice con la perspectiva general de que la prueba 3 es la de mejor desempeño.

4.2 Análisis mediante tablas de contingencia

Un método útil para el análisis de estos resultados es utilizar una tabla de contingencia, de esta manera se puede comparar los datos por tramos, es decir se puede desglosar los datos y obtener el % de acierto de acuerdo a cierto tramo. Como se ha visto en el capítulo 1, los tramos pueden ser clasificados por niveles de concentración, tal como lo muestra la tabla 1.3 en bueno (B), regular (R), alerta (A), preemergencia (PE) y emergencia (E). La tabla de contingencia es un método poderoso para obtener la información del % de predictibilidad de la red, ya que índica de manera explícita el funcionamiento de la red.

En la figura 4.3 se puede observar el % de predictibilidad general de las tablas de contingencia

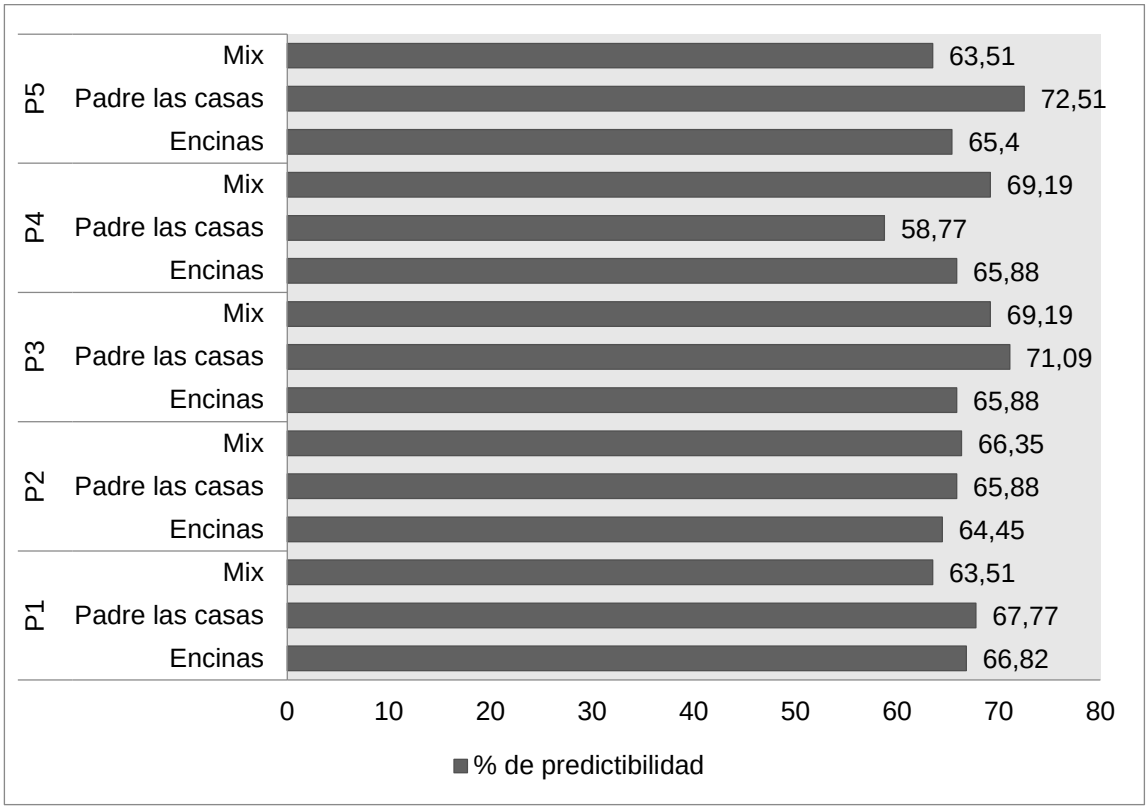

Figura 4.3: % de predictibilidad de las diferentes redes neuronales utilizando tablas de contingencia

De la figura 4.3 se observa que los modelos con un mayor porcentaje de predictibilidad, con especial énfasis en los datos obtenidos de estación padre las casas son los modelos basados en las pruebas P3 y P5. Por lo tanto, se procedió a descartar los modelos basados de la prueba P1, P2 y P4. Adicionalmente, se descartarán los modelos entrenados en estación las encinas por obtener bajos porcentajes de predictibilidad en estas pruebas.

Así, se procederá a analizar los modelos restantes mediante su tabla de contingencia completa, donde se comparan datos observados con los datos pronosticados.

Tabla 4.1: Tabla de contingencia modelo P3 para Padre las Casas

		Datos PRONOSTICADOS						
		B	**R**	**A**	**PE**	**E**	**Total**	**%**
Datos OBS	**B**	50	3	0	0	0	53	94,3
	R	12	32	7	3	0	54	59,3
	A	0	4	21	18	1	44	47,7
	PE	0	0	2	33	11	46	71,7
	E	0	0	0	0	14	14	100,0
	Total	62	39	30	54	26	211	
	%	80,65	82,05	70,00	61,11	53,85		**71,09**

Tabla 4.2: Tabla de contingencia modelo P3 para datos mixtos

		Datos PRONOSTICADOS						
		B	**R**	**A**	**PE**	**E**	**Total**	**%**
Datos OBS	**B**	48	3	0	0	0	51	94,1
	R	14	33	7	4	0	58	56,9
	A	0	2	18	20	1	41	43,9
	PE	0	0	5	31	9	45	68,9
	E	0	0	0	0	16	16	100,0
	Total	62	38	30	55	26	211	
	%	77,42	86,84	60,00	56,36	61,54		**69,19**

Tabla 4.3: Tabla de contingencia modelo P5 para Padre las Casas

		Datos PRONOSTICADOS						
		B	**R**	**A**	**PE**	**E**	**Total**	**%**
Datos OBS	**B**	51	4	0	0	0	55	92,7
	R	11	33	3	0	0	47	70,2
	A	0	2	23	19	1	45	51,1
	PE	0	0	4	33	12	49	67,3
	E	0	0	0	2	13	15	86,7
	Total	62	39	30	54	26	211	
	%	82,26	84,62	76,67	61,11	50,00		**72,51**

Tabla 4.4: Tabla de contingencia modelo P5 para datos mixtos

		Datos PRONOSTICADOS						
		B	**R**	**A**	**PE**	**E**	**Total**	**%**
Datos OBS	**B**	50	3	0	0	0	53	94,3
	R	12	33	14	8	0	67	49,3
	A	0	2	14	22	3	41	34,1
	PE	0	0	2	25	11	38	65,8
	E	0	0	0	0	12	12	100,0
	Total	62	38	30	55	26	211	
	%	80,65	86,84	46,67	45,45	46,15		**63,51**

Como se puede observar en las tablas anteriores, el modelo con mejor desempeño en general es el modelo de P3. Se observa claramente que el modelo P5 con entrenamiento de datos mixtos es inferior en porcentaje de predictibilidad del PM2,5, por lo tanto se descarta este modelo.

Ya que se busca un modelo efectivo pero sencillo, utilizar datos mixtos como se observan en las tablas de la 4.1 a la 4.3 son los de menor efectividad general y no generan una diferencia significativa, y además son los que más cantidad de datos de entrada requiere, por lo tanto, por sencillez descartaremos el modelo P3 con entrenamiento de datos mixtos.

4.3 Resumen de resultados

De acuerdo a los resultados obtenidos en la sección 4.2, se procederá a discriminar entre el modelo de entrenamiento P3 y P5 con datos de estación Padre las Casas. Para esto, ya que lo importante del modelo predictivo es determinar la predictibilidad de episodios críticos (ya que es en estos donde se requieren tomar medidas o acciones de mitigación), se compararán en el índice de predictibilidad de alertas, preemergencias y emergencias.

En resumen, se tiene para P3 y P5 los siguientes parámetros de entrada:

- PM2,5 promedio horario y promedio móvil 24h de las 18:00h.
- Temperatura ambiente promedio horario y promedio móvil 24h de las 18:00h.
- Velocidad del viento promedio horario y promedio móvil 24h de las 18:00h.
- Temperatura ambiente mínima del día siguiente.
- PM10 promedio horario y promedio móvil 24h de las 18:00h.
- PM2,5 promedio horario y promedio móvil 24h de las 17:00h (Solo P5).

Así, se procede a comparar estos dos modelos neuronales en la tabla 4.5.

Tabla 4.5: Comparación final de modelos de redes neuronales con mejor desempeño

Parámetro	*P3*	*P5*	*Δ%*	*Δepisodios*
Parámetros de Entrada	9 neuronas	11 neuronas	-	-
% Alertas Observadas	47,7	51,1	3,4	1,0
% Alertas Pronosticadas	70,0	76,7	6,7	3,0
% PE Observadas	71,7	65,8	5,9	3,2
% PE Pronosticadas	61,1	45,5	15,6	7,2
% Emergencias Observadas	100	86,7	13,6	3,5
% Emergencias Pronosticadas	53,8	50,0	3,8	0,5
% Predictibilidad Total	71,1	72,5	1,4	2,9
Δepisodios totales críticos	-	-	-	*10,4*

Donde Δ% corresponde a la diferencia porcentual del porcentaje de aciertos del modelo neuronal, y Δepisodios nos indica cuantos episodios de diferencia se pronosticaron correctamente entre cada una de las redes neuronales. Esta última información nos indica el peso real entre la diferencia de porcentajes. El Δepisodios totales críticos índica la diferencia de episodios críticos entre cada uno de las redes neuronales, donde el modelo P3 aventaja en 10,4 episodios críticos pronosticados en comparación con el modelo P3.

Finalmente, el % de alertas, preemergencias (PE) y emergencias observadas corresponde a la cantidad de episodios correctamente pronosticados sobre la cantidad total de episodios reales (por ejemplo, alertas correctas sobre el total de alerta de los datos reales). En cambio el % de alertas, preemergencias (PE) y emergencias pronosticadas corresponde a la cantidad de episodios correctamente pronosticados sobre la cantidad total de episodios pronosticados (por ejemplo, alertas correctas sobre el total de alertas pronosticadas por el modelo).

Con estos dos valores se puede obtener un % de falsas predicciones de episodios, el cual en promedio es del 26,3% para el modelo P3 y del 27,6% para el modelo P5

Se observa de la tabla 4.5 que el modelo P3 entrenado con datos de la estación de monitoreo Padre las Casas es el que mejor desempeño tiene con los episodios críticos y cuenta con menos parámetros de entrada, por lo tanto es el recomendado para concluir resultados sobre un modelo alternativo predictivo.

Para comparar estos resultados con el modelo WRF-MMA es necesario agrupar los datos en episodios (sobre los 80μg/m^3) y no episodio (bajo los 80μg/m^3), así se obtiene que el el porcentaje de predictibilidad de episodios es del 91% de predictibilidad de episodios y de un 4% de falsas alarmas (ver tabla 4.1).

4.4 Análisis de red utilizando datos de gases criterio NO_2 y CO

Un análisis interesante para este caso es utilizar los datos de gases criterio para la red neuronal. Dos de los contaminantes medidos en la estación Padre las Casas II son el dióxido de nitrógeno NO_2 y el monóxido de carbono CO. Para el caso del NO_2 , siendo precursor de la formación de partículas secundarias (nitratos), su mayor o menor presencia en horas previas a la hora del pronóstico, podría influir en las concentraciones observadas de MP2.5.

Para el caso de los gases, debido a la disponibilidad de datos entregados por SINCA, no es factible considerar los datos del año 2014 para el entrenamiento ya que existe un periodo de tiempo extenso en el cual no se realizó la medición de estos datos. Por lo tanto, se utilizó los datos del año 2015 para entrenamiento y validación y el año 2016 para prueba de la red.

Para ingresar los datos de las concentraciones de NO_2 y CO al modelo de la red, se utilizó el dato promedio móvil 10 horas de las 18:00h. Esto debido a que las concentraciones de NO_2 y CO se mantienen estables en ese período de tiempo (entre las 9:00 y las 19:00 horas). Estos datos fueron añadidos al modelo tipo P3 con datos de estación Padre las Casas II descrito en los capítulos anteriores, dando como resultado tres configuraciones nuevas del modelo de red neuronal.

- Modelo con concentración de CO en promedio móvil horario 10h.
- Modelo con concentración de NO2 en promedio móvil horario 10h.
- Modelo con concentración de CO y NO_2 en promedio móvil horario 10h.

Los resultados resumidos en estos modelos se pueden observar en la figura 4.4. De esta se puede observar que los resultados de predictibilidad bajan notoriamente al utilizar esta configuración (menos datos de entrenamiento y la utilización de gases criterio); Por lo que bajo estas condiciones, no es recomendable utilizar las concentraciones de gases para predecir la concentración de PM2,5.

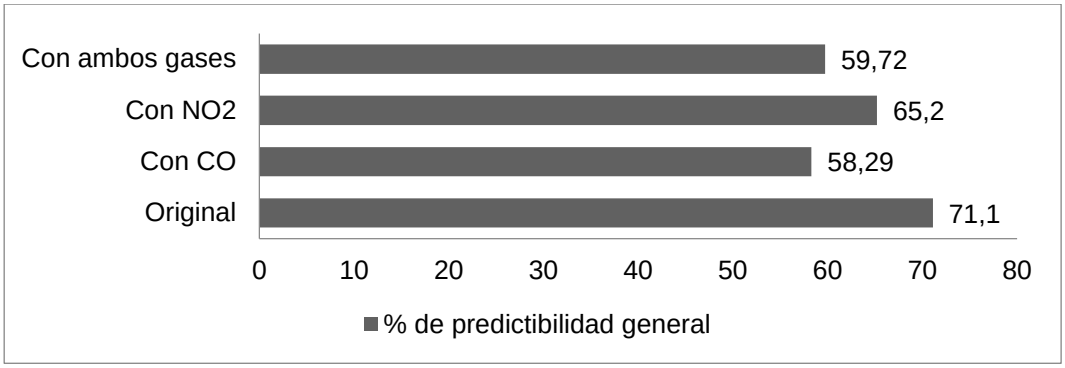

Figura 4.4: % de predictibilidad general de las redes neuronales obtenidas usando gases criterio mediante tablas de contingencia.

Para analizar de mejor manera la información, se analizará la predictibilidad en caso de episodios críticos, en las figuras 4.5, 4.6 y 4.7 se muestra la predictibilidad de los episodios de emergencia, preemergencia y alerta respectivamente.

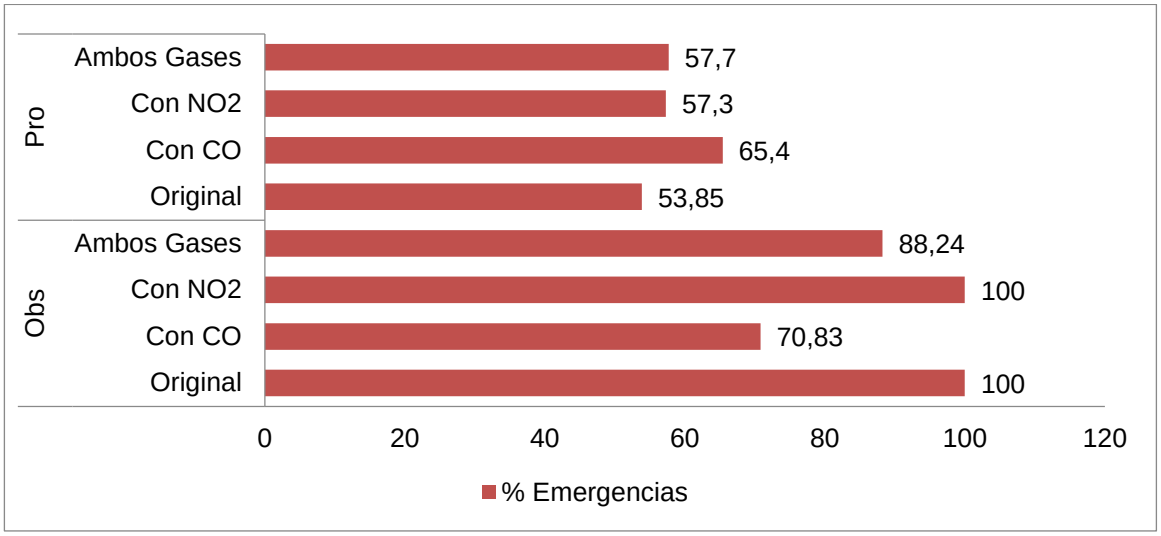

Figura 5: % de predictibilidad de emergencias de las redes neuronales obtenidas usando gases criterio mediante tablas de contingencia.

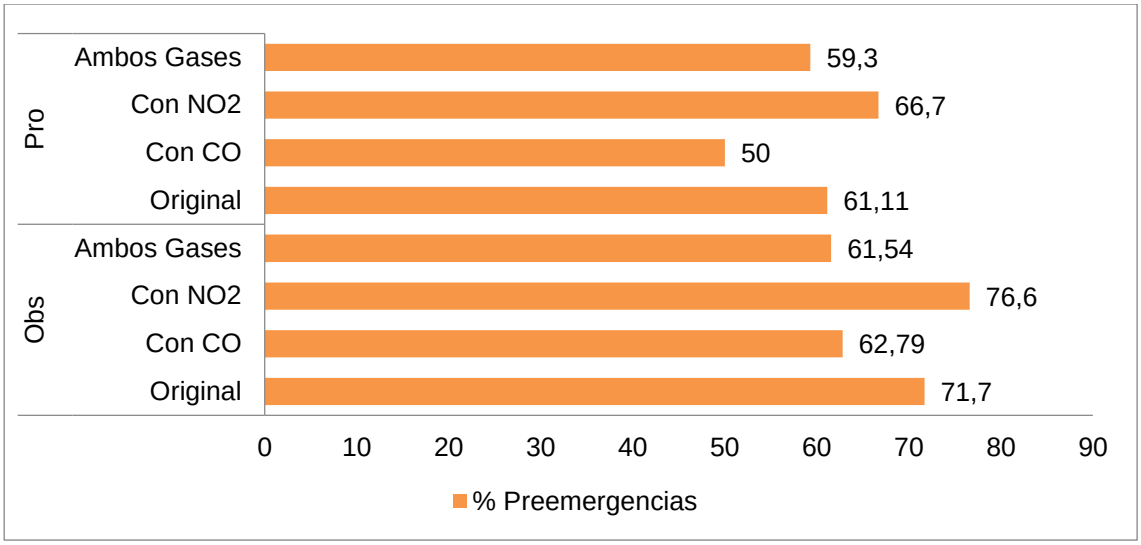

Figura 6: % de predictibilidad de preemergencias de las redes neuronales obtenidas usando gases criterio mediante tablas de contingencia.

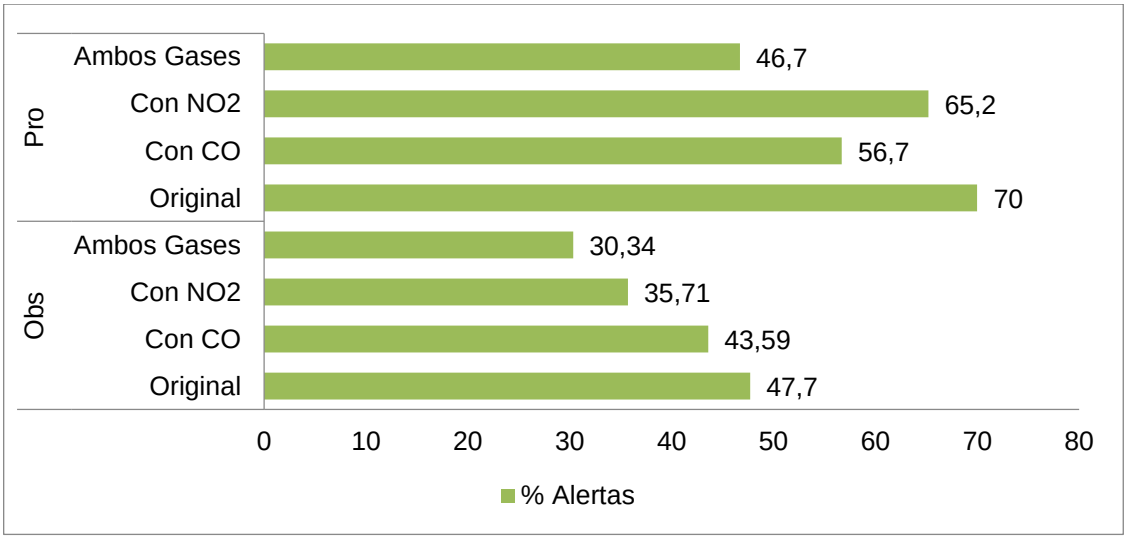

Figura 7: % de predictibilidad de alertas de las redes neuronales obtenidas usando gases criterio mediante tablas de contingencia.

De las figuras anteriores se observa que el porcentaje de predicción para emergencias y preemergencias se mantiene dentro de rangos cercanos, inclusive en algunos casos superando al modelo original (Como el caso de las preemergencias y emergencias utilizando NO2), sin embargo el porcentaje de predictibilidad de las alertas baja. Aun así, el mejor resultado de los gases se obtiene utilizando el NO_2 en el modelo predictivo. Sin embargo, para este resultado solo se considerará el NO_2 como posible caso de estudio futuro debido a la baja predictibilidad general.

Conclusiones

“It has come to my attention, that air pollution is polluting the air!”

--George W. Bush

Se logró construir un modelo predictivo de redes neuronales artificiales para la concentración de material particulado fino PM2,5 en la ciudad de Temuco. Los mejores resultados, y los más relevantes para el estudio se obtuvieron entrenando la red con datos de la estación de monitoreo *“Padre las Casas”,* la cual está ubicada en la comuna de Padre las Casas en Temuco. Esta relevancia está justificada por la cantidad de episodios críticos medidos por esta estación, ya que la mayor cantidad de episodios críticos para esta ciudad fueron determinados mediante esta estación.

La mejor arquitectura determinada en este estudio para la red neuronal artificial fue de 9 neuronas de entradas, con los siguientes datos de la red: PM2,5 promedio horario y promedio móvil 24h de las 18:00h; Temperatura ambiente promedio horario y promedio móvil 24h de las 18:00h; Velocidad del viento promedio horario y promedio móvil 24h de las 18:00h; Temperatura ambiente mínima del día siguiente; PM10 promedio horario y promedio móvil 24h de las 18:00h.

Con respecto a la capa oculta de la red, se determinó mediante el criterio de información de Akaike que la cantidad óptima de neuronas artificiales en la capa oculta es de 18 neuronas, con esto se evita el sobre-entrenamiento o sub-entrenamiento de la red. Otro parámetro importante que se determinó es un delay o información previa a utilizar en la red es d=1. Esto quiere decir que se comparan los datos del día anterior y el día presente para pronosticar el día siguiente. Donde el parámetro de salida es el mayor promedio móvil 24h del día siguiente.

Con estos parámetros, se logró construir una red con un porcentaje de predictibilidad total del 71,1%, con especial énfasis de la predictibilidad de episodios críticos donde el error de predictibilidad varía entre el 52,2% (para episodios de alerta) como máximo y el 0% (para episodios de emergencia), esta información está detallada en las tablas 4.1 y 4.5.

Finalmente, en esta última red de mayor porcentaje de predictibilidad se probó incorporar los datos de gases CO y NO_2 de manera individual y simultánea, sin embargo, los resultados empeoraron el porcentaje de predictibilidad total. Cabe destacar que para la prueba de gases solo se utilizó para entrenamiento y validación los datos del año 2015, esto debido a la disponibilidad de datos de la estación Padre las Casas II, por lo tanto la utilización de gases para la predicción de la calidad del aire se desechó. Sin embargo, la utilización del NO_2 mostró

una buena predictibilidad en episodios críticos de contaminación, por lo que es un precedente a utilizar para futuros estudios.

Si bien, la red neuronal construida tiene una predictibilidad aceptable y los resultados indican un modelo alternativo para la ciudad de Temuco, no es posible afirmar que este modelo tenga un mejor rendimiento que el modelo oficial existente (WRF-MMA) el cual tiene un porcentaje de predictibilidad global de episodios para la estaciones de Padre las Casas II de un 73%[5] con un porcentaje de falsas alarmas de un 23%. Estos resultados correponden a un análisis de los datos de 2014 reportados en Saide et al, 2016. Al comparar los episodios del mismo modo que el modelo WRF-MMA (episodio sobre 80µg/m^3 independiente de la magnitud del episodio), la red construida en este estudio alcanza un 91% de predictibilidad de episodios y solo un 4% de falsas alarmas. Constituyendo una mejora notable en los resultados. Para poder hacer una comparación más concluyente, sería necesario correr los dos modelos simultáneamente en línea.

Por lo tanto, se recomienda este modelo como una alternativa o un complemento para las decisiones políticas y los decretos de episodios críticos para la ciudad de Temuco.

Debido a las limitaciones de este estudio, queda abierto la posibilidad en el futuro (cuando exista la disponibilidad de datos de medición), repetir este estudio incorporando un mayor número de datos al entrenamiento y series más completas de los gases CO y NO_2. Y también la incorporación al estudio de los datos de la nueva estación “Cerro Ñielol”.

[5] Valor de la persistencia de la predictibilidad en el tiempo.

Referencias Bibliográficas

Barai, S., Dikshit, A., & Sharma, S. Neural Network Models for Air Quality Prediction: A Comparative Study. Advances In Soft Computing, 290-305. http://dx.doi.org/10.1007/978-3-540-70706-6_27

Comisión Nacional del Medio Ambiente de la Araucanía. (2017). Actualización del inventario de emisiones atmosféricas en las comunas de Temuco y Padre las Casas (pp. 204-218). Temuco: Ingeniería DICTUC.

Dirección de planificación territorial de la comuna de Temuco. (2017). Diagnóstico sistémico territorial (pp. 2-9). Temuco: Municipalidad de Temuco.

Dockery, D. (1993). Epidemiologic Study Design for Investigating Respiratory Health Effects of Complex Air Pollution Mixtures. Environmental Health Perspectives, 101, 187. http://dx.doi.org/10.2307/3431676

Fenger, J. (1999). Urban air quality. Atmospheric Environment, 33(29), 4877-4900. http://dx.doi.org/10.1016/s1352-2310(99)00290-3

Freeman, J., & Skapura, D. (1994). Neural networks. Reading, Mass. [u.a.]: Addison-Wesley.

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment, 32(14-15), 2627-2636. http://dx.doi.org/10.1016/s1352-2310(97)00447-0

Guevara Díaz, José Manuel; (2014). Uso correcto de la correlación cruzada en Climatología: el caso de la presión atmosférica entre Taití y Darwin. *Terra Nueva Etapa,* Enero-Junio, 79-102.

Kandel, E., & Mack, S. Principles of neural science.

Martín del Brío, B., & Sanz Molina, A. (2006). Redes neuronales y sistemas borrosos. Madrid: Ra-Ma.

Mc. Cullogh y Pitts, W. (1943). The linear theory of neuron networks: The dynamic problem. *The Bulletin Of Mathematical Biophysics*, *5*(1), 23-31. http://dx.doi.org/10.1007/bf02478116

MacKay, Neural Computation, Vol. 4, No. 3, 1992, pp. 415–447

Morales Segura, R. (2006). Contaminación atmosférica urbana. Santiago de Chile: Universidad de Chile, Centro de Química Ambiental.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. IEEE Transactions On Neural Networks, 5(6), 865-872. http://dx.doi.org/10.1109/72.329683

Organización mundial de la Salud. (2017). Guías de calidad del aire de la OMS relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre. (pp. 4-7). Ginebra, Suiza: Ediciones OMS.

Panchal, G., Ganatra, A., Kosta, Y., & Panchal, D. (2010). Searching Most Efficient Neural Network Architecture Using Akaike's Information Criterion (AIC). International Journal Of Computer Applications, 1(5), 54-57. http://dx.doi.org/10.5120/126-242

Peña Sánchez de Rivera, D. (1992). Estadística modelos y métodos. Vol. 1, Fundamentos. Madrid: Alianza Editorial.

Pérez, P., Trier, A., & Reyes, J. (2000). Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment, 34(8), 1189-1196. http://dx.doi.org/10.1016/s1352-2310(99)00316-7.

Pope, C., & Dockery, D. (1992). Acute Health Effects of PM10Pollution on Symptomatic and Asymptomatic Children. American Review Of Respiratory Disease, 145(5), 1123-1128. http://dx.doi.org/10.1164/ajrccm/145.5.1123

Saide, P., Mena-Carrasco, M., Tolvett, S., Hernandez, P., & Carmichael, G. (2016). Air quality forecasting for winter-time PM2.5 episodes occurring in multiple cities in central and southern Chile. Journal Of Geophysical Research: Atmospheres, 121(1), 558-575. http://dx.doi.org/10.1002/2015jd023949

Sohn, S., Oh, S., & Yeo, Y. (1999). Prediction of air pollutants by using an artificial neural network. Korean Journal Of Chemical Engineering, 16(3), 382-387. http://dx.doi.org/10.1007/bf02707129

Ulriksen P., Merino M. y Llano A.; (2001) Informe Técnico, Análisis comparativo de condiciones meteorológicas asociadas a episodios de contaminación atmosférica en Santiago durante los períodos otoño-invierno 1997, 1998, 1999, 2000 y 2001. Laboratorio de Modelación y Contaminación Atmosférica, CENMA.