

INFORME TESIS

Modelo predictivo para hospitalizaciones por enfermedades respiratorias.

Pablo Alfonso Torrejón Aguirre

En el siguiente documento se encuentra el informe final del proyecto de tesis para el título de magister de Economía y Políticas Públicas.

Profesor Guía: Luis Herskovic.

Profesora Correctora: María Nieves Valdés

Profesor Corrector: Claudio Aracena

Resumen

Uno de los grandes desafíos que enfrenta el mundo hoy en día y sus futuras generaciones son los efectos del cambio climático. Sus consecuencias están ligadas a varios factores siendo la contaminación ambiental uno de gran importancia no sólo por su responsabilidad en el aumento de las temperaturas en el mundo sino también por los problemas que trae a corto, mediano y largo plazo para la salud y economía de las personas. Hoy en día, las investigaciones sobre las consecuencias atribuidas a la contaminación atmosférica en la salud han logrado avances significativos en entender como ambas interactúan.

Esta investigación pretende entender si es posible, a partir de modelos de Machine Learning, predecir la cantidad de hospitalizaciones a causa de enfermedades respiratorias en comunas de Santiago de Chile utilizando variables de contaminación ambiental y de temperatura obtenidas por medio de una interpolación espaciotemporal. Los modelos seleccionados para el trabajo son CART y Random Forest que para efectos de este trabajo son explorados utilizando la comuna de Puente Alto en función de los siguientes predictores: tipo de enfermedad según, grupo etario, niveles de MP2.5, niveles de MP10 y grados de temperatura.

Aunque los modelos predictivos no necesariamente permiten concluir algo sobre la causalidad en sí misma de la contaminación sobre las enfermedades respiratorias, estos modelos pueden ayudar a estimar el impacto causal de una intervención que reduzca los niveles de contaminación cuando se realice. Sin embargo, quedan aún desafíos para poder seguir explorando estos mecanismos y en este caso en particular con mayores estaciones de medición, mejoras en las mediciones de otras variables medioambientales y en el gobierno de datos de salud con el fin de poder obtener mejores datos y así ir mejorando la precisión de los modelos.

1. Introduccion

Uno de los grandes desafíos que enfrenta el mundo hoy en día y sus futuras generaciones son los efectos del cambio climático. Sus consecuencias están ligadas a varios factores siendo la contaminación ambiental uno de gran importancia no sólo por su responsabilidad en el aumento de las temperaturas en el mundo sino también por los problemas que trae a corto, mediano y largo plazo para la salud y economía de las personas.

Según datos de la Organización Mundial de la Salud (WHO) en 2016 cerca del 92% de la población mundial vive en lugares que están expuesto a un nivel superior al recomendado de $10 \mu\text{g} / \text{m}^3$ de MP2.5 promedio anual. La misma organización estimó que en 2012, una de cada nueve muertes estaba relacionada con las condiciones de contaminación ambiental. De esas muertes, cerca de 3 millones alrededor del mundo se relacionan con la polución y para Chile aproximadamente 3 mil (WHO, 2016 Ambiente air pollution). Es más, las proyecciones no son optimistas y para el 2060 se estima que alrededor del mundo mueran entre 6 y 9 millones de personas por causas relacionadas costando anualmente cerca del 1% de PIB global por enfermedades, costos médicos y pérdidas de cosechas (OCDE, 2019; "The Air That We Breathe: Addressing the Risks of Global Urbanization on Health", 2012)).

Los efectos mencionados anteriormente son solo una dimensión de los daños producidos por contaminación ambiental ya que hay efectos en salud a nivel individual que no requieren hospitalización y que si lo vemos a nivel agregado causan un impacto económico global afectando el mercado laboral por medio del ausentismo y el capital humano perjudicando a los individuos desde los meses de gestación, en el aprendizaje temprano de los escolares e incluso en la etapa adulta (Zivin & Neidell, 2018).

La contaminación del aire incluye una amplia variedad de componentes químicos y biológicos siendo los contaminantes atmosféricos riesgosos para la salud el material particulado y compuestos químicos gaseosos como el monóxido de carbono, ozono, dióxido de azufre y dióxido de carbono. Las personas suelen estar expuestos a la mezcla de ellos, y pueden incluso potenciarse entre sí mismos para afectar la salud de las personas (Oyarzún G, 2010). Si se considera la mortalidad relacionada a la exposición de MP2.5, las variables que muestran una influencia significativa en los resultados son la electricidad limpia, consumo de gasolina para motores, combustible para cocinar, densidad de población y educación (Requia et al., 2018).

El material particulado se define según su tamaño, aquellas partículas que tienen un diámetro menor o igual a $10 \mu\text{m}$ como MP10 y de $2.5 \mu\text{m}$ como MP2.5, y se denominan respirables dado la capacidad que tienen de introducirse en las vías respiratorias. El ozono, es un contaminante formado por la reacción química entre hidrocarburos y óxidos de nitrógeno en presencia de calor y luz solar, es un componente mayor de la contaminación urbana y un potente irritante que puede aumentar la reacción de los niños ante otros contaminantes (Gavidia, Pronczuk & Sly, 2009).

La ciudad de Santiago de Chile es una de las ciudades que más contaminación atmosférica presenta en Latinoamérica, su particular geografía genera condiciones desfavorables para la ventilación del aire (Perez & Reyes, 2006; Oyarzún G & Valdivia C, 2012) haciendo que los niveles de ciertos contaminantes sobrepasen los niveles determinados por la OMS provocando que durante los meses de invierno se genere alta demanda por atenciones de urgencia como consecuencia de problemas respiratorios y cardiovasculares siendo las personas más vulnerables, principalmente los niños y adultos mayores los más

propensos a estos efectos (Gavidia, Pronczuk & Sly, 2009; Oyarzún G, 2010) y los que soportan la mayor carga de las consecuencias. En otoño e invierno se caracteriza por la presencia de altos niveles MP10 y MP2.5, mientras que en primavera y verano por altos niveles de oxidantes fotoquímicos como el Ozono (Academia medicina de Chile, 2012). La exposición a contaminantes del aire, especialmente PM2.5, NO2 y O3, exacerba los síntomas respiratorios tanto en niños asmáticos como no asmáticos (Prieto-Parra et al., 2017). Por otro lado, un incremento de 10 $\mu\text{g}/\text{m}^3$ de MP2.5 se relaciona con un aumento en alrededor de 2% las hospitalizaciones por causas respiratorias con 1 a 2 días de rezago desde la exposición y que dicho porcentaje aumentó al 5% del total de las hospitalizaciones por enfermedades respiratorias en niños, cuando la exposición de produjo con 8 días de rezago (Matus C. & Oyarzún G., 2019). Por otro lado, un cambio en el nivel promedio de MP10 de 10 $\mu\text{g} / \text{m}^3$ se relaciona con un cambio en la mortalidad en un 1% tras un día de rezago siendo en los meses de frío los más complicados, con los adultos mayores como aquellos que mayor riesgo corren (Cakmak, Dales & Vidal, 2007). Esta mortalidad es cargada en mayor medida por aquellos individuos de menor nivel educativo, que viven en áreas de menores ingresos y que experimenten de cerca la pobreza (Cakmak, Dales, Angelica Rubio & Blanco Vidal, 2011).

Durante las últimas décadas la digitalización de registros, la expansión de sensores de medición y la computación de la ciudad ha producido una mayor riqueza de datos en las ciudades con mayor frecuencia temporal (Glaeser et al., 2018). Los principales intentos presentes en la literatura por introducir métodos predictivos en base a niveles de contaminación se enfocan en predecir los propios niveles de contaminación ambiental (Perez & Salini, 2008), en lugar de las enfermedades respiratorias, producto de las dificultades presentes en la recopilación y procesamiento de datos (Bellinger et al., 2017; Perez & Gramsch, 2016). Sin embargo, una investigación en la ciudad de Lanzhou en China obtiene buenos resultados empleando técnicas de data mining para predecir atenciones por enfermedades respiratorias (Zhu, Wang, Zhang & Sun, 2012). En Chile, actualmente la Universidad de Concepción trabaja en un proyecto para la región del Bio Bio de nombre “respira bio bio” el cual persigue la misma lógica de entregar elementos predictivos para mejorar la gestión en salud (Remaggi S., Mardones P. & Jiménez R., 2019).

2. Objetivo

Las investigaciones sobre las consecuencias atribuidas a la contaminación atmosférica en la salud han logrado avances significativos en entender como ambas interactúan. Esta investigación pretende entender si es posible, a partir de modelos de Machine Learning, predecir la cantidad de hospitalizaciones a causa de enfermedades respiratorias en comunas de Santiago de Chile utilizando variables de contaminación ambiental y de temperatura obtenidas por medio de una interpolación espaciotemporal. El trabajo continua de la siguiente manera, en la sección 3 se explican los datos y sus orígenes, en las secciones 4 y 5 se mencionan las variables utilizadas y el proceso de interpolación espaciotemporal necesario para utilizarlas. Luego en las secciones 6 y 7 se elige el modelo predictivo y se presentan los resultados respectivamente. Finalmente, en la sección 8 se discute sobre lo investigado.

3. Datos

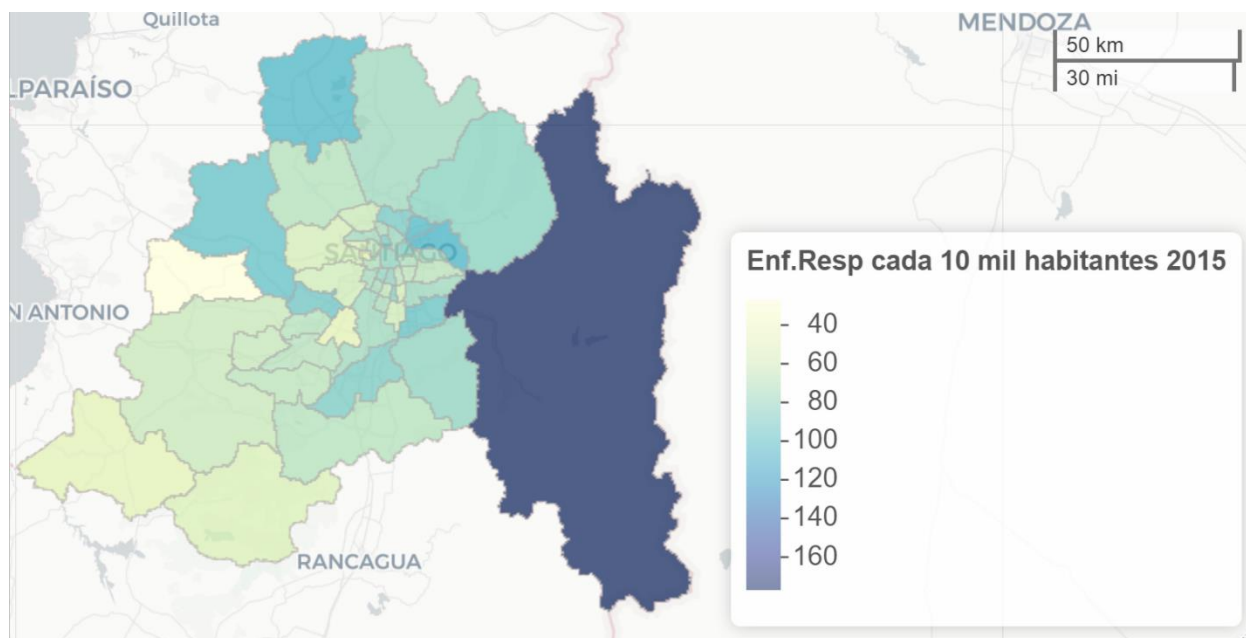
Para los fines de este trabajo los datos utilizados son extraídos desde el SINCA (Sistema de Información Nacional de Calidad del Aire del ministerio del medio ambiente en Chile) del cual se pueden obtener las

siguientes mediciones por hora en distintos estados sobre las emisiones de Dióxido de Azufre, Monóxido de nitrógeno, Dióxido de Nitrógeno, Monóxido de carbono, Ozono, Óxidos de Nitrógeno, MP10 y MP2.5 además de datos de condiciones meteorológicas de humedad, temperatura e incluso viento. Los datos que provienen desde este sistema de monitoreo pueden estar en hasta 11 estaciones de monitoreo, ubicadas en 11 comunas distintas. Los valores pueden presentar registros no validados que corresponden a datos no representativos dado a situaciones como cortes de luz en las estaciones o re calibraciones. En algunas ocasiones estas mediciones reciben posteriormente un proceso de validación con el fin de mejorar información, pasando a ser registros preliminares. Una vez obtenidos los registros preliminares, estos pasan por un segundo proceso de validación, en el cual se verifica si la información es representativa y puede ser utilizada para evaluar las normas de calidad del aire primarias, estos datos tienen el nombre de registros oficiales Para efectos de este trabajo son utilizadas las mediciones de MP2.5, MP10 y Temperatura entre los años 2010 y 2018 (**Ver Anexo 1**).

También son utilizados datos del DEIS (Departamento de Estadísticas e Información de Salud) de donde se obtiene información acerca de los egresos hospitalarios con variables determinantes sobre los pacientes: fecha de egreso del paciente, días de estadía, sexo, edad, previsión de salud, comuna de origen de los pacientes, región, nacionalidad, diagnóstico y hospital donde fue atendido entre los años 2010 y 2018. En la **Figura 1** es posible ver el promedio de enfermedades respiratorias por cada mil habitantes para el año 2015. Los egresos de los pacientes son obtenidos de hospitales públicos y clínicas. Luego de ser codificados en cada establecimiento centralizadamente se realizan procesos masivos de calidad de datos y consistencia de la información. De existir inconsistencia se solicita a cada establecimiento corregir cuando corresponde. Los diagnósticos están ordenados según CIE-10 (Clasificación Internacional de Enfermedades y Problemas Relacionados con la Salud) que constituye uno de los mecanismos más usados para elaborar estadísticas de morbilidad y mortalidad en el mundo. Este es un código alfanumérico que cuenta con una letra en su primera posición y números desde la segunda hasta la cuarta posición (A000-Z000). Para el propósito de este trabajo se utilizaron aquellas enfermedades clasificadas como enfermedades del sistema respiratorio, las cuales tienen la letra J en su primera posición. Esta clasificación a su vez está dividida en otros 10 tipos de enfermedades respiratorias: Infecciones agudas de las vías respiratorias superiores (J00-J06), Influenza y neumonía (J10-J18), Otras infecciones agudas de las vías respiratorias inferiores (J20-J22), Otras enfermedades de las vías respiratorias superiores (J30-J39), Enfermedades crónicas de las vías respiratorias inferiores (J40-J47), Enfermedades del pulmón debida a agentes externos (J60-J70), Otras enfermedades respiratorias que afectan principalmente al intersticio (J80-J84), Afecciones supurativas y necróticas de las vías respiratorias inferiores (J85-J86), Otras enfermedades de la pleura (J90-J94) y Otras enfermedades del sistema respiratorio (J95-J99). Para más detalles sobre cuáles son los diagnósticos especificados en cada grupo de clasificación ver (**Anexo 2**).

Con el fin de poder implementar las estrategias necesarias que requieren los objetivos planteados, se han agregado a las estaciones de monitoreo y a las 52 comunas de origen de los pacientes datos de geolocalización.

Figura 1. Enfermedades respiratorias cada 10 mil habitantes durante el año 2015.



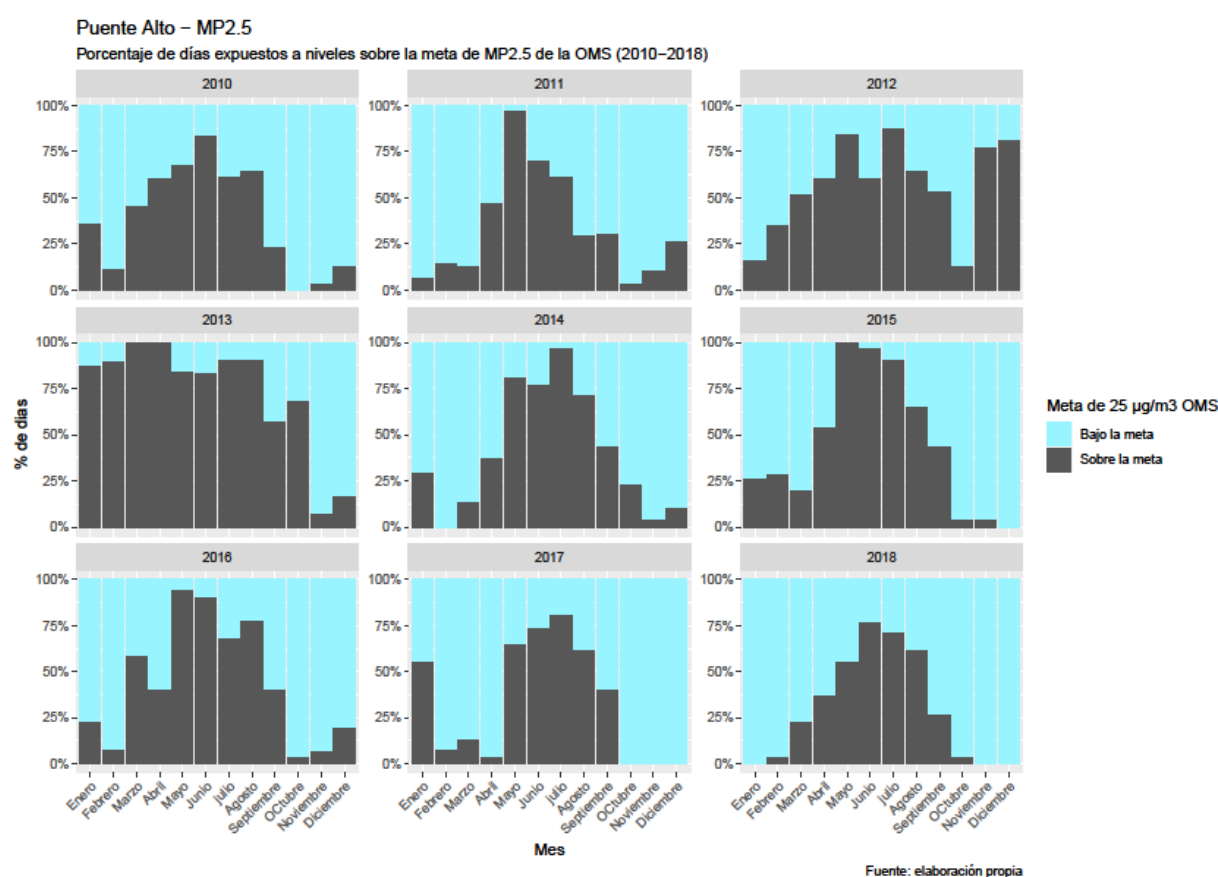
4. Variables utilizadas

Para efectos de este trabajo tanto los modelos como las variables son explorados utilizando la comuna de Puente Alto. La variable a predecir corresponde a la cantidad de ingresos hospitalarios diarios. Dentro de los predictores utilizados se encuentran: tipo de enfermedad según clasificación CIE-10, grupo etario, niveles de MP2.5, niveles de MP10 y grados de temperatura. Tanto los predictores como la cantidad de ingresos diarios utilizados son desde el año 2010 hasta el año 2018.

Las clasificación de enfermedades CIE-10 fueron categorizadas en 10 tipos de enfermedades respiratorias: Infecciones agudas de las vías respiratorias superiores (J00-J06)(5), Influenza y neumonía (J10-J18)(8), Otras infecciones agudas de las vías respiratorias inferiores (J20-J22)(7), Otras enfermedades de las vías respiratorias superiores (J30-J39)(2), Enfermedades crónicas de las vías respiratorias inferiores (J40-J47)(6), Enfermedades del pulmón debida a agentes externos (J60-J70)(3), Otras enfermedades respiratorias que afectan principalmente al intersticio (J80-J84)(10), Afecciones supurativas y necróticas de las vías respiratorias inferiores (J85-J86)(1), Otras enfermedades de la pleura (J90-J94)(4) y Otras enfermedades del sistema respiratorio (J95-J99)(9). Los grupos etarios están divididos entre: Menores de 1 año(1), entre 1 a 4 años(2), entre 5 a 14 años(3), entre 15 y 64 años(4), y mayores de 65 años(5). Las variables medioambientales medidas por nivel de microgramos por metro cubico de MP2.5, microgramos por metro cubico de MP10 y grados centígrados de temperatura son extraídos desde de los sistemas de monitoreo del SINCA. Las estaciones de monitoreo del sistema solo están disponibles en algunas de las 52 comunas de donde provienen los pacientes y para poder lograr el objetivo planteado de determinar los niveles de polución a los que se exponen los pacientes provenientes de cada comuna, se aprovechan los datos geo estadísticos que se obtienen a través del tiempo en las estaciones de monitoreo para realizar una **interpolación espacio temporal** de los datos de contaminación (**ver sección 5**). Por este medio se obtiene una aproximación de los niveles de contaminación diarios para las 52 comunas de la región metropolitana. Además se tomaron en consideración a partir de estos resultados de la interpolación rezagos de hasta 8 días (Matus C. & Oyarzún G., 2019) en las tres variables

medioambientales por lo que cada variable cuenta con nueve mediciones por observación. Por tanto para predecir la cantidad de ingresos diarios por enfermedades respiratorias por comuna se cuenta con 29 predictores, 1 por tipo de enfermedad, 1 por el rango de edad y 27 de niveles de variables medioambientales y sus rezagos en un total de 407.868 observaciones. El promedio de observaciones por comuna es de 7.844 con un máximo de 28.440 en la comuna de Puente Alto (**Ver Anexo 3**). En la misma comuna los habitantes habrían estado expuestos durante más del 50% de los días del mes a niveles diarios por sobre lo recomendado por la OMS de 25 $\mu\text{g}/\text{m}^3$ de MP2.5 durante los meses de mayo, junio y julio entre los años 2010 y 2018, esto es posible de observar en la **Figura 2**. Aquellos que más han cargado con enfermedades respiratorias durante los meses de invierno en la región metropolitana son los habitantes menores de 1 año (**Figura 3**), entre 1 y 4 años (**Figura 4**) y aquellos mayores de 65 años (**Figura 5**).

Figura 2. Porcentaje de días expuesto a niveles sobre la meta de MP2.5 en Puente Alto



La **Figura 6** demuestra como en la comuna de Puente Alto, el promedio diario de ingresos por enfermedades respiratorias presentan una estacionalidad similar al promedio de contaminación diaria por mes entre 2010 y 2018. Al mismo tiempo, en la **Figura 7** se pueden ver los ingresos por enfermedades respiratorias tienen una correlación negativa con la temperatura y positiva con los niveles de MP2.5 para los habitantes entre 1 y 4 años.

Figura 3. Ingresos Hospitalarios por enfermedades respiratorias en la RM según rango de edad entre 2010-2018

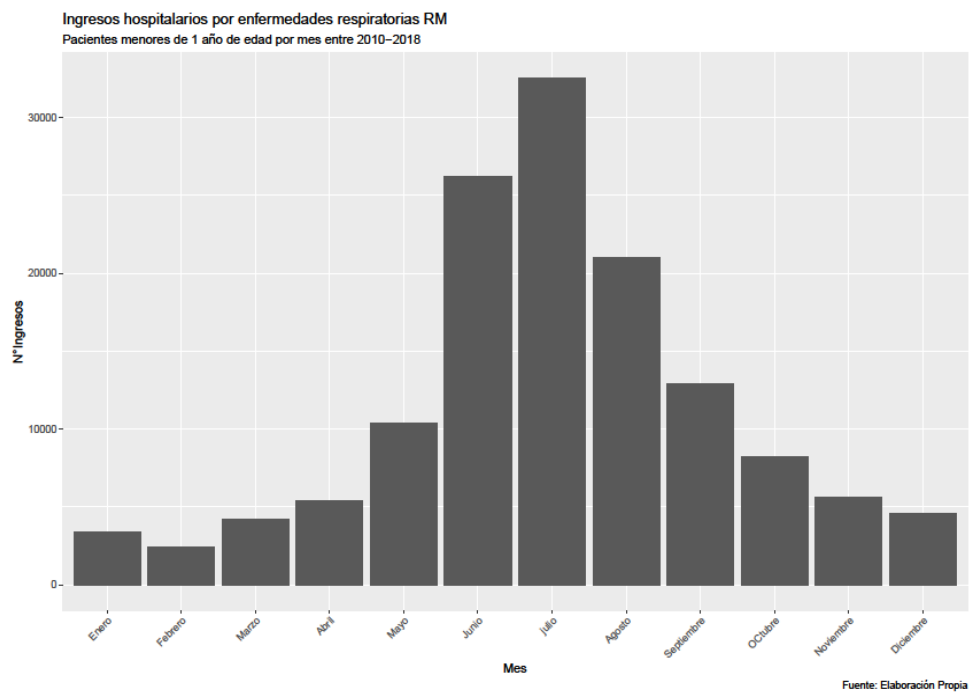


Figura 4. Ingresos Hospitalarios por enfermedades respiratorias en la RM según rango de edad entre 2010-2018

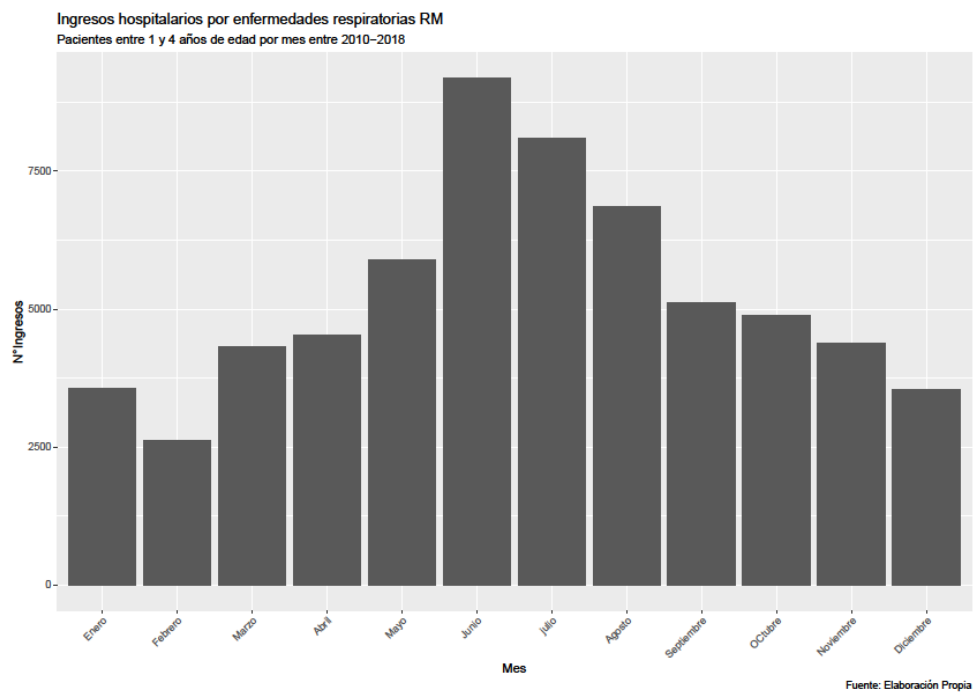


Figura 5. Ingresos Hospitalarios por enfermedades respiratorias en la RM según rango de edad entre 2010-2018

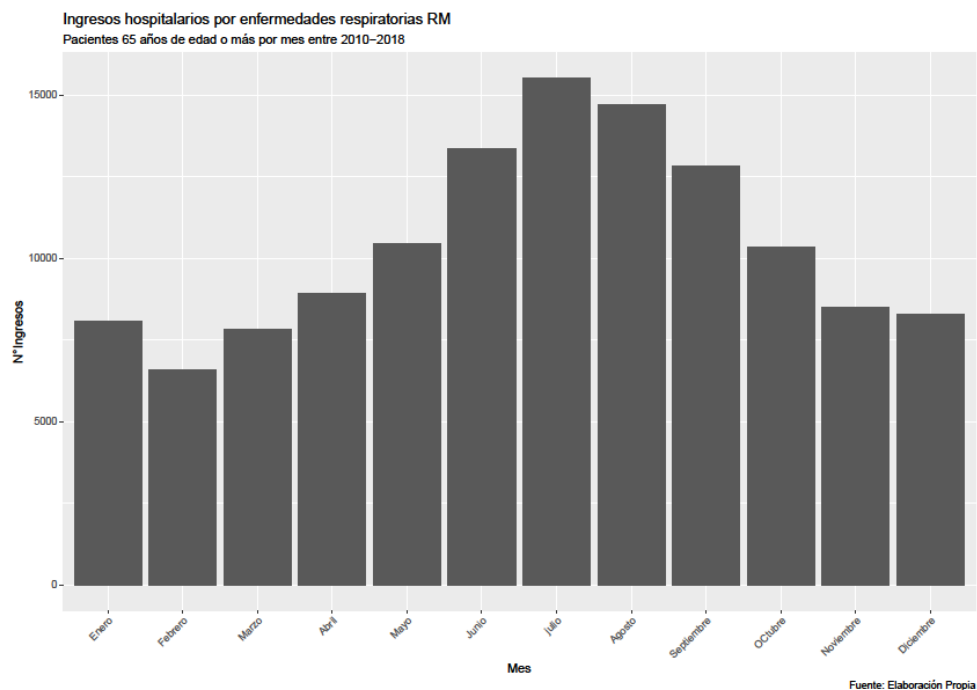


Figura 6. Promedio diario de ingresos por enfermedades respiratorias y mediciones medioambientales por mes entre 2010-2018

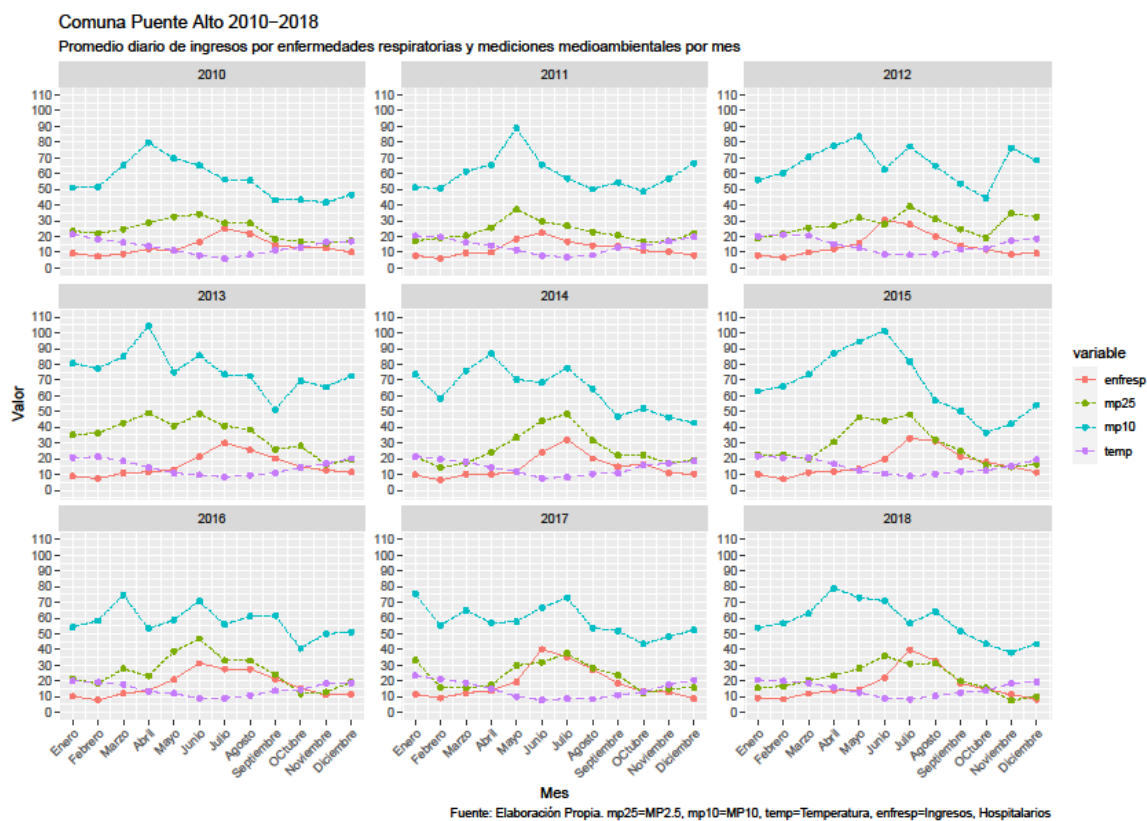
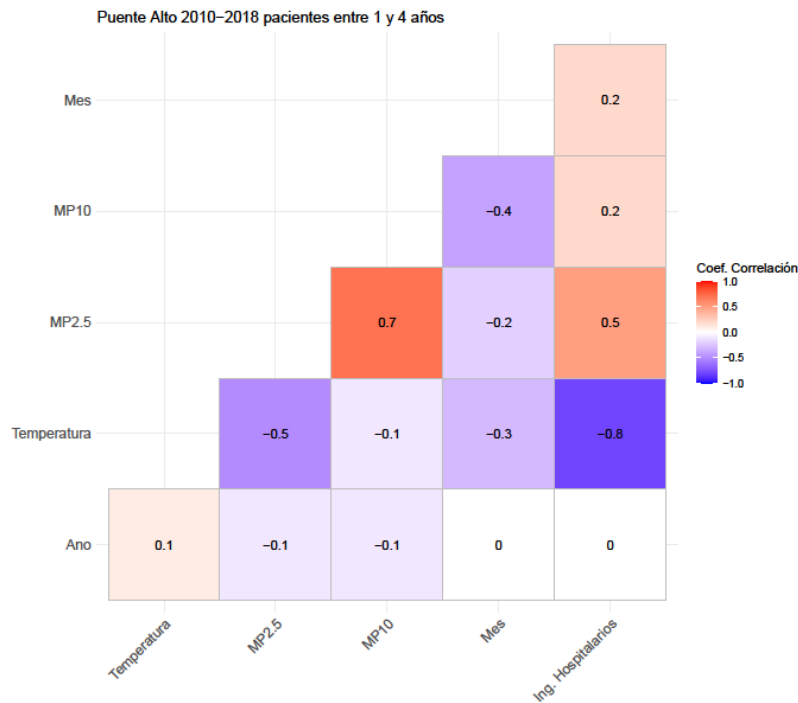


Figura 7. Correlación entre los ingresos por enfermedades respiratorias y las variables medioambientales en pacientes entre 1 y 4 años.



5. Interpolación espaciotemporal

Lo que se busca en esta sección es conocer la relación entre las condiciones medioambientales en diferentes puntos de Santiago y cómo podemos aprovechar esa relación para hacer predicciones en lugares en los cuales no tenemos una muestra sobre las condiciones exploradas. Los ejercicios de interpolación espacial para puntos en un campo son comunes en la geo estadística. Usualmente la metodología de kriging ha sido utilizada para hacer frente a problemas en minería, combustibles fósiles, contaminación y salud pública (Singh et al., 2011). Kriging da una solución a la estimación espacial en base a un modelo continuo de variación espacial estocástica tomando en cuenta que un factor puede variar en el espacio a través de un modelo de variograma. Este método consiste básicamente hacer estimaciones a partir de combinaciones lineales ponderadas de datos. Las ponderaciones son asignadas a datos dentro del entorno del punto que se quiere estimar de tal manera que se minimice la varianza del kriging y los estimadores sean insesgados. El método de kriging tiene distintos tipos de interpolación dentro de los cuales se distingue algunos como Ordinary Kriging y Ordinary Co-Kriging, entre otros. (Webster & Oliver, 2009).

Hoy en día, los sensores permiten tener bases de datos espaciotemporales lo que abre la oportunidad de ampliar la teoría en cuanto a los procedimientos para realizar interpolación (Gräler, et al., 2013). Para esto la interpolación espacio temporal permite tener mejor resultados al utilizar distintas mediciones a lo largo del tiempo para realizar la interpolación, adicionalmente permite hacer predicciones en ubicaciones o campos enteros entre o después de los momentos de observación (Gräler, Pebesma & Heuvelink, 2016).

En la interpolación espacio temporal como plantean Gräler, Pebesma & Heuvelink (2016) puede entenderse asumiendo en un comienzo un campo aleatorio de espacio tiempo Z definido en un dominio espacial S y un dominio temporal T donde se observa una muestra $z = (z(s_1, t_1), \dots, z(s_n, t_n))$ en un set de ubicaciones espacio temporales $(s_1, t_1), \dots, (s_n, t_n) \in S \times T \subseteq \mathbb{R}^2 \times \mathbb{R}$ que puede incluir mediciones repetidas en las mismas ubicaciones o en múltiples ubicaciones. Lo que se busca es modelar Z a partir de z con el fin de predecir muestras en ubicaciones no observadas. El campo Z se puede caracterizar por medio de un promedio μ y una función de covarianza espacio temporal C_{st} (Pebesma & Gräler, 2020) donde esta depende solamente de la distancia de separación en el espacio $h \in \mathbb{R}_{\geq 0}$ y el tiempo $u \in \mathbb{R}_{\geq 0}$. La covarianza está dada por $C_{st}(h, u) = Cov(Z(s, t), Z(\tilde{s}, \tilde{t}))$ para distancia espacial h , distancia temporal u y cualquiera par de puntos $(s, t), (\tilde{s}, \tilde{t}) \in S \times T$ con $h = \|s - \tilde{s}\|$ y $\|t - \tilde{t}\| = u$. Esta covarianza es difícil de calcular, sin embargo el paquete **gstat** del software estadístico **R** presenta los medios para lograr con sus respectivos variogramas espacio-temporales $\gamma_{st}(h, u) = C_{st}(0, 0) - C_{st}(h, u)$ y es la herramienta que para los propósitos de este trabajo se utiliza para lograr la interpolación. Algunos de los variogramas presentes son los modelos *separable*, *product-sum*, *metric* y *sum-metric* y se basan en el variograma muestral (*sample*) obtenido a partir de la muestra (**Ver Figura 8 y Anexo 4**), el cual es calculado a través de la función **variogramST()** del paquete **spacetime** del mismo software R (Pebesma, 2012). Luego de estimar el modelo muestral, se ajustan a través de la función **fit.StVariogram()** para posteriormente identificar el modelo que se ajuste mejor a la muestra y realizar la interpolación a través de la función **krigeST()**. Para efectos de este trabajo se ha utilizado el modelo *sum-metric* que obtuvo para MP2.5, MP10 y Temperatura la menor raíz del error cuadrático medio (RMSE) para lograr la grilla con las predicciones. Los variogramas realizados para MP2.5 y los resultados de la interpolación pueden ser vistos en la **figuras 8 y 9** respectivamente.

Para este caso se cuenta con información tanto espacial como temporal. Sin embargo, se presenta información en un espacio temporal amplio pero en pocos puntos lo que significa conclusiones menos acertadas respecto a la relaciones espacio temporales

6. Selección de metodología de predicción

La gran acumulación de datos gracias al aumento exponencial de información generada a través del tiempo y espacio nos invita a explorar herramientas más poderosas en la manipulación de estos. Este fenómeno a su vez permite que posiblemente nos encontremos con una mayor cantidad de potenciales predictores que los apropiados para una estimación y que por tanto se necesiten formas de selección de variables. Por otro lado, estas largas bases de datos pueden permitir estudiar relaciones más flexibles que los modelos lineales y las técnicas de machine learning como los arboles de decisión, máquinas de vectores de soporte, redes neurales, deep learning, entre otros pueden ayudar a modelar relaciones más complejas entre los datos. En el sentido más práctico de un problema estadístico de predicción, lo que estamos interesados en entender es la distribución condicional de alguna variable y dado otras variables $x = (x_1, \dots, x_p)$. En machine learning, las variables x son conocidas como predictores y lo que se busca es un modelo que permita encontrar una “buena” predicción de y dado nuevos valores de x . Una buena predicción se caracteriza por que minimiza alguna función de pérdida como la raíz del error cuadrático medio (RMSE) asociada a las nuevas observaciones fuera de muestra de x . (Varian, 2014).

Figura 8. Variogramas de muestra y modelos de para interpolación espaciotemporal de MP2.5

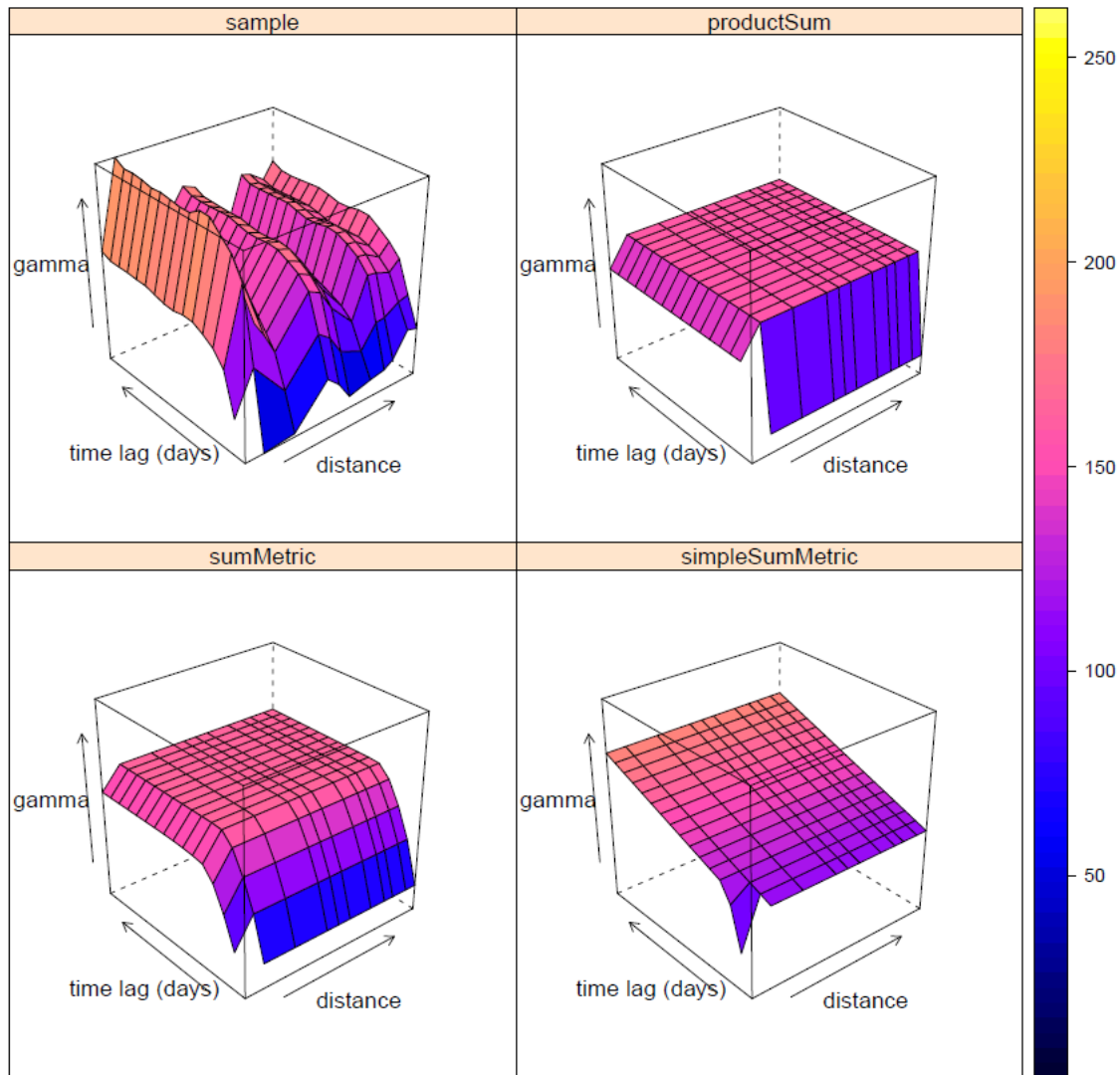
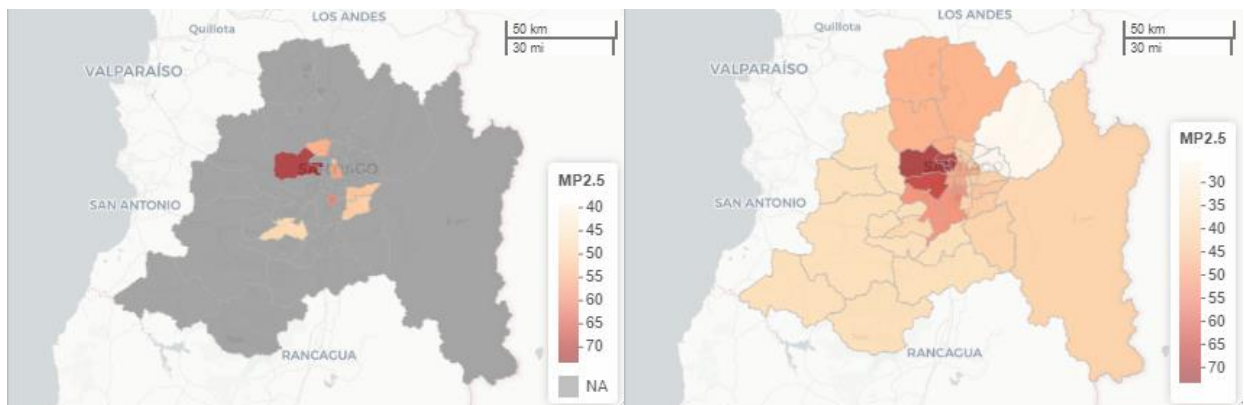


Figura 9. Interpolación para MP2.5: 04 de mayo de 2012 (antes y después)



Dentro de las técnicas de Machine Learning, los algoritmos de “Classification and Regression Trees” (CART) (Breiman et al., 1984) y “Random Forest” (RF) (Breiman, 2001a) se han vuelto métodos bastantes populares y efectivos para estimar de manera flexible funciones de regresión en entornos donde el poder predictivo fuera de la muestra es importante (Athey & Imbens, 2019). Además de esto, el modelo de árboles de decisión ha sido reconocido como dentro de los top 10 algoritmos en data mining. Por su lado Random Forest es vista como una las técnicas de aprendizaje estadístico más certera disponible, y ha sido utilizada en diferentes campos de investigación como la física y la genética (David, 2017) y no requiere de muchas habilidades estadísticas e incluso es bastante familiar para economistas que están acostumbrados a representaciones de árbol (Varian, 2014). Por tanto y tomando en consideración lo anterior para el proposito de esta investigación se utilizarán las tecnicas de CART y RF.

6.1 CART

Árboles de clasificación y de regresión. Estos métodos por lo general pueden ser usados para clasificaciones con múltiples resultados (“árboles de clasificación”) o con variables continuas dependientes (“árboles de regresión”) (Varian, 2014). El motivo de esta investigación representa un planteamiento para arboles de regresión. La intuición básica de este modelo consiste en dividir los espacios del predictor en diferentes sub espacios y aproximar la respuesta que se busca predecir por su media empírica en cada sub espacio. El espacio inicial de los arboles es conocido como “*nodo raíz*”. Los otros espacios se conocen como nodos (t) o “*hojas*” y cualquier otra parte del árbol de decisión puede llamarse “*rama*” (T_t) (Ver figura 7). La función es posible representarla como una suma de constantes, donde $\hat{y}(x)$ es el valor predicho o ajustado, D_{t_*} corresponde a una variable dummy que toma el valor de 1 si el nodo terminal t_* es considerado, $\hat{a}_{t_*}$ es la respuesta predicha en ese nodo y N_t corresponde al número de casos por casa nodo (David, 2017):

$$\hat{y}(x) = \sum_{t_*} D_{t_*} \hat{a}_{t_*}$$

$$D_{t_*} = \begin{cases} 1, & \text{if } x \in t_* < 0 \\ 0, & \text{Caso contrario} \end{cases}$$

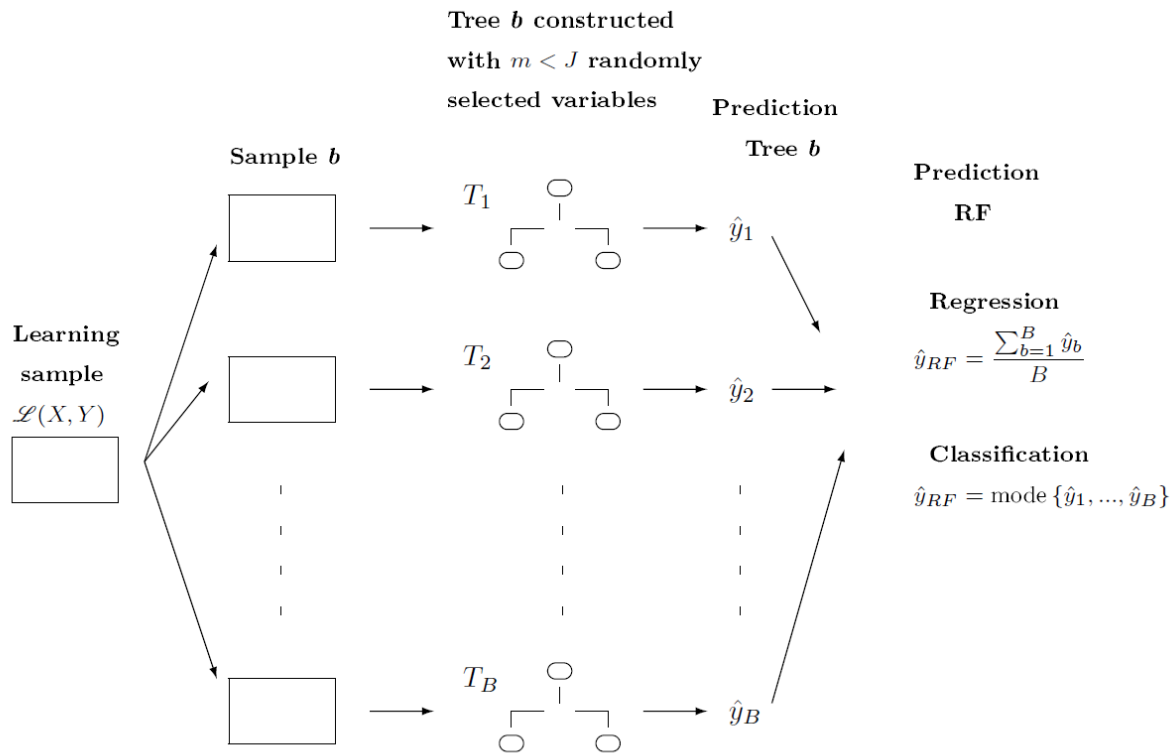
$$\hat{a}_t = \bar{y}_t = \frac{1}{N_t} \sum y_n$$

El modelo CART permite que todas las variables incluidas sean tomadas en cuenta como potenciales variables de interacciones. En cada nodo evalúa la posibilidad de cada variable de ser una variable divisora y prueba todas las posibles interacciones para construir el árbol de decisiones.

Un problema de los árboles de decisión es que tienden a sobre ajustar (“overfit”) los datos. Un árbol con muchas ramas puede lograr ajustar a los datos de entrenamiento de forma precisa, sin embargo las predicciones hechas a partir del set de testeo, es decir, fuera de la muestra pueden ser muy deficientes causando inestabilidad en el modelo (David, 2017).

6.2 Random Forest

“Random Forest” (RF) puede resolver la posible inestabilidad de los modelos CART y tiene considerablemente mejor poder predictivo que los árboles de decisiones (Athey & Imbens, 2019). Esta metodología consiste en usar múltiples árboles. En primer lugar elige una muestra “bootstrap” de las observaciones para comenzar a crear un árbol. Luego, a cada nodo del árbol, se elige una muestra aleatoria de predictores para tomar la siguiente decisión y así sucesivamente hasta crear un bosque de árboles (Varian, 2014). Para poder determinar la clasificación de la nueva observación es considerada cada variable que son tomadas por cada árbol para ser clasificadas. Luego, siendo un caso de regresión, los valores predichos son las respuestas promedio de todos los árboles. La función es representada por David (2017) de la siguiente manera:



Fuente: David, 2017

7. Resultados

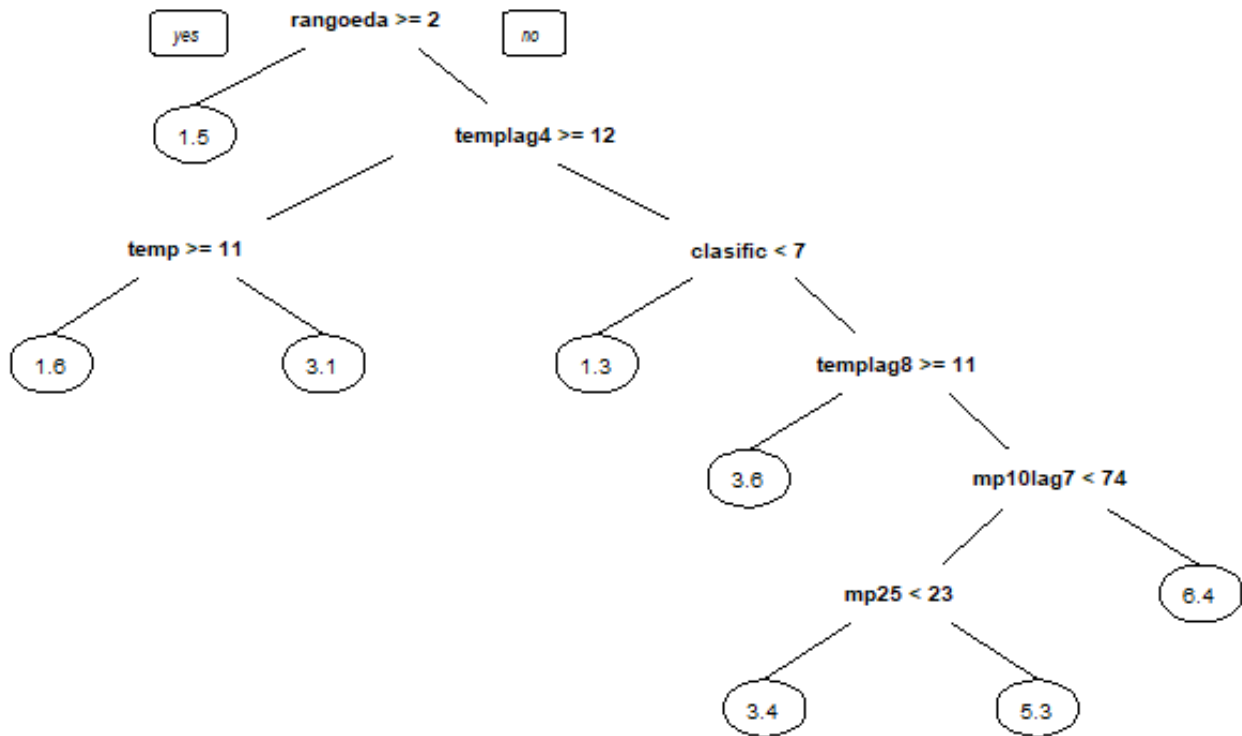
En la siguiente sección se utilizan los modelos CART y Random Forest para obtener los resultados de las predicciones de esta investigación.

7.1 Resultados CART

El modelo de árboles de decisión es implementado por el medio del el paquete estadístico de **R** llamado **rpart**. Para mostrar los resultados se toma en consideración la comuna de Puente Alto donde el promedio

de ingresos hospitalarios por día según tipo de enfermedad y rango de edad es de 1.77 con un máximo de 24 ingresos (**Ver anexo 3**).

Figura 7 Árbol de decisiones: Comuna Puente Alto



Para el modelo de árboles de decisión de la **figura 7** se observa un RMSE de 1.34, o sea falla en promedio por 1.34 en la predicción de ingresos hospitalarios diarios producto de enfermedades respiratorias para la comuna de Puente Alto.

7.2 Resultados Random Forest

El modelo de árboles de decisión es implementado por medio del paquete estadístico de **R** llamado **randomForest()** (Liaw & Matthew, 2001). Para mostrar los resultados se toma en consideración la comuna de Puente Alto donde el promedio de ingresos hospitalarios por día según tipo de enfermedad y rango de edad es de 1.77 con un máximo de 24 ingresos.

```

Call:
  randomForest(formula = cantidadenf ~ mp25 + mp10 + temp + mp25lag1 +      mp10lag1 + templag1 +
    mp25lag2 + mp10lag2 + templag2 + mp25lag3 +      mp10lag3 + templag3 + mp25lag4 + mp10lag4 + temp
    lag4 + mp25lag5 +      mp10lag5 + templag5 + mp25lag6 + mp10lag6 + templag6 + mp25lag7 +      mp1
    0lag7 + templag7 + mp25lag8 + mp10lag8 + templag8 + rangoedad +      clasificacion, data = egreso
    s.rango7)
  Type of random forest: regression
    Number of trees: 500
No. of variables tried at each split: 9

  Mean of squared residuals: 1.582205
    % Var explained: 40.55
  
```

	Overall
mp25	1462.314
mp10	1458.833
temp	2186.849
mp25lag1	1428.877
mp10lag1	1407.633
temp1ag1	1930.333
mp25lag2	1294.571
mp10lag2	1267.758
temp1ag2	2178.456
mp25lag3	1599.745
mp10lag3	1319.357
temp1ag3	2355.337
mp25lag4	1463.202
mp10lag4	1367.988
temp1ag4	2176.190
mp25lag5	1505.283
mp10lag5	1408.386
temp1ag5	2415.461
mp25lag6	1453.039
mp10lag6	1276.368
temp1ag6	2086.509
mp25lag7	1464.964
mp10lag7	1489.451
temp1ag7	2407.891
mp25lag8	1746.751
mp10lag8	1460.302
temp1ag8	2285.329
rangoedad	13350.198
clasificacion	8558.852

Para el modelo de RF se observa que el promedio al cuadrado de los residuos es de 1.582, o sea se falla en promedio por 1.582 en la predicción de ingresos hospitalarios diarios producto de enfermedades respiratorias para la comuna de Puente Alto. El porcentaje de varianza explicado es de 40.55%. En este modelo, es posible apreciar la importancia que tomaron las variables. El rango de edad y la clasificación del tipo de enfermedad cobraron mayor importancia para la predicción seguidas por las temperaturas con un rezago de 5 y 7 días.

8. Discusión

Aunque los modelos predictivos no necesariamente permiten concluir algo sobre la causalidad en sí misma de la contaminación sobre las enfermedades respiratorias, estos modelos pueden ayudar a estimar el impacto de una intervención que reduzca los niveles de contaminación. Tomando en consideración la premisa de que las políticas que buscan prevenir el cambio climático podrían ser aún más efectivas si toman en consideración sus beneficios a la salud (Markandya et al., 2018) y bajo lo expuesto en esta investigación, con big data se podría mejorar el manejo de las ciudades sobre sus niveles de contaminación ambiental pensando en la calidad de vida de las personas. Al establecer operaciones guiadas por los datos, la ciudades pueden ir ajustando regulaciones, mejorar la focalización de recursos escasos, predecir necesidades futuras y por medio de los esfuerzos de digitalización atraer emprendimiento e innovación (Glaeser et al., 2018). Es este sentido que los economistas pueden jugar un rol crucial en resolver políticas que involucren problemas de predicción aumentando las nuevas preguntas que se pueden hacer (Mullainathan & Spiess, 2017). Sin embargo, quedan aún desafíos para poder seguir explorando estos mecanismos, en este caso en particular con mayores estaciones de medición, mejoras en las mediciones de otras variables medioambientales y en el gobierno de datos de salud con el fin de poder obtener mejores datos y así ir mejorando la precisión de los modelos.

Anexo 1

Detalle de los datos extraídos desde las estaciones de monitoreo del SINCA

Estación	Contaminante	Fecha datos extraídos	Intervalo	Datos validados	Datos Ausentes	% DA	Datos Preliminares	% DP	Datos No Validados	% DNV	Datos Totales
Parque Ohiggins	MP10	Enero 2010-diciembre 2019	Hora	79661	929	1,1%	6570	7,5%	488	0,6%	87648
Quilicura	MP10	Enero 2010-diciembre 2019	Hora	78305	1263	1,4%	7792	8,9%	288	0,3%	87648
Puente Alto	MP10	Enero 2010-diciembre 2019	Hora	79849	580	0,7%	6993	8,0%	226	0,3%	87648
Pudahuel	MP10	Enero 2010-diciembre 2019	Hora	79321	1235	1,4%	6744	7,7%	348	0,4%	87648
La Florida	MP10	Enero 2010-diciembre 2019	Hora	78806	1707	1,9%	6779	7,7%	356	0,4%	87648
Independencia	MP10	Enero 2010-diciembre 2019	Hora	79476	1516	1,7%	6314	7,2%	342	0,4%	87648
El Bosque	MP10	Enero 2010-diciembre 2019	Hora	79093	1314	1,5%	6702	7,6%	539	0,6%	87648
Cerro Navia	MP10	Enero 2010-diciembre 2019	Hora	78835	1456	1,7%	6584	7,5%	773	0,9%	87648
Cerrillos	MP10	Enero 2010-diciembre 2019	Hora	75717	4733	5,4%	6342	7,2%	856	1,0%	87648
Las Condes	MP10	Enero 2010-diciembre 2019	Hora	84797	2051	2,3%	297	0,3%	503	0,6%	87648
Parque Ohiggins	MP2.5	Enero 2010-diciembre 2019	Hora	78936	1251	1,4%	6623	7,6%	838	1,0%	87648
Quilicura	MP2.5	Enero 2010-diciembre 2019	Hora	78421	2346	2,7%	6473	7,4%	408	0,5%	87648
Puente Alto	MP2.5	Enero 2010-diciembre 2019	Hora	79158	1296	1,5%	6580	7,5%	614	0,7%	87648
Pudahuel	MP2.5	Enero 2010-diciembre 2019	Hora	76627	1805	2,1%	8456	9,6%	760	0,9%	87648
La Florida	MP2.5	Enero 2010-diciembre 2019	Hora	78831	2174	2,5%	6077	6,9%	566	0,6%	87648
Independencia	MP2.5	Enero 2010-diciembre 2019	Hora	78357	2074	2,4%	6367	7,3%	850	1,0%	87648
El Bosque	MP2.5	Enero 2010-diciembre 2019	Hora	78823	1808	2,1%	6333	7,2%	684	0,8%	87648
Cerro Navia	MP2.5	Enero 2010-diciembre 2019	Hora	78995	1046	1,2%	6568	7,5%	1039	1,2%	87648
Cerrillos	MP2.5	Enero 2010-diciembre 2019	Hora	65427	7837	8,9%	14149	16,1%	235	0,3%	87648
Las Condes	MP2.5	Enero 2010-diciembre 2019	Hora	84980	1857	2,1%	95	0,1%	716	0,8%	87648
Talagante	MP2.5	Enero 2010-diciembre 2019	Hora	76125	3535	4,0%	15	0,0%	7973	9,1%	87648
Parque Ohiggins	Temperatura	Enero 2010-diciembre 2019	Hora	86516	1132	1,3%	0	0,0%	0	0,0%	87648
Puente Alto	Temperatura	Enero 2010-diciembre 2019	Hora	86700	948	1,1%	0	0,0%	0	0,0%	87648
La Florida	Temperatura	Enero 2010-diciembre 2019	Hora	76643	11005	12,6%	0	0,0%	0	0,0%	87648
Independencia	Temperatura	Enero 2010-diciembre 2019	Hora	84617	3031	3,5%	0	0,0%	0	0,0%	87648
El Bosque	Temperatura	Enero 2010-diciembre 2019	Hora	86459	1189	1,4%	0	0,0%	0	0,0%	87648
Cerro Navia	Temperatura	Enero 2010-diciembre 2019	Hora	86730	918	1,0%	0	0,0%	0	0,0%	87648
Talagante	Temperatura	Enero 2010-diciembre 2019	Hora	84980	1443	1,7%	0	0,1%	0	0,0%	87648

Anexo 2

Tipo Enfermedad	Clasificación CIE	Nombre Diagnóstico
Afecciones supurativas y necróticas de las vías respiratorias inferiores	J85-J86	AbscesoDelMediastino
		AbscesoDelPulmonConNeumonia
		AbscesoDelPulmonSinNeumonia
		GangrenaYNecrosisDelPulmon
		PitoraxConFistula
Enfermedades crónicas de las vías respiratorias inferiores	J40-J47	PitoraxSinFistula
		Asma,NoEspecificado
		AsmaMixta
		AsmaNoAlergica
		AsmaPredominantementeAlergica
		Bronquiectasia
		Bronquitis,NoEspecificadaComoAgudaOCronica
		BronquitisCronicaMixtaSimpleYMucopurulenta
		BronquitisCronicaMucopurulenta
		BronquitisCronicaNoEspecificada
		BronquitisCronicaSimple
		EnfermedadPulmonarObstructivaCronica,NoEspecificada
		EnfermedadPulmonarObstructivaCronicaC/Infecc.Ag.DeViasRespiratoriasInferiores
		EnfermedadPulmonarObstructivaCronicaConExacerbacionAguda,NoEspecificada
		Enfisema,NoEspecificado
		EstadoAsmatico
		OtrasEnfermedadesPulmonaresObstructivasCronicasEspecificadas
Enfermedades del pulmón debida a agentes externos	J60-J70	Afecc.Respirat.NoEspecific.,Deb.AlnhalacionDeGases,Humos,VaporesYSustanciasQuimicas
		AfeccionesRespirat.CronicasDeb.Alnhalac.DeGases,Humos,VaporesYSustanciasQuimicas
		AfeccionesRespiratoriasDebidasAAgentesExternosNoEspecificados
		AfeccionesRespiratoriasDebidasAOtrosAgentesExternosEspecificados
		Aluminosis(DelPulmon)
		Bagazosis
		Beriliosis
		Bisinosis
		BronquitisYNeumonitisDebidasAlnhalacionDeGases,Humos,VaporesYSustanciasQuimicas
		EdemaPulmonarAgudoDebidoAlnhalacionDeGases,Humos,VaporesYSustanciasQuimicas
		EnfermedadDeLasViasAereasDebidaAOtrosPolvosOrganicosEspecificos
		Fibrosis(DelPulmon)DebidaABauxita
		Fibrosis(DelPulmon)DebidaAGrafito
		Inflamac.Respirat.Sup.Deb.Alnhalac.Gases,Humos,VaporesYSust.Quimicas,Ncop
		Neumoconiosis,NoEspecificada
		NeumoconiosisDebidaAlAsbestoYAOTrasFibrasMinerales
		NeumoconiosisDebidaAOtrosPolvosInorganicosEspecificados
		NeumoconiosisDebidaAOtrosPolvosQueContienenSilice
		NeumoconiosisDebidaAPolvoDeTalco(Silice)
		NeumoconiosisDeLosMinerosDelCarbon
		NeumonitisDebidaAAspiracionDeAceitesYESencias
		NeumonitisDebidaAAspiracionDeAlimentoOVomito
		NeumonitisDebidaAAspiracionDeOtrosSolidosYLIquidos
		NeumonitisDebidaAHipersensibilidadAPolvoOrganicoNoEspecificado
		NeumonitisDebidasAHipersensibilidadAOtrosPolvosOrganicos
		NeumonitisDeLaVentilacionDebidaAlAcondicionadorYHumidificadorDelAire
		OtrasAfecc.Respirat.Ag.YSubag.Deb.Alnhalac.DeGases,Humos,VaporesYSust.Quimicas
		OtrasAfecc.Respirat.Deb.AlnhalacionDeGases,Humos,VaporesYSustanciasQuimicas
		PulmonDelGranjero
		PulmonDelOmitofilo
		Siderosis

Tipo Enfermedad	Clasificacion CIE	Nombre Diagnostico
Infecciones agudas de las vías respiratorias superiores	J00-J06	Amigdalitis Aguda, No Especificada
		Amigdalitis Aguda Debida A Otros Microorganismos Especificados
		Epiglotitis Aguda
		Faringitis Aguda, No Especificada
		Faringitis Aguda Debida A Otros Microorganismos Especificados
		Faringitis Estreptococica
		Infeccion Aguda De Las Vias Respiratorias Superiores, No Especificada
		Insuficiencia Respiratoria, No Especificada
		Insuficiencia Respiratoria Aguda
		Insuficiencia Respiratoria Cronica
		Laringitis Aguda
		Laringitis Cronica
		Laringitis Obstructiva, Aguda (Crup)
		Laringotraqueitis Aguda
		Laringotraqueitis Cronica
		Otras Infecciones Agudas De Sitios Múltiples De Las Vias Respiratorias Superiores
		Pansinusitis Aguda
		Rinofaringitis Aguda (Resfriado Común)
		Rinofaringitis Cronica
		Sinusitis Aguda, No Especificada
		Sinusitis Esfenoidal Aguda
		Sinusitis Etmoidal Aguda
		Sinusitis Frontal Aguda
		Sinusitis Maxilar Aguda
		Traqueitis Aguda
Influenza (gripe) y neumonía	J10-J18	Trastorno Respiratorio, No Especificado
		Bronconeumonía, No Especificada
		Bronquitis Aguda Debida A Haemophilus Influenzae
		Bronquitis Aguda Debida A Mycoplasma Pneumoniae
		Bronquitis Aguda Debida A Virus Parainfluenza
		Influenza, Con Otras Manifestaciones, Debida A Virus De La Influenza Identificado
		Influenza Aviar (Gripe Aviar), Influenza A (H1N1)
		Influenza Con Neumonía, Debida A Virus De La Influenza Identificado
		Influenza Con Neumonía, Virus No Identificado
		Influenza Con Otras Manifestaciones, Virus No Identificado
		Influenza Con Otras Manifestaciones Respiratorias, Debida A Virus De La Influenza Identificado
		Influenza Con Otras Manifestaciones Respiratorias, Virus No Identificado
		Influenza Debida A Ciertos Virus De La Influenza Identificados
		Neumonía, No Especificada
		Neumonía Bacteriana, No Especificada
		Neumonía Debida A Adenovirus
		Neumonía Debida A Chlamidias
		Neumonía Debida A Escherichia Coli
		Neumonía Debida A Estafilococos
		Neumonía Debida A Streptococos Del Grupo B
		Neumonía Debida A Haemophilus Influenzae
		Neumonía Debida A Klebsiella Pneumoniae
		Neumonía Debida A Metaneumovirus Humano
		Neumonía Debida A Metapneumovirus Humano
		Neumonía Debida A Mycoplasma Pneumoniae
		Neumonía Debida A Otras Bacterias Aerobias Gramnegativas
		Neumonía Debida A Otros Streptococos
		Neumonía Debida A Otros Microorganismos Infecciosos Especificados
		Neumonía Debida A Otros Virus
		Neumonía Debida A Pseudomonas
		Neumonía Debida A Streptococcus Pneumoniae
		Neumonía Debida A Virus Parainfluenza
		Neumonía Debida A Virus Sincicial Respiratorio
		Neumonía Hipostática, No Especificada
		Neumonía Lobar, No Especificada
		Neumonía Viral, No Especificada (Virus)
		Otras Neumonías, De Microorganismo No Especificado
		Otras Neumonías Bacterianas

Tipo Enfermedad	Clasificacion CIE	Nombre Diagnostico
Otras enfermedades de la pleura	J90-J94	AfeccionPleural,NoEspecificada DerramePleuralNoClasificadoEnOtraParte Fibrotorax Hemotorax Neumotorax,NoEspecificado NeumotoraxEspontaneoAPresion OtrasAfeccionesEspecificadasDeLaPleura OtrosNeumotorax OtrosTiposDeNeumotoraxEspontaneo PaquipleuritisConAsbestosis PaquipleuritisSinAsbestosis
Otras enfermedades de las vías respiratorias superiores	J30-J39	Absceso,FurunculoYCarbuncoDeLaNariz AbscesoPeriamigdalino AbscesoRetrofaringeoYParafaringeo AmigdalitisCronica AmigdalitisEstreptococica DegeneracionPolipoideDeSenoParanasal DesviacionDeITabiqueNasal EdemaDeLaringe EnfermedadCronicaDeLasAmigdalasYDeLasAdenoides,NoEspecificada EnfermedadDeLasViasRespiratoriasSuperiores,NoEspecificada EnfermedadesDeLaTraqueaYDeLosBronquios,NoClasificadasEnOtraParte Estenosis EstenosisLaringea EstenosisSubgloticaConsecutivaAProcedimientos FaringitisCronica HipertrofiaDeLasAdenoides HipertrofiaDeLasAmigdalas HipertrofiaDeLasAmigdalasConHipertrofiaDeLasAdenoides HipertrofiaDeLosCornetesNasales NodulosDeLasCuerdasVocales OtrasEnfermedadesCronicasDeLasAmigdalasYDeLasAdenoides OtrasEnfermedadesDeLaFaringe OtrasEnfermedadesDeLaLaringe OtrasEnfermedadesDeLasCuerdasVocales OtrasEnfermedadesEspecificadasDeLasViasRespiratoriasSuperiores OtrasRinitisAlergicas OtrasSinusitisAgudas OtrasSinusitisCronicas OtrosAbscesosDeLaFaringe OtrosPoliposDeLosSenosParanasales OtrosTiposDeEnfisema OtrosTrastornosEspecificadosDeLaNarizYDeLosSenosParanasales PansinusitisCronica ParalisisDeLasCuerdasVocalesYDeLaLaringe PolipoDeLaCavidadNasal PolipoDeLasCuerdasVocalesYDeLaLaringe PolipoNasal,NoEspecificado QuisteYMucoceleDeSenoParanasal RinitisAlergica,NoEspecificada RinitisAlergicaDebidaAlPolen RinitisCronica RinitisVasomotora SinusitisCronica,NoEspecificada SinusitisEsfenoidalCronica SinusitisEtmoidalCronica SinusitisFrontalCronica SinusitisMaxilarCronica

Tipo Enfermedad	Clasificación CIE	Nombre Diagnóstico
Otras enfermedades del sistema respiratorio	J95-J99	Colapso Pulmonar
		Enfermedades Del Mediastino, No Clasificadas En Otra Parte
		Enfisema Centrolobular
		Enfisema Panlobular
		Espasmo Laringeo
		Funcionamiento Defectuoso De La Traqueostomía
		Insuficiencia Pulmonar Aguda Consecutiva A Cirugía Extratorácica
		Insuficiencia Pulmonar Aguda Consecutiva A Cirugía Torácica
		Insuficiencia Pulmonar Crónica Consecutiva A Cirugía
		Laringofaringitis Aguda
		Manifestaciones Pulmonares Agudas Debidas A Radiación
		Manifestaciones Pulmonares Crónicas Y Otras Manifestaciones Debidas A Radiación
		Neumoconiosis Asociada Con Tuberculosis
		Otra Rinitis Alergica Estacional
		Otros Trastornos Respiratorios Consecutivos A Procedimientos
		Otros Trastornos Respiratorios Especificados
		Quilotórax
		Reacción De Hipersensibilidad De Las Vías Respiratorias Superiores, Sitio No Especificado
		Síndrome De Macleod
		Síndrome De Mendelson
		Suberosis
		Trastorno No Especificado Del Sistema Respiratorio, Consecutivo A Procedimientos
		Trastornos Del Diafragma
		Trastornos Pulmonares Intersticiales Agudos Inducidos Por Drogas
		Trastornos Pulmonares Intersticiales Crónicos Inducidos Por Drogas
		Trastornos Pulmonares Intersticiales No Especificados Inducidos Por Drogas
Otras enfermedades respiratorias que afectan principalmente al intersticio	J80-J84	Afecciones Alveolares Y Alveoloparietales
		Edema Pulmonar
		Enfermedad Pulmonar Intersticial, No Especificada
		Enfisema Intersticial
		Eosinofilia Pulmonar, No Clasificada En Otra Parte
		Otras Enfermedades Pulmonares Intersticiales Con Fibrosis
		Otras Enfermedades Pulmonares Intersticiales Especificadas
		Síndrome De Dificultad Respiratoria Del Adulto
Otras infecciones agudas de las vías respiratorias inferiores	J20-J22	Bronquiolitis Aguda, No Especificada
		Bronquiolitis Aguda Debida A Metaneumovirus Humano
		Bronquiolitis Aguda Debida A Metapneumovirus Humano
		Bronquiolitis Aguda Debida A Otros Microorganismos Especificados
		Bronquiolitis Aguda Debida A Virus Sincitial Respiratorio
		Bronquitis Aguda, No Especificado
		Bronquitis Aguda Debida A Streptococos
		Bronquitis Aguda Debida A Otros Microorganismos Especificados
		Bronquitis Aguda Debida A Rinovirus
		Bronquitis Aguda Debida A Virus Coxsackie
		Bronquitis Aguda Debida A Virus Echo
		Bronquitis Aguda Debida A Virus Sincitial Respiratorio
		Infección Aguda No Especificada De Las Vías Respiratorias Inferiores
		Otros Trastornos Del Pulmón

Anexo 3

Estadísticas descriptivas de las 4 variables principales: MP2.5, MP10, Temperatura y cantidad de enfermedades

```
summary(egresos.rango$mp25)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00  16.52   23.99   28.98  37.65  193.00
```

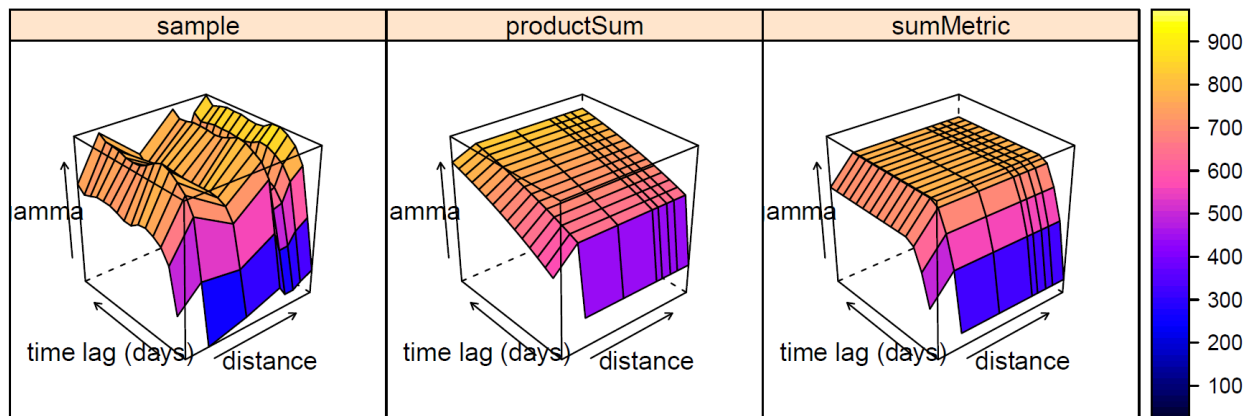
```
summary(egresos.rango$mp10)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.22  46.73   63.11   69.45  86.78  286.03
```

```
summary(egresos.rango$temp)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.670   9.691  13.391  13.907  18.444  29.100
```

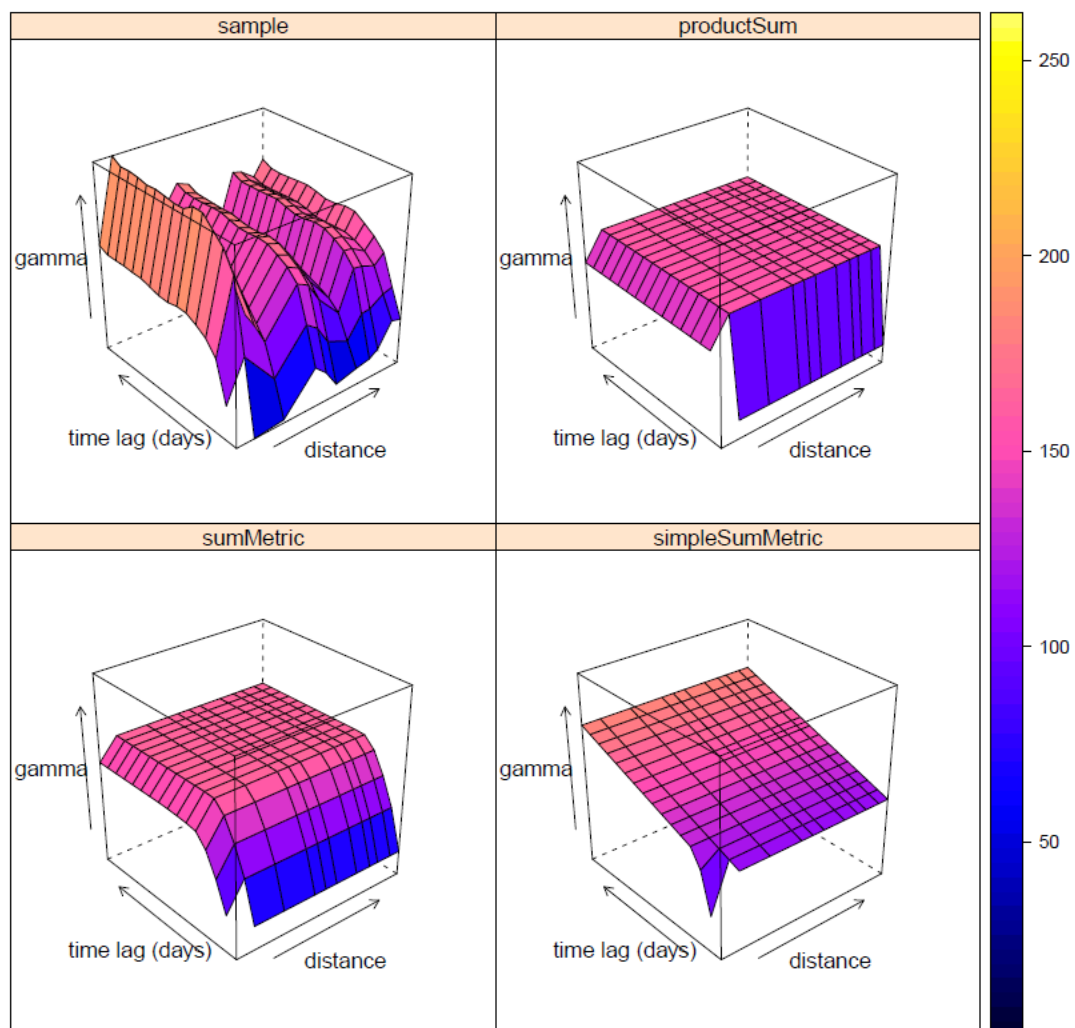
```
summary(egresos.rango$cantidadenf)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.000   1.000   1.000   1.292   1.000  24.000
```

Anexo 4

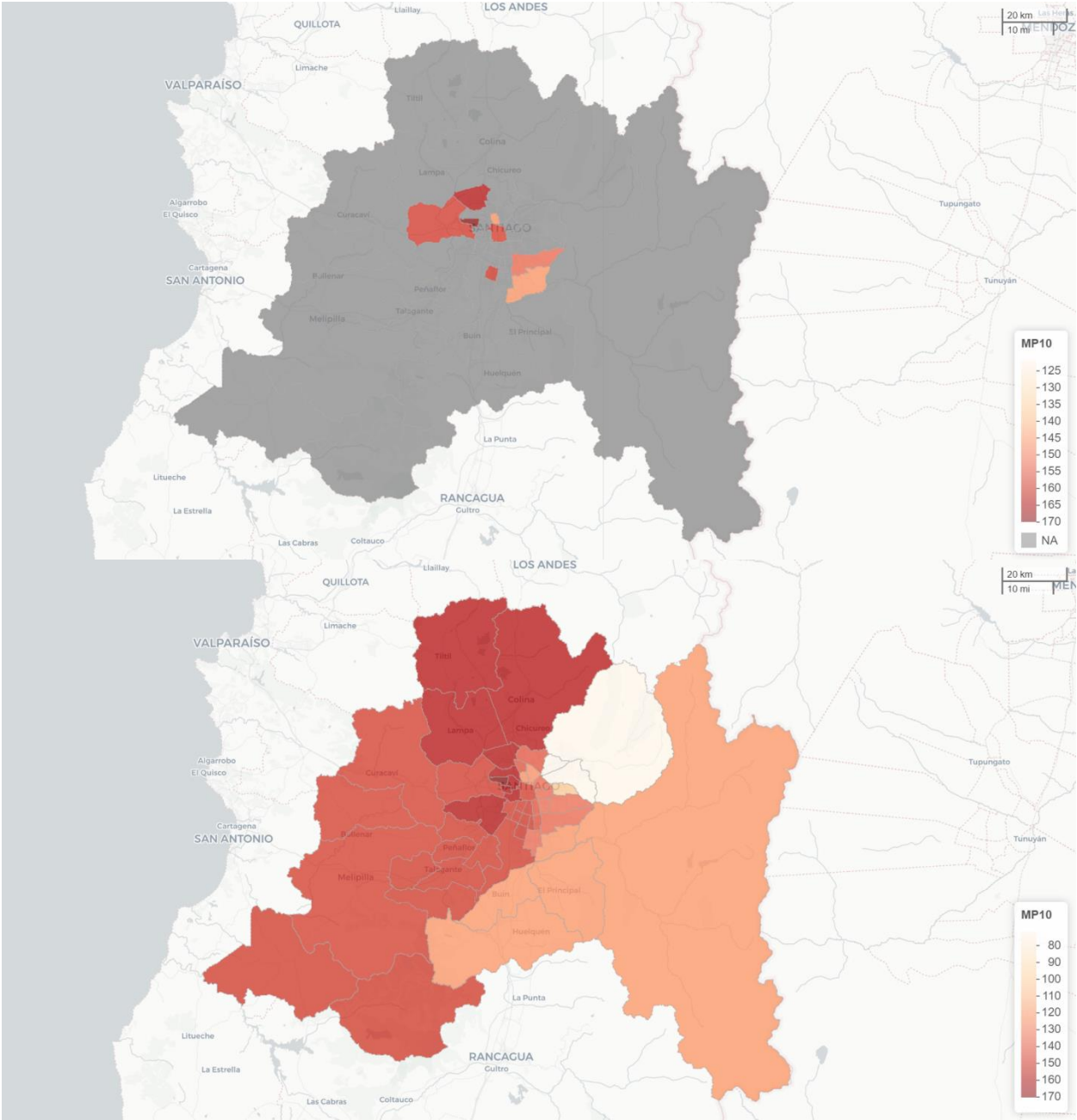
Variogramas MP10



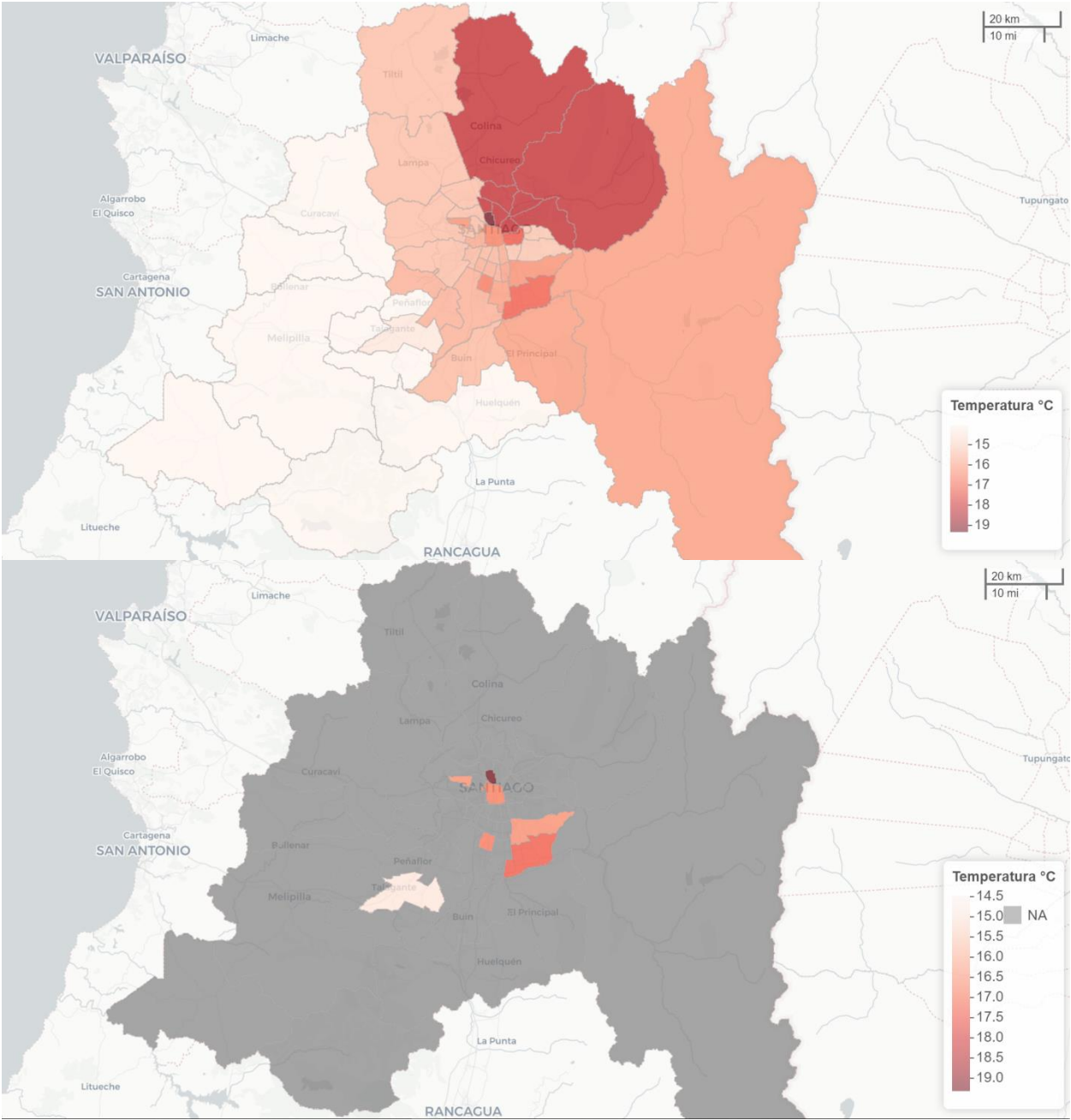
Variogramas MP2.5



Inteporlacion MP10 04 de mayo de 2012



Interpolación temperatura 04 de mayo de 2012



Bibliografía

- Athey, Susan & Imbens, Guido. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*. 11. 1-41. 10.1146/annurev-economics-080217-053433.
- Bellinger, C., Mohamed Jabbar, M., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, 17(1).
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.
- L. Breiman. Random forests. *Machine Learning*, 45(1):5,32, 2001a.
- Cakmak, S., Dales, R., & Vidal, C. (2007). Air Pollution and Mortality in Chile: Susceptibility among the Elderly. *Environmental Health Perspectives*, 115(4), 524-527. doi: 10.1289/ehp.9567
- Cakmak, S., Dales, R., Angelica Rubio, M., & Blanco Vidal, C. (2011). The risk of dying on days of higher air pollution among the socially disadvantaged elderly. *Environmental Research*, 111(3), 388-393. doi: 10.1016/j.envres.2011.01.003
- David, B. (2017). Model economic phenomena with CART and Random Forest algorithms.
- Gavidia, T., Pronczuk, J., & Sly, P. (2009). Impactos ambientales sobre la salud respiratoria de los niños: Carga global de las enfermedades respiratorias pediátricas ligada al ambiente. *Revista Chilena De Enfermedades Respiratorias*, 25(2).
- Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." (pdf) *Economic Inquiry* 56, no. 1 (January 2018): 114–137.
- Graeler, Benedikt & Gerharz, Lydia & Pebesma, Edzer. (2013). Spatio-temporal analysis and interpolation of PM 10 measurements in Europe.
- Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. *The R Journal*, 8(1), 204. doi: 10.32614/rj-2016-014
- Liaw, Andy & Wiener, Matthew. (2001). Classification and Regression by RandomForest. *Forest*. 23.
- Markandya, A., Sampedro, J., Smith, S., Van Dingenen, R., Pizarro-Irizar, C., Arto, I., & González-Eguino, M. (2018). Health co-benefits from air pollution and mitigation costs of the Paris Agreement: a modelling study. *The Lancet Planetary Health*, 2(3), e126-e133. doi: 10.1016/s2542-5196(18)30029-9
- Matus C., P., & Oyarzún G., M. (2019). Impacto del Material Particulado aéreo (MP2,5) sobre las hospitalizaciones por enfermedades respiratorias en niños: estudio caso-control alterno. *Revista Chilena De Pediatría*, 90(2), 166.

Mullainathan, Sendhil & Spiess, Jann. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*. 31. 87-106. 10.1257/jep.31.2.87.

Oyarzún G, M., & Valdivia C, G. (2012). Contaminación atmosférica: Declaración de la Academia de Medicina, Instituto de Chile. *Revista Chilena De Enfermedades Respiratorias*, 28(3), 171-173.

Oyarzún G, M. (2010). Contaminación aérea y sus efectos en la salud. *Revista Chilena De Enfermedades Respiratorias*, 26(1).

Pebesma, Edzer. (2012). Spacetime: Spatio-Temporal Data in R. *Journal of Statistical Software*. 51. 1-30. 10.18637/jss.v051.i07.

Pebesma, Edzer & Graler, Benedict. (2020). Introduction to Spatio-Temporal Variography. Institute for Geoinformatics, University of Munster. <https://cran.r-project.org/web/packages/gstat/vignettes/st.pdf>

Perez, P., & Gramsch, E. (2016). Forecasting hourly PM_{2.5} in Santiago de Chile with emphasis on night episodes. *Atmospheric Environment*, 124, 22-27. doi: 10.1016/j.atmosenv.2015.11.016

Perez, P., & Salini, G. (2008). PM_{2.5} forecasting in a large city: Comparison of three methods. *Atmospheric Environment*, 42(35), 8219-8224.

Perez, P., & Reyes, J. (2006). An integrated neural network model for PM₁₀ forecasting. *Atmospheric Environment*, 40(16), 2845-2851. doi: 10.1016/j.atmosenv.2006.01.010

Prieto-Parra, L., Yohannessen, K., Brea, C., Vidal, D., Ubilla, C., & Ruiz-Rudolph, P. (2017). Air pollution, PM 2.5 composition, source factors, and respiratory symptoms in asthmatic and nonasthmatic children in Santiago, Chile. *Environment International*, 101, 190-200. doi: 10.1016/j.envint.2017.01.021

Remaggi S., J., Mardones P., C., & Jiménez R., J. (2019). Impacto de la contaminación por material particulado sobre las atenciones de urgencias por causas respiratorias en Chillán, Concepción y Los Ángeles. *Revista Chilena De Enfermedades Respiratorias*, 35(3), 181-190.

Requia, W., Adams, M., Arain, A., Papatheodorou, S., Koutrakis, P., & Mahmoud, M. (2018). Global Association of Air Pollution and Cardiorespiratory Diseases: A Systematic Review, Meta-Analysis, and Investigation of Modifier Variables. *American Journal Of Public Health*, 108(S2), S123-S130.

Singh, V., Carnevale, C., Finzi, G., Pisoni, E., & Volta, M. (2011). A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations. *Environmental Modelling & Software*, 26(6), 778-786.

Varian, Hal. (2014). Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives*. 28. 10.1257/jep.28.2.3.

Webster, R., & Oliver, M. (2009). *Geostatistics for environmental scientists*. Chichester: John Wiley & Sons.

Zhu, W., Wang, J., Zhang, W., & Sun, D. (2012). Short-term effects of air pollution on lower respiratory diseases and forecasting by the group method of data handling. *Atmospheric Environment*, 51, 29-38.

Zivin, J., & Neidell, M. (2018). Air pollution's hidden impacts. *Science*, 359(6371), 39-40. doi: 10.1126/science.aap7711