Estimación de la concentración de ozono troposférico mediante la regresión por mínimos cuadrados parciales PLS a partir de mediciones de la RMCAB y de imágenes satelitales Landsat en la ciudad de Bogotá D.C. para el periodo 2020-2022.

YERSON STIVEN GALINDO HUESO

Código 11792219987

Trabajo para optar al título de especialista en sistemas de Información Geográfica

UNIVERSIDAD ANTONIO NARIÑO

ESPECIALIZACIÓN EN SISTEMAS DE INFORMACIÓN GEOGRÁFICA

FACULTAD DE INGENIERÍA AMBIENTAL Y CIVIL.

Bogotá D.C, Colombia.

2022

Estimación de la concentración de ozono troposférico mediante la regresión por mínimos cuadrados parciales PLS a partir de mediciones de la RMCAB y de imágenes satelitales Landsat en la ciudad de Bogotá D.C. para el periodo 2020-2022.

YERSON STIVEN GALINDO HUESO

Código 11792219987

Proyecto de frado presentado como requisito para optar al título de especialista en sistemas de Información Geográfica

UNIVERSIDAD ANTONIO NARIÑO

ESPECIALIZACIÓN EN SISTEMAS DE INFORMACIÓN GEOGRÁFICA

FACULTAD DE INGENIERÍA AMBIENTAL Y CIVIL.

Bogotá D.C, Colombia.

2022

Contenido

Tablas.

Tabla de Figuras

Resumen.

En este trabajo se desarrolló una regresión lineal multivariable para la predicción de ozono troposférico basada en la metodología de mínimos cuadrados parciales (PLS) a partir de información de mediciones tomadas por la RMCAB y de índices ambientales de escenas satelitales Landsat 8-9 en la ciudad de Bogotá para el año 2020, 2021 y 2022. Se obtuvieron una totalidad de 49 observaciones para 23 variables que fueron O3, CO, NO2, SO2, PM 2.5, Radiación solar, y temperatura; Temperatura de superficie terrestre y los índices espectrales $EVI, NDVI, Pv, SAVI\ NDMI\ y\ NBR$; los valores de reflectancia de las bandas 1 a 7 Landsat y la ubicación de las estaciones de monitoreo (Longitud y Latitud). Para la regresión se implementó la librería pls en el entorno de programación R modelando la predicción para cada uno de los años considerados y validando el mismo por medio de validación cruzada (CV). Los resultados en la modelación fueron comparados con la metodología de componentes principales (PCR) obteniendo un error cuadrático medio de predicción ($RMSEP$) ligeramente menor para los tres años donde también se consideraron las medidas estadísticas de varianza de las variables independientes (X), la varianza de la variable dependiente u Ozono (Y) y el coeficiente de determinación (R^2) cuyos valores fueron de 86% 87.9% y 71.5% para X; un 65.37%, 51.49% y 45.61% para Y y un R^2 de 0.65, 0.52 y 0.46 para los años 2020, 2021 y 2022, respectivamente. Finalmente, los resultados obtenidos fueron representados como concentración en $\mu g/m^3$ de ozono troposférico medido y predicho concluyendo como zonas más afectadas las localidades de Kennedy y Usaquén en Bogotá D.C.

Introducción.

El ozono es un gas formado en la estratosfera por la reacción entre moléculas de oxígeno cuya presencia para la vida en el planeta tierra es fundamental, dado que hace parte de la capa que evita la entrada de radiación dañina directa sobre el suelo y los seres vivos, sin embargo, el ozono también se genera a nivel de suelo. Conocido como ozono troposférico se forma no necesariamente al entrar en reacción con moléculas de oxígeno sino con contaminantes generados por la actividad humana, principalmente la combustión fósil, óxidos nitrosos y compuestos orgánicos volátiles los cuales en presencia de radiación solar generan un gas dañino constituyente del smog y que afecta la salud de diversos grupos sociales, principalmente el sistema respiratorio. (Jones & Wigley, 1991)

En niños pequeños o personas de la tercera edad las afecciones respiratorias por ozono troposférico pueden ir desde irritación de los pulmones hasta daños permanentes de los mismos, de igual manera puede empeorar enfermedades como el asma en jóvenes quienes al ser una población constantemente activa en entornos al aire libre al realizar ejercicio o la práctica de deportes tienden a exponerse en una mayor frecuencia a este gas. Además, los síntomas de afecciones respiratorias no siempre son evidentes y el daño puede llegar a darse por exposiciones prolongadas a bajas concentraciones sin si quiera notar que algo malo puede estar sucediendo en el organismo. (EPA, 2000)

La elaboración de este trabajo de grado nació de la consideración de la existencia de diversos métodos para medir y presentar valores de concentración de contaminantes donde es predominante la medición in situ. De esta manera, se observó que los avances tecnológicos en obtención de información por medio de percepción remota y los registros de imágenes junto con sus propiedades espectrales representan una herramienta con bastante potencial para cualquier campo de aplicación como lo puede ser la estimación de contaminantes atmosféricos tanto antropogénicos como naturales; teniendo en cuenta que no toda presencia de un contaminante está asociada necesariamente a las actividades humanas.

Por esta razón, el enfoque en el contaminante ozono troposférico y la ciudad de Bogotá fue debido a su comportamiento en ausencia de intervención directa del ser humano, así como en el desarrollo de su química en áreas suburbanas. La idea también fue observar el comportamiento del contaminante en un momento histórico como lo fue la cuarentena debido a la pandemia por covid-19 en el año 2020 donde la actividad humana se vio drásticamente reducida. Con estos aspectos en mente se definió estimar las concentraciones de ozono por medio de mediciones de contaminantes y variables meteorológicas e información espectral del área de estudio, desarrollar un código y metodología de regresión PLS validando los resultados obtenidos para una posterior representación visual de los mismos.

1. Marco teórico

1.1. Ozono troposférico

El ozono troposférico es un contaminante secundario lo cual significa que es generado a partir de la presencia de otros contaminantes primarios, este se genera en condiciones de altas temperaturas y radiación solar afectando la salud humana, en su mayoría por afecciones respiratorias y de especies vegetales. Sus principales precursores son los óxidos de nitrógeno y los compuestos orgánicos volátiles, representados como NOx y COVs, respectivamente y que están presentes en la combustión incompleta de combustibles fósiles y en la evaporación de disolventes químicos, no obstante, la presencia de COVs no es exclusiva de estos compuestos pues hay estudios en los cuales se ha observado que la vegetación misma los genera, dando consigo la generación misma de ozono a nivel de suelo. (Caicedo, Tomás, & Antonio, 2010)

Los dos principales precursores del ozono troposférico reaccionan bajo condiciones de clima seco, altas temperaturas, días soleado y presencia alta de vehículos; su generación es además favorecida bajo momentos de inversión térmica conteniendo la mezcla a nivel del suelo y evitando su dilución. (Seijas, 2004)

1.1.1. Química del ozono

El ozono troposférico puede formarse en dos maneras, en primer lugar, partiendo de los óxidos de nitrógeno, los cuales en presencia de radiación se disocian y el átomo de oxígeno se une a la molécula de oxígeno del aire. El ciclo termina al reaccionar el ozono nuevamente con el monóxido de nitrógeno generando oxígeno y dióxido de nitrógeno. Sin embargo, este tipo de reacción se da en ausencia de intervención antropogénica, con dióxido de nitrógeno formado por procesos naturales en equilibrio con el ozono formado, dando que este se forme, pero disocie de manera muy rápida sin efectos notables en la salud humana. (Seijas, 2004)

$$NO_2 + hv \rightarrow NO + O$$

$$O + O_2 \rightarrow O_3$$

$$O_3 + NO \rightarrow O_2 + NO_2$$

Por otra parte, los dióxidos de nitrógeno formados por actividades antropogénicas forman radicales libres que oxidan el NO haciendo que no pueda reaccionar de manera equilibrada con el O3 e iniciando una reacción en cadena donde cada molécula de COV, por ejemplo, un alcano RH, se oxide en la troposfera, genere dos moléculas de dióxido de nitrógeno y a su vez dos moléculas de O3. (Dominguez, 2012)

$$Radical\ hidroxilo\ (HO) + RH \rightarrow R + H_2O$$
$$R + O_2 \rightarrow RO_2$$
$$RO_2 + NO \rightarrow RO + \boldsymbol{NO_2}$$

$$RO \rightarrow Compuestos\ carbonilicos + H_2O$$
$$H_2O + NO \rightarrow HO + \boldsymbol{NO_2}$$

$$\boldsymbol{2NO_2} + hv \rightarrow 2NO + 2O$$
$$2O + 2O_2 \rightarrow \boldsymbol{2O_3} \uparrow$$

Las principales fuentes de dióxidos de nitrógeno son los vehículos representando un 50% de su totalidad, un 30% corresponde a fuentes fijas como las plantas termoeléctricas y el restante a calderas industriales, incineradores, turbinas de gas, motores estacionarios de Diesel y de encendido por chispa, fábricas de hierro y acero, manufactura de cemento, manufactura de vidrio, refinerías de petróleo, y manufactura de ácido nítrico. (Seijas, 2004). De igual manera las fuentes naturales o biogénicas de óxidos de nitrógeno incluyen los relámpagos, incendios forestales, incendios de pastos, árboles, arbustos, pastos, y levaduras; como especies relevantes se encuentran el Eucalipto,

Encinos y cultivos como el maíz. Estas fuentes diversas producen diferentes cantidades de cada óxido. (Caicedo, Tomás, & Antonio, 2010)

1.2. Percepción remota.

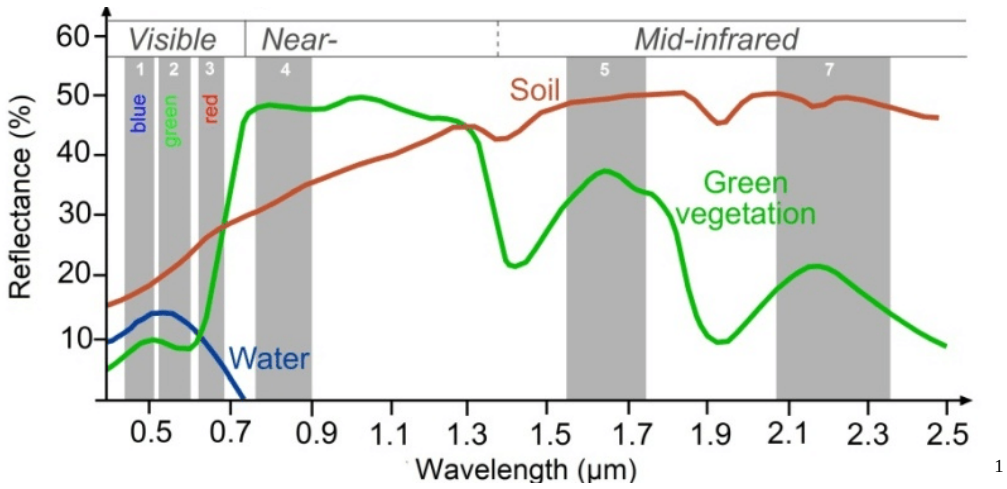
La percepción remota es una técnica en la cual se realiza la captura, tratamiento y análisis de imágenes de la superficie terrestre obtenidas por medio de satélites artificiales, la técnica se enfoca en la obtención de información geográfica. El fundamento principal es la interacción entre la radiación electromagnética y la superficie terrestre, con el sol como fuente de energía. (Santos, 2017)

1.2.1. Radiancia y reflectancia.

La magnitud obtenida por los sensores luego de la energía haber sido reflejada por la superficie terrestre es definida como radiancia, cuando esta se refiere a una porción específica del espectro electromagnético se le denomina radiancia espectral. Por su parte, la reflectancia es la parte de la irradiancia que refleja la superficie receptora, se considera adimensional y normalmente representada en porcentajes, es diferente para cada superficie receptora y para cada cuerpo.

1.2.2. Superficies reflectantes

Las superficies del planeta en el contexto de teledetección se pueden agrupar en tres grandes categorías en función de su química, temperatura, humedad, textura e incluso pendiente, estas serían: suelo, vegetación y agua. Esta importante propiedad de las superficies permite que puedan ser fácilmente diferenciadas, bien sea entre vegetación y suelo e incluso entre tipos de vegetación. El grafico entre tipo de superficie y porcentaje de reflectancia es conocido como firma espectral. (Santos, 2017)



[1]

Figura 1. Firmas espectrales.

1.2.3. Resolución de sensores

Los sensores para obtención de imágenes satelitales cuentan con resolución temporal, espectral, espacial, radiométrica y angular. La resolución temporal hace referencia a la frecuencia con la cual el satélite toma los datos, cuanto tiempo tarda en sobrepasar de nuevo un área; la resolución espectral se relaciona con las zonas del espectro electromagnético que es capaz de detectar el sensor; la resolución espacial se relaciona con el tamaño del pixel y determina la escala en la cual sería necesario trabajar y la resolución angular a la capacidad del sensor para tomar imágenes oblicuas. (Santos, 2017)

1.3. Procesamiento digital de imágenes satelitales.

Este procedimiento consiste básicamente en la mejora de las imágenes a analizar, es decir, recuperación de pixeles o corrección atmosférica. Con el procesamiento digital se busca obtener una imagen lo más cercano a la manera en la cual habría de verse la zona de interés. (Coy, 2021)

Obtener una imagen cuyo ruido radiométrico y geométrico haya sido solucionado es primordial para posterior análisis visual o digital permitiendo mayor enfoque en objetivos y obtención de

[1] Minotti P, 2017 Fluvial wetland classification in De La Plata Basin (mainly in Argentina). Recuperado de: https://onx.la/46551

índices ambientales. De esta manera se hace útil la corrección geométrica, radiométrica y atmosférica como parte de la anterior.

En primer lugar, la corrección geométrica de una imagen satelital busca reducir las distorsiones espaciales, inherentes al proceso de georreferenciación de la imagen; es decir, reducir la diferencia de ubicación de un punto sobre la imagen respecto al mismo punto sobre el mapa de referencia. Por su parte, la corrección radiométrica es una técnica que permite corregir los valores de los niveles digitales de tal manera que demuestren valores en la manera más cercana a cómo debería ser en una toma de datos ideal. También es útil en la recuperación de pixeles perdidos por errores que pueda tener el sensor para lo cual se toman valores de los niveles digitales vecinos. (Coy, 2021) Por otra parte, se realiza la corrección radiométrica del bandeado que corrige el efecto de desplazamiento del histograma de la imagen con el fin de obtener el mismo valor promedio y una misma desviación típica para todas las bandas. (Santos, 2017)

Por su parte, la energía incidente sobre la tierra tiende a ser absorbida o dispersa debido a la presencia de diversos gases atmosféricos y aerosoles como lo son el oxígeno, el ozono atmosférico o el metano y debido a esto al realizar percepción remota, la información obtenida por los sensores satelitales tiende a ser diferente a aquella información real de los objetos terrestres. Dada esta situacion en el preprocesamiento digital se busca reducir dicha interferencia en la información debido a la presencia de la información y obtener así la información más acertada.

1.4. Red de monitoreo de calidad de aire de Bogotá (RMCAB)

La red de monitoreo de calidad de aire de Bogotá cuenta actualmente con 20 estaciones distribuidas en la zona urbana de la capital. En años anteriores contaba con una menor cantidad de estaciones y de instrumentos, pero estos han sido progresivamente añadidos a la red. A continuación, se presenta el mapa de ubicación de las estaciones de monitoreo para el año 2022 así como también los parámetros medibles por la misma. La RMCAB cuenta de igual manera con un sitio web donde se puede visualizar las mediciones más recientes, los pronósticos e informes de contaminación de aire,

así como también presenta la interfaz de solicitud y descarga de datos históricos, pudiendo ser obtenidos en un archivo Excel o su visualización en graficas. (SDA, 2022)

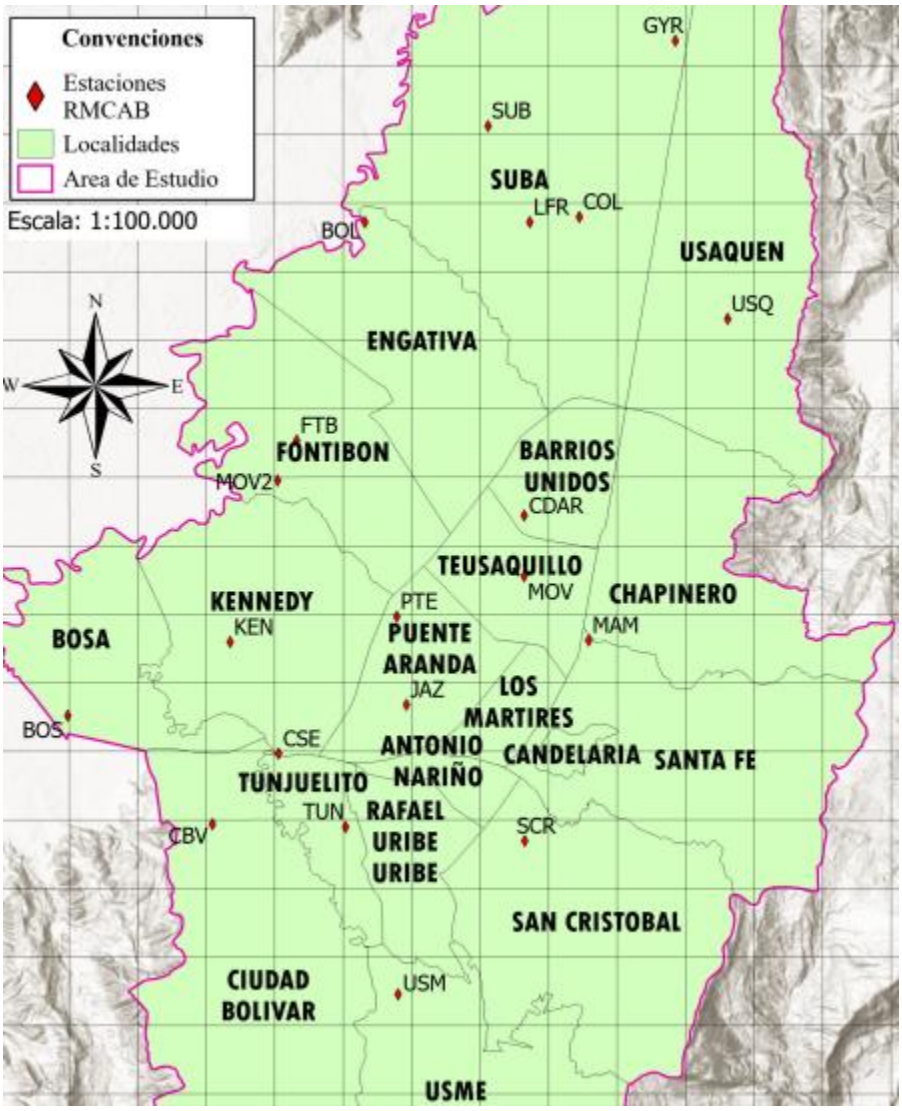


Figura 2. RMCAB año 2022

	Ubicación					Contaminantes						Variables meteorológicas							
Nombre	**Sigla**	**LAT**	**LON**	**Loc**	**T Z**	**PM10**	**PM2.5**	**O3**	**NO2**	**CO**	**Sox**	**BC**	**VV**	**DV**	**T**	**P**	**RS**	**HR**	**PA**
Guaymaral	GYR	4°47'1.52"	74°2'39.06"	Suba	SU	x	x	x	x	x	x		x	x	x	x	x	x	x
Usaquén	USQ	4°42'37.26"	74°1'49.50"	Usaquén	U	x	x	x	x	x	x		x	x	x				x
Suba	SUB	4°45'40.49"	74° 5'36.46"	Suba	SU	x	x	x	x	x	x		x	x	x				x
Bolivia	BOL	4°44'9.12"	74°7'33.18"	Engativá	SU	x	x	x	x	x	x		x	x	x				x
Las Ferias	LFR	4°44'9.12"N	74°4'56.94"W	Engativá	U	x	x	x	x	x	x	x	x	x	x		x	x	x
Centro de alto rendimiento	CDAR	4°39'30.48"N	74°5'2.28"W	Engativá	U	x	x	x	x	x	x	x	x	x	x	x	x		x
Estación móvil 7ma	MOV	4°38'32.75"N	74°5'2.28"W	Chapinero	U	x	x	x	x	x	x		x	x	x	x	x	x	x
Min Ambiente	MAM	4°37'31.75"N	74°4'1.13"W	Santa Fe	U	x	x	x	x	x	x	x	x	x	x				x
Fontibón	FTB	4°40'41.67"N	74°8'37.75"W	Fontibón	U	x	x	x	x	x	x	x	x	x	x		x		x
Puente Aranda	PTE	4°37'54.36"N	74°7'2.94"W	Puente Aranda	U	x	x	x	x	x	x		x	x	x				x
Kennedy	KEN	4°37'30.18"N	74°9'40.80"W	Kennedy	U	x	x	x	x	x	x	x	x	x	x	x	x		x
Carvajal	CSE	4°35'44.22"N	74°8'54.90"W	Kennedy	U	x	x	x	x	x	x	x	x	x	x		x		x
Tunal	TUN	4°34'34.41"N	74°7'51.44"W	Tunjuelito	U	x	x	x	x	x	x	x	x	x	x	x	x	x	x
San Cristóbal	SCR	4°34'21.19"N	74°5'1.73"W	Tunjuelito	U	x	x	x	x	x	x	x	x	x	x	x	x		x
El Jazmín	JAZ	4°36'30.6"N	74°06'53.8"W	Puente Aranda	U	x	x	x	x	x	x		x	x	x		x	x	x
Usme	USM	4°31'55.4"N	74°07'01.7"W	Usme	U	x	x	x	x	x	x		x	x	x		x	x	x
Bosa	BOS	4°36'20.2"N	74°12'14.6"W	Bosa	U	x	x	x	x	x	x		x	x	x		x	x	x
Ciudad Bolívar	CBV	4°34'40.1"N	74°04'10.0"W	Ciudad Bolívar	U	x	x	x	x	x	x		x	x	x		x	x	x
Colina	COL	4°44'14.1"N	74°04'10.0"W	Suba	U	x	x	x	x	x	x		x	x	x		x	x	x
Móvil Fontibón	MOV2	4°40'03.7"N	74°08'55.9"W	Fontibón	U	x	x	x	x	x	x		x	x	x		x	x	x[2]

Tabla 1. Red de Calidad del Aire de Bogotá 2022, SDA.

[2] TZ (Tipo de Zona. U: Urbana, SU: Suburbana); BC (Black Carbon); VV (Velocidad del Viento); DV (Dirección del viento); T (Temperatura); P (Precipitación); RS (Radiación Solar); HR (Humedad Relativa); PA (Presión Atmosférica)

1.5. Análisis multivariado.

En el análisis de datos se hace necesaria la implementación de métodos de regresión para la determinación y estimación de relación entre variables, métodos que permiten la predicción y definición de valores de interés. En la regresión estadística se analiza la relación entre una variable dependiente y una variable independiente. No obstante, es importante considerar que hay situaciones en diferentes áreas de las ciencias en las cuales hay multi variabilidad es decir no una sola variable independiente indica el comportamiento de la variable dependiente.

Por ejemplo, en las ciencias médicas se pueden dar dos casos, en primer lugar, se busca determinar si una persona es propensa a tener altos niveles de colesterol en la sangre; no sería propicio considerar una única variable como lo puede ser el consumo de ciertos alimentos puesto que la edad, peso, composición de la sangre o actividad física influye en la variable de interés. Como segundo ejemplo se puede dar la situacion en la cual se intenta clasificar un grupo de pacientes con infecciones por virus o por bacterias. Una única variable independiente no sería suficiente para poder clasificar dichos pacientes y entre más variables y cantidad de información correlacionada haya será más fácil.

Sin embargo, demasiada información y variables pueden distorsionar el estudio por lo cual se ha implementado el concepto y método de componentes principales en el cual en esencia se busca reducir el número de variables y su dimensionalidad a lo más importante.

En el análisis multivariado se tienen las variables independientes, generalmente representadas por la matriz X y las variables dependientes con la matriz representada por Y de las cuales se busca establecer un modelo lineal como lo puede ser $Y=mX + b$. Siendo Y la variable dependiente, X la matriz de variables independientes, m el vector solución y b los residuos o vector error. Al vector solución también se le conoce como los coeficientes de regresión siendo el valor multiplicable con la matriz independiente que soluciona la multicolinealidad. De esta manera, partiendo de la matriz de variables independientes Y, se toma un porcentaje de valores como datos de entreno del modelo

para su ejecución y el restante porcentaje como conjunto de prueba de validación. El método generalmente utilizado es la validación cruzada indicando de manera asertiva la desviación que podría esperarse al usar el modelo en un nuevo conjunto de pruebas. (Alciaturi, Escobar, De La Cruz, & Rincon, 2003)

La regresión por componentes principales, así como también por mínimos cuadrados parciales busca reducir la multicolinealidad, es decir situaciones en las cuales las variables independientes no difieren lo suficiente entre sí, dificultando distinguir la influencia individual de estas sobre la variable dependiente. Los dos métodos de regresión son útiles, sin embargo, se enfocan a la reducción de multicolinealidad de manera distinta, dado que, por una parte, la regresión por componentes principales (PCR) se enfoca en reducir la dimensionalidad de las variables X mientras que, de otra manera, la regresión por mínimos cuadrados parciales (PLS) se enfoca en la relación entre las variables X e Y. De esta manera el método PLS tiene una mayor utilidad en la predicción. (Aparicio & Caballero, 2009)

1.5.1. Regresión por mínimos cuadrados parciales (PLS).

La regresión por mínimos cuadrados parciales es una extensión de la regresión por componentes principales, otra manera de entender esta técnica es partiendo de esta segunda técnica. La regresión por componentes principales (PCA) se utiliza especialmente para superar el problema de la colinealidad, es decir la correlación entre variables en la regresión lineal, mediante la combinación de variables explicativas o bien, independientes, en un conjunto más pequeño de variables no correlacionadas. Con un ejemplo usando esta metodología se puede apreciar su funcionamiento, suponiendo que hay un conjunto de variables y datos conformados por la edad y el nivel de colesterol como variables independientes y donde se busca estimar el nivel de presión sanguínea como variable dependiente. El problema de colinealidad se evidencia al observar que las variables edad y colesterol se encuentran linealmente correlacionadas y, que no pudiendo ser removida alguna variable debido a contar con importante información o que a partir de bien sea una o la otra

variable se busque hallar la presión sanguínea; ambas entonces deban ser tenidas en cuenta para el análisis. De esta manera las dos variables pueden ser "combinadas", representadas como PC1 o componente principal 1. (Tilevik, 2022)

	PS	**Colesterol**	**Edad**
1	120	126	38
2	125	128	40
3	130	128	42
4	121	130	42
5	135	130	44
6	140	132	46

Tabla 2. Base de datos ejemplo 1.

$$PS = intercepto + colesterol + edad$$

$$PS = intercepto + PC1$$

Partiendo de estos valores de entrada el modelo permite calcular el primer par de valores o pesos relacionadas con cada variable independiente denominado como vector propio o "eigenvector" (α_a) cuyos cuadrados sumados equivalen a 1. De esta manera los pesos de cada variable permiten el cálculo de PC1 siendo reemplazado en la tabla y obteniendo un valor de las variables que combinado representa la máxima varianza de información posible:

$$\alpha_{a1} = 0{,}589$$

$$\alpha_{a2} = 0{,}808$$

$$PC1 = 0{,}589 * colesterol + 0{,}808 * edad$$

$$PC1 = 0{,}589 * 126 + 0{,}808 * 38 = 105$$

$$\alpha_{b1} = 0{,}589$$

$$\alpha_{b2} = 0{,}808$$

Al realizar PCA a un par de variables deberían obtenerse 2 componentes principales, cuando se realiza a 3 variables serian 3 componentes principales y así sucesivamente. Continuando se obtienen

los valores de PC2 a partir de los pesos para cada variable conforme al segundo eigenvector(α_b). Es importante notar que los valores de PC normalmente se presentan como valores centrados, es decir, la media obtenida es restada a cada valor y la nueva media de los valores es 0.

	PS	Colesterol	Edad	PC1	PC2
1	120	126	38	104,918	-79,426
2	125	128	40	107,712	-79,864
3	130	128	42	109,328	-78,686
4	121	130	42	110,506	-80,302
5	135	130	44	112,122	-79,124
6	140	132	46	114,916	-79,562

Tabla 3. Componentes principales PCA

Al calcularse la varianza de los componentes principales se obtienen los valores 12,08 y 0,32 indicando que el primer componente principal aporta la mayor cantidad de información y dado que la idea del PCA es reducir el número de variables o de dimensiones el PC2 puede simplemente ser descartado. Ahora la regresión por PCA se reduce a una regresión lineal donde se buscar hallar el intercepto y la pendiente de la nueva ecuación, cuyos valores pueden ser reemplazados en el modelo PCA:

$$PS = \sigma_0 + \sigma_1 * PC1$$

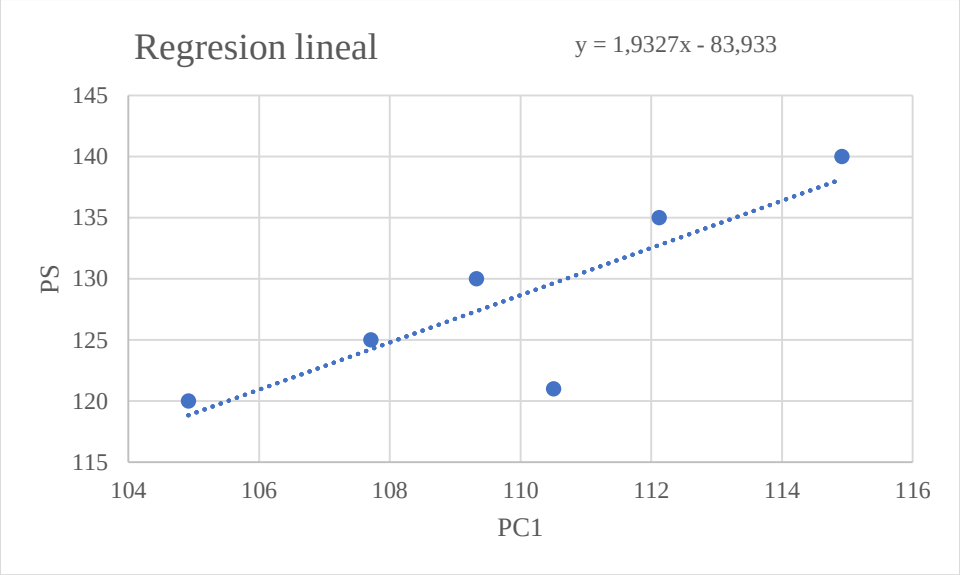


Figura 3. Regresión lineal PCA

$$PS = -83{,}93 + 1{,}932(0.589 * colesterol + 0.808 * edad)$$

$$PS = -83{,}93 + 1{,}14 * colesterol + 1{,}56 * edad$$

De esta manera el modelo PCA queda definido y permite la estimación de información, por ejemplo, identificar la presión sanguínea de una persona de 60 años y un colesterol de 125

$$PS = -83{,}93 + 1{,}14 * 125 + 1{,}56 * 60 = 152{,}17$$

Error cuadrático medio de predicción (RMSEP)

Ahora bien, suponiendo un análisis en el cual se incluya un mayor número de variables e información como el siguiente, donde se añade el peso y la altura como variables independientes se tendrían 4 componentes principales, pero más importante aún es determinar el número de componentes ideal a extraer o usar como anteriormente se hizo al usar solamente uno. Teniendo demasiada información se puede hacer uso de un método de validación automático sin embargo teniendo únicamente 6 individuos puede realizarse un método de validación cruzada para su comprensión.

	PS	Colesterol	Edad	Peso	Altura
1	120	126	38	60	165

2	125	128	40	80	180
3	130	128	42	70	170
4	121	130	42	85	185
5	135	130	44	90	190
6	140	132	46	87	187

Tabla 4. Base de datos ejemplo 2.

En primer lugar, se calcula el modelo PCA para todos los individuos sin tener en cuenta la primera fila de datos, con el modelo obtenido se estima la PS de dicha primera fila $\hat{y}$, este valor se calcula teniendo 0,1,2 y 3 componentes principales extraídos. Luego se toma la medición real de PS de la primera fila y. Esta resta observada en el numerador de la sumatoria de la ecuación muestra que tan lejano se encuentra el valor predicho del real. Posteriormente se realiza el mismo procedimiento tomando como base de datos de partida todos los datos excluyendo la fila número 2, luego la numero 3 y así hasta la fila 6. De esta manera, al tomar la raíz cuadrada del promedio de la sumatoria del cuadrado de estos 6 valores residuales se obtiene el error cuadrático medio de predicción registrado en la siguiente tabla. (Tilevik, 2022)

$$RMSEP = \sqrt{\sum \frac{(\hat{y} - y)^2}{n}}$$

Donde $y = Valor\ real\ de\ la\ variable\ dependiente$

$\hat{y} = Valor\ estimado\ de\ la\ variable\ dependiente$

$n = tamaño\ de\ la\ muestra.$

Componentes extraídos	**RMSEP**
0	8,729
1	8,069
2	5,205
3	6,196

Tabla 5. RMSEP PCR

De esta manera se puede observar que el usar únicamente 2 componentes principales arroja un error de predicción notablemente menor y que no necesariamente al usar más componentes habrá un mejor resultado además que el número de variables o dimensiones no se reduciría.

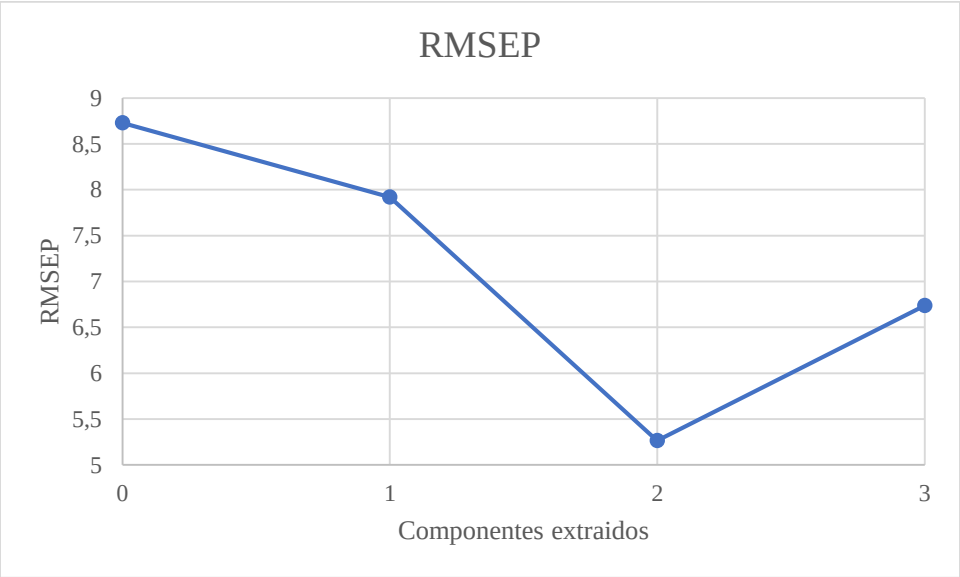


Figura 4. RMSEP PCR.

Tomando entonces dos componentes principales el modelo quedaría de la siguiente manera usando los valores de los vectores principales para cada componente, así como aplicando la regresión lineal para obtener la pendiente e intercepto:

$$PS = \sigma_0 + \sigma_1 * PC1 + \sigma_2 * PC2$$

$$PC1 = 0.120 * colesterol + 0.145 * edad + 0.743 * peso + 0.642 * altura$$

$$PC2 = 0.445 * colesterol + 0.867 * edad - 0.050 * peso - 0.221 * altura$$

$$PS = -63.9189 + 0.2995 * PC1 + 2.6690 * PC2$$

$$PS = -63.918 + 0.2995(0.120 * colesterol + 0.145 * edad + 0.743 * peso + 0.642 * altura) + 2.6690(0.445 * colesterol + 0.867 * edad - 0.050 * peso - 0.221 * altura)$$

$$PS = -63.918 + 1.223 * colesterol + 2.356 * edad + 0.089 * peso - 0.397 * altura$$

De esta manera se obtiene el modelo para la predicción de presión sanguínea para 4 variables y 2 componentes principales por medio de PCA, sin embargo, es importante notar que su cálculo se

realizó combinando variables que poco tienen que ver respecto a la variable dependiente. Por ejemplo, los coeficientes en los componentes principales asignan un mayor peso a la variable *peso* sin ser esta necesariamente la variable que mejor explica la presión sanguínea, como bien lo puede ser el nivel de colesterol, cuyo peso en los componentes es menor. Por esta razón la regresión por componentes principales representa una desventaja frente a la regresión por mínimos cuadrados parciales PLS la cual si tiene en cuenta el problema de pesos.

Ahora ingresando a la explicación de PSL se puede hacer uso de la misma base de datos usada anteriormente y observando la anterior ecuación, los pesos de cada coeficiente relacionado con el colesterol y edad se obtuvieron con el fin de obtener la máxima varianza posible al combinarse en el componente principal ambas variables, no obstante, los valores de cada uno no se relacionan lo suficiente con la variable dependiente y por lo tanto no sería lo más indicado para predecir esta misma. Lo útil seria entonces buscar pesos que también maximicen la predicción de la variable dependiente. En la regresión por mínimos cuadrados parciales se combina en primer lugar el par de variables en una sola, denominada variable latente o bien, componente PLS donde los coeficientes expliquen las variables independientes de la mejor manera:

$$PS = intercepto + LV1$$

$$LV1 = \theta_1 * colesterol + \theta_2 * edad$$

La suma de los coeficientes beta al cuadrado deben sumar 1, justo como en PCA, no obstante, los valores de beta para este caso no corresponderán necesariamente al primer eigenvector como en PCA pues la idea es identificar los coeficientes que representen una mayor covarianza entre la variable dependiente y la variable latente LV1, no la mayor varianza. El cálculo de estos valores óptimos se realiza generalmente por medio de un algoritmo llamado SIMPLS dada la cantidad de datos, pero partiendo de la base de datos de ejemplo básicamente se hace lo siguiente:

$$\theta_1^2 + \theta_2^2 = 1$$

$$\theta_1 = \sqrt{1 - \theta_2^2}$$

Si a θ_2 se le asigna, por ejemplo, el valor de 0,1 entonces $\theta_1 = 0{,}995$ donde reemplazando se tendría lo siguiente:

$$LV1 = 0{,}995 * colesterol + 0{,}1 * edad$$

Ahora se calcula la covarianza entre la PS y este parcial LV1, para estos primeros valores de beta seria 14.14. Entonces, tomando θ_2 como 0,5 θ_1 tomaría un valor de 0,886 y el valor de covarianza para estos nuevos pesos seria de 20.56. Entre estas dos pruebas la segunda sería la mejor opción para la predicción de datos, pero al tomar cualquier valor entre 0 y 1 para θ_1 se podría observar que la máxima covarianza se encontraría en 0,52 y por lo tanto 0,85 para θ_2 obteniendo el modelo la siguiente forma:

$$LV1 = 0{,}52 * colesterol + 0{,}85 * edad$$

	PS	Colesterol	Edad	LV1
1	120	126	38	97,82
2	125	128	40	100,56
3	130	128	42	102,26
4	121	130	42	103,3
5	135	130	44	105
6	140	132	46	107,74

Tabla 6. Valores variable latente LV1

$$PS = \beta_0 + \beta_1 LV1$$

Ahora haciendo uso de regresión lineal, el intercepto sería igual a $\beta_0 = -72.781$ y la pendiente $\beta_2 = 1.958$ entonces:

$$PS = -72.781 + 1.958(0.52 * colesterol + 0.85 * edad)$$

$$PS = -72.781 + 1.02 * colesterol + 1.66 * edad$$

De esta manera se obtiene el modelo de predicción para dos variables explicativas y tomando el mismo ejemplo, el valor para un paciente supuesto de 60 años y un colesterol de 125 el valor de presión sanguínea seria 154,3

Incrementando el número de variables, justo como en el análisis PCA, es importante determinar cuántas variables latentes o componentes sería útil extraer, al realizar el cálculo del RMSEP tomando los valores obtenidos por medio del modelo PLS se obtienen los siguientes valores y se puede observar que el uso de solo dos componentes arrojaría la mejor información:

Componentes extraídos	RMSEP
0	8,729
1	7.920
2	5.265
3	6,737

Tabla 7. RMSEP PSL

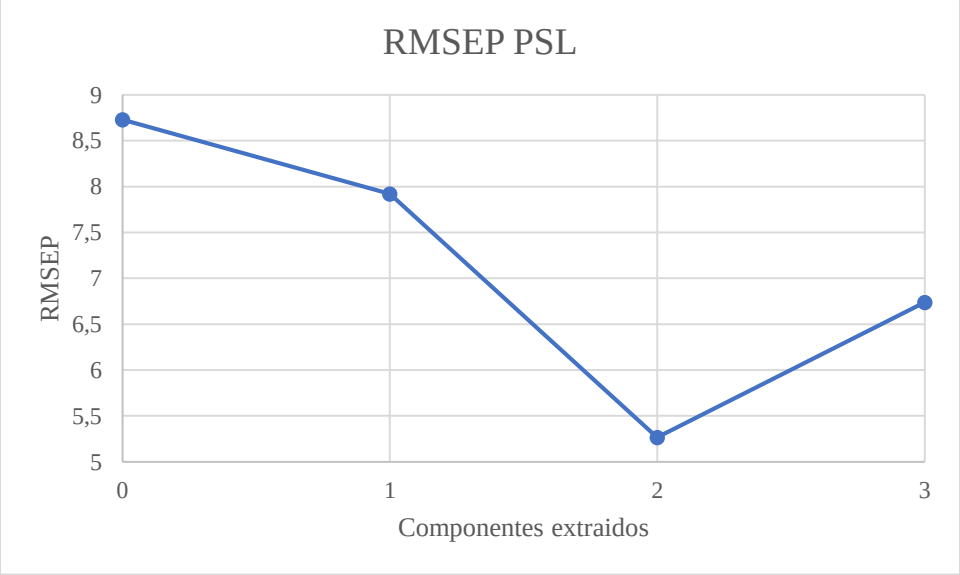


Figura 5. RMSEP PSL

$$PS = \beta_0 + \beta_1 LV1 + \beta_2 LV2$$

El modelo PLS con dos componentes extraídos sería el siguiente:

$$PS = -46.756 + 0.964 * colesterol + 2.546 * edad + 0.033 * peso + 0.327 * altura$$

2. Estado del conocimiento

La regresión por mínimos cuadrados parciales y por componentes principales fue generada en el campo de la medicina, sin embargo, se han llevado a cabo investigaciones similares en las que se busca obtener valores de concentración de contaminantes ambientales, más específicamente en la predicción de la calidad del aire. De esta manera la implementación de PLS y PCR no se restringe a un solo campo de aplicación pues de hecho esta regresión multivariable ha sido relacionada con una alta variedad de conjuntos de información, yendo desde aplicaciones en la industria automotriz identificando predicciones en el comportamiento de un motor conforme a los caballos de fuerza hasta la calidad de la gasolina respecto a su octanaje medido en la reflectancia difusa de la misma. (Alonso, 2017)

De esta manera pueden ser mencionados diversas investigaciones relacionadas con el campo de la contaminación ambiental, en este caso del aire. En primer lugar se tiene la investigación titulada *"Estimación de la concentración de ozono troposférico mediante análisis geoespacial de imágenes satelitales y mínimos cuadrados parciales, para las parroquias urbanas del cantón Quito*" en la cual se tomaron mediciones de diversas variables ambientales relacionadas con la generación de ozono troposférico como SO2, NO2, NOx, humedad, radiación y valores de pixel de índices espectrales así como también de la temperatura de brillo y por medio de una matriz multivariable, compilando información de 3 años, se determinó un modelo PLS de predicción de ozono troposférico. Se obtuvieron salidas gráficas y pudo determinarse las zonas más afectadas, las variables con más relevancia, así como también el grado de error del modelo. (Michelle & Copo, 2017)

Por otra parte, se realizó una investigación similar en el mismo país titulada "Valoración de la concentración de ozono y dióxido de azufre a través de sensores remotos en el área urbana en la ciudad de Cuenca" en el cual los investigadores no realizaron una regresión por mínimos cuadrados

parcial sino por componentes principales, haciendo uso de diferentes 24 variables pero considerando como variable medida por la red de monitoreo de calidad del aire de cuenca únicamente el ozono troposférico y dióxido de azufre; el resto de variables fueron tomadas de las imágenes Landsat 8 y Sentinel 2 , respectivamente. Aproximadamente 500 imágenes satelitales fueron tomadas en consideración y el análisis realizado únicamente a 17 para e ozono, 14 para el dióxido de azufre. El modelo PCR permitió la predicción y graficacion del ozono troposférico validando el modelo con métricas estadísticas como el RMSEP, % de varianza y R2. (Sinchi & Sagal, 2018)

De manera similar, en China se realizó una investigación relacionada con la regresión PLS y el ozono troposférico. En esta investigación se predijo el ozono troposférico haciendo uso de aprendizaje automatizado extremo Kernel o Kernel Extreme Learning Machine (KELM) y de support vector regression machine (SVR) haciendo un pretratamiento a la información por medio de transformada ondícula o wavelet transformation (WT) y PLS; el estudio se realizó para datos de 2014 a 2016 en la zona industrial de Nanjing. Tomaron valores horarios promedio de O3, NO, NO2, CO, COVs, temperatura, presión atmosférica, humedad relativa, velocidad del viento y dirección del viento. La mayor correlación se dio entre el ozono y la humedad relativa. Finalmente concluyeron que la metodología WT y PLS mejoran las habilidades de predicción de KELM y SVR para concentraciones demasiado altas en un 21% y 35%; además que el error absoluto promedio (MAE), el error medio de porcentaje absoluto (MAPE), la raíz del error cuadrático medio (RMSE), la raíz del error cuadrático medio normalizada (NRMSE) y el coeficiente de determinación (R2) fueron 7.71 ppb, 0.37, 9.75 ppb, 11.83% y 0.78, respectivamente. (Xiaoqian, Junlin, Yuxin, Ping, & Bin, 2020)

3. Objetivos.

3.1. Objetivo general

Estimar la concentración de ozono troposférico a partir de datos de imágenes Landsat y mediciones de estaciones de calidad del aire en Bogotá D.C. con el uso de la regresión de mínimos cuadrados parciales PLS.

3.2. Objetivos específicos.

Generar una matriz multivariable a partir de datos de contaminantes ambientales, variables meteorológicas e índices espectrales para el entrenamiento y ejecución de la regresión PLS.

Determinar el modelo de estimación de ozono troposférico para la ciudad de Bogotá D.C con el uso de la regresión por mínimos cuadrados parciales.

Representar espacialmente el ozono troposférico estimado por medio de modelación PLS

4. Metodología.

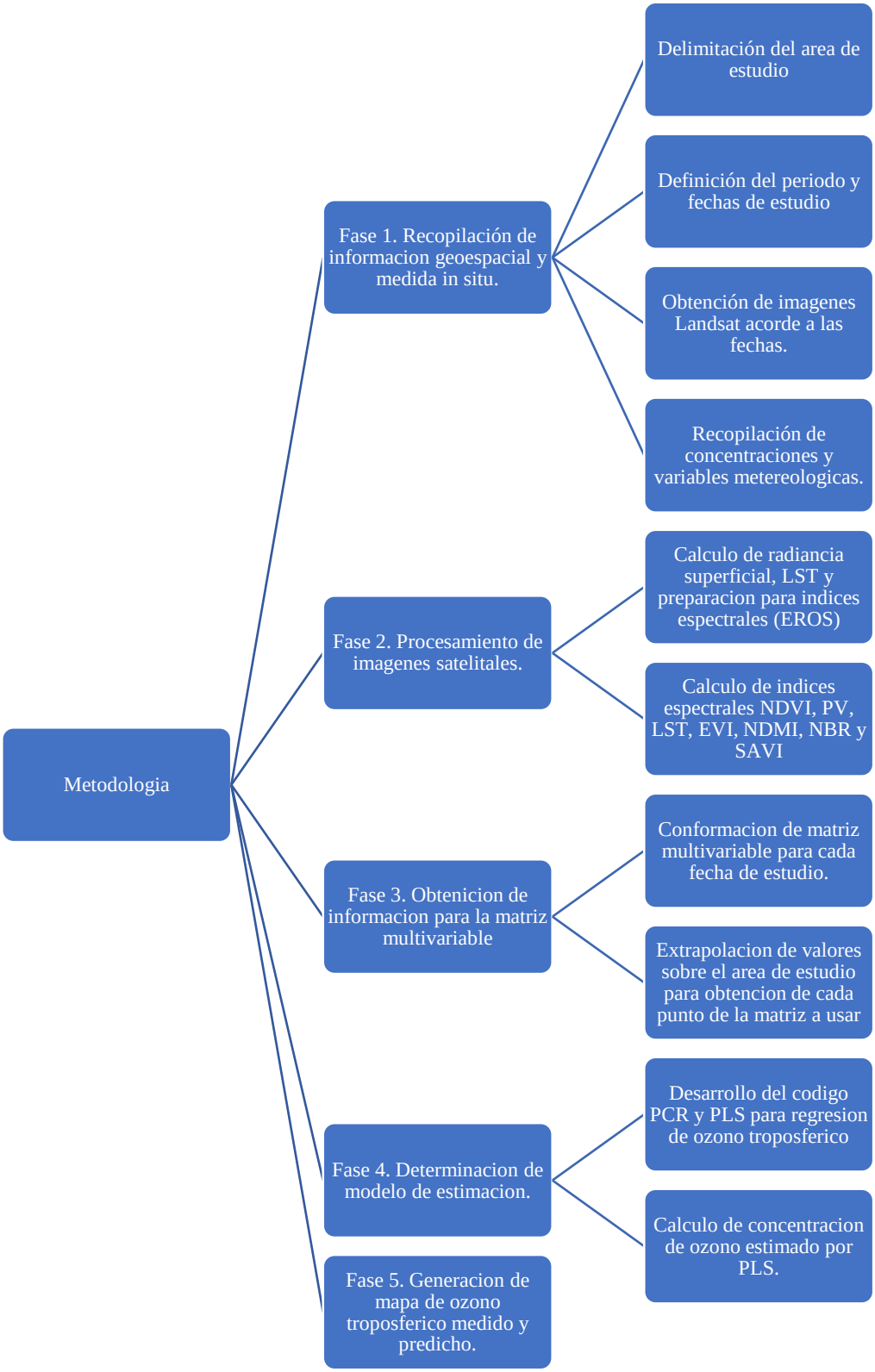


Figura 6. Metodología.

La metodología de este trabajo se conformó por 5 fases en las que de manera general se inició con la recopilación de información geoespacial, índices espectrales y de estaciones de monitoreo de calidad del aire obteniendo así la información base para la matriz multivariable usada para la ejecución de la regresión por mínimos cuadrados parciales que permitió la representación gráfica del ozono estimado en Bogotá D.C. (Figura 6)

4.1. Recopilación de información geoespacial y de mediciones in situ.

En primer lugar, se determinó como área de estudio el área total del distrito capital de Bogotá D.C. la cual cuenta con 20 localidades y la presencia de 20 estaciones de monitoreo de calidad del aire distribuidas principalmente en la urbe, de estas 20 estaciones dos son móviles y para el año 2021 se instalaron las estaciones Colina y Móvil Fontibón. De esta manera se buscó obtener información para los años 2020, 2021 y 2022, teniendo en cuenta que la información medida por la RMCAB para el año 2020 se obtendría de únicamente 12 estaciones para comenzar.

4.1.1. Área de estudio.

El área de estudio incluyó 19 localidades de Bogotá en las cuales se cuenta con estaciones de monitoreo de la calidad de aire, siendo excluida por esta razón y la ausencia de un entorno urbano la localidad de Sumapaz. El área total de estudio fue de 855.400.803,7 metros cuadrados, es decir 855.4 km^2 de los cuales 307 km^2 corresponden a zonas urbanas.

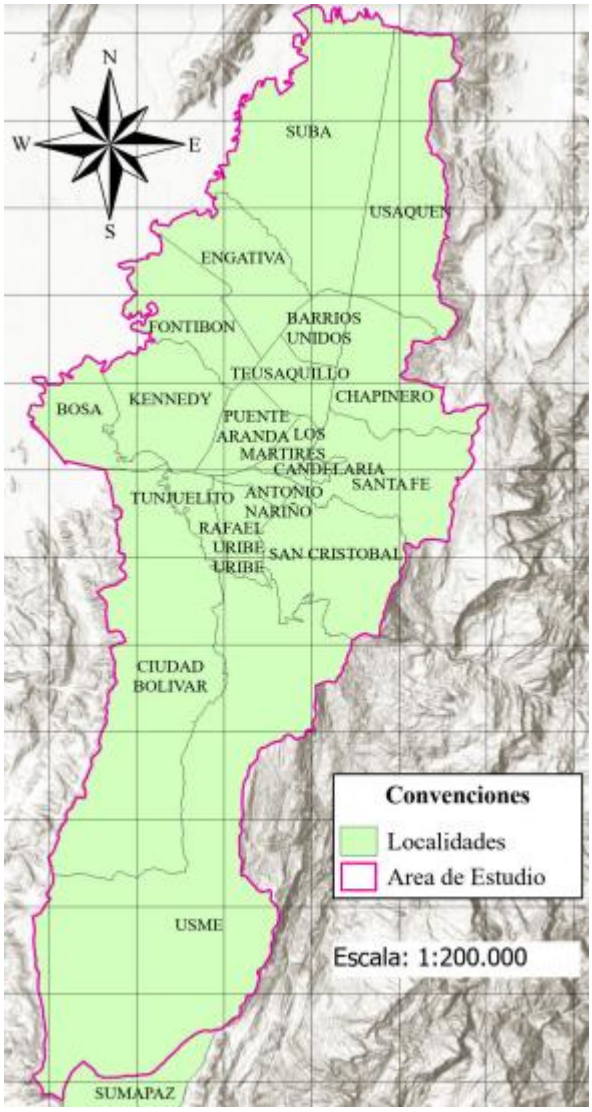


Figura 7. Área de estudio. Elaboración propia.

4.1.2. Imágenes satelitales Landsat

Para la recopilación de imágenes Landsat 8 y 9 se definió como criterio de búsqueda un porcentaje de nubes menor al 20% y las imágenes descargadas hicieron parte de la colección numero dos y nivel 2 presentadas en el catálogo del Servicio Geológico de Estados Unidos USGS, estos productos Landsat se caracterizan por contar con valores de reflectancia superficial y de temperatura superficial en Kelvin, normalmente se denominan por la abreviación en el nombre del producto *SP* (Science Product). Se buscó obtener este tipo de productos debido a la actualización en la

tecnología y a la recomendación de la USGS de migrar la información y metodologías a este formato debido a que la colección número 1 de Landsat dejara de funcionar en el año 2023.

Las imágenes seleccionadas son del 29 de agosto de 2020, 14 de diciembre de 2021 y 31 de enero de 2022. Estas imágenes cuentan con el menor porcentaje de nubes sobre el distrito capital tanto por parte del satélite Landsat 8 como Landsat 9 (Tabla 8).

ID imagen	Fecha	Hora (CST)	% nubes de escena	% nubes área de Estudio
LC08_L2SP_008057_20200829_20200906_02_T1	29/08/2020	15:12	27.57	<15
LC09_L2SP_008057_20211214_20220120_02_T1	14/12/2021	15:13	28.80	<10
LC09_L2SP_008057_20220131_20220202_02_T1	31/01/2022	15:13	19.54	0

Tabla 8. Imágenes Landsat 8/9 seleccionadas.

Previo procesamiento digital a las imágenes, la información requerida para el área de estudio fue obtenida con el uso de la herramienta "*Extract Data*" en el software ArcGIS Pro en la cual se ingresaron las bandas relevantes para el estudio y el polígono del área de estudio, teniendo de esta manera solamente la información a usar. De igual manera se realizó la composición de bandas 1 a 7 para obtener una visualización multiespectral de las imágenes. Es importante notar que la paleta de colores usada en la imagen del año 2022 debido a la ausencia de nubes se reorganizan los rangos de coloración por lo tanto la zona urbana y rural toman estos colores en comparación con las imágenes de los dos años anteriores (Figura 8).

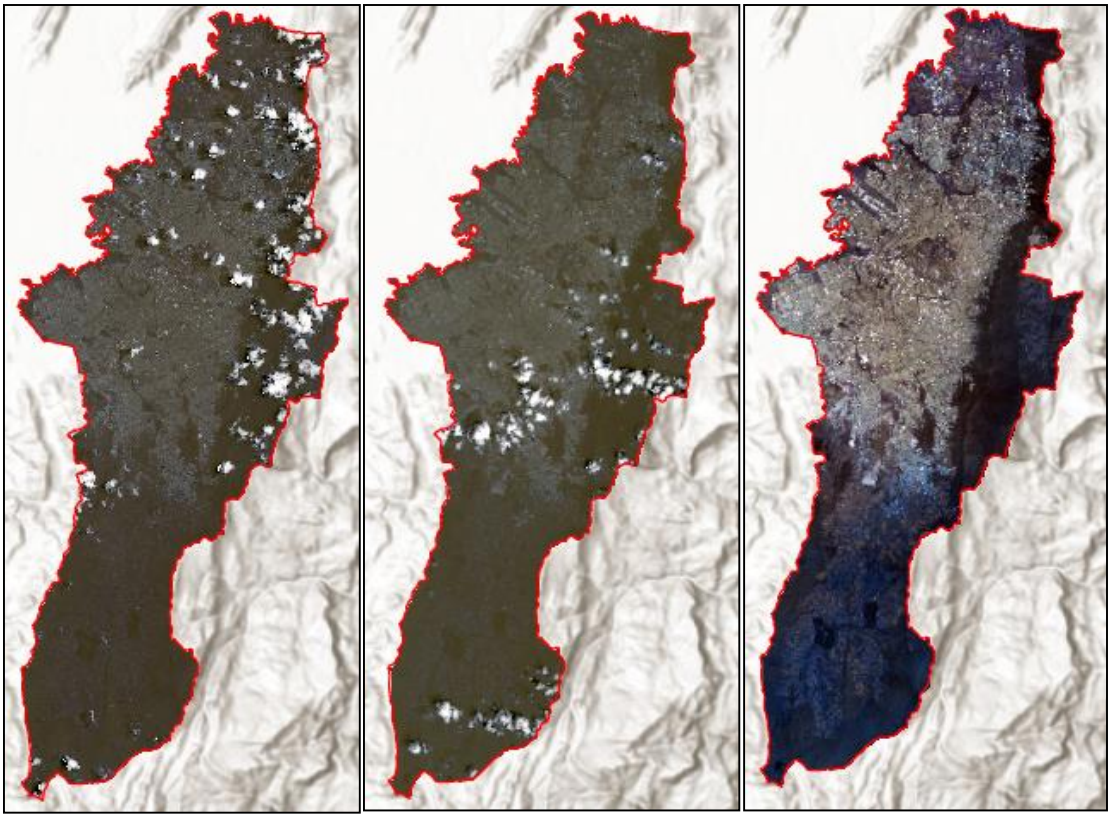


Figura 8. Composición multiespectral 2020, 2021 y 2022

4.1.3. Mediciones de concentración de ozono (RMCAB).

Definidas las fechas de las imágenes satelitales, se procedió a obtener la información medida en las estaciones distribuidas en la ciudad para dichos días haciendo uso del sitio web de la red de monitoreo de la calidad del aire de Bogotá. Como variables para tener en cuenta se definieron el ozono troposférico O_3, la radiación solar, el monóxido de carbono O, dióxido de nitrógeno NO_2, dióxido de azufre SO_2 y la temperatura. Dado que la mayor influencia de la radiación sobre la generación de ozono troposférico se da cerca del mediodía se tomaron los valores promedio de las variables entre las 06:00 am y las 17:00 pm, esta hora final se definió debido a que la imagen Landsat fue tomada entre las 13:00 y 14:00 CST, obteniéndose los valores más cercanos al momento de la toma de la escena sin ignorar el proceso de generación del ozono previamente iniciado en las horas de la mañana.

En primer lugar, para el día 29 de agosto de 2020, las estaciones Tunal y Fontibón no contaban con datos para este día, pero tienen información referente a otras variables para el estudio, además, 4 estaciones entraron en funcionamiento solamente hasta octubre del mismo año, estas fueron Usme, Bosa, Ciudad Bolívar y Jazmín. Por su parte, la estación Bolivia estuvo fuera de funcionamiento a finales de septiembre del 2019 por obras de reubicación volviendo a la toma de datos el 16 de octubre de 2020 y la estación Fontibón presento fallas a lo largo del año, precisamente para esta fecha. Debido a que los porcentajes de datos válidos para estas últimas estaciones es menor al 75%, además de no presentar datos para la fecha requerida, no fueron tenidas en cuenta. (SDA S. D., 2021)

En adición, para el año 2020 algunas de las estaciones tenidas en cuenta contaban con medición parcial de variables, tanto meteorológicas como de contaminantes, debido a esto algunos datos no pudieron ser establecidos por mediciones directamente, no obstante el porcentaje de estos datos no es demasiado alto, además en el informe de calidad del aire del año 2020 de la secretaria de ambiente de Bogotá se presentan algunos valores promedio mensuales por estación, valores que fueron tomados de manera complementaria.

Para aquellos datos faltantes se realizó un procedimiento de interpolación IDW y una posterior extracción de valores para las estaciones cuya infraestructura o situaciones de contingencia no permitió la medición de datos, pero manteniendo el mismo número de estaciones consideradas para el año 2020, es decir 12 estaciones para este caso. A continuación, se presentan los datos obtenidos en la RMCAB y se resaltan en azul los datos obtenidos del informe de calidad del aire de la secretaria de ambiente para el 2020 y en rojo aquellos datos faltantes, posteriormente determinados por medio de interpolación (Tabla 9).

	CO ppm	O_3 ppb	NO_2 ppb	SO_2 ppb	Temp. °C	PM2.5 µg/m3	Rad. Solar W/m²
Carvajal sevillana	1,173	20,919	29,476	3,391	17,3	40	133,6
CDAR	0,397	30,285	14,201	1,191	17,7	19,4	399
Guaymaral	0,556	27,318	9,574	4,564	16,4	20,3	537
Kennedy	0,734	37,818	16,510	2,555	17,4	33,1	132,8
Las Ferias	0,498	20,848	12,879	3,404	16,9	16,8	367,50
Min Ambiente	0,873	17,111	15,760	2,531	17,2	19,3	223,54
Móvil 7ma	1,075	25,171	15,001	1,713	17,7	30	94
Puente Aranda	0,961	14,235	18,694	1,99	15,8	31,8	227,79
San Cristóbal	0,829	23,125	17,553	2,804	15,8	23,2	493
Suba	0,437	25,556	12,234	1,948	17,2	18,4	410,05
Tunal	0,699	29,618	14,894	2,609	15,6	24,7	349
Usaquén	0,520	29,618	17,021	11,738	17,1	20,1	345,94

Tabla 9. Datos 29 agosto de 2020 RMCAB.

Para el año 2021 la cobertura y equipos de mediciones aumentaron en el distrito capital, sin embargo, debido a diversas situaciones algunas estaciones obtuvieron un porcentaje de datos validados menor al 75% por lo cual no se tuvieron en cuenta; dichas estaciones fueron Puente Aranda y Bosa con información incompleta debido a eventos climáticos y a disturbios sociales, respectivamente, para este año se consideraron entonces solamente 18 estaciones. (SDA, 2021)

De manera similar al año anterior, algunas estaciones contaron con equipos de medición, pero había datos faltantes que debieron ser completados con el uso del informe anual de calidad del aire de Bogotá del año 2021 y en aquellas estaciones con información relevante, pero datos faltantes debido a falta de equipos, se obtuvo los valores interpolados con la herramienta "*Extract Multi Values to Points*" del software ArcGIS pro. A continuación, se presentan los datos obtenidos en la RMCAB y se resaltan en azul los datos obtenidos del informe de calidad del aire de la secretaria de ambiente para el 2021 y en rojo aquellos datos faltantes, posteriormente determinados por medio de interpolación (Tabla 10)

	CO ppm	O_3ppb	NO_2ppb	SO_2ppb	Temp. °C	PM 2.5$\mu g/m^3$	Rad. Solar W/M²
Carvajal sevillana	2,784	10,660	26,596	6,935	17,4	35	350,64
CDAR	0,639	21,378	16,047	1,677	18,3	18,3	407,00
Guaymaral	0,446	25,774	4,465	5,289	17,3	25,2	446,00
Kennedy	1,342	16,065	26,766	4,659	16,5	38,8	375,35
Las Ferias	1,080	20,681	18,691	5,890	16,8	20,6	254,11
Min Ambiente	1,029	19,531	24,422	3,641	17,9	21,3	237,87
Móvil 7ma	1,846	20,320	5,319	2,701	18,7	24,8	45,00
San Cristóbal	0,746	16,176	20,326	4,111	16,3	26	517,00
Suba	0,524	24,504	12,661	6,666	17,1	23	310,02
Tunal	2,003	18,607	31,966	4,144	17,2	40,9	328,00
Usaquén	0,854	21,583	18,668	1,908	16,9	19,7	311,81
Bolivia	0,338	14,898	19,369	6,984	17,0	21,4	341,28
Ciudad Bolívar	1,156	23,206	30,645	4,646	14,9	35,4	351,00
Colina	0,501	10,538	16,339	5,345	16,3	12	235,00
Fontibón	0,495	27,831	19,149	8,578	17,0	25,4	488,70
Jazmín	1,127	21,013	30,098	2,164	16,7	27,2	339,00
Móvil Fontibón	0,495	17,059	22,486	8,578	17,0	27,1	515,00
Usme	0,493	11,456	16,170	3,361	17,8	23,8	446,00

Tabla 10. Datos 14 de diciembre de 2021 RMCAB

Por su parte, para enero del año 2022, la estación de monitoreo de Puente Aranda volvió a su funcionamiento regular después de la sobrecarga eléctrica debida a la caída de un rayo en el año 2021; sin embargo, las demás estaciones con datos faltantes continuaron siendo las mismas y de esta manera se consideraron únicamente 19 estaciones, dejando por fuera la estación Bosa en los 3 años. Debido a esto los valores debieron ser nuevamente interpolados para su obtención y también ciertos datos debieron ser tomados del informe, en este caso mensual, de calidad de aire de enero del 2022 de la secretaria distrital de ambiente (SDA, 2022) (Tabla 11).

	$CO\ ppm$	$O_3 ppb$	$NO_2 ppb$	$SO_2\ ppb$	$Temp\ °C$	$PM\ 2.5\ \mu g/m^3$	Rad Solar
Carvajal sevillana	0,770	14,026	17,893	4,28	18	23,5	538,8
CDAR	0,623	21,788	25,111	1,857	17,1	21,9	490,5
Guaymaral	0,303	29,804	3,208	1,873	17,1	17,9	611
Kennedy	0,494	20,231	15,621	1,352	16,2	23,4	533,9
Las Ferias	1,621	21,029	24,762	1,815	15,9	34,5	560,1
Min Ambiente	0,823	25,509	21,509	1,961	16,9	22,4	503,2
Móvil 7ma	1,319	22,500	6,975	1,829	18	26,3	475
San Cristóbal	0,491	24,283	16,531	2,092	15,4	21,3	577
Suba	0,485	22,233	9,746	1,876	16,7	18,9	561,2
Tunal	0,787	24,398	14,672	1,571	17,1	17,3	553
Usaquén	0,502	30,414	18,118	1,915	16,8	20,4	543,6
Bolivia	0,646	15,858	17,492	1,435	16,3	20,8	543,4
Ciudad Bolívar	1,037	25,003	19,724	2,534	14,3	30,3	557
Colina	0,504	9,851	17,886	1,841	15,7	15,7	563
Fontibón	0,076	20,924	24,491	2,295	16,2	31	551,2
Jazmín	0,702	19,724	18,954	2,29	15,9	21,2	492
Móvil Fontibón	0,254	22,91	15,259	2,190	16,2	34,4	557
Usme	0,607	25,409	20,215	1,932	15,6	21,6	492
Puente Aranda	0,784	21,695	24,932	1,055	15,6	20,4	536

Tabla 11. Datos 31 de enero de 2022 RMCAB

4.2. Procesamiento de imágenes satelitales

Una vez recortadas las imágenes Landsat conforme al área de estudio se realizó la re-proyección y re-escalamiento con el fin de garantizar un correcto sistema de referencia, para este caso UTM zona

18N EPSG 32618 sin embargo manteniendo el mismo tamaño de píxel de origen, es decir 30 m. Respecto al rescalamiento se obtuvo imágenes de 16 a 8 unsigned bits con lo cual el rango de valores que serían almacenados sería menor y con valores positivos. Este re-escalamiento es útil para el espacio en memoria usado con cada proceso realizado en el software ArcGIS, no obstante, al requerir el cálculo de diferentes índices espectrales el tipo de píxel y su profundidad debió ser floating y 32 bits, respectivamente, esto debido a los valores de los índices, generalmente entre -1 y 1.

4.2.1. Cálculo de índices espectrales como insumos para la matriz multivariable

Al contar con datos Landsat 8 y 9 de la colección 2 y nivel 2, science products (SP), las escenas descargadas ya contaban con corrección atmosférica, con valores de reflectancia y temperatura superficiales por lo cual no se realizó el cálculo de radiancia o valores a nivel de atmósfera o superficial. De igual manera los niveles digitales con los cuales se descargó la imagen requieren de escalamiento con el uso de factores multiplicativos y valores aditivos que se encuentran en los archivos MTL de las escenas Landsat mostrados a continuación (Tabla 12).

Banda	**Unidades**	**Factor multiplicativo de escala**	**Factor de compensación aditivo**
ProductID_SR_B1-B7	Reflectancia	0.00341802	-0.2
ProductID_ST_B10	Kelvin	0.00341802	149

Tabla 12. Factores de corrección de bandas, Landsat 8/9 C2L2

Los índices ambientales pueden ser obtenidos haciendo uso de la herramienta "*Ráster Calculator*" en ArcGIS, introduciendo las fórmulas de cada índice y cada banda con el rango de espectro electromagnético asociado. Al realizar este procedimiento con los valores calculados de reflectancia para cada banda los resultados no eran lógicos inicialmente y se encontraban fuera de las categorías propias de cada índice por lo cual se pudo aclarar que para el cálculo de índices espectrales con bandas de la colección y nivel 2 de Landsat, o science products, no es necesario este escalamiento pues la imagen ya cuenta con los valores necesarios para su cálculo directo. Esto pudo ser

confirmado al hacer uso de la herramienta EROS Science Processing Architecture (ESPA) del Servicio Geológico de Estados Unidos que permite solicitar índices espectrales calculados directamente por esta institución. A continuación, se presenta la manera en la cual se solicitó dicha información en la página web https://espa.cr.usgs.gov/index/:

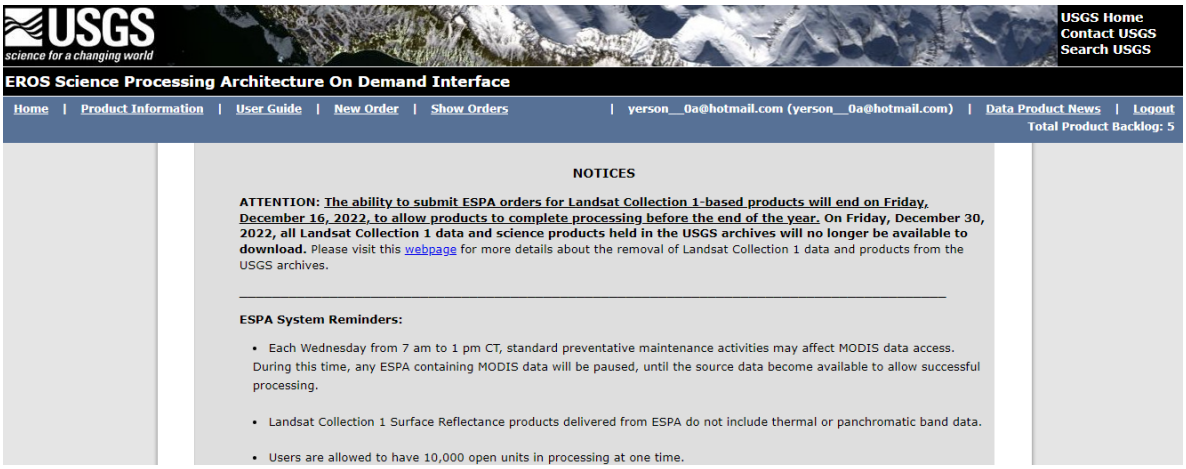


Figura 9. Página de inicio EROS. USGS

En primer lugar, es necesario contar con una cuenta en el USGS y encontrarse en una sesión abierta en la página web para accedes a sus servicios. La información "*on demand*" o a pedido se encuentra en la opción "*bulk ordering – order data*" y posteriormente se solicitó el ID de la escena o lista de escenas a usar en un archivo de texto, así como también los parámetros y la información o índices solicitados.

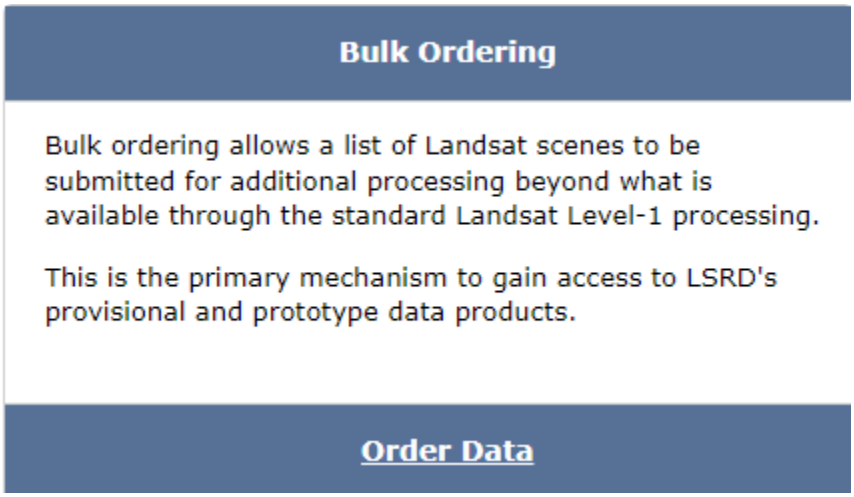


Figura 10. Opción de solicitudes on demand EROS

Add Input Products (Show Available Products)

Scene List

Seleccionar archivo Ninguno archivo selec.

Select Product Contents

Source Products

☐ Input Products

Additional Processing

Landsat Level-2 Products

☐ Surface Reflectance - *Not available for thermal or panchromatic bands*

☐ Provisional Aquatic Reflectance - *Only Available for Landsat 8 & 9*

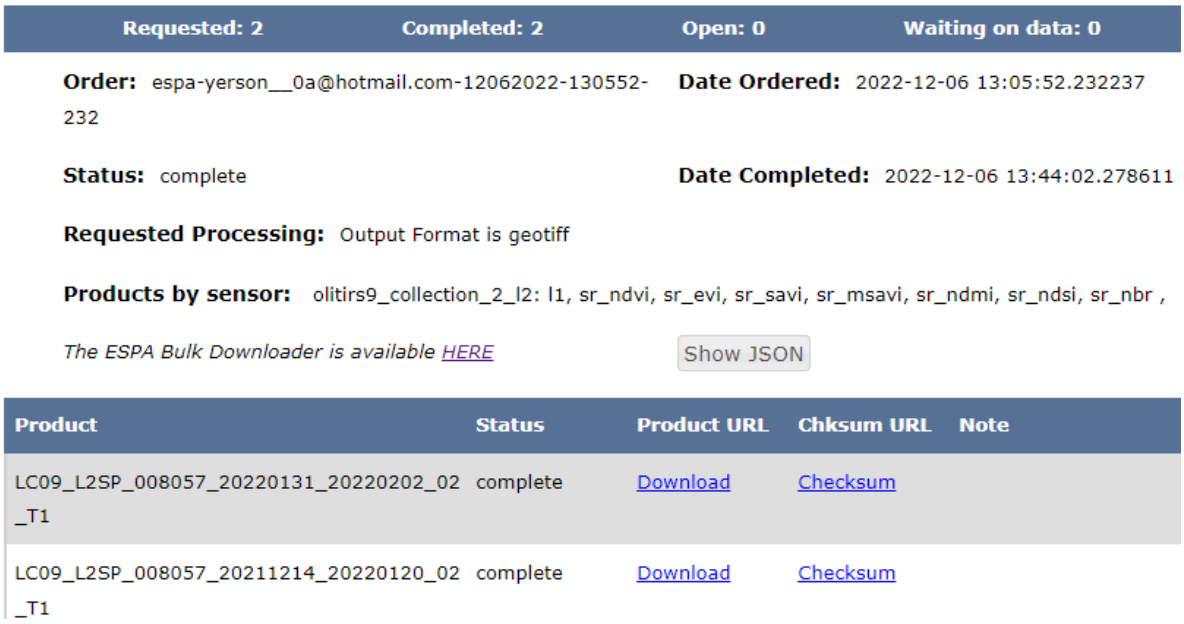
☐ Top of Atmosphere Reflectance

☐ Brightness Temperature - *Thermal band TOA processing*

☐ Spectral Indices

Figura 11. Parámetros de ordenes EROS

Una vez solicitada la informacion y dependiendo de la cantidad de escenas puede tardar algunos minutos en ser completada, por medio de correo electronico se notifica este proceso y se encontrara asi lista la informacion para ser descargada.



Requested: 2	Completed: 2	Open: 0	Waiting on data: 0

Order: espa-yerson__0a@hotmail.com-12062022-130552-232 **Date Ordered:** 2022-12-06 13:05:52.232237

Status: complete **Date Completed:** 2022-12-06 13:44:02.278611

Requested Processing: Output Format is geotiff

Products by sensor: olitirs9_collection_2_l2: l1, sr_ndvi, sr_evi, sr_savi, sr_msavi, sr_ndmi, sr_ndsi, sr_nbr ,

The ESPA Bulk Downloader is available HERE Show JSON

Product	Status	Product URL	Chksum URL	Note
LC09_L2SP_008057_20220131_20220202_02_T1	complete	Download	Checksum	
LC09_L2SP_008057_20211214_20220120_02_T1	complete	Download	Checksum	

Figura 12. Estado de ordenes EROS.

LC09_L2SP_008057_20211214_20220120_02_T1_ST_NDVI.TIF

Arriba se indica la sintaxis en la cual se presentan los productos solicitados y el índice asociado al final de este.

Una vez obtenidos los índices espectrales para cada año se procedió a realizar la extracción de la información conforme a la extensión del área de estudio. Posteriormente, dados los rangos de la información obtenida del USGS fue necesario usar el valor multiplicativo para cada índice presentado en la guía del usuario de la interfaz de solicitud del EROS Science Processing Architecture (ESPA), de manera general este factor fue 0,0001, a continuación, se muestra el procedimiento para la obtención del NDVI; además, justo como con las variables y contaminantes ambientales, se extrajeron los valores de cada ráster en cada estación considerada. Para esto se hizo uso de la herramienta "*Extract Multi Values to Points"*. (USGS U. S., 2022)

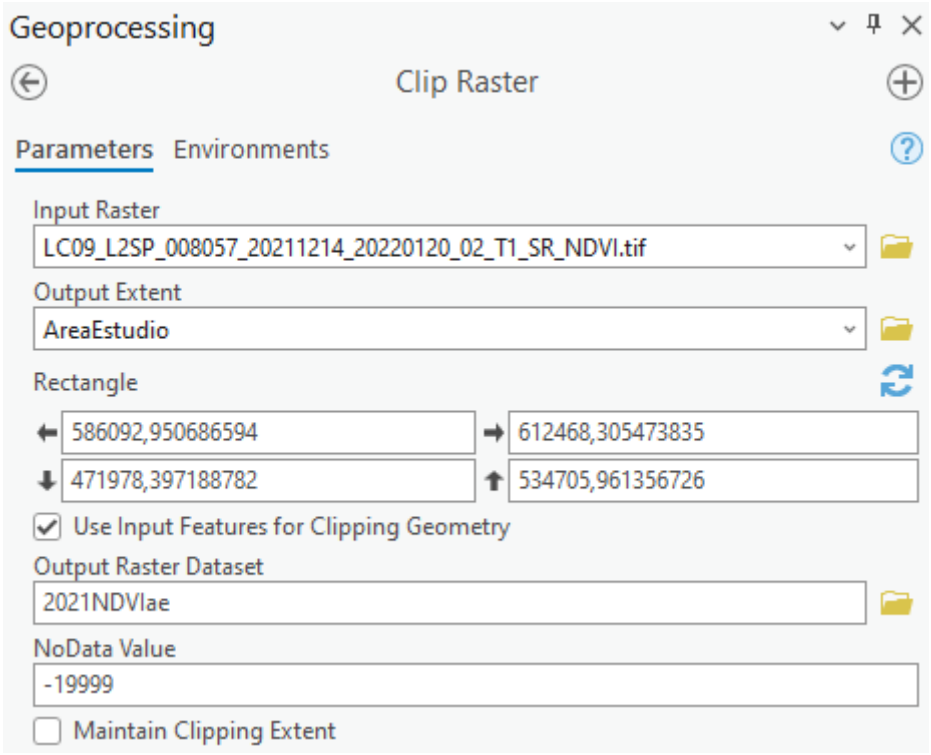


Figura 13. Herramienta Clip ArcGIS pro.

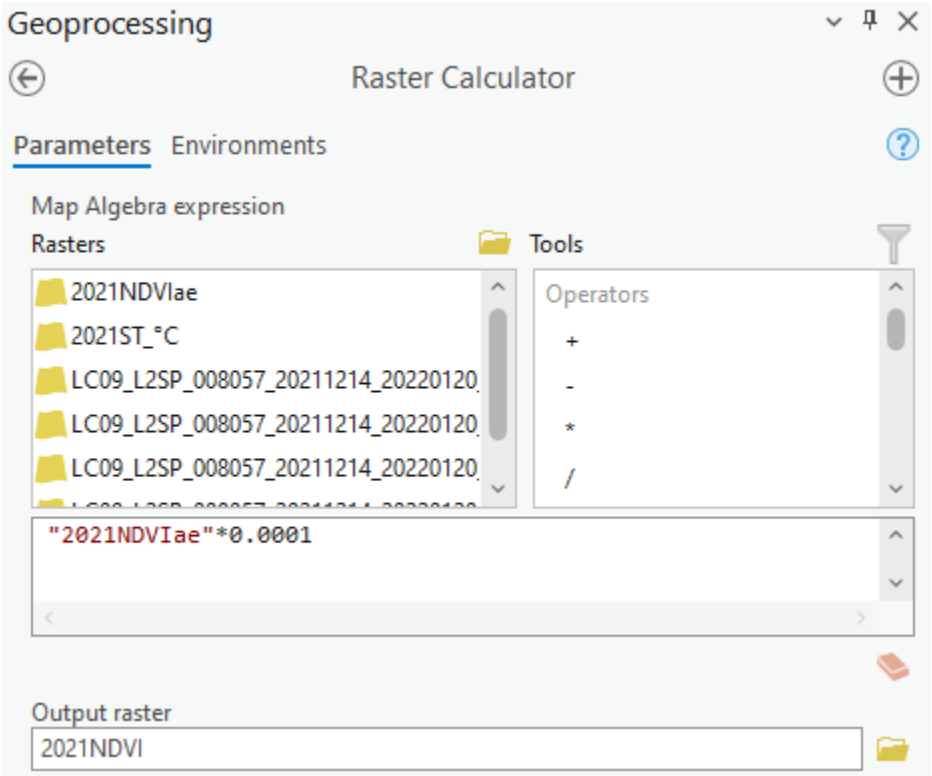


Figura 14. Herramienta Ráster Calculator ArcGIS pro

Figura 15. Diciembre 2021. Índice de vegetación de diferencia normalizad (NDVI)

Los índices espectrales disponibles por el USGS son el Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Soil Adjusted Vegetation Index (SAVI), Modified Soil Adjusted Vegetation Index (MSAVI), Normalized Difference Moisture Index (NDMI), Normalized Burn Ratio (NBR), Normalized Burn Ratio 2 (NBR2), Normalized Difference Snow Index (NDSI) (Collection 2 only) y para el análisis se utilizaron únicamente **el NDVI, Pv, EVI, NDMI, SAVI** y **NBR**. Por esta razón solamente el índice de proporción de vegetación (Pv) tuvo que ser calculado manualmente de la siguiente manera:

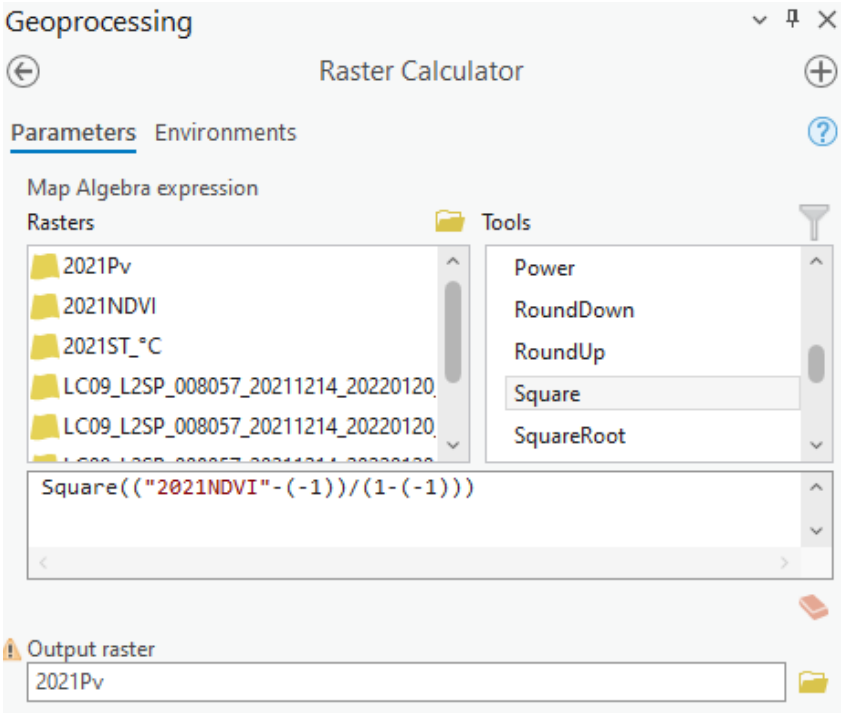


Figura 16. Cálculo del Pv. 2021

Figura 17. Índice de proporción de vegetación. 2021

A continuación, se presentan los demás índices espectrales obtenidos para el año 2021

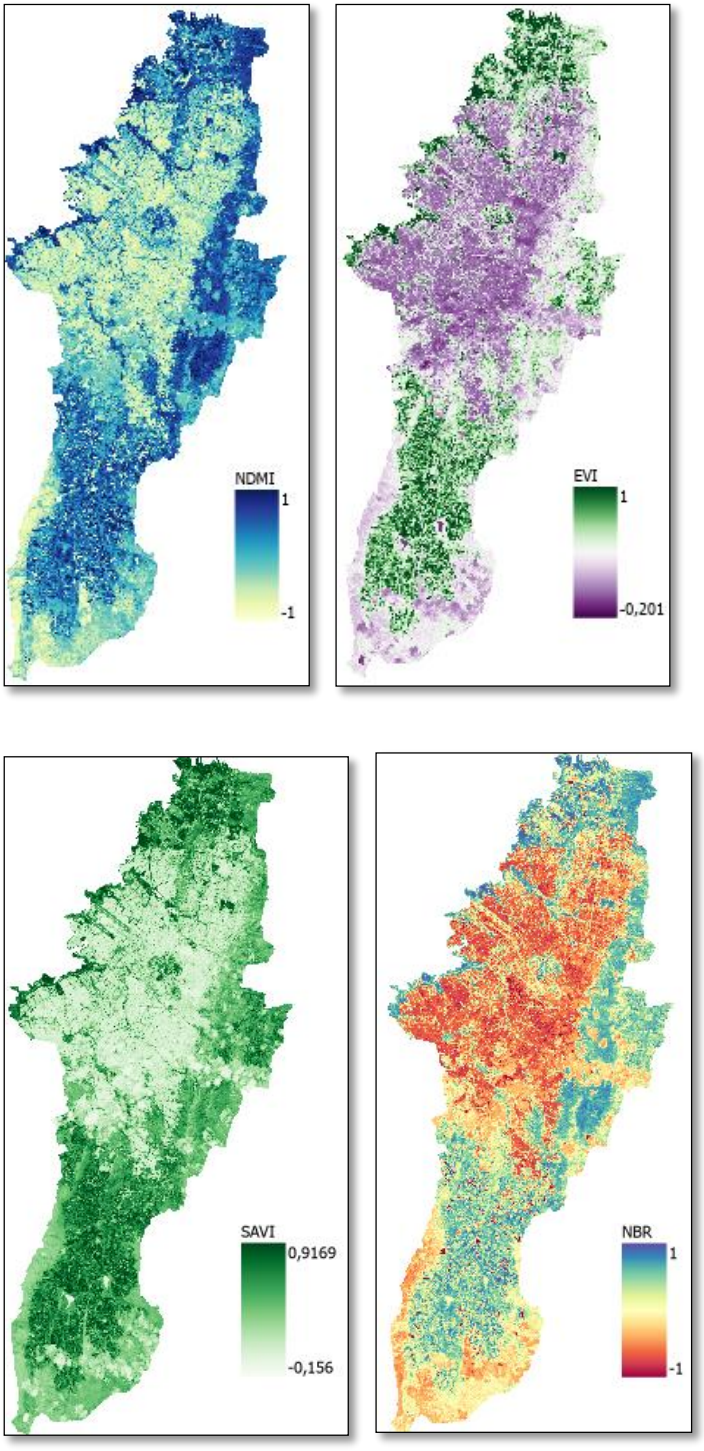


Figura 18. Índices espectrales año 2021

De manera similar para el cálculo de la temperatura superficial, obtenida con la banda 10_ST fue necesario realizar el escalamiento con los factores correspondientes presentados en la tabla 12, pudiendo ser obtenidos los valores en Kelvin o grados centígrados para una mejor visualización.

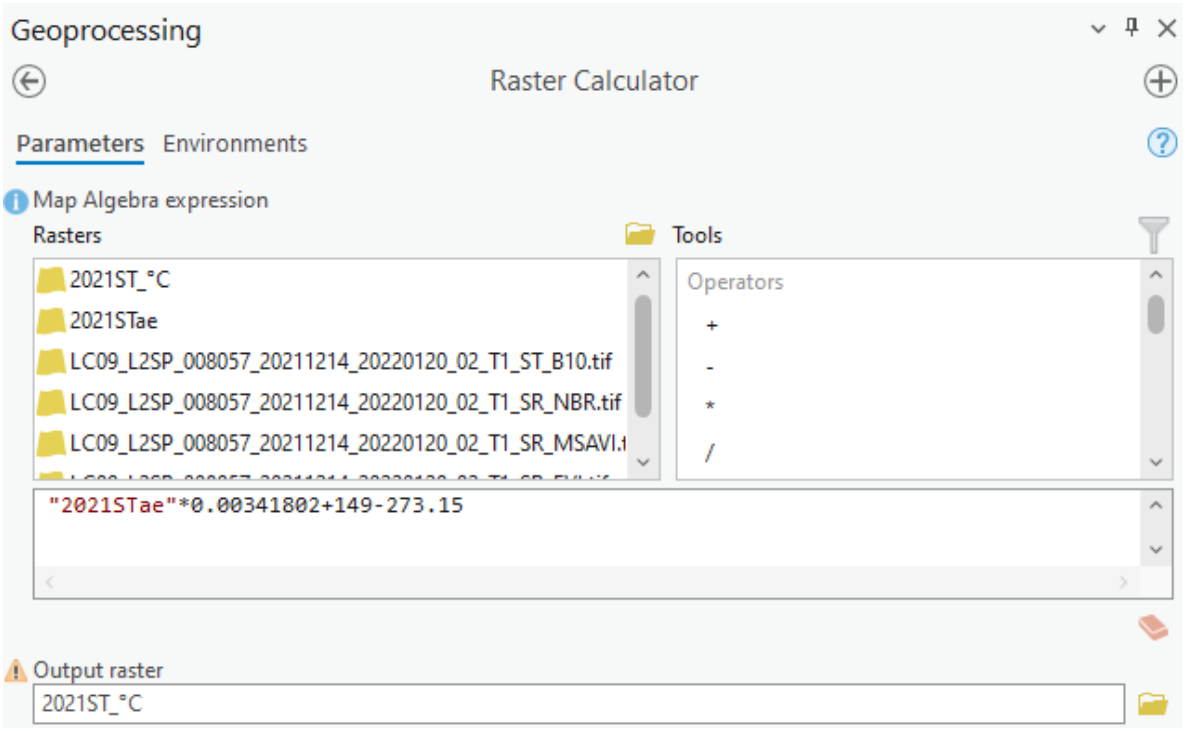


Figura 19. Calculo temperatura superficial en grados centígrados.

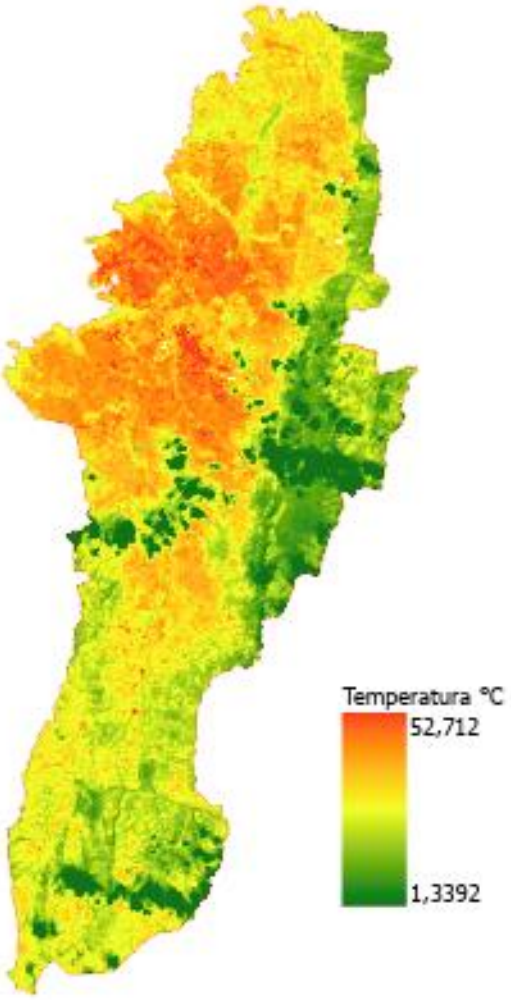


Figura 20. Temperatura superficial diciembre 2021 (°C)

Los valores de reflectancia superficial también fueron determinados con el uso de factores como se muestra a continuación.

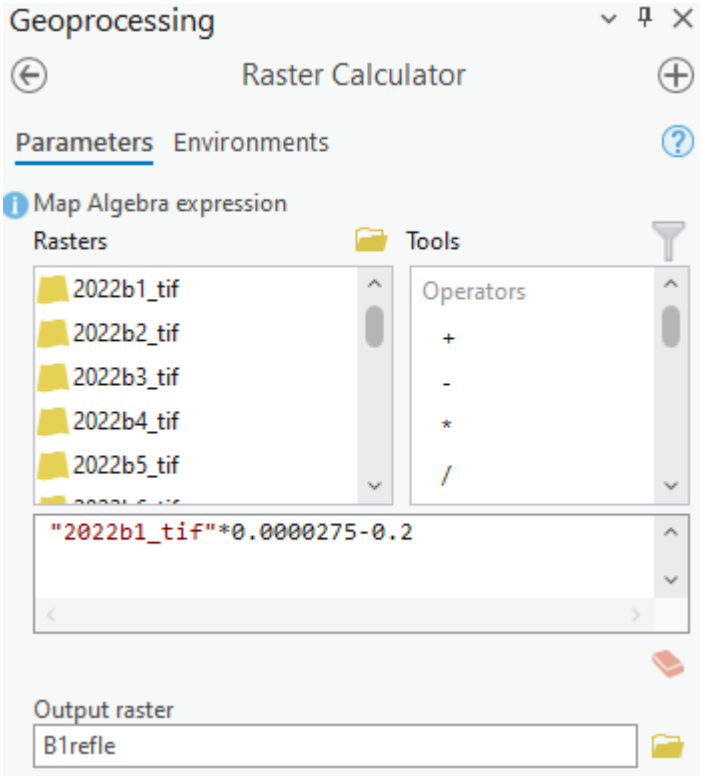


Figura 21. Cálculo de reflectancia para cada banda

4.3. Matriz multivariable.

Para la obtención de los datos a usar en el modelo de predicción PLS el propósito fue obtener datos de cada uno de los contaminantes, ubicación geográfica, índices espectrales y valores de reflectancia en toda el área de estudio para cada uno de los años. En primer lugar, se generó una red de puntos sobre el área de estudio cada 30 metros conforme a la resolución espacial de las imágenes Landsat con el uso de la herramienta “Create Fishnet” coincidiendo con el centro de cada píxel de un ráster de referencia previamente ajustado al área de estudio, por ejemplo, los valores de reflectancia de la banda 1 del año 2021. Luego, con la herramienta de selección por atributos se exportaron aquellos puntos de la red que estuvieran solamente dentro del área de estudio.

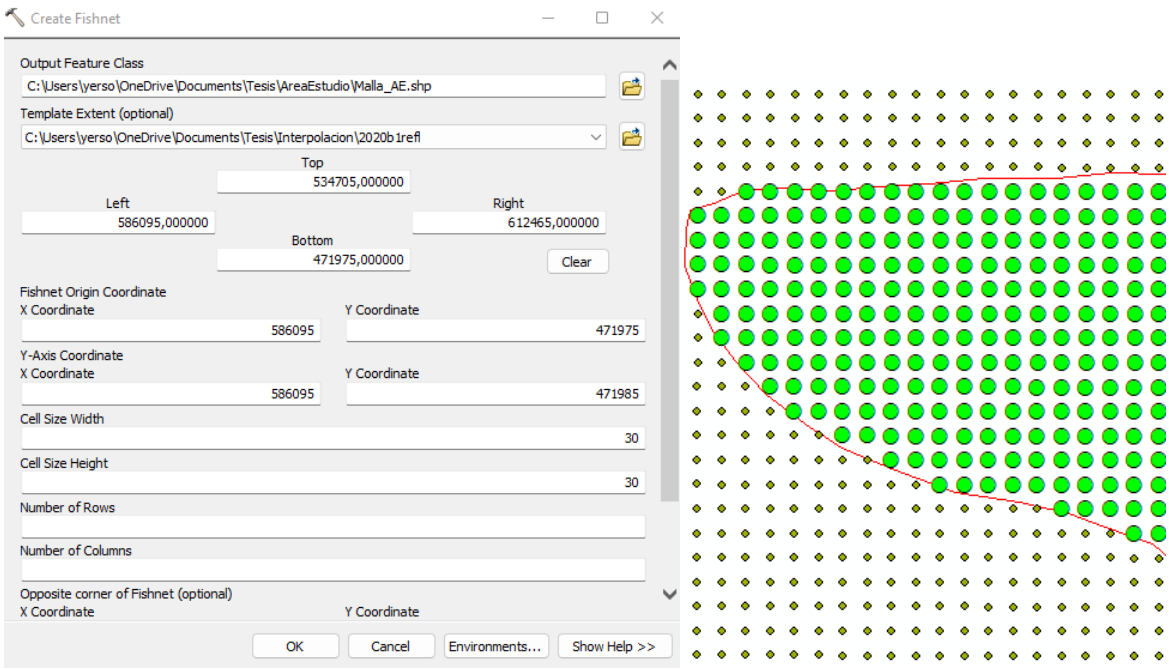


Figura 22. Herramienta Create Fishnet. Desktop ArcMap

De igual manera, fue necesario calcular la posicion de cada uno de los puntos en la malla con el uso de la herramienta “Add XY Coordinates” de ArcGIS pro obteniendo asi una capa de puntos, sobre el centro de todos los raster con un total de 948852 puntos. Esta malla fue usada tambien para los procesos de los años 2021 y 2022.

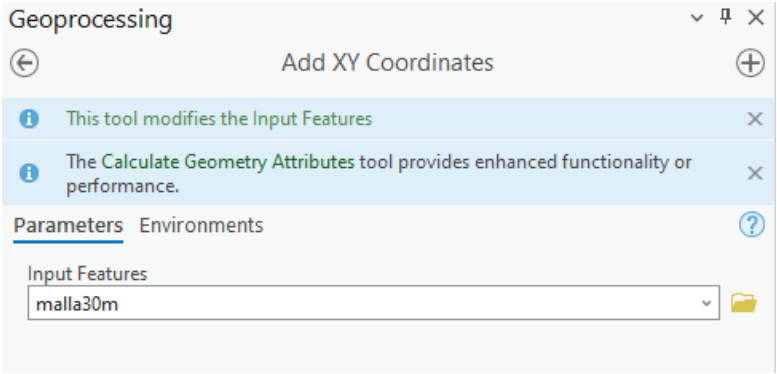


Figura 23. Herramienta Add XY Coordinates. ArcGIS pro

Por su parte, a partir de los datos de contaminantes medidos se realizó una interpolación IDW con el uso de la herramienta de ArcGIS pro “Geostatistical Wizard” obteniendo valores solo para el área cubierta por aquellas estaciones en funcionamiento para cada año. Luego, con la herramienta “GA

Layer to Points” se extrapolaron los valores de cada contaminante conforme la malla del área de estudio.

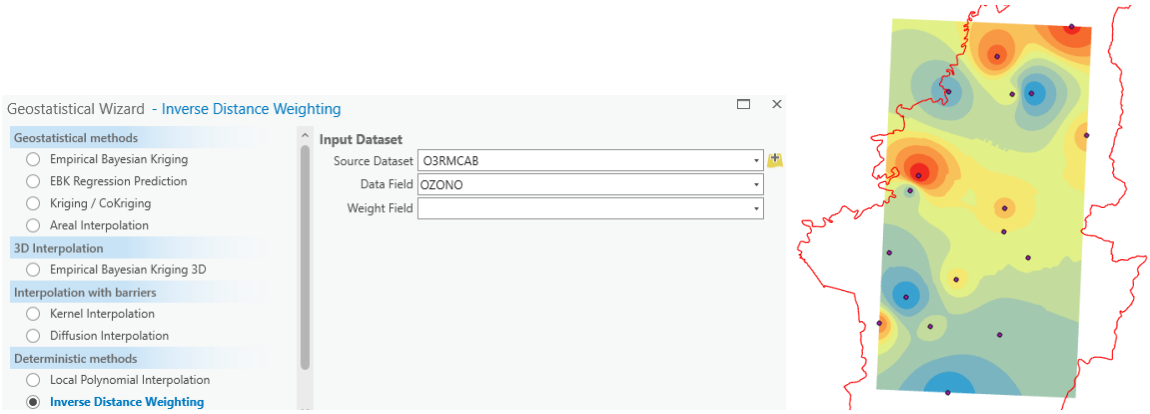


Figura 24. Herramienta Geostatistical Wizard ArcGIS pro. Ozono RMCAB año 2021

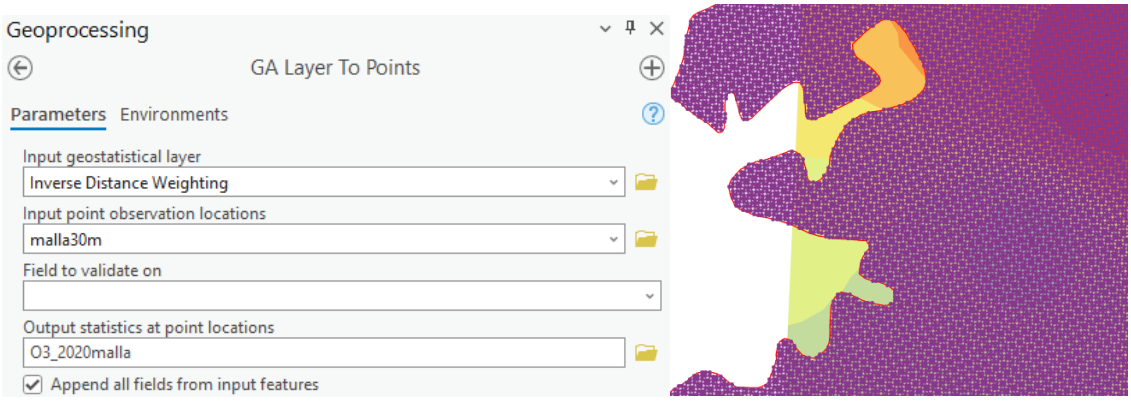


Figura 25. Herramienta de predicción/validación GA Layer to Point. ArcGIS pro. O3 RMCAB año 2021

Finalmente, se realizo nuevamente una interpolacion IDW con esta nueva capa de puntos, obteniendo asi un raster de cada contaminante cubriendo en su totalidad el area de estudio. Este proceso fue realizado para los contaminantes en cada año. A continuacion se presenta la extrapolacion del ozono para el año 2021.

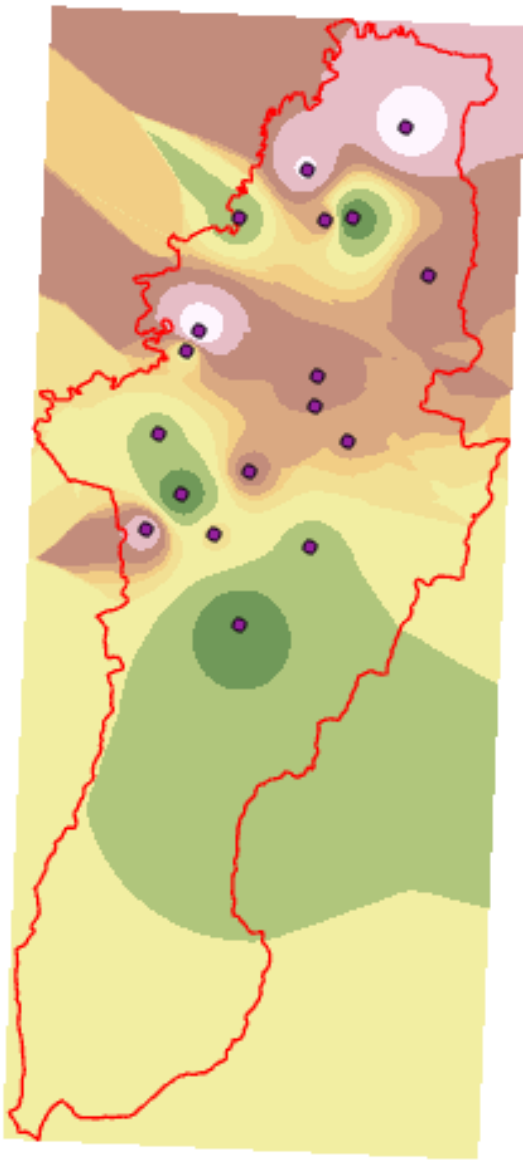


Figura 26. Ozono RMCAB extrapolado año 2021

Con la información de contaminantes, índices espectrales, localización geográfica y valores de reflectancia se obtuvieron los valores para cada uno de los puntos de la malla en cada uno de los años haciendo uso de la herramienta "*Extract Multi Values to Points*" de la siguiente manera:

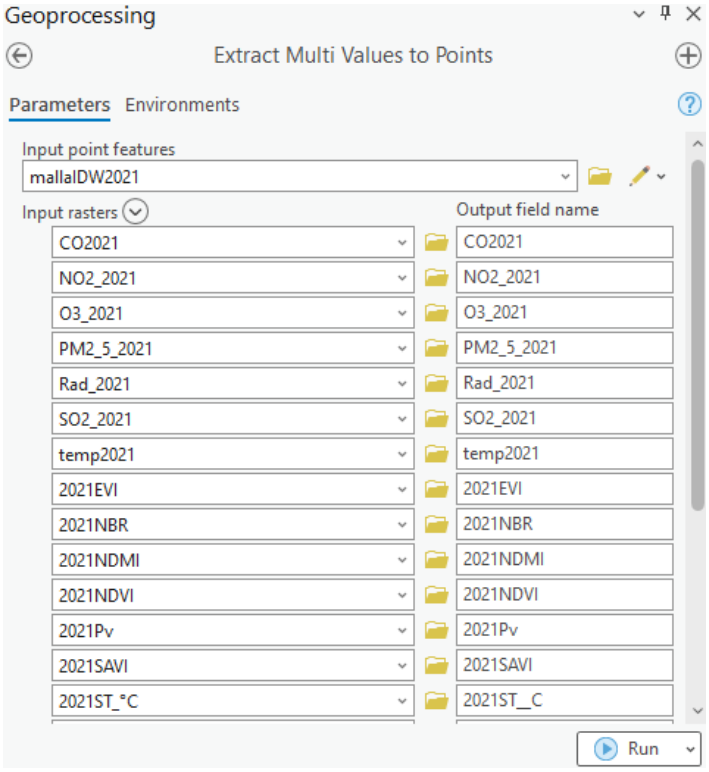


Figura 27. Herramienta Extract Multi Values to Points. ArcGIS pro

Estas 3 matrices multivariables, cada una con 948852 puntos, fueron las usadas para la regresión por mínimos cuadrados sin embargo a continuación se presenta una matriz multivariable parcial anual con las variables y los valores tenidos en cuenta en cada una de las estaciones (Tabla 13 a 18).

	CO	O3	NO2	SO2	T	PM2.5	Rad	LST	NDVI	Pv	EVI	NDMI	SAVI	NBR
Carvajal sevillana	1,173	20,919	29,476	3,391	17,3	40,0	133,6	35,58	0,1478	0,3294	0,1329	-0,0544	0,1065	-0,1274
CDAR	0,397	30,285	14,201	1,191	17,7	19,4	399,0	16,49	0,2638	0,3993	0,2195	0,0937	0,2204	0,2218
Guaymaral	0,556	27,318	9,574	4,564	16,4	20,3	537,0	28,22	0,4048	0,4934	0,2997	0,0581	0,2936	0,2311
Kennedy	0,734	37,818	16,510	2,555	17,4	33,1	132,8	37,47	0,4398	0,5183	0,3002	0,1022	0,2982	0,2498
Las Ferias	0,498	20,848	12,879	3,404	16,9	16,8	367,5	18,50	0,3151	0,4324	0,1077	0,0584	0,1263	0,0775
Min Ambiente	0,873	17,111	15,760	2,531	17,2	19,3	223,5	23,49	0,2163	0,3698	0,1300	0,1148	0,1174	0,2492
Móvil 7ma	1,075	25,171	15,001	1,713	17,7	30,0	94,0	28,42	0,8012	0,8111	0,7227	0,4189	0,6303	0,6640
Puente Aranda	0,961	14,235	18,694	1,990	15,8	31,8	227,8	38,07	0,1676	0,3408	0,1227	-0,0525	0,1079	-0,0623
San Cristóbal	0,829	23,125	17,553	2,804	15,8	23,2	493,0	27,66	0,4722	0,5418	0,3118	0,1586	0,3051	0,3253
Suba	0,437	25,556	12,234	1,948	17,2	18,4	410,0	37,76	0,2421	0,3857	0,1530	-0,0621	0,1487	-0,0363
Tunal	0,699	29,618	14,894	2,609	15,6	24,7	349,0	36,05	0,4524	0,5274	0,2947	0,1346	0,2758	0,2690
Usaquén	0,520	29,618	17,021	11,738	17,1	20,1	345,9	30,34	0,1897	0,3538	0,1407	-0,0018	0,1294	0,0874

Tabla 13. Matriz multivariable 2020 CO ~ NBR

	LAT	LON	reflB1	reflB2	reflB3	reflB4	reflB5	reflB6	reflB7
Carvajal sevillana	4,596	-74,149	0,1065	0,1549	0,1758	0,1971	0,2655	0,2960	0,3430
CDAR	4,658	-74,084	0,0766	0,1196	0,2192	0,2314	0,3972	0,3292	0,2530
Guaymaral	4,784	-74,044	0,0613	0,0779	0,1262	0,1393	0,3288	0,2927	0,2054
Kennedy	4,625	-74,161	0,0507	0,0640	0,1135	0,1155	0,2969	0,2418	0,1782
Las Ferias	4,736	-74,082	-0,0202	0,0215	0,0592	0,0624	0,1199	0,1067	0,1026
Min Ambiente	4,625	-74,067	0,0787	0,0879	0,1183	0,1110	0,1723	0,1368	0,1036
móvil 7ma	4,642	-74,084	0,0341	0,0396	0,0781	0,0548	0,4965	0,2033	0,1002
Puente Aranda	4,632	-74,117	0,0872	0,1167	0,1411	0,1566	0,2196	0,2439	0,2488
San Cristóbal	4,573	-74,084	0,0477	0,0593	0,0904	0,0999	0,2786	0,2023	0,1418
Suba	4,761	-74,093	0,0736	0,0842	0,1021	0,1313	0,2153	0,2438	0,2315
Tunal	4,576	-74,131	0,0572	0,0663	0,1053	0,0938	0,2487	0,1897	0,1433
Usaquén	4,710	-74,030	0,0855	0,1142	0,1515	0,1689	0,2479	0,2488	0,2081

Tabla 14. Matriz multivariable 2020 Latitud ~ Reflectancia B7

	CO	O3	NO2	SO2	T	PM2.5	Rad	LST	NDVI	Pv	EVI	NDMI	SAVI	NBR
Carvajal sevillana	2,784	10,660	26,596	6,935	17,4	35,0	350,6	32,8	0,1221	0,3148	0,0779	-0,0678	0,0734	-0,0043
CDAR	0,639	21,378	16,047	1,677	18,3	18,3	407,0	30,0	0,7886	0,7998	0,5659	0,3734	0,5300	0,6395
Guaymaral	0,446	25,774	4,465	5,289	17,3	25,2	446,0	30,4	0,3334	0,4445	0,1978	0,0463	0,2074	0,2159
Kennedy	1,342	16,065	26,766	4,659	16,5	38,8	375,3	33,7	0,4422	0,5200	0,2751	0,0790	0,2698	0,2252
Las Ferias	1,080	20,681	18,691	5,890	16,8	20,6	254,1	28,0	0,2442	0,3870	0,1304	-0,0129	0,1304	0,0494
Min Ambiente	1,029	19,531	24,422	3,641	17,9	21,3	237,9	26,5	0,1627	0,3380	0,1241	0,0690	0,1081	0,1775
Móvil 7ma	1,846	20,320	5,319	2,701	18,7	24,8	45,0	30,5	0,7333	0,7511	0,5488	0,2655	0,5077	0,5322
San Cristóbal	0,746	16,176	20,326	4,111	16,3	26,0	517,0	30,1	0,6533	0,6834	0,4486	0,2682	0,4397	0,4440
Suba	0,524	24,504	12,661	6,666	17,1	23,0	310,0	33,7	0,1052	0,3054	0,0592	-0,0694	0,0753	-0,0379
Tunal	2,003	18,607	31,966	4,144	17,2	40,9	328,0	21,8	0,2268	0,3763	0,1414	-0,0448	0,1444	0,0596
Usaquén	0,854	21,583	18,668	1,908	16,9	19,7	311,8	28,4	0,1688	0,3415	0,0959	-0,0782	0,1017	0,0985
Bolivia	0,338	14,898	19,369	6,984	17,0	21,4	341,3	30,5	0,5076	0,5682	0,3598	0,2195	0,3365	0,3096
Ciudad Bolívar	1,156	23,206	30,645	4,646	14,9	35,4	351,0	30,5	0,6081	0,6465	0,4400	0,1515	0,4062	0,3883
Colina	0,501	10,538	16,339	5,345	16,3	12,0	235,0	28,6	0,7728	0,7857	0,5744	0,3577	0,5331	0,5781
Fontibón	0,495	27,831	19,149	8,578	17,0	25,4	488,7	38,0	0,0770	0,2900	0,0460	-0,1408	0,0453	-0,1474
Jazmín	1,127	21,013	30,098	2,164	16,7	27,2	339,0	33,8	0,3226	0,4373	0,2624	0,1600	0,2405	0,3456
Movil Fontibón	0,495	17,059	22,486	8,578	17,0	27,1	515,0	35,6	0,2743	0,4060	0,1702	-0,0698	0,1611	0,0133
Usme	0,493	11,456	16,170	3,361	17,8	23,8	446,0	34,5	0,4749	0,5438	0,2967	0,0755	0,3027	0,2606

Tabla 15. Matriz multivariable 2021 CO ~ NBR

	LAT	LON	reflB1	reflB2	reflB3	reflB4	reflB5	reflB6	reflB7
Carvajal sevillana	4,596	-74,149	0,0843	0,1011	0,1345	0,1469	0,1878	0,2151	0,1894
CDAR	4,658	-74,084	0,0215	0,0275	0,0610	0,0429	0,3630	0,1656	0,0798
Guaymaral	4,784	-74,044	0,0520	0,0602	0,0929	0,1181	0,2362	0,2153	0,1523
Kennedy	4,625	-74,161	0,0525	0,0591	0,0910	0,0956	0,2473	0,2111	0,1564
Las Ferias	4,736	-74,082	0,0463	0,0673	0,1086	0,1044	0,1719	0,1764	0,1557
Min Ambiente	4,625	-74,067	0,0936	0,1235	0,1485	0,1666	0,2313	0,2015	0,1616
Movil 7ma	4,642	-74,084	0,0305	0,0377	0,0763	0,0572	0,3714	0,2156	0,1134
San Cristóbal	4,573	-74,084	0,0274	0,0371	0,0778	0,0705	0,3364	0,1941	0,1295
Suba	4,761	-74,093	0,0536	0,0601	0,0745	0,2045	0,2526	0,2902	0,2725
Tunal	4,576	-74,131	0,0775	0,0804	0,1519	0,1426	0,2263	0,2475	0,2008
Usaquén	4,710	-74,030	0,0712	0,0740	0,1227	0,1396	0,1963	0,2297	0,1611
Bolivia	4,736	-74,126	0,0513	0,0649	0,0961	0,0975	0,2984	0,1910	0,1573
Ciudad Bolívar	4,577	-74,166	0,0433	0,0544	0,0941	0,0787	0,3227	0,2378	0,1422
Colina	4,737	-74,069	0,0229	0,0314	0,0571	0,0484	0,3773	0,1785	0,1009
Fontibón	4,678	-74,144	0,0776	0,0957	0,1206	0,1489	0,1738	0,2307	0,2338
Jazmín	4,609	-74,115	0,0969	0,1082	0,1529	0,1673	0,3267	0,2366	0,1589
Movil Fontibón	4,668	-74,149	0,0648	0,0812	0,1043	0,1167	0,2049	0,2357	0,1995
Usme	4,532	-74,117	0,0344	0,0501	0,1030	0,0970	0,2725	0,2343	0,1598

Tabla 16. Matriz multivariable 2021 Latitud ~ Reflectancia B7

	CO	O3	NO2	SO2	T	PM2.5	Rad	LST	NDVI	Pv	EVI	NDMI	SAVI	NBR
Carvajal sevillana	0,771	14,026	17,893	4,280	18,0	23,5	538,8	31,6673	0,0888	0,0983	0,0613	-0,0305	0,0563	-0,0171
CDAR	0,624	21,788	25,111	1,857	17,1	21,9	490,5	22,5959	0,6820	0,5782	0,4249	0,3229	0,4137	0,5643
Guaymaral	0,303	29,804	3,208	1,874	17,1	17,9	611,0	28,4885	0,4117	0,3100	0,3202	0,1349	0,2860	0,3203
Kennedy	0,495	20,231	15,621	1,352	16,2	23,4	533,9	31,6092	0,4444	0,3380	0,2673	0,0750	0,2705	0,2043
Las Ferias	1,621	21,029	24,762	1,816	15,9	34,5	560,2	24,4929	0,1817	0,1471	0,1135	-0,0334	0,1063	-0,0073
Min Ambiente	0,823	25,509	21,509	1,962	17,0	22,4	503,3	18,5352	0,1373	0,1225	0,0904	0,0379	0,0837	0,1934
Móvil 7ma	1,320	22,500	6,975	1,829	18,0	26,3	475,0	22,5548	0,6960	0,5944	0,5459	0,3255	0,4929	0,5703
San Cristóbal	0,492	24,283	16,531	2,093	15,4	21,3	577,0	22,1720	0,5459	0,4328	0,3324	0,2055	0,3246	0,3642
Suba	0,485	22,233	9,746	1,876	16,7	18,9	561,3	31,8587	0,0831	0,0956	0,0469	-0,1246	0,0602	-0,1275
Tunal	0,788	24,398	14,672	1,571	17,1	17,3	553,0	30,4949	0,2101	0,1639	0,1329	-0,0672	0,1298	0,0404
Usaquén	0,502	30,414	18,118	1,916	16,8	20,4	543,6	27,1042	0,2384	0,1816	0,1471	-0,0185	0,1396	0,0734
Bolivia	0,646	15,858	17,492	1,435	16,4	20,8	543,4	31,7596	0,4499	0,3429	0,2930	0,1142	0,2835	0,2602
Ciudad Bolívar	1,038	25,003	19,724	2,534	14,3	30,3	557,0	29,2371	0,3272	0,2432	0,2370	-0,0215	0,2432	0,1336
Colina	0,504	9,851	17,886	1,842	15,7	15,7	563,0	25,2072	0,5213	0,4088	0,3758	0,0831	0,3601	0,3388
Fontibón	0,076	20,924	24,491	2,295	16,2	31,0	551,2	34,3060	0,1001	0,1037	0,0509	-0,1343	0,0512	-0,1165
Jazmín	0,702	19,724	18,954	2,290	15,9	21,2	492,0	27,8904	0,3561	0,2651	0,2954	0,1935	0,2670	0,3693
Móvil Fontibón	0,254	22,910	15,259	2,190	16,2	34,4	557,0	33,6395	0,2463	0,1868	0,1290	-0,0588	0,1290	-0,0166
Usme	0,608	25,409	20,215	1,932	15,6	21,6	492,0	25,4807	0,2686	0,2016	0,1730	0,0126	0,1741	0,1366
Puente Aranda	0,785	21,695	24,932	1,055	15,6	20,4	536,0	30,1907	0,0691	0,0892	0,0410	-0,1195	0,0389	-0,1162

Tabla 17. Matriz multivariable 2022 CO ~ NBR

	LAT	LON	reflB1	reflB2	reflB3	reflB4	reflB5	reflB6	reflB7
Carvajal sevillana	4,5956	-74,1486	0,0916	0,1166	0,1400	0,1666	0,1991	0,2116	0,2060
CDAR	4,6585	-74,0840	0,0262	0,0330	0,0601	0,0540	0,2855	0,1461	0,0795
Guaymaral	4,7838	-74,0442	0,0735	0,0906	0,1249	0,1269	0,3045	0,2321	0,1567
Kennedy	4,6251	-74,1613	0,0349	0,0529	0,0947	0,0949	0,2466	0,2122	0,1629
Las Ferias	4,7359	-74,0825	0,0744	0,0925	0,1279	0,1309	0,1891	0,2021	0,1918
Min Ambiente	4,6255	-74,0670	0,0755	0,1040	0,1208	0,1476	0,1946	0,1804	0,1315
Móvil 7ma	4,6424	-74,0840	0,0400	0,0482	0,0849	0,0680	0,3792	0,1929	0,1038
San Cristóbal	4,5726	-74,0838	0,0321	0,0471	0,0769	0,0746	0,2538	0,1673	0,1183
Suba	4,7612	-74,0935	0,0577	0,0625	0,0818	0,2138	0,2526	0,3245	0,3264
Tunal	4,5762	-74,1310	0,0685	0,0877	0,1160	0,1383	0,2118	0,2423	0,1953
Usaquén	4,7104	-74,0304	0,0593	0,0843	0,1147	0,1220	0,1983	0,2058	0,1712
Bolivia	4,7359	-74,1259	0,0503	0,0626	0,0945	0,0996	0,2626	0,2087	0,1541
Ciudad Bolívar	4,5770	-74,1660	0,0544	0,0829	0,1323	0,1653	0,3260	0,3403	0,2491
Colina	4,7373	-74,0694	0,0511	0,0610	0,0952	0,1021	0,3246	0,2747	0,1603
Fontibón	4,6782	-74,1438	0,0644	0,0760	0,1046	0,1163	0,1422	0,1863	0,1797
Jazmín	4,6085	-74,1149	0,1001	0,1064	0,1618	0,1609	0,3388	0,2289	0,1560
Móvil Fontibón	4,6677	-74,1489	0,0587	0,0658	0,0978	0,1011	0,1671	0,1880	0,1728
Usme	4,5321	-74,1171	0,0671	0,0800	0,1214	0,1392	0,2413	0,2353	0,1834
Puente Aranda	4,6318	-74,1175	0,0776	0,0981	0,1172	0,1400	0,1607	0,2043	0,2030

Tabla 18. Matriz multivariable 2022 Latitud ~ Reflectancia B7

4.4. Modelo de regresión.

Una vez toda la información requerida estaba completa se dio inicio al desarrollo del código en R para la regresión multivariable. Se usó el software RStudio y a partir de códigos previamente descritos, así como el uso de la herramienta de ayuda de R se definieron las líneas necesarias para la predicción del ozono en los 3 años considerados. El paquete PLS de R permite ejecutar regresión por componentes principales (PCA) y por mínimos cuadrados parciales (PLS) por esta razón se ejecutaron ambas metodologías pudiendo comparar el error cuadrático medio de predicción entre ellas. En términos generales, el código consistió en la descarga y puesta a disposición de los paquetes necesarios para la ejecución de la regresión seguido del importe de las bases de datos, 3 en este caso, para después año por año, definir conjuntos de entrenamiento y prueba del modelo, en ambas metodologías fueron asignados un 75% y 25% de los datos respectivamente, de manera aleatoria. Luego se estableció el modelo y determinación de componentes a usar y finalmente con el conjunto de datos de prueba se realizan las predicciones cuyo resultado siendo comparado con los valores medidos representan el RMSEP. (Alonso, 2017)

El dataframe de estudio contó con 948852 filas de observaciones cada una con 23 variables, es decir 21823596 valores a considerar. A continuación, se presentarán algunas líneas de código necesarias para su ejecución con base al año 2021. El código completo se presenta en los anexos.

4.4.1. Modelo PCR.

En primer lugar, se realizó la prueba del modelo asignando 10 componentes principales como máximo, validando el modelo por medio de validación cruzada (CV) y calculando una estandarización de variables al dividir cada variable entre su desviación estándar haciendo la media 0:

```
>2021<-pcr (O3~.,ncomp=10,data=dataTrain2021, scale=TRUE,
validation="CV")
```

Con el resumen del modelo (summary ()) se puede observar el error cuadrático medio de predicción y como conforme el número de componentes aumenta este reduce, no obstante, no lo hace con la misma proporción que con los primeros componentes. Esto se puede igualmente visualizar al graficar el RMSEP y la librería PLS también aporta un método de visualización para sugerir el numero óptimo de componentes primarios que deberían ser usados.

```
>plot(RMSEP(2021),main="Gráfico de validación PCR 2021",xlab="número de componentes",legendpos="topright")
```

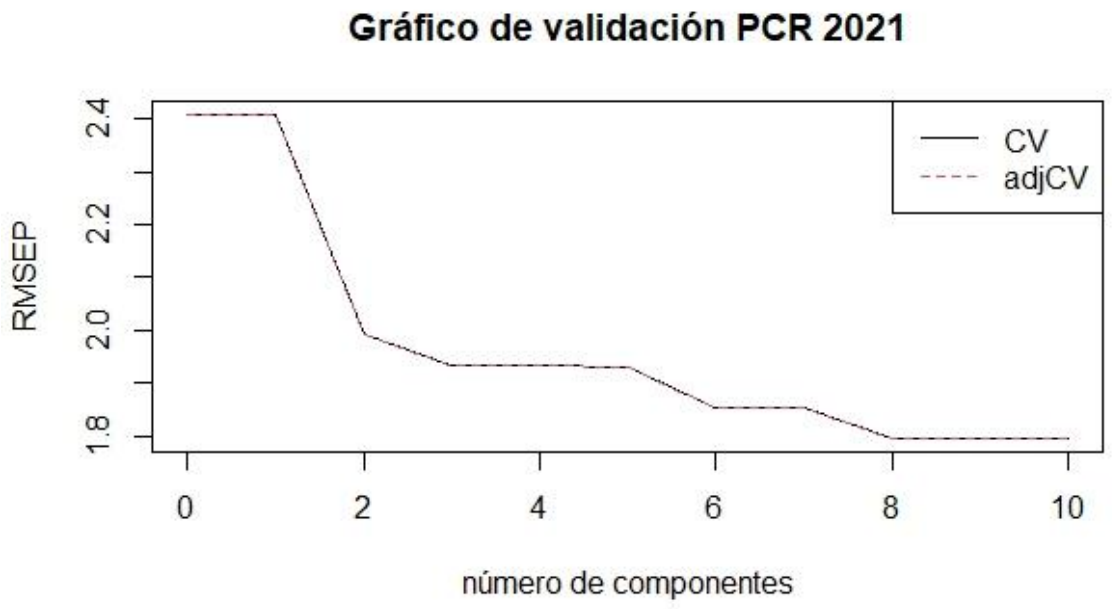


Figura 28. RMSEP PCR año 2021

```
>ncomp.onesigma<-selectNcomp(modelo2021,method="onesigma", plot=TRUE,ylim=c(1.7, 2.5))
```

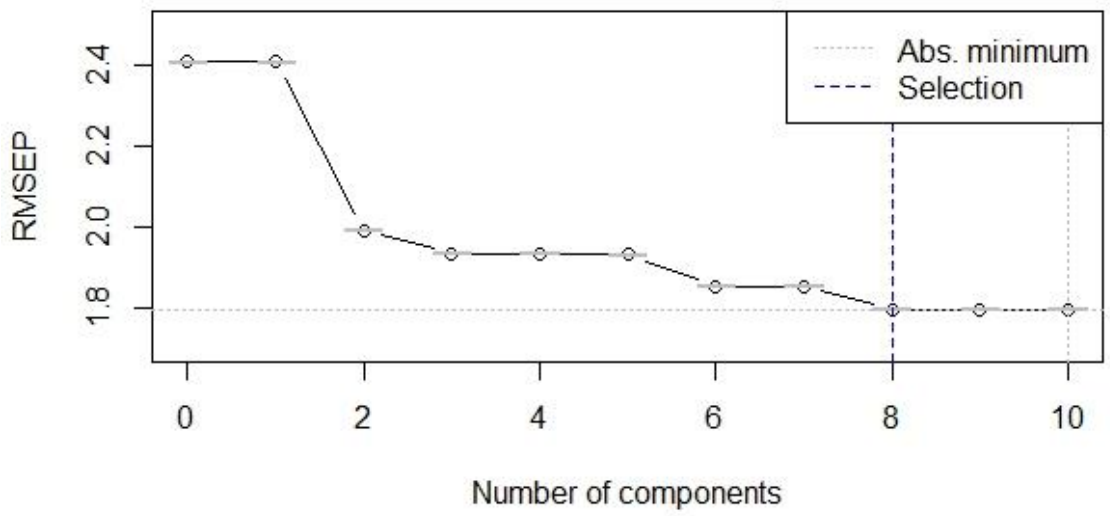


Figura 29. Componentes óptimos PCR año 2021

Como se puede observar en la imagen anterior, el numero óptimo de componentes principales escogido para la regresión PCR año 2021 fueron 8, así se puede entonces ajustar el modelo de regresión con los componentes que se usaran. Posteriormente se calcularon las varianzas acumuladas que describen el porcentaje de variabilidad de los datos que cada componente va explicando, en este caso, por ejemplo, un 89% de variabilidad de los datos es explicable con el uso de 8 componentes principales.

```
>modelo2021<-pcr(O3~.,ncomp=8,data=dataTrain2021, scale=TRUE,
validation="CV")

>explvar(modelo2021)

  Comp 1    Comp 2    Comp 3    Comp 4    Comp 5    Comp 6    Comp 7    Comp 8
21.559133 18.642828 12.874389 11.314354  9.533408  7.356566  4.545999  3.430577

>cumsum(explvar(modelo2021))

Comp 1   Comp 2   Comp 3   Comp 4   Comp 5   Comp 6   Comp 7   Comp 8
21.55913 40.20196 53.07635 64.39070 73.92411 81.28068 85.82668 89.25725
```

Finalmente, con este modelo y el conjunto de prueba se pudieron calcular los valores de predicción de ozono, posteriormente fue usado para calcular el valor predicho en todos los puntos del área de estudio considerados y permite igualmente observar la relación entre los valores medidos y predichos en el conjunto de prueba usado inicialmente, así como también el valor de RMSEP obtenido.

```
>pred2021<-predict(modelo2021,ncomp=8,newdata=dataTest2021)
>plot(modelo2021,ncomp=8,asp=1,line=TRUE)
```

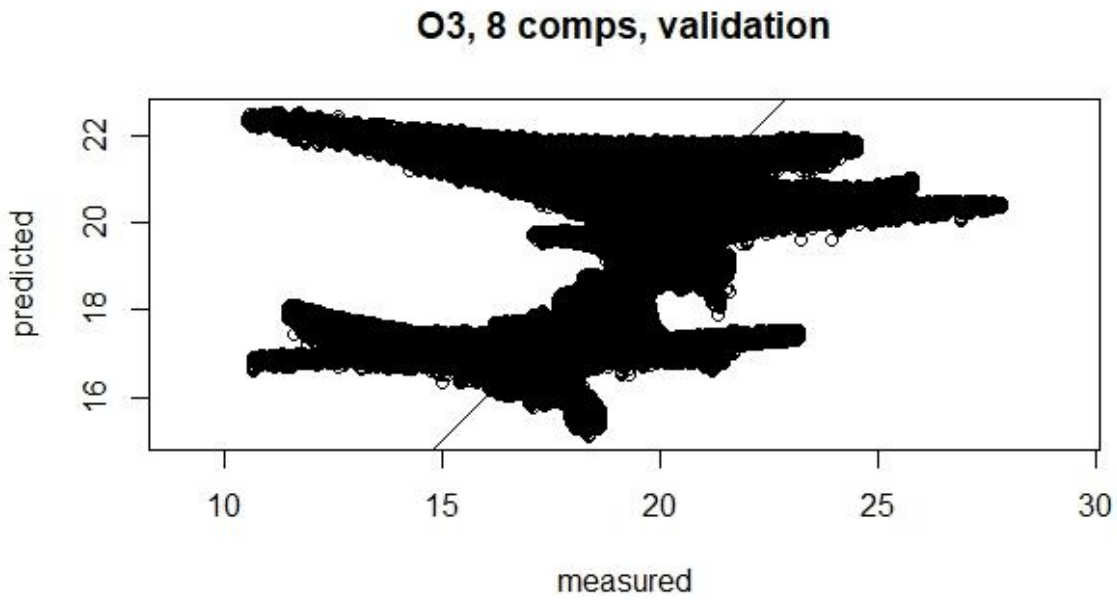


Figura 30. Predicciones vs mediciones PCR año 2021

```
>RMSEP(modelo2021,newdata=dataTest2021)
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps
      2.414        2.414        1.994        1.936        1.936        1.933
    6 comps      7 comps      8 comps
      1.854        1.854        1.799
```

4.1.2. Modelo PSL

Respecto a la regresión por mínimos cuadrados parciales se usó el mismo conjunto de datos de entrenamiento y de prueba que con PCR sin embargo es necesario ajustar el modelo nuevamente, visualizar el error de predicción y determinar el número óptimo de componentes principales a usar, esto se llevó a cabo con las siguientes líneas de código.

```
>modelopls<-
plsr(O3~.,ncomp=10,data=dataTrain2021,scale=TRUE,validation="CV")
```

```
>plot(RMSEP(modelopls),main="Gráfico de validación PLS 2021",xlab="número
de componentes",legendpos="bottomright")
```

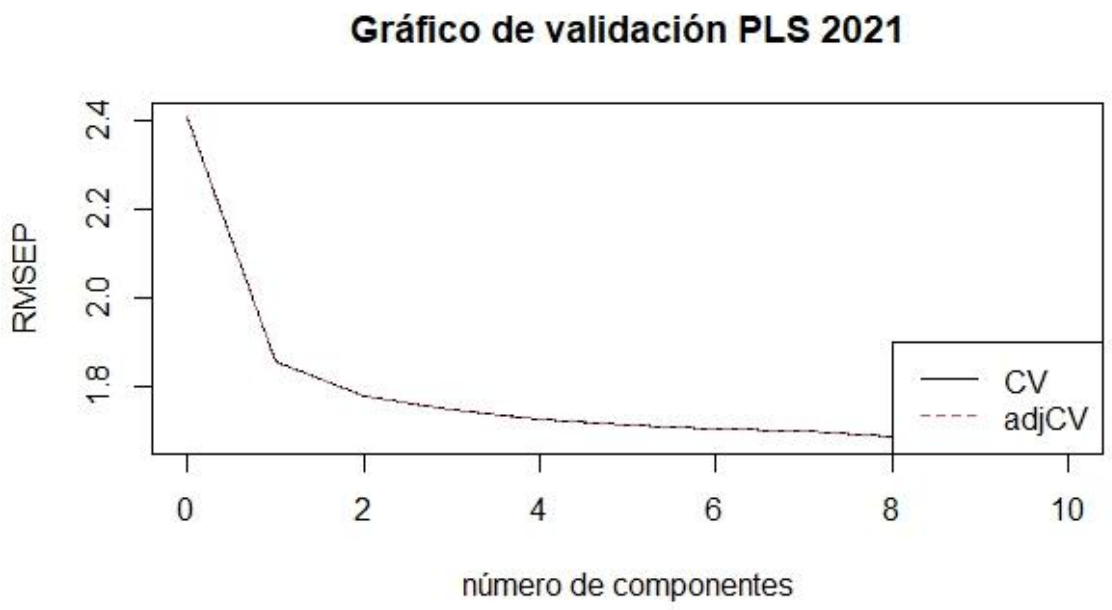


Figura 31. RMSEP PLS año 2021

```
>ncomp.onesigma<-
selectNcomp(modelopls,method="onesigma",plot=TRUE,ylim=c(1.5,2.4))
```

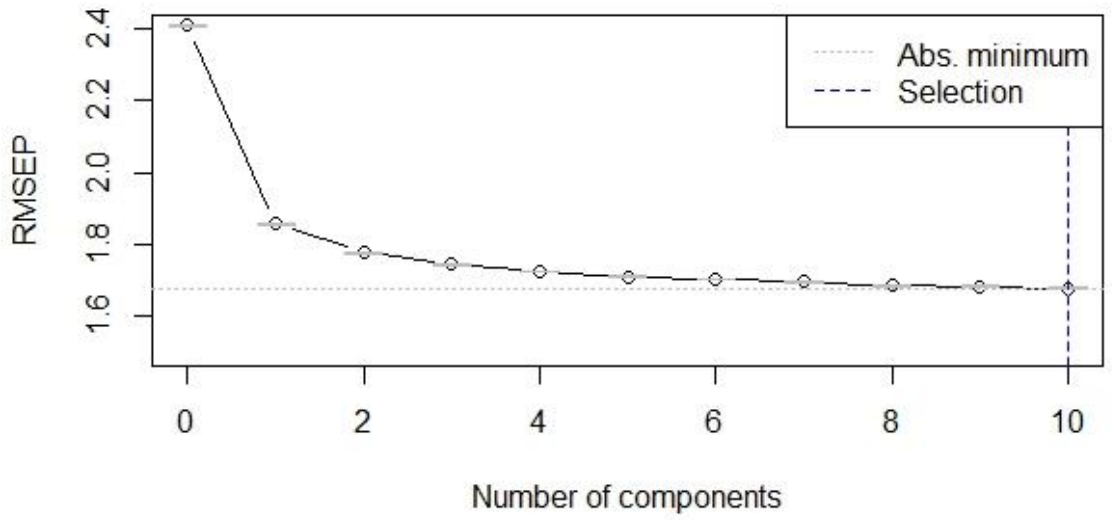


Figura 32. Componentes óptimos PLS año 2021

De esta manera el modelo PLS se definió para usar 10 componentes principales con la información del año 2021 en la cual se observó la siguiente distribución entre los valores medidos y predichos, así como también los valores de varianza acumulada.

```
> explvar(modelopls)
   Comp 1    Comp 2    Comp 3    Comp 4    Comp 5    Comp 6    Comp 7    Comp 8
17.988529  8.889917  8.504750  6.010738  7.148024 16.032222  8.292576  2.959208
   Comp 9   Comp 10
 8.926136  3.172762
> cumsum(explvar(modelopls))
  Comp 1   Comp 2   Comp 3   Comp 4   Comp 5   Comp 6   Comp 7   Comp 8   Comp 9
17.98853 26.87845 35.38320 41.39393 48.54196 64.57418 72.86676 75.82596 84.75210
 Comp 10
87.92486
```

El modelo de predicción quedó entonces definido como:

```
>predichopls<-predict(modelopls,ncomp=10,newdata=dataTest2021)

>plot(modelopls,ncomp=10,asp=1,line=TRUE)

> RMSEP(modelopls,newdata=dataTest2021)

(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps

      2.414        1.857        1.777        1.745        1.723        1.708

    6 comps      7 comps      8 comps      9 comps     10 comps

      1.699        1.692        1.682        1.680        1.674
```

Como se puede observar el RMSEP obtenido haciendo uso de la metodología PSL es ligeramente menor al de PCR. Finalmente se realizó el cálculo de ozono troposférico predicho para el conjunto total de los datos el cual fue exportado para la visualización en ArcGIS.

4.5. Generación de mapa de ozono troposférico predicho

Las 3 matrices multivariables con puntos cada 30 m sobre el área de estudio fueron usadas para la regresión PCR y PLS, dado que la regresión PLS presenta resultados con un menor error de predicción respecto a la metodología PCR esta fue usada para predecir los valores de ozono troposférico para la totalidad de observaciones en cada año. La información fue exportada de R a Excel y de allí importada a ArcMap con la opción “Excel to Table” dado que la información csv no puede ser directamente añadida al SIG debido a que excede el valor de filas que pueden ser guardadas en este formato, siendo el límite cercano a 80000 filas únicamente.

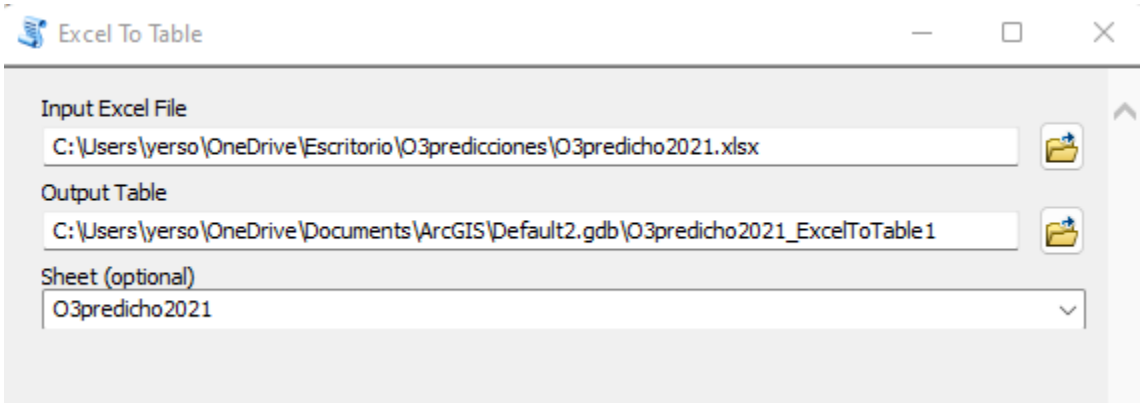


Figura 33. Herramienta Excel to Table. Desktop ArcMap

Con la tabla en el software se requirió georreferenciar cada punto con los valores X y Y con que contaba la tabla desde Excel y una vez ubicados los puntos se procedió a realizar finalmente la interpolación IDW del ozono predicho obtenido en R. El tamaño de celda puede ser modificado para mejorar la resolución del ráster resultante.

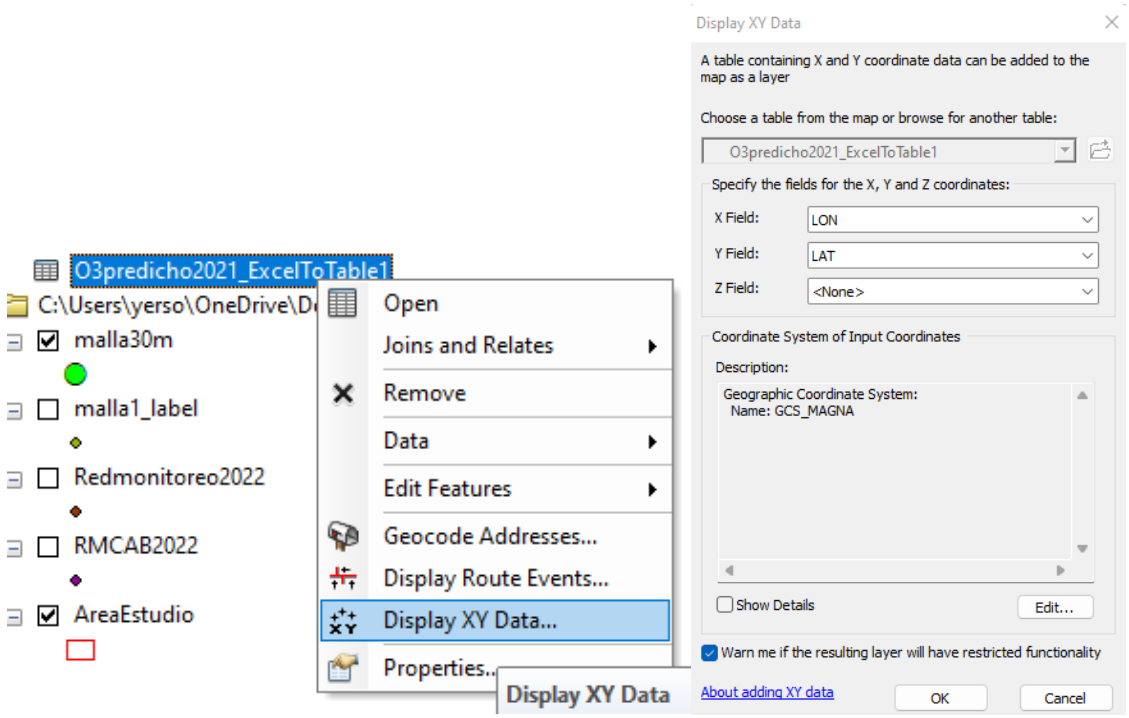


Figura 34. Georreferenciación de valores predichos. Desktop ArcMap

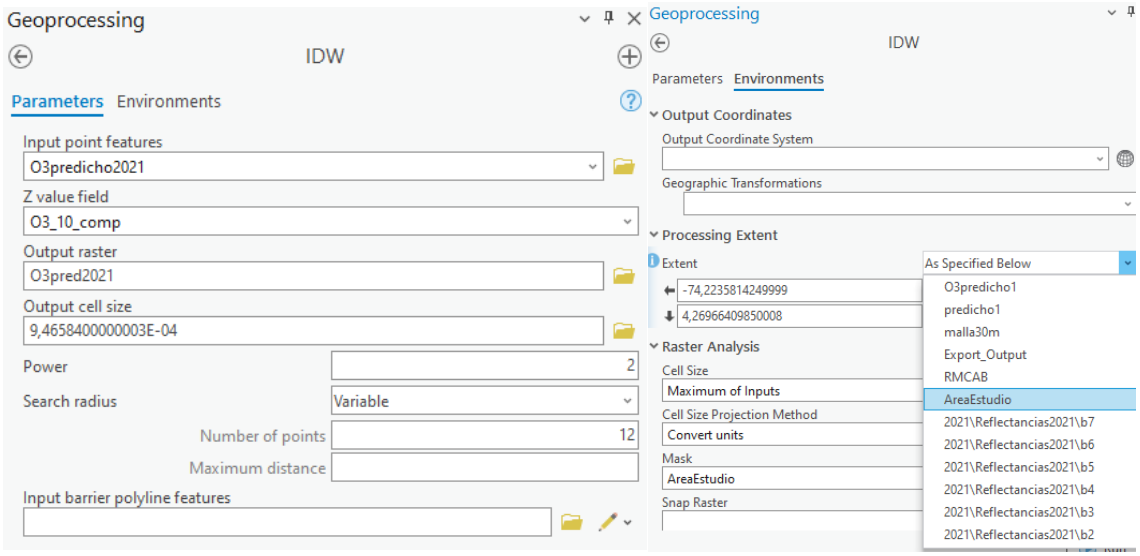


Figura 35. Herramienta IDW para valores predichos. ArcGIS pro

De esta manera se obtuvieron los raster de ozono troposferico medido y predicho en cada año de estudio haciendo uso de la herramienta IDW, los mapas son presentados en los resultados.

5. Resultados y discusión.

En términos generales, la regresión PLS para los años 2020, 2021 y 2022 se ejecutaron haciendo uso de 9, 10 y 7 componentes principales, respectivamente; valores asignados dependiendo del valor sugerido obtenido por medio de la gráfica de validación del paquete R, pero cuya variación depende de los rangos de los valores mismos considerados.

La regresión por mínimos cuadrados parciales arrojó resultados ligeramente menores respecto a la regresión por componentes principales en cada año de estudio. Dentro de los resultados obtenidos con esta modelación se cuenta con varios términos importantes, como lo son la varianza, loadings o cargas de X, la correlación y los coeficientes de regresión. Estos elementos de la regresión lineal llevada a cabo permitieron observar el grado de error y comportamiento del modelo desarrollado.

En primer lugar, las puntuaciones o scores de X observadas para cada uno de los años permitieron la identificación de valores, agrupaciones o comportamientos atípicos en el conjunto de datos

usados. Debido a la cantidad de observaciones consideradas esta gráfica demostró agrupación en todos los componentes, no obstante, es debido a la escala de representación. Los porcentajes relacionados con cada componente, sin embargo, son los valores de varianza en los cuales se puede observar que tanta variabilidad de la totalidad de las variables independientes (X) fue explicada por cada componente principal. Es decir, se explicó un 86% de las variables independientes con 9 componentes para el año 2020, un 87,9% con 10 componentes para el año 2021 y un 71,5% con 7 componentes para el año 2022; alcanzando respectivamente una varianza de la variable respuesta ozono troposférico (O_3) de 65.37%, 51.49% y 45.61%.(Tabla 19, 20 y 21)

	Comp 1	**Comp 2**	**Comp 3**	**Comp 4**	**Comp 5**	**Comp 6**	**Comp 7**	**Comp 8**	**Comp 9**
Varianza X	18,404	16,774	24,925	5,106	12,209	2,533	2,298	2,936	1,744
Varianza acumulada X	18,404	35,178	60,102	65,208	77,417	79,951	82,249	85,184	86,929
O3	21,650	41,900	45,020	52,200	54,760	59,860	63,520	64,870	65,370

Tabla 19. Varianzas modelo PLS año 2020

	Comp 1	**Comp 2**	**Comp 3**	**Comp 4**	**Comp 5**	**Comp 6**	**Comp 7**	**Comp 8**	**Comp 9**	**Comp 10**
Var X	17,989	8,890	8,505	6,011	7,148	16,032	8,293	2,959	8,926	3,173
Var acum X	17,989	26,878	35,383	41,394	48,542	64,574	72,867	75,826	84,752	87,925
O3	40,590	45,580	47,520	48,740	49,660	50,030	50,380	51,000	51,130	51,490

Tabla 20. Varianzas modelo PLS año 2021

	Comp 1	**Comp 2**	**Comp 3**	**Comp 4**	**Comp 5**	**Comp 6**	**Comp 7**
Varianza X	31,956	9,149	10,398	5,732	7,883	2,089	4,285
Varianza Acumulada X	31,956	41,105	51,503	57,234	65,117	67,206	71,491
O3	16,19	29,6	34,96	40,72	43,64	45,54	45,61

Tabla 21. Varianza de variables independientes modelo PLS año 2022

Los valores del coeficiente de determinación (R^2) obtenidos para cada modelo fueron los siguientes:

	2020	**2021**	**2022**
Componentes	9	10	7
R^2	0,653	0,515	0,459

Figura 36. Coeficientes de determinación.

Las cargas de X o de las variables independientes “Loadings” son útiles para buscar picos y está relacionado con la manera en la que las variables “cargan” contribuyeron a la creación de cada componente. Es decir, una gráfica de cargas es útil para determinar variables sin importancia que pueden ser removidas de un estudio porque representarían pesos diminutos en todos los componentes. Debido a esto dichas variables pueden ser removidas dado que no tienen ninguna correlación con ninguno de los componentes o con alguna variable (Dunn, 2022).

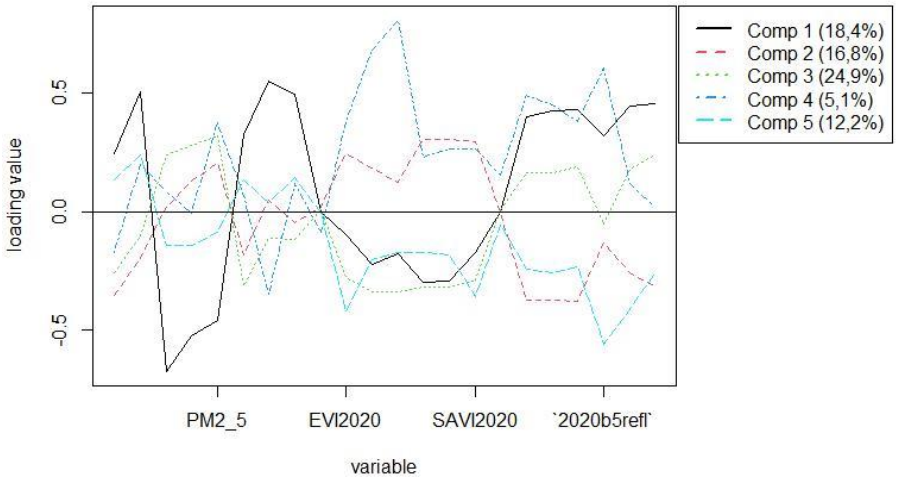


Figura 37. Cargas modelo 2020 componentes 1 a 5

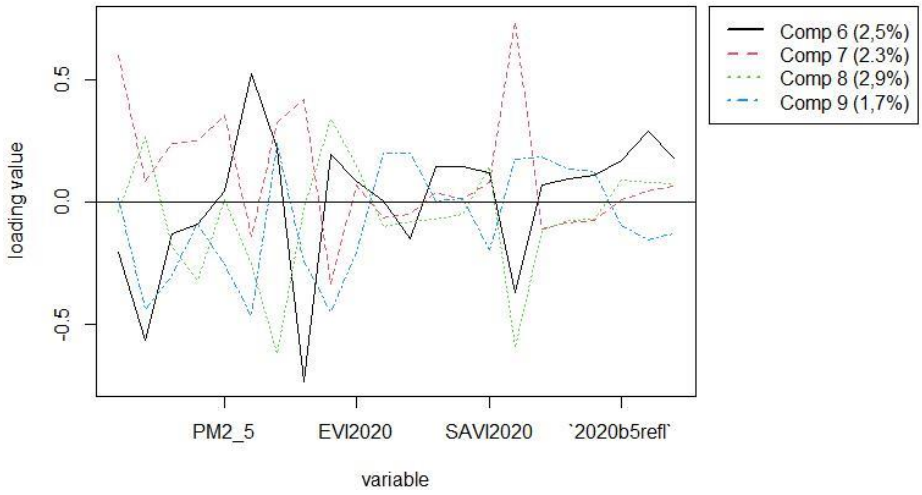


Figura 38. Cargas modelo 2020 componentes 6 a 9

Por ejemplo, las cargas del modelo para el año 2020 no muestran variables cuyo aporte a la creación de ninguno de los 9 componentes usados haya sido mínima o nula, sin embargo, componentes como el número 3, tuvieron aportes bajos en el conjunto de variables atmosféricas medidas y en algunos índices espectrales considerados, no obstante, no fue removida debido a que en otras variables su aporte fue significativo; en los otros 2 años dado el comportamiento de las cargas tampoco fue removida ninguna variable. En los anexos se presentan los resultados de estos elementos para los otros dos años

La plataforma R también permite visualizar la correlación entre las variables y los componentes por medio de un gráfico de dispersión en el cual cada punto representa una variable independiente y la distancia al cuadrado entre su ubicación y el centro del grafico representan la varianza de éstas. Por ejemplo, se representan los primeros 5 componentes del modelo para el año 2020 y la distribución de cada una de las 23 variables respecto a cada uno de los componentes considerados. Como se puede observar, en los componentes 1 a 5, la mayoría de las variables toman valores alejados del centro mientras que los componentes 6 a 9 concentran las variables en mayor cantidad cercanas a 0, esto significa que los primeros componentes representan una mayor correlación respecto a las

variables y los últimos componentes una menor correlación por lo cual conforme una mayor cantidad de componentes se use menor varianza en la información se da (Bjorn-Helge & Wehrens, 2022) (Figura 39).

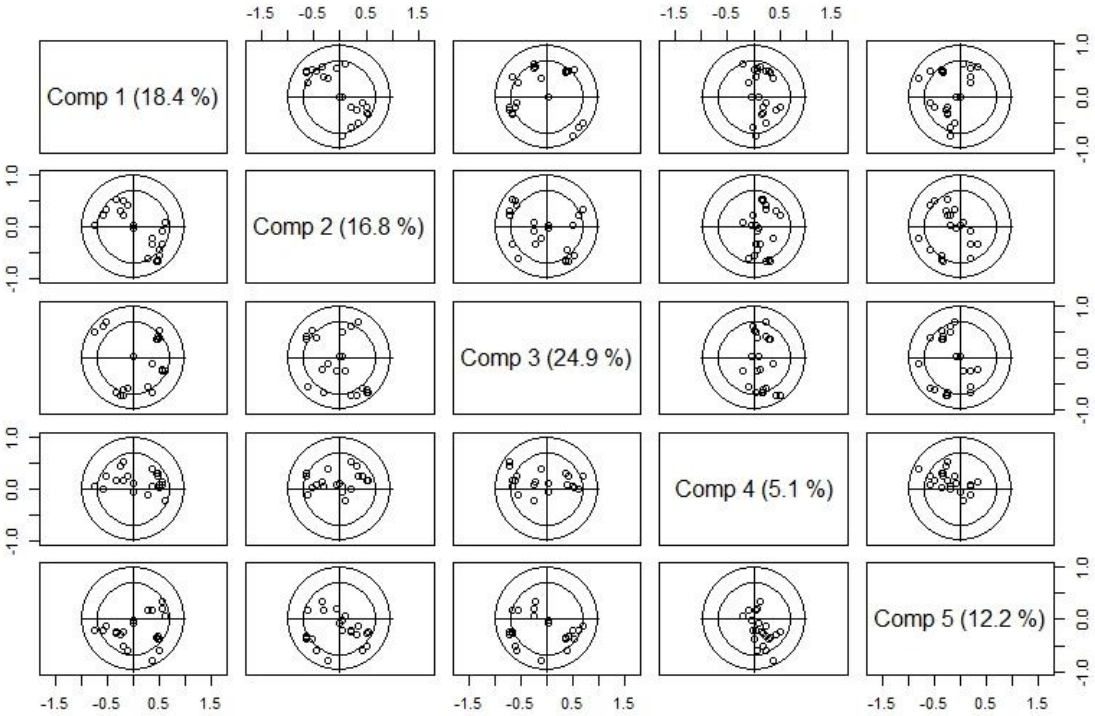


Figura 39. Correlación de variables modelo 2022. Componentes 1 a 5

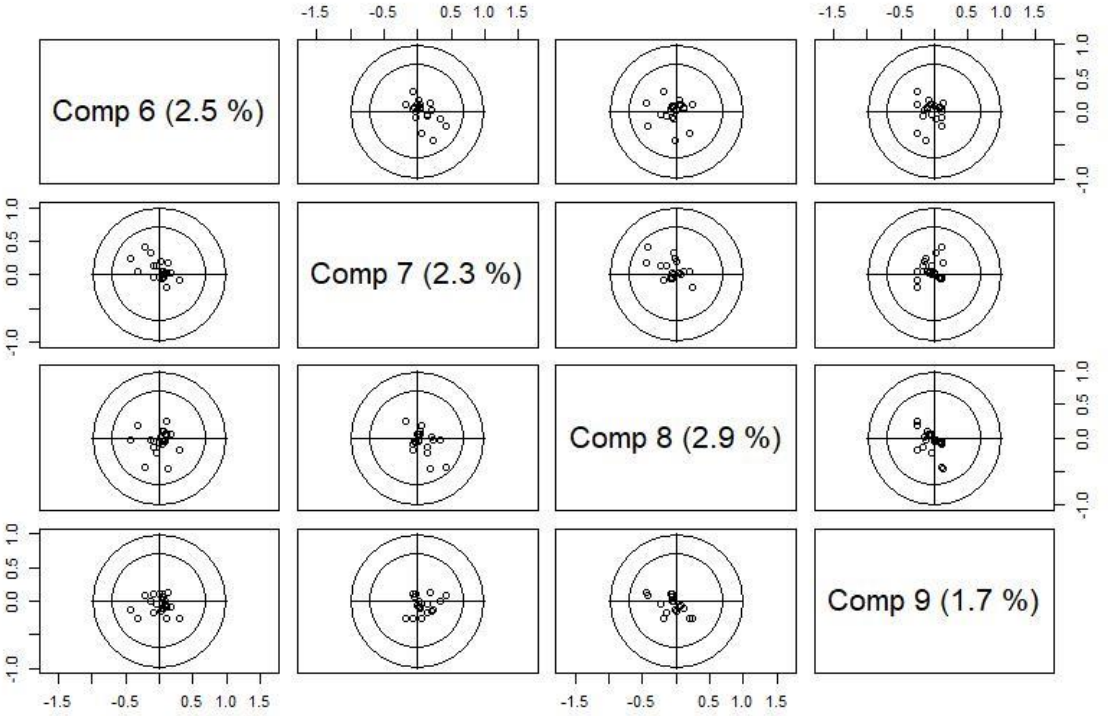


Figura 40. Correlación de variables modelo 2022. Componentes 6 a 9

Finalmente, los coeficientes de regresión fueron obtenidos gráficamente para cada modelo y la información obtenida se encuentra asociada con la contribución de los componentes a la regresión, así como también a la importancia de las variables independientes. Como se puede apreciar en la figura a continuación, desde el quinto componente en adelante el comportamiento es similar indicando que las predicciones con estos componentes lo serían igualmente y también se puede observar que las variables con mayor importancia se encuentran en el grupo de contaminantes y variables meteorológicas medidas por las RMCAB, además del aporte de índices espectrales relacionados con la presencia de vegetación (NDMI, NDBI, NBR) (Alonso, 2017).

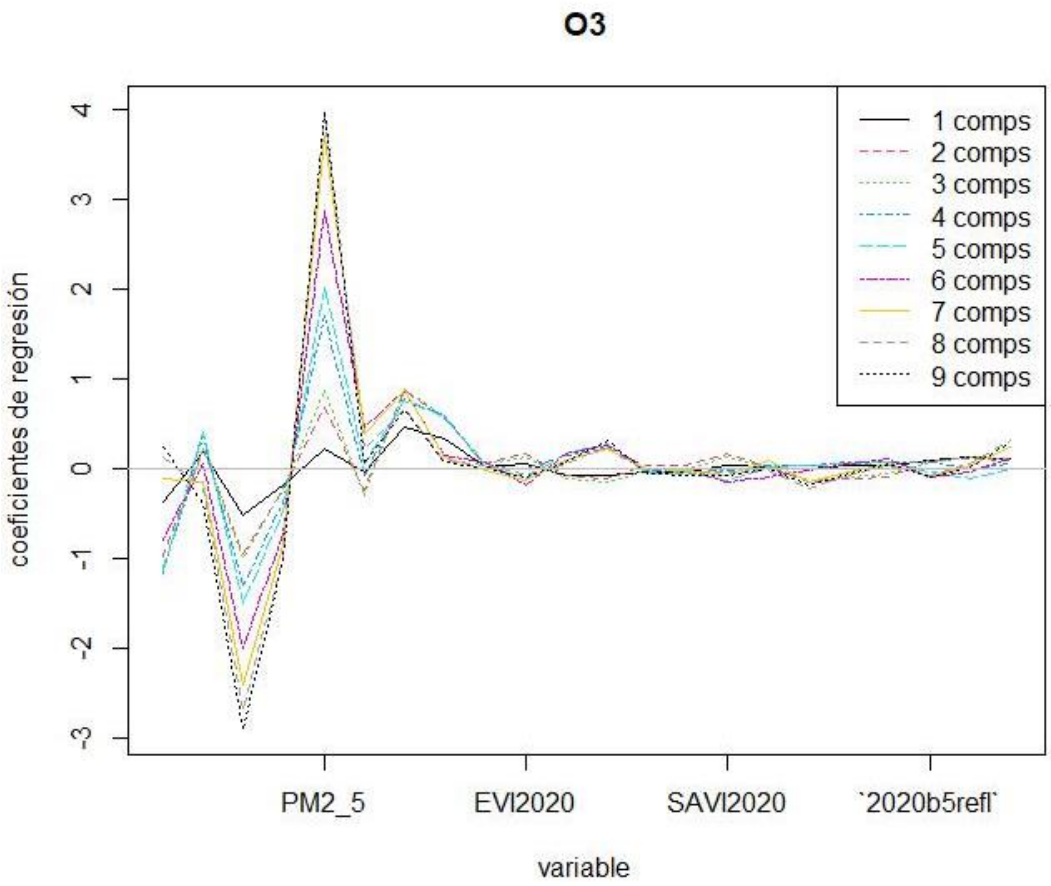


Figura 41. Coeficientes de regresión, modelo año 2020

Las mediciones usadas para el entrenamiento del modelo y las predicciones realizadas una vez ajustado con 9 componentes, para el caso del año 2020 fueron graficadas de igual manera mostrando la relación lineal que el modelo permitió establecer (Figura 44).

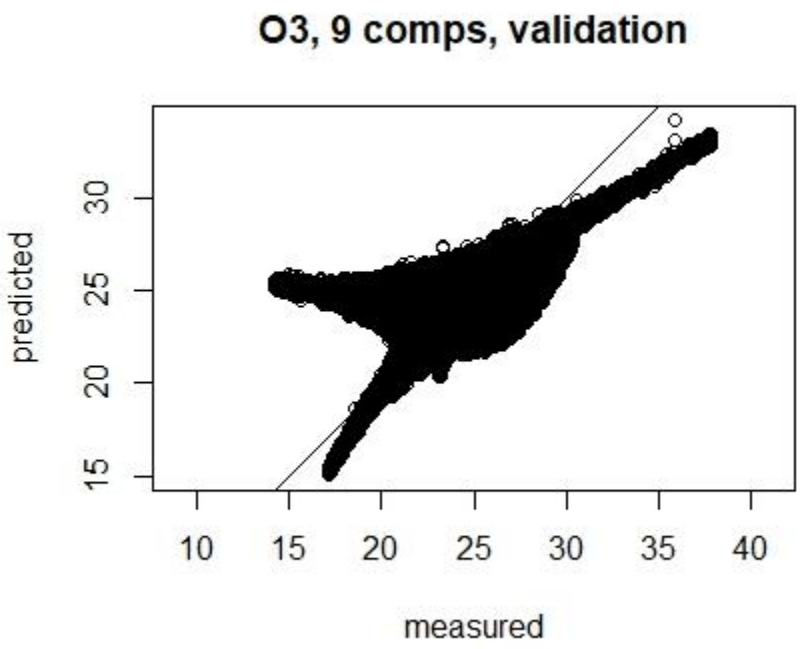


Figura 42. Mediciones vs Predicciones O3 (ppb) año 2020

Como se mencionaba anteriormente un aspecto importante al momento de ajustar el modelo PLS, previa identificación del número de componentes optimo es la validación del modelo. El paquete pls de R cuenta con la metodología de validación cruzada en la cual el modelo es corrido dejando una de las observaciones por fuera para luego ser evaluado con esta misma y realizando el mismo procedimiento con todas las observaciones. Este método es también el método por defecto utilizado, es decir, al no ser especificado en la línea de código, R realizara una validación cruzada, el otro método disponible es LOO (Leave One Out) similar a CV (cross validation) o bien, puede indicarse que no se realice validación con el comando validation=”none”. Realizando una comparación con investigación realizada por Burgos y Copos en el 2017 en la ciudad de Quito Ecuador es posible notar que en dicha investigación no se llevó a cabo algún método de validación por lo cual los resultados respecto del presente estudio no presentaban ningún número de componentes sugerido y dicha selección se hizo de manera visual. Además, los valores medidos vs predichos se ajustaron demasiado bien a la línea de tendencia (Burgos & Copo, 2017).

Respecto a la información considerada en dicho estudio en Ecuador, el número de estaciones consideradas fue únicamente de 4, mientras que para Bogotá se usaron de 12 a 20 estaciones con lo cual la información y cobertura fue considerablemente mayor. También, las variables ambientales usadas incluyeron el monóxido de nitrógeno y óxidos de nitrógeno, no obstante se incluyeron solamente 2 índices espectrales (NDVI y Pv) así como también la temperatura de brillo de las imágenes Landsat; mientras que en la actual investigación no se incorporó el monóxido de nitrógeno ni óxidos de nitrógeno como precursores importantes del ozono troposférico, aunque se consideró el dióxido de nitrógeno y además se incluyeron en total 6 índices espectrales relacionados con coberturas vegetales, humedad e incendios así como también la temperatura superficial.

De igual manera, en la investigación de Burgos & Copo (2017) se realizó únicamente una modelación, incluyendo información de los 3 años tenidos en cuenta concluyendo que un 96.39% de varianza de la variable dependiente ozono se logró con el uso de 10 componentes, mientras que en el actual estudio se realizaron 3 modelos con diferentes números de componentes sugeridos por el paquete de R lograron un porcentaje de varianza de la variable respuesta de solo 65.37%, 51.49% y 45.61% con 9, 10 y 7 componentes respectivamente para cada año. Los porcentajes son evidentemente menores sin embargo es importante considerar que el modelo acá descrito fue validado por validación cruzada y los componentes principales asignados de igual manera, mientras que en la investigación en Ecuador el numero de componentes fue escogido precisamente observando los valores de varianza que se explicarían al no considerar ninguna metodología de validación (Burgos & Copo, 2017).

Por su parte, Sinchi & Sagal (2018) lograron un 87,8% de varianza explicada con 12 componentes en una matriz multivariable conformada por 16 observaciones y 24 variables, sin embargo, a diferencia de Burgos y Copo la variable considerada medida por la red de monitoreo en Cuenca, Ecuador fue únicamente el ozono troposférico, el resto de las variables fueron índices espectrales e información de escenas Landsat 8. De manera similar, no se contó con un método de validación lo

cual explicaría el valor del porcentaje de varianza. Además, se menciona el uso de más de 100 imágenes Landsat para la obtención de los índices espectrales lo cual representa una mayor cantidad de información disponible para el modelo. La medida de error R2 también es importante, en el estudio se menciona, referenciando a Chang & Hanna (2004) que un modelo de calidad de aire para ser confiable debe encontrarse entre 0,5 y 0,9, logrando ellos un valor de 0.87 y en comparación, este estudio obtuvo un valor de R2 de 0.65, 0.51 y 0.46 para los años 2020, 2021 y 2022, respectivamente, quedando este último ligeramente por debajo del valor indicado y pudiendo estar asociado a la cantidad de componentes principales seleccionados (7), en comparación con los otros dos años (9 y 10). (Sinchi & Sagal, 2018)

La investigación realizada en Nanjing, China, a diferencia de las dos anteriores no incluyó ningún tipo de información satelital, no obstante, se usaron 531 muestras donde las variables incorporadas y la importancia de las variables meteorológicas, como la humedad relativa y los compuestos orgánicos volátiles en la matriz multivariable representaron un aporte relevante en la modelación. Además, la regresión PLS fue considerada más como una herramienta de pretratamiento de datos que posteriormente fue usada en la metodología KELM y SVR, más allá de eso realizaron combinaciones de metodologías de pretratamiento y de aprendizaje automatizado concluyendo que aquella incluyendo PLS (KELM-WT-PLS) arrojo el mejor R2, es decir 0,78. De esta manera se puede inferir que el incluir información relacionada más directamente con el ozono troposférico y una metodología de modelación superior a PLS puede mejorar los valores de R^2 obtenidos en este estudio. (Xiaoqian, Junlin, Yuxin, Ping, & Bin, 2020)

A continuación, se presentan los mapas obtenidos con la realización del estudio, se incluyeron los mapas de ozono medido obtenido a partir de las mediciones directas de la RMCAB para su comparación con los mapas de ozono troposférico predicho a partir de regresión PLS. Un aspecto notable al realizar la interpolación IDW de los valores predichos fue que gráficamente era posible distinguir tanto los valores de concentración de ozono, así como también coberturas como

construcciones o zonas rurales, esto debido a la información proveniente de las imágenes Landsat. En términos generales se puede observar que, en ninguna de las 3 fechas consideradas tanto para el ozono medido como el predicho, la normativa de ozono troposférico llego a ser excedida siendo esta una concentración de 100 $\mu g/m^3$ en un promedio de 8 horas. El valor máximo de concentración se presentó en agosto del 2020 con 75,6 $\mu g/m^3$ en la localidad de Kennedy y la concentración mínima en enero de 2022 en la localidad de Suba con un valor de 19,7 $\mu g/m^3$ (Figura 43 y 47)

Para el año 2020, las mayores concentraciones medidas se dieron en las localidades de Kennedy, Barrios Unidos, Tunjuelito y Usaquén mientras que aquellas predichas se presentaron en las mismas mostrando concentraciones más leves en Tunjuelito. En el año 2021 Usaquén, Fontibón y Suba presentaron las concentraciones más altas medidas mientras que en el modelo de predicción dichas concentraciones se concentraron más en la localidad de Teusaquillo, Suba y Usaquén; las concentraciones más bajas medidas fueron Usme y Tunjuelito y en la regresión se redistribuyó por el centro de la ciudad y Usme. Por su parte, a inicios del año 2022 las mayores concentraciones medidas se dieron nuevamente en Usaquén y aquellas predichas se distribuyeron por el oriente de la ciudad con un alto porcentaje en Usaquén mientras que las concentraciones más bajas, se presentaron en el occidente de la ciudad (Figuras 45 y 46).

Considerando únicamente los resultados obtenidos por mediciones se puede evidenciar que la zona más afectada es el norte de la ciudad pudiendo estar influenciado por la generación de contaminantes precursores en el occidente y centro de la ciudad que terminan siendo transportados. Así mismo, realizando la comparación de la información predicha se puede notar concentraciones más homogéneas pudiéndose fácilmente identificar puntos críticos, pero también ver el comportamiento a una escala mayor, por ejemplo, en agosto del año 2020, debido a la presencia de vientos, el centro de Bogotá tuvo las concentraciones más bajas aun cuando en Kennedy se dieron las más altas. Respecto al año 2021 el norte se encuentra nuevamente afectado pero la

concentración disminuye gradualmente sin afectación por vientos o presencia de nuevos puntos críticos hacia el sur. Finalmente, en el año 2022, se dieron resultados interesantes que indican mayores concentraciones en el oriente, de donde provienen los vientos y donde hay presencia relevante de vegetación. Dado que la vegetación genera compuestos orgánicos volátiles puede considerarse que precisamente, debido a condiciones meteorológicas tal vez ausentes en enero de 2022, la vegetación tuvo un aporte significativo al ozono troposférico o bien que fue generado en la urbe como Kennedy o Usaquén, pero se transportó y concentró en el oriente dispersándose en el occidente para las horas de la noche (Figuras 43, 45 y 47)

También, es importante considerar el comportamiento que se dio en la ciudad durante el año 2020 y 2021, donde iniciaron las series de cuarentenas a comienzos del 2020 y tomando su comportamiento normal a mediados del 2021 comportamiento reflejado en las concentraciones del 2020 y 2021, pudiéndose observar que aunque el punto crítico de Kennedy se mantuvo, en términos generales el centro de Bogotá D.C., generalmente concurrido por vehículos y los trabajadores de la ciudad presentó concentraciones bajas. Situación opuesta presentada en el año 2021 donde el centro de la capital y el norte, incluyendo Usaquén, probablemente volvieron a su generación normal de óxidos de nitrógeno, por vehículos y fábricas y, por lo tanto, de ozono troposférico.

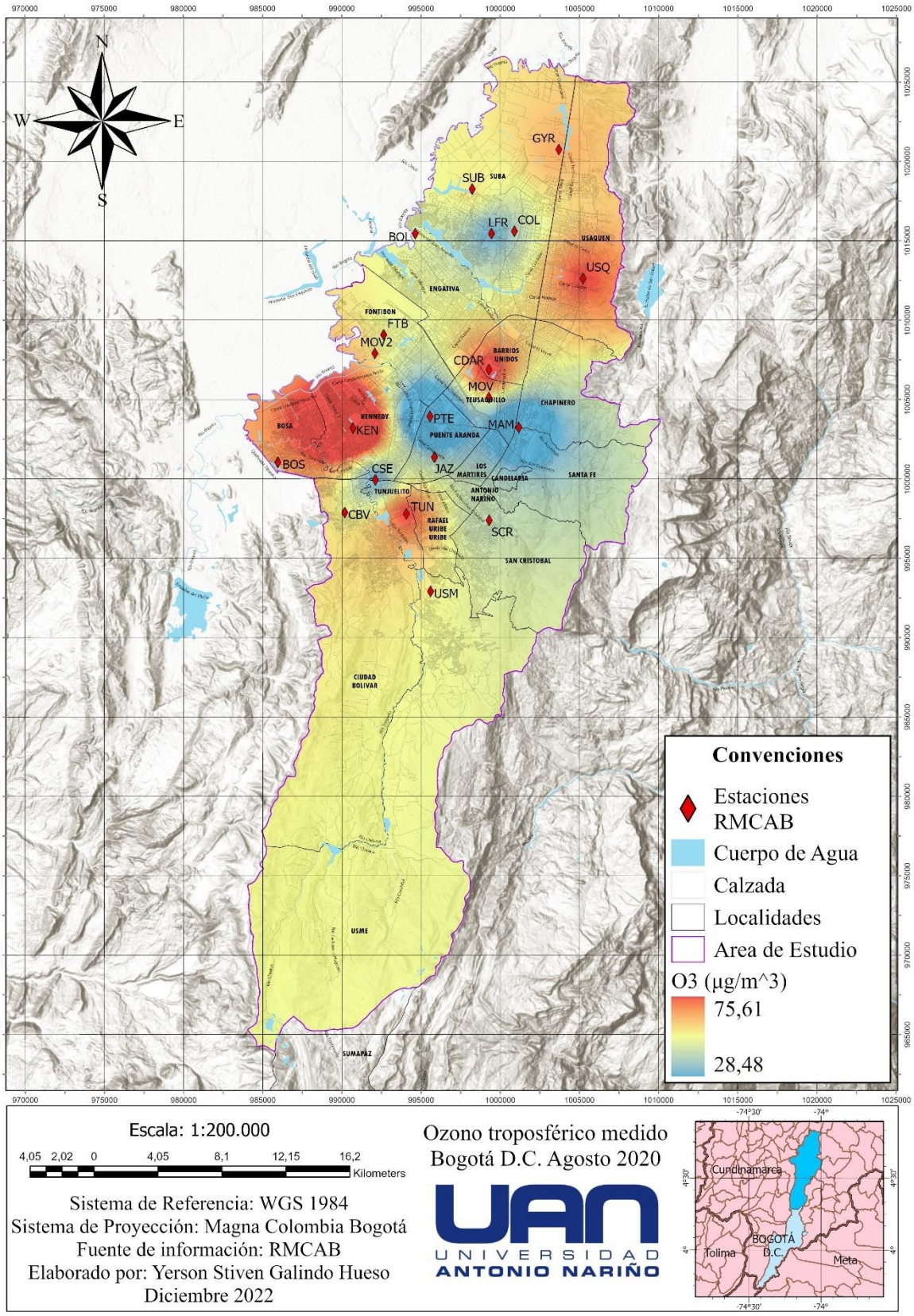


Figura 43. Mapa de ozono troposférico medido año 2020.

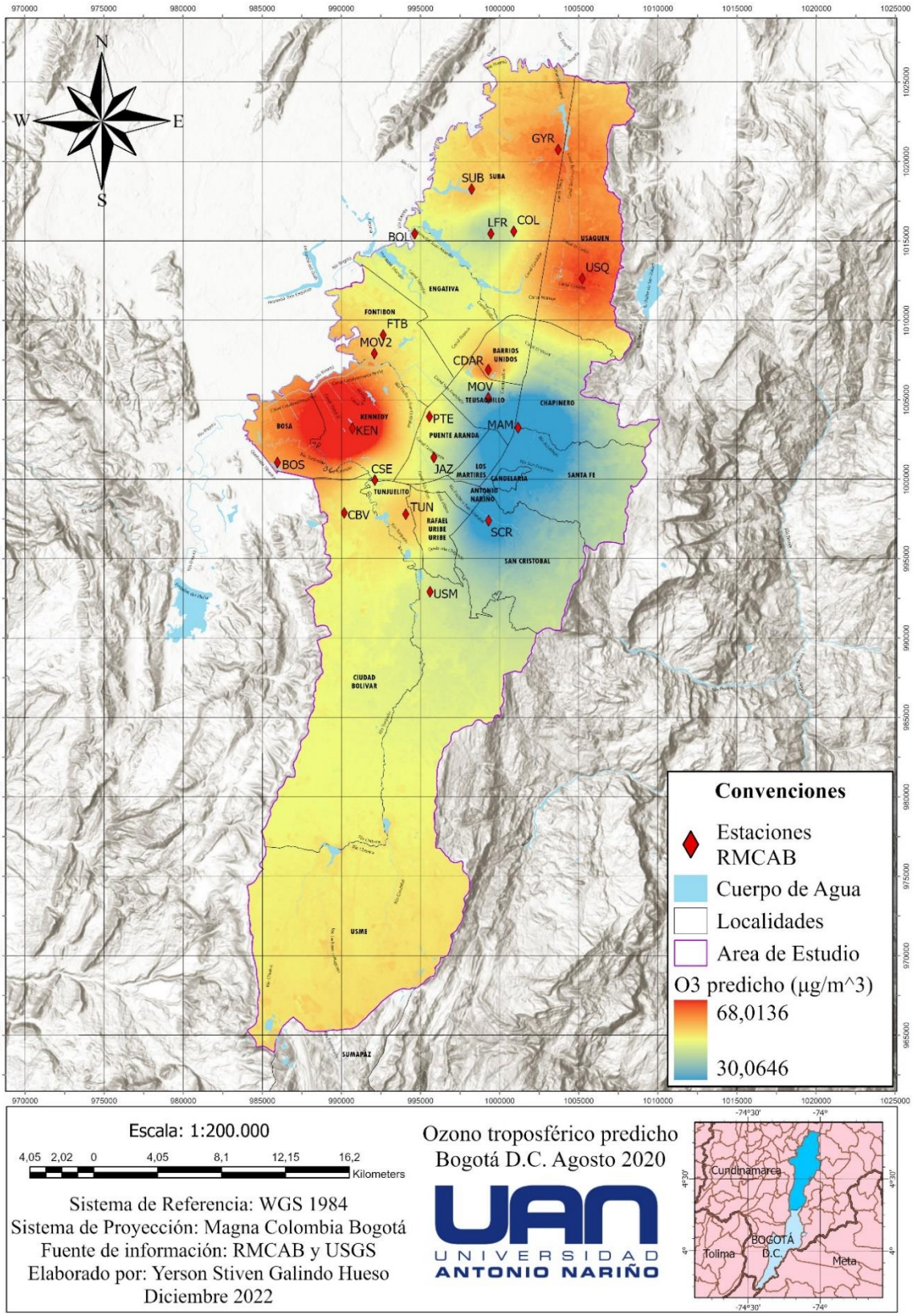


Figura 44. Mapa de ozono troposférico predicho año 2020.

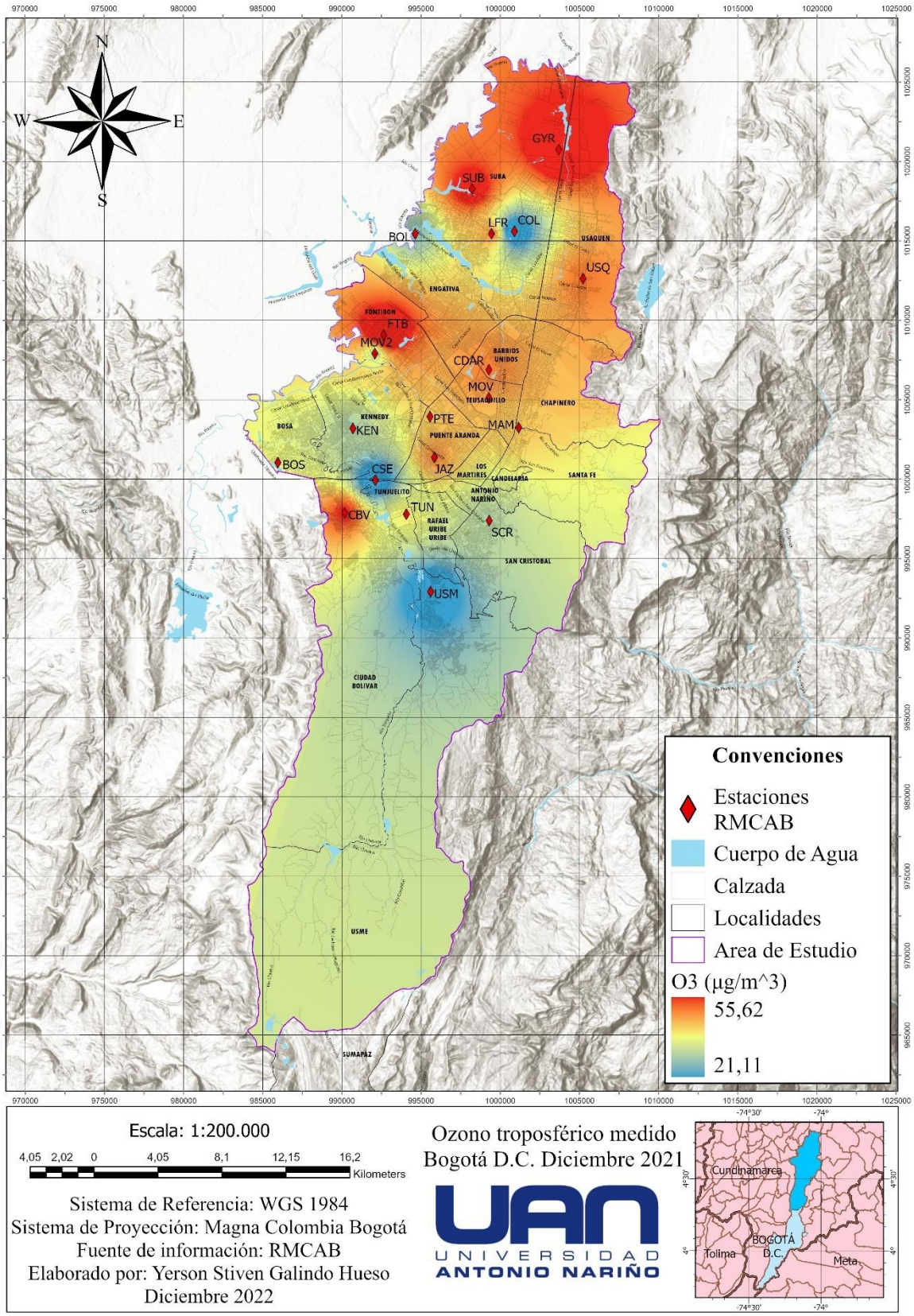


Figura 45. Mapa de ozono troposférico medido año 2021

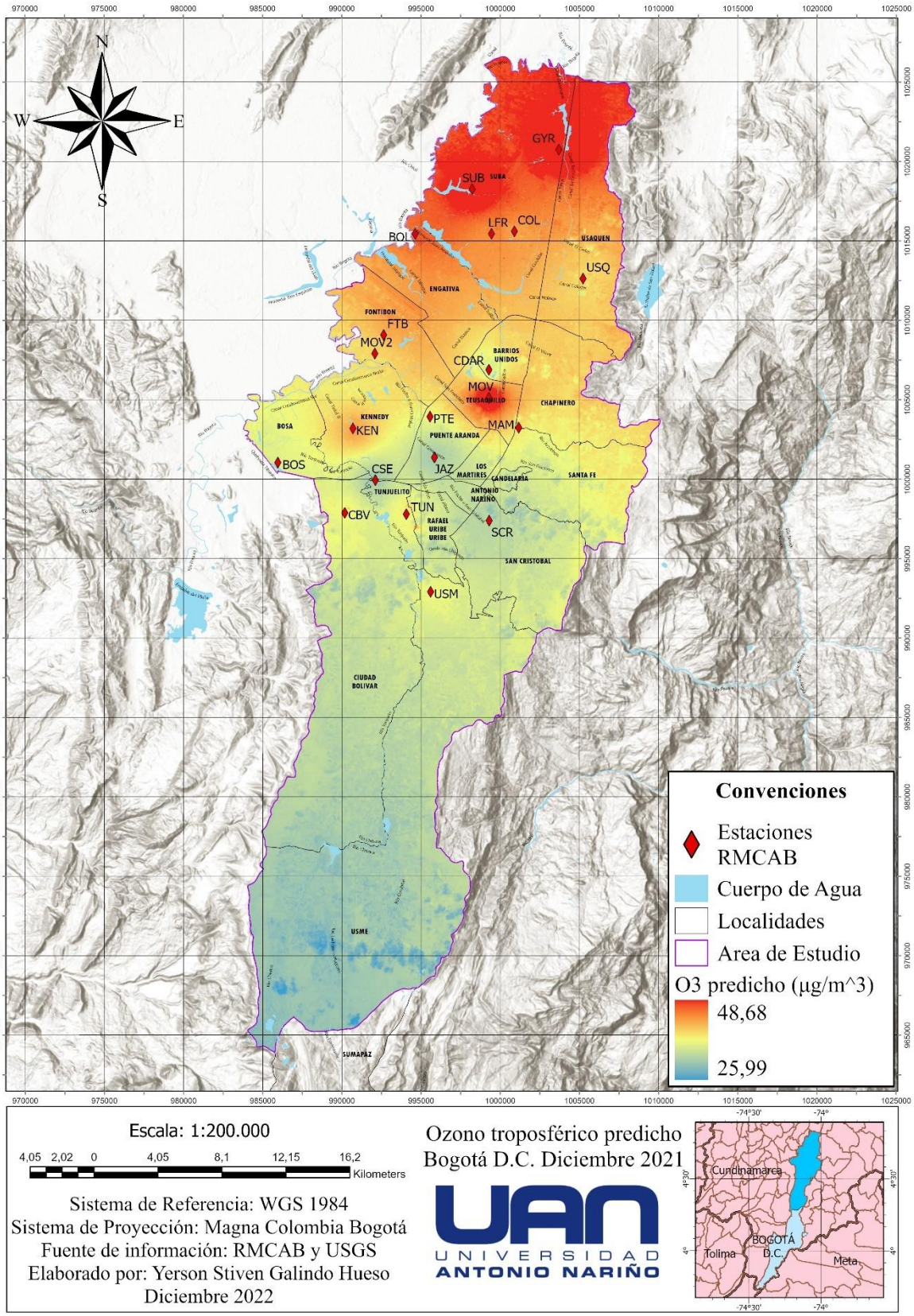


Figura 46. Mapa de ozono troposférico predicho año 2021.

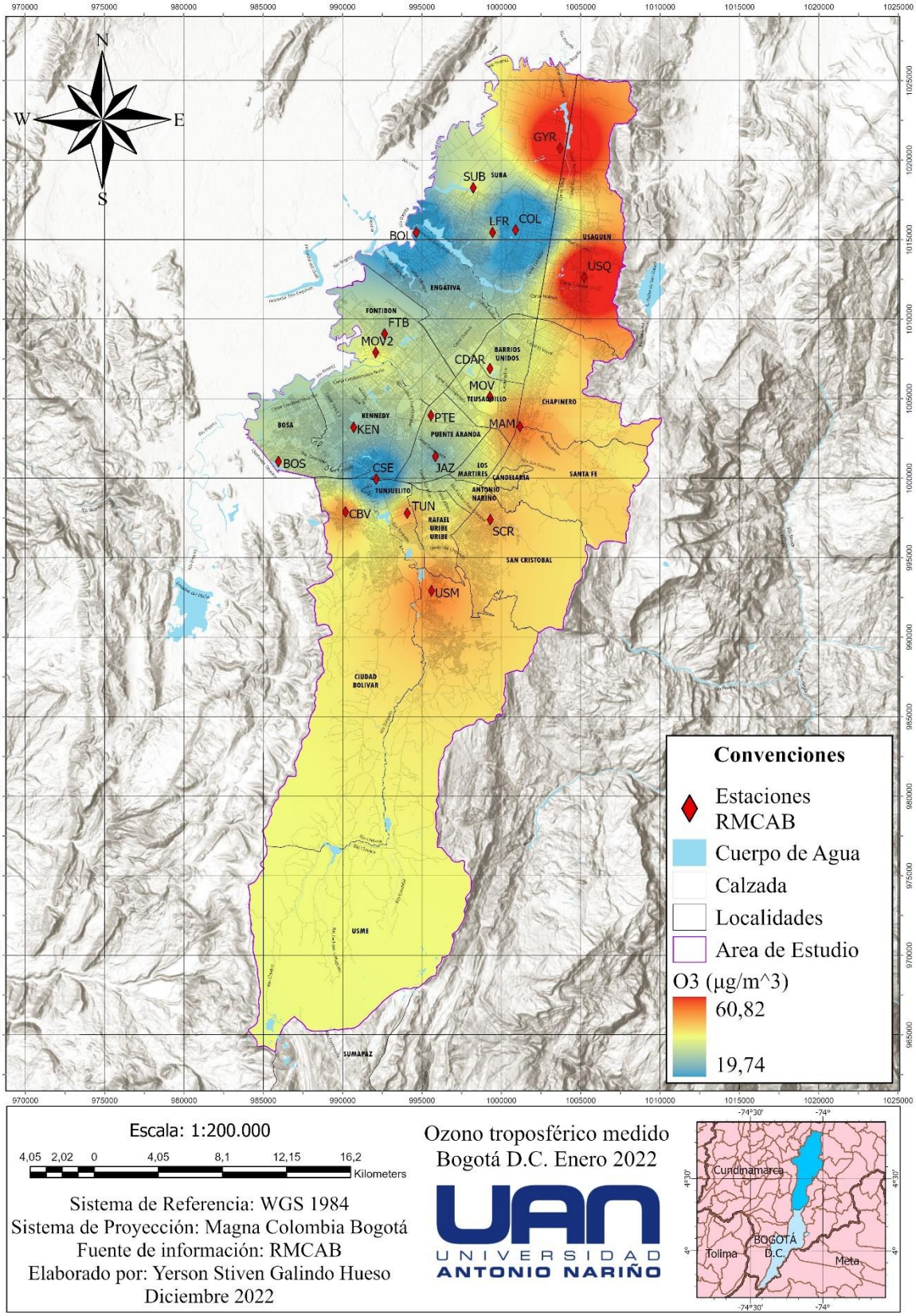


Figura 47. Mapa de ozono troposférico medido año 2022

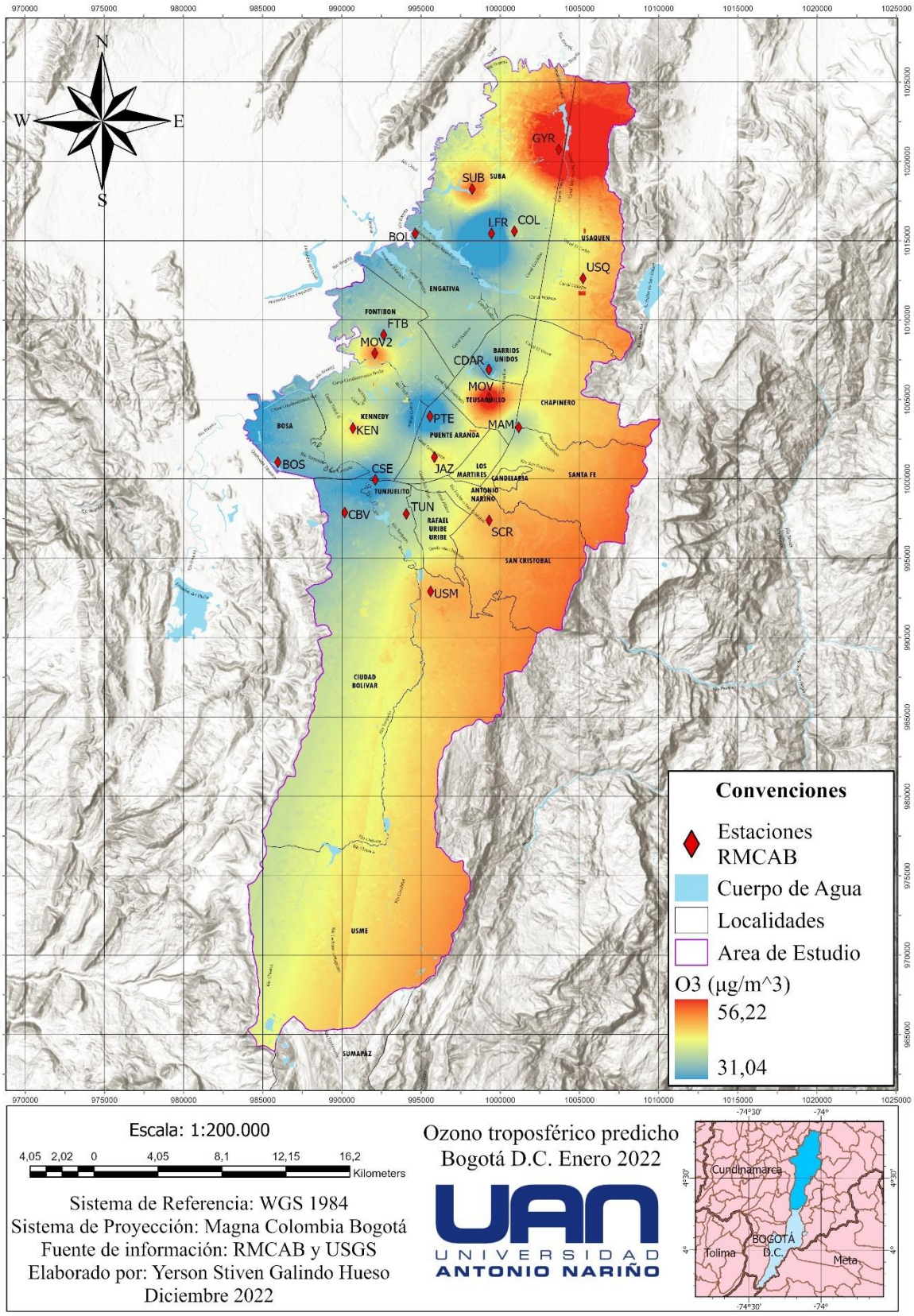


Figura 48. Mapa de ozono troposférico predicho año 2022

6. Conclusiones

La información recopilada, tanto de mediciones como de información geoespacial, representó diversidad en las variables consideradas para el estudio, yendo desde la radiación solar hasta índices espectrales como el NDMI y su relación con la humedad del área de estudio. No obstante, el objetivo de la investigación se basaba en que una regresión lineal incluyendo demasiadas variables altamente correlacionadas representa una problemática; el previamente llamado problema de colinealidad, donde la influencia individualizada de cada variable independiente sobre la variable dependiente no logra ser distinguida. Por esta razón, la implementación de la regresión PLS represento una ventaja sobre el PCR donde ambas son útiles cuando hay un alto número de variables y se logran reducir el número de éstas a componentes primarios; sin embargo, la segunda no toma en consideración la relación entre las variables explicativas X y el vector respuesta Y a diferencia de la primera metodología que sí lo hace.

Respecto a la información base recopilada, se logró la compilación de datos promedio horarios de contaminantes y variables atmosféricas para 3 años diferentes considerando de 12 a 19 estaciones de monitoreo, es decir, 49 observaciones que fueron extrapoladas a aproximadamente 3 millones de observaciones en total. Por otra parte, la información en los últimos 3 años de imágenes satelitales Landsat fue escasa pudiendo obtener únicamente 3 escenas cuyo porcentaje de nubes no interfería en la investigación siendo menores al 20% sin embargo, la información obtenida se encontró en el formato mas reciente de productos Landsat Science Product (SP) con lo que se pudo profundizar en la metodología e información asociada a productos Landsat de este tipo y que será la manera en la cual se presentaran al público en el futuro.

En adición, los índices espectrales NDVI, Pv, EVI, NDMI, SAVI y NBR pudieron ser obtenidos por dos procesos diferentes, el calculo manual y la solicitud a demanda de la información al Servicio Geológico de Estados Unidos. Junto con la información obtenida previamente se pudieron igualmente obtener los valores de cada índice considerado en cada pixel para cada uno de los años

junto con la información meteorológica y de contaminantes asociada, permitiendo así la generación de las matrices multivariable conformadas inicialmente por 12, 18 y 19 observaciones por 23 variables cada una que posteriormente contaron con 948852 observación por 23 variables cada una y representaron la información base con la cual se modeló la regresión PLS en cada año.

El modelo de regresión PLS por medio de validación cruzada logró determinar el numero de componentes optimo para cada año siendo 9, 10 y 7 componentes para el 2020, 2021 y 2022 representando con ello una varianza de la variable X o independientes del 86% 87.9% y 71.5%; de la variable respuesta ozono troposférico (O_3) un 65.37%, 51.49% y 45.61% y un coeficiente de determinación (R2) de 0.65, 0.52 y 0.46 respectivamente. De esta manera que los modelos para el año 2020 y 2021 se ajustan en una manera adecuada al comportamiento del ozono para su predicción mientras que el modelo para el año 2022 debió contar con datos fuera de rango o que incluso, el valor óptimo de componentes sugerido por R fue insuficiente para su desarrollo.

Con el desarrollo del modelo de regresión por mínimos cuadrados parciales se pudo finalmente calcular la concentración de ozono troposférico para la totalidad de los datos de cada año, se realizaron mapas de concentración medida y predicha en la ciudad de Bogotá para su comparación y con esto se pudo identificar que las zonas con las más altas concentraciones se encuentran en el norte y occidente de la capital con valores entre los 55 y los 75 μ/m^3, sin embargo, para las fechas consideradas, la norma de ozono troposférico no llegó a ser excedida siendo esta de 100 μ/m^3 para un promedio de 8 horas. Kennedy y Usaquén fueron las localidades mas afectadas donde la primera cuenta con un sector comercial e industrial al occidente mientras que en el norte, las concentraciones altas de Usaquén pueden deberse a ser una zona suburbana donde los compuestos orgánicos junto con el parque automotriz influyen en dichas concentraciones. La información más atípica fue también obtenida para el año 2022 donde las mayores concentraciones se dieron sobre los cerros orientales con coberturas vegetales y vientos favorables, siendo la información obtenida de meses con alta incidencia de radiación como lo es enero esto pudo deberse a ausencia de vientos

o probablemente el ozono se generó en la urbe, Kennedy o Puente Aranda y en las horas de la noche se transportó hacia esta zona. La segunda hipótesis puede estar relacionada con los solamente 7 componentes usados en la regresión que influyeron en los puntos críticos medidos que fueron Usaquén y Suba y la distribución consecuente hacia el oriente con los datos predichos.

La metodología de regresión por mínimos cuadrados parciales representa una herramienta importante en investigaciones ambientales para la calidad del aire, porque su uso no se restringe a campos específicos de aplicación, permitiendo así realizar predicciones de concentraciones de contaminantes en casos donde se cuente con únicamente ciertas variables o incluso para determinar el grado de influencia que variables independientes tienen sobre otra dependiente; además de considerar el importante valor agregado que tiene la percepción remota como fuente de información ambiental.

7. Recomendaciones

Dada la naturaleza del ozono para investigaciones similares se recomienda incluir variables ambientales mucho mas relacionadas con este contaminante como los compuestos orgánicos volátiles, la velocidad y dirección del viento, así como también la humedad relativa, determinada como altamente correlacionada con el ozono en estudios similares. También se podría considerar realizar mediciones manualmente garantizando la totalidad de datos requeridos. De igual manera tomar mediciones en rangos mayores de tiempo más allá de un solo día para un más amplio espectro de análisis e interpretación incluida la cantidad de escenas geo satelitales, además se recomienda incorporar información satelital proveniente de diferentes misiones, con mayores resoluciones espaciales y temporales.

El área de estudio también puede ser revisada pudiendo mejorarse los resultados con puntos de mediciones mejor distribuidos sobre una zona y en donde la falta de información en específicos limites no influya en resultados sensibles a incorrectas interpretaciones.

Con este trabajo se puede predecir el comportamiento del ozono en zonas con características ambientales similares a Bogotá en las cuales se pretenda expandir las zonas urbanas o aumentar el flujo de automóviles pudiendo indicar como se comportaría el ozono si se llegase a ciertos niveles de NO2, de COVs o de especificas coberturas relacionadas con coberturas vegetales o de suelos desnudos que influyan en el ozono troposférico. Sería útil como herramienta de planeación urbana y de alerta de calidad de aire

Tener en cuenta las características de validación con las que cuentan los modelos de regresión es importante pues representan ayudas estadísticas para su entendimiento e interpretación. Realizar modelación en entornos de lenguaje y programación como R indica que los procesos llevados a cabo son lo suficientemente extensos para no poder ser realizados manualmente y contando con procesos validación se garantiza que la información de entrada fue analizada y revisada de manera

optima obteniendo los resultados mas acorde a la modelación, independientemente de la posibilidad de escoger manualmente aspectos del modelo las recomendaciones tienen fundamentación en la matemática del modelo.

Como se mencionaba en un estudio realizado en china la regresión pls no fue la metodología central sino una herramienta de pretratamiento de datos por lo cual se recomienda para estudios similares el desarrollo de códigos de aprendizaje automático avanzados como KELM y SVR que incluyen PLS como paso intermedio.

Bibliografía

Alciaturi, C., Escobar, M., De La Cruz, C., & Rincon, C. (2003). *Partial least squares (PLS) regression and its application to coal analysis.* Venezuela: Universidad del Zulia.

Alonso, L. E. (2017). *Modelo PLS.* Sevilla, España: Departamento de estadistica e investigacion operativa .

Aparicio, G., & Caballero, A. (2009). *La regresion por minimos cuadrados parciales: origenes y evolucion.* España: Historia de la probabilidad y estadistica.

Bjorn-Helge, M., & Wehrens, R. (2022). *Introduction to the PLS package.* Oslo, Norway: University Center for Information Technology.

Burgos, M., & Copo, K. (2017). *Estimacion de la concentracion de Ozono troposferico mediante analisis geoespacial de imagenes satelitales y minimos cuadrados parciales para las parroquias urbanas del canton Quito.* Sangolqui: Universidad de las fuerzas armadas ESPE.

Caicedo, Y. C., Tomás, B., & Antonio, M. (2010). EMISIONES DE COMPUESTOS ORGÁNICOS VOLÁTILES DE ORIGEN BIOGENICO. *Revista Intropica*, 10.

Cortez, J. (2013). *Estimacion de emisiones de metano del Relleno Sanitario Bordo Poniente por medio de imagenes de satelite.* Mexico D.F: Centro Interdisciplinario de investigaciones y Estudios sobre Medio Ambiente y Desarrollo (CIIEMAD).

Coy, Y. A. (2021). *Análisis Multitemporal de Cambio en el Espejo de Agua del Río Cusiana Mediante.* Bogota D.C.: Universidad Antonio Nariño.

Dominguez, P. P. (2012). *Estudio multifractal de la influencia de factores meteorologicos y quimicos en la concentracion de ozono troposferico.* España: Universidad de Cordoba.

Dunn, K. (2022). *Process Improvement Using Data.* Ontario, Canada: Creative Commons Contribution.

EPA. (2000). El Smog. Lo que usted necesita saber acerca del Ozono y su salud. *United States Environmental Agency*, 10.

IGAC. (2013). *Descripcion y correccion de productos Landsat 8* . Bogota, Colombia: Instituto Geografico Agustin Codazzi.

Jones, R., & Wigley, T. (1991). *Algunos datos sobre el ozono y la salud.* Boletin de la oficina sanitaria panamericana.

Michelle, B., & Copo, K. (2017). *Estimacion de la concentracion de ozono troposferico mediante analisis geoespacial de imagenes satelitales y minimos cuadrados parciales, para las parroquias urbanas del canton Quito.* Sangolqui, Ecuador: ESPE.

Santos, D. P. (2017). *ESTUDIO DE LA ISLA DE CALOR URBANO.* Quito: Universidad Internacional SEK.

SDA. (2021). *Informe Anual de la Calidad del Aire de Bogotá año 2021.* Bogotá: RMCAB.

SDA. (2022). *Informe mensual de calidad del aire de bogota Enero 2022.* Bogota D.C.: SDA.

SDA. (2022). *Red de monitoreo de calidad del aire de bogota, RMCAB.* Bogota D.C: SDA.

SDA, S. D. (2021). *Informe anual de calidad de aire de bogota Año 2020.* Bogota D.C. : RMCAB.

Seijas, S. L. (2004). *El ozono troposferico.* Obtenido de Academia.edu.

Sinchi, A., & Sagal, C. (2018). *Valoracion de la concentracion de ozono y dioxido de azufre a traves de sensores remotos en el area urbana en la ciudad de cuenca.* Cuenca, Ecuador: Universidad Politecnica Salesiana.

Tek, K. (2018). *NDVI, NDBI & NDWI Calculation Using Landsat 7, 8.* Calgary, Canada.

Tilevik, A. (Febrero de 2022). *Partial least squares regression (PLSR).* Obtenido de TileStats: https://www.tilestats.com/our-story/

USGS. (2020). *Landsat 8-9 Operational Land Imager (OLI) - Thermal Infrared Sensor (TIRS) Collection 2 Level 2 (L2) Data Format Control Book (DFCB).* South Dakota: United States Geological Service.

USGS. (2022). *Landsat 8-9 Collection 2 (C2) Level 2 Science Product (L2SP) Guide.* South Dakota: Unites States Geological Service.

USGS, U. S. (2022). *EROS Science Processing Architecture (ESPA) On-Demand Interface User Guide.* South Dakota: USGS.

Vega, J., & Josue, G. (2010). *Regresion PLS y PCA como solucion al problema de multicolinealidad en regresion multiple.* Puerto Rico: Revista de matematica: Teoria y aplicaciones.

Xiaoqian, S., Junlin, A., Yuxin, Z., Ping, Z., & Bin, Z. (2020). *Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods.* China: Turkish National Committee for Air Pollution Research and Control.

Zhuzhingo, C. (2017). *Efecto del Ozono Troposferico en la Fisiologia de Quinua (Polylepis Reticulata) en el Parque Nacional Cajas.* Cuenca, Ecuador: Universidad de Cuencas.