SERIES DE TIEMPO Y RANDOM FOREST REGRESSION EN EL PERIODO 2015 - 2019, MODELAMIENTO DE LA TEMPERATURA EN BOGOTÁ D.C.

TIME SERIES AND RANDOM FOREST REGRESSION IN THE PERIOD 2015 - 2019, TEMPERATURE MODELING IN BOGOTÁ D.C.

Jhon Alexander Fonseca Garzón

Especialista en Estadística Aplicada Fundación Universitaria Los Libertadores.
Ingeniero Industrial.
Bogotá, Colombia.
Contacto jafonsecag@libertadores.edu.co
Edgar Javier López Moreno.

RESUMEN

En este documento presentamos un análisis predictivo de la temperatura con datos capturados por la Red de Monitoreo de Calidad del Aire de Bogotá, utilizando técnicas estadísticas y de machine learning, con la aplicación de estas técnicas y de otras técnicas descriptivas han demostrado las interacciones históricas y futuras que tienen las variables meteorológicas, material particulado, gases de efecto invernadero y contaminantes, con la temperatura. Nuestros resultados demuestran como estas relaciones se dan a lo largo del tiempo y el poder predictivo que tienen, además del uso de algoritmos no supervisados para determinar similitudes en las series de tiempo, análisis estadísticos para determinar los impactos que tiene la temperatura con relación a las demás variables. Nuestros resultados presentan una visión diferente ante el entendimiento del clima, sus interacciones y la problemática medioambiental y la calidad del aire en la ciudad de Bogotá.

Palabras clave: *Temperatura, Modelo, Machine Learning, Pronostico, Calidad del aire*

ABSTRACT

In this document we present a predictive analysis of temperature with data captured by the Bogotá Air Quality Monitoring Network, using statistical and machine learning techniques. The application of these techniques and other descriptive techniques have demonstrated the historical and future interactions that meteorological variables, particulate matter, greenhouse gases and pollutants have with temperature. Our results show how these relationships occur over time and the predictive power they have, in addition to the use of unsupervised algorithms to determine similarities in time series, statistical analysis to determine the impacts that temperature has in relation to other variables. Our results present a different view of the understanding of climate, its interactions and the environmental problems and air quality in the city of Bogotá.

Keywords: *Temperature, Model, Machine Learning, Forecast, Air Quality*

INTRODUCCIÓN

El presente artículo hace referencia a las interacciones entre algunas variables meteorológicas en la ciudad de Bogotá. El efecto invernadero y el cambio climático es una realidad que afecta a los seres vivos en el planeta, por ello la necesidad de integrar y analizar dichos temas. Este artículo pretende contribuir mediante el uso de modelos estadísticos de Series de Tiempo y la técnica de Machine Learning Random Forest Regression, modelar la temperatura y sus interacciones con (1) calidad del aire (2) cambio climático. A lo largo del tiempo en la ciudad de Bogotá se han llevado a cabo largos estudios sobre variables que tienen gran influencia en el comportamiento de la temperatura. La Red de Monitoreo de Calidad del Aire de Bogotá - RMCAB ha venido recolectando información sobre la concentración de material particulado de gases contaminantes en la atmosfera que afectan directamente en la salud de la población,

así mismo el clima es uno de los factores ambientales que impacta en diversa forma y nivel la vida del ser humano, según el Instituto de Hidrología Meteorología y Estudios Ambientales – IDEAM el clima es el conjunto fluctuante de las condiciones atmosféricas y se describe a partir de variables como la temperatura y la precipitación. Sin embargo, los bogotanos desconocen su realidad más cercana y como estas variables interactúan y que importancia tienen. En este sentido se desea aportar a la ciudad de Bogotá, un modelo predictivo que permita entender mejor el comportamiento de la temperatura, aplicando (1) metodología Box Jenkins (2) técnica Machine Learning, Random Forest Regression a las series de tiempo, con el fin de comprender mejor las interacciones y/o comportamientos en el efecto invernadero y/o la calidad del aire con la temperatura registrada en la ciudad de Bogotá.

REFERENTES TEÓRICOS

Caracterización climática

La temperatura es un elemento constitutivo del clima, es el indicador de la cantidad de energía calorífica acumulada en el aire. En el año 2017 se registró en Bogotá la temperatura más alta con 25.1 grados Celsius, según el IDEAM esto se dio por la interacción de diferentes variables como disminución de la humedad, vientos fuertes, poca nubosidad y alta radiación solar. En el mismo año, el Instituto presento una proyección de la temperatura media del aire frente al cambio climático en Bogotá, con un horizonte de tiempo a 2100 utilizando los diferentes RCP (trayectoria de concentración representativa) mostrando un incremento en la temperatura del 2.2°C entre el año 2071 al 2100 revelando la influencia de los gases de efecto invernadero y la continuación del aumento de emisiones durante todo el siglo XXI.

En efecto, la Secretaria Distrital de Ambiente – SDA a través de la RMCAB evidencio que la concentración de material particulado PM10 y PM2.5 son los contaminantes que mayor predominan en la ciudad de Bogotá junto con los gases O3, CO, SO2 y NO2. Como bien

señala la Unión Europea en el informe de ExternE (2005) los materiales PM10 y PM2.5 son los que tienen más impacto en la contaminación atmosférica, otros contaminantes del aire como NOx y SO2 su dispersión atmosférica es significativa durante cientos de miles de km, así mismo señalan que separar los roles de SO2, NO2 Y PM10 es problemático ya que tienden a variar juntos en la mayoría de los lugares, de igual forma refieren que el impacto de O3 en función de una de las emisiones NO es el más extremo dado por la fuerte dependencia de la curva en el otro. En efectos de salud, el impacto de los contaminantes del aire tiene lugar a nivel de morbilidad principalmente en el sistema respiratorio y cardiovascular y conduce a la mortalidad prematura.

Metodología Box - Jenkins.

Esta estrategia de modelamiento y análisis de series de tiempo inicialmente presentada en (Box, 1970), la cual tiene dos principales enfoques que son mínimos cuadrados no lineales y la estimación de máxima verosimilitud, básicamente consiste en generar modelos dentro de un clico iterativo donde se pueden configurar modelos tipo estructura general SARIMA $(p,d,q)\times(P,D,Q)$ o cual quiera de sus variantes, este ciclo iterativo se describe generalmente de la siguiente manera ;

1. Proponer una clase de modelos (AR, MA, ARMA, ARIMA, etc.)
2. Elegir el modelo dentro de la clase propuesta, que mejor se ajuste a datos
3. Especificar los parámetros del modelo seleccionado
4. Verifique la calidad del ajuste

La metodología Box – Jenkins, se puede representar como un flujograma, en la figura 1. Si el modelo cumple con la verificación de los supuestos se puede pasar a la etapa de predicción.

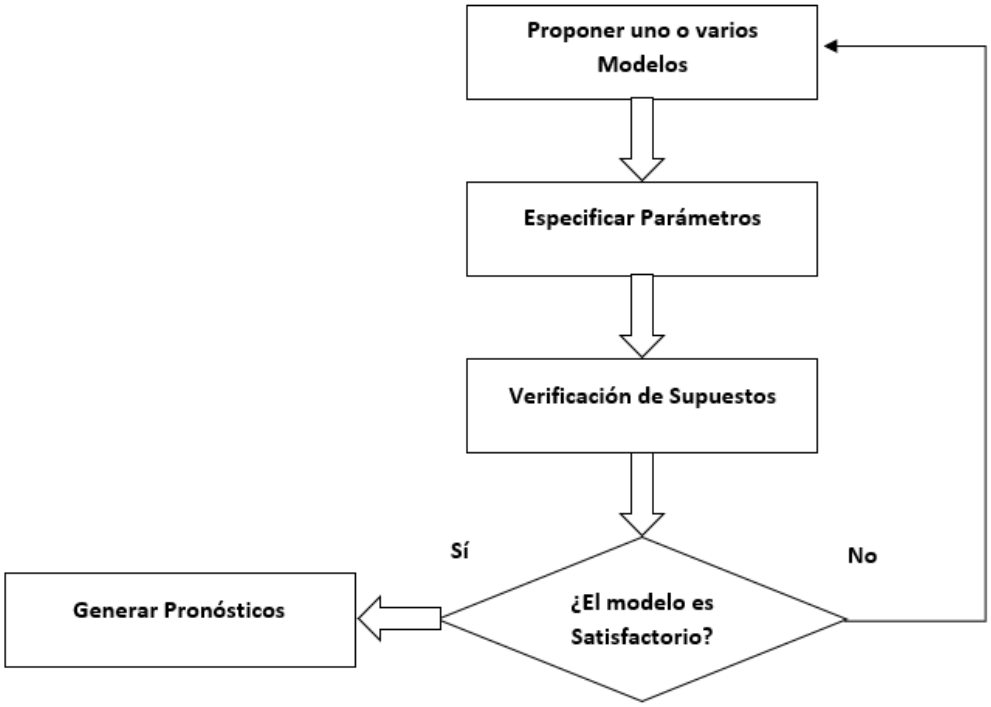

Figura 1: Metodología de Box – Jenkins

Esta metodología se planteó en este proyecto con algunas modificaciones y solo para las técnicas de modelamiento estadísticas, para este caso se utilizará el siguiente criterio que se tendrá en cuenta en la etapa de especificación del modelo, y para elegir el mejor modelo entre técnicas estadísticas y técnicas de machine learning se evaluarán los pronósticos con algunas medidas de error que se describirán más adelante.

Medidas Error de Pronósticos.

Para determinar que modelo es mejor se evaluaran los pronósticos generados por los modelos propuestos entre las diferentes técnicas, el análisis se realizó de la siguiente manera. Al tener una serie de tiempo con J observaciones, se extraen las ultimas n observaciones. Se generan los modelos con $J - n$ observaciones, de los cuales se obtienen pronósticos donde se evaluarán las diferencias con las n observaciones extraídas inicialmente, esto se puede representar como: $$e_i = y_i - \hat{y}_i$$

Donde $\hat{y}_i$ es el pronóstico de y_i.

El poder predictivo de los modelos propuestos se evaluará con dos medias de error, donde $k = J - n$.

1. Error absoluto medio (MAE): $MAE = \frac{1}{n} \sum_{i=k+1}^{J} |e_i|$

2. Raíz de error cuadrático medio (RMSE): $RMSE = \sqrt{\frac{1}{n} \sum_{i=k+1}^{J} e_i^2}$

Modelo VAR (p).

El modelo vectorial auto regresivo de orden (p), se considera un estándar dentro de las técnicas de la econometría, ya que permite analizar series temporales multivariadas, dos de sus bondades es que se puede determinar la interdependencia y las relaciones dinámicas entre las variables (Pfaff, 2008). El modelo VAR se caracteriza porque en su estructura anida y generaliza a los modelos auto regresivos AR(p) univariados, esto se puede representar como:

$$y_t = A_1 y_{t-1} + \cdots A_p y_{t-p} + v + u_t$$

Donde $y_t = (y_{1t} \dots y_{kt})$ es un vector aleatorio de dimensión (K×1), donde se encuentran un conjunto de variables exógenas, estas están definidas por A_i que son matrices (K × K) de coeficientes, $v = (v_1 \dots v_k)$ es el vector de interceptos de dimensión (K ×1). Por último u_t es un proceso de ruido blanco K - dimensional con una variante temporal positiva matriz de covarianza $E(u_t u_t') = \sum u$. Especificando el operador de rezago polinomial $A(L) = I_k - A_1 - \cdots A_p$, el proceso anterior puede definirse como: $A(L)y_t = v + e_t$

Random Forest Regression.

El bosque aleatorio es un método basado en un conjunto de arboles de decisión, este método fue defendido por Leo Breiman y Adele Cutler (Brett Lantz, 2013). Dentro de esta técnica de

machine learning hay una combinación de predictores, donde cada uno de los árboles depende de los valores de un vector aleatorio v muestreado de forma independiente y con la misma distribución para todo el conjunto de árboles, de esta manera el predictor del árbol $h\ (x, v)$, toma valores diferentes a los de la etiqueta de clase. Por otra parte, los valores de salida con son numéricos y se asume que el conjunto de entrenamiento trazo los datos independientemente a la distribución del vector aleatorio y, x. El error de generalización del cuadrado medio para todo los predictores numéricos $h\ (x, v)$, se expresa de la siguiente forma: $E_{x,y}\ (Y - h(X))^2$

El predictor aleatorio en el caso del Random Forest Regresion se forma tomando el promedio sobre k de los árboles $\{h(x, v_k)\}$ (Breiman, 2001), debido a que la selecciona los predictores de forma aleatoria para dividir cada nodo genera tasas de error favorables en comparación con otras técnicas.

Fortalezas

- Un modelo multiuso que funciona bien en la mayoría de los problemas
- Puede manejar datos ruidosos o faltantes; características categóricas o continuas
- Selecciona sólo las características más importantes

Debilidades

- A diferencia de un árbol de decisión, el modelo no es fácilmente interpretable
- Puede requerir algo de trabajo para ajustar el modelo a los datos (Brett Lantz, 2013).

Metodología.

Este documento contribuye a la literatura analizado los datos que captura la Red de Monitoreo de Calidad del Aire de Bogotá – RMCAB, generando un pronóstico de la temperatura y explicando las interacciones con las demás variables, en esta sección se discute

como se integraron los datos, y como se obtuvo una medida mensual de cada variable para toda la ciudad de Bogotá, además del proceso de selección y creación de otras variables para la realización del modelo VAR (p) y el Random Forest Regression.

Datos

Los datos sobre calidad del aire en Bogotá fueron consolidados a partir de la información disponible en la página web http://rmcab.ambientebogota.gov.co/report/MonitorReport, en el propósito de calidad del aire seleccionando todas la zonas, seleccionando los monitores (variables) que estuviesen en la mayoría de la estaciones, los reportes se descargaron en formato Excel, por periodos mensuales que van desde enero del año 2015 hasta diciembre del año 2019, y los registros son promedios por hora del día de cada variable medida por cada estación, en la Figura 2 se muestran las estaciones seleccionadas y en la Figura 3 se muestra el diccionario de datos.

Número	Estación
1	Carvajal – Sevillana
2	Centro de Alto Rendimiento
3	Fontibón
4	Guaymaral
5	Kennedy
6	Las Ferias
7	Puente Aranda
8	San Cristóbal
9	Tunal
10	Usaquén
11	Min Ambiente
12	Suba

Figura 2: Estaciones RMCAB

Variable	Métrica	Descripción
CO	Ppm	Monóxido de carbono, es un gas inodoro, incoloro, inflamable y altamente tóxico
NO	Ppb	Óxido de nitrógeno, es un gas incoloro, que se produce por la quema de combustibles fósiles
NO2	Ppb	Dióxido de nitrógeno, es uno de los principales contaminantes entre los varios óxidos de nitrógeno
NOX	Ppb	Se refiere a la combinación de gases (NO, NO2) que presentan cuando hay oxígeno, el principal es el NO2
HR	%	Denomina humedad ambiental a la cantidad de vapor de agua presente en el aire, en este caso de forma relativa
SO2	Ppb	Dióxido de azufre, es un gas incoloro con un característico olor asfixiante
Temperatura	°C	La temperatura es una magnitud referida a las nociones comunes de calor o frío, que físicamente es energía
Vel Viento	m/s	El viento es el movimiento del aire en la troposfera, generado por causas naturales
Rad Solar	W/M²	La radiación solar es el conjunto de radiaciones electromagnéticas emitidas por el Sol
OZONO	Ppb	Es un gas altamente reactivo y contaminante, que genera preocupación ya que es altamente oxidante y afecta a los tejidos vivos
PM10	µg/m3	Son pequeñas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera
PM2.5	µg/m3	Son aerosoles que pueden ser de origen natural o debida a la actividad humana
Precipitación	Mm	Es un fenómeno atmosférico que se inicia con la condensación del vapor de agua contenido en las nubes
Presión Baro	mmHg	Es la presión ejercida por el aire en cualquier punto de la atmósfera

Figura 3: Diccionario de Datos

Preparación de los Datos.

Para esta fase se utilizó el lenguaje de programación Python 3 del software Anaconda, utilizando principalmente las librerías Pandas y Numpy, para obtener una medida mensual por cada una de las variables, primero se analizó como se comportaba cada una de las variables a lo largo del tiempo, por medio de diagramas de caja mensuales Figura 4, se

decidió tomar la mediana, ya que no tiene problemas con los datos atípicos debido a sus características.

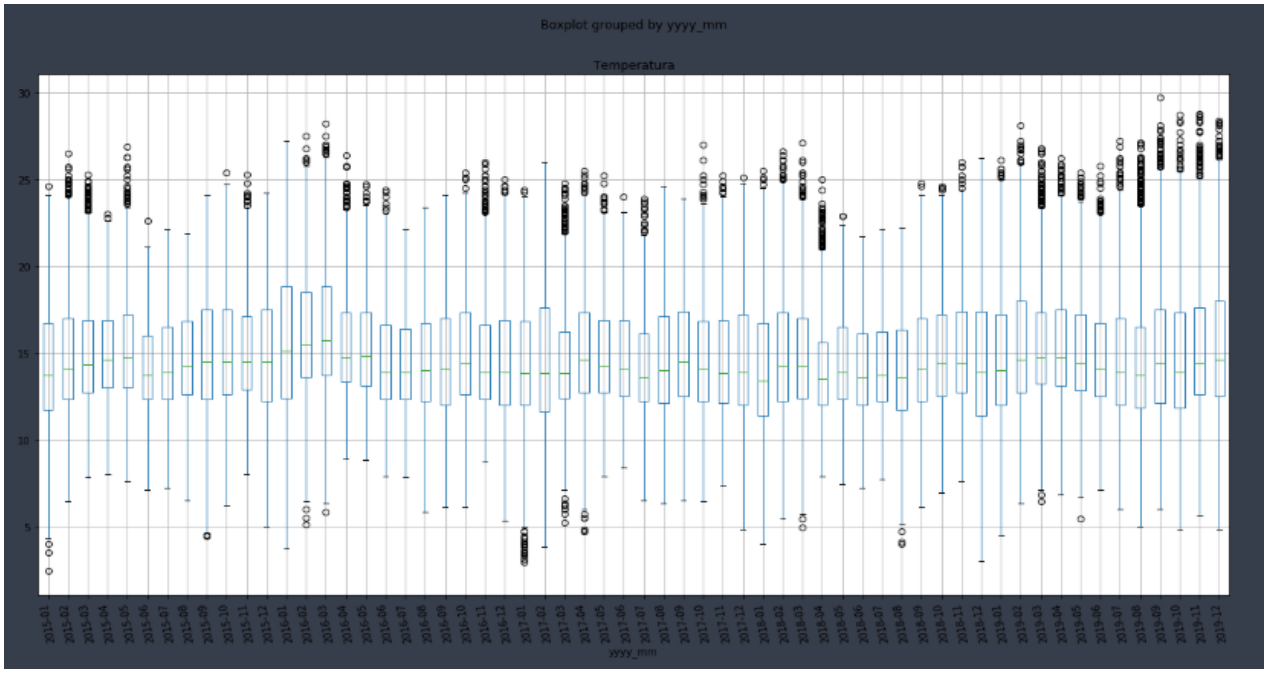

Figura 4: Diagrama de Caja Temperatura

Para tomar la mediana de cada variable se hizo un análisis de completitud por cada una de las variables capturadas por las estaciones, observando la completitud mensual de variable/estación, el criterio que se utilizo para crear una “marca” fue que al menos tuviese el 50% de historia, con esta marca se puedo obtener la cantidad de meses que la cumplen por variable/estación, posteriormente se decidió seleccionar los campos variable/estación que cumplieran con al menos el 96% de los meses con el criterio de completitud, con esto se garantizó tener suficiente historia en cada serie y que hubieran al menos dos estaciones por variable, así se seleccionaron los campos variable/estación para obtener la mediana de cada una de las 13 variables aptas a lo largo del tiempo, con la ayuda de la librería Scikit-Learn, se imputaron algunos datos con la mediana en las series que en el peor casos serian tres datos debido al proceso anteriormente descrito.

Selección de Variables Modelo VAR (p).

Se aplico el algoritmo de Dynamic Time Warping (DTW), el cual permite clasificar y agrupar series de tiempo midiendo sus alineaciones no lineales entre dos series temporales para acomodar secuencias que son similares, pero localmente desfasados (Ratanamahatana, 2004), se aplicó este algoritmo a una matriz que contiene las 13 series de tiempo ya transformadas con retornos logarítmicos, esta matriz se analizó utilizando la técnica de Análisis de Componentes Principales (ACP), pasando la variable de la temperatura como suplementaria, con la ayuda del ambiente de programación libre R en su versión 3.6.0 para la plataforma Windows, y el editor Rstudio, se utilizaron las librerías dtw y FactoMineR, en la Figura 5 se muestra la salida del análisis del ACP.

.

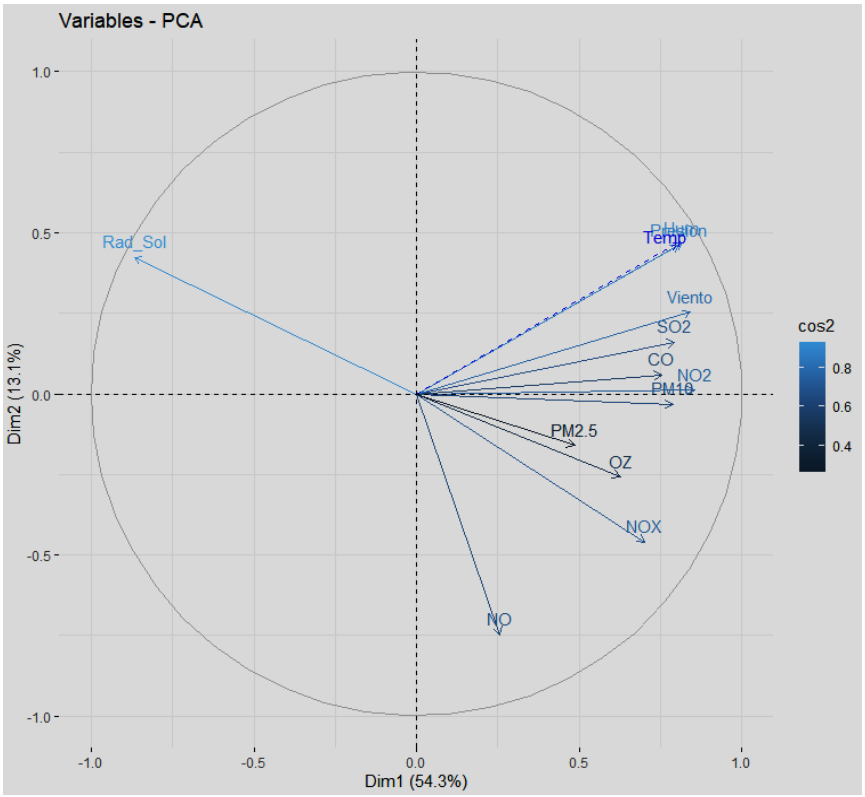

Figura 5: ACP Variables

Se puede observar en la Figura 5 que las series de tiempo que están mas correlacionadas a la temperatura, son la Humedad Relativa y la Presión Barométrica, esto tiene mucho sentido ya

que responde a un proceso natural, el cual básicamente se puede explicar que a mayor temperatura mayor humedad y que la presión barométrica baja cuando el aire caliente aumenta.

Con esta base del ACP, se aplicó posteriormente la técnica de agrupación K-medias la cual es bastante eficiente y funciona bien al dividir los datos en grupos útiles (Brett Lantz, 2013), este algoritmo no supervisado se implementó utilizando el paquete stats de R, y arrojo la salida que se muestra en la Figura 6, donde se pueden observar los tres clúster que se formaron, siendo el tercer clúster el que contiene las variables que se utilizaron para modelar el VAR (p), las series de tiempo seleccionadas fueron; Temperatura, Humedad Relativa, Presión Barométrica, Velocidad del Viento, SO2, CO, NO2 y PM10.

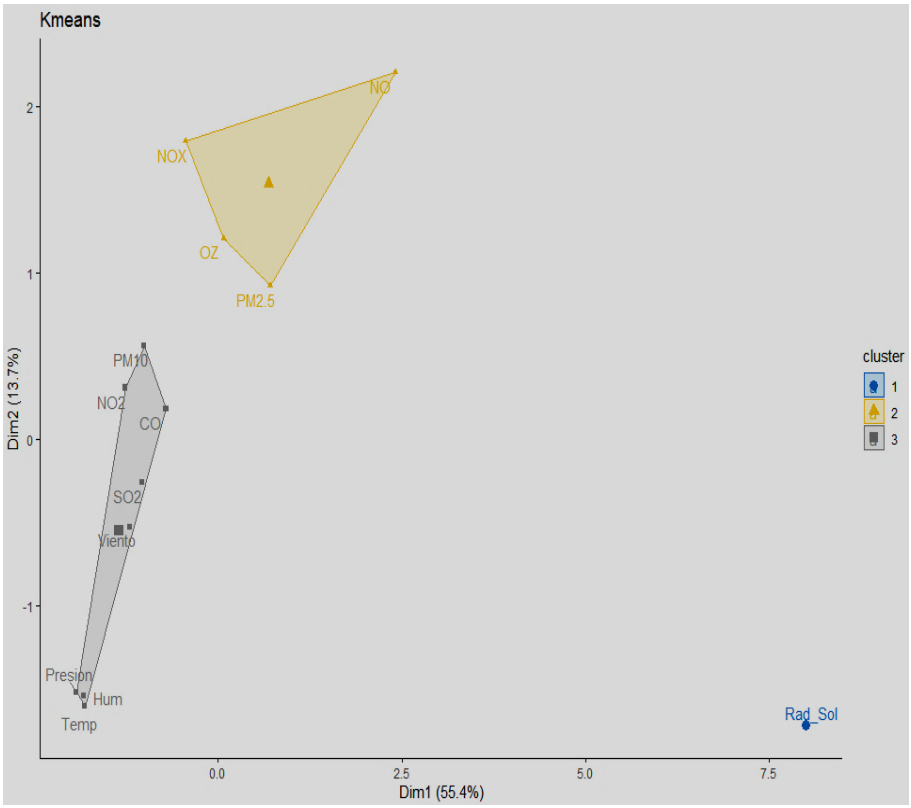

Figura 6: K-medias

Transformación Series de Tiempo Random Forest Regression.

En cuanto a las variables que se utilizaron para modelar la temperatura con la técnica de machien learning, se hicieron diferenciaciones a las 11 series de tiempo de uno hasta seis meses atrás, con esto se garantiza que el algoritmo tome la información del pasado para explicar el comportamiento de la temperatura y así pueda predecir, en esta etapa esa fue la única transformación que se le hicieron a los datos, esto es una ventaja en cuanto a los modelos estadísticos, ya que después de hacer las predicciones hay que hacer una transformación inversa para obtener el dato, además que el algoritmo tiene un proceso interno que evalúa la importación de las variables, con esto se pueden seleccionar las variables más predictivas.

Resultados.

Para la etapa de modelamiento de las series de tiempo, se utilizo un esquema de prueba por fuera de la muestra, es decir, que se particiono la base en dos grupos uno de entrenamiento y otro de prueba, los datos de entrenamiento tienen un histórico desde enero del año 2015 hasta agosto del año 2019, y los datos de prueba son septiembre hasta diciembre del año 2019, esto para determinar qué modelo es más preciso, por medio de las medidas de error de pronóstico.

Modelo VAR.

El primer paso es identificar la relación que tienen las series de tiempo consideradas, se usó el método explicado en la sección de metodología “Selección de Variables Modelo VAR (p)”, donde se obtuvieron diez series (Temperatura, Humedad Relativa, Presión Barométrica, Velocidad del Viento, SO2, CO, NO2 y PM10), con estas series ya transformadas en una matriz, se realizó un proceso iterativo empleado el método de selección de variables *forward*, el cual consiste en ir agregando variables de acuerdo a la correlación que tienen con la variable dependiente (temperatura), para este caso se tomó el dato de las distancias entre la

temperatura y las variables del clúster obtenido con el método de K-means, al final se obtuvo un modelo con seis series de tiempo Figura 7, que cumplía con la fase de especificación con AIC de -4.279402e+01 ,el AIC fue introducido por Hirotsugu Akaike en (H. Akaike, 1974), el cual permite elegir el modelo que minimiza la perdida de información, en este caso brinda un modelo VAR con parámetro de rezago p de 6 periodos, se realizo el ajuste del modelo VAR (6), posteriormente se realizó la prueba de raíces unitarias, donde cada módulo del modelo dieron menor a 1, es decir el proceso es estacionario.

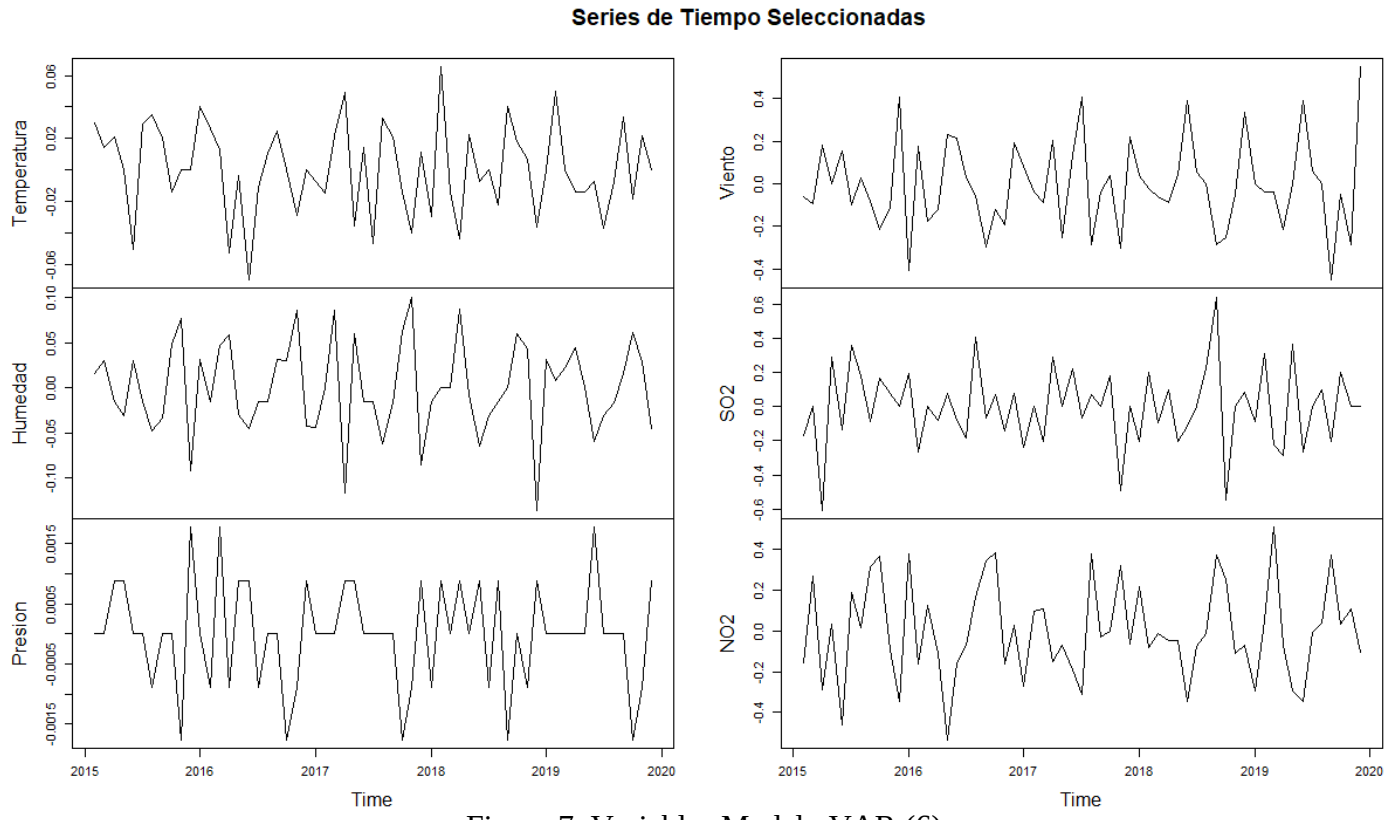

Figura 7: Variables Modelo VAR (6)

Se aplicaron las pruebas (con una significancia estadística del 5%) de Portmanteau (H0: Residuos no correlacionados) donde se obtuvo un p-valor de 0.1095, Jarque-Bera multivariado (H0: Distribución Normal en los Residuos) donde se obtuvo un p-valor de 0.8505 y ARCH (H0: Residuos homoscedásticos) donde se obtuvo un p-valor de 1, por lo se

tuvo suficiente información estadística para decir que no se tiene ningún problema con los residuos del modelo VAR (6).

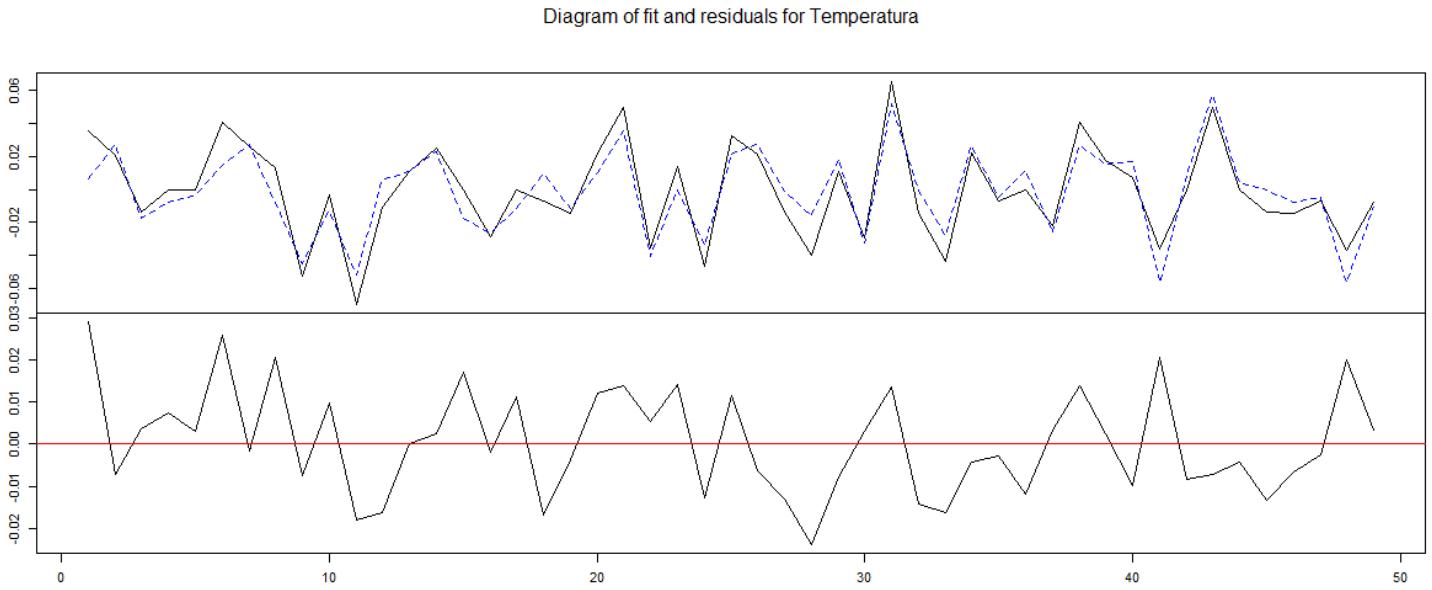

Figura 8: Ajuste vs Real y Residuos

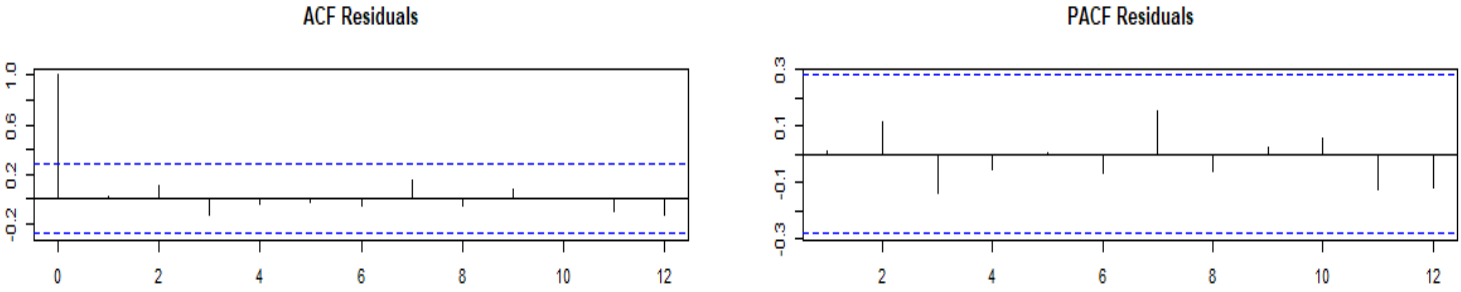

Figura 9: Autocorrelogramas Residuos VAR (6)

Funciones de Impulso.

En las Figuras 10 – 15, se muestran las funciones de impulso respuesta (FIR) asociadas al modelo propuesto, tomando la variable temperatura como variable dependiente, en la Figura 10 se observa los impulsos de la temperatura con si misma, se pueden observar fluctuaciones a lo largo del tiempo, hay una relación cíclica variación, atenuándose al final, lo cual se puede ver como la temperatura depende de si misma para su comportamiento futuro.

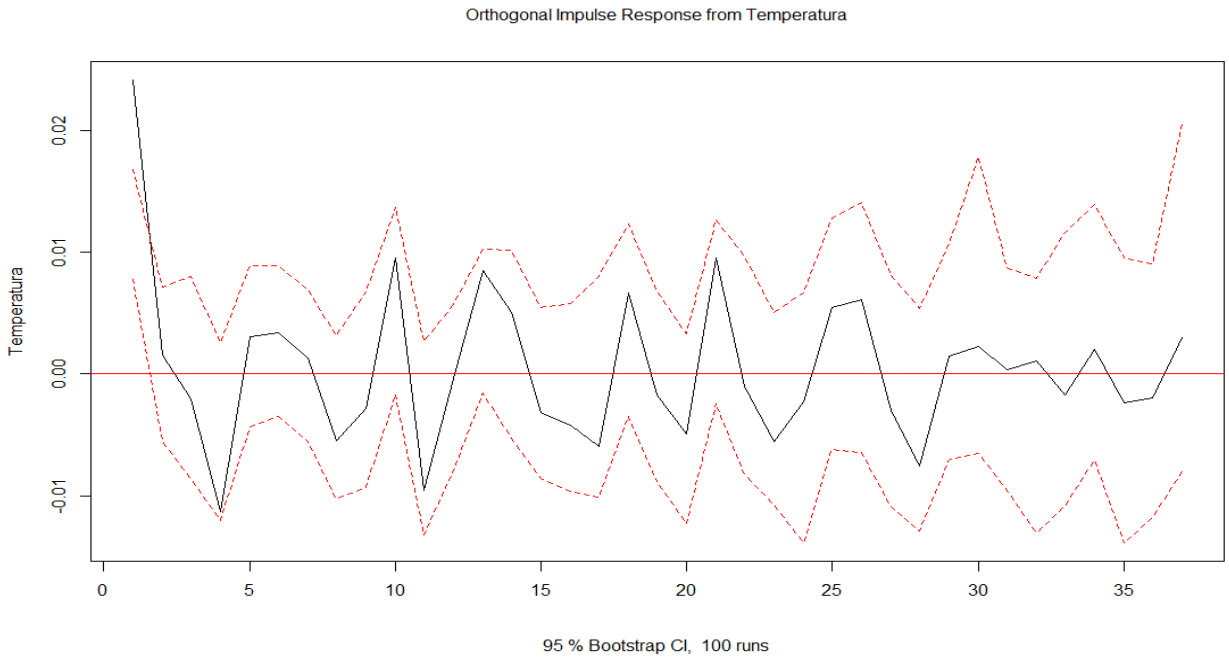

Figura 10: FIR Temperatura - Temperatura

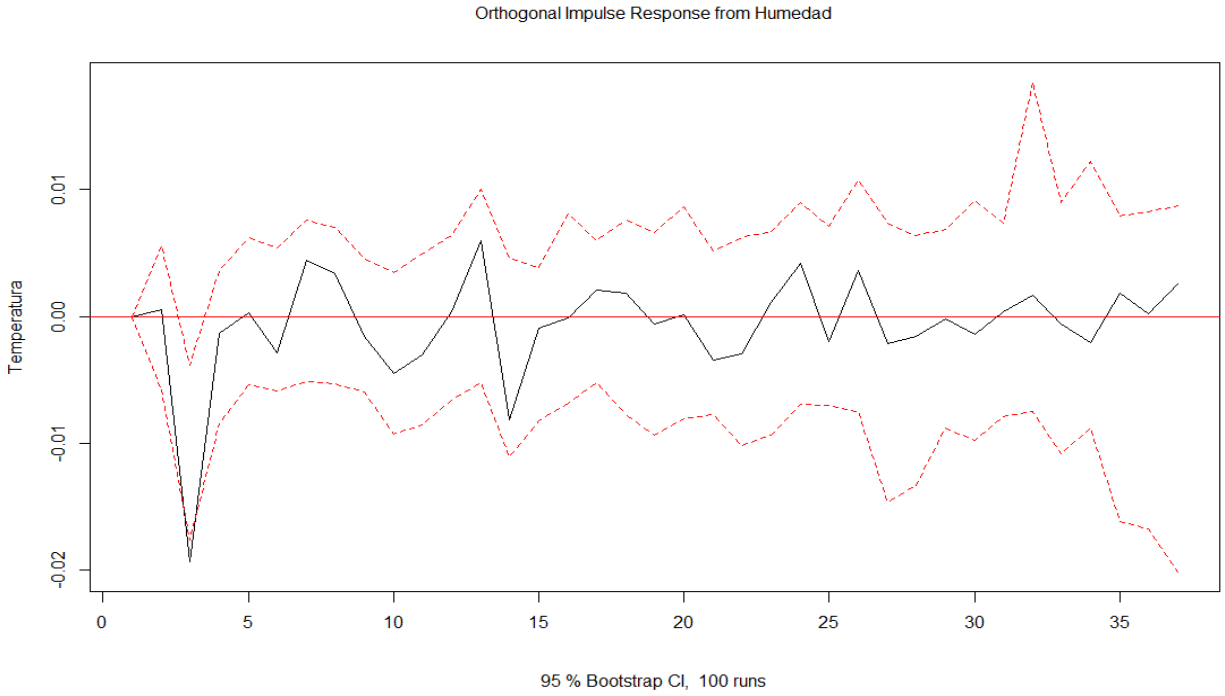

Figura 11: FIR Humedad – Temperatura

En la Figura 11 se muestra la FIR entre la Humedad Relativa a la Temperatura, inicialmente se observa un impacto negativo fuerte lo cual presenta un comportamiento de temperaturas bajas, pero después del periodo cinco tiene un comportamiento fluctuante sin mucho impacto.

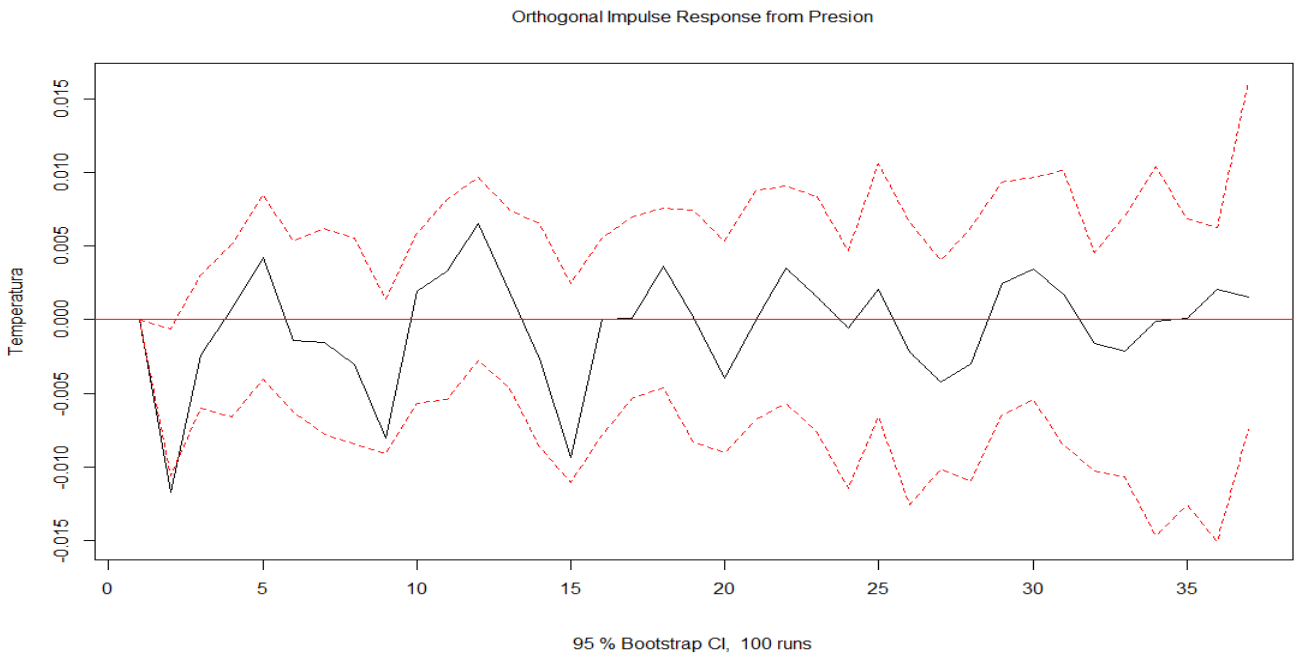

Figura 12: FIR Presión - Temperatura

En la Figura 12 se muestra la FIR entre la Presión Barométrica a la Temperatura, se observa un comportamiento cíclico con un rango de más de cinco periodos entre las dos primeras fluctuaciones, lo cual nos indica un comportamiento sostenido en el tiempo, pero no tan intenso en la variación de la temperatura al futuro, atenuándose la final.

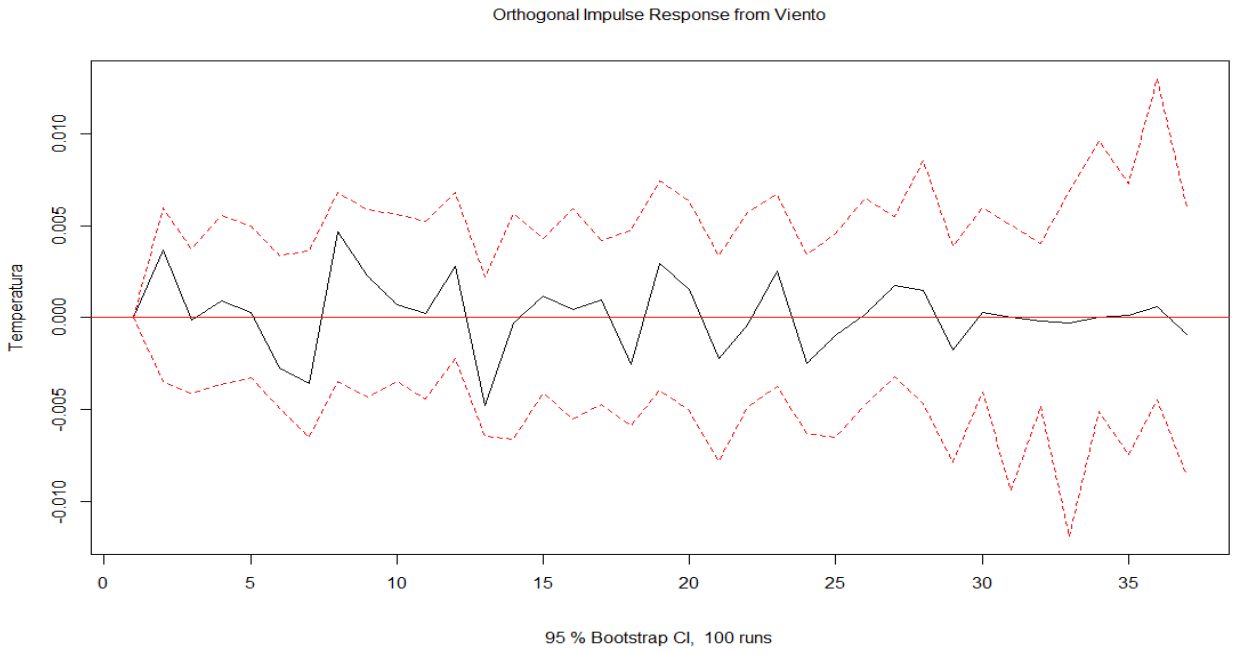

Figura 13: FIR Viento - Temperatura

En la Figura 13 se muestra la FIR entre la Velocidad del Viento a la Temperatura, se observa fluctuaciones con impactos débiles a lo largo del tiempo, alrededor del séptimo periodo un incremento sostenido, lo cual indica incremento de la temperatura en ese futuro periodo, pero después pierde fuerza en impacto.

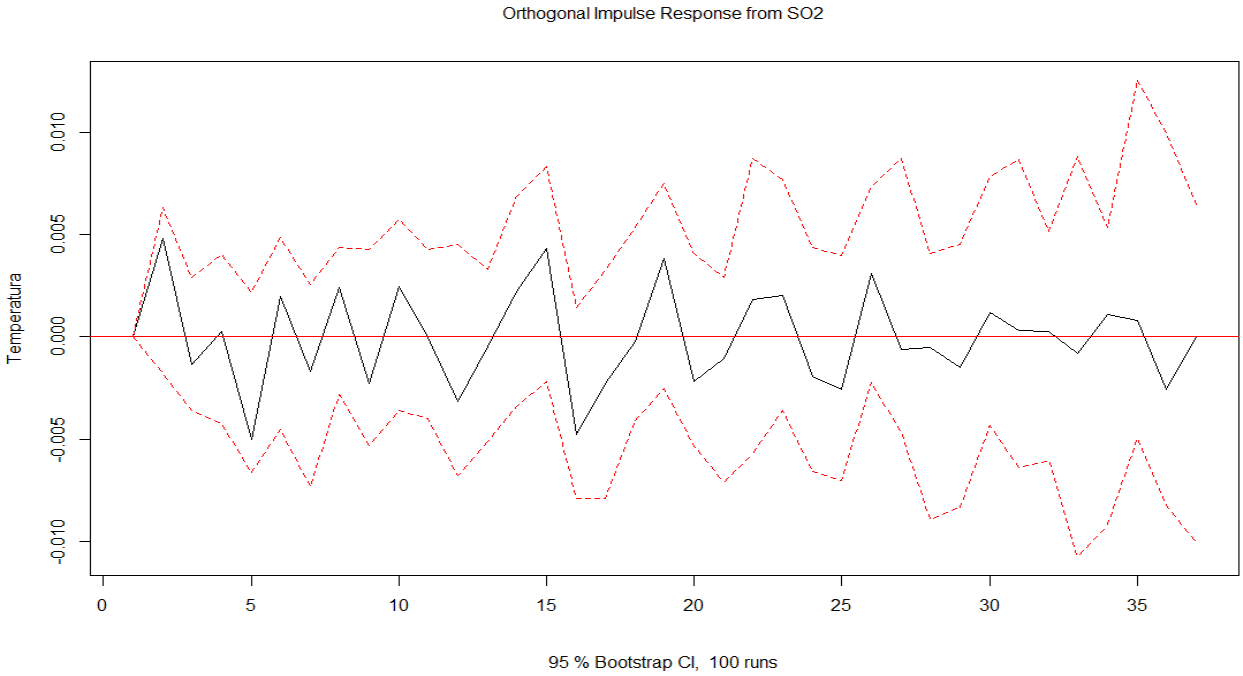

Figura 14: FIR SO2 – Temperatura

En la Figura 14 se muestra la FIR entre el Dióxido de Azufre (SO2) a la Temperatura, se observa un comportamiento cíclico en impacto, que en principio tiene caída de impacto negativo que llega hasta el periodo cinco, y entre el periodo 10 hasta el 20 se presentan las mayores fluctuaciones, lo cual muestra que este contaminante del aire que produce problemas respiratorios, tiene una interacción entre sus emisiones y la variación de la temperatura, es decir, afecta su comportamiento futuro.

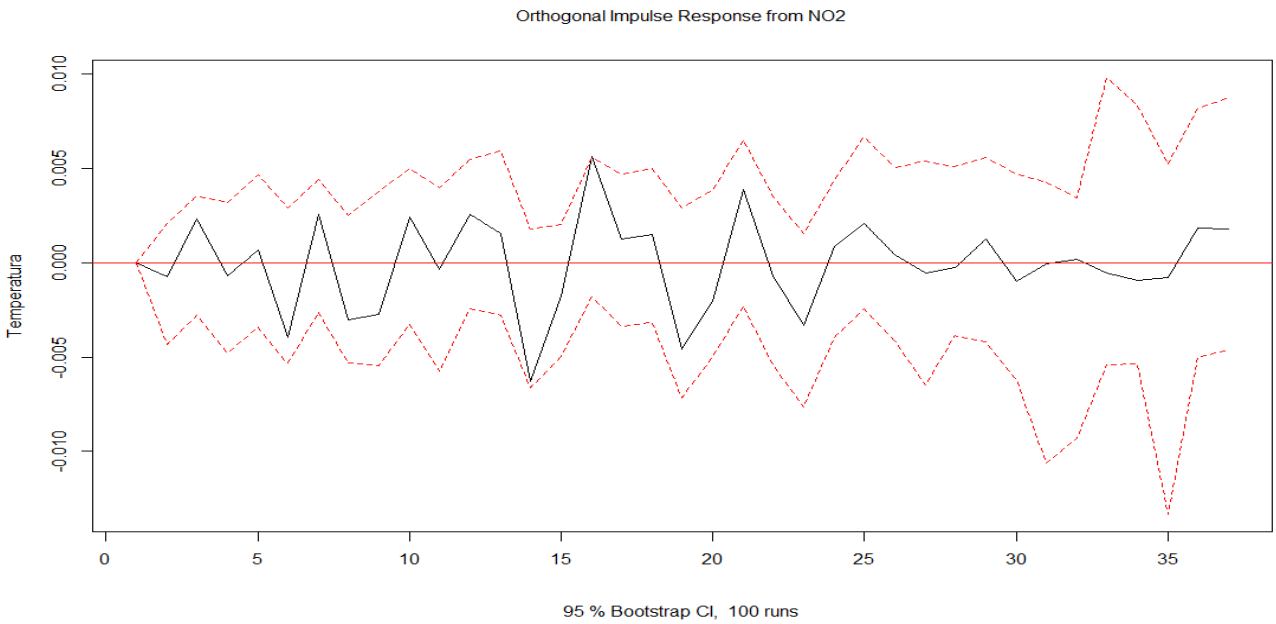

Figura 15: FIR NO2 – Temperatura

En la Figura 15 se muestra la FIR entre el Dióxido de nitrógeno (NO2) a la Temperatura, se observa un comportamiento cíclico que en principio tiene un impacto leve, pero entre el periodo 14 y 23 se presenta sus mayores fluctuaciones, lo cual muestra que este contaminante del aire que además es fuente indirecta de los gases de efecto invernadero y que produce problemas respiratorios (IDEAM, 2007), también provoca variaciones en el comportamiento futuro de la temperatura.

Descomposición de la Varianza.

En la Figura 16 se presenta las descomposiciones de la varianza en los pronósticos realizados para la temperatura en la ciudad de Bogotá, se puede observar que, en cada paso siguiente en el tiempo, la Presión Barométrica y la Humedad Relativa explican la varianza del pronostico entre un 20% hasta un 45%, mientras que las demás variables explican alrededor del 10%, esto se debe al proceso natural que existen entre estas variables, cabe resaltar que el SO2 es el mas presencia tiene de las variables de calidad del aire.

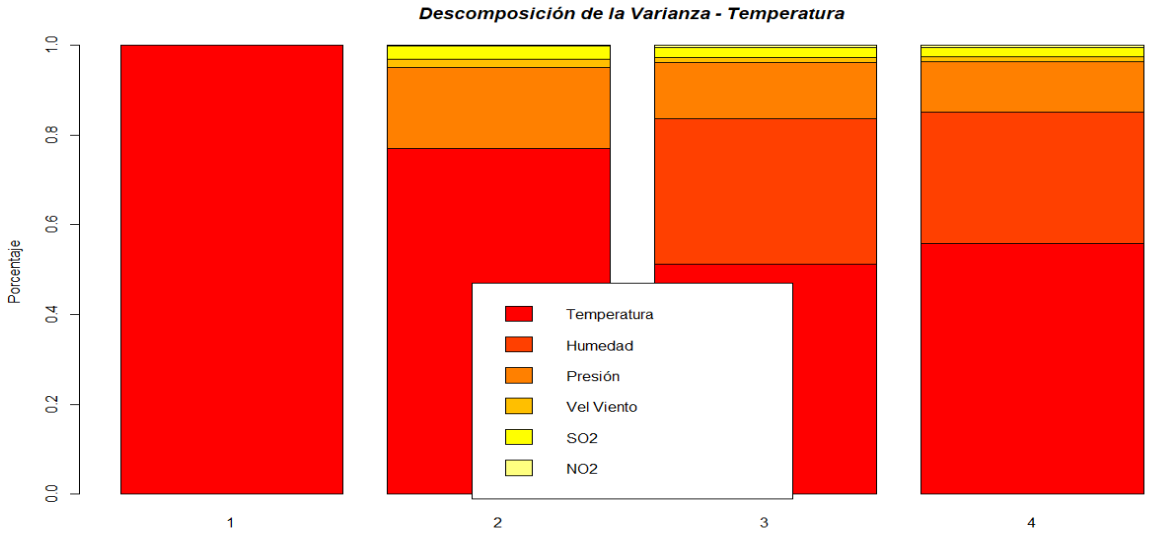

Figura 16: FEVD Pronóstico Temperatura

Random Forest Regression.

Después de crear las características con hasta 6 periodos de diferenciación, se implemento un primer ajuste teniendo en cuenta los siguientes parámetros;

- *ntree*: cantidad de árboles en el bosque. Se quiere estabilizar el error, pero usar demasiados árboles puede ser innecesariamente ineficiente.
- *mtry*: cantidad de variables aleatorias como candidatas en cada ramificación.

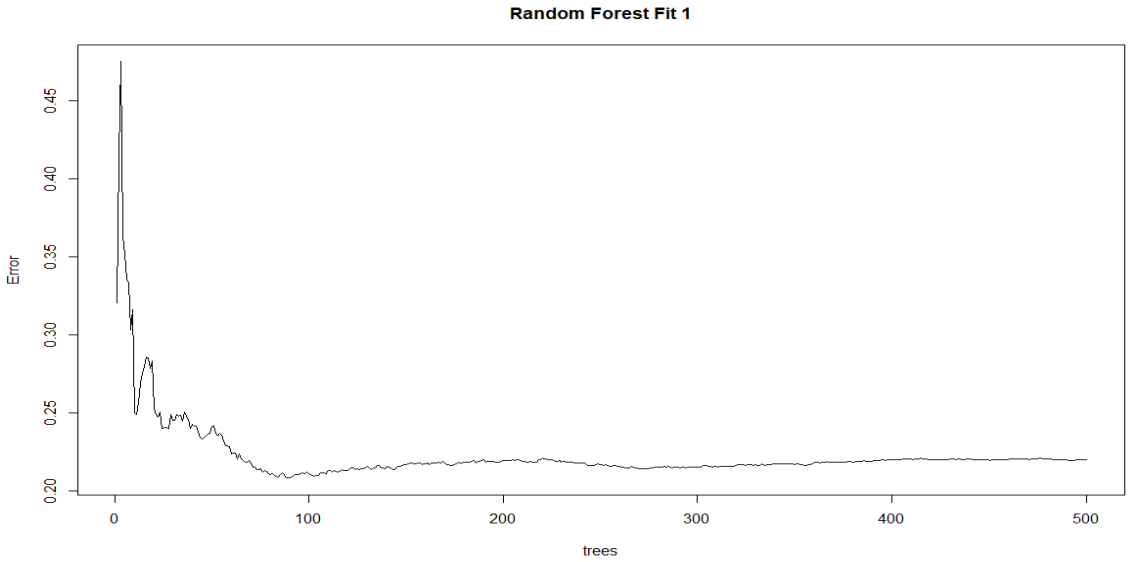

Figura 17: Error – ntree

Para este primer ajuste el parámetro de *ntree* fue de 89 Figura 17, y el *mtry* de 8 Figura 18,

estos fueron los mejores parámetros para reducir el error del modelo.

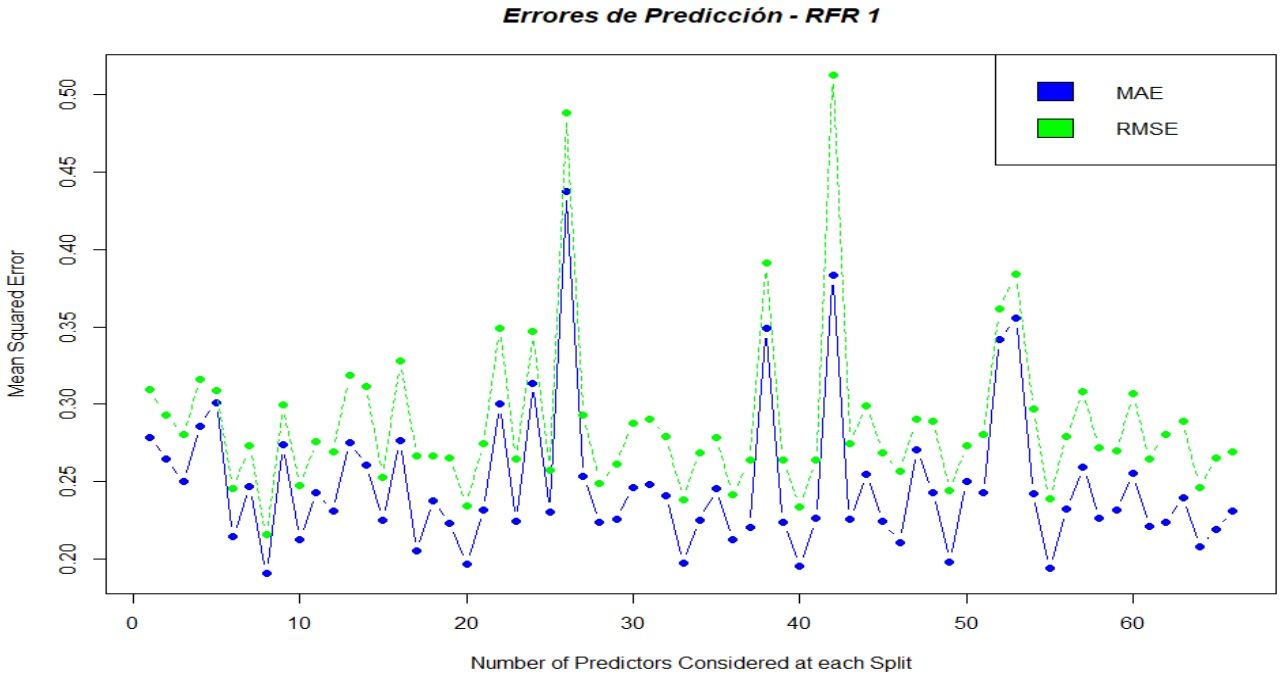

Figura 18: Errores - mtry

El objetivo de este primer ajuste fue obtener las variables más importantes, para un segundo ajuste, la importancia se calcula como el promedio de sus actuaciones permutadas en todo el bosque aleatorio, en la Figura 19 se muestra el top 16 del segundo ajuste.

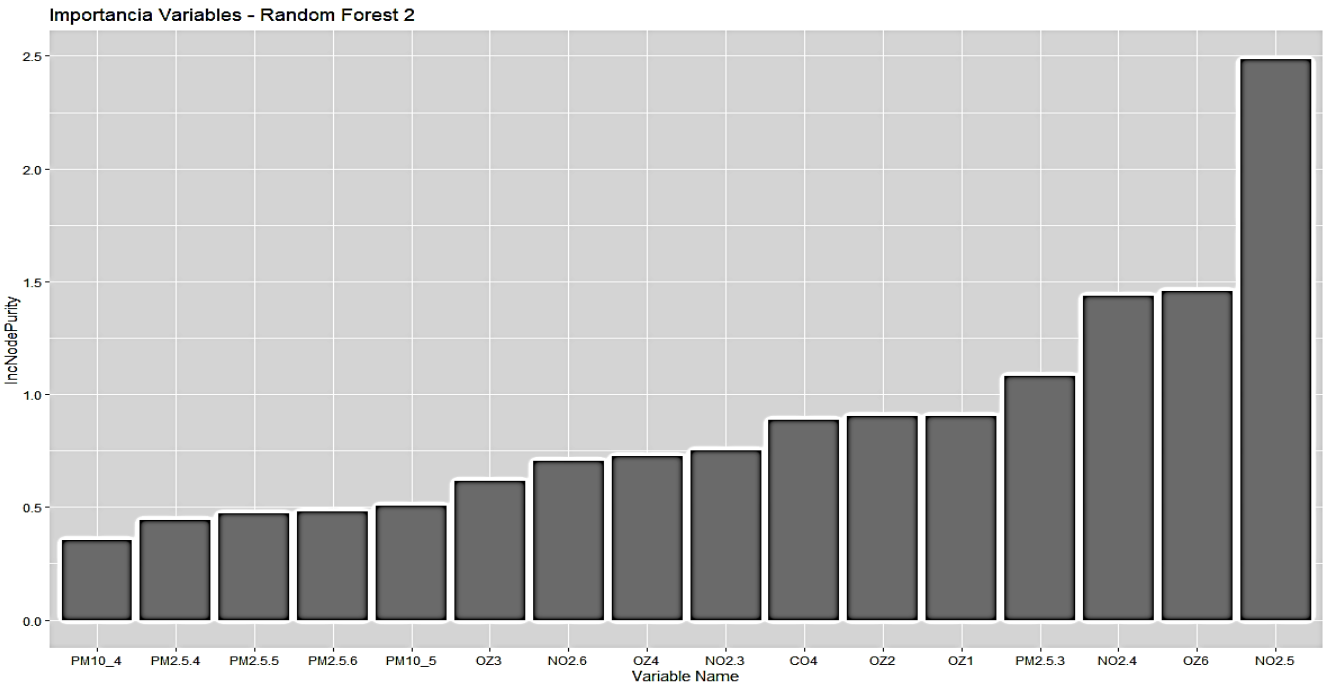

Figura 19: Importancia Variables Fit 2

En el segundo ajuste también aplicando la misma parametrización anteriormente descrita, obteniendo un valor para el *ntree* de 159 y para el *mtry* de 9, estas 16 variables se pueden agrupar en cinco series iniciales las cuales son: Dióxido de nitrógeno (N02), Ozono, PM 2.5, Monóxido de carbono (CO) y PM 10, en orden descendiente de importancia, lo cual muestra la interacción que tienen los gases de efecto invernadero como el NO2, el Ozono, y el CO, en el comportamiento histórico y futuro de la temperatura, y en un nivel menor el material particulado pero no menos importante ya que estos compuestos y partículas afectan de varias formas la salud humana e incrementan los problemas medioambientales en cuanto al cambio climático, la Figura 20 muestra el ajuste del modelo a los datos reales de entrenamiento.

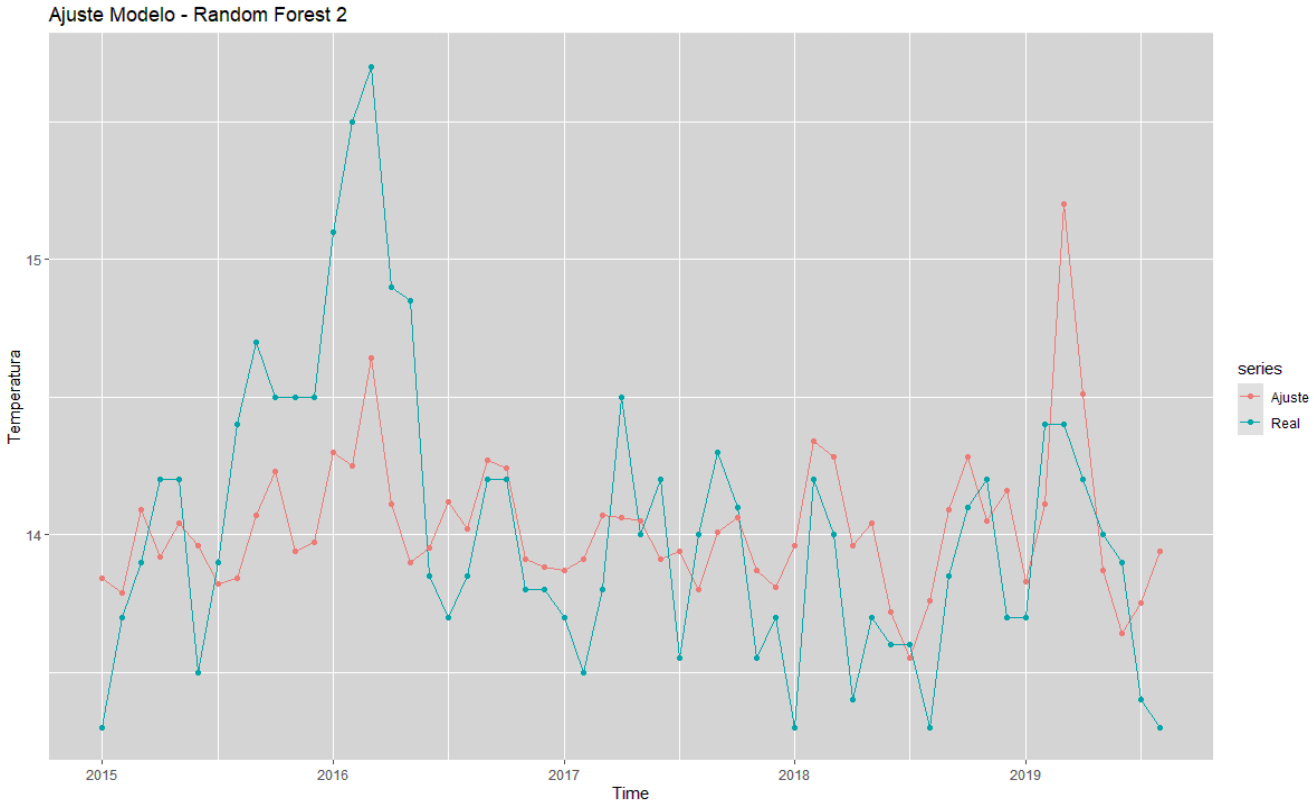

Figura 20: Ajuste Fit 2 vs Real

Comparación de Pronósticos.

En esta sección evaluaremos la precisión de los pronósticos del modelo estadístico VAR (6) y el modelo de machine learning Random Forest Regression, en la Figura 21 se muestra el

ajuste del modelo VAR (6), en los datos de prueba que contienen el ultimo cuatrimestre del año 2019.

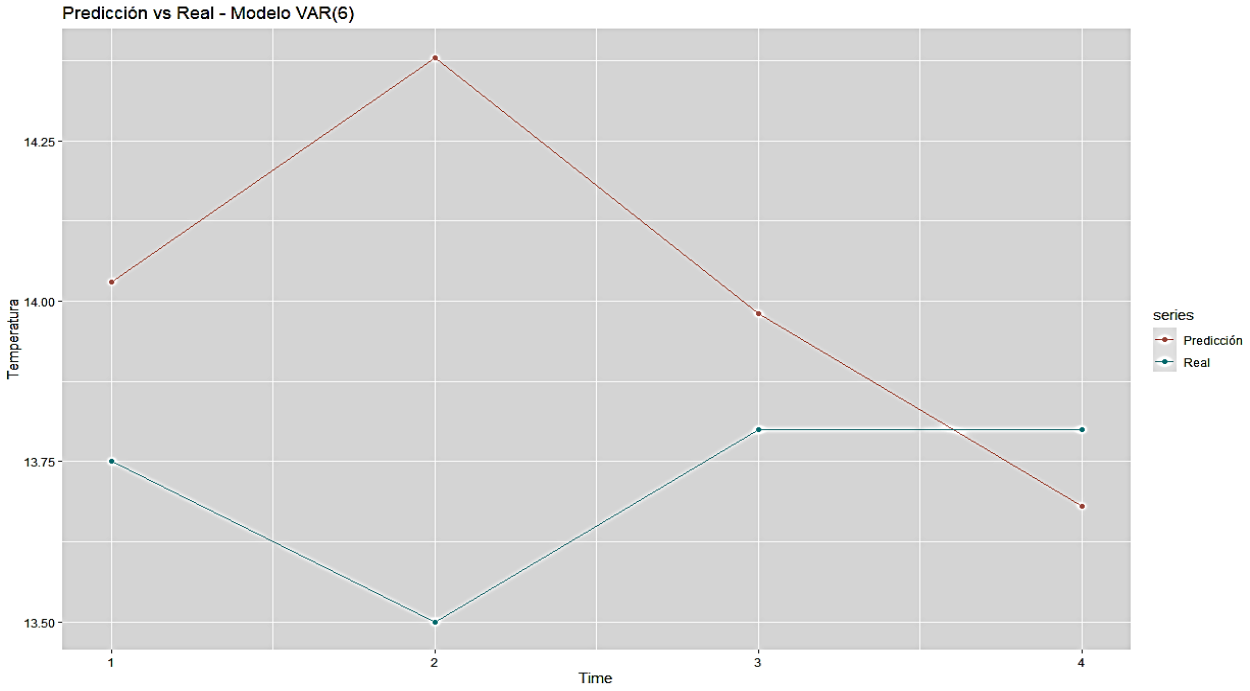

Figura 21: Ajuste Pronostico VAR (6)

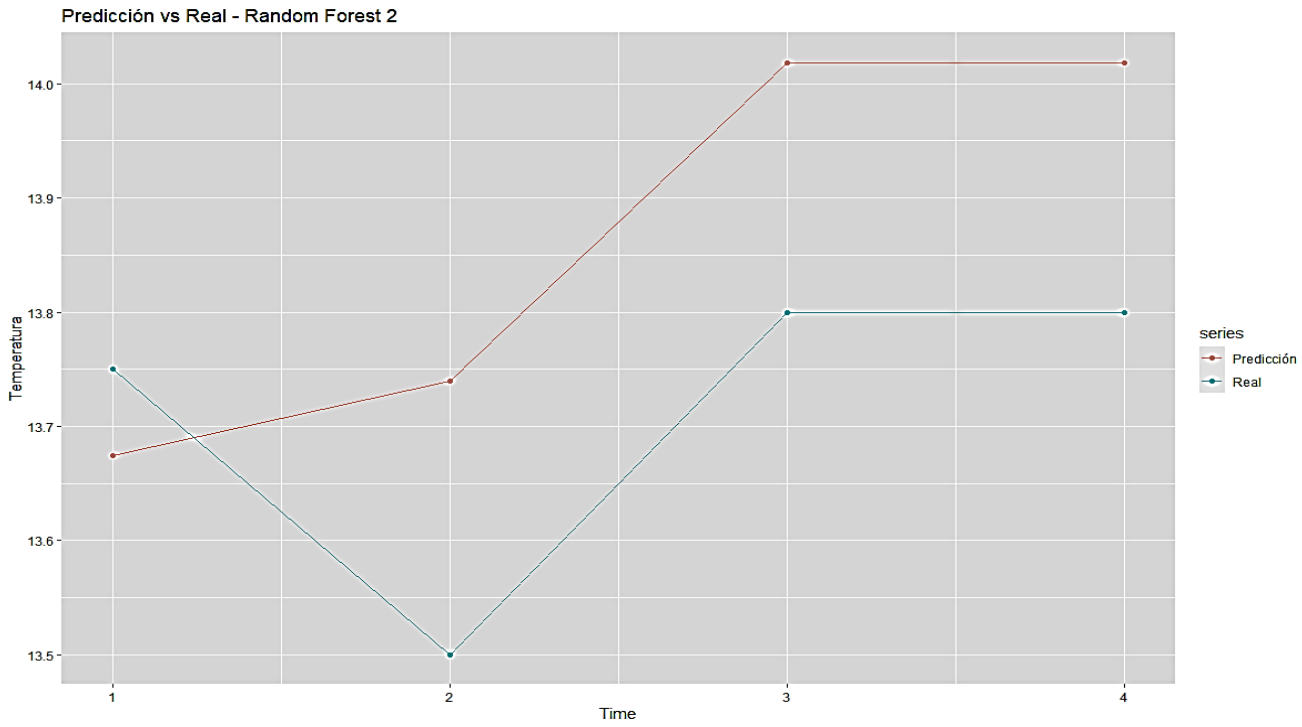

Figura 22: Ajuste Random Forest Regression

Modelo	MAE	RMSE
VAR (6)	0.36	0.47
Random Forest	0.19	0.20

Figura 23: Medidas Error de Pronóstico

En la Figura 23 se muestra la evaluación de los pronósticos realizados por ambas técnicas, lo cual muestra un mejor ajuste del modelo de Random Forest Regression, esto también se puede observar en la Figura 22 donde el pronóstico se ajusta también a la tendencia de los datos de prueba, lo que no logra hacer el modelo VAR (6).

Conclusiones.

El objetivo principal de este proyecto de investigación es la obtener un modelo que permita predecir la temperatura en la ciudad de Bogotá, además que se pueda explicar la interacción entre contaminantes y variable de calidad del aire.

- Fue creado un modelo estadístico VAR (p), con el cual se generaron pronóstico para contrastar con los datos de prueba, el cual no tuvo buenos rendimientos en la mediadas de error.
- También se creo un modelo de machine learning Random Forest Regression, con el cual se obtuvieron pronósticos, los cuales fueron evaluados con las medidas de error, se obtuvieron resultados competitivos superando el modelo VAR, y capturando la tendencia de los datos de prueba.
- En el proceso de selección de series aplicando ACP y K-means, se obtuvieron resultados en cuanto a la explicación de las variables analizadas, principalmente entre las variables meteorológicas.
- Con los análisis estadísticos del modelo VAR, se pudieron explicar varias interacciones entre las variables meteorológicas, contaminantes que afectan la salud pública, gases de efecto invernadero y la temperatura.

- En la evaluación del modelo Random Forest Regression, se puedo obtener la importancia de las características, identificando interacciones importantes en comportamiento histórico y futuro de la temperatura, ya que tiene un poder predictivo y explicativo, entre esas variables destacan los gases de efecto invernadero y el material particulado, variables que tiene un impacto en la calidad del aire de Bogotá y que son reconocidas mundialmente como causantes del cambio climático.
- Para futuros trabajos, como se puedo ver en la implementación del VAR, existe un comportamiento estacional, en este proyecto no se alcanzo a implementar un modelo SVAR, por otra parte, este proyecto es una buena base para llegar hacer predicciones en una escala de tiempo más pequeña, en días.

REFERENCIAS BIBLIOGRÁFICAS.

Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, (2017). Informe del estado de la calidad del aire en Colombia 2016. Recuperado de: http://www.ideam.gov.co/web/contaminacion-y-calidad-ambiental/informes-del-estado-de-la-calidad-del-aire-en-colombia

Instituto Distrital de Gestión de Riesgos y Cambio Climático IDIGER, (2018). Caracterización Genera del Estado de Cambio Climático para Bogotá. Recuperado de: https://www.idiger.gov.co/rcc

Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, (2017). Tercera comunicación nacional de Colombia a la convención marco de las Naciones Unidas sobre cambio climático. Recuperado de: http://documentacion.ideam.gov.co/openbiblio/bvirtual/023731/TCNCC_COLOMBIA_CMNUCC_2017_2.pdf

Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, (2017). Bogotá supero su pico histórico de temperatura. Recuperado de: http://www.ideam.gov.co/web/intranet/noticias//asset_publisher/gO37c5HXVo8L/content/bogota-supero-su-pico-historico-de-temperatura?

ExternE, (2005). Externalities of Energy: Methodology 2005 Update. Belgica, Europea Commission. Recuperado de: https://www.researchgate.net/publication/232075838_ExternE_Externalities_of_Energy_Methodology_2005_Update

Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, (2018). Metodología de la operación estadística variables meteorológicas. Recuperado de:

http://www.ideam.gov.co/documents/11769/72085840/Documento+metodologico+variables+meteorologicas.pdf/8a71a9b4-7dd7-4af4-b98e-9b1eda3b8744

Secretaria Distrital de Ambiente, (2011). Informe de Avance. Programa Distrital de Acción Frente al cambio Climático Línea base emisiones GEI, escenarios emisiones referente 2008, proyecciones 2019,2038 y 2050 de Bogotá D.C. Recuperado de: http://ambientebogota.gov.co/es/c/document_library/get_file?uuid=d3e6b6be-ecf2-4ee2-a9a8-9a403c7243af&groupId=10157

La Red de Monitoreo de Calidad del Aire de Bogotá – RMCAB, (2018). Informe anual de Calidad del Aire en Bogotá. Recuperado de: http://rmcab.ambientebogota.gov.co/Pagesfiles/IA%20Informe%20Anual%202018%20RMCAB.pdf

Box, G.E.P.; Jenkins, G. (1970). Time series analysis: forecasting and control. Holden-Day series in time series analysis. Holden-Day.

Pfaff, B. (2008). Analysis of Integrated and Cointegrated Time Series with R (Use R). Springer, 2nd edition.

Brett Lantz. 2013. Machine Learning with R. Packtpub

H. Akaike, "A new look at the statistical model identification," in IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, December 1974

Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, (2007). Información Técnica sobre Gases de Efecto Invernadero y el Cambio Climático. Recuperado de: http://www.ideam.gov.co/documents/21021/21138/Gases+de+Efecto+Invernadero+y+el+Cambio+Climatico.pdf

Ratanamahatana, C. y Keogh, EJ (2004). Making Time-series Classification More Accurate Using Learned Constraints. SDM