Afectación por datos faltantes en un modelo PLS ajustado a los registros de Ozono en Santiago de Cali para el año 2019

Víctor Hugo Cifuentes Rodríguez

Juan David Espinosa Maca

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2021

Afectación por datos faltantes en un modelo PLS ajustado a los registros de Ozono en Santiago de Cali para el año 2019

Víctor Hugo Cifuentes Rodríguez
Juan David Espinosa Maca

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director:
Ph.D. Javier Olaya Ochoa
Codirector:
Ph.D. Víctor Manuel González Rojas

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2021

Dedicatoria

A Dios por darme la vida, la Salud, la sabiduría, paciencia y fuerza. A mi padre, pilar de mi vida por su amor incondicional y sacrificio constante. A mi hermana, mi compañera de vida por extender su mano y apoyarme en cada momento. A mis tías Luz, Anita y Yolanda por ser mis guías y maestras de vida, sin ustedes no sería quien soy hoy en día. A Abel por su compañía y entrega. A Daniela, mi novia, por ser mi bastón y mi refugio, por no dejarme desfallecer ni en los momentos más duros. A Beatríz por su infinito apoyo y enorme corazón. A todos y cada uno de los que aportaron de una u otra forma para alcanzar esta meta tan importante. Por Siempre, ¡Gracias!

Víctor Hugo Cifuentes Rodríguez

Este logro se lo dedico a mis padres que siempre creyeron en mí y los que nunca me hicieron dudar de lo orgullosos que se sienten, y a mi hermana Laura Daniela, la que siempre me apoyo con sus consejos, y me guió en los momentos más difíciles dándome aliento, haciendo que me esforzara día a día, además de enseñarme que por más errores que uno cometa, siempre se tendrá la oportunidad de redimirse.

Juan David Espinosa Maca

Agradecimientos

Agradecemos a nuestras familias, quienes nos han brindado todo apoyo y compañía en este largo proceso, quienes nos han visto crecer y con múltiples esfuerzos han logrado convertirnos en los seres que somos hoy en día, quienes no nos permitieron desfallecer y creyeron en nuestras capacidades incluso en los momentos más difíciles.

Agradecemos a la Universidad del Valle por brindarnos la oportunidad de estudiar y alcanzar el titulo de Estadísticos y a la Escuela de Estadística por brindarnos todas las herramientas para cumplir cada etapa de nuestro aprendizaje.

A nuestros profesores y directores de trabajo de grado Javier Olaya Ochoa y Víctor Manuel González Rojas quienes con su conocimiento, experiencia, sabiduría, tiempo, interés y paciencia nos guiaron en cada etapa de este proyecto.

A nuestros profesores quienes nos transmitieron día a día todo su conocimiento y ponen lo mejor de sí para fojar profesionales de alta calidad.

Al Departamento Administrativo de Gestión del Medio Ambiente (DAGMA), por apoyarnos en la consecución y entendimiento de los datos usados para llevar a cabo este trabajo de grado.

Resumen

La estación de monitoreo Compartir, en Cali, Colombia, registra datos horarios del Ozono troposférico y las variables climáticas: temperatura, humedad relativa, lluvia, velocidad del viento y radiación solar. Sin embargo esta base de datos contiene datos faltantes sistemáticos. Usando los registros del año 2019 se construye un conjunto de datos con todas las filas que tienen datos completos y se ajusta un modelo de regresión PLS con una componente y MCO del Ozono en función de las variables climáticas, encontrando que la diferencia de R^2 entre los modelos es de 0.022, no obstante, los intervalos de confianza para los coeficientes de regresión en el PLS tienen menos amplitud, ya que corrige la multicolinealidad presentada en las variables predictoras. Posteriormente, se contamina la base con el 5 %, 10 %, 15 %, 20 %, 30 %, 40 %, 50 % de datos faltantes aleatorios en las variables predictoras, simulando 10 mil escenarios de contaminación con cada porcentaje y se ajusta el modelo PLS en cada uno, encontrando que el modelo mantiene buenas métricas hasta el 15 % de datos faltantes aleatorios.

Palabras clave: Ozono, Variables Climáticas, PLS, Datos Faltantes

Abstract

The Compartir monitoring station in Cali, Colombia, records hourly data for tropospheric ozone and the climatic variables: temperature, relative humidity, rainfall, wind speed and solar radiation. However, this database contains systematic missing data. Using the records of the year 2019, a dataset is constructed with all the rows that have complete data and a PLS and OLS regression model of Ozone as a function of the climatic variables is fitted, finding that the R^2 difference between the models is 0.022. However, the confidence intervals for the regression coefficients in the PLS have less amplitude since it corrects the multicollinearity presented in the predictor variables. Subsequently, the base is contaminated with 5 %, 10 %, 15 %, 20 %, 30 %, 40 %, 50 % of random missing data in the predictor variables, simulating 10 thousand contamination scenarios with each percentage and the PLS model is adjusted in each one, finding that the model maintains good metrics up to 15 % of missing data.

Keywords: Ozone, climatic variables, PLS, Missing Data

Contenido

Lista de Figuras

Lista de Tablas

Declaración

Nos permitimos afirmar que hemos realizado el presente Trabajo de Grado de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el propio trabajo. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los hemos reconocido. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de Tesis o Trabajo de Grado.

Igualmente declaramos que los datos utilizados en este trabajo están protegidos por las correspondientes cláusulas de confidencialidad.

Santiago de Cali, 07.09.2021

_____

(Víctor Hugo Cifuentes Rodríguez)

_____

(Juan David Espinosa Maca)

1 Introducción

El Ozono a nivel del suelo se considera un contaminante secundario de gran importancia, que trae consigo afectaciones a la salud de los seres vivos e impactos negativos en el sector industrial. Debido a ello es importante registrar constantemente mediciones de este contaminante en cualquier parte del mundo.

Santiago de Cali es una ciudad ubicada en el suroccidente colombiano, con una población aproximada de 2.4 millones de personas en el 2018, cuenta con instituciones encargadas de vigilar la calidad medio ambiental en la ciudad, como el DAGMA, CVC, entre otros. Este municipio tiene actualmente nueve estaciones de monitoreo de la calidad del aire, en algunas de ellas registra datos del Ozono troposférico (O_3) además de variables climáticas como temperatura, humedad relativa, lluvia, velocidad del viento y radiación solar cuyas mediciones son un factor importante para entender el comportamiento, dilución, creación y transporte del Ozono, lo que permite construir una ecuación matemática a partir de estas variables climáticas y obtener estimaciones del O_3.

No obstante, por motivos de falla en el suministro de la red eléctrica o calibración periódica de los medidores, se dan las condiciones para la presencia de datos faltantes, incluso se obtienen rachas seguidas de ausencia de datos durante cierto periodos, lo que genera inconvenientes para usar métodos de regresión clásicos y obtener estimaciones de Ozono en función de las variables climáticas. Adicionalmente la naturaleza de las variables genera correlaciones considerables entre ellas, lo que se traduce en un problema de multicolinealidad que afecta la interpretabilidad e inferencia de modelos como la regresión por mínimos cuadrados ordinarios.

Una posible solución frente a los datos faltantes (NA) podría ser la imputación. Sin embargo, existen métodos estadísticos que permiten ajustar modelos cuando se tiene presencia de datos faltantes e incluso multicolinealidad, como es el caso del modelo de regresión PLS (Partial Least Squares) (Tenenhaus 1998), que utiliza el principio de datos disponibles. Lo anterior, da lugar a la pregunta ¿Cómo afectan los datos faltantes a un modelo de regresión PLS ajustado a los datos de Ozono en función de las variables climáticas?. Para responder, se usan los registros de Ozono troposférico y variables climáticas en la Estación Compartir para el año 2019, donde se observa que la base cuenta con datos faltantes con un comportamiento sistemático, por lo que no es posible trabajar directamente con los NA

originales.

Debido a lo anterior, se construye inicialmente una base de datos omitiendo las filas que contengan al menos un dato faltante en la base de Compartir 2019, lo que permite obtener una base con datos completos y a partir de ella evaluar el ajuste del modelo de regresión MCO y PLS. Para comparar los resultados obtenidos en ambos modelos, se obtienen los coeficientes de determinación R^2 e intervalos de confianza para los coeficientes de regresión en ambos modelos. Sin embargo, dado que aún no se conocen las propiedades que tienen las estimaciones de los coeficientes de regresión del modelo PLS, se utiliza el método Bootstrap.

Posteriormente la base de datos es contaminada con diferentes porcentajes de registros faltantes aleatorios en el conjunto de variables predictoras y se generan diez mil escenarios diferentes de contaminación para cada porcentaje, para ajustar el modelo de regresión PLS en presencia de datos faltantes, lo que posibilita la construcción de intervalos de confianza para el R^2 y los coeficientes de regresión en cada porcentaje.

Finalmente, se compara el R^2 obtenido en PLS con datos completos y se analiza su variación en función del aumento progresivo de la presencia de datos faltantes. Adicionalmente, se evalúa el cambio que sufre la amplitud del intervalo de confianza para los coeficientes de regresión en cada porcentaje de NA.

Esta investigación permite ajustar un modelo de regresión para estimar el Ozono troposférico en función de las variables climáticas en Santiago de Cali teniendo en cuenta la presencia de datos faltantes aleatorios sin recurrir a técnicas de imputación, además de establecer un porcentaje aceptable de datos faltantes en el conjunto de observaciones, sin tener afectaciones importantes en las métricas del modelo ni la eficacia del mismo.

2 Planteamiento del problema

El Ozono a nivel de la estratosfera protege a los seres vivos de la radiación solar, formando la Capa de Ozono. Sin embargo, a nivel del suelo, el Ozono se considera un serio contaminante del aire (EPA 2003). Surge a partir de otros productos, principalmente Óxidos de Nitrógeno (NO_x) y Compuestos Orgánicos Volátiles (VOC), en presencia de abundante luz solar, por lo que se le define como contaminante secundario. Los NO_x se generan en los procesos de combustión y especialmente por los medios de transporte tradicionales, los VOC se generan a partir de un número de fuentes variado, transporte por carretera, refinerías, pintura, limpieza en seco de tejidos y otras actividades que implican el uso de disolventes (Gómez 2016).

Según la Organización Mundial de la Salud (OMS 2003), la contaminación producida por Ozono provoca irritación de ojos, tos, dolores de cabeza, disminución de la función pulmonar, asma, falta de aliento, dolor en el pecho, entre otras afectaciones en la salud. Además, en grandes cantidades también es perjudicial para el resto de los seres vivos, puesto que afecta a las paredes celulares, disminuye la actividad fotosintética y perjudica el crecimiento de las plantas, provocando una disminución de la vegetación natural y de la producción agrícola.

El Boletín Mensual de Calidad del Aire de Santiago de Cali, para el mes de marzo de 2018 (DAGMA 2018), informa que los niveles de Ozono Troposférico (O_3) sobrepasan los niveles máximos permisibles (100μ g/m^3) 8 y 6 veces en el mes, en las estaciones Pance y Univalle, respectivamente. Las excedencias se registraron en horas cercanas al mediodía, por lo que la alta radiación solar puede presentar una influencia en el incremento de este contaminante. Según dicho informe, la causa de esta excedencia es la ubicación de la estación ambiental, ya que al encontrarse retiradas de las vías, las concentraciones de NO_x son menores, lo que hace que disminuyan las reacciones secundarias que consumen el Ozono.

En Santiago de Cali, al igual que en otras ciudades de Colombia, hay periodos en los que no es posible obtener la completitud de datos de las variables ambientales y/o de Ozono en las estaciones de monitoreo, esto se debe a que los medidores deben ser calibrados al menos una vez al año o deben repararse. Por otro lado, el registro de datos depende especialmente del fluido eléctrico, es decir, en las ocasiones en que no se cuenta con energía eléctrica es imposible tomar mediciones del Ozono y de las variables climáticas. Al año, se presentan cortes en el fluido eléctrico en sectores de la ciudad debido a lluvias, tormentas eléctricas,

daños en el sistema, entre otros factores. Es necesario tener en cuenta que los equipos de medición también necesitan de una temperatura óptima para no tomar mediciones erróneas, por ende, deben estar regulados por medio de un aire acondicionado que también puede presentar fallas por cortes eléctricos u otros factores, lo que conlleva a tener rachas de datos faltantes durante algunos periodos de tiempo.

Las variables climáticas juegan un papel muy importante en la creación y transporte del Ozono troposférico, debido a que los contaminantes precursores de este se ven influenciados directamente por la lluvia, humedad relativa, radiación solar, dirección y velocidad del viento, entre otros (Velázquez de Castro 2003). Cabe resaltar que, por su naturaleza las variables pueden presentar alta correlación entre sí, esto quiere decir que las variables predictoras pueden presentar multicolinealidad, lo que dificulta realizar una estimación de un modelo usando el método de Mínimos Cuadrados Ordinarios, según Montgomery et al. (2006).

Teniendo en cuenta estos factores, resulta necesario buscar métodos que permita estimar el Ozono troposférico usando las variables climáticas como predictores, cuando se tienen datos faltantes y multicolinealidad entre las variables predictoras, como el Partial Least Squares (PLS).

3 Justificación

El Ozono a nivel del suelo se considera un contaminante secundario de gran importancia que trae consigo afectaciones a la salud de los seres vivos e impactos negativos en distintos sectores económicos e industriales. Santiago de Cali cuenta con nueve estaciones de monitoreo de la calidad del aire, de las cuales hasta el momento, seis miden el Ozono troposférico, pero sólo la Estación Compartir registra mediciones de las variables climáticas en conjunto con este contaminante.

Es común que se puedan presentar fallas en la toma de datos o incluso que no se registren, debido a diversos factores climáticos, logísticos y técnicos. Estos datos faltantes pueden presentarse en la variable Ozono y las variables climáticas, incluso por largos periodos, por lo que resulta necesario estimar el Ozono cuando hacen falta datos de las variables climáticas, con el fin de tomar decisiones tempranas que permitan mantener una óptima calidad en el aire de la ciudad de Santiago de Cali.

Dado que los métodos clásicos presentan limitaciones para trabajar con variables correlacionadas y/o datos faltantes, surgen metodologías que permitan superar estos dos inconvenientes. Sin embargo, algunas de ellas no son interpretables fácilmente. Es aquí donde resulta pertinente usar técnicas como la Regresión PLS (Partial Least Squares) que permiten estimar un modelo con datos faltantes y variables correlacionadas, además de permitir una interpretación sencilla sobre el fenómeno. Cabe resaltar que con un modelo de regresión PLS óptimo, es posible realizar un acercamiento a la imputación de datos faltantes.

4 Objetivos

4.1. Objetivo General

Determinar los niveles de afectación de los datos faltantes en un modelo PLS ajustado con los datos de Ozono en función de las variables climáticas registradas en la Estación Compartir en Santiago de Cali para el año 2019.

4.2. Objetivos Específicos

- Caracterizar el comportamiento de los datos faltantes en los registros de la Estación Compartir para el año 2019.
- Estimar el modelo de Regresión PLS para el Ozono en Santiago de Cali durante el año 2019 en función de las variables climáticas registradas en la Estación de Monitoreo Compartir, con datos completos en las variables predictoras y de respuesta.
- Estimar el modelo de Regresión PLS para el Ozono en Santiago de Cali durante el año 2019 en función de las variables climáticas registradas en la Estación de Monitoreo Compartir, con distintos porcentajes de datos faltantes aleatorios en las variables predictoras.
- Evaluar el nivel de afectación del modelo de Regresión PLS con los distintos porcentajes de datos faltantes aleatorios en las variables predictoras.

4.3. Pregunta de investigación

¿Cómo afectan los datos faltantes a un modelo de regresión PLS ajustado a los datos de Ozono en función de las variables climáticas registradas en la Estación Compartir en Santiago de Cali para el año 2019?

5 Antecedentes

Para modelar la concentración del Ozono han surgido diversas propuestas y enfoques como modelos de regresión cuantílica, múltiple, no lineal, jerárquica, logística y series de tiempo. Por otra parte, existen algunos acercamientos a la Regresión PLS en diversos campos, como solución a la multicolinealidad y datos faltantes. Algunos de los trabajos más destacados se presentan a continuación:

Vega Vilca & Guzmán (2011) proponen dos métodos para solucionar el problema de multicolinealidad en regresión múltiple; estos dos métodos son regresión por componentes principales (PCA) y regresión por componentes desde mínimos cuadrados parciales (PLS), permitiendo realizar una comparación entre estos e ilustrar metodologías con ejemplos de aplicación. Este artículo es fundamental para el presente trabajo de grado ya que posibilita la comprensión del algoritmo PLS en presencia de multicolinealidad entre las variables de estudio usadas.

González Rojas (2016) escribió el artículo 'Inter-Battery Factor Analysis via PLS: The Missing' con el objetivo de desarrollar el Análisis Factorial Interbaterias mediante el uso de métodos PLS con datos faltantes. El autor propone modificar la fase iterativa del algoritmo NIPALS e implementar el principio de datos disponibles, con esto se permite realizar análisis de los resultados. El software implementado para este artículo es R, el cual también será usado para la presente tesis. Al finalizar el artículo, el autor compara los resultados obtenidos con presencia del 7 % de datos faltantes y los obtenidos con los datos completos, llegando a la conclusión que los resultados son similares y que el uso del método PLS fue efectivo para el problema mencionado.

Naranjo & Ortíz (2016) en 'Estimación del nivel de Ozono troposférico en el aire a partir de un índice de temperatura-humedad, de la radiación solar y del nivel de Dióxido de Nitrógeno utilizando análisis de regresión con datos funcionales en la ciudad de Santiago de Cali', presenta un modelo de regresión funcional para la estimación de contaminación horaria del Ozono a partir de muchas de las variables climáticas usadas en el presente trabajo, las cuales fueron obtenidas de distintas Estaciones de Monitoreo de la calidad del Aire en Santiago de Cali, usando registros tomados entre febrero de 2010 a diciembre del 2011. Debido a que se plantea colinealidad entre las variables temperatura y humedad relativa, el autor propone calcular un índice de temperatura-humedad (ITH) con el fin de

tratar este problema. En conclusión, este antecedente aborda el mismo problema tratado en el presente trabajo de grado, desde la perspectiva de los modelos funcionales, además aborda la multicolinealidad a través de una propuesta que mide el nivel de inconformidad en los seres vivos.

El objetivo de Al-Shammari (2018), en el artículo 'Towards an accurate ground-level ozone prediction' es encontrar la técnica más apropiada para predecir el Ozono troposférico en la estación Al Jahra, Kuwait. Para esto se tomaron datos sobre las variables climáticas: temperatura del aire, humedad relativa, radiación solar, dirección y velocidad del viento, además de la concentración de siete contaminantes del medio ambiente (SO_2, NO_2, NO, CO_2, CO, NMHC y CH_4). Se implementaron los métodos de regresión PLS, máquina de soporte vectorial (SVM) y regresión múltiple por mínimos cuadrados para realizar dicha estimación y tras evaluar cada uno, se llegó a la conclusión que cualquiera de los tres podría usarse para predecir las concentraciones de Ozono a nivel del suelo en la estación Al Jahra en Kuwait.

Por último, Abdul-Wahab et al. (2005) utilizaron las concentraciones de 7 contaminantes ambientales (SO_2, NO_2, NO, CO_2, CO, NMHC y CH_4) y variables meteorológicas (velocidad y dirección del viento, temperatura del aire, humedad relativa y radiación solar) para predecir las concentraciones de ozono utilizando los métodos de regresión lineal múltiple y regresión por componentes principales (PCA); además se hicieron análisis separados para las concentraciones de ozono durante el día y la noche. Los autores descubrieron que alta temperatura y alta energía solar tienden a aumentar las concentraciones durante el día, mientras que, durante la noche las concentraciones de ozono fueron influenciadas predominantemente por los óxidos de nitrógeno ($NO+NO_2$). Sin embargo, durante la noche el modelo no predijo las concentraciones de ozono con tanta precisión como lo hizo durante el día, por lo que ellos consideran que podría deberse a algunos factores que no se consideraron en su estudio.

En los antecedentes aquí presentados, se evidencia que el método PLS ha sido un buen mecanismo para tratar con problemas de multicolinealidad, y es posible realizar modificaciones a este método para trabajar con datos faltantes, algo que es muy recurrente en temas como la estimación de Ozono troposférico en cualquier ciudad. Por otra parte, se observa que las variables climáticas y los principales contaminantes del medio ambiente son esenciales para la estimación del Ozono a nivel del suelo, ya que influyen directamente en su creación y transporte.

6 Marco Teórico

6.1. Ozono troposférico

El Ozono troposférico ($\mu g/m^3$) es denotado como O_3 y es medido en las estaciones de monitoreo a través de fotometría ultravioleta. El Ozono es un gas, generalmente incoloro, compuesto por tres átomos de oxígeno (Straif et al. 2017). Este elemento se encuentra en la estratósfera (Capa de Ozono) y tropósfera. En esta última, es el resultado de las reacciones entre óxidos de Nitrógeno (N_2O, NO, N_2O_3, N_2O_4, NO_2, N_2O_5) e hidrocarburos emitidos principalmente por medios convencionales de transporte e industrias, en presencia de radiación solar (de Ambiente Vivienda y Desarrollo Territorial 2010).

Las variables climáticas juegan un papel muy importante en la creación y transporte del Ozono troposférico, ya que según Velásquez de Castro (2003), 'existe una *correlación* preferente entre los valores máximos de Ozono diarios y los valores máximos de temperatura del aire'; además la dirección del viento es influyente en el transporte de este gas (Wackter & Bayly 1988).

El Ozono troposférico es perjudicial para la salud del ser humano y del medio ambiente. En el ser humano, la exposición a altos niveles de Ozono, durante un período considerable de tiempo, produce irritación en los ojos, fosas nasales, garganta, bronquios e inflamación en las mucosas (Velázquez de Castro 2003). En las plantas, la exposición a Ozono puede producir reducción en el proceso de fotosíntesis, lo que conlleva a disminución en el rendimiento de los cultivos (Friedman et al. 1988). Por otro lado, también se ha descubierto que altos niveles de Ozono puede ser causa de afectaciones en objetos como el caucho, algodón, entre otros (Shaver et al. 1983).

6.1.1. Niveles permitidos de Ozono

La organización Mundial de la Salud (2003), recomienda que los niveles de Ozono troposférico no deben superar los 200 $\mu g/m^3$ durante una hora, ni 120 $\mu g/m^3$ durante ocho horas, ya que, si se superan estos niveles se puede desencadenar consecuencias cómo las ya mencionadas. Por otro lado, en Colombia, el Ministerio de Ambiente, Vivienda y Desarrollo Territorial, en la Resolución 2254 del 01 de Noviembre de 2017, Artículo 02

(de Ambiente Vivienda y Desarrollo Territorial 2017); establece como niveles máximos permitidos en todo el territorio nacional 100 $\mu g/m^3$ durante 8 horas.

6.2. Variables climáticas

La topografía y meteorología local deben ser consideradas para realizar predicciones de cualquier contaminante del aire (Şen et al. 2006); en este sentido, las variables climáticas: temperatura del aire, humedad relativa, radiación solar, dirección y velocidad del viento pueden ser empleadas en la estimación del Ozono. Estas variables son registradas en la ciudad de Santiago de Cali en diversas estaciones de monitoreo y se definen de la siguiente manera:

- **Temperatura del aire (°C):** Es una medida numérica que indica el calentamiento o enfriamiento de la atmósfera, causado principalmente por la radiación del sol (Rodríguez et al. 2004); los valores numéricos de la temperatura permiten, entre otras cosas, diferenciar una región climática de otra (Trapasso 2005). La temperatura del aire dependerá en parte de la radiación solar y ambas son necesarias para que ocurran las reacciones que dan origen al Ozono. Esta variable es medida con un termómetro de mercurio y debe estar protegido de la lluvia y la radiación solar.

- **Humedad relativa (%):** Es el cociente entre cantidad de vapor de agua contenida en el aire (humedad absoluta) y la máxima cantidad que el aire sería capaz de contener a esa temperatura (humedad absoluta de saturación), expresada en porcentaje. La saturación ocurre cuando el aire húmedo a presión P y la temperatura T pueden equilibrarse con agua pura o hielo a la misma temperatura y presión (Oliver 2005). Esta variable se mide gracias a un higrómetro instalado en la estación de monitoreo (Sarochar 2014).

- **Radiación solar (W/m^2):** Según el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM 2014), la radiación solar es la energía emanada por el sol que se propaga en todas las direcciones a través del espacio mediante ondas electromagnéticas. Es de vital importancia para el planeta tierra ya que sin ella, la tierra sería fria, oscura y sin vida, además de impulsar, junto a la rotación de la tierra, la circulación oceánica y atmosférica (McArthur 2005). La radiación solar es afectada por la lluvia y las nubes, ya que estos factores aumentan o disminuyen la energía solar.

  Esta variable es especialmente importante en la formación de Ozono troposférico ya que permite elevarlo o disminuirlo, según los niveles presentes. El elemento usado para

medir esta variable es el piranómetro.

- **Dirección (grados) y Velocidad del viento (m/s):** El viento es definido como el movimiento del aire que sirve como equilibrio térmico primario que ayuda a compensar el desequilibrio energético latitudinal persistente de la Tierra (Balling & Cerveny 2005). La velocidad y dirección del viento, son los encargados del transporte del Ozono a nivel del suelo, es decir, velocidades altas lo diluyen rápidamente (Gómez 2016) y la dirección determina de donde provienen los causantes de este contaminante. La medición de la velocidad del viento se obtiene por un anemómetro y la dirección del viento se mide mediante una veleta.

- **Lluvia (mm):** Según la Organización Meteorológica Mundial (Organization 1993), la lluvia es precipitación de gotas de agua que caen a la tierra y provienen de una nube formada en la atmósfera. Para ser considerada lluvia, la precipitación de agua debe superar los 0.5 mm, ya que si es inferior a este valor es considerada llovizna. La lluvia se ve afectada por la presión atmosférica, la temperatura y la humedad atmosférica (Tucker 2005). En las Estaciones de monitoreo del aire, es posible cuantificar la cantidad de lluvia en una hora gracias a pluviómetros.

6.3. Regresión lineal múltiple

En general, se puede relacionar la variable de respuesta y con p regresores o variables predictoras, en un modelo de tipo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x2 + \cdot + \beta_p x_p + e$$

A este modelo se le llama modelo de regresión lineal múltiple con p regresores, en el que se ha supuesto:

- $E[e] = 0$
- $Var[e] = \sigma^2$
- $Cov[e_i, e_j] = 0 \quad , \quad \forall \quad i \neq j$
- $e \sim N(0, \sigma^2 I)$

Los parámetros β_j, $j = 0, 1, \ldots, p$ se llaman coeficientes de regresión. El parámetro β_j representa el cambio esperado en la respuesta y por cambio unitario en x_j cuando todas las variables regresoras $x_i (i \neq j)$ se mantienen constantes.

6.3.1. Estimación por Mínimos Cuadrados Ordinarios (MCO)

Según Montgomery et al. (2006), se puede usar el método de mínimos cuadrados ordinarios para estimar los coeficeintes de regresión. En general, y es un vector de $n \times 1$ observaciones, X es una matriz de $n \times p$ de las variables regresoras, β es un vector de $p \times 1$ de los coeficientes de regresión y e es un vector de $n \times 1$ de errores aleatorios.

Se desea determinar el vector $\hat{\beta}$ de estimadores de mínimos cuadrados que minimice:

$$S(\beta) = \sum_{i=1}^{n} e_i^2 = e^{'}e = (y - X\beta)^{'}(y - X\beta)$$

Operando un poco, se tiene que $S(\beta)$ se puede expresar como:

$$S(\beta) = y^{'}y - \beta^{'}X^{'}y - y^{'}X\beta + \beta^{'}X^{'}X\beta = y^{'}y - 2\beta^{'}X^{'}y + \beta^{'}X^{'}X\beta$$

Ya que $\beta^{'}X^{'}y$ es una matriz de 1×1, es decir, un escalar, por lo que su transpuesta es el mismo escalar. Los estimadores de mínimos cuadrados ordinarios deben satisfacer:

$$\frac{\partial S}{\partial \beta}|_{\hat{\beta}} = -2X^{'}y + 2X^{'}X\hat{\beta} = 0$$

Cuya ecuación se puede simplificar hasta obtener las ecuaciones normales de mínimos cuadrados.

$$X^{'}X\hat{\beta} = X^{'}y \tag{6-1}$$

Para resolver las ecuaciones normales se multiplican a ambos lados de 6-1 por la inversa de $X^{'}X$. Así, el estimador de β por mínimos cuadrados es:

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}y \tag{6-2}$$

Siempre y cuando exista la matriz inversa $(X^{'}X)^{-1}$. Esta matriz siempre existe si los regresores son linealmente independientes, dicho de otra forma, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

6.4. Multicolinealidad

Se entiende como multicolinealidad a las dependencias casi lineales entre las variables regresoras, es decir, la alta correlación existente entre dos o más variables predictoras. La multicolinealidad grave puede causar inferencias engañosas o erróneas en el modelo de

regresión. La presencia de multicolinealidad hace que el modelo de regresión por mínimos cuadrados no sea adecuado. (Montgomery et al. 2006)

Algunas de las fuentes principales de multicolinealidad son:

- El método de recolección de datos que se empleó.
- Restricciones en el modelo o en la población.
- Especificación del modelo.
- Un modelo sobre definido (más parámetros que datos).
- La naturaleza de las variables predictoras y su asociación.

Supongamos un modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, se toma una muestra aleatoria de tamaño n, si se estandarizan las variables entonces:

$$W_{i1} = \frac{X_{i1}-\bar{X_1}}{\sqrt{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})^2}} \qquad ; \qquad W_{i2} = \frac{X_{i2}-\bar{X_2}}{\sqrt{\sum_{i=1}^{n}(X_{i2}-\bar{X_2})^2}}$$

El modelo estaría definido como: $Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + e$, por lo que se sigue el procedimiento por mínimos cuadrados para la estimación de los β's.

$$\mathrm{W}= \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ \vdots & \vdots \\ W_{n,1} & W_{n,2} \end{pmatrix} \qquad W^TW= \begin{pmatrix} \sum_{i=1}^{n}(W_{i1})^2 & \sum_{i=1}^{n} W_{i1}W_{i2} \\ \\ \sum_{i=1}^{n} W_{i2}W_{i1} & \sum_{i=1}^{n}(W_{i2})^2 \end{pmatrix}$$

$$\sum_{i=1}^{n}(W_{i1})^2 = \sum_{i=1}^{n}(\frac{X_{i1}-\bar{X_1}}{\sqrt{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})^2}})^2 = \frac{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})^2}{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})^2} = 1$$

$$\sum_{i=1}^{n} W_{i1}W_{i2} = \sum_{i=1}^{n}(\frac{X_{i1}-\bar{X_1}}{\sqrt{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})^2}})(\frac{X_{i2}-\bar{X_2}}{\sqrt{\sum_{i=1}^{n}(X_{i2}-\bar{X_2})^2}})$$

$$\sum_{i=1}^{n} W_{i1}W_{i2} = \frac{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})(X_{i2}-\bar{X_2})}{\sqrt{\sum_{i=1}^{n}(X_{i1}-\bar{X_1})^2(X_{i2}-\bar{X_2})^2}} = r_{12} = r_{21}$$

Como se puede observar, la matriz W^TW corresponde a una matriz de correlaciones, donde en la diagonal principal serán 1's, ya que esto es la correlación de la variable W_i consigo misma, mientras que la diagonal secundaria quedará la correlación entre la variable W_1 y

W_2. Por lo tanto:

$$W^TW = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \quad \Rightarrow \quad (W^TW)^{-1} = \frac{1}{1-r_{12}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

$$W^T\vec{Y} = \begin{pmatrix} \vec{W_1^T} \\ \vec{W_2^T} \end{pmatrix}\vec{Y} = \begin{pmatrix} \vec{W_1^T}\vec{Y} \\ \vec{W_2^T}\vec{Y} \end{pmatrix}$$

$$\hat{\vec{\beta}} = (W^TW)^{-1}W^T\vec{Y} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

$$\Rightarrow \hat{\vec{\beta}} = \frac{1}{1-r_{12}^2}\begin{pmatrix} \vec{W_1^T}\vec{Y} - r_{12}\vec{W_2^T}\vec{Y} \\ -r_{12}\vec{W_1^T}\vec{Y} + \vec{W_2^T}\vec{Y} \end{pmatrix} \tag{6-3}$$

De la Ecuación 6-3, se puede observar que las estimaciones para los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ serán:

$$\hat{\beta}_1 = \frac{1}{1-r_{12}^2}(\vec{W_1^T}\vec{Y} - r_{12}^2\vec{W_2^T}\vec{Y}) \qquad \hat{\beta}_2 = \frac{1}{1-r_{12}^2}(\vec{W_2^T}\vec{Y} - r_{12}^2\vec{W_1^T}\vec{Y})$$

Si X_1 y X_2 están altamente correlacionadas linealmente, entonces $|r_{12}| \approx 1$. Esto tiene un efecto en los parámetros de la siguiente manera:

$$\hat{\vec{\beta}} = \frac{1}{1-1^2}\begin{pmatrix} \vec{W_1^T} - r_{12}^2\vec{W_2^T}\vec{Y} \\ -r_{12}^2\vec{W_1^T} + \vec{W_2^T}\vec{Y} \end{pmatrix} = \frac{1}{1-1}\begin{pmatrix} \vec{W_1^T} - r_{12}^2\vec{W_2^T}\vec{Y} \\ -r_{12}^2\vec{W_1^T} + \vec{W_2^T}\vec{Y} \end{pmatrix} \to \infty$$

$$Var(\hat{\beta}_j) = C_{jj}\sigma^2 \qquad C = (W^TW)^{-1}$$

$$Var(\hat{\beta}_1) = C_{11}\sigma^2 = \frac{1}{1-r_{12}^2}\sigma^2 = \frac{1}{1-1^2}\sigma^2 = \frac{1}{1-1}\sigma^2 \to \infty$$

$$Var(\hat{\beta}_2) = C_{22}\sigma^2 = \frac{1}{1-r_{12}^2}\sigma^2 = \frac{1}{1-1^2}\sigma^2 = \frac{1}{1-1}\sigma^2 \to \infty$$

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 = \frac{-r_{12}}{1-r_{12}^2}\sigma^2 = \pm\frac{1}{1-1^2}\sigma^2 = \pm\frac{1}{1-1}\sigma^2 \to \pm\infty$$

Por lo tanto, mientras más correlacionadas estén las variables, más se incrementan las estimaciones de los parámetros, las varianzas y covarianzas, lo que deterioran el uso y la interpretación del modelo, así como las inferencias que se puedan hacer. Lo anterior, implica que al tomar una nueva muestra, las estimaciones de los parámetros del modelo podrían ser muy diferentes.

Este problema se puede observar de manera equivalente desde el álgebra lineal. Para eso supongamos un modelo de la forma $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + e$

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & X_{14} \\ 1 & X_{21} & X_{22} & X_{23} & X_{24} \\ 1 & X_{31} & X_{32} & X_{33} & X_{34} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & X_{n,2} & X_{n,3} & X_{n,4} \end{pmatrix} \qquad Rank(X) = P = 5$$

El rango de una matriz se define como el número de filas (o columnas) linealmente independientes. Ahora supongamos que la variable X_1 es combinación lineal de las demás variables predictoras, de ahí que:

$$Rank(X^T X) \leq Rank(X) < 5$$

Por lo que $Ran(X^T X) < P$, esto hace que sea una matriz de rango incompleto y por propiedades el determinante vale 0. En este mismo sentido, una matriz solo es invertible si su determinante es distinto de 0, ya que:

$$A^{-1} = \frac{1}{det(A)} adj(A^T)$$

Usando el método de regresión por Mínimos Cuadrados para estimar los β's, resulta necesario $(X^T X)^{-1}$, pero no es posible invertir esta matriz cuando existe multicolinealidad, haciendo que el sistema de ecuaciones normales (SEN) no tenga solución única, por lo tanto, no existe una única estimación de los β's.

6.5. Datos faltantes

Según Dagnino (2014), los datos faltantes son los valores no disponibles por diversos factores, que serían útiles o significativos para el análisis de los resultados. Este problema puede ocurrir de forma aleatoria, es decir, que todos los individuos tienen la misma probabilidad de verse afectados, o sistemática, esto es, que tienen una forma repetitiva evidente y se debe a alguna razón especial o algún tipo de sesgo. Cuando faltan datos, se ve alterada la estimación de las variables de respuesta en un modelo de regresión, para esto existen diversos métodos, siendo el más conocido la imputación.

6.6. Partial Least Square (PLS)

El análisis de componentes principales (ACP), permite establecer relaciones existentes entre individuos y variables, a partir de la creación de una matriz $X_{n \times p}$ (Hotelling 1933). Tiene como objetivo describir p variables con un subconjunto q de variables incorrelacionadas entre sí, es decir, reducir la dimensionalidad de los datos (Márquez Ruiz 2017). Al conjunto

q se le conoce como **componentes principales** y es evidente que $q \leq p$.

El PLS, se basa en los conceptos del Análisis de Componentes Principales y Regresión, desarrollados inicialmente por Herman Wold en 1975 (Vega Vilca & Guzmán 2011). El método PLS permite trabajar con datos faltantes, multicolinealidad y mayor número de individuos que variables (GonzÃ¡lez Rojas 2016), de no tener estos problemas en el conjunto de datos, entonces no se considera necesario usar este método.

6.6.1. PLS1

Wold (1975), realizó una modificación al algoritmo NIPALS para obtener componentes basadas en regresión, conocida después como regresión PLS (PLS-R). Esta modificación mantiene el beneficio de trabajar con datos faltantes.

Se conocen dos tipos de PLS: PLS1 y PLS2, el primero es el método PLS cuando se tiene una variable de respuesta y el segundo es cuando se tienen dos o más variables a explicar.
La regresión que se quiere estimar, está dada por:

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + e \tag{6-4}$$

En este modelo de regresión se requiere explicar una sola variable (Y) con base a todas las variables explicativas $X_1, ..., X_p$, que pueden estar altamente correlacionadas (multicolinealidad) y presentar sobredimensionalidad, es decir, $p > n$. Es necesario que la matriz X y el vector Y se consideren estandarizados o centrados. Se buscan componentes ortogonales $t = Xw$ altamente correlacionadas con Y en el espacio de las variables predictoras , tal que se realice la regresión:

$$Y = c_1 t_1 + c_2 t_2 + ... + c_p t_p + Y_H$$

Para luego, mediante el desdoblamiento de las $t = f(x)$ (presentado en la Ecuación 6-11), que son combinación lineal de las X_i gracias al w, estimar el modelo expresado inicialmente en 6-4. Continuando con la misma idea que en el ACP, se busca proyectar cada vector X_i sobre un vector w, sabiendo que $w'w = 1$. Lo anterior, se presenta de manera gráfica mediante la Figura **6-1**.
De aquí, se puede observar un triángulo rectángulo en específico formado por el vector Y, los residuos y_1 y la componente t_1 (Figura **6-1**), en el qué, por el teorema de Pitágoras y despejando t_1 se obtiene:

$$Y^2 = t_1^2 + y_1^2 \quad \Rightarrow \quad Y^2 - y_1^2 = t_1^2$$

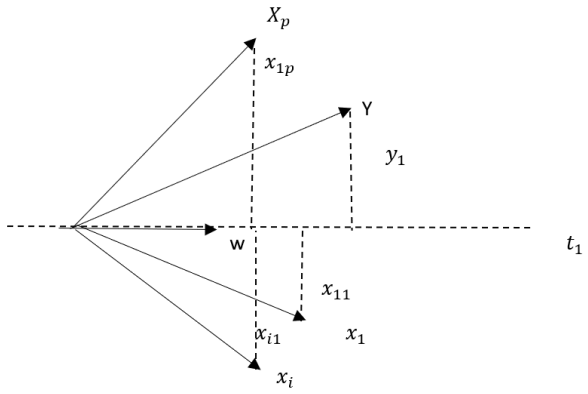

Figura 6-1: Proyección de cada vector X_i e Y sobre un vector w.

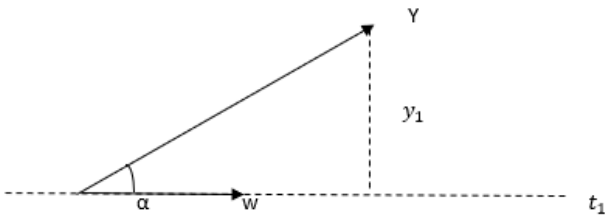

Figura 6-2: Triángulo rectángulo formado por Y, y_1 y t1

Se busca minimizar y_1^2 y maximizar t_1^2, de manera geométrica esto se consigue haciendo que la componente t_1 esté altamente correlacionada con Y, el coseno en el plano geométrico es equivalente a la correlación por lo que si el ángulo de $\alpha = 90°$ entonces el $\cos(\alpha) = 0$ lo que indica que Y es ortogonal a w, es decir, Y es independiente de w. Por otro lado, si $\alpha = 0°$ entonces el $\cos(\alpha) = 1$, es decir, Y está superpuesto sobre w, por lo tanto Y tiene correlación perfecta con w, por lo que, al maximizar el coseno o la correlación es lo mismo que maximizar t_1. De ahí que,

$$t_1^2 = (Cor(t_1, Y))^2 = (Cor(Xw, Y))^2 \tag{6-5}$$

La covarianza entre dos variables estandarizadas se establece como

$$Cov(X,Y) = \sum_{i=1}^{n}(\frac{(X_i-\bar{X})}{\sigma_X} - 0)(\frac{(Y_i-\bar{Y})}{\sigma_y} - 0) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sigma_X \sigma_Y} = Cor(X,Y)$$

Por propiedades de la covarianza y suponiendo una constante k, se puede demostrar que $Cov(kX, Y) = kCov(X, Y)$, aplicando esto en la Ecuación 6-5 se obtiene $t_1^2 = Cor(Xw, Y))^2 = (Cov(Xw, Y))^2 = (w'Cov(X, Y))^2$.

Sea V la matríz de orden $p \times 1$, el vector de covarianzas de X e $Y(V = X'Y)$, entonces

$(w'Cov(X,Y))^2 = (w'V)^2 = w'VV'w$. La función lagrangiana $\varnothing$ que maximiza $w'VV'w$, sujeta a la restricción de ortonormalidad de w, es $\varnothing = w'VV'w - \lambda(w'w - 1)$, por lo tanto, $\frac{\partial}{\partial w}\varnothing = 2VV'w - 2\lambda w = 0$, lo que conlleva a obtener la Ecuación 6-6.

$$VV'w = \lambda w \tag{6-6}$$

De aquí que λ y w son el valor y el vector propio de VV' respectivamente, además VV' es simétrica. Ahora, se multiplica la Ecuación 6-6 a ambos lados por w', resultando:

$$w'VV'w = \lambda \tag{6-7}$$

También se multiplica la Ecuación 6-6 por V' a ambos lados, de tal modo que $V'VV'w - \lambda V'w = 0$, luego de factorizar se obtiene como resultado $(V'V - \lambda)V'w = 0$, donde se puede observar que $(V'V - \lambda) = 0$ o $V'w = 0$, no obstante, se descarta la segunda opción debido a que V es el vector que se busca maximizar y w por su naturaleza, tampoco puede ser cero. Por consiguiente, $(V'V - \lambda) = 0 \Rightarrow \lambda = V'V$

$$\lambda = \|V\|^2 \tag{6-8}$$

De la expresión 6-8 se puede deducir que $\lambda^2 = (V'V)(V'V) = \lambda\|V\|^2$, por lo tanto,

$$\frac{V'}{\|V\|}VV'\frac{V}{\|V\|} = \lambda \tag{6-9}$$

Finalmente, al igualar las ecuaciones 6-7 y 6-9, se obtiene

$$\frac{V'}{\|V\|}VV'\frac{V}{\|V\|} = w'VV'w \tag{6-10}$$

Por consiguiente, el vector w que maximiza la covarianza al cuadrado es:

$w = \frac{V}{\|V\|} = \frac{X'Y}{\|X'Y\|}$

Entonces la primera componente estará determinado por: $t_1 = Xw$ y luego se realiza la regresión de Y sobre t_1.

$Y = c_1t_1 + Y_1 = c_1w_{11}x_1 + c_1w_{12}x_2 + ... + c_2w_{1p}x_p + Y_1$

Donde $c_1 = Y^T t_1$ es el coeficiente de regresión e Y_1 el vector de residuos, se calcula además $P = X^T t1$ que es el coeficiente de regresión de X_j sobre t_1, y se actualiza la matriz X y el vector Y, para poder calcular una segunda componente en caso de que el poder explicativo de la regresión $Y = t_1$ no sea suficiente y se procede, de manera análoga a la primera, a estimar el modelo $Y = t_1 + t_2$.

El desdoblamiento de las componentes $t = f(x)$ para presentar el modelo de regresón PLS con los coeficientes de regresión estimados en función de las variables originales, se realiza a través de la Ecuación 6-11

$$\beta_{PLS} = W(P^T W)^{-1} C \tag{6-11}$$

Este procedimiento se repite hasta encontrar las componentes necesarias, aunque según diversos autores, con las tres primeras es suficiente, ya que de ahí en adelante el poder explicativo del modelo no aumentará significativamente (González Rojas 2016). Víctor González (2014), expone el siguiente algoritmo de PLS1 que resume lo explicado anteriormente:

ALGORITMO PLS1

1. Se parte de $X_0 = X$ e $y_0 = y$
2. Para h=1,2,...,a (a es el rango de X, $a \leqslant p$)
   - *a)* $w_{hi} = X'_{h-1} y_{h-1} / \|y_{h-1}\|$
   - *b)* $w_h = w_{hi} / \|w_{hi}\|$
   - *c)* $t_h = X_{h-1} w_h / w'_h w_h$
   - *d)* $P_h = X'_{h-1} t_h / t'_h t_h$
   - *e)* $X_h = X_{h-1} - t_h P'_h$
   - *f)* $c_h = y'_{h-1} t_h / t'_h t_h$
   - *g)* $u_h = y_{h-1} / c_h$
   - *h)* $y_h = y_{h-1} - t_h c_h$

Cuando se presentan datos faltantes, el algoritmo PLS1 funciona de la misma manera pero teniendo en cuenta el principio de datos disponibles antes de empezar con el paso a) y en los pasos c), d) y f).

Ya que sólo se tiene una variable de respuesta (Y), el algoritmo no busca convergencia por medio de iteraciones. Por otro lado, las coordenadas de los vectores de los ítem *b*, *c*, *d* y *f*

pueden ser calculadas con datos faltantes.

Las propiedades matemáticas de la regresión PLS1 pueden ser consultadas en la página 50 del texto 'Análisis conjunto de múltiples Tablas de datos mixtos mediante PLS' de Víctor González (2014).

6.7. Principio de datos disponibles

Según Ochoa Muñoz (2018), el principio de datos disponibles, consiste en que es posible realizar operaciones entre matrices y/o vectores sin importar que se tenga presencia de datos faltantes, trabajando con los datos disponibles en la matriz y/o vector. Para ejemplificar este principio, sea A una matriz de tamaño 5×3 y, sea B un vector de tamaño 3×1, ambos con datos NA, cuyo producto interno es AB.

$$AB = \begin{pmatrix} a_{11} & a_{12} & NA \\ NA & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & NA & a_{43} \\ a_{51} & NA & NA \end{pmatrix} \begin{pmatrix} b_1 \\ NA \\ b_3 \end{pmatrix} = \begin{pmatrix} a_{11}b_1 \\ a_{23}b_3 \\ a_{31}b_1 + a_{33}b_3 \\ a_{41}b_1 + a_{43}b_3 \\ a_{51}b_1 \end{pmatrix}$$

Cabe resaltar que en caso de que una fila de la matriz A contenga datos faltantes en su totalidad, aún es posible realizar el producto AB, ya que se elimina esta fila y la matriz A reduce su dimensión a 4×3.

7 Metodología

7.1. Santiago de Cali

El municipio de Santiago de Cali es la capital del Departamento del Valle del Cauca, una de las regiones más importantes de Colombia. Está situada en el suroccidente colombiano, a 1.070 metros sobre el nivel del mar; cuenta con 261.7 kilómetros de superficie total y con 2.445.281 habitantes hasta el 2018. Territorialmente, está dividida en 22 comunas en la zona urbana y 15 corregimientos en la zona rural, para un total de 249 barrios según la Alcaldía de Santiago de Cali (DAP 2019).

El clima de la capital vallecaucana es considerado tropical con verano seco según la clasificación climática de Koppen (DAP 2019), con un clima promedio de 24.3 $°C$, la precipitación anual es de 775.4 mm y el promedio de humedad relativa anual es de 77,2 %. Las estaciones secas van de diciembre a febrero y de junio a agosto y las estaciones de lluvias de marzo a mayo y de septiembre a noviembre.

7.2. Medición de la calidad del aire en Santiago de Cali

La calidad del aire en Santiago de Cali, es registrada y evaluada a través de nueve estaciones de monitoreo a lo largo y ancho de la ciudad, estas estaciones registran datos horarios las 24 horas al día, los 365 días del año. La Figura **7-1** ilustra la ubicación geográfica de las estaciones de monitoreo en la ciudad y en la zona rural.

No todas las estaciones de monitoreo registran datos de los mismos contaminantes ni variables ambientales. Según la Tabla **9-1** , sólo la estación de monitoreo Compartir registra mediciones de Ozono troposférico, en conjunto a las variables climáticas (velocidad del viento, dirección del viento, lluvia, humedad relativa, radiación solar, temperatura y presión barométrica). En esta Tabla, los contaminantes y variables meteorológicas están representadas como:
PM: Material particulado, SO_2: Dióxido de Azufre, NO_2: Dióxido de Nitrógeno, O_3: Ozono troposférico, H_2**S**: Ácido Sulfhídrico. **VV**: Velocidad del viento, **DV**: Dirección del viento, **DV**: Dirección del viento, **LL**: Lluvia, **HR**: Humedad relativa, **RS**: Radiación solar, **T**: Temperatura, y **PB**: Presión Barométrica.

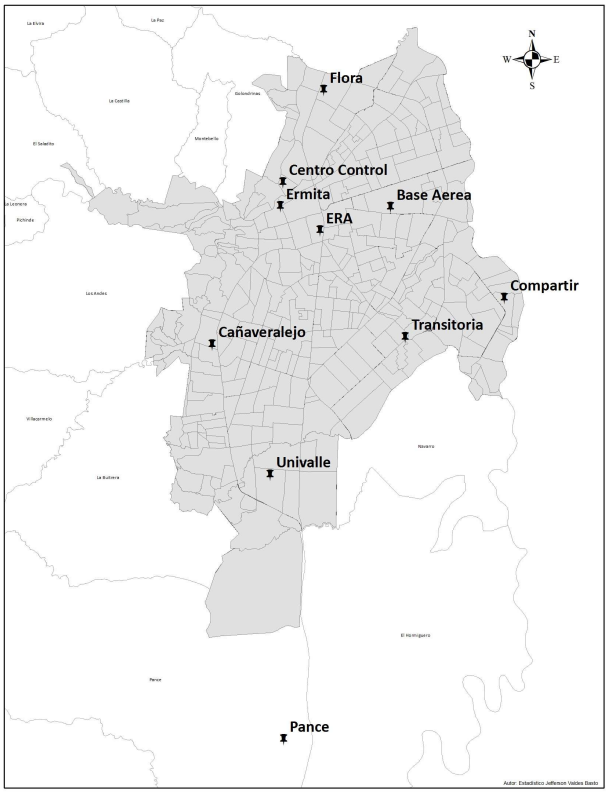

Figura 7-1: Distribución de las estaciones de monitoreo en Santiago de Cali

Tabla 7-1: Variables medidas por cada estación

Estación	**Contaminantes**	**Meteorología**
Base Aérea	$PM_{2,5}$, SO_2, O_3	
Cañaveralejo	PM_{10}, SO_2,	VV, DV, LL
Compartir	PM_{10}, $PM_{2,5}$, O_3	VV, DV, LL, HR, RS, T, PB
ERA	PM_{10}, O_3, H_2S	LL, RS
La Ermita	PM_{10}, SO_2	LL
La Flora	PM_{10}, O_3, H_2O	LL, RS
Pance	PM_{10}, O_3	LL, HR, RS
Transitoria	PM_{10}	
Univalle	$PM_{2,5}$, NO_2, O_3	

7.3. Datos

Santiago de Cali cuenta con nueve estaciones de monitoreo de la calidad del aire, de las cuales cuatro registran datos del Ozono troposférico y variables climáticas (Base Aérea, Pance, Univalle y Compartir) (DAGMA 2018). Sin embargo, sólo la Estación

Compartir registra datos de las variables climáticas: temperatura, humedad relativa, lluvia, velocidad del viento y radiación solar. Las estaciones de monitoreo en Santiago de Cali son supervisadas por la Alcaldía Municipal, a través de entidades como el DAGMA, IDEAM y el SVCASC, por lo tanto brindan datos confiables, ya que se encuentran bajo control y vigilancia constantemente. Estas estaciones aportan información representativa de la zona de la ciudad en la cual está ubicada, no obstante, se encuentran dispersas a lo largo y ancho del municipio, logrando así una representatividad adecuada de los niveles de Ozono troposférico en la capital vallecaucana. Por lo anterior, se decide tomar como base de datos los registros recolectados por esta estación, delimitándola al año 2019 (datos disponibles más actuales al momento de realizar el presente trabajo). Esta base de datos fue compartida directamente por funcionarios del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA), en el transcurso del semestre Agosto-Diciembre 2020.

Los datos registrados en la Estación Compartir para el año 2019 se componen de nueve columnas: Fecha & Hora, Ozono, Temperatura, Humedad, Lluvia, Velocidad de viento, Dirección del viento, Radiación Solar y Presión Barométrica. Se registran un total de 8760 horas, desde la 01:00 del 01 de enero de 2019 hasta las 24:00 horas del 31 de diciembre de 2019. Las definición teórica de las variables está descrita en la Sección 6.3. y 6.4, no obstante la clasificación estadística de estas variables se presenta a continuación:

- **Fecha** & **Hora:** Variable cualitativa nominal, que describe la fecha y la hora en la que se presenta el la información.
- **Ozono:** Variable cuantitativa continua, que describe el valor de Ozono troposférico ($\mu g/m^3$) registrado en Santiago de Cali.
- **Temperatura:** Variable cuantitativa continua, que describe la temperatura (°C) registrada en Santiago de Cali.
- **Humedad relativa:** Variable cuantitativa continua, corresponde al porcentaje de humedad relativa registrada en Santiago de Cali. Al ser un porcentaje, esta variable no puede ser inferior a cero ni superior a 100.
- **Lluvia:** Variable cuantitativa continua, correspondiente a la cantidad de lluvia (mm) registrada en Santiago de Cali.
- **Velocidad del viento:** Variable cuantitativa continua que corresponde la velocidad (en m/s) registrada en Santiago de Cali.
- **Dirección del viento:** Variable cualitativa nominal que corresponde la dirección del viento (Grados).
- **Radiación Solar:** Variable cuantitativa continua, correspondiente al valor de radiación solar (Watt/m^2) registrado en Santiago de Cali.

- **Presión Barométrica:** Variable cuantitativa continua que cuantifica la presión barométrica (mm/Hg) registrada en Santiago de Cali.

Según DAGMA (2018), la presión Barométrica no influye en la creación, dilución o transporte de Ozono troposférico, por lo tanto será omitida en el resto del informe. Por otro lado, el modelo PLS propuesto contiene sólo variables cuantitativas debido que incluir la variable dirección del viento trae consigo una recategorización, lo que conlleva a calcular un modelo para cada categoría e interpretarlo, además de aumentar significativamente la carga computacional, es por ello que dicha variable no será tenida en cuenta. En adelante, las variables serán llamadas: Ozono (O_3), temperatura (Temp), humedad relativa (HR), lluvia (Ll), velocidad del viento (VV) y radiación solar (RS).

Una vez identificados y delimitados los datos a usar, se procede a realizar el análisis exploratorio de datos, apoyándose en las estadísticas descriptivas que se consideren relevantes, que permita conocer y describir más a fondo la situación del Ozono y las variables climáticas en la ciudad durante el período establecido, además, se debe evaluar la relación existente entre ellas.

7.4. Datos Faltantes

Al evaluar la base de datos proporcionada por el DAGMA se evidenció que el porcentaje de datos faltantes para las variables climáticas es exactamente el mismo para todas. Al profundizar más, se determinó que cada vez que falta un dato en una variable climática, lo hará en el resto de variables; esto es debido a que los medidores climatológicos están conectados a una misma fuente y cuando esta falla por factores como cortes de energía o daño en el sistema de aire acondicionado, no se registra información de ninguna variable climática. Lo anterior indica que en la base original, no es posible aplicar el principio de datos disponibles, ya que al no existir datos en una fila de variables predictoras, no hay manera de tomar ese registro en consideración. Cabe resaltar que la metodología usada para ajustar el modelo PLS con datos faltantes está pensada teniendo solo ausencia de datos en las variables predictoras, es por ello que en el Ozono se debe tener la completitud de los registros.

Del conjunto de datos, se obtienen las filas en las que no hace falta ni un solo registro en ninguna variable, obteniendo así una base con datos completos. Ahora, es posible dividir los datos en la variable de respuesta (Ozono) y las variables predictoras (variables climáticas), teniendo en cuenta que: X_1 = Temperatura, X_2 = Humedad, X_3 = Lluvia, X_4 = Velocidad del viento, X_5 = Radiación solar e Y= Ozono.

7.5. Ajuste del modelo de Regresión lineal

Con la matriz de datos X y el vector Y encontrados anteriormente, se ajustará un modelo de Regresión lineal usando Mínimos Cuadrados Ordinarios (MCO), presentado por Montgomery et al. (2006) de la forma:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_5 X_5 + e \tag{7-1}$$

Posterior a ello, se deben evaluar el cumplimiento de los supuestos del MCO: los errores deben seguir una distribución normal, la media de los errores debe ser igual a cero, los errores deben tener varianza constante (Homocedasticidad) y los errores deben ser independientes entre sí. Además, se debe evaluar la bondad del ajuste de la regresión a través del coeficiente de determinación R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7-2}$$

De acuerdo a lo hallado, se determinará si es posible ajustar el modelo de regresión lineal MCO cumpliendo los supuestos y qué valor de R^2 se obtendría. Además se cuantifica la intensidad de la multicolinealidad a través del índice de factor de inflación de la varianza (VIF por sus siglas en inglés) que mide hasta qué punto la varianza se incrementa a causa de la multicolinealidad (Backhaus et al. 2016) y está dado por:

$$VIF = \frac{1}{1 - R_j^2} \tag{7-3}$$

Donde R_j^2 es el coeficiente de determinación de un modelo donde x_j es la variable dependiente y el resto de variables $X's$ son predictoras.

7.6. Ajuste de PLS con datos completos

Con el mismo conjunto de datos usados para ajustar el MCO, se procede a estimar un Modelo de Regresión por Mínimos Cuadrados Parciales (PLS), empleando el Ozono como variable de respuesta y las variables climáticas como predictoras, usando el algoritmo PLS1 propuesto en la Sección 6.6.1 y el algoritmo 'PLSGEN' (incluido en el anexo 11.1) en el programa R (RStudio Team 2021). Una vez hallado el coeficiente de determinación, es posible compararlo con el R^2 obtenido en el modelo de regresión (MCO) y observar sus

diferencias. Cabe resaltar que el PLSGEN usa la librería 'far' (Serge 2015) para ortogonalizar las componentes T y ortonormalizar las W.

De otro lado, es posible determinar un intervalo de confianza para los coeficientes de regresión y observar si se disminuye la variabilidad respecto al MCO. Sin embargo, el intervalo de confianza en PLS no se puede hallar a través de un pivote (como se hace en MCO), ya que aún no se han establecido propiedades distribucionales del PLS. Por lo anterior, se propone usar la metodología Bootstrap (Efron 1992) que consiste en tomar múltiples muestras con reemplazo de una sola muestra aleatoria, es decir, de la base con datos completos se muestrean aleatoriamente 7543 filas con reemplazo, a esta muestra le ajusta el modelo PLS y se obtienen sus coeficientes de regresión, posteriormente se repite el proceso 100 mil veces, obteniendo 100 mil vectores de coeficientes de regresión, lo que permite estudiar su distribución muestral y construir el intervalo de confianza a partir de los percentiles 2.5 y 97.5. El Bootstrap se implementa en el programa R (RStudio Team 2021) usando la librería 'rsample' (Silge et al. 2021).

7.7. Modelo PLS con datos faltantes

Con el objetivo de ajustar un modelo de Regresión PLS con datos faltantes, es indispensable contaminar la base generando registros NA aleatorios en las variables climáticas, para ello se simulan distintos escenarios de contaminación con diversos porcentajes de datos ausentes, especificamente con el 5 %, 10 %, 15 %, 20 %, 30 %, 40 % y 50 %. El proceso de introducir datos faltantes aleatorios en la matriz X se repite en 10.000 veces para para cada porcentaje, partiendo desde una semilla con el fin de realizar una investigación reproducible y mantener la comparabilidad entre cada porcentaje de datos ausentes aleatorios.

El siguiente paso es realizar el método de regresión PLS con datos faltantes para cada escenario, para luego desdoblar los $\beta's$ y mantener su interpretabilidad, además, hallar los R^2 de cada modelo para analizar la varianza y consistencia de estos valores. Es posible realizar un contraste entre el modelo de regresión PLS con y sin datos faltantes, con el fin de determinar qué tan semejantes fueron los resultados obtenidos y así evaluar la eficiencia del modelo con ausencia de datos. Con esto se finaliza la parte procedimental del proyecto y se prosigue a realizar el informe escrito, reportando los resultados obtenidos.

8 Análisis descriptivo de los registros horarios

Con el fin de analizar el comportamiento de cada una de las variables mencionadas anteriormente, se realiza un análisis exploratorio de datos, además se evalúan las correlaciones existentes entre ellas. Los descriptivos como máximo, mínimo, media, etc fueron contrastados con los datos obtenidos en el DAGMA, lo que confirma que estos datos, que en ocasiones pueden ser atípicos, son originales y no corresponden a errores de digitación o manipulación.

8.1. Ozono

Inicialmente se evalúan los datos de Ozono troposférico encontrados en la base de datos de la estación Compartir en Cali, encontrando que el valor mínimo de Ozono registrado en dicha estación es de 2.95 $\mu g/m^3$ y se presenta el 3 de enero a las 6:00 de la mañana, hora en que no hay radiación solar y no se presentaron lluvias. Por otro lado, se tiene que el valor máximo de este contaminante es de 152.81 $\mu g/m^3$, registrado el 23 de enero a las 15:00 horas, donde no hubo presencia de lluvia y se tenía una temperatura de $36.5°C$. Para analizar más a fondo esta variable y entender mejor su comportamiento se presentan los datos de Ozono respecto a la radiación solar y, respecto a los meses y estaciones del año, en las Figuras **8-1** y **8-2** respectivamente. Es evidente que los niveles de Ozono dependen de las horas del día, es decir, las horas en la que la luz solar es mayor los niveles de Ozono registrados son mayores, mientras que en las horas en que la luz solar es nula o baja, es cuando se registran los niveles más bajos de O_3; además, en las horas en las que amanece y anochece se tienen niveles intermedios.

Santiago de Cali tiene dos temporadas durante el año: las estaciones secas y las lluviosas. La primera va desde diciembre hasta febrero y de junio hasta agosto, mientras que la segunda se presenta entre los meses de marzo a mayo y de septiembre a noviembre, siendo este último mes el más lluvioso del año generalmente. Para comprender esto, primero se observa el comportamiento del Ozono a lo largo de cada uno de los meses del año 2019 en la Figura **8-2**. En los primeros meses del año los niveles de O_3 alcanzan picos altos, los cuales van disminuyendo paulatinamente a medida que llegan los primeros meses lluviosos. En junio vuelven a elevarse los niveles máximos de Ozono ya que regresa la temporada seca. Llama la atención que los meses de septiembre y noviembre tienen picos altos, sin

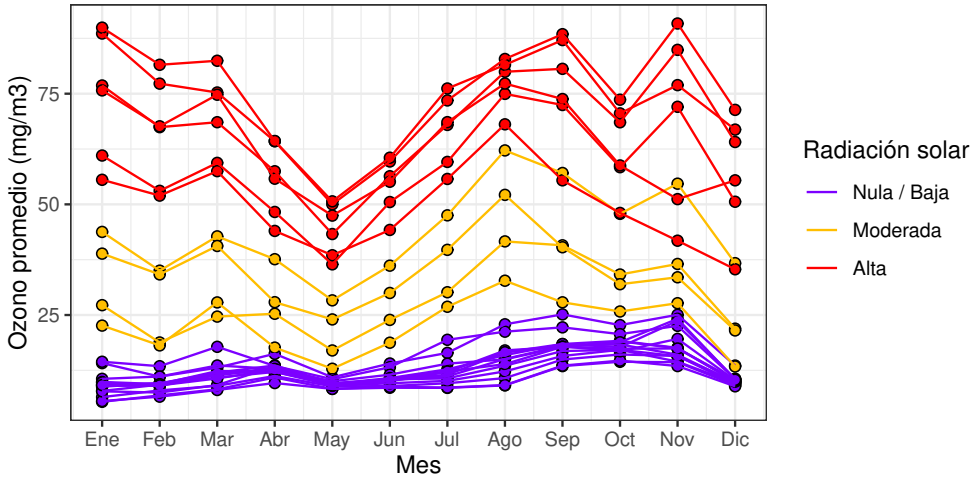

Figura 8-1: Ozono respecto a la radiación solar

embargo, es importante recordar que estos meses presentan gran cantidad de datos faltantes, particularmente noviembre en que sólo se tienen datos de los primeros cinco días, lo que puede explicar este comportamiento.

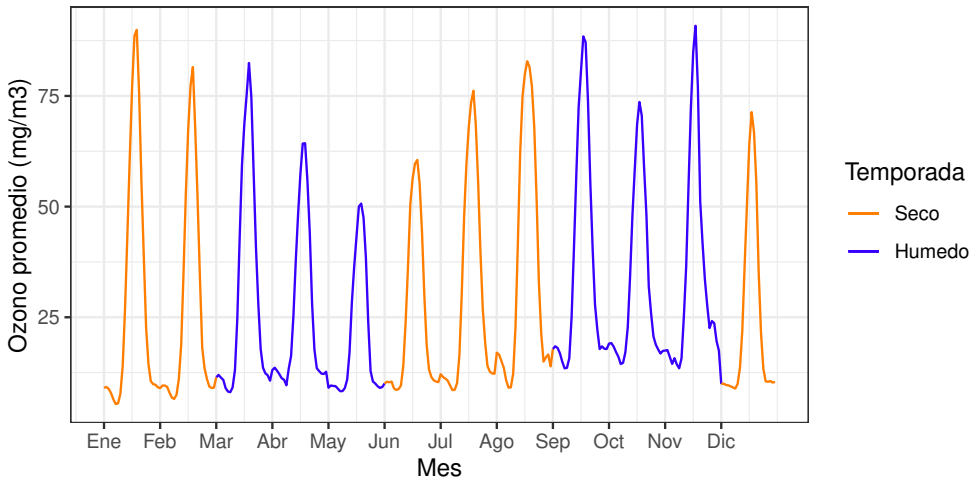

Figura 8-2: Ozono por mes y época del año

8.2. Temperatura

La temperatura promedio en el año 2019 es superior a las temperaturas registradas en los dos años anteriores tal como se evidencia en la Figura **8-3**. Se puede pensar que este aumento se debe al constante incremento en la temperatura de la tierra debido a factores como el

calentamiento global, teniendo un incremento medio de $0.56°C$ y $0.17°C$ para los años 2018 y 2019, respectivamente. Esto se puede observar claramente en los picos altos que se presentan en las horas alrededor de las 14:00 horas, que es cuando la temperatura alcanza su punto máximo, no obstante, en las horas de la noche se aprecian valles en los que la temperatura también aumenta año a año.

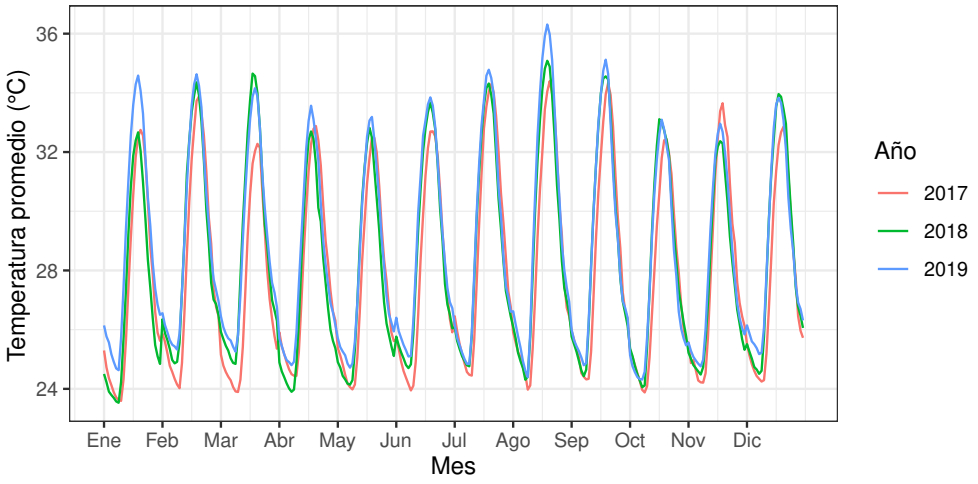

Figura 8-3: Temperatura para los años 2017, 2018 y 2019

8.3. Humedad Relativa

Según datos oficiales, la Humedad relativa promedio en Santiago de Cali es del 77.2 %. Al obtener las estadísticas descriptivas de esta variable, se tienen valores incluso del 12.8 %, lo cual lleva a pensar que estos valores no son coherentes con la realidad climatológica y geográfica de la ciudad, sin embargo, estos valores fueron consultados y confirmados con la entidad a cargo.

En la Figura **8-4**, se puede observar el comportamiento de la humedad relativa durante el 2019 en contraste con los años anteriores, lo que permite apreciar que durante el último año se tienen valores que están por debajo del 20 % en agosto y septiembre que son meses secos y en los que la temperatura tiende a ser más alta y, en consecuencia, la humedad relativa tiende a ser baja. Cabe destacar que en todos los años se aprecia que la mediana de esta variable es más alta en los meses lluviosos y más baja en los meses secos.

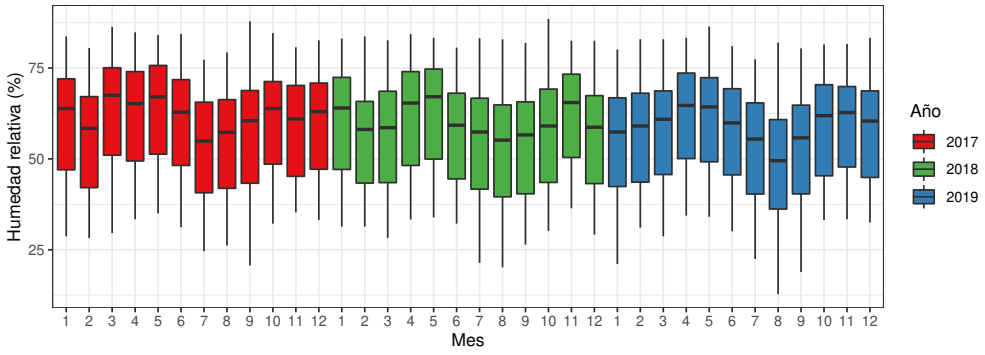

Figura 8-4: Humedad relativa para el 2017, 2018 y 2019

8.4. Lluvia

En la Figura **8-5**, se puede observar la cantidad de horas en cada mes en las cuales hubo y no hubo precipitación. Cabe resaltar que esta variable mide la cantidad de lluvia registrada en una (1) hora, por lo tanto cuando el registro es cero indica que no ha llovido durante esa hora. Evidentemente, en los meses de diciembre, enero, febrero, junio, julio y agosto es cuando menos llueve en Santiago de Cali, mientras que abril, mayo y octubre son los meses con más precipitación en este mismo año. Ya que es el valor acumulado de lluvia durante un periodo de una (1) hora, sin embargo, para ser considerada lluvia, la precipitación debe superar los 0.5 mm, de lo contrario es considerada llovizna. Evidentemente, en los meses secos es cuando menos precipitación hay y en especial agosto el cual es un mes seco, presenta la menor cantidad de precipitación del año 2019. Además, en los meses húmedos es cuando más precipitación hay y el mes de abril, que es un mes húmedo es el mes con más precipitación del año 2019.

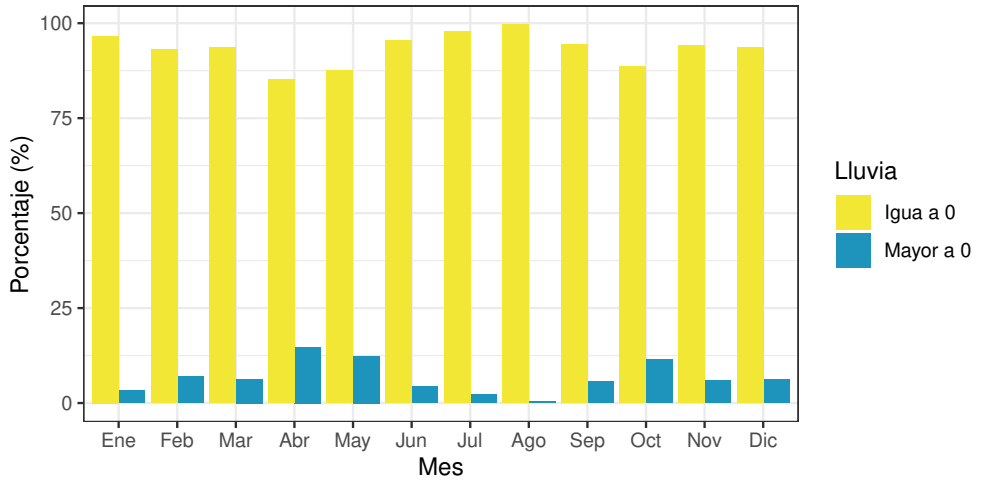

Figura 8-5: Lluvia durante el 2019

8.5. Velocidad del viento

Al revisar las estadísticas descriptivas de la variable, se encontró que hay un total de 4 ceros durante todo el año 2019, este valor se presenta a las 02:00, 03:00, 07:00 y 19:00 horas en distintos días. En la Figura **8-6**, se puede observar que, la velocidad del viento es más alta entre las 15:00 y las 18:00 horas, pero también son más variables, por lo que el viento tiende a moverse más en esas horas y a ser más calmado entre las 22:00 y las 08:00. Lo anterior, puede estar relacionado con la condición topográfica de Santiago de Cali, ya que esta ciudad se encuentra en custodiada por los Farallones de Cali, una formación montañosa que hace parte de la Cordillera Occidental.

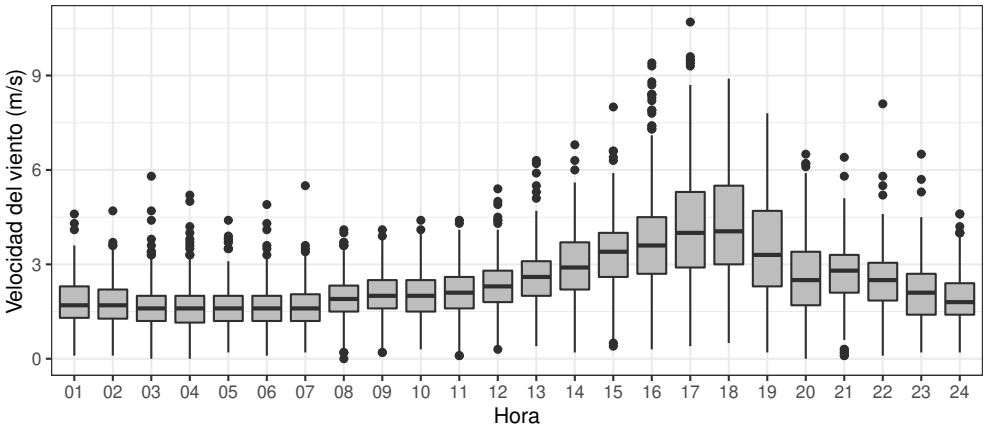

Figura 8-6: Velocidad del viento por horas

8.6. Radiación solar

La radiación solar es una variable indiscutible a la hora de estimar el Ozono, esta influye directamente en la creación de O_3, ya que cuando se presentan valores altos, el Ozono registrado también es alto. En las horas de la noche y la madrugada, hay ausencia total de luz solar, por ende, en estas horas el registro de radiación solar debe ser de 0 W/m^2. Dada la ubicación geográfica de Santiago de Cali, se tiene que la duración del día es aproximadamente la misma durante todo el año, esto es porque se encuentra muy cerca de la línea del Ecuador, por lo tanto, el día inicia entre las 05:00 y las 06:00 horas y la noche inicia alrededor de las 19:00 y las 19:30.

Al comparar los datos de la radiación solar registrados entre 2017 y 2019 en la Figura **8-7**, se observa que el comportamiento para este último no concuerda con lo presentado en los dos años anteriores. Esto se debe a que, según el DAGMA, en el año 2019 hubo un ajuste de la ecuación del sensor de radiación solar, no obstante, con este cambio se registran valores

que incluso pueden llegar a 1000 W/m^2 lo que llama la atención, sin embargo, estos datos fueron confirmados por la entidad. Con el fin de comparar la radiación solar entre los años 2017, 2018 y 2019 se estandariza la variable y se presenta en la Figura **8-8**.

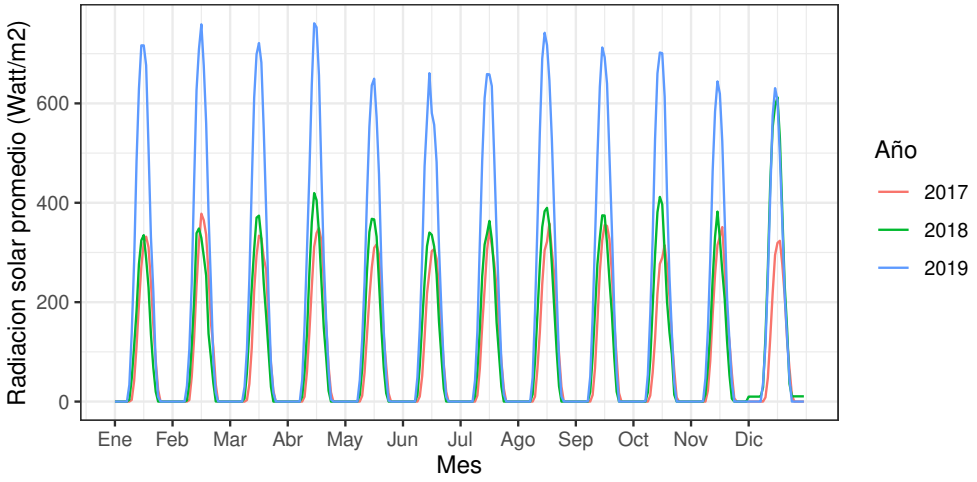

Figura 8-7: Radiación solar para el 2017, 2018 y 2019

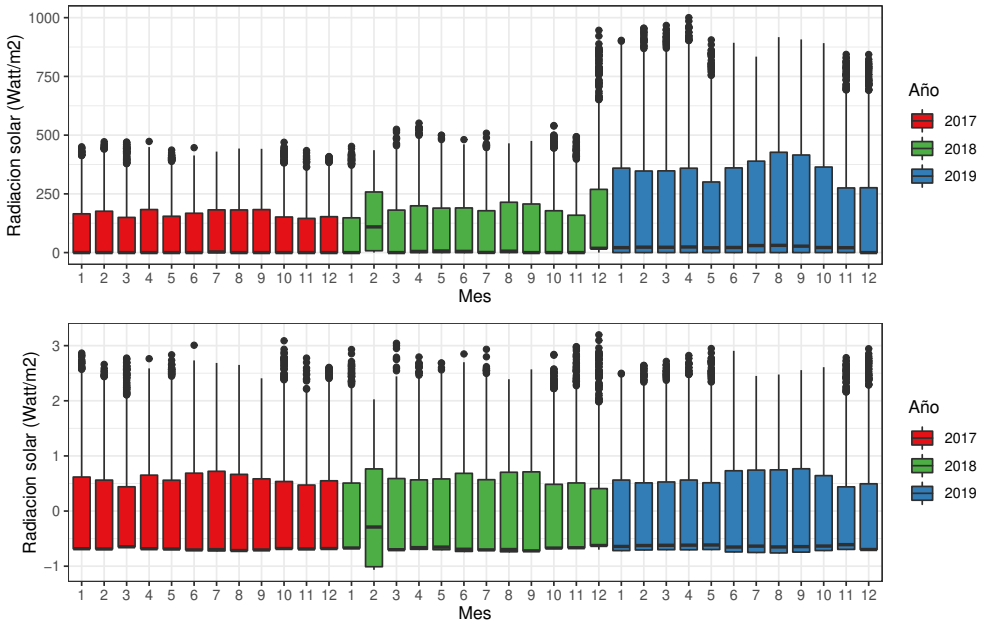

Figura 8-8: Radiación solar para el 2017, 2018 y 2019 estandarizado

8.7. Correlación

Finalmente, es necesario evaluar la relación existente entre las variables, para esto, se presenta la Figura **8-9** en la que se observa que la radiación solar y el O_3 tienen una correlación del 75 %, lo que indica que a medida que incrementa la radiación aumenta la concentración de Ozono en la ciudad de Cali, esto mismo ocurre con la temperatura y el Ozono que tienen una relación del 84 % y, la humedad relativa presenta una correlación inversa de $-0{,}8$ con el Ozono, indicando que con porcentajes bajos de esta variable se tienen niveles altos de O_3. Por otro lado, la relación existente entre temperatura y radiación solar es de 0.71, así mismo, la humedad relativa y la temperatura tienen una relación inversa casi perfecta, lo que indica que posiblemente haya fuerte multicolinealidad.

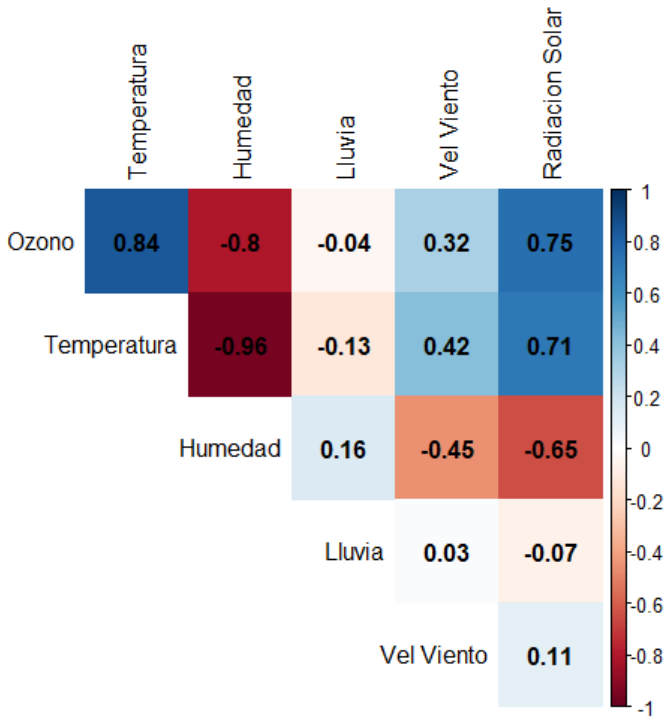

Figura 8-9: Correlación entre variables

8.8. Datos faltantes

Al revisar la base de datos, es evidente la presencia de datos faltantes en cada variable a analizar, esto se debe a distintas situaciones, tales como: falta en el suministro de energía eléctrica en la zona, fallas en los equipos de medición, fallas en el aire acondicionado, hurto de equipos o cableado, mantenimiento o cambio de equipos, entre otros. Por otro lado, debido a las características técnicas de los elementos de medición, cuando no es posible tomar

registros de alguna variable climática, tampoco se toman de las demás y esto se evidencia en la igualdad de la cantidad de datos faltantes para cada variable climática; no obstante, si es posible tomar información del Ozono así no se registren las variables climáticas y viceversa.

En la Tabla **8-1**, se evidencia que el Ozono presenta un porcentaje de datos faltantes del 13,7 %, mostrando que la ausencia de datos es más frecuente que en las variables climáticas. Este contaminante se registra a través del medidor 'Modelo T400 con software Numa View', las recomendaciones del fabricante respecto al mantenimiento de este medidor es que deben realizarse con regularidad desde cada catorce días, hasta cada año dependiendo la duración de cada elemento; esto acarrea que se dejen de tomar mediciones por largos periodos debido a que las calibraciones no se realizan en la ciudad de Cali sino que se debe enviar el medidor a Medellín.

Tabla 8-1: Cantidad de datos faltantes por variables

Variable	No. datos faltantes	No. datos disponibles
O3	1202 (13.7 %)	7558 (86.3 %)
Temp	125 (1.4 %)	8635 (98.6 %)
HR	125 (1.4 %)	8635 (98.6 %)
Ll	125 (1.4 %)	8635 (98.6 %)
VV	125 (1.4 %)	8635 (98.6 %)
RS	125 (1.4 %)	8635 (98.6 %)

Tabla 8-2: Cantidad de horas con datos faltantes por mes

Mes	O3	Temp	HR	Ll	VV	RS	Total
Enero	6	3	3	3	3	3	24
Febrero	5	2	2	2	2	2	17
Marzo	20	1	1	1	1	1	26
Abril	85	8	8	8	8	8	133
Mayo	150	1	1	1	1	1	156
Junio	5	2	2	2	2	2	17
Julio	6	18	18	18	18	18	114
Agosto	11	3	3	3	3	3	29
Septiembre	43	38	38	38	38	38	271
Octubre	161	49	49	49	49	49	455
Noviembre	604	0	0	0	0	0	604
Diciembre	106	0	0	0	0	0	106
Total	1202	125	125	125	125	125	1952

En la Tabla **8-2**, se observa que en noviembre se presenta gran cantidad de datos faltantes en la variable Ozono, esto se debe a que desde el 5 de noviembre hasta el 5 de diciembre el medidor se envió a calibración, una vez se calibró, se registraron mediciones de manera continua, además, las variables climáticas se registraron completas durante todo noviembre y diciembre. En los tres primeros meses del año 2019 la cantidad de datos faltantes son bajos para cada variable al igual que en junio y agosto, no obstante, en septiembre y octubre se presentan gran número de ausencia de datos en las variables climáticas, en comparación con el resto de los meses.

Los 1202 datos faltantes de la variable O_3 se encuentran repartidos en 87 días diferentes, además, se logró determinar que hay un total de 39 días, en los que hay ausencia total de datos de la variable. Como se menciona anteriormente, las ubicaciones de los datos faltantes en las variables climáticas son las mismas, distribuidas en 19 días diferentes.

En el mes de octubre de 2019, no se tomaron datos de ninguna variable desde el día 16 a las 16:00 horas hasta el 18 a esa misma hora, siendo en total 49 datos faltantes de variables climáticas y Ozono. Para esta última no se registraron datos entre el 24 de octubre a las 22:00 horas, hasta el 29 a las 10:00 de manera continua. Por otro lado, en mayo se tienen ausencia de datos de Ozono, de forma continua, desde el día 15 a las 11:00, hasta el 17 a las 9:00 y del 19 a las 13:00, hasta el 23 a las 16:00. El 2 de abril hay datos faltantes en todas las variables, de manera continua, durante ocho horas, además, del Ozono no se registran valores desde el 08 de abril a las 11:00, hasta el 11 a las 13:00.

9 Resultados

La base de datos con las mediciones horarias de Ozono, temperatura, humedad relativa, lluvia, velocidad del viento y radiación solar obtenidas en la Estación de Monitoreo Compartir para el año 2019, cuenta con un total de 8760 filas, al omitir las 1248 horas en las se que tiene al menos un dato faltante, se obtiene una matriz de datos con 6 columnas y 7543 filas. Con esta base de datos se ajusta un modelo del Ozono en función de las variables climáticas usando Regresión por Mínimos Cuadrados Ordinarios y Regresión por Mínimos Cuadrados Parciales, además, se simulan los escenarios de contaminación de datos faltantes sobre esta base con el objetivo de ajustar un modelo PLS a cada uno.

9.1. Mínimos Cuadrados Ordinarios (MCO)

Inicialmente, con el conjunto de datos se ajusta un modelo de Regresión por Mínimos Cuadrados, obteniendo como resultado una estimación de los coeficientes de regresión para cada variable que permiten construir el siguiente modelo:

$$Ozono = \text{-}71.85 + 3.60 * Temp - 0.19 * HR + 1.39 * Ll + \ 0.48 * VV + 0.03 * RS \qquad (9\text{-}1)$$

Para permitir la comparabilidad entre los coeficientes y sus aportes al modelo, se estandarizan las variables, lo que conlleva a obtener coeficientes estandarizados. Cabe resaltar que cuando se realiza este procedimiento, el intercepto es cero; teniendo esto presente, el modelo queda expresado de la siguiente forma:

$$Ozono = 0.503 * Temp - 0.103 * HR + 0.059 * Ll + 0.024 * VV + 0.331 * RS \qquad (9\text{-}2)$$

Para determinar qué tan precisos son las estimaciones de los coeficientes estandarizados del modelo, se calcula el error estándar asociado a cada uno, y con ello es posible hallar un intervalo de confianza, los resultados se presentan en la Tabla **9-1**. Cabe resaltar que el valor p de las variables climáticas incluídas en el modelo tienen valores menores a 0.001 ($valor-p < 0{,}001$), lo que permite inferir que todas las variables predictoras son significativas en el modelo de regresión lineal por mÃnimos cuadrados ordinarios.

Tabla 9-1: Intervalos de confianza para los β's estandarizados del MCO

Variable	Error estándar	LI	LS	Amplitud
Temp	0.02	0.459	0.546	0.087
HR	0.02	-0.143	-0.063	0.080
Ll	0.01	0.048	0.070	0.022
VV	0.01	0.011	0.037	0.036
RS	0.01	0.314	0.348	0.034

9.1.1. Cumplimiento de los supuestos

A continuación, se realiza la evaluación de los cuatro supuestos sobre los errores del modelo MCO (media cero, varianza constante, se distribuyen normal y son independientes entre ellos), con ello es posible evaluar qué tan efectivo es el MCO para el conjunto de datos presentado anteriormente.

Normalidad de los errores

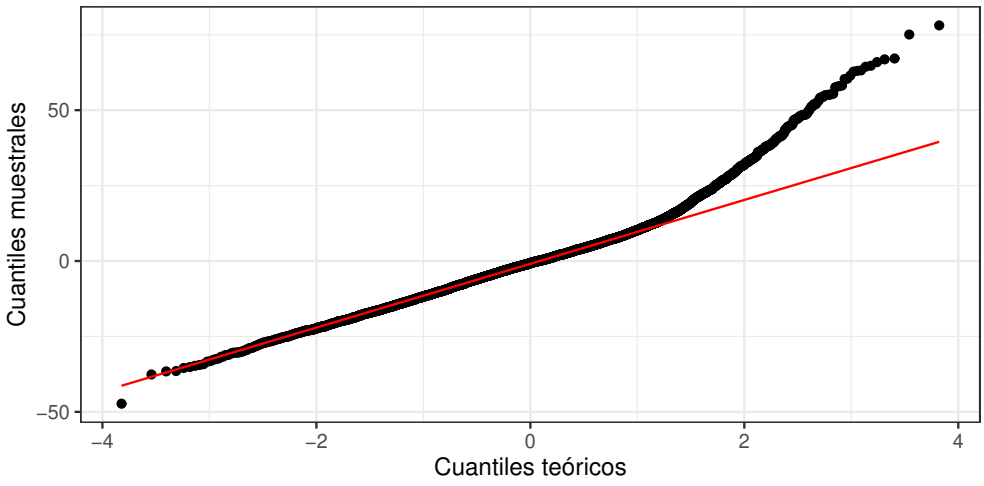

Figura 9-1: Normalidad de los errores - Modelo MCO

Los errores del modelo MCO deben distribuirse normal, con media cero y varianza $\sigma^2 I$, es decir, $e \sim N(0, \sigma^2 I)$. Se realiza una prueba gráfica cuantil-cuantil (Q-Q plot) que compara los valores ordenados de una variable con los cuantiles de una distribución normal (Figura **9-1**), evidenciando que la función empírica de los errores del modelo no son similares a los teoricos, por lo tanto no siguen una distribución normal. Adicional a esto, se decide implementar la prueba Kolmogorov Smirnov para evaluar si los errores se distribuyen normal, obteniendo como resultado que el valor $p < 0.001$, lo que corrobora el resultado

obtenido en la prueba gráfica

Varianza constante en los errores

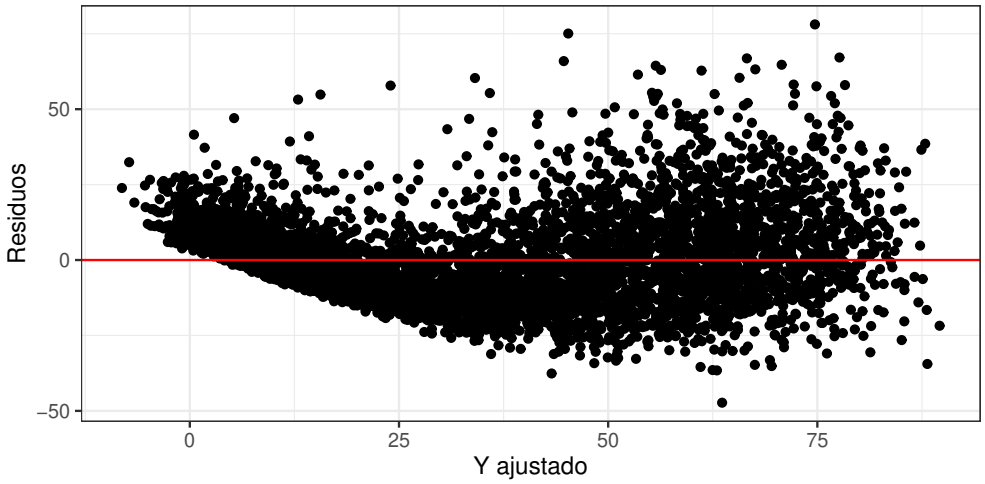

Figura 9-2: Varianza de los errores - Modelo MCO

Para evaluar que la varianza de los errores es constante (homocedasticidad) se implementa la prueba formal llamada Breusch-Pagan, debido a que esta se realiza cuando la muestra es grande y no se requiere conocer la forma funcional de la heterocedasticidad, así mismo, no exige normalidad en los errores (Breusch & Pagan 1980), la hipótesis nula planteada en la prueba es que los errores son homocedásticos. Adicionalmente, se tiene la gráfica **9-2**, donde se evidencia que la varianza de los errores no es constante.

Al implementar el test Breusch-Pagan en R, se obtiene cómo resultado un valor $p < 0.001$, lo que indica que se rechaza la hipotesis nula, por lo tanto, los errores del modelo MCO no presentan varianza constante, bajo cualquier nivel de significancia comunmente usado.

Independencia de los errores

Cabe resaltar que los datos usados en el ajuste del modelo de regresió MCO corresponden a datos horarios, por ende presentan autocorrelación temporal, es decir, la medición obtenida en una hora estará condicionada por la medición de la hora anterior y a su vez condicionará la siguiente hora, por lo que es evidente que no se cumple el supuesto de independencia en los errores.

Multicolinealidad

Para evaluar la presencia de multicolinealidad basta con observar la Figura **8-9** de la Sección 8.7, en ella se observa que las variables temperatura y humedad relativa tienen una correlación inversa casi perfecta, no obstante, se cuantifica la intensidad de la multicolinealidad a través del índice de factor de inflación de la varianza y los resultados obtenidos fueron que para Temp: 15.45, HR: 13.39, Ll: 1.05, VV: 1.38 y RS: 2.28, confirmando así los indicios de mulcolinealidad presentados por las variables temperatura y humedad relativa, ya que sus VIF son mayores a 10. El número de condición obtenido es de 102166.9, lo que indica que la multicolinealidad es grave.

A pesar que con el modelo de regresión por Mínimos cuadrados Ordinarios se obtuvo un R^2 de 0.76, es claro que no es una solución viable ya que no cumple con los supuestos de este modelo, además que se presenta multicolinealidad grave. Por otro lado, este modelo no permite trabajar con datos faltantes bajo ninguna circunstancia, por lo anterior, resulta necesario acudir a otros métodos de regresión para ajustar un modelo a los datos de Ozono troposférico en Cali.

9.2. Mínimos Cuadrados Parciales (PLS) con datos completos

Con el mismo conjunto de datos usados en el modelo de regresión MCO, se ajusta un modelo de regresión por Mínimos Cuadrados Parciales (PLS), usando el algoritmo presentado en la sección 6.6.1. Para realizar el algoritmo en R se crea una función llamada PLSGEN, incluida en los anexos 11.1 del presente trabajo.

Inicialmente, se estandarizan los datos respecto a la media y varianza en cada variable, ahora se buscará el número de componentes a usar en el modelo, para ello se presenta la Figura **9-3**, donde se observa que el cambio más notorio en el coeficiente de determinación R^2 se tiene en la segunda componente, sin embargo, sólo es de aproximadamente 0.02 y en adelante los cambios son muy sutiles. No obstante, con una sola componente, es posible mantener los signos de los coeficientes de correlación y con ello su interpretación, por ende, se determina que es apropiado usar la primera componente.

En el modelo PLS se busca componentes ortogonales t en el espacio de las variables predictoras, de máxima covarianza con la variable de respuesta, como se detalla en la sección 6.6.1. La estimación del coeficiente de la primera componente t es 0.52 y el error estándar es 0.004, posterior a ello se desdobla este coeficiente y se obtiene la estimación de los β's para cada variable, dando como resultado el modelo:

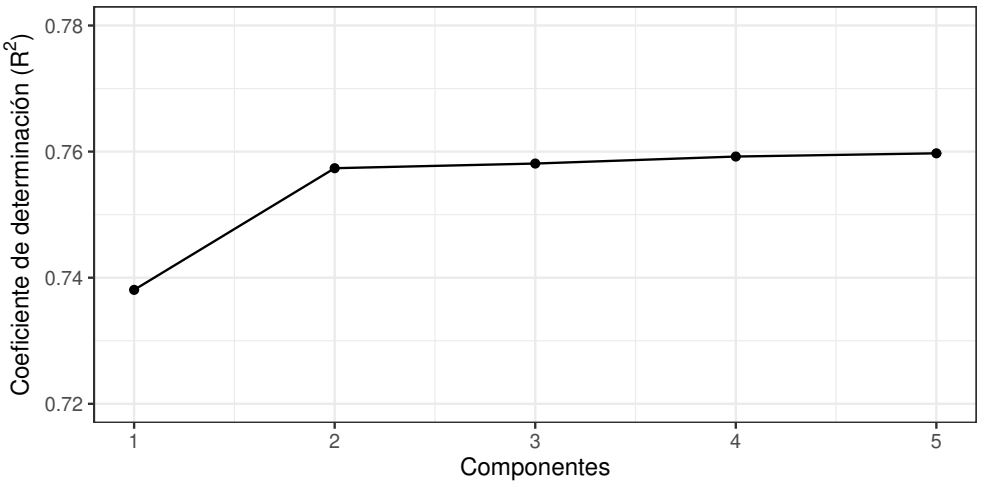

Figura 9-3: R^2 respecto al No. de componentes en PLS con datos completos

$$Ozono = 0.307 * Temp - 0.292 * HR - 0.016 * Ll + 0.118 * VV + 0.275 * RS \quad (9\text{-}3)$$

Las variables temperatura, velocidad del viento y radiación solar tienen coeficientes positivos, lo que indica que tienen relación directamente proporcional en la creación de Ozono troposférico en la ciudad de Santiago de Cali, mientras que para las variables humedad relativa y lluvia se tienen coeficientes negativos, es decir, tienen una relación inversamente proporcional en la producción del O_3. Se tiene que el coeficiente de determinación es R^2 = 0.7382, esto quiere decir que el modelo PLS1 explica el 73.82 % de la variabilidad del Ozono troposférico, de otro lado, este modelo corrige la multicolinealidad generada por la naturaleza de las variables climáticas, lo que permite tener coeficientes de regresión menos variables, además se mantienen los signos de la correlación en todas las variables, mientras que en el modelo MCO el signo del coeficiente de regresión de la variable lluvia es distinto.

Mediante una técnica de remuestreo con reemplazo conocida como Bootstrap, es posible estimar intervalos de confianza para cada coeficiente de regresión PLS a través de los percentiles 2.5 y 97.5 de 100 mil remuestreos realizados, obteniendo como resultado la Tabla **9-2**. Al comparar los intervalos para los coeficientes de regresión del MCO y del PLS, se aprecia que los del PLS presentan menor amplitud, debido a que los estimadores de los coeficientes no sufren por la inflación de varianza debido a la multicolinealidad presentada en las variables predictoras.

Cabe resaltar que la matriz de datos usada en el modelo fue previamente estandarizada, así que los betas presentados en la Ecuación 9-3 son estandarizados. Es posible presentar los coeficientes de regresión en la escala original de cada variable para mantener la

Tabla 9-2: Intervalos de confianza para los β's del PLS con datos completos

Variable	LI	LS	Amplitud
Temp	0.304	0.309	0.005
HR	-0.295	-0.290	0.005
Ll	-0.020	-0.012	0.008
VV	0.112	0.124	0.012
RS	0.272	0.279	0.007

interpretabilidad del modelo, en pocas palabras 'desestandarizar' el beta, multiplicando cada coeficiente por $\frac{s_y}{s_{x_j}}$ y el intercepto se obtiene a través de la ecuación 9-4, teniendo en cuenta que β_j es el coeficiente de regresión 'desestandarizado' de cada variable. Con lo anterior, se obtiene como resultado el modelo presentado en la ecuación 9-5.

$$\beta_{intercepto} = \bar{y} - \sum_{j=1}^{5} (\bar{x}_j * \beta_j) \tag{9-4}$$

$$Ozono = \text{-}14.93 + 2.20 * Temp - 0.54 * HR - 0.37 * Ll + 2.33 * VV + 0.03 * RS \tag{9-5}$$

9.3. Contaminación de la base de datos

Con el objetivo de ajustar un modelo de regresión PLS con datos faltantes, se contamina la base de datos con el 5 %, 10 %, 15 %, 20 %, 30 %, 40 % y 50 % de NA's de manera aleatoria únicamente en el conjunto de variables predictoras, para evaluar su consistencia, se generan 10000 escenarios de contaminación en cada porcentaje y, para mantener la reproducibilidad del ejercicio y la comparación adecuada entre los porcentajes, se fija la semilla en 19970828. Para llevar a cabo esta contaminación, se crea la función en R llamada 'contaminador', incluida en los anexos.

La generación de datos faltantes se realiza de manera aleatoria en todas las columnas, es decir, la matriz X tiene 5 columnas y 7543 filas, equivalente a 37715 valores en total, para realizar la contaminación con el 5 % de datos faltantes, se procede a seleccionar aleatoriamente 1886 valores, en ellos se cambiará el dato correspondiente por "NA", los cuales estarán distribuidos por todas las columnas y no necesariamente de manera equitativa en cada una, de tal modo que una misma fila pueda presentar más de un registro faltante. Se plantea como restricción que, como máximo, se pueden presentar cuatro valores ausentes en una misma fila, ya que si se tienen 5 valores simplemente la fila sería completamente nula, lo que conlleva al problema de principal de la fuente de datos original e imposibilita

usar dicha observación en cualquier modelo de regresión.

Para contaminar la matriz X con el 10 % de datos faltantes, se mantienen los registros contaminados con el 5 % y sobre ellos se contamina el 5 % restante, de tal manera que sea completamente comparable la matriz obtenida inicialmente y la matriz con el 10 % de NA's. Lo mismo se realiza con el 15 %, 20 %, 30 %, 40 % y 50 %, de tal modo que esta última matriz tendrá los registros con 'NA' de los porcentajes anteriores.

9.3.1. PLS con datos faltantes

A cada uno de los escenarios de contaminación, se le ajusta el modelo de regresión PLS con una componente, ya que según lo descrito en la Figura **9-3**, incluir una segunda componente incrementa levemente el R^2 pero no se conservan los signos de las estimaciones. El ajuste del modelo usando el software R se lleva a cabo con la función 'PLSGEN' incluida igualmente en los anexos y cuyo pseudoalgoritmo es presentado en la sección 6.6.1. Una vez desarrollado el proceso de contaminación y el ajuste de un modelo PLS a cada matriz resultante, se obtienen los resultados descriptivos para el R^2, acompañado de un intervalo de confianza construido para el parámetro, a partir de los quantiles 2.5 y 97.5 y presentados en la Tabla **9-3**.

Tabla 9-3: R^2 para las matrices contaminadas con datos faltantes

	5 %	**10 %**	**15 %**	**20 %**	**30 %**	**40 %**	**50 %**
Mínimo	0.113	0.053	0.010	0.010	0.008	0.007	0.004
Q25	0.732	0.721	0.704	0.678	0.479	0.114	0.046
Q50	0.732	0.724	0.709	0.687	0.602	0.333	0.089
Q75	0.733	0.725	0.713	0.693	0.623	0.473	0.202
Máximo	0.737	0.732	0.724	0.710	0.659	0.588	0.469
Media	0.732	0.722	0.700	0.660	0.508	0.299	0.132
Desv. Estandar	0.009	0.024	0.058	0.106	0.180	0.181	0.111
Límite Inferior	0.729	0.712	0.652	0.190	0.052	0.024	0.012
Límite Superior	0.735	0.728	0.718	0.701	0.642	0.544	0.391
Amplitud	0.006	0.016	0.065	0.510	0.591	0.520	0.379

Inicialmente se tiene un porcentaje de 5 % de datos faltantes y paulatinamente se va escalando hasta llegar al extremo del 50 %. Al revisar detalladamente el cuantil 25, se tiene que el R^2 se mantiene por encima de 0.70 hasta el 15 % de datos faltantes, de ahí en adelante sufre un descenso notorio y a partir del 30 % los valores del cuantil descienden en gran medida hasta llegar al 0.046 cuando se tiene ausencia del 50 % de los datos. Se evidencia además que existen muestras desafortunadas donde el R^2 escasamente sobrepasa el 0.10 y puede ser cada vez más bajo a medida que este porcentaje aumenta, sin embargo, son casos

con baja ocurrencia.

Respecto a la variación del coeficiente de determinación, se observa que con porcentajes pequeños de datos faltantes la desviación estándar es baja, mientras que en presencia de 20 %, 30 % y 40 % de datos ausentes la desviación se incrementa. Este fénomeno ocurre ya que al tener mayor completitud de datos, se tiene la mayor información posible de la variable, y aunque se tengan datos faltantes se conserva la suficiente información como para que el R^2 sea alto sin importar la ubicación que se tenga de los NA's; esto mismo ocurre cuando se tienen porcentajes muy altos de datos ausentes, ya que al haber tanta perdida de información poco afecta la posición del dato faltante.

Dado que se tienen 10.000 simulaciones, es posible estimar un intervalo de confianza para los coeficientes de cada variable en cada porcentaje de datos faltantes propuestos, para ello se tienen las Tablas **9-4**, **9-5**, **9-6** y **9-7**. En todas las variables se tiene que el intervalo presenta menor amplitud en el 5 % de NA's y aumenta hasta el 30 %, sin embargo disminuye nuevamente a partir del 40 %. Para ejemplificar este comportamiento, se ilustra el gráfico **11-4**, donde se aprecia la densidad para la estimacón del coeficiente de regresión para la variable Radiación Solar; en este, se evidencia que en porcentajes bajos de datos faltantes, la estimación del coeficiente está por encima de 0.25, no obstante, en el 30 % la cola izquierda ya no se encuentra tan pegada al eje x, indicando que hay más frecuencia de estimaciones inferiores a 0.20 y cuando aumenta al 40 % se hacen más frecuente valores cercanos a cero. Ya en el 50 % la distribución del coeficiente tiende a cero, indicando que la variable pierde significancia en el modelo. Cabe resaltar que este comportamiento se tiene en todas las estimaciones de los coeficientes de las variables, demostrando así que con porcentajes de 20 % o menos de datos faltantes, es posible obtener un buen ajuste del modelo. Las gráficas para las densidades de los demás coeficientes de regresión PLS están incluidas en los anexos.

Es importante aclarar que la amplitud de los intervalos de confianza para los coeficientes de regresión en el modelo PLS con los distintos porcentajes de datos faltantes aleatorios no es comparable con la amplitud del intervalo obtenido en el modelo PLS con datos completos, ya que en el segundo caso se realizó mediante Bootstrap, mientras que en PLS con datos faltantes el intervalo de confianza para los coeficientes de regresión se construyeron a partir de la distribución de de las 10.000 simulaciones en cada uno de los porcentajes de contaminación con datos faltantes aleatorios.

Tabla 9-4: Intervalos de confianza para los $\beta's$ con 5 % de NA's

	5 %		
	2.5	**97.5**	**Amplitud**
Temperatura	0.303	0.306	0.003
Humedad	-0.291	-0.288	0.003
Lluvia	-0.017	-0.014	0.002
Vel Viento	0.115	0.119	0.004
Radiación Solar	0.272	0.275	0.003

Tabla 9-5: Intervalos de confianza para los $\beta's$ con 10 y 15 % de NA's

	10 %			**15 %**		
	2.5	**97.5**	**Amplitud**	**2.5**	**97.5**	**Amplitud**
Temperatura	0.296	0.303	0.008	0.271	0.299	0.028
Humedad	-0.289	-0.282	0.007	-0.285	-0.258	0.027
Lluvia	-0.017	-0.014	0.003	-0.017	-0.013	0.004
Vel Viento	0.112	0.118	0.006	0.103	0.117	0.013
Radiación Solar	0.266	0.273	0.007	0.243	0.269	0.026

Tabla 9-6: Intervalos de confianza para los $\beta's$ con 20 y 30 % de NA's

	20 %			**30 %**		
	2.5	**97.5**	**Amplitud**	**2.5**	**97.5**	**Amplitud**
Temperatura	0.078	0.292	0.214	0.021	0.269	0.247
Humedad	-0.278	-0.075	0.204	-0.256	-0.020	0.236
Lluvia	-0.017	-0.004	0.013	-0.016	-0.001	0.015
Vel Viento	0.030	0.114	0.084	0.008	0.105	0.097
Radiación Solar	0.070	0.263	0.193	0.019	0.242	0.222

Tabla 9-7: Intervalos de confianza para los $\beta's$ 40 y 50 % de NA's

	40 %			**50 %**		
	2.5	**97.5**	**Amplitud**	**2.5**	**97.5**	**Amplitud**
Temperatura	0.010	0.229	0.220	0.005	0.168	0.163
Humedad	-0.219	-0.009	0.210	-0.160	-0.005	0.156
Lluvia	-0.013	-0.001	0.013	-0.009	0.000	0.009
Vel Viento	0.004	0.089	0.086	0.002	0.065	0.063
Radiación Solar	0.009	0.206	0.197	0.004	0.151	0.147

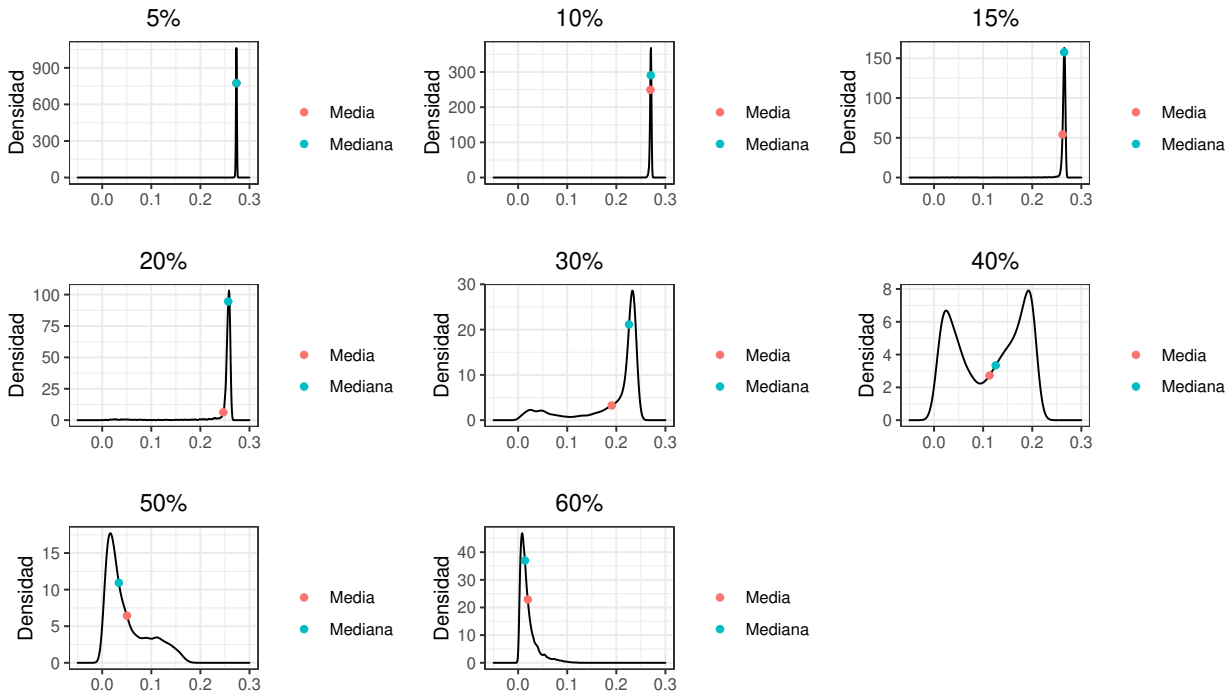

Figura 9-4: Densidad del coeficiente de regresión PLS para la Radiación Solar

10 Conclusiones y recomendaciones

10.1. Conclusiones

Debido a fallas en el sistema eléctrico, aire acondicionado o mantenimiento en los medidores, se presenta ausencia de información en la base de datos recopilada en la Estación de Monitoreo Compartir durante el año 2019. En los registros horarios de las variables, temperatura, humedad relativa, lluvia, velocidad del viento y radiación solar, se tienen datos faltantes de manera sistemática por columna, es decir, cuando en una hora hace falta un dato en una variable climática, también falta en todas las demás variables, lo que implica que se tenga ausencia total de datos en esta hora; adicionalmente, la falta de información suele presentarse en horas seguidas lo que conlleva a tener bloques de datos faltantes en las variables predictoras. Por lo anterior, no es posible usar la estructura de los datos faltantes originales para ajustar un modelo matemático que busque determinar la relación entre el Ozono en función de las variables climáticas registradas en la Estación de Monitoreo.

El modelo de regresión por Mínimos Cuadrados Parciales con una componente ajustado en la base con datos completos, tiene un R^2 de 0.7382, adicionalmente todos los coeficientes de regresión mantienen los signos de la correlación, mientras que en el MCO, la variable lluvia tiene signo contrario, así mismo, los coeficientes del modelo PLS presentan menor amplitud en los intervalos de confianza debido a que se corrige la multicolinealidad generada por algunas variables predictoras, mientras que estas altas correlaciones hacen que las estimaciones de los coeficientes del modelo MCO tengan intervalos mayores y por ende menos precisión.

Dado que no es posible usar los datos ausentes originales debido a su carácter sistemático, se recrean escenarios de simulación que contenga diversos porcentajes de datos faltantes, encontrando que a medida que este porcentaje aumenta y se ajusta el modelo PLS, se incrementa la amplitud de los intervalos de confianza propuestos para los coeficientes de regresión, hasta contaminar la base con el 40 % de datos faltantes, donde la amplitud del intervalo nuevamente empieza a disminuir pero la estimación tiende a cero, indicando que las variables pierden significancia en el modelo. De otro lado, el coeficiente de determinación sufre un descenso moderado hasta el 15 % de datos faltantes, de ahí en adelante este coeficiente disminuye considerablemente, lo que conlleva a que el modelo PLS ajustado a los datos de Ozono en función de las variables climáticas en la Estación Compartir para el año

2019 mantiene buenas métricas hasta el 15 % de datos faltantes aleatorios en las variables predictoras.

10.2. Recomendaciones

Con el fin de complementar esta investigación, se recomienda realizar estudios con datos posteriores al año 2019, teniendo en cuenta que las mediciones horarias para las variables climáticas y Ozono troposférico tienen dependencia temporal, lo que abre paso a un modelo PLS con estructura de autocorrelación, en la cual se implementa una matriz de datos rezagada.

Se recomienda además estudiar la influencia de variables cualitativas como dirección del viento en el modelo PLS y evaluar la posible afectación del modelo con datos faltantes al tener presencia de este tipo de variables. Adicionalmente, se podría evaluar algunos supuestos que permitan agregarle eficiencia al modelo y que garanticen que las estimaciones de los coeficientes sean insesgadas, es decir, la esperanza de los errores es igual a cero, de otro lado, que el modelo tenga la misma capacidad predictiva en el rango de las variables predictoras, esto es, varianza constante en los errores. Además, se podría evaluar la distribución Skew Normal en los residuos del modelo PLS y revisar un posible ajuste en el modelo para corregir la asimetría

Se recomienda también evaluar la viabilidad de usar métricas como la distancia de Cook para determinar el cambio en la estimación de los coeficientes de regresión obtenidos en modelo PLS con datos completos y los coeficientes de regresión obtenidos con datos faltantes.

Se recomienda

Bibliografía

Abdul-Wahab, S. A., Bakheit, C. S. & Al-Alawi, S. M. (2005), 'Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations', *Environmental Modelling & Software* **20**(10), 1263–1271.

Al-Shammari, E. T. (2018), 'Towards an accurate ground-level ozone prediction.', *International Journal of Electrical & Computer Engineering (2088-8708)* **8**(2).

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016), *Multivariate analy semethoden*, Springer.

Balling, R. C. & Cerveny, R. S. (2005), *Winds and Wind Systems*, Springer Netherlands, Dordrecht, pp. 813–819.

Breusch, T. S. & Pagan, A. R. (1980), 'The lagrange multiplier test and its applications to model specification in econometrics', *The review of economic studies* **47**(1), 239–253.

DAGMA (2018), *Boletín Mensual de Calidad del Aire de Santiago de Cali. período de Análisis: Marzo de 2018.*

Dagnino, J. (2014), 'Datos faltantes (missing values)', *Rev Chil Anest* **43**, 332–334.

DAP (2019), *Cali en Cifras 2018-2019*.
URL: *https://www.cali.gov.co/planeacion/*

de Ambiente Vivienda y Desarrollo Territorial, M. (2010), 'Resolución 610 de 2010. por la cual se modifica la resolución 601 del 4 de abril de 2006.', *Diario Oficial No. 47.672 de 6 de abril de 2010* .

de Ambiente Vivienda y Desarrollo Territorial, M. (2017), 'Resolución 2254 de noviembre 01 de 2017. por la cual se adopta la norma de calidad del aire ambiente y se dictan otras disposiciones'.

Efron, B. (1992), Bootstrap methods: another look at the jackknife, *in* 'Breakthroughs in statistics', Springer, pp. 569–593.

EPA (2003), 'Bueno arriba alto malo cerca - ¿qué es el ozono?', *Environmental Protection Agency EE.UU* .

Friedman, R., Milford, J., Rapoport, R., Szabo, N., Harrison, K., Van Aller, S., Niblock, R. & Andelin, J. (1988), 'Urban ozone and the clean air act: Problems and proposals for change', *Staff Paper, Office of Technology Assessment* p. 79.

Gómez, I. C. (2016), Estudio de la concentración de compuestos orgánicos volátiles, óxidos de nitrógeno y ozono en el núcleo urbano de la ciudad de Cartagena y evaluación de la exposición de la población, PhD thesis, Universidad de Murcia.

González Rojas, V. (2016), 'Inter-battery factor analysis via pls: The missing data case', *Revista Colombiana de Estadística* **39**(2), 247–266.

GonzÃ¡lez Rojas, V. M. (2016), 'Inter-Battery Factor Analysis via PLS: The Missing Data Case', *Revista Colombiana de EstadÃstica* **39**(2), 247–266.

Hotelling, H. (1933), 'Análisis de un complejo de variables estadísticas en componentes principales.', *Revista de psicología educativa* **24**(6), 417.

IDEAM (2014), 'Radiación solar - ideam'.
URL: *http://www.ideam.gov.co/web/tiempo-y-clima/radiacion-solar*

Márquez Ruiz, C. (2017), Modelo de regresión pls, Tesis de grado en estadística, Universidad de Sevilla.

McArthur, L. J. B. (2005), *Solar Radiation*, Springer Netherlands, Dordrecht, pp. 667–673.

Montgomery, D., Peck, E. & Vining, G. G. (2006), 'Introducción al análisis de regresión lineal', *México: Limusa Wiley* .

Naranjo, R. & Ortíz, A. (2016), 'Estimación del nivel de ozono troposférico en el aire a partir de un índice de temperatura-humedad, de la radiación solar y del nivel de dióxido de nitrógeno utilizando análisis de regresión con datos funcionales en la ciudad de santiago de cali', *Proyecto de grado Universidad del Valle* .

Ochoa Muñoz, A. F. (2018), 'Análisis de correspondencias múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo nipals (acmpdd)', *Universidad del Valle* .

Oliver, J. E. (2005), *Relative Humidity*, Springer Netherlands, Dordrecht, pp. 617–618.

OMS (2003), 'Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide', *Organización Mundial de la Salud OMS* .

Organization, W. M. (1993), *Atlas Internacional de Nubes*.

Rodríguez, R., Capa, a. & Adelaida, P. (2004), *Meteorología y Climatología*.

RStudio Team (2021), *RStudio: Integrated Development Environment for R*, RStudio, PBC, Boston, MA.
URL: *http://www.rstudio.com/*

Sarochar, H. E. (2014), 'Introducción a la meteorología general', *Facultad de Ciencias Astrónomicas y Geofísicas Universidad Nacional de la Plata* .

Şen, Z., Altunkaynak, A. & Özger, M. (2006), 'Space-time interpolation by combining air pollution and meteorologic variables', *Pure and Applied Geophysics* **163**(7), 1435–1451.

Serge, D. J. G. (2015), *far: Modelization for Functional AutoRegressive Processes*. R package version 0.6-5.
URL: *https://CRAN.R-project.org/package=far*

Shaver, C. L., Cass, G. R. & Druzik, J. R. (1983), 'Ozone and the deterioration of works of art', *Environmental science & technology* **17**(12), 748–752.

Silge, J., Chow, F., Kuhn, M. & Wickham, H. (2021), *rsample: General Resampling Infrastructure*. R package version 0.1.0.
URL: *https://CRAN.R-project.org/package=rsample*

Straif, K., Cohen, A., Samet, S. et al. (2017), 'Air pollution and cancer'.

Tenenhaus, M. (1998), *La régression PLS: théorie et pratique*, Editions technip.

Trapasso, L. M. (2005), *Temperature Distribution*, Springer Netherlands, Dordrecht, pp. 711–716.

Tucker, D. (2005), *Precipitation*, Springer Netherlands, Dordrecht, pp. 574–576.

Vega Vilca, J. C. & Guzmán, J. (2011), 'Regresión pls y pca como solución al problema de multicolinealidad en regresión múltiple', *Revista de Matemática: Teoría y Aplicaciones Vol. 18 Núm. 1 2011* .

Velázquez de Castro, F. (2003), Modelización y análisis de las concentraciones de ozono troposférico, PhD thesis, Universidad Complutense de Madrid, Servicio de Publicaciones.

Wackter, D. & Bayly, P. (1988), 'The effectiveness of emission controls on reducing ozone levels in connecticut from 1976 through 1987'.

Wold, H. (1975), 'Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach', *Journal of Applied Probability* **12**(S1), 117–142.

11 Anexos

11.1. Función en R para ajuste del modelo de regresión PLS

```
PLSGEN<-function(X,Y,NC){

  W <- matrix(0 , ncol(X) , NC)
  rownames(W)<- colnames(X)
  colnames(W)<- paste(1:NC,"Componente")

  WN = W

  t <- matrix(0 , nrow(X) , NC)
  colnames(t)<- paste(1:NC,"Componente")

  P <- matrix(0 , ncol(X) , NC)
  rownames(P)<- colnames(X)
  colnames(P)<- paste(1:NC,"Componente")

  C <- matrix(0 , 1 , NC)
  colnames(C)<- paste(1:NC,"Componente")

  U <- matrix(0 , nrow(Y) , NC)
  colnames(U)<- paste(1:NC,"Componente")

  B <- matrix(0 , ncol(X) , NC)
  rownames(B)<- colnames(X)
  colnames(B)<- paste(1:NC,"Componente")

  for(h in 1:NC){

    ##############################################################################
```

```
### Calcular W = Cor(X,Y) ----------------------------------------------

for(j in 1:ncol(X)){

  PDD<-na.omit(cbind(X[,j],Y))
  W[j,h]<-(t(PDD[,1])%*%PDD[,2])/(PDD[,2]%*%PDD[,2]) } # (X'Y)/Y'Y

### Ortonormar el Vector W ---------------------------------------------

if(any(!is.finite(X))){
  WN[,h]<-orthonormalization(W[,1:h], basis = F, norm=T)[,h]
}else{
  WN[,h]<-W[,h]/sqrt(sum(W[,h]^2)) }

##############################################################################
### Componente T = XW --------------------------------------------------

for(j in 1:nrow(X)){

  PDD    <- na.omit(cbind(X[j,],WN[,h]))
  t[j,h] <- PDD[,1]%*%PDD[,2]/(PDD[,2]%*%PDD[,2]) } # (XW)/W'W

if (any(!is.finite(X))){
  t[,h]<-orthonormalization(scale(t[,1:h], center=T, scale=F),
                            basis = F, norm=F)[,h] }

##############################################################################
### Coeficiente de Regresion X ~ T -------------------------------------

for(j in 1:ncol(X)){

  PDD    <- na.omit(cbind(X[,j],t[,h]))
  P[j,h] <- t(PDD[,1])%*%PDD[,2]/(PDD[,2]%*%PDD[,2]) } # (X'T)/T'T

##############################################################################
###-----------------------------------------------------------------------

X<- X - t[,h]%*%t(P[,h])        # Actualizar X

PDD<-na.omit(cbind(Y,t[,h]))
```

```
    C[,h]<- t(PDD[,1])%*%PDD[,2]/(t(PDD[,2])%*%PDD[,2]) # Cof Regr Y~T

    U[,h]<- Y%*%C[,h]^-1            # Componente Y

    Y<- Y - t(C[,h]%*%t[,h])        # Actualizar Y

    B[,h]<- WN[,1:h]%*%solve(t(P[,1:h])%*%WN[,1:h])%*%C[,1:h] }

  return(list("W"=WN , "T"=t , "P" =P , "C"=C , "U"=U , "Beta Des"=B))

}          # PLS con Correciones
```

11.2. Función en R para contaminar con datos faltantes

```
Contaminador<-function(PCTNA,X,S){

  TDatos <-dim(X)[1]*dim(X)[2] # Total de datos
  CNAT   <-round(TDatos*PCTNA) # Total NA  para el total de datos
  MM<-rep(1:dim(X)[1],4)       # "MARCO MUESTRAL"

  set.seed(S)                  # Semilla para la contaminacion
  FS<-sample(MM,CNAT) ; FS<-FS[order(FS)] # Filas seleccionadas

  Cant<-table(FS) ;   CS<-c() # Cantidad NA por cada Fila
  SFS<-FS[!duplicated(FS)]    # Semilla conservacion %

  for(i in 1:length(Cant)){

    set.seed(S+SFS[i])
    CS<- c(CS,sample(1:5,Cant[i])) } # Columnas Seleccionadas


  PosCont<-cbind(FS,CS)            # Posicionaes para contaminar
  colnames(PosCont)<-c("Filas","Colum")

  Xcon<-X

  for(i in 1:length(PosCont[,1])){
```

```
    Xcon[PosCont[i,1], PosCont[i,2]]<-NA }

  return(list("Posi Cont"=PosCont, "Xcon"=Xcon, "Cant NA"=CNAT))
} # Contaminador de NA
```

11.3. Densidades de las estimaciones de los coeficientes de regresión PLS

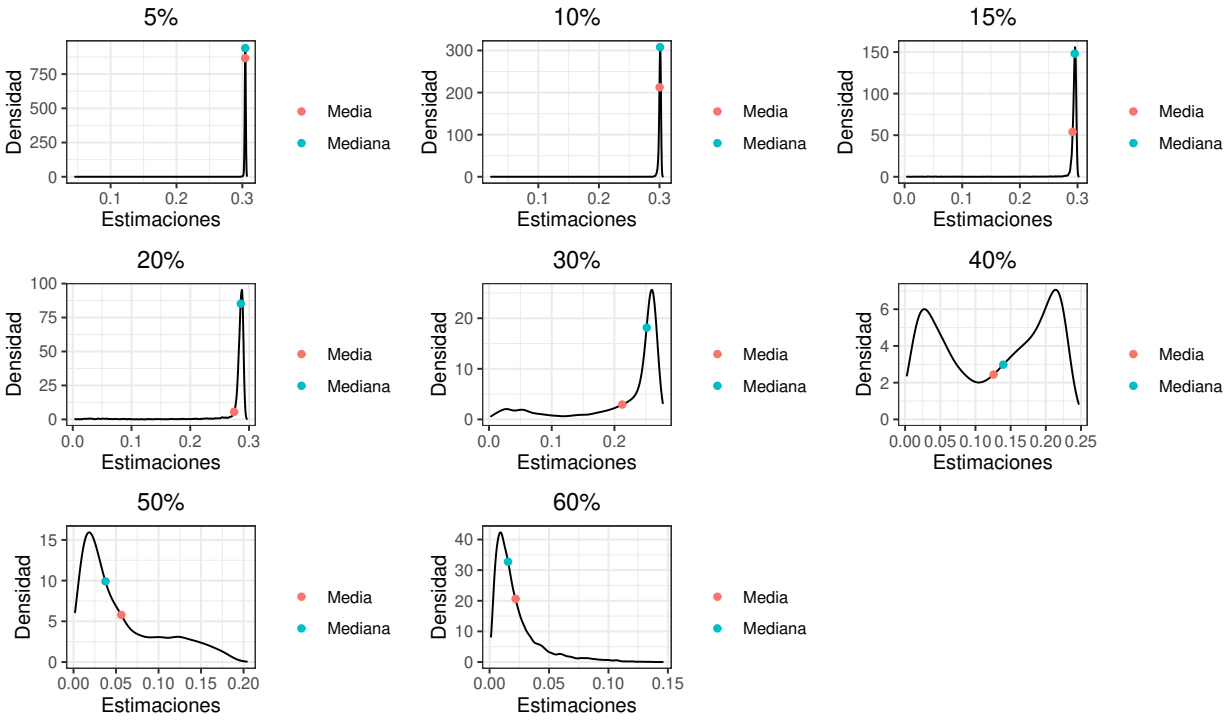

Figura 11-1: Densidad del coeficiente de regresión PLS para la temperatura

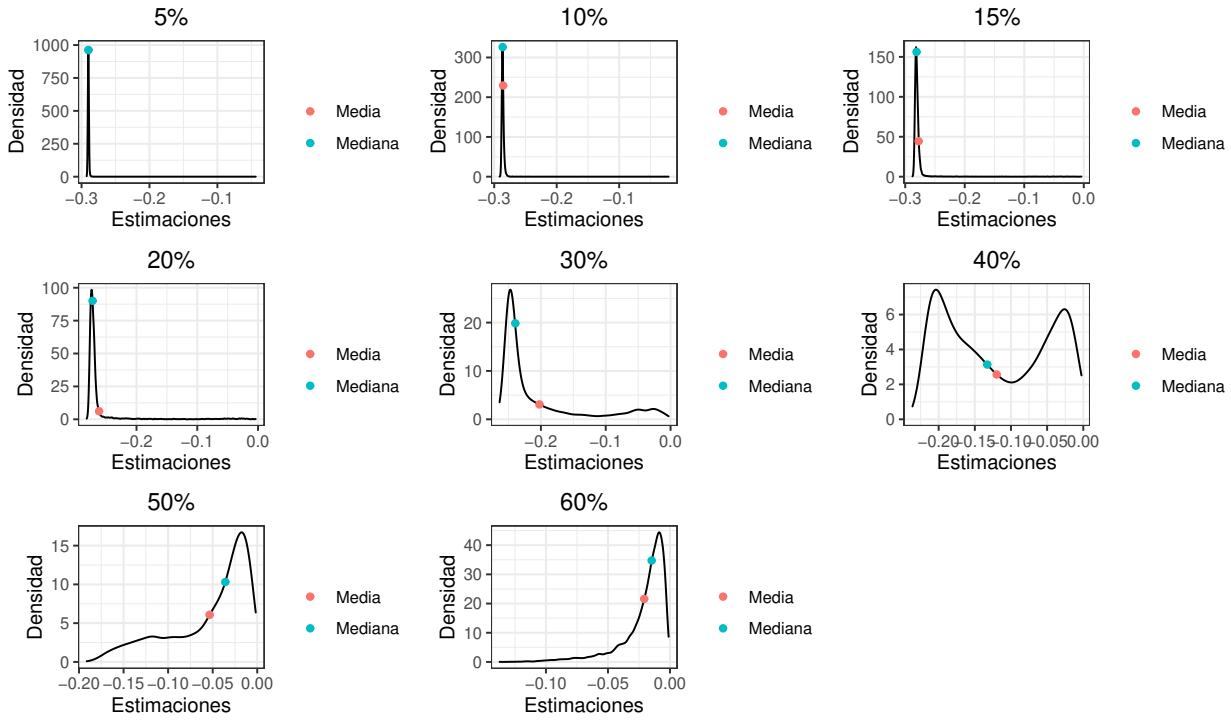

Figura 11-2: Densidad del coeficiente de regresión PLS para la humedad relativa

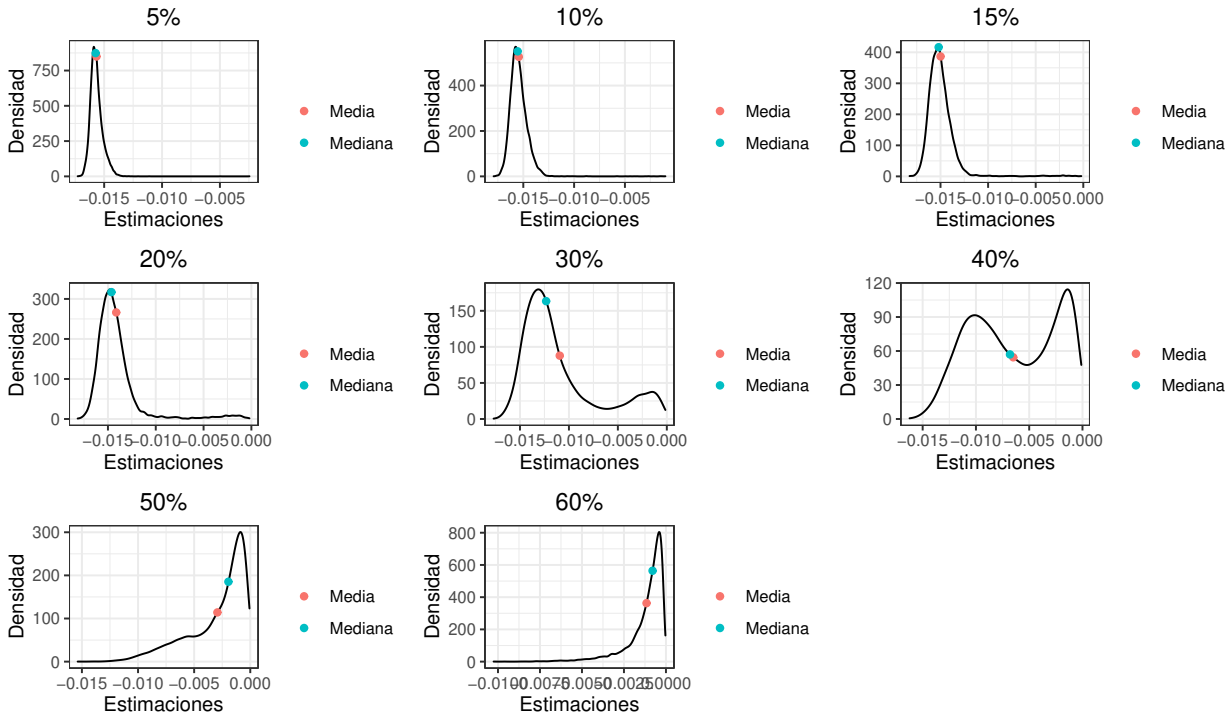

Figura 11-3: Densidad del coeficiente de regresión PLS para la lluvia

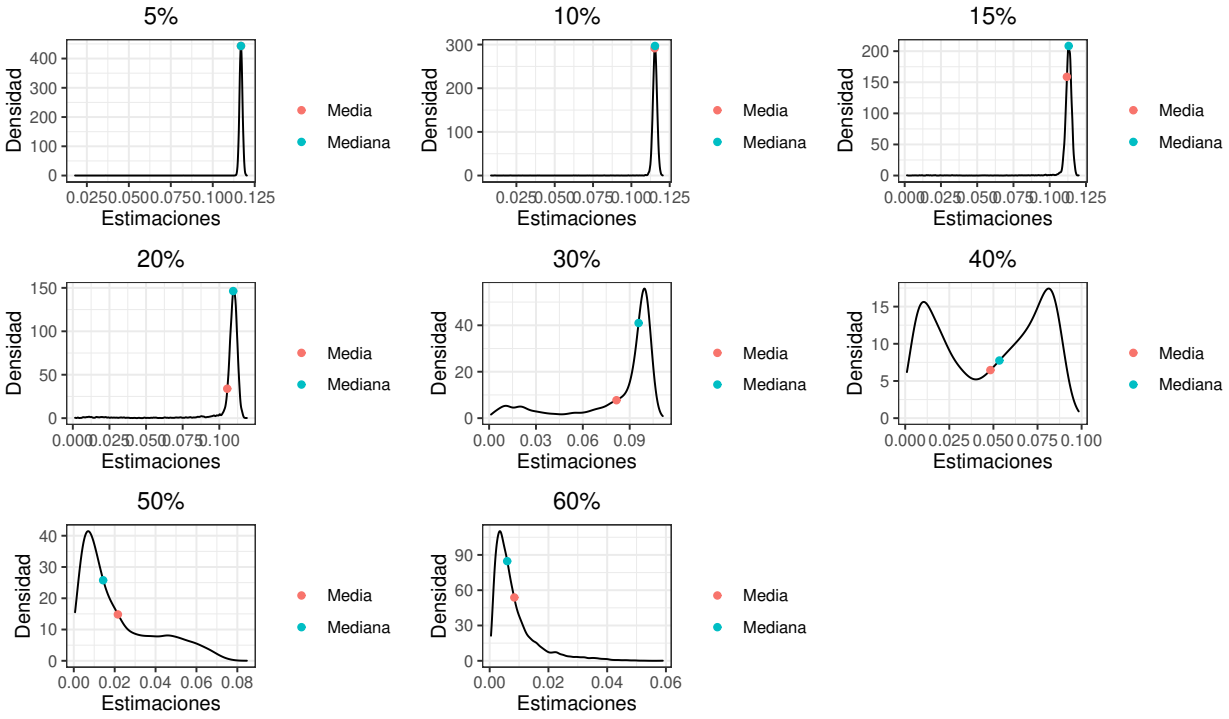

Figura 11-4: Densidad del coeficiente de regresión PLS para la velocidad del viento