UNIVERSIDAD DE TALCA

FACULTAD DE INGENIERÍA

ESTIMACIÓN DE LA CALIDAD DEL AIRE EN LA CIUDAD DE TALCA UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Eduardo Alejandro Zapata González

Tesis para optar al grado de
Magíster en Gestión de Operaciones

Prof. Supervisor César Astudillo H.

Junio, 2017

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su encargado Biblioteca Campus Curicó certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Curicó, 2019

A mí querida esposa y mis amadas pequeñas...

RESUMEN EJECUTIVO

Actualmente, muchas ciudades del centro y sur de Chile presentan un problema estacional (otoño-invierno) de contaminación del aire por material particulado, asociado a las características meteorológicas de los emplazamientos, y por otra parte, a altas emisiones de calefactores domiciliarios a leña. Sin embargo, los estudios de predicción de la calidad del aire se han concentrado en la ciudad de Santiago de Chile, observándose sólo dos estudios en ciudades de tamaño medio del centro y sur (Saide, Mena-Carrasco, et al. 2016).

El presente trabajo se ha enfocado en generar un modelo de predicción de la calidad del aire en la ciudad de Talca (35°26'S; 71°44'W), ciudad de tamaño medio, localizada en la zona central de Chile. El objetivo del estudio es predecir con un día de anticipación el material particulado de diámetro igual o menor a 2.5 micrómetros ($MP_{2,5}$), para lo cual hemos utilizando tres modelos de aprendizaje automático: Perceptrón Multicapa (MLP), redes neuronales semisupervisadas basado en Self-organizing map (SOM) y Random Forest (RF).

Los modelos utilizan una base de datos construida a partir de los registros históricos de calidad del aire, además de registros meteorológicos de velocidad del viento y temperatura en tres estaciones de medición de la ciudad. En tanto, los procesos de modelación consideran una búsqueda exhaustiva de parámetros.

Nuestros resultados comprueban la alta capacidad de Random Forest como algoritmo predictor de episodios de emergencia ambiental en las tres estaciones de monitoreo de la ciudad, superando a los algoritmos de redes neuronales artificiales, reportados con los mejores resultados en la bibliografía. El modelo es una alternativa para las autoridades locales de la ciudad, ya que permitiría mejorar el sistema actual de pronósticos de calidad del aire.

Palabras Claves: Estimación de calidad del aire, Material particulado, Aprendizaje automático, Talca.

EXTENDED ABSTRACT

At present, many cities in central and southern Chile have a seasonal (autumn-winter) problem of air pollution by particulate matter, associated with the meteorological characteristics of the sites and high emissions of domestic wood-burning stoves. However, air quality forecasts studies have been concentrated in Santiago-Chile, observing only two studies in medium-sized cities in the center and south from Chile (Saide, Mena-Carrasco, et al. 2016).

The present paper has focused on generating a model of air quality forecasts in the city of Talca (35°26'S; 71°44'W), medium-sized city, located in the central zone from Chile. The aim of the study is to predict one-day-ahead particulate matter with a diameter equal to or less than 2.5 micrometers ($PM_{2.5}$), for which we have used three models of machine learning: Multi-layer Perceptron (MLP), semi-supervised neural networks based on Self-organizing map (SOM) and Random Forest (RF).

The models use a database constructed from the historical air quality records and meteorological records (wind speed and temperature) in three measurement stations of the city. Modeling processes consider a comprehensive search for parameters.

Our results verify the high capacity of Random Forest as an algorithm predictor of environmental emergency episodes in the three monitoring stations of the city, surpassing the algorithms artificial neural networks, reported with the best results in the literature. The model is an alternative for local authorities in the city, as it would improve the current system of air quality forecasts.

Keywords: Air quality forecasting, Particulate matter, machine learning, Talca.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

1.1. Contexto del Problema

La mala calidad del aire por altas concentraciones de Material Particulado afecta la calidad de vida de la población de las comunas de Talca y Maule. Los efectos de la contaminación atmosférica en las ciudades del centro y sur de Chile se han asociado significativamente con la mortalidad prematura y otros efectos negativos para la salud (Díaz-Robles, Cortés, et al. 2015), (Sanhueza, et al. 2009).

El material particulado con diámetro inferior a los 10 micrometros (MP_{10}) corresponde a un contaminante primario que afecta la calidad del aire, y por lo reducido de su tamaño puede ingresar al sistema respiratorio. Un subconjunto de este contaminante corresponde al material particulado con diámetro inferior a los 2,5 micrometros ($MP_{2,5}$), el cual genera mayores daños a la salud que el MP_{10}, por su menor diámetro y su capacidad de penetrar los alvéolos pulmonares.

En la zona central y de sur de Chile, la mala calidad del aire se observa en las ciudades del valle central, principalmente por las características meteorológicas, donde se observan períodos invernales con una alta estabilidad atmosférica en los días que registran las menores temperaturas. En tanto, la mayor fuente emisora de $MP_{2,5}$ en las ciudades al sur de Santiago es la calefacción domiciliaria a leña.

En la ciudad de Talca, el año 2004 se inicia el monitoreo de la calidad del aire, con la instalación de una estación de medición discreta de MP_{10}. El año 2010 se declararon como zonas saturadas por MP_{10} las comunas de Talca y Maule (Decreto N°12 de 04-02-2010 del Ministerio Secretaria General de la Presidencia), por la superación de las normas primarias de concentración de MP_{10} como concentración anual los años 2007 y 2008, como promedio aritmético de tres años calendarios consecutivos para los años 2005, 2006, 2007 y 2008; y excedida como concentración de 24 horas para los mismos años.

En Abril de 2013, se inauguran tres estaciones de monitoreo de calidad del aire, ubicadas en Universidad de Talca (35°24'24''S; 71°37'60''W), Universidad Católica del Maule (35°24'24''S; 71°37'60''W) y Cesfam La Florida (35°26'07''S; 71°40'41''W), las cuales capturan en forma continua los parámetros contaminantes MP_{10} y $MP_{2,5}$ y en frecuencia horaria los parámetros meteorológicos: presión atmosférica, humedad relativa del aire, temperatura ambiental, dirección y velocidad del viento. Además, en la Estación Cesfam La Florida se miden de forma continua los parámetros contaminantes: Dióxido de azufre (SO2), monóxido de nitrógeno (NO), dióxido de nitrógeno (NO2), monóxido de carbono (CO), ozono (O3) y óxidos de nitrógeno (NOx).

Durante el primer año de medición, se contabilizaron un total de 86 días con superación de la norma de $MP_{2,5}$; existiendo 38 días con calidad del aire regular, 25 días con alerta, 17 días con preemergencia y 6 días con emergencia ambiental, por lo que el Ministerio de Salud emite las Resoluciones Sanitarias Exentas N° 2093, de fecha 7 de Mayo de 2014, y N° 1742, de fecha 14 de

Abril de 2015, donde se establece que en períodos de pre-emergencia y emergencia ambiental se prohíbe el funcionamiento de fuentes fijas comerciales, industriales, residenciales, domiciliarias y comunitarias, que usen como combustible leña y otros derivados tales como aserrín, viruta, piñas, etc., cuya combustión genere emisiones atmosféricas de Material Particulado, sea a través de descargas directas o ductos, desde las 18:00 hrs. hasta las 23:59 hrs, en tres polígonos circundantes a las estaciones de monitoreo de la ciudad de Talca y a la ciudad de Curicó.

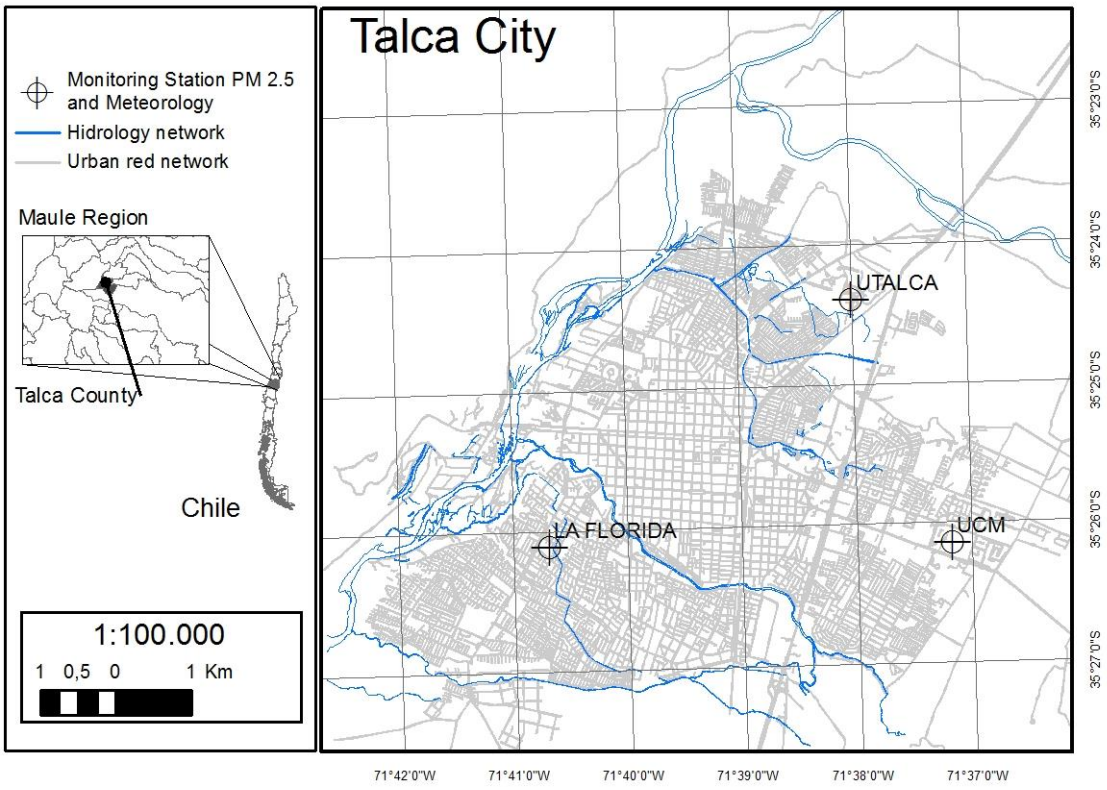


Figura 1: Estaciones de monitoreo de calidad del aire en la ciudad de Talca

Según el valor de concentración de $MP_{2,5}$, la calidad del aire se puede clasificar según los niveles críticos por la norma primaria de calidad ambiental para material particulado no respirable $MP_{2,5}$ (DS 12 de 01-01-2012 del Ministerio de Medio Ambiente), detallado en la Tabla 1:

Tabla 1: Medidas de emergencia ambiental según norma primaria de calidad

Nivel	Concentración $MP_{2,5}$ ($\mu g/m^3$)
Alerta	80-109
Preemergencia	110-169
Emergencia	170 o más

De acuerdo a este criterio, existen tres niveles de concentración $MP_{2,5}$ donde las autoridades realizan medidas de contingencia (ver Tabla 1). El primer nivel denominado “Alerta”, se produce cuando la concentración de $MP_{2,5}$ se encuentra entre 80 y 109 $\mu g/m^3$. La siguiente categoría denominada “Preemergencia” ocurre entre los 110 y 169 $\mu g/m^3$. Finalmente la categoría más dañina para el organismo, denominada “Emergencia” ocurre cuando la concentración es mayor o igual a 170 $\mu g/m^3$.

Para cualquier medida de contingencia, es necesario estimar cuál será la condición del aire, al menos con un día de anticipación, y de esta forma, reducir los efectos adversos en la salud de la población.

1.2. Objetivos

1.2.1. Objetivo General

- Predecir la concentración del promedio móvil 24 horas de $MP_{2,5}$ en el aire, monitoreado en las estaciones Universidad de Talca, Universidad Católica del Maule y Cesfam La Florida de la comuna de Talca.

1.2.2. Objetivos Específicos

a) Recopilar la literatura asociada a estudios de estimación de calidad del aire en ciudades chilenas.
b) Recopilar la literatura asociada a algoritmos de aprendizaje automático utilizados para resolver problemas de calidad de aire.
c) Modelar la contaminación atmosférica del aire por $MP_{2,5}$ en la ciudad de Talca, para el período 2014-2015, utilizando tres algoritmos de aprendizaje automático.
d) Evaluar los resultados en las tres estaciones de monitoreo de la ciudad de Talca, comparando el desempeño entre modelos.

1.3. Contribuciones de la Tesis

a) A través de tres algoritmos, estimamos la calidad del aire de $MP_{2,5}$ en la ciudad de Talca.
b) Utilizamos tres algoritmos de aprendizaje automático: perceptrón multicapa, un algoritmo basado en Self-organizing maps, denominado TTOSOM, y Random Forest para enfrentar la naturaleza no lineal del problema.
c) Construimos un modelo simple, ya que sólo incorpora dos variables meteorológicas: los promedios de temperatura y de velocidad del viento.
d) El algoritmo Random Forest obtiene buenos resultados de regresión, superando significativamente los resultados obtenidos por MLP y TTOSOM, algoritmos reportados con los mejores resultados en la bibliografía.

1.4. Organización de la Tesis

El trabajo se encuentra organizado de la siguiente forma. En las secciones 1.5 y 1.6 se presenta el problema de estimación de la calidad del aire en la ciudad de Talca y un resumen de la literatura acerca de la utilización de algoritmos de aprendizaje automático en la estimación de problemas de calidad del aire. La sección 1.7 muestra la metodología de trabajo propuesta, iniciando en la construcción de la base de datos, luego la modelación del problema de estimación utilizando tres algoritmos de aprendizaje automático. La sección 1.8 compara los resultados de los tres modelos, y finalmente la sección 1.9 las conclusiones.

1.5. El problema de Estimación de calidad del aire

Este problema se refiere al cambio de la composición química del aire por emisiones antropogénicas a la atmósfera. El problema de predicción de calidad del aire ha incrementado su importancia en los últimos años debido a que afecta en el corto y largo plazo el bienestar humano (Carmichaela, et al. 2008). El problema de predicción de calidad del aire presenta grandes incertidumbres asociadas a información incompleta, emisiones inexactas y procesos pobremente parametrizados (Carmichaela, et al. 2008).

Los problemas de predicción de la calidad de aire se han resuelto utilizando modelos deterministas y estadísticos. Los modelos deterministas no demandan una gran cantidad de datos históricos, sin embargo exigen un conocimiento de las condiciones meteorológicas y las fuentes contaminantes, la cantidad de emisiones en tiempo real, la descripción explícita de las principales reacciones químicas y procesos físicos en la capa inferior de la atmósfera.

Los modelos estadísticos requieren una gran cantidad de datos históricos en diversas condiciones atmosféricas. Diferentes funciones se pueden utilizar para establecer las respectivas relaciones entre los datos de contaminantes y los predictores utilizando métodos de regresión y de aprendizaje automático. El principal inconveniente de este enfoque es que un modelo representa sólo una estación específica y no se puede extender a otras regiones con diferentes condiciones meteorológicas. Sin embargo, el enfoque estadístico es generalmente más apropiado para el descubrimiento de dependencias complejas de un sitio específico entre las concentraciones de contaminantes atmosféricos y predictores potenciales y a menudo tienen una mayor precisión, que los modelos deterministas.

Los métodos estadísticos más utilizados incluyen la regresión lineal múltiple (MLR), ANN, Support Vector Machine (SVM), Fuzzy logic, filtro de Kalman (KF) y el modelo oculto de Markov (HMM) (Feng, et al. 2015).

Un inconveniente común con estos modelos, es que durante los días con altas concentraciones de material particulado (MP), los errores de previsión tienden a ser mucho mayores, y las concentraciones de MP son subestimadas sistemáticamente (Feng, et al. 2015).

1.5.1.El problema de estimación de calidad del aire en la ciudad de Talca

En Chile, la ciudad que ha centralizado la mayor atención con respecto a los problemas de calidad del aire, es la ciudad de Santiago, donde el gobierno ha desarrollado planes de descontaminación del aire desde el año 1997. Esta política ha generado una activa investigación de modelos de pronóstico de calidad del aire, siendo ejemplo de ello (Cassmassi 1999), (Perez and Reyes 2002), (Pérez and Salini 2008), (Saide, Carmichael, et al. 2011).

Según lo señalado en (Saide, Mena-Carrasco, et al. 2016), las ciudades de tamaño medio al sur de Santiago también han presentado episodios de alta polución del aire, sin embargo, han recibido una menor atención. En estas ciudades, los modelos de pronóstico de calidad del aire han sido limitados, solo reportándose hasta el año 2016 un estudio en la ciudad de Temuco (Díaz-Robles, Ortega, et al. 2008).

Por las razones anteriormente expuestas, en (Saide, Mena-Carrasco, et al. 2016), se desarrolla un modelo de pronóstico para nueve ciudades con monitoreo regular de la calidad del aire: Santiago, Rancagua, Curicó, Talca, Chillán, Los Ángeles, Temuco, Valdivia y Osorno. El modelo tiene como base el modelo químico y meteorológico en línea “Weather Research and Forecasting with Chemistry” (WRF-Chem) (Saide, Carmichael, et al. 2011), el cual predice concentración de CO. Los autores aproximan la predicción de $MP_{2,5}$ con la alta correlación de concentración de CO en los episodios críticos.

El sistema utiliza el modelo WRF-Chem, configurado para una grilla con celdas de 2 km para predecir el tiempo y las concentraciones de $MP_{2,5}$ por hora, utilizando una calibración basada en la observación de variables locales, lo que permite un modelo menos intensivo computacionalmente que un modelo netamente químico.

Las variables locales observadas incluyen una mayor probabilidad de ocurrencia de episodios críticos durante los fines de semana y los días más fríos, esto último relacionado con el aumento de las emisiones de las estufas a leña.

El modelo pronostica eventos con tres días de anticipación, sin embargo a menudo se presentan una gran variabilidad entre los pronósticos, debido a diferentes inicializaciones meteorológicas.

Para la estación La Florida de la ciudad de Talca, en el período entre abril y agosto de 2014, los resultados reportados del modelo señalan la capacidad de predecir episodios críticos en un 66% con un día de anticipación, 64% con dos días de anticipación y 59% con tres días de anticipación. En

tanto, las falsas alarmas corresponden a 39% para un día de anticipación, 35% para dos días de anticipación y 37% para tres días de anticipación.

1.5.2. Modelación de calidad del aire con algoritmos de aprendizaje automático

Los algoritmos de aprendizaje automático han sido ampliamente utilizados como herramientas en los estudios de predicción atmosférica y de calidad del aire. En especial, los algoritmos de redes neuronales artificiales (ANN) se han utilizado en este tipo de problemas, por su capacidad de resolver problemas de naturaleza no lineal (Feng, et al. 2015).

En (Gardner and Dorling 1998) se realizó una revisión bibliográfica de las aplicaciones de ANN en ciencias atmosféricas, especialmente relacionadas al perceptrón multicapa. Se señalan las ventajas de ANN en la manipulación de sistemas no lineales, sobre todo cuando los modelos teóricos son difíciles de construir.

En (Perez and Reyes 2002) se aplicó un modelo Perceptron Multicapa (MLP) y un modelo lineal para predecir el máximo en 24 horas de las concentraciones promedio de MP_{10} en Santiago, Chile. Argumentan que aunque MLP da un resultado ligeramente mejor que el modelo lineal, la selección de potenciales predictores es más importante que la selección de modelos (MLP o lineales).

En (Kolehmainen, Martikainen and Ruuskanen 2001) se comparan dos familias de redes neuronales, Self-organizing maps (SOM) y Perceptrón Multicapa (MLP). Además, se evalúa el efecto de eliminar componentes periódicos con respecto a las redes neuronales. Los métodos se evaluaron utilizando series de tiempo por hora de NO2 y variables meteorológicas básicas recogidas en la ciudad de Estocolmo en 1994 hasta 1998. Los valores estimados para la predicción se calcularon en tres formas: el uso de los componentes periódicos, la aplicación de métodos de redes neuronales a los valores residuales después de la eliminación de los componentes periódicos, y la aplicación de redes neuronales a los datos originales. Los resultados mostraron que las mejores estimaciones se lograron aplicando directamente MLP a los datos originales, y por lo tanto, que una combinación del método de regresión periódica y algoritmos neuronales no da ninguna ventaja sobre una aplicación directa de los algoritmos neuronales.

En (Kukkonen, et al. 2003) se evaluaron cinco modelos ANN comparándolos con un modelo lineal y uno determinista para predecir las concentraciones de MP_{10} y NO2 en Helsinki, Finlandia. Los modelos ANN dieron mejores resultados que los otros modelos, especialmente para los modelos de ANN que se construyeron con varianza no constante.

En (Jiang, et al. 2004) se utiliza un modelo MLP para predecir el promedio diario del índice de calidad del aire, modificando el método de entrenamiento y la estructura del modelo ANN mejoró significativamente la precisión de la predicción. Los autores también sugieren que una estructura más

simple del modelo MLP con paradas entrenamiento tempranas entrega mejores resultados en los datos de prueba.

En (Hooyberghs, et al. 2005) se mejoró la precisión de la predicción de los promedios diarios de material particulado con la incorporación de la altura de la capa límite como una de las variables de entrada en el modelo ANN. Llegaron a la conclusión de que las condiciones meteorológicas juegan un papel importante en las fluctuaciones diarias de concentración de MP_{10} en Bélgica.

En (Lu, Hsieh and Chang 2006) se emplearon dos etapas con modelos ANN para pronosticar las concentraciones de ozono. Las condiciones meteorológicas se agruparon en primer lugar en diferentes regímenes meteorológicos utilizando un mapa auto-organizativo (SOM). A continuación, se utilizó un modelo MLP supervisado para aproximar la relación no lineal ozono-meteorológico en cada régimen meteorológico. Los resultados del modelo híbrido SOM-MLP puede explicar, al menos el 60% de la variación en las concentraciones de ozono.

En (Brunelli, et al. 2007) se investigó la aplicabilidad de redes neuronales recurrentes (modelo de Elman) para predecir las concentraciones máximas diarias de distintos contaminantes. Encontraron mejor consistencia en las concentraciones pronosticadas por las redes de Elman, en comparación con MLP.

En (Díaz-Robles, Ortega, et al. 2008) se propone un modelo híbrido ARIMA-ANN para mejorar la precisión de los pronósticos de MP_{10} en Temuco, Chile. Los resultados experimentales muestran que el modelo híbrido mejora la precisión de predicción comparando con los modelos utilizados por separado. El modelo híbrido fue capaz de capturar 100% y 80% de episodios de alertas y pre-emergencia, respectivamente.

En (Pérez and Salini 2008) se comparan tres modelos para la estimación del máximo promedio móvil de concentración de $MP_{2,5}$. Los modelos utilizados en el estudio son: el lineal, el perceptrón multicapa y un algoritmo de clustering denominado “hybrid clustering algorithm” (HCA). Las variables de entrada son concentraciones pasadas de cuatro estaciones de monitoreo e información meteorológica real y pronosticada en Santiago de Chile entre los años 2004 y 2007. Los resultados señalan que si bien los tres métodos pueden ser utilizados operativamente, el algoritmo HCA fue el que entrego los mejores resultados.

En (Hrust, et al. 2009) se seleccionaron los intervalos promedios para variables de entrada. En el informe se seleccionan períodos promedios óptimos para cada predictor potencial mediante la comparación de valores de coeficiente de correlación entre las concentraciones modeladas y medidas. Los autores de (Hrust, et al. 2009) también realizaron un análisis de sensibilidad para cada variable de entrada.

En (Kurt and Oktay 2010) se propone un modelo geográfico para predecir las concentraciones medias diarias de SO2, CO y MP_{10} con tres días de anticipación utilizando MLP. Se emplearon tres tipos de modelos geográficos: single-site neighborhood model, two-site neighborhood model y el modelo basado en la distancia. Los resultados experimentales mostraron que los modelos basados en variables geográficas superaron el modelo simple, especialmente el modelo basado en distancia. Se espera un aumento en la mejora si se añaden las variables meteorológicas al modelo geográfico.

En (Siwek and Osowski 2012) se aplicó la transformación Wavelet en conjunto con ANN para predecir las concentraciones medias diarias de MP_{10}. En el estudio combinaron distintos tipos de ANN en un conjunto para generar una predicción final en una red neuronal adicional. Los resultados mostraron la utilidad de transformación Wavelet en los pronósticos de contaminación del aire al descomponer el problema, simplificando las tareas y aumentando la predicción final.

En (Feng, et al. 2015) se presenta un modelo híbrido que combina análisis de trayectoria de masa de aire y la transformación Wavelet para mejorar la exactitud de predicción de la red neuronal (ANN) en el pronóstico de concentraciones promedios diarias de $MP_{2,5}$ con dos días de anticipación. Se utilizaron variables meteorológicas y contaminantes diarias pronosticadas con un modelo Perceptrón multicapa (MLP). Los días con altos índices de $MP_{2,5}$ son anticipados mediante el uso descomposición wavelet y la tasa de detección (DR) para un umbral de alerta determinado de modelo híbrido puede alcanzar el 90% en promedio.

1.6. Algoritmos de Aprendizaje Automático

En el presente trabajo se han seleccionado tres algoritmos de aprendizaje automático para estimar la concentración del promedio 24 horas de $MP_{2,5}$ en el aire de la ciudad de Talca. Nuestra selección reúne dos algoritmos ANN, el Perceptrón multicapa, algoritmo referenciado en la literatura como el mayormente utilizado en problemas de estimación de calidad del aire, y un modelo basado en SOM, denominado TTOSOM, algoritmo capaz de procesar vectores con información faltante de manera directa. Además, hemos incorporado el algoritmo Random Forest, uno de los algoritmos más certeros existentes en la actualidad.

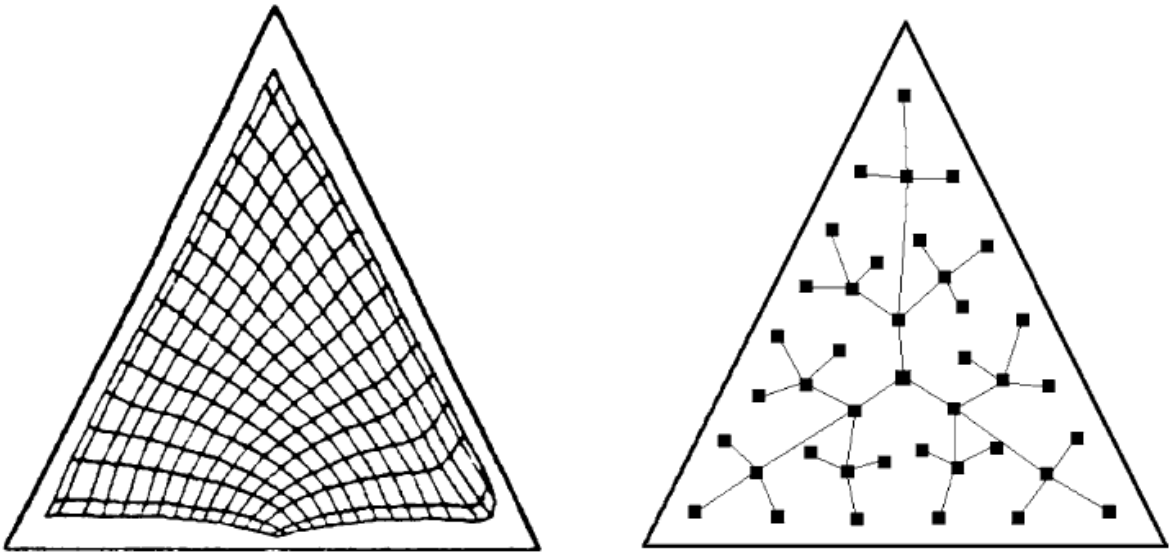
1.6.1. Modelación con Algoritmo TTOSOM

Los autores de (Astudillo and Oommen 2011) proponen un algoritmo basado en SOM, al cual denominan Tree-based Topology Oriented SOM (TTOSOM).

Para explicar el funcionamiento del algoritmo TTOSOM, antes es necesario presentar el algoritmo SOM, el cual es un tipo de red neuronal que utiliza aprendizaje no supervisado para el entrenamiento de neuronas, cuya meta es representar todas las instancias del conjunto de entrada por un grupo de prototipos llamados neuronas, de tal manera que puntos cercanos en el conjunto de entrada sean representados por neuronas que también se encuentran cercanas y análogamente puntos lejanos

sean representados por neuronas lejanas. Es decir, las neuronas intentan preservar las propiedades topológicas del espacio de entrada (Kohonen 1995).

TTOSOM mantiene las capacidades de SOM, pero que a diferencia de SOM, utiliza una estructura de árbol en vez de una estructura de malla. La elección de TTOSOM permite un nivel de resolución mayor con respecto a SOM, ya que la estructura de árbol representa de mejor forma la estructura original de datos que la estructura de malla, y al igual que SOM, preservando la topología (Astudillo and Oommen 2011). Un ejemplo se puede apreciar en la Figura 2.



(a) La grilla aprendida por SOM (b) Árbol aprendido por TTOSOM

Figura 2: Cómo aprenden una distribución en forma de triángulo (Astudillo and Oommen 2011)

TTOSOM asume que el usuario tiene la capacidad de definir un árbol con anterioridad.

1.6.2.Perceptrón Multicapa (MLP)

El perceptrón multicapa consiste en un sistema de neuronas, o nodos, interconectadas, en un grafo dirigido, que permite modelar relaciones no lineales entre un vector de entrada y un vector de salida. Los nodos están conectados por pesos y señales de salida, que están en función de la suma de las entradas al nodo modificada por una función no lineal simple de transferencia o activación. Es la superposición de muchas funciones de transferencia no lineal simples que permite al perceptrón multicapa aproximarse a funciones extremadamente no lineales. Si la función de transferencia es lineal, entonces el perceptrón multicapa sólo es capaz de modelar funciones lineales.

La salida de un nodo es escalada por el peso de conexión y alimentada hacia delante, para ser una entrada a los nodos en la siguiente capa de la red. Esto implica una dirección de procesamiento de la información, por lo tanto, el perceptrón multicapa se conoce como una red neural de alimentación hacia adelante.

La arquitectura de un perceptrón multicapa es variable, pero en general consiste en varias capas de neuronas. La capa de entrada no desempeña ningún papel computacional, sino que simplemente sirve para ingresar el vector de entrada a la red, luego puede contener una o más capas ocultas y finalmente una capa de salida (Gardner and Dorling 1998).

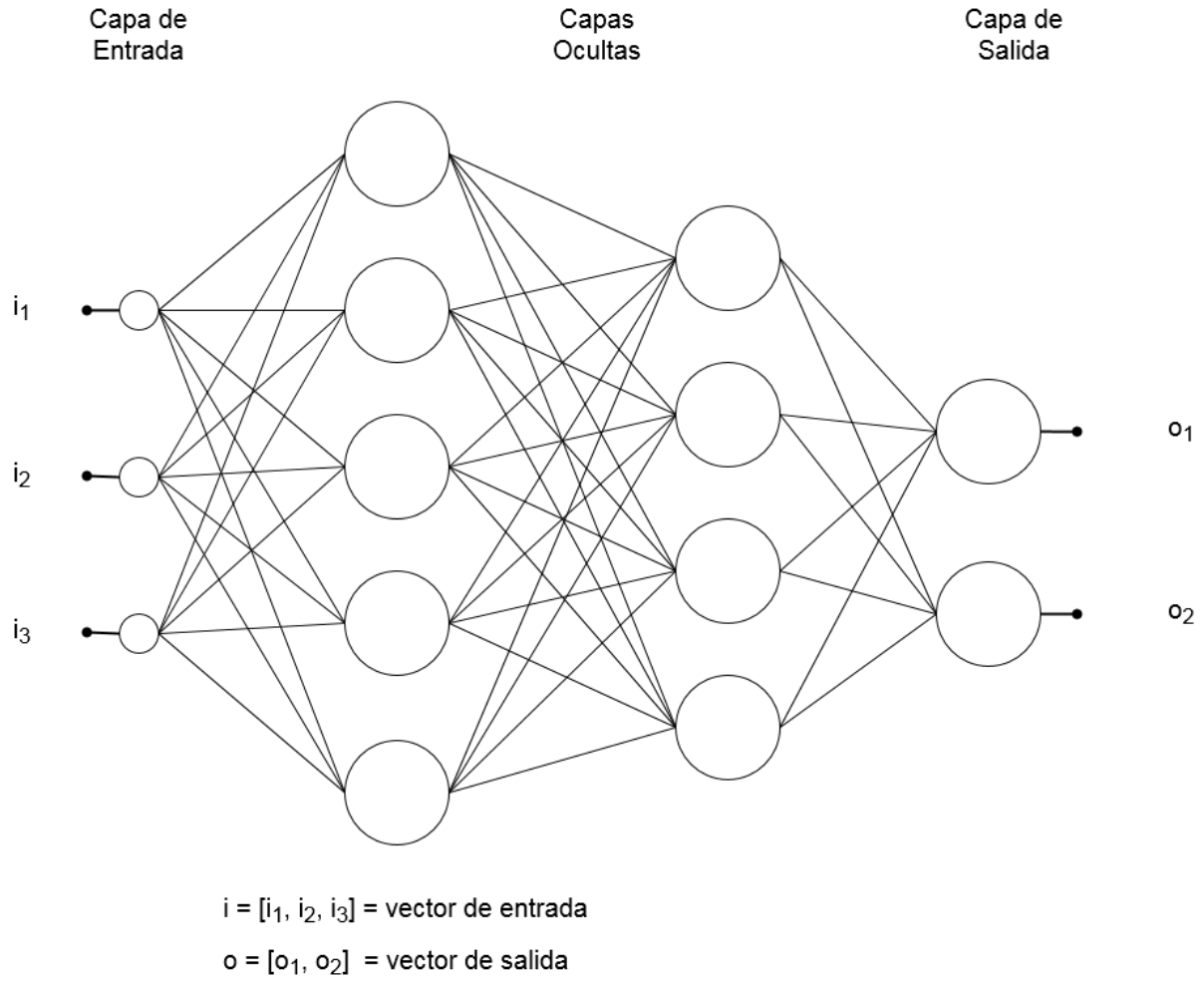


Figura 3: Ejemplo de un perceptrón multicapa con dos capas ocultas (Gardner and Dorling 1998)

1.6.3.Random Forest (RF)

Random Forest es un algoritmo propuesto por (Breiman 2001), el cual combina el método random subspace y bagging. El algoritmo es uno de los más certeros en la actualidad y ha sido utilizado para resolver una gran cantidad de tareas (Verikas, Gelzinis and Bacauskiene 2011).

Inicialmente tenemos un conjunto de datos de entrenamiento X_t = $\{f(x_m, y_m)$, m =(1,…, m)}, donde x_m es una variable de entrada y y_m una variable de salida. Un aprendiz débil puede ser creado utilizando el conjunto de entrenamiento X_t, donde el aprendizaje débil es un predictor $f(x, X_t)$ con bajo sesgo y alta varianza. En RF un árbol de decisión es utilizado como aprendiz débil.

Mediante un muestreo aleatorio del conjunto X_t, un conjunto de árboles de decisión $f(x, X_t, \theta_k)$ puede ser creado, con $f(x, X_t, \theta_k)$ el k-ésimo árbol de decisión y θ_k es el vector aleatorio que selecciona los puntos de datos para el k-ésimo árbol de decisión. Al aplicar el muestreo bootstrap para generar θ_k, por ejemplo, se utiliza dos tercios de los datos para cada árbol de decisión, y cerca de un tercio quedan fuera del muestreo bootstrap.

Una de las características de θ_k es ser independiente e idénticamente distribuidos (iid), por lo que al combinar los árboles aleatorios a través de un promedio, el sesgo se mantiene prácticamente inalterado, en tanto la varianza se reduce por un factor de $\bar{\rho}$ el valor promedio de correlación entre los árboles de aprendizaje.

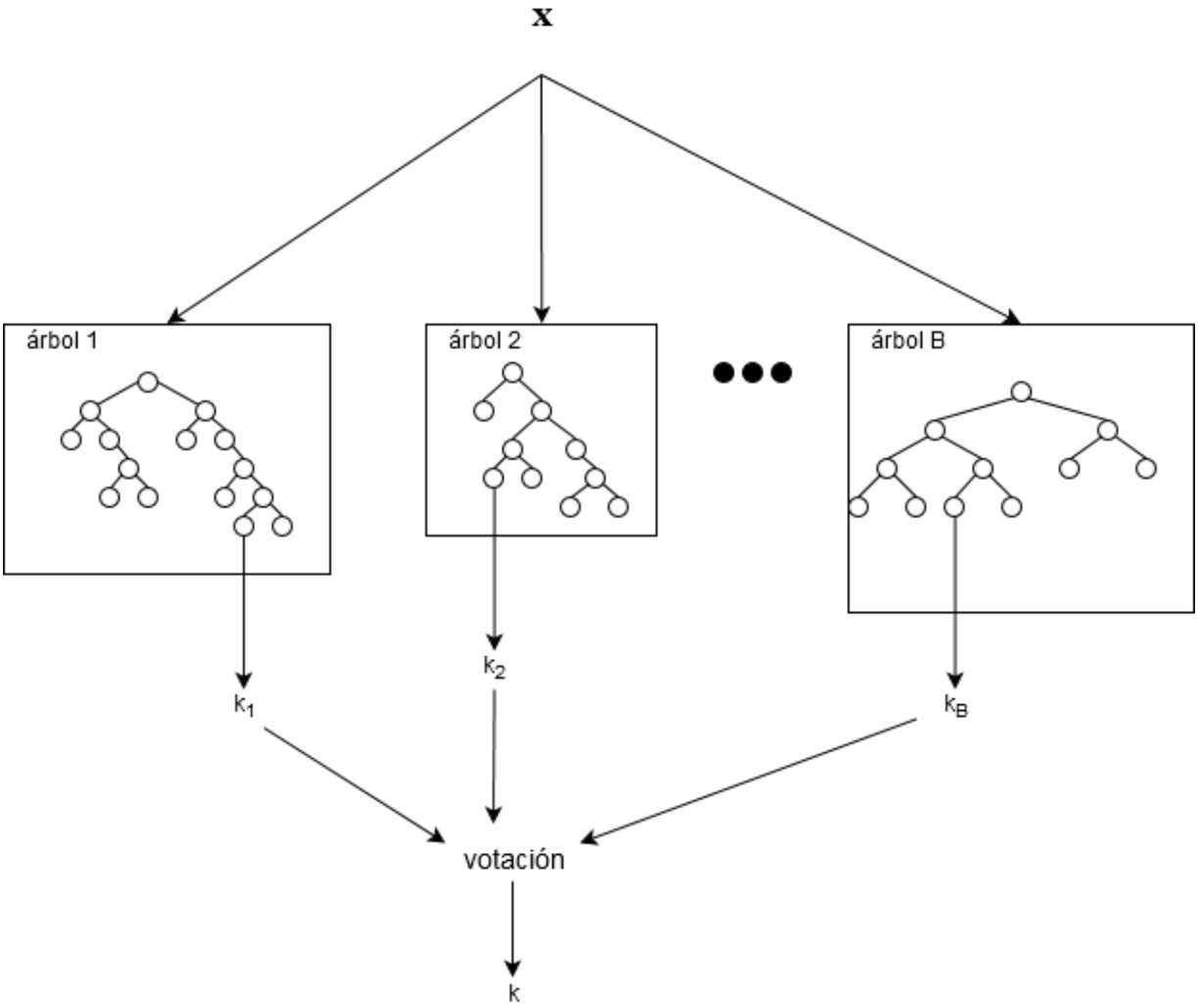


Figura 4: Arquitectura general de Random Forest (Verikas, Gelzinis and Bacauskiene 2011)

La Figura **4** presenta una estructura general de RF, donde B es el número de árboles en RF y k_1, k_2,…, k_b son las etiquetas de las clases. A medida que aumenta el número de árboles, las tasas de error convergen a un límite, por lo que no hay sobreajuste por grandes números de árboles.

1.7. Metodología

Se ha preparado la base de datos con parámetros contaminantes y meteorológicos de tres estaciones de la ciudad de Talca, para predecir la calidad del aire utilizando algoritmos de aprendizaje automático. Los resultados de los tres algoritmos serán evaluados y comparados según los valores de las medidas de desempeño. En la Figura 5 se presenta la metodología general propuesta:

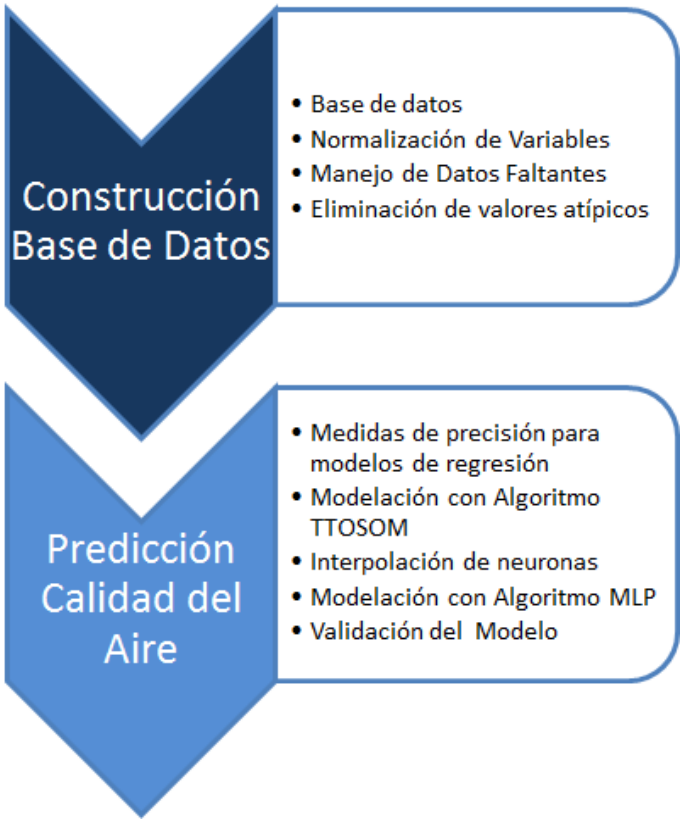


Figura 5: Metodología Experimental

La construcción de la base de datos ha sido realizada a partir de los registros oficiales de contaminación del aire en las estaciones de la ciudad de Talca. La base construida presenta los promedios diarios móviles de concentración de $MP_{2,5}$ de las estaciones Cesfam La Florida, Universidad Católica del Maule y Universidad de Talca. Todos los valores han sido normalizados llevando los valores al rango [0,1], manipulando los valores faltantes y eliminando valores atípicos.

Luego, para predecir la calidad del aire, se han seleccionado tres algoritmos de aprendizaje automático, el Perceptrón Multicapa, una variante del algoritmo Self-organizing maps (SOM) llamado Tree-based Topology Oriented SOM (TTOSOM), y por último Random Forest.

Los tres algoritmos han sido entrenados con los registros entre 1 de Abril de 2014 y el 30 de Junio de 2015.

Para medir y comparar las capacidades de predicción de los tres algoritmos, se han seleccionado una lista de medidas de precisión para modelos de regresión.

La validación propuesta, determina la capacidad de predicción de los métodos en datos futuros, es decir, los modelos son entrenados con valores hasta el 30 de Junio de 2015 y la validación se desarrolla con los valores restantes desde el 1 de Julio de 2015.

1.7.1.Base de Datos Original

En Chile, el Ministerio de Medio Ambiente es el organismo público responsable de la gestión del ambiente y de los recursos renovables naturales. Para la administración de la información de calidad del aire, el ministerio ha desarrollado el Sistema de Información Nacional de Calidad del Aire (SINCA), el cual administra una red de estaciones de monitoreo de calidad del aire y meteorológicas. A través de su portal web, se presentan las mediciones de calidad del aire en línea, el seguimiento histórico de las mediciones de calidad del aire y meteorología, antecedentes de las estaciones de monitoreo, documentación relacionada con calidad del aire y monitoreo, y enlaces a sitios web nacionales e internacionales.

La base de datos confeccionada corresponde a los datos de las estaciones La Florida, U.C. Maule y Universidad de Talca, con información entre los meses de Abril y Agosto de los años 2014 y 2015.

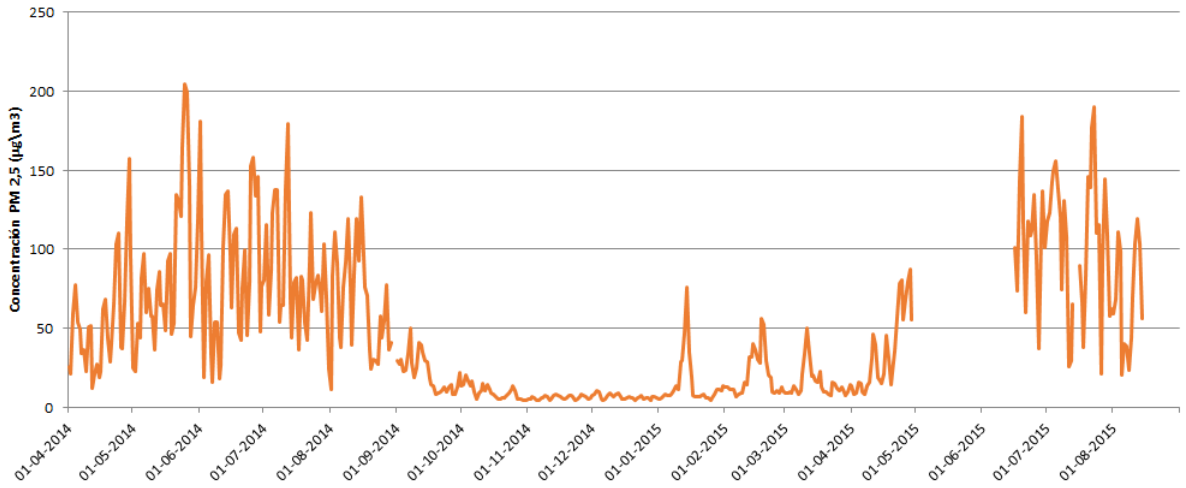
En las tres estaciones, los registros han sido capturados por las técnicas de medición señaladas en la Tabla 2.

Tabla 2: Sensores de medición de parámetros contaminantes y meteorológicos de las estaciones La Florida, U.C. Maule y Universidad de Talca (Ministerio de Medio Ambiente 2009-2015).

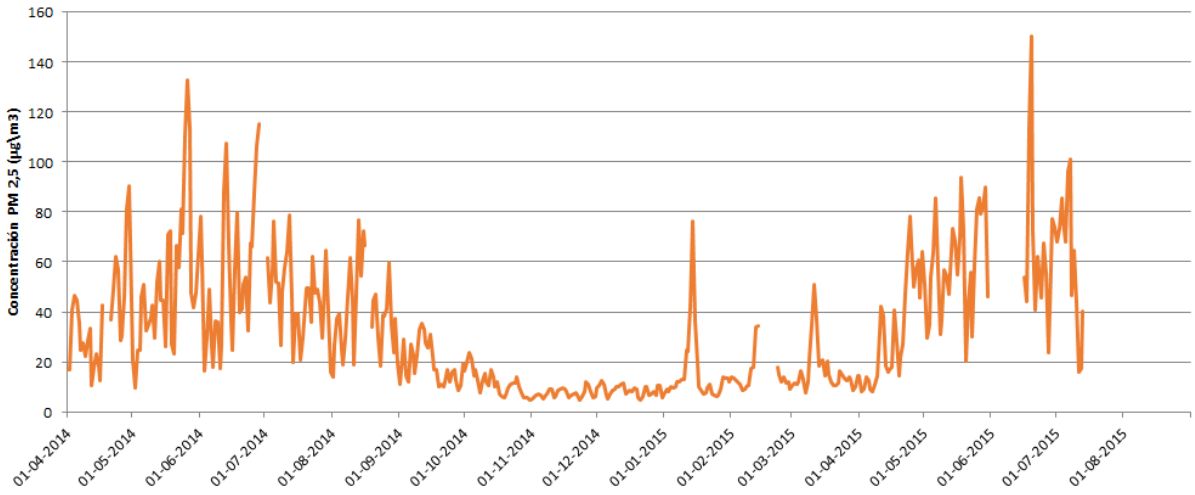
Parámetro	Unidades	Frecuencia	Altura de Medición	Técnica de Medición
$MP_{2,5}$	$\mu g/m^3$	Continua	-	Atenuación de Radiación Beta MET ONE BAM1020
Temperatura Ambiental	°C	Horaria	10 m	Sensor VAISALA HMP155
Velocidad del Viento	m/s	Horaria	10 m	Sensor MET ONE 010C

La Figura 6 presenta los registros de contaminación de $MP_{2,5}$ registrados en las tres estaciones de la ciudad entre los meses de Abril y Agosto de los años 2014 y 2015. En los meses de invierno se

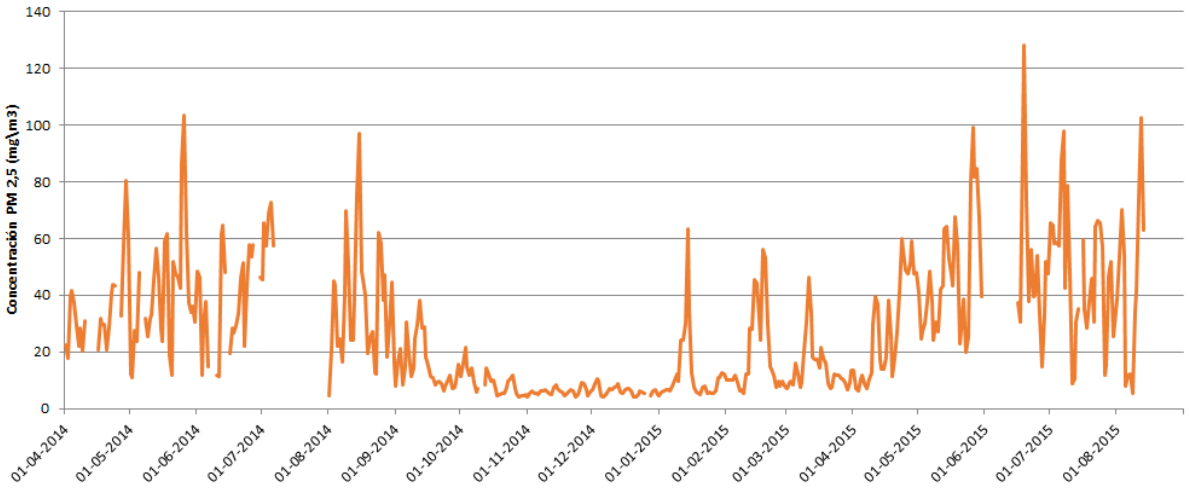
observan los mayores registros de contaminación en todas las estaciones, en cambio los meses de verano presentan registros más bajos, debido a la ausencia de emisiones de calefactores a leña. No obstante, en el período de verano se observan en las tres estaciones peaks de concentración, siendo estos registros consecuencia de incendios forestales cercanos a la ciudad. Los períodos no graficados corresponden a sectores con datos faltantes en la base de datos.



(a) Estación Cesfam La Florida



(b) Estación Universidad Católica del Maule



(c) Estación Universidad de Talca

Figura 6: Promedio diarios de $MP_{2,5}$ entre los meses de Abril de 2014 y Agosto de 2015

Las variables meteorológicas utilizadas en la modelación son: la temperatura ambiental, variable asociada al comportamiento de las personas de la ciudad, donde una mayor cantidad de emisiones se relaciona a bajas temperaturas y viceversa. Otra variable es la velocidad del viento, la cual representa la capacidad de dispersión de la contaminación atmosférica en el lugar, es decir a mayor velocidad del viento menor es la concentración de $MP_{2,5}$. Por último, la contaminación promedio de 24 horas por $MP_{2,5}$ del día anterior, representa la contaminación base en el sistema.

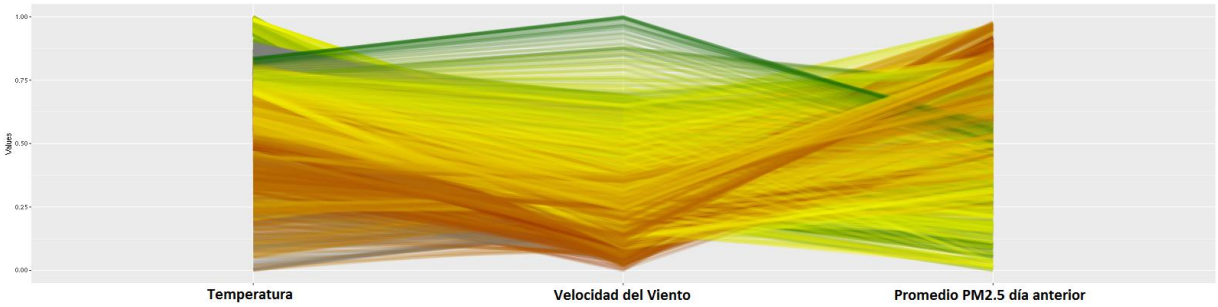
La Figura 7 muestra un gráfico de coordenadas paralelas, el cual presenta en forma desagregada las relaciones entre las variables explicativas y la estimada para cada estación de monitoreo. El gráfico representa en color verde los valores bajos de contaminación del aire por $MP_{2,5}$ y los valores altos en color rojo. En tanto, a mayor altura en cada línea de cada eje, representa valores mayores de las variables temperatura ambiental, velocidad del viento y contaminación promedio de 24 horas por $MP_{2,5}$ del día anterior, y viceversa.

Tabla 3: Correlación entre la concentración 24 horas de $MP_{2,5}$ con las variables temperatura promedio 24 horas, velocidad del viento 24 horas y concentración 24 horas de $MP_{2,5}$ del día anterior en las estaciones Cesfam La Florida, U.C. Maule y Universidad de Talca.

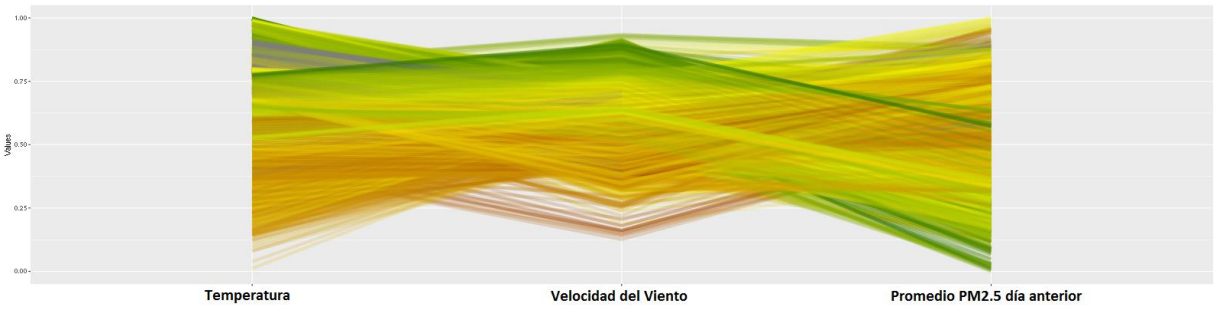
Estación de Monitoreo	Variables de Modelación		
	Temperatura	Velocidad del Viento	MP2,5 del día anterior
Cesfam La Florida	-0,62	-0,52	0,50
UCM	-0,47	-0,58	0,48
UTALCA	-0,42	-0,60	0,43

Complementando la Figura 7, la Tabla 3 refleja una relación inversa entre la concentración 24 horas de $MP_{2,5}$ con las variables temperatura promedio 24 horas, velocidad del viento 24 horas y una relación directa con la variable concentración 24 horas de $MP_{2,5}$ del día anterior. En la estación Cesfam La Florida, la temperatura promedio 24 horas es la variable más correlacionada a los índices de contaminación, en tanto, en las estaciones UCM y UTAL, la velocidad del viento promedio 24 horas es la variable que mayor explica los índices de contaminación.

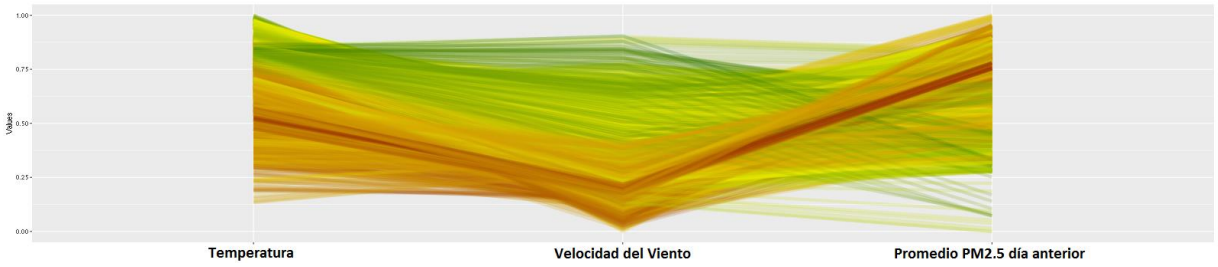
Para cada estación, se ha construido una base de datos con 7.344 filas, donde cada fila presenta los promedios diarios móviles de las variables de contaminación, entre el 01 de Abril de 2014 y el 31 de Agosto de 2015.



(a) Estación Cesfam La Florida



(b) Estación Universidad Católica del Maule

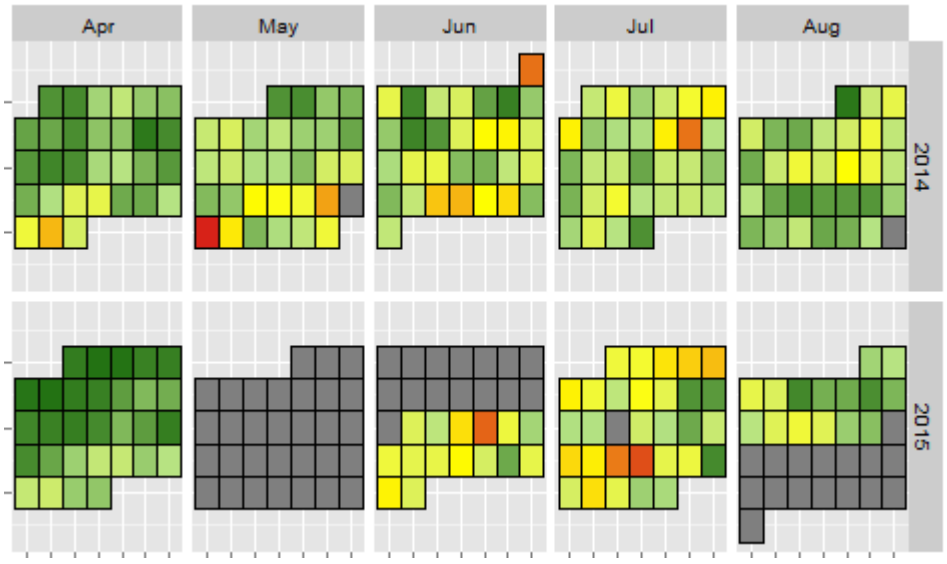


(c) Estación Universidad de Talca

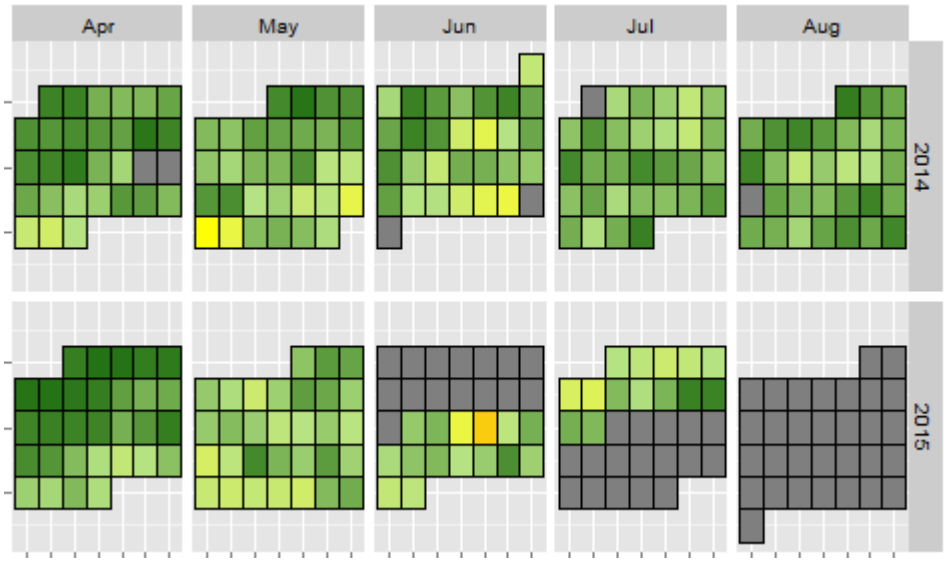
Figura 7: Comportamiento de las variables meteorológicas y de contaminación con respecto a la variable a estimar, entre los meses de Abril de 2014 y Agosto de 2015.

El promedio móvil se calcula a partir del registro de contaminación de un día y hora con los anteriores 23 registros de contaminación. Según lo señalado en DS N° 61 de 2008 del Ministerio de Salud, los promedios diarios son obtenidos cuando existe al menos 18 horas de registros horarios en un día.

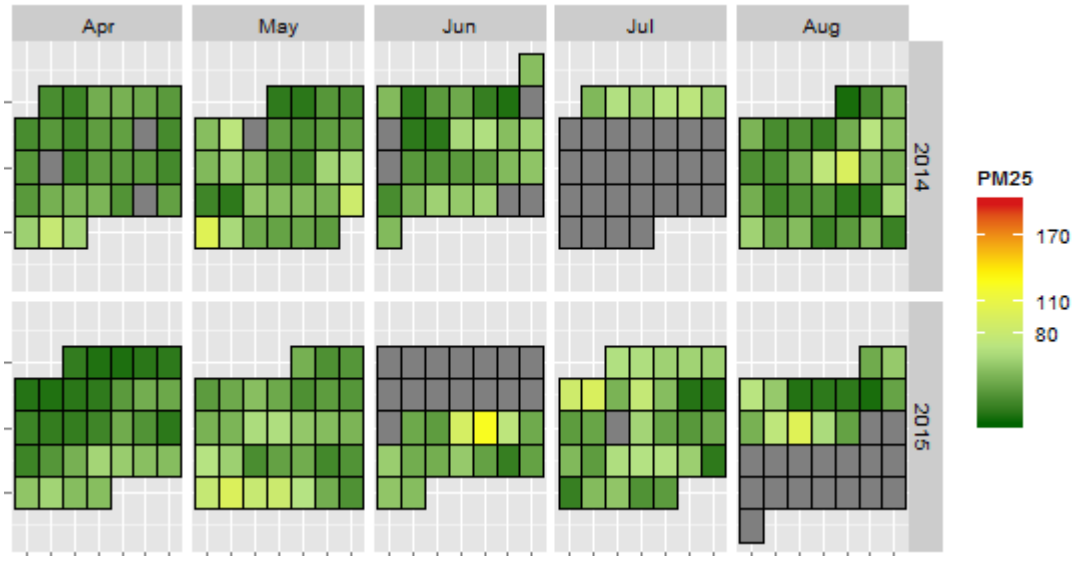
La Figura **8** presenta el promedio diario de $MP_{2,5}$ para las tres estaciones entre los meses de Abril y Agosto de los años 2014 y 2015. Los colores representan la contaminación de cada día, siendo el verde las concentraciones más bajas, el verde claro las concentraciones asociadas a episodios de alerta ambiental, los amarillos asociados a preemergencias ambientales y naranjos y rojos en las emergencias. Los casilleros en gris representan los días sin información.



(a) Estación Cesfam La Florida



(b) Estación Universidad Católica del Maule



(c) Estación Universidad de Talca

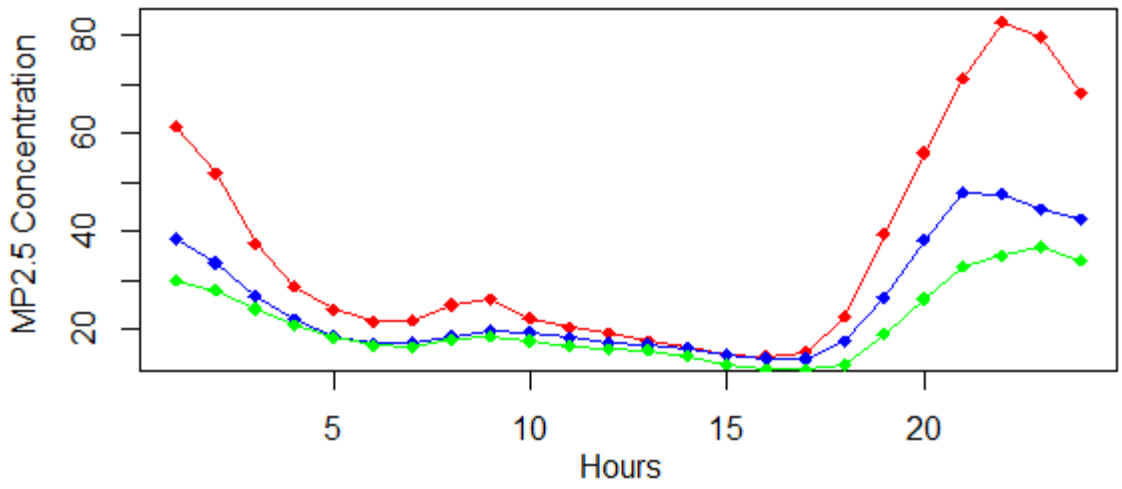
Figura 8: Promedio Diario de $MP_{2,5}$ entre los meses de Abril y Agosto de los años 2014 y 2015

En el período entre Abril y Agosto de los años 2014 y 2015, en la estación Cesfam La Florida se contabilizaron 145 días sin problemas, 41 “Alertas”, 49 “Preemergencias”, 7 “Emergencias” y 64 registros nulos. En la estación Universidad Católica del Maule, se contabilizan 215 días normales, 16 “Alertas”, 6 “Preemergencias”, 0 “Emergencias” y 69 registros nulos. Por último, en la estación Universidad de Talca, se contabilizan 227 días sin problemas, 11 “Alertas”, 1 “Preemergencia”, 0 “Emergencias” y 67 registros nulos.

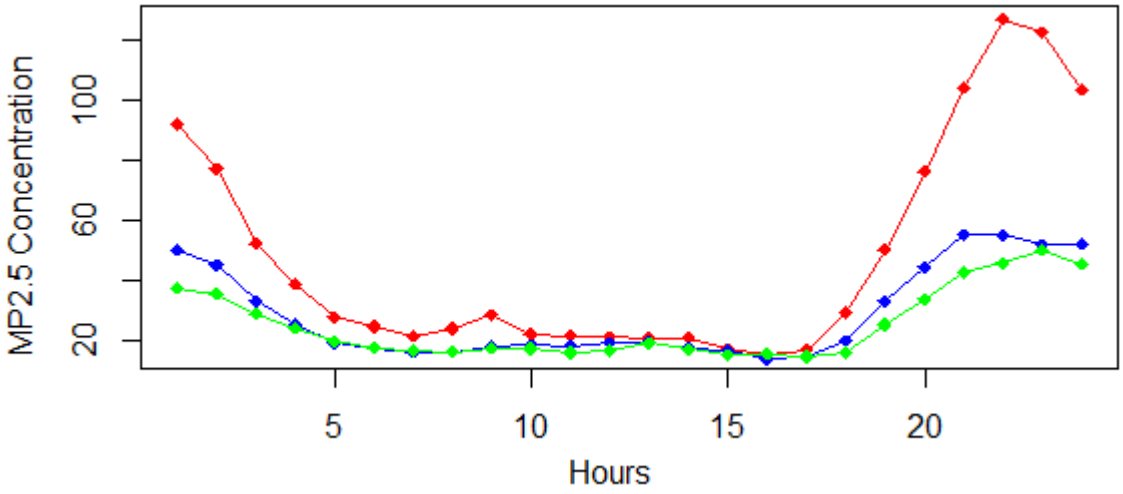
Como se puede apreciar en las cifras del párrafo anterior y en la Figura **8**, en general, los días con registros bajo 80 $\mu g/m^3$, son los valores mayormente registrados en las tres estaciones, sin embargo la estación Cesfam La Florida presenta los peores registros de contaminación en la ciudad, siendo la única estación que presenta períodos de “Emergencias”.

En la Figura 9 se presentan los promedios y la desviación estándar de los registros horarios de $MP_{2,5}$ para las tres estaciones, entre los meses de Abril y Agosto de los años 2014 y 2015. El rojo representa la estación Cesfam La Florida, en azul la estación Universidad Católica del Maule (UCM) y en verde la estación Universidad de Talca (UTalca).

En la Figura 9(a), en las tres estaciones se aprecia que a partir de las 17 horas empieza el aumento de la concentración promedio de $MP_{2,5}$, con un peak a las 22 horas. Este aumento está directamente relacionado a las emisiones de calefactores a leña de las residencias de la ciudad, disminuyendo las emisiones en el período nocturno.



(a) Promedio



(b) Desviación Estándar

Figura 9: Registros horarios de MP2,5 entre los meses de Abril y Agosto de los años 2014 y 2015

En la Figura 9(a) también se observa un peak a las 9 horas, el cual se encuentra asociado a la salida de vehículos en el horario punta mañana.

En la Figura 9(b) se presenta la desviación estándar de las concentraciones de $MP_{2,5}$, donde al igual que la Figura 9(a), se presentan peaks a las 22 hrs y las 9 hrs. Ambos gráficos nos señalan que existen períodos del día que presentan un comportamiento homogéneo, en cambio, en los peaks se registran períodos con bajos y altos registros durante la temporada de invierno.

1.7.2.Normalización de Datos

Según lo señalado por (U.S., Environmental Protection Agency 2003), las variables de concentraciones de partículas y registros meteorológicos presentan una distribución log-normal con un marcado comportamiento estacional (U.S., Environmental Protection Agency 2003).

Por lo que, se ha utilizado la función logaritmo natural en nuestras variables para obtener una estructura normal de distribución de datos.

Luego, los parámetros de las variables de la base de datos original son llevados al rango [0,1], restando a cada valor el valor mínimo del parámetro y dividiendo por la resta entre el valor máximo y mínimo de cada parámetro.

$$Norm(MP_{2,5})_i = \frac{Registro_i - Registro_{min}}{Registro_{max} - Registro_{min}}$$

1.7.3.Manipulación de Datos Faltantes

El problema de datos faltantes es un problema común en las bases de datos de investigación de contaminación del aire (Junninen, et al. 2004). Las razones principales del problema de datos faltantes es por muestreos insuficientes, errores de medición o por fallas en la adquisición de datos.

En la Tabla **4**, se presentan cuatro enfoques para la resolución de missing data, presentados en (Bougoudis, Demertzis and Iliadis 2015), mostrando sus ventajas y desventajas.

En los registros de concentración de $MP_{2,5}$, del período de tiempo utilizado para la modelación del problema, la estación Cesfam La Florida presenta un 21,56% de datos faltantes, la estación UCM un 22,85% y la estación UTAL un 21,90 %. Los registros utilizados corresponden a los señalados como registros oficiales, descartando los registros preliminares y no validados.

Según las características de los algoritmos, se han construido dos bases de datos para cada estación de monitoreo, donde una base se ha construido eliminando los registro que presentan datos faltantes y, la segunda base de datos ha sido construida pensando en métodos con la capacidad de procesar datos faltantes de manera directa, es decir, métodos que no alteren el vector original, sino que, ventajosamente, algoritmos capaces de procesar vectores con información faltante de manera directa.

Tabla 4: Enfoques, ventajas y desventajas en missing data (Bougoudis, Demertzis and Iliadis 2015)

Enfoque	Ventajas	Desventajas
Reemplazar missing data por media muestral o moda	Puede utilizar los métodos de análisis de casos completos	Reduce la variabilidad Debilita estimaciones de covarianza y correlación en los datos (porque ignora relación

		entre las variables)
Ajuste variable dummy	Utiliza toda la información disponible acerca de la observación faltante	Resultados en estimaciones sesgadas No impulsada teóricamente
La sustitución de missing data con valores pronosticados por una ecuación de regresión	Utiliza la información de los datos observados	Sobreestima el ajuste del modelo y las estimaciones de correlación Debilita la varianza
Identificación de un conjunto de parámetros que produce la más alta probabilidad	Utiliza información completa (ambos casos completos y casos incompletas) para calcular la probabilidad Estima valores perdidos con imparcialmente con datos aleatorios	Complejidad del modelo
Descartar todos los missing data	Simplicidad Comparabilidad entre los análisis	Reduce la potencia estadística No utiliza toda la información

1.7.4.Eliminación de valores atípicos

En (James, et al. 2013) se presentan problemas comunes de los modelos de regresión, donde existen dos problemas que se refieren específicamente a valores presentes en las bases de datos que disminuyen los resultados de las medidas de precisión. Estos casos se identifican como valores atípicos (outliers) y puntos de alto apalancamiento (high leverage).

Un outlier es un punto donde el valor medido (y_i) de la variable a predecir está lejos de ser el valor predicho por el modelo. En tanto, high leverage presenta valores inusuales para las variables dependientes (x_i).

En general, la eliminación de outliers tiene poco impacto en la pendiente de la función, pero un alto impacto en el valor de las medidas de precisión. En cambio, los puntos con alto apalancamiento tienden a tener un impacto considerable en la regresión estimada.

Para medir outliers, se utiliza una prueba llamada residuos estudentizados (studentized residual). Estos residuos se determinan dividiendo cada residuo por su error estándar estimado. En nuestro caso, los valores considerados como datos atípicos son los valores mayores y menores al absoluto de 3.

En tanto, para medir los puntos con alto apalancamiento se realiza a través de una prueba llamada Hat Values, a través de la siguiente formula:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x'_i - \bar{x})^2}$$

Las observaciones que superen con creces (p+1)/n, entonces el punto es sospechoso de alto apalancamiento.

Una medición conjunta de outliers y high leverage es realizada a través de la Distancia de Cook (CD_i), técnica que mide la influencia conjunta (combinada) en el caso de ser un valor atípico en Y y en el espacio de los predictores. En (Stevens 1984) se formula la Distancia de Cook (CD_i) en función de studentized residual (r_i^2) y hat values ($h_i i$).

$$CD_i = \frac{1}{(p+1)} \cdot r_i^2 \cdot \frac{h_i i}{(1 - h_i i)}$$

Para determinar los datos atípicos de las variables explicativas, se utilizó la Distancia de Cook, donde los datos mayores a 4/(n-k-1) son considerados atípicos (Fox 1991).

1.7.5.Medidas de precisión para modelos de regresión

El error de un modelo, se mide como la diferencia entre el valor predicho y el valor medido, en especial en las etapas de entrenamiento y validación.

Para medir el grado de ajuste de los datos en nuestro modelo, utilizaremos las pruebas señaladas en la Tabla **5**.

Tabla 5: Medidas de Precisión.

Métrica	Fórmula	Referencia
R^2	$\frac{\sum(\hat{y_i} - \bar{y_i})^2}{\sum(y_i - \bar{y_i})^2}$	(James, et al. 2013)
MAPE	$\frac{\sum\frac{\lvert y_i - \bar{y_i}\rvert}{\lvert y_i\rvert}}{n}$	(Grillenzoni 1998)
RSME	$\sqrt{\frac{\sum\lvert y_i - \bar{y_i}\rvert^2}{n}}$	(Florinsky 2016)
RSE	$\frac{\sum\lvert y_i - \bar{y_i}\rvert}{n}$	(James, et al. 2013)

El estadístico R^2 representa la proporción de varianza explicada, presentando valores entre 0 y 1, independiente de la escala de la variable explicada. Un valor del estadístico cercano a 1, indica que

una gran proporción la variable dependiente es explicada por la regresión. En tanto, un valor cercano a 0 indica que la regresión no explica gran parte de la variable dependiente, esto debido a un mal modelo lineal o que el error σ^2 es alto, o ambas causas.

En tanto, el error porcentual absoluto medio (MAPE), expresa el error medio de una estimación, presentando valores porcentuales. Valores cercanos a 0% indican un bajo error medio de la regresión.

RSE y RSME son estimaciones de la desviación estándar del error en las mismas unidades del modelo. Valores pequeños de ambas pruebas se consideran como un buen ajuste del modelo a los datos, en tanto, valores grandes indican que el modelo no se ajusta bien a los datos. La RSE y la RSME ofrecen una medida absoluta de la falta de ajuste del modelo a los datos, pero como se mide en las unidades de la variable estimada, no siempre está claro el valor que constituye un buen resultado.

1.7.6.Modelación con Algoritmo TTOSOM

Para modelar la contaminación del aire, se utilizó la propuesta realizada en (Astudillo and Oommen 2011) llamado algoritmo Tree-based Topology Oriented SOM (TTOSOM), técnica descrita en la sección 1.6.1 y que representa los datos utilizando una topología de árbol, intentando descubrir la distribución subyacente de los datos pertenecientes a un conjunto de entrada, y al igual que SOM, preservando la topología.

Para maximizar la capacidad de predicción del algoritmo TTOSOM, se realizó una búsqueda en grilla de parámetros (Tabla 6) para cada base de datos buscando el conjunto de parámetros con mejores resultados según las medidas de precisión. En total, se realizan 168 pruebas para cada base de datos.

Tabla 6: Grilla de Parámetros del modelo TTOSOM

N° Iteraciones	:	500k
Radio Inicial	:	15 – 10
Radio Final	:	1 – 0
Tasa Inicial	:	0,5 – 0,9
Tasa Final	:	0,1 – 0,0
Árboles (Profundidad; N° de Neuronas)	:	4;43 – 42;42 – 4;39 3;37 – 3;40 – 5;125 3;11 – 3;16 – 4;69

Para determinar el valor de concentración de $MP_{2,5}$, los valores de las neuronas resultantes del proceso de entrenamiento con el algoritmo TTOSOM son interpoladas, comparando la distancia entre las neuronas con las variables dependientes. El método de interpolación utilizado se presenta en la Tabla 7.

Tabla 7: Algoritmo de Interpolación de neuronas de TTOSOM

Interpolación de neuronas TTOSOM
Entrada:
i) N, lista de neuronas con m elementos entrenadas con el algoritmo TTOSOM
ii) D, matriz de distancia entre neuronas entrenadas con el algoritmo TTOSOM para cada período j, con m elementos
Salida:
R, concentración de $MP_{2,5}$ con l elementos.
Método:
1: Se genera una lista con el orden de la neurona m según distancia al elemento j
2: Se seleccionan las k primeras neuronas
3: Para cada neurona seleccionada, se obtiene el inverso de la división de la distancia del elemento m por la suma total de las distancias de las k neuronas seleccionadas.
4: Se multiplica el valor del paso 3 por valor de concentración de $MP_{2,5}$ de la neurona m.
5: Se suman los valores del punto 4.
Fin método

La interpolación considera los valores de las n neuronas más cercanas, obteniéndose el valor de la regresión a través de un promedio inversamente proporcional a la distancia neural, es decir, a menor distancia a una neurona, mayor será la influencia del valor de esa neurona en el valor de la regresión, esto según lo señalado en las siguientes ecuaciones.

$$y_j = \frac{\sum_{i=1}^{n} x_i}{x_j}$$

$$\sum_{j=1}^{n} \frac{y_j \cdot \omega_j}{\sum_{k=1}^{n} y_k}$$

El valor de la variable x_i corresponde a la distancia de la neurona j, en tanto, ω_j corresponde al peso de la neurona j.

Se realiza una búsqueda de parámetros, donde se selecciona el número de neuronas más cercanas utilizadas en la interpolación, las que maximizan las medidas de precisión del modelo. La Tabla 8 presenta los parámetros utilizados que maximizan las medidas de precisión en cada base de datos.

Tabla 8: Parámetros que optimizan el resultado del algoritmo TTO-SOM

Parámetro	Cesfam La Florida	UCM	UTalca
N° Iteraciones	500k	500k	500k
Radio Inicial	10	10	10
Radio Final	1	1	0
Tasa Inicial	0,9	0,5	0,9
Tasa Final	0,1	0,1	0,0
N° de Niveles del Árbol	3	3	3
N° de Neuronas del Árbol	11	40	11
Número de neuronas interpoladas	2	2	2

1.7.7.Modelación con Algoritmo Perceptrón Multicapa (MLP)

En la sección 1.6.2, se presentó el estado del arte de publicaciones que utilizaron algoritmos de aprendizaje automático para resolver problemas de estimación de calidad del aire. En la mayoría de las publicaciones, el algoritmo seleccionado correspondió al Perceptrón Multicapa (MLP), en especial en (Perez and Reyes 2002), (Jiang, et al. 2004), (Lu, Hsieh and Chang 2006), (Díaz-Robles, Ortega, et al. 2008), (Kurt and Oktay 2010) y (Feng, et al. 2015).

Para la modelación con MLP, se ha utilizado la misma base de datos resultante, sin embargo, MLP corresponde a un método supervisado, por lo que no pueden existir datos faltantes en la base de datos utilizada para la modelación. Para eliminar los datos faltantes, se seleccionó el método de descartar todos los registros faltantes, descrito en la Tabla 4.

Para maximizar la capacidad de predicción del algoritmo MLP, se realizó una búsqueda en grilla de parámetros del algoritmo, seleccionando el conjunto de parámetros con mejores resultados. Los valores de la grilla de parámetros se presentan en la Tabla 9.

Tabla 9: Grilla de Parámetros de MLP

Decaimiento	N° de Neuronas
0,0001 – 0,001 – 0,01 – 0,1	4 – 5 – 6 – 7 – 8 – 9 – 10 – 11 – 12 – 13 – 14 -15

Del proceso de búsqueda de parámetros del algoritmo MLP, los valores que obtienen los mejores resultados varían dependiendo de la estación, según el registro de la Tabla 10.

Tabla 10: Parámetros que optimizan el resultado del algoritmo MLP

Estación de Monitoreo	Decaimiento	Unidades escondidas
Cesfam La Florida	0,001	15
Universidad Católica del Maule	0,0001	15
Universidad de Talca	0,001	15

1.7.8.Modelación con Algoritmo Random Forest (RF)

Para maximizar la capacidad de predicción del algoritmo RF, se realizó una búsqueda en grilla de parámetros del algoritmo, realizando pruebas con 100, 500 y 1000 árboles y 2 y 3 variables. Del proceso de búsqueda de parámetros del algoritmo RF, en las tres estaciones de monitoreo, los parámetros que obtuvieron los mejores fueron 1000 árboles y 3 variables.

Para la modelación con RF, se utiliza la base con datos faltantes eliminados, según lo descrito en la Tabla **4**.

1.7.9.Validación del Modelo

En las tres bases de datos, utilizaremos los índices de calidad del aire registrados en cada vector para estimar la calidad del aire de instancias obtenidas desde el mes de Julio de año 2015, donde conocemos parámetros de contaminación.

Nuestro modelo ha sido entrenado con los valores registrados desde el 1 de Abril de 2014 hasta el 30 de Junio de 2015.

Si bien, la validación del modelo se realizará con los registros a partir del 1 de Julio de 2015 al 30 de Agosto de 2015, no todas las estaciones contienen la misma cantidad de registros entre estas fechas, por lo que el porcentaje de registros de validación es distinto para cada estación. La Tabla 11 presenta el número de registros utilizados para validar en cada estación y el porcentaje con respecto al total de los registros de las bases de datos.

Tabla 11: Número de registros y porcentajes para validación de los modelos en las tres estaciones de monitoreo de calidad del aire.

Estación	N° de Registros para validar	Porcentaje de validación
Cesfam La Florida	1024	14,88
Universidad Católica del Maule	298	5,08
Universidad de Talca	891	13,70

1.8. Resultados Experimentales

En la presente sección se analizan y comparan los resultados de los modelos de regresión de los tres algoritmos de aprendizaje automático. Además, analizamos las capacidades de los algoritmos de predecir episodios de emergencia ambiental, resultados presentados a través de tablas de contingencia.

1.8.1. Resultados de regresión

Iniciamos el análisis de resultados del algoritmo TTOSOM, los que se presentan en la Tabla 12.

Tabla 12: Resultados de regresión del algoritmo TTOSOM

Métrica	Cesfam La Florida	UCM	UTALCA
R^2	0,6053	0,5667	0,4871
MAPE	44,59%	35,65%	42,98%
RSME (μgm-3)	26,78	16,20	16,08
RSE (μgm-3)	20,11	13,02	11,46

La Tabla 12 evidencia que las medidas de ajuste R^2 obtienen su valor más reducido en la estación Universidad de Talca, en tanto, la estación Cesfam La Florida obtienen el mayor ajuste. En tanto, la medida MAPE evidencia menores errores en la estación Universidad Católica del Maule.

Las medidas RSME y RSE presentan mejores resultados en las estaciones UCM y UTAL, presentando mayores errores en la Estación La Florida. Esta diferencia de resultados se debe que esta última estación de monitoreo presenta mediciones con peaks más pronunciados, lo que aumenta el rango entre las mediciones de concentración de $MP_{2,5}$.

En tanto, los resultados de la modelación con el algoritmo MLP se presentan en la Tabla 13.

Tabla 13: Resultados de regresión del algoritmo MLP

Métrica	Cesfam La Florida	UCM	UTALCA
R^2	0,7135	0,5384	0,5115
MAPE	26,58%	33,05%	37,03%
RSME (μgm-3)	22,82	16,74	15,70
RSE (μgm-3)	16,72	13,17	10,86

Los resultados del algoritmo MLP son similares a los obtenidos con el algoritmo TTOSOM, obteniendo mejores resultados en la estación Cesfam La Florida. La Tabla 13 evidencia que las medidas de ajuste R^2 y MAPE obtienen su valor más reducido en la estación Universidad de Talca. Las medidas RSME y RSE presentan mejores resultados en las estaciones UCM y UTAL, presentando mayores errores en la Estación La Florida.

Por último, los resultados de la modelación con el algoritmo RF se presentan en la Tabla 14.

Tabla 14: Resultados de regresión del algoritmo RF

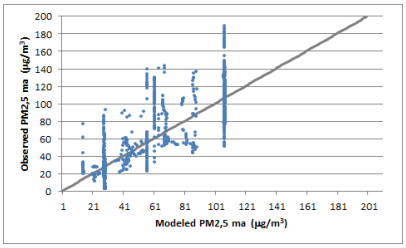
Métrica	Cesfam La Florida	UCM	UTALCA
R^2	0,9747	0,9447	0,9337
MAPE	6,38%	8,18%	9,57%
RSME (μgm^{-3})	6,78	5,79	5,78
RSE (μgm^{-3})	4,09	3,63	3,29

De las tres estaciones de monitoreo, los resultados del algoritmo RF son significativamente superiores a los obtenidos por los dos algoritmos neuronales artificiales.

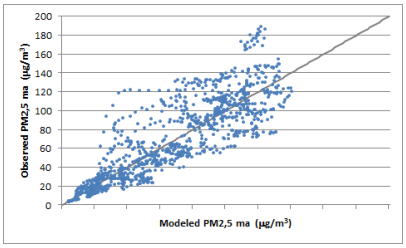
La Tabla 14 evidencia que las medidas de ajuste R^2 y MAPE obtienen su valor más reducido en la estación Universidad de Talca, en tanto, la estación Cesfam La Florida obtienen el mayor ajuste.

En la Figura 10 se comparan los valores de validación observados y modelados para los promedios diarios móviles de $MP_{2,5}$. En términos generales, en los tres algoritmos se observan una subvaloración de las estimaciones. Con respecto a los rendimientos, al igual que lo observado en los cuadros de

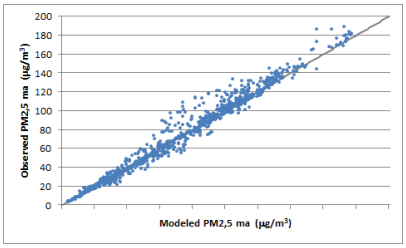
medidas de precisión, se evidencia la capacidad del algoritmo RF sobre los dos algoritmos neuronales artificiales.



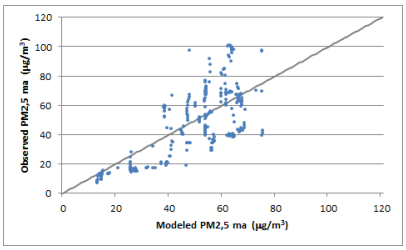
(a) Cesfam La Florida - TTOSOM



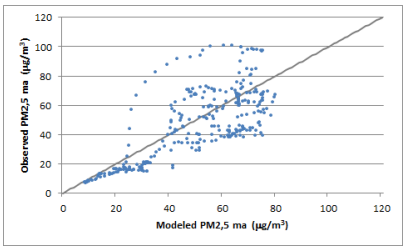
(b) Cesfam La Florida - MLP



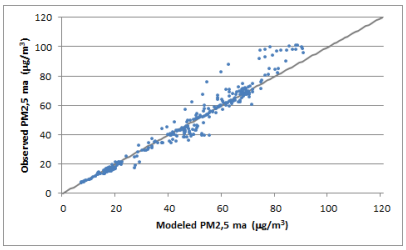
(c) Cesfam La Florida - RF



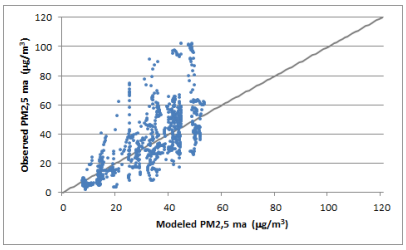
(d) UCM - TTOSOM



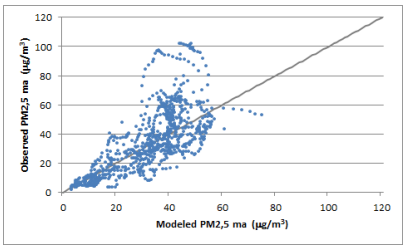
(e) UCM - MLP



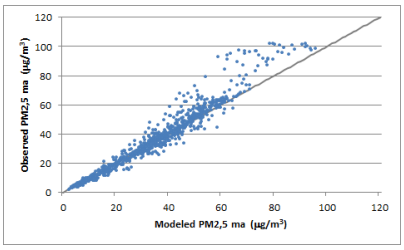
(f) UCM - RF



(g) UTALCA - TTOSOM



(h) UTALCA - MLP



(i) UTALCA - RF

Figura 10: Rendimientos de modelación en los datos de validación para el promedio diario móvil de $MP_{2,5}$.

1.8.2.Predicción de Episodios de Emergencia

En esta sección evaluamos las capacidades de los tres algoritmos para estimar episodios de emergencia ambiental. Cada episodio es dependiente del valor estimado de concentración, y etiquetado según lo señalado en la Tabla 1, donde se consideraron los 24 promedios móviles de un día, seleccionando el registro con mayor valor.

La Tabla **15** presenta las capacidades de los tres algoritmos de predecir episodios de emergencia en la estación Cesfam La Florida. El algoritmo TTOSOM obtiene un acierto del 52% y un 5% de falsas alarmas, tendiendo el modelo a subestimar los episodios de emergencia. El algoritmo MLP obtiene un acierto del 62% y un 14% de falsas alarmas, mejorando las capacidades de predicción de TTOSOM,

sin embargo, sin la capacidad de pronosticar los episodios más críticos. Por último, el algoritmo RF obtiene un 83% de aciertos y un 7% de falsas alarmas, presentando el algoritmo la capacidad de estimar los episodios más críticos.

Tabla 15: Tabla de contingencia de la estación Estación La Florida entre 01/07/2015 al 15/08/2015

		Medido			
		Normal	Alerta	Preemergencia	Emergencia
TTOSOM	Normal	16	2	2	
	Alerta	2	6	12	2
	Preemergencia				
	Emergencia				
MLP	Normal	16	3	1	
	Alerta	2	1	4	
	Preemergencia		2	11	2
	Emergencia				
RF	Normal	17	1		
	Alerta	1	5	3	
	Preemergencia		2	11	
	Emergencia				2

La Tabla 16 presenta los resultados de la estación Universidad Católica del Maule, donde el algoritmo TTOSOM obtiene un acierto del 82% y un 0% de falsas alarmas, ya que el modelo solo pronostica condiciones normales de calidad del aire. El algoritmo MLP obtiene un acierto del 73% y un 9% de falsas alarmas, errando en el pronóstico de un registro al señalar una condición de alerta. Por último, el algoritmo RF obtiene un 100% de aciertos y un 0% de falsas alarmas.

Tabla 16: Tabla de contingencia de la estación Universidad Católica del Maule entre 01/07/2015 al 14/07/2015

		Medido			
		Normal	Alerta	Preemergencia	Emergencia
TTOSOM	Normal	9	2		
	Alerta				
	Preemergencia				
	Emergencia				
MLP	Normal	8	2		
	Alerta	1			
	Preemergencia				
	Emergencia				
RF	Normal	9			
	Alerta		2		
	Preemergencia				
	Emergencia				

La Tabla 17 presenta los resultados de la estación Universidad de Talca, donde los algoritmos TTOSOM y MLP obtienen aciertos del 91% y un 0% de falsas alarmas, ya que ambos modelos solo pronostican condiciones normales de calidad del aire. El algoritmo RF obtiene un 97% de aciertos y un 0% de falsas alarmas, siendo la única falla una condición de alerta pronosticada como normal.

En el caso de los algoritmos TTOSOM y MLP, los resultados de las tablas de contingencia confirman lo planteado en la bibliografía, es decir, estos modelos presentan errores de subestimación de los valores extremos superiores. Este hecho representa una gran debilidad de ambos modelos, ya que el objetivo de los modelos de pronóstico es determinar episodios de emergencia ambiental, con el objetivo de tomar medidas y reducir la exposición de la población a altos índices de contaminación.

Por otra parte, los buenos resultados del algoritmo RF para estimar episodios de emergencia, nos permiten señalar que este algoritmo ha corregido la tendencia de presentar errores de subestimación, observados en los resultados de la regresión. Por ejemplo, cabe destacar la capacidad del algoritmo de estimar dos días de “Emergencia” ambiental en la estación Cesfam La Florida, donde el algoritmo TTOSOM predijo ambos episodios como “Alertas” y MLP como episodios de “Preemergencia”. Es importante mencionar que la mayoría de los artículos científicos han reportado las bondades de los algoritmos neuronales artificiales como regresor en problemas de calidad del aire, sin embargo,

nuestros resultados nos permiten señalar que RF supera significativamente las capacidades de los algoritmos neuronales artificiales en este tipo de problemas.

Tabla 17: Tabla de contingencia de la estación Universidad de Talca entre 01/07/2015 al 14/08/2015

		Medido			
		Normal	Alerta	Preemergencia	Emergencia
TTOSOM	Normal	31	3		
	Alerta				
	Preemergencia				
	Emergencia				
MLP	Normal	31	3		
	Alerta				
	Preemergencia				
	Emergencia				
RF	Normal	31	1		
	Alerta		2		
	Preemergencia				
	Emergencia				

1.9. Conclusiones

Si bien, la ciudad de Talca, una ciudad de la zona central de Chile, presenta altos índices de contaminación atmosférica, existen pocos estudios desarrollados en el área, más específicamente, estudios de predicción de calidad del aire. En el estudio propuesto en (Saide, Mena-Carrasco, et al. 2016), se desarrolla un modelo determinístico de pronóstico para nueve ciudades de Chile con monitoreo regular de la calidad del aire, entre ellas la ciudad de Talca. Los resultados de este estudio, señalan una capacidad de predicción en un 66% de episodios críticos con un día de anticipación, y un 39% de falsas alarmas, en la estación Cesfam La Florida.

Nuestro estudio propone utilizar tres algoritmos de aprendizaje automático para resolver el problema de pronóstico de la concentración de $MP_{2,5}$ en tres estaciones de monitoreo de la ciudad de Talca: Cesfam La Florida, Universidad Católica del Maule y Universidad de Talca.

Los algoritmos seleccionados fueron: MLP, el algoritmo neuronal artificial más utilizado, TTOSOM, un algoritmo basado en SOM con capacidad de procesar vectores originales, es decir, sin la necesidad

de modificar la distribución original de los datos modelados, y Random Forest uno de los algoritmos más precisos de la actualidad.

Nuestros resultados han sido entrenados con registros de un año y medio de mediciones y validados con datos posteriores a los de entrenamiento, de tal manera que reflejan un entorno bastante realista respecto a un escenario real.

De los resultados reportados de los modelos de regresión, podemos señalar que el algoritmo Random Forest obtuvo resultados significativamente superiores a los dos algoritmos neuronales artificiales. Incluso, Random Forest superó uno de los mayores problemas señalados en la literatura, la pobre representación de los modelos estadísticos del extremo superior de la base de datos, tendiendo a generar errores de subestimación.

Si bien, el modelo utiliza sólo dos variables meteorológicas, los errores de estimación son bajos, especialmente las estimaciones del algoritmo Random Forest, por lo que podemos señalar que este modelo es una alternativa para el pronóstico de calidad del aire en la ciudad de Talca, complementando el modelo meteorológico WRF-Chem, propuesto en (Saide, Mena-Carrasco, et al. 2016), o utilizando otros modelos meteorológicos para pronosticar las variables velocidad del viento y temperatura.

2. Artículo: Predicción de calidad del aire con el algoritmo Random Forest en la ciudad de Talca, Chile.

2.1. Introducción

La mala calidad del aire por altas concentraciones de $MP_{2,5}$ afecta la calidad de vida de la población de la ciudad de Talca. Los efectos de la contaminación atmosférica en las ciudades del centro y sur de Chile se han asociado significativamente con la mortalidad prematura y otros efectos negativos para la salud (Díaz-Robles, Cortés, et al. 2015), (Sanhueza, et al. 2009).

El material particulado con diámetro inferior a los 10 micrometros (MP_{10}) corresponde a un contaminante primario que afecta la calidad del aire, y por lo reducido de su tamaño puede ingresar al sistema respiratorio. Un subconjunto de este contaminante corresponde al material particulado con diámetro inferior a los 2,5 micrometros ($MP_{2,5}$), el cual genera mayores daños a la salud que el MP_{10}, por su menor diámetro y su capacidad de penetrar los alvéolos pulmonares.

En la zona central y de sur de Chile, la mala calidad del aire se observa en las ciudades del valle central, principalmente por las características meteorológicas, donde se observan períodos invernales una alta estabilidad atmosférica en los días que registran las menores temperaturas. En tanto, la mayor fuente emisora de $MP_{2,5}$ en las ciudades al sur de Santiago es la calefacción domiciliaria a leña. Las variables meteorológicas que más influyen en la concentración de material particulado respirable fino $MP_{2,5}$ en las ciudades del centro sur de Chile son la temperatura mínima y la velocidad del viento (Yañez, et al. 2017), relacionados con las emisiones desde calefactores a leña y la dispersión de los contaminantes, respectivamente.

En la ciudad de Talca, el año 2004 se inicia el monitoreo de la calidad del aire, con la instalación de una estación de medición discreta de MP_{10}. El año 2010 se declararon como zonas saturadas por MP_{10} las comunas de Talca y Maule (Decreto N° 12 de 4-2-2010 del Ministerio Secretaria General de la Presidencia), por la superación de las normas primarias de concentración de MP_{10} como concentración anual los años 2007 y 2008, como promedio aritmético de tres años calendarios consecutivos para los años 2005, 2006, 2007 y 2008; y excedida como concentración de 24 horas para los mismos años.

En Abril de 2013, se inauguran tres estaciones de monitoreo de calidad del aire, ubicadas en Universidad de Talca (35°24’24’’S; 71°37’60’’W), Universidad Católica del Maule (35°24’24’’S; 71°37’60’’W) y Cesfam La Florida (35°26’07’’S; 71°40’41’’W), las cuales capturan en forma continua los parámetros contaminantes MP_{10} y $MP_{2,5}$ y en frecuencia horaria los parámetros meteorológicos: presión atmosférica, humedad relativa del aire, temperatura ambiental, dirección y velocidad del viento. Además, en la Estación Cesfam La Florida se miden de forma continua los parámetros

contaminantes: Dióxido de azufre (SO2), monóxido de nitrógeno (NO), dióxido de nitrógeno (NO2), monóxido de carbono (CO), ozono (O3) y óxidos de nitrógeno (NOx).

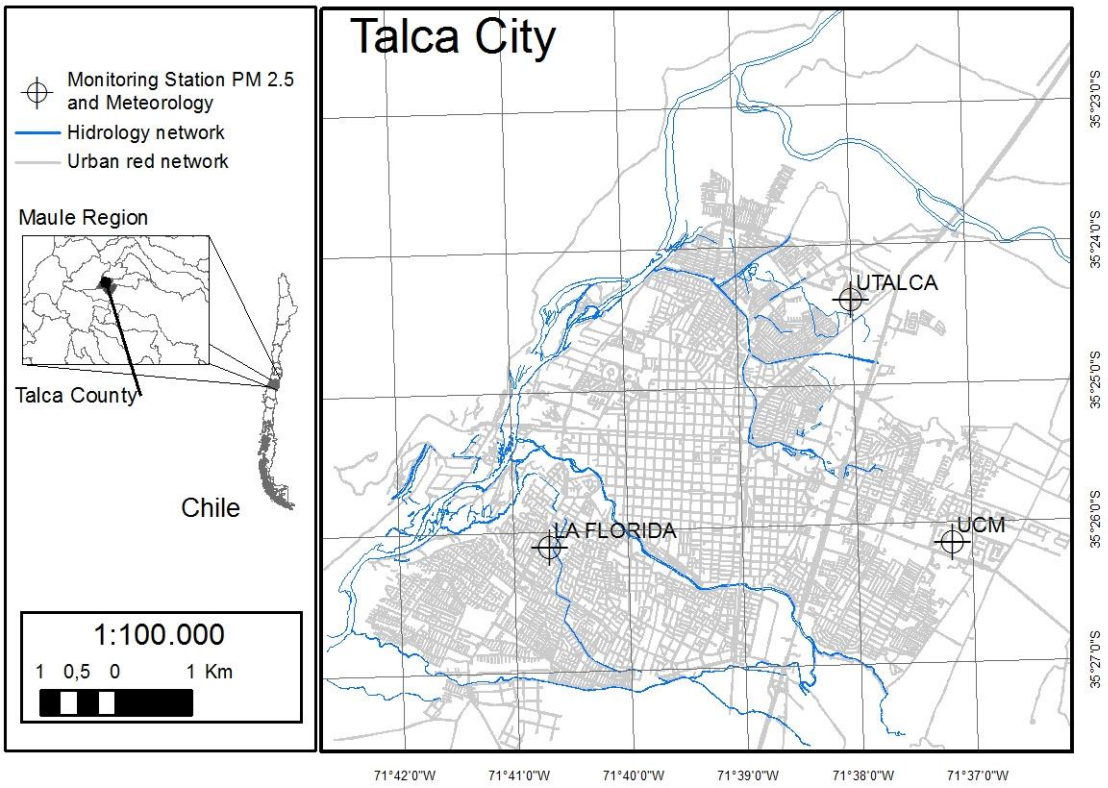


Figura 11: Estaciones de monitoreo de calidad del aire en la ciudad de Talca

Durante el primer año de medición, se contabilizaron un total de 86 días con superación de la norma de $MP_{2,5}$; existiendo 38 días con calidad del aire regular, 25 días con alerta, 17 días con preemergencia y 6 días con emergencia ambiental.

Según el valor de concentración de $MP_{2,5}$, la calidad del aire se puede clasificar según los niveles críticos por la norma primaria de calidad ambiental para material particulado fino respirable $MP_{2,5}$ (DS 12 de 01-01-2012 del Ministerio de Medio Ambiente), existiendo tres niveles de concentración de $MP_{2,5}$. El primer nivel denominado “Alerta”, se produce cuando la concentración de $MP_{2,5}$ se encuentra entre 80 y 109 $\mu g/m^3$. La siguiente categoría denominada “Preemergencia” ocurre entre los 110 y 169 $\mu g/m^3$. Finalmente la categoría más dañina para el organismo, denominada “Emergencia” ocurre cuando la concentración es mayor o igual a 170 $\mu g/m^3$.

Para adoptar medidas ante cualquier categoría de emergencia ambiental, es necesario estimar cuál será la condición del aire, al menos con un día de anticipación, y de esta forma, reducir los efectos adversos en la salud de la población. El desafío es predecir la concentración del promedio móvil 24 horas de $MP_{2,5}$ en el aire, a partir de información estadística de estaciones de monitoreo de calidad del aire.

2.1.1. Contribución del paper

1. Utilizamos uno de los algoritmos de aprendizaje automático existentes más certeros en la actualidad, denominado Random Forest para enfrentar la naturaleza no lineal del problema.

2. Construimos un modelo simple, ya que sólo incorpora dos variables meteorológicas: el promedio de temperatura y de velocidad del viento.

3. El algoritmo Random Forest obtiene buenos resultados de regresión.

2.1.2. Organización del paper

El trabajo se encuentra organizado de la siguiente forma. En la sección 2.2 se presenta el problema de estimación de la calidad del aire en la ciudad de Talca y un resumen de la literatura acerca de la utilización de algoritmos de aprendizaje automático en la estimación de problemas de calidad del aire.

La sección 2.3 muestra los materiales y métodos, iniciando en la construcción de la base de datos, luego la modelación del problema de estimación utilizando el algoritmo Random Forest. En la sección 2.4 se presentan los resultados, y finalmente la sección 2.5 las conclusiones.

2.2. El problema de estimación de calidad del aire

Este problema se refiere al cambio de la composición química del aire por emisiones antropogénicas a la atmósfera. El problema de predicción de calidad del aire ha incrementado su importancia en los últimos años debido a que afecta en el corto y largo plazo el bienestar humano (Carmichaela, et al. 2008).

El problema de predicción de calidad del aire presenta grandes incertidumbres asociadas a información incompleta, emisiones inexactas y procesos pobremente parametrizados (Carmichaela, et al. 2008).

Los problemas de predicción de la calidad de aire se han resuelto utilizando modelos deterministas y estadísticos. Los modelos deterministas no demandan una gran cantidad de datos históricos, sin embargo exigen un conocimiento de las condiciones meteorológicas y las fuentes contaminantes, la cantidad de emisiones en tiempo real, la descripción explícita de las principales reacciones químicas y procesos físicos en la capa inferior de la atmósfera (Feng, et al. 2015).

Los modelos estadísticos requieren una gran cantidad de datos históricos en diversas condiciones atmosféricas. El principal inconveniente de este enfoque es que un modelo representa sólo una estación específica y no se puede extender a otras regiones con diferentes condiciones meteorológicas, sin embargo, el enfoque estadístico es generalmente más apropiado para el descubrimiento de dependencias complejas de un sitio específico entre las concentraciones de

contaminantes atmosféricos y predictores potenciales y a menudo tienen una mayor precisión, que los modelos deterministas (Feng, et al. 2015).

Los métodos estadísticos más utilizados incluyen la regresión lineal múltiple (MLR), ANN, Support Vector Machine (SVM), Fuzzy logic, filtro de Kalman (KF) y el modelo oculto de Markov (HMM) (Feng, et al. 2015). Un inconveniente común con estos modelos, es que durante los días con altas concentraciones de material particulado (MP), los errores de previsión tienden a ser mucho mayores, y las concentraciones de MP son subestimadas sistematicáticamente (Feng, et al. 2015).

En Chile, la ciudad que ha centralizado la mayor atención con respecto a los problemas de calidad del aire, es la ciudad de Santiago, donde el gobierno ha desarrollado planes de descontaminación del aire desde el año 1997.

Esta política ha generado una activa investigación de modelos de pronóstico de calidad del aire, siendo ejemplo de ello (Cassmassi 1999), (Perez and Reyes 2002), (Pérez and Salini 2008), (Saide, Carmichael, et al. 2011).

En cambio, las ciudades de tamaño medio al sur de Santiago también han presentado episodios de alta polución del aire, sin embargo, han recibido una menor atención. En estas ciudades, los modelos de pronóstico de calidad del aire han sido limitados (Díaz-Robles, Ortega, et al. 2008).

Por las razones anteriormente expuestas en (Saide, Mena-Carrasco, et al. 2016), se desarrolla un modelo de pronóstico para nueve ciudades con monitoreo regular de la calidad del aire: Santiago, Rancagua, Curicó, Talca, Chillán, Los Ángeles, Temuco, Valdivia y Osorno. El modelo tiene como base el modelo químico y meteorológico en línea “Weather Research and Forecasting with Chemistry” (WRF-Chem) (Saide, Carmichael, et al. 2011), el cual predice concentración de CO. Los autores aproximan la predicción de $MP_{2,5}$ con la alta correlación de concentración de CO en los episodios críticos.

El sistema utiliza el modelo WRF-Chem, configurado para una grilla con celdas de 2 km para predecir el tiempo y las concentraciones de $MP_{2,5}$ por hora para tres días de anticipación, utilizando una calibración basada en la observación de variables locales, lo que permite un modelo menos intensivo computacionalmente que un modelo netamente químico. Las variables locales observadas incluyen una mayor probabilidad de ocurrencia de episodios críticos durante los fines de semana y los días más fríos, esto último relacionado con el aumento de las emisiones de las estufas a leña.

El modelo presenta estimaciones similares de temperatura para distintas inicializaciones atmosféricas, por lo que se evalúa eliminar la variable local de días más fríos, en tanto, la capacidad de predecir la velocidad del viento mejora desde Curicó al sur.

Para la estación La Florida de la ciudad de Talca, en el período entre abril y agosto de 2014, los resultados reportados del modelo señalan la capacidad de predecir episodios críticos en un 66% con

un día de anticipación, 64% con dos días de anticipación y 59% con tres días de anticipación. En tanto, las falsas alarmas corresponden a 39% para un día de anticipación, 35% para dos días de anticipación y 37% para tres días de anticipación.

2.3. Materiales y métodos

El comportamiento y la estimación de la calidad del aire en la ciudad de Talca se basan en el análisis de un conjunto de datos de parámetros contaminantes.

Pare ello, se preparó este conjunto de datos de parámetros contaminantes normalizando sus variables, manejando los datos faltantes y eliminando valores atípicos. Una vez construido el conjunto de datos, fue utilizado para predecir la calidad del aire, utilizando el algoritmo de aprendizaje automático denominado Random Forest y comparando resultados según las medidas de desempeño en los datos para validación.

2.3.1. Conjunto de Datos

En Chile, el Ministerio de Medio Ambiente es el organismo público responsable de la gestión del ambiente y de los recursos renovables naturales. Para la administración de la información de calidad del aire, el ministerio ha desarrollado el Sistema de Información Nacional de Calidad del Aire (SINCA), el cual administra una red de estaciones de monitoreo de calidad del aire y meteorológicas. A través de su portal web, se presentan las mediciones de calidad en línea, el seguimiento histórico de las mediciones de calidad del aire y meteorología, antecedentes de las estaciones de monitoreo, documentación relacionada con calidad del aire, monitoreo y enlaces a sitios web nacionales e internacionales.

El conjunto de datos confeccionado corresponde a los datos de las estaciones Cesfam La Florida, U.C. Maule y Universidad de Talca, con información entre los meses de Abril y Agosto de los años 2014, 2015, 2016 y 2017.

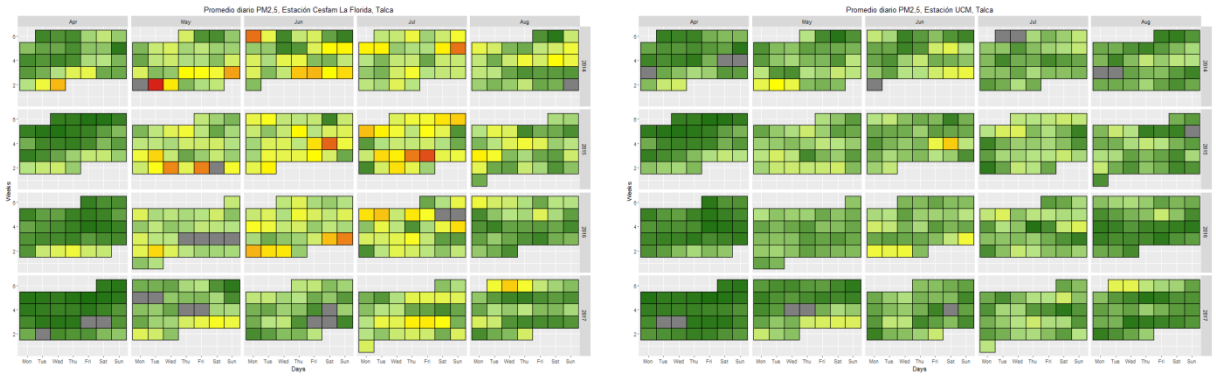
La elección de los registros en las fechas señaladas, se debe a que la ciudad de Talca presenta el problema de contaminación del aire en la temporada invernal. Además, al eliminar los registros de calidad del aire de temporadas estivales, se eliminan registros con altos índices de contaminación asociados a otros factores, por ejemplo incendios forestales.

Para cada estación, se ha construido un conjunto de datos con 14.070 filas, donde cada fila representa los promedios diarios móviles de las variables de contaminación, entre el 01 de Abril de 2014 y el 31 de Agosto de 2017.

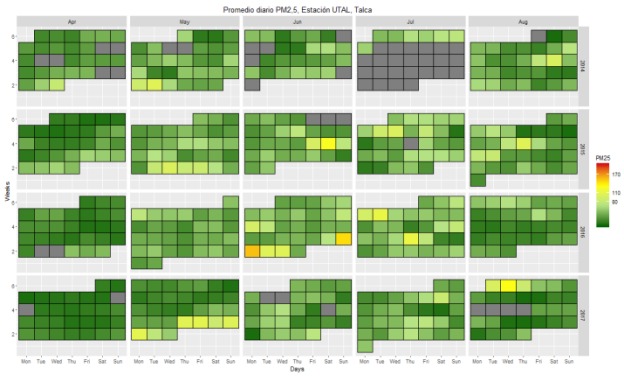
El promedio móvil se calcula a partir del registro de contaminación de un día y hora con los 23 registros anteriores de contaminación. Los promedios diarios son obtenidos cuando existe al menos

18 horas de registros horarios en un día, esto según la metodología señalada en DS N° 61 de 2008 del Ministerio de Salud.

La Figura 12 presenta el comportamiento del promedio diario de $MP_{2,5}$ para las tres estaciones entre los meses de Abril y Agosto de los años 2014, 2015, 2016 y 2017. Los colores representan la contaminación de cada día, siendo el verde las concentraciones más bajas, el verde claro las concentraciones asociadas a episodios de alerta ambiental, los amarillos asociados a preemergencias ambientales y naranjos y rojos en las emergencias. Los casilleros en gris representan los días sin información.



(a) Estación Cesfam La Florida (b) Estación Universidad Católica del Maule



(c) Estación Universidad de Talca

Figura 12: Promedio Diario de $MP_{2,5}$ entre los meses de Abril y Agosto entre los años 2014 y 2017

Como se puede apreciar en la Figura 12, en general, los días con registros bajo 80 $\mu g/m^3$, son los valores mayormente registrados en las tres estaciones, sin embargo la estación Cesfam La Florida presenta los peores registros de contaminación en la ciudad, siendo la única estación que presenta períodos de “Emergencias”.

Por otra parte, el comportamiento diario de las concentraciones en la ciudad, es representado por la Figura 13, donde se presentan los promedios de los registros horarios de $MP_{2,5}$ para las tres estaciones, entre los meses de Abril y Agosto de los años 2014 y 2017. El rojo representa la estación Cesfam La Florida, en azul la estación UCM y en verde la estación UTalca.

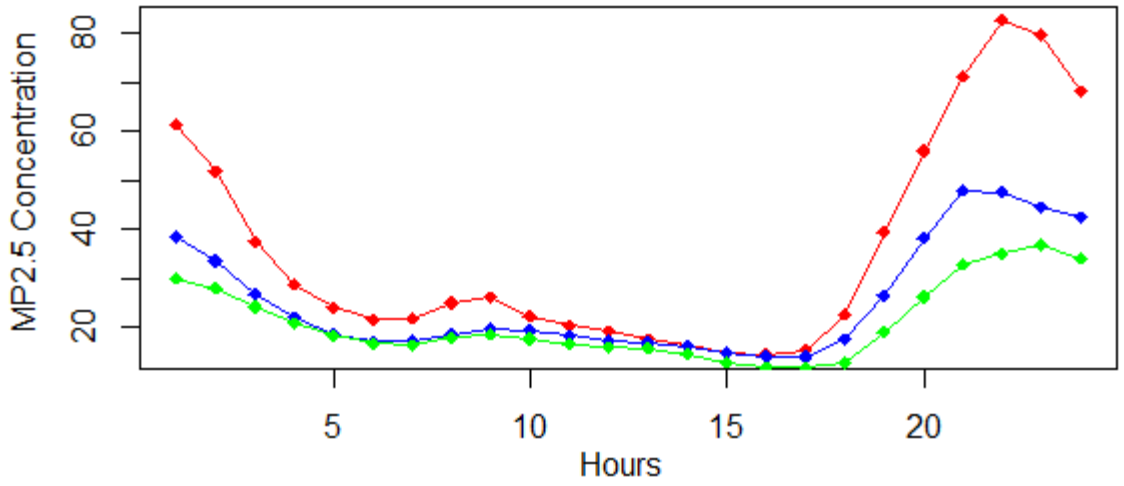


Figura 13: Promedios horarios de registros de $MP_{2,5}$ entre los meses de Abril de 2014 y Agosto de 2017

En la Figura 13, en las tres estaciones se aprecia que a partir de las 17 horas empieza el aumento de la concentración promedio de $MP_{2,5}$, con un peak a las 22 horas. Este aumento está directamente relacionado a las emisiones de calefactores a leña de las residencias de la ciudad, disminuyendo las emisiones en el período nocturno. En la Figura 13 también se observa un peak a las 9 horas, el cual se encuentra asociado a la salida de vehículos en el horario punta mañana.

La Tabla 18 refleja una relación inversa entre la concentración 24 horas de $MP_{2,5}$ con las variables temperatura promedio 24 horas, velocidad del viento 24 horas y una relación directa con la variable concentración 24 horas de $MP_{2,5}$ del día anterior. En la estación Cesfam La Florida, la temperatura promedio 24 horas es la variable más correlacionada a los índices de contaminación, en tanto, en las estaciones UCM y UTAL, la velocidad del viento promedio 24 horas es la variable que mayor explica los índices de contaminación.

Tabla 18: Correlación entre la concentración 24 horas de $MP_{2,5}$ con las variables temperatura promedio 24 horas, velocidad del viento 24 horas y concentración 24 horas de $MP_{2,5}$ del día anterior en las estaciones Cesfam La Florida, U.C. Maule y Universidad de Talca.

Estación de Monitoreo	Variables de Modelación		
	Temperatura	Velocidad del Viento	MP2,5 del día anterior
Cesfam La Florida	-0,62	-0,52	0,50
UCM	-0,47	-0,58	0,48
UTALCA	-0,42	-0,60	0,43

2.3.2.Procesamiento del Conjunto de Datos

Según lo señalado en (U.S., Environmental Protection Agency 2003), las variables de concentraciones de partículas y registros meteorológicos presentan una distribución log-normal con un marcado comportamiento estacional, (U.S., Environmental Protection Agency 2003). Por lo que, se ha utilizado la función logaritmo natural en todas nuestras variables para obtener una estructura normal de distribución de datos. Luego, los parámetros de las variables del conjunto de datos original son llevados al rango [0,1].

Las bases de datos han sido construidas pensando en métodos con la capacidad de procesar datos faltantes de manera directa.

2.3.3.Random Forest (RF)

Random Forest es un algoritmo propuesto por (Breiman 2001), el cual combina el método random subspace y bagging. El algoritmo es uno de los más certeros en la actualidad y ha sido utilizado para resolver una gran cantidad de tareas (Verikas, Gelzinis and Bacauskiene 2011).

Inicialmente tenemos un conjunto de datos de entrenamiento X_t = $\{f(x_m, y_m)$, m =(1,…, m)}, donde x_m es una variable de entrada y y_m una variable de salida. Un aprendiz débil puede ser creado utilizando el conjunto de entrenamiento X_t, donde el aprendizaje débil es un predictor $f(x, X_t)$ con bajo sesgo y alta varianza. En RF un árbol de decisión es utilizado como aprendiz débil.

Mediante un muestreo aleatorio del conjunto X_t, un conjunto de árboles de decisión $f(x, X_t, \theta_k)$ puede ser creado, con $f(x, X_t, \theta_k)$ el k-ésimo árbol de decisión y θ_k es el vector aleatorio que selecciona los puntos de datos para el k-ésimo árbol de decisión. Al aplicar el muestreo bootstrap para generar θ_k, por ejemplo, se utiliza dos tercios de los datos para cada árbol de decisión, y cerca de un tercio quedan fuera del muestreo bootstrap.

Una de las características de θ_k es ser independiente e idénticamente distribuidos (iid), por lo que al combinar los árboles aleatorios a través de un promedio, el sesgo se mantiene prácticamente inalterado, en tanto la varianza se reduce por un factor de $\bar{\rho}$ el valor promedio de correlación entre los árboles de aprendizaje.

La Figura 14 presenta una estructura general de RF, donde B es el número de árboles en RF y k_1, k_2,…, k_b son las etiquetas de las clases. A medida que aumenta el número de árboles, las tasas de error convergen a un límite, por lo que no hay sobreajuste por grandes números de árboles.

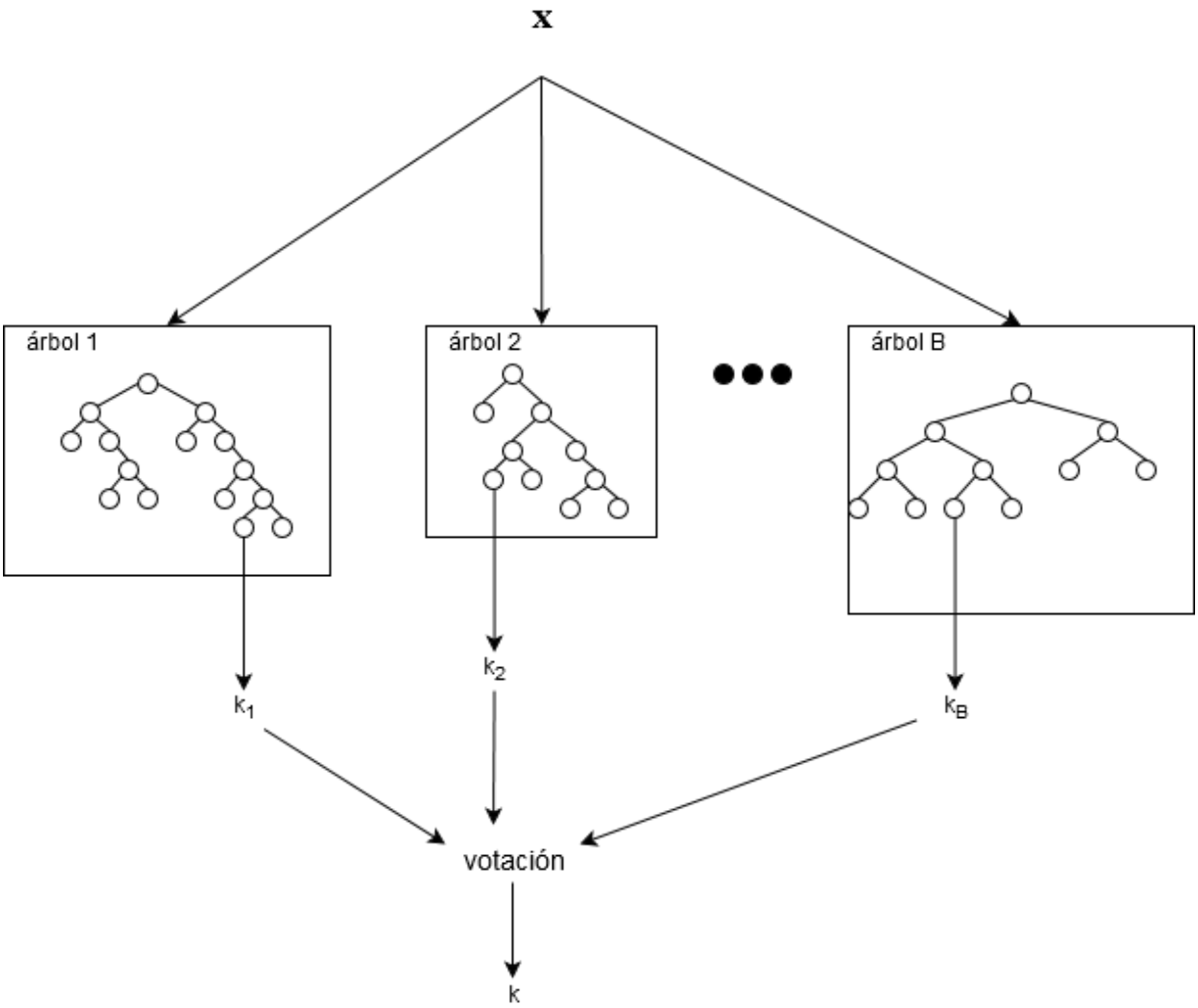


Figura 14: Arquitectura general de Random Forest (Verikas, Gelzinis and Bacauskiene 2011).

2.3.4.Medidas de precisión para modelos de regresión

El resultado de las regresiones fue medida utilizando los estadísticos: Coeficiente de determinación (R^2), Mean Absolute Percentage Error (MAPE), Root Mean Error (RSE) y Root Mean Squared Error (RSME).

El estadístico R^2 representa la proporción de varianza explicada, presentando valores entre 0 y 1, independiente de la escala de la variable explicada.

En tanto, el error porcentual absoluto medio (MAPE), expresa el error medio de una estimación, presentando valores porcentuales. Valores cercanos a 0% indican un bajo error medio de la regresión.

RSE y RSME son estimaciones de la desviación estándar del error en las mismas unidades del modelo. Valores pequeños de ambas pruebas se consideran como un buen ajuste del modelo a los datos, en tanto, valores grandes indican que el modelo no se ajusta bien a los datos.

2.3.5. Validación de Modelos

Nuestros algoritmos han sido entrenados con los valores registrados desde el 1 de Abril de 2014 hasta el 31 de Agosto de 2016.

La validación de los algoritmos se realizará con los registros a partir del 1 de Abril de 2017 al 31 de Agosto de 2017, por lo tanto, los algoritmos fueron entrenados con el 75% de los datos y validados con el 25% de los datos restantes.

2.4. Resultados

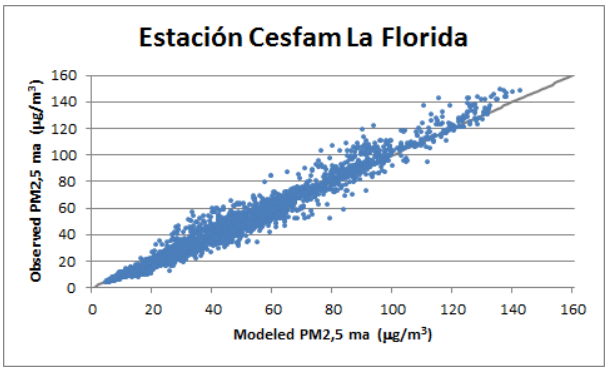
En la presente sección se presentan y analizan los resultados de la regresión con el algoritmo Random Forest. Además, analizamos la capacidad del algoritmo de predecir episodios de emergencia ambiental, resultados presentados a través de tablas de contingencia. La Tabla 19, presenta los resultados de la regresión utilizando el modelo.

Las tres estaciones de monitoreo obtienen buenos resultados de estimación, obteniendo el valor más reducido en la estación Universidad de Talca, en tanto, la estación Cesfam La Florida obtienen el mayor ajuste.

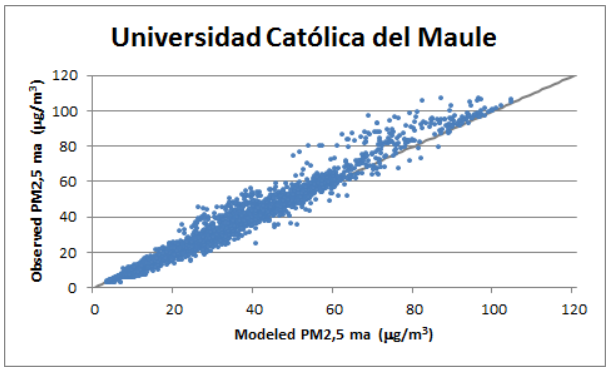
Tabla 19: Resultados de regresión del algoritmo RF

Métrica	Cesfam La Florida	UCM	UTALCA
R^2	0,9725	0,9670	0,9569
MAPE	8,98%	8,37%	12,44%
RSME (μgm^{-3})	5,21	3,93	4,52
RSE (μgm^{-3})	3,33	2,32	2,43

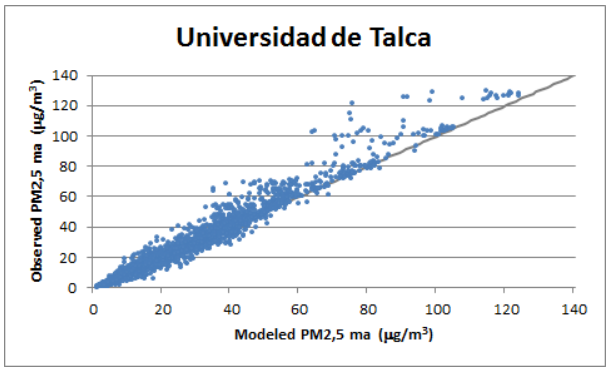
En la Figura 15 se comparan los valores de validación observados y modelados para los promedios diarios móviles de $MP_{2,5}$. En términos generales, en las tres estaciones se observan una subvaloración de las estimaciones.



(a) Estación Cesfam La Florida



(b) Estación Universidad Católica del Maule



(c) Estación Universidad de Talca

Figura 15: Rendimientos de modelación en los datos de validación para el promedio diario móvil de $MP_{2,5}$

2.4.1.Predicción de Episodios de Emergencia

Con respecto a la capacidad de predicción de episodios de emergencia, la Tabla 20 presenta los resultados obtenidos a partir del modelo de regresión, donde en la estación Cesfam La Florida se obtiene un 97,64% de aciertos y un 1,58% de falsas alarmas, presentando el algoritmo la capacidad de estimar los episodios más críticos; en la estación UCM un 98,12% de aciertos y un 0,09% de falsas alarmas y por último, en la estación Universidad de Talca un acierto del 98,84% y un 0,09% de falsas alarmas.

Los buenos resultados del algoritmo RF para estimar episodios de emergencia, nos permiten señalar que este algoritmo ha corregido la tendencia de presentar errores de subestimación, observados en los resultados de la regresión.

Tabla 20: Tabla de Contingencia de niveles de calidad del aire

		Estimado		
Estación	Medido	Normal	Alerta	Preemergencia
Cesfam La Florida	Normal	2861	21	
	Alerta	35	296	5
	Preemergencia		18	113
UCM	Normal	3346	3	
	Alerta	63	95	
	Preemergencia			
UTALCA	Normal	3344	3	
	Alerta	28	56	
	Preemergencia	3	6	14

2.5. Conclusiones

La ciudad de Talca presenta altos episodios de contaminación del aire en el período invernal, sin embargo existen pocos estudios desarrollados donde se estime la calidad del aire en la ciudad y las herramientas de pronóstico son limitadas. En el presente estudio desarrollamos y evaluamos un modelo de pronóstico capaz de predecir episodios con un día de anticipación para la ciudad de Talca.

Al igual que varias ciudades del centro y sur del país, en la ciudad de Talca el problema de calidad del aire por concentraciones de $MP_{2,5}$ se encuentran mayormente asociado a las emisiones de calefactores a leña, evidenciado en los mayores índices de contaminación en el período invernal y con un incremento diario de los contaminantes a partir de las 17 horas, con un peak a las 23 horas de cada día.

En el período 2014 - 2017, la estación Cesfam La Florida registra los mayores índices de contaminación de la ciudad. En esta estación, las concentraciones de $MP_{2,5}$ se encuentran mayormente correlacionadas con la variable temperatura ambiental, donde el descenso de la temperatura ambiental genera mayores emisiones por parte de calefactores a leña.

En tanto, las estaciones Universidad Católica del Maule y Universidad de Talca se registran menores registros de contaminación, donde las concentraciones de $MP_{2,5}$ se encuentran mayormente correlacionadas con la variable velocidad del viento, comportamiento similar al reportado en otros estudios para ciudades de Curicó al sur.

El algoritmo Random Forest fue utilizado para estimar la calidad del aire en la ciudad, cuyos resultados superaron uno de los mayores problemas señalados en la literatura, la pobre representación de los modelos estadísticos del extremo superior del conjunto de datos, tendiendo a generar errores de subestimación.

Si bien, el modelo utiliza sólo dos variables meteorológicas (velocidad del viento y temperatura ambiental promedio), los errores de estimación son bajos, por lo que podemos señalar que este modelo es una alternativa para el pronóstico de calidad del aire en la ciudad de Talca, complementando el modelo meteorológico WRF-Chem, propuesto en (Saide, Mena-Carrasco, et al. 2016), ya que simplifica la calibración de variables locales.

Creemos que los buenos resultados obtenidos del modelo de regresión pueden ser extrapolados a las ciudades desde Curicó al sur, ya que también la contaminación del aire en estas ciudades se encuentra mayormente asociada a las emisiones de calefactores a leña y además, desde el punto de vista operativo, el sistema de pronóstico actual (Saide, Mena-Carrasco, et al. 2016) entrega buenas estimaciones de temperatura ambiental y velocidad del viento, variables utilizadas por nuestro modelo.

2.6. Agradecimientos

Agradezco a mi esposa e hija, por el apoyo en este gran esfuerzo que hemos realizados durante estos años de estudio.

A mis padres, quienes gracias a su esfuerzo me permitieron crecer profesionalmente y como persona.

A mis suegros, la tita Eva y el tata José, por apoyarnos en los cuidados de nuestra pequeña.

A Rodrigo Fica, profesional de la Seremi de Medio Ambiente de la región del Maule, por el recibimiento y la buena disposición de apoyar este trabajo.

A mi profesor guía, César Astudillo, por creer en mi propuesta de trabajo, apoyar con su conocimiento y guiar esta investigación.

El trabajo de esta Tesis fue parcialmente financiada por el proyecto FONDECYT 11121350 “Tree-Based Pattern Recognition”.

2.7. Referencias Bibliográficas

Astudillo, César, and John Oommen. "Imposing tree-based topologies onto self organizing maps." *Information Sciences*, no. 181 (2011): 3798–3815.

Bougoudis, I., K. Demertzis, and L. Iliadis. "HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens." *Neural Computing and Applications*, 2015.

Breiman, Leo. «Random Forest.» *Machine Learning* 45, nº 1 (2001): 5-32.

Brunelli, U, V Piazza, L Pignato, F Sorbello, and S Vitabile. "Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy." *Atmospheric Environment* 41, no. 14 (2007): 2967–2995.

Carmichaela, Gegrory, Adrian Sandub, Tianfeng Chaia, Dacian Daescuc, Emil Constantinescub, and Youhua Tang. "Predicting air quality: Improvements through advanced methods to integrate models and measurements." *Journal of Computational Physics* 227, no. 7 (2008): 3540–3571.

Cassmassi, , Joseph. *Improvement of the forecast of air quality and of the knowledge of the local meteorological conditions in the metropolitan region.* Informe Final, CONAMA, 1999.

Díaz-Robles, Luis, et al. "A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile." *Atmospheric Environment*, no. 42 (2008): 8331-8340.

Díaz-Robles, Luis, et al. "A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile." *Atmospheric Environment* 42, no. 35 (2008): 8331-8340.

Díaz-Robles, Luis, Samuel Cortés, Alberto Vergara-Fernández, and Juan Carlos Ortega. "Short Term Health Effects of Particulate Matter: A Comparison between Wood Smoke and Multi-Source Polluted Urban Areas in Chile." *Aerosol and Air Quality Research* 15, no. 1 (2015): 306-318.

Feng, Xiao, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, and Jingjie Wang. "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation." *Atmospheric Environment* 107 (2015): 118–128.

Florinsky, Igor. *Digital terrain analysis in soil science and geology.* Pushchino: Elsiever, 2016.

Fox, John. *Regression diagnostics: An introduction.* SAGE Publications, 1991.

Gardner, M.W., and S.R. Dorling. "Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences." *Atmospheric Environment* 32, no. 14-15 (1998): 2627–2636.

Grillenzoni, Carlo. «Forecasting unstable and nonstacionary time series.» *Internacional Journal of Forecasting* 14, nº 4 (1998): 469-482.

Hooyberghs, Jef, Clemens Mensink, Gerwin Dumont, Frans Fierens, and Olivier Brasseur. "A neural network forecast for daily average PM10 concentrations in Belgium." *Atmospheric Environment* 39, no. 18 (2005): 3279–3289.

Hrust, Lovro, Zvjezdana Bencetić Klaićb, Josip Križan, Oleg Antonić, and Predrag Hercog. "Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations." *Atmospheric Environment* 43, no. 35 (2009): 5588–5596.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* New York: Springer, 2013.

Jiang, Dahe, Yang Zhang, Xiang Hu, Yun Zeng, Jianguo Tan, and Demin Shao. "Progress in developing an ANN model for air pollution index forecast." *Atmospheric Environment* 38, no. 40 (2004): 7055–7064.

Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen. "Methods for imputation of missing values in air quality data sets." *Atmospheric Environment* 38, no. 18 (2004): 2895–2907.

Kohonen, T. *Self-Organizing Maps.* New York, Inc., Secaucus, NJ, USA: Springer-Verlag, 1995.

Kolehmainen, Mikko, H Martikainen, and J Ruuskanen. "Neural networks and periodic components used in air quality forecasting." *Atmospheric Environment* 35, no. 5 (2001): 815-825.

Kukkonen, Jaakko, et al. "Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki." *Atmospheric Environment* 37, no. 32 (2003): 4539–4550.

Kurt, Atakan, and Ayşe Betül Oktay. "Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks." *Expert Systems with Applications* 37, no. 12 (2010): 7986–7992.

Lu, Hsin-Chung, Jen-Chieh Hsieh, and Tseng-Shuo Chang. "Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network." *Atmospheric Research* 81, no. 2 (2006): 124–139.

Ministerio de Medio Ambiente. *Sistema de Información Nacional de Calidad del Aire.* 2009-2015. https://sinca.mma.gob.cl (último acceso: 01 de 07 de 2018).

Pérez, Patricio, and Giovanni Salini. "PM2.5 forecasting in a large city: Comparison of three methods." *Atmospheric Environment* 45, no. 35 (2008): 8219-8224.

Perez, Patricio, and Jorge Reyes. "Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile." *Atmospheric Environment* 36, no. 28 (2002): 4555–4561.

Saide, Pablo, et al. "Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model." *Atmospheric Environment* 45, no. 16 (2011): 2769-2780.

Saide, Pablo, Marcelo Mena-Carrasco, Sesbastian Tolvett, Pablo Hernandez, and Gregory Carmichael. "Air quality forecasting for winter-time PM2.5 episodes occurring in multiple cities in central and southern Chile." *Journal of Geophysical Research: Atmospheres* 121, no. 1 (2016): 558–575.

Sanhueza, Pedro, Monica Torreblanca, Luis Diaz-Robles, Nicolas Schiappacasse, Maria Silva, and Teresa Astete. "Particulate Air Pollution and Health Effects for Cardiovascular and Respiratory Causes in Temuco, Chile: A Wood-Smoke-Polluted Urban Area." *Journal of the Air & Waste Management Association* 59, no. 12 (2009): 1481-1488.

Siwek, K, and S Osowski. "Improving the accuracy of prediction of PM10 pollution by the wavelet transformation and an ensemble of neural predictors." *Engineering Applications of Artificial Intelligence* 25, no. 6 (2012): 1246–1258.

Stevens, J.P. "Outliers and influential data points in regression analysis." *Psychological Bulletin* 95, no. 2 (1984): 334-344.

U.S., Environmental Protection Agency. *Guidelines for Developing an Air Quality (Ozone and PM2.5) Forecasting Program.* Triangle Park, North Carolina, 2003.

Verikas, Antanas, Adas Gelzinis, and Marija Bacauskiene. "Mining data with random forests: A survey and results of new tests." *Pattern Recognition* 44, no. 2 (2011): 330-349.

Yañez, Marco, Ricardo Baettig, Jorge Cornejo, Francisco Zamudio, Jorge Guajardo, and Rodrigo Fica. "Urban airborne matter in central and southern Chile: Effects of meteorological conditions on fine and coarse particulate matter." *Atmospheric Environment* 161 (Julio 2017): 221-234.