ANÁLISIS DE LA CALIDAD DEL AIRE

Anna Cabrerizo Requena, Luis Miguel Rioja Gallo

Máster de Ciencia de Datos, UV

14 de novembre de 2023

Índex

Introducción y objectivo

Procesamos datos del Estudio Estadístico del INE sobre Calidad del Aire, empleando un visor que categoriza en 6 niveles. La normativa de septiembre de 2020 de la Dirección General de Calidad y Evaluación Ambiental detalla el método para evaluar el Índice de Calidad del Aire (ICA).

Procesamiento de datos

Formato tidy

Combinamos mediciones diarias de agentes químicos del aire de múltiples archivos en un conjunto de datos organizado, facilitando el análisis de variables durante cinco años.

PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUESTREO	ANNO	MES	DIA	H01	H02
1	22	1	8	01022001_8_8	2018	1	1	2	0
1	22	1	8	01022001_8_8	2018	1	2	2	3
1	22	1	8	01022001_8_8	2018	1	3	1	1
1	22	1	8	01022001_8_8	2018	1	4	3	3
1	22	1	8	01022001_8_8	2018	1	5	1	1
1	22	1	8	01022001_8_8	2018	1	11	2	1

Figura: Formato Inicial

Procesamiento de datos

Formato tidy

Fecha	MUNICIPIO	PROVINCIA	ESTACION	C6H6.num	CO.num	NO2.num	NOx.num	O3.num
2022-01-01	5	12	5	*NA*	0.1	26.00	*NA*	63.00
2022-01-01	9	3	6	*NA*	0.1	22.00	*NA*	60.00
2022-01-01	9	12	7	*NA*	0.2	46.00	*NA*	58.00
2022-01-01	10	46	1	*NA*	0.2	15.00	*NA*	48.00
2022-01-01	14	3	6	2.1	0.5	64.00	*NA*	72.00
2022-01-01	14	3	8	*NA*	*NA*	61.00	*NA*	72.00

Figura: Unimos datasets

N_PROVINCIA	N_MUNICIPIO	LATITUD_G	LONGITUD_G	TIPO_AREA	Fecha	ESTACION	C6H6.num
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-07-29	1	*NA*
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-02-15	1	*NA*
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-11-14	1	*NA*
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-08-02	1	*NA*
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-11-24	1	*NA*
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-08-27	1	*NA*
CASTELLÓN/CASTELLÓ	TORRE ENDOMÉNECH	40.26944	-0.07889	RURAL	2018-06-24	1	*NA*

Figura: Metadatos

Procesamiento de datos

Limpieza de datos - Outliers

Chemical_Element <chr>	IQR <int>	Hampel <int>	Percentil <int>
C6H6.num	550	4660	6373
CO.num	1658	6215	789
NO.num	2191	2169	7247
NO2.num	961	20733	12282
NOx.num	1750	8298	353
O3.num	2435	2336	10336
PM10.num	399	1374	2525
PM25.num	6661	6048	8916
SO2.num	3644	3496	3768

Figura: Modelo Lineal Múltiple

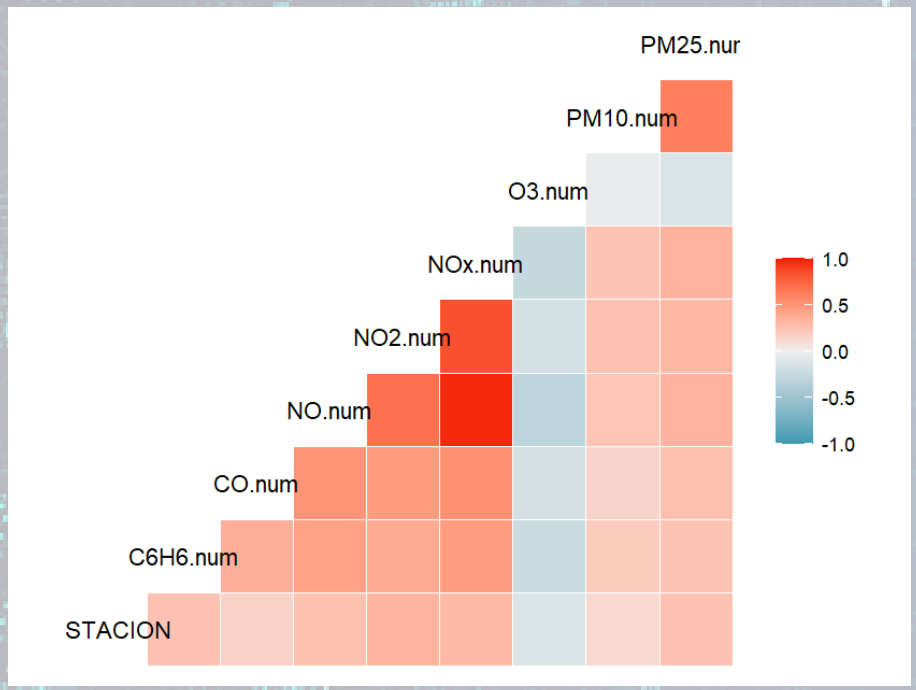
SO_2	PM25	PM10	O_3	NO_2	Cat.índice
751-1250	76-800	151-1200	381-800	341-1000	Ext.Desfavorable

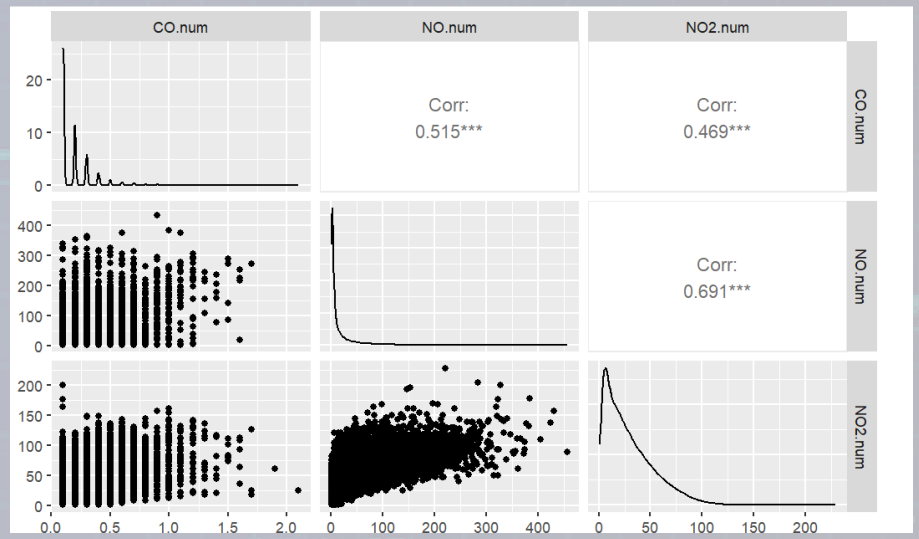
Taula: Criterio para comparar

Procesamiento de datos

Limpieza de datos - Datos Faltantes NA

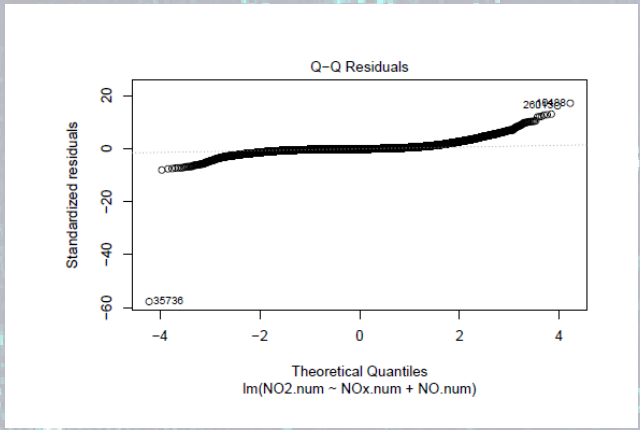
En primer lugar analizamos las covarianzas y correlaciones entre los datos para poder encontrar posibles predictores

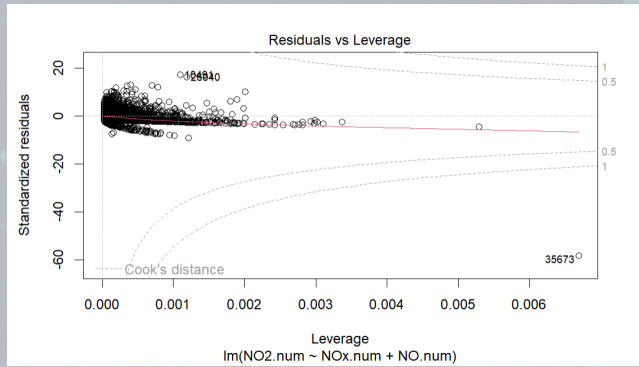


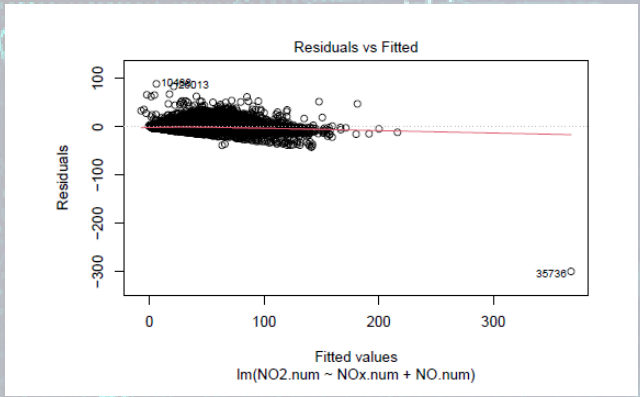


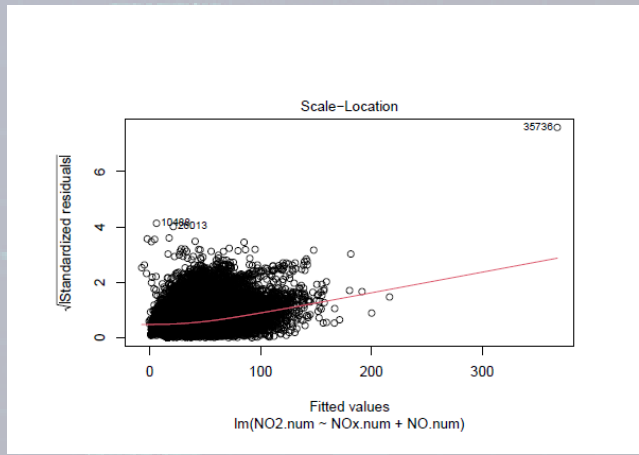
Procesamiento de datos

Limpieza de datos - Datos Faltantes NA - Modelo Predictivo









Procesamiento de datos

Limpieza de datos - Datos Faltantes NA

Imputamos los valores MNAR entre estaciones mediante KNN tras transformar coordenadas y fechas, y categorizar Tipo de Área.

O3.num <dbl>	PM10.num <dbl>	PM25.num <dbl>	SO2.num <dbl>	Dias <dbl>	X_UTM30N <dbl>	Y_UTM30N <dbl>
98	28	24	4	209	748380.2	4461758
68	16	10	3	45	748380.2	4461758
89	22	15	3	317	748380.2	4461758
99	49	44	4	213	748380.2	4461758
84	4	3	3	327	748380.2	4461758
84	22	14	4	238	748380.2	4461758

Figura: Modelo Lineal Múltiple

Cáculo del ICA

Limpieza de datos

Evaluamos la calidad del aire mediante el ICA, el cual se obtiene al categorizar las variables numéricas y seleccionar la categoría más desfavorable.

NO2.cat	O3.cat	PM10.cat	PM25.cat	SO2.cat	ICA
Buena	Razonablemente Buena	Razonablemente Buena	Regular	Buena	Regular
Buena	Razonablemente Buena	Buena	Buena	Buena	Razonablemente Buena
Buena	Razonablemente Buena	Razonablemente Buena	Razonablemente Buena	Buena	Razonablemente Buena
Buena	Razonablemente Buena	Regular	Desfavorable	Buena	Desfavorable
Buena	Razonablemente Buena	Buena	Buena	Buena	Razonablemente Buena
Buena	Razonablemente Buena	Razonablemente Buena	Razonablemente Buena	Buena	Razonablemente Buena

Figura: Categoricas para calcular el ICA

Análisis Univariante

Estadísticos

En primer lugar proporcionamos un resumen estadístico de los datos

	Fecha	NO_2	O_3	PM10	PM25	SO_2
Mínimo	2018-01-01	1.00	1.00	1.00	0.00	1.00
Q1	2019-04-07	9.00	74.00	12.00	7.00	3.00
Mediana	2020-07-06	21.00	88.00	22.00	12.00	4.00
Media	2020-07-04	27.91	88.60	29.34	15.42	6.359
Q3	2021-10-04	41.00	103.00	37.00	19.00	6.00
Máximo	2022-12-31	229.00	207.00	757.00	296.00	239.00

Taula: Estadísticos

Análisis Univariante

Visión gráfica de los componentes

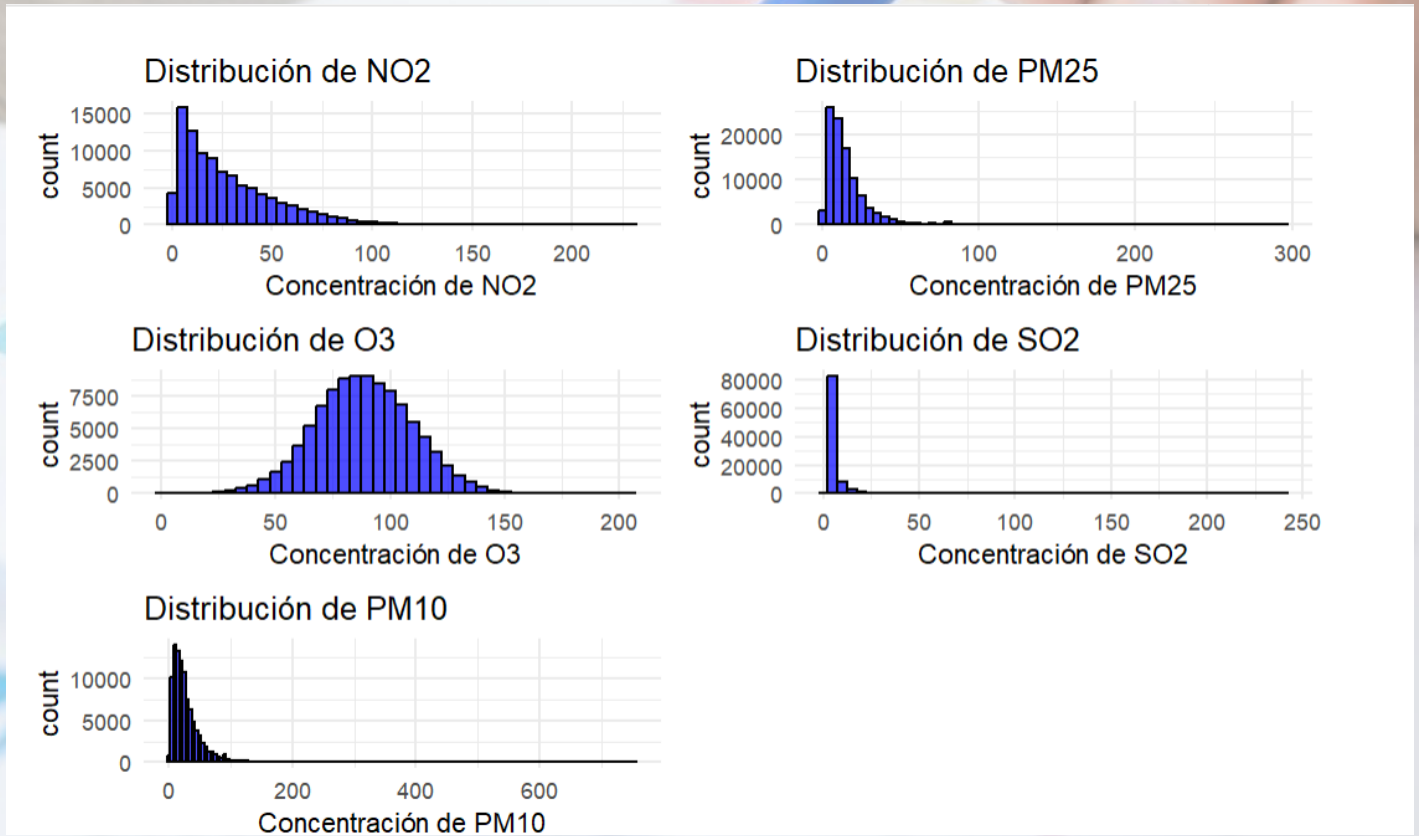


Figura: Densidad partículas

Análisis Univariante

Visualizamos nuestras partículas a lo largo de estos años

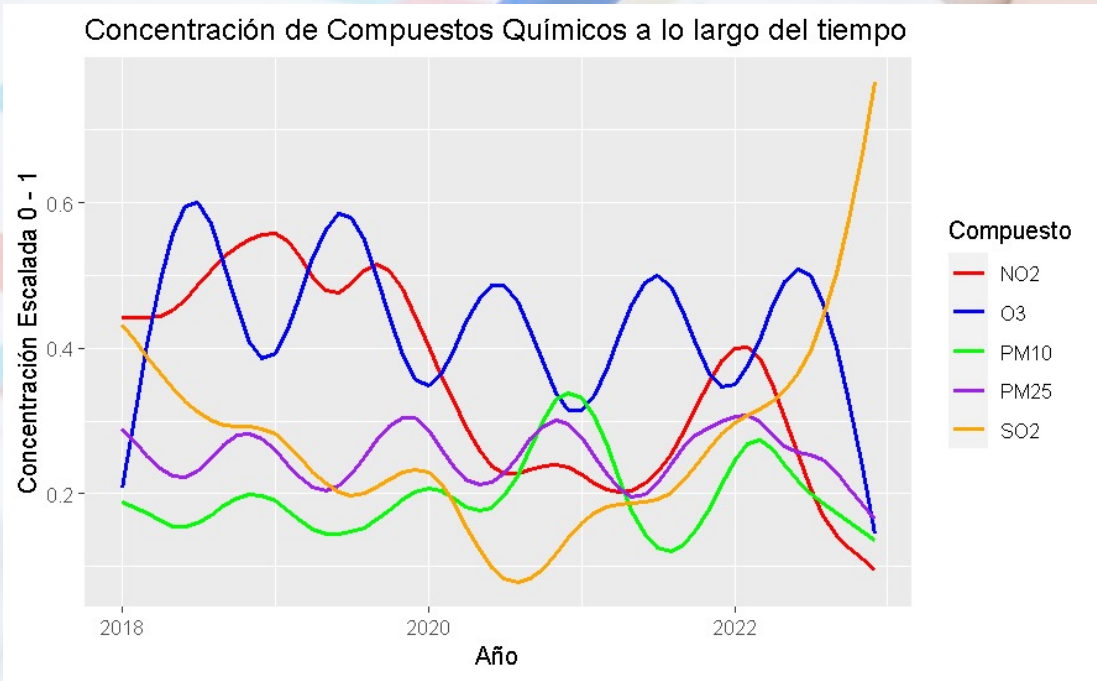
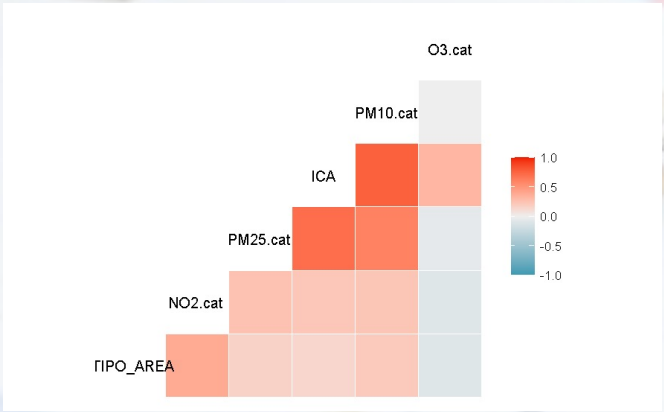
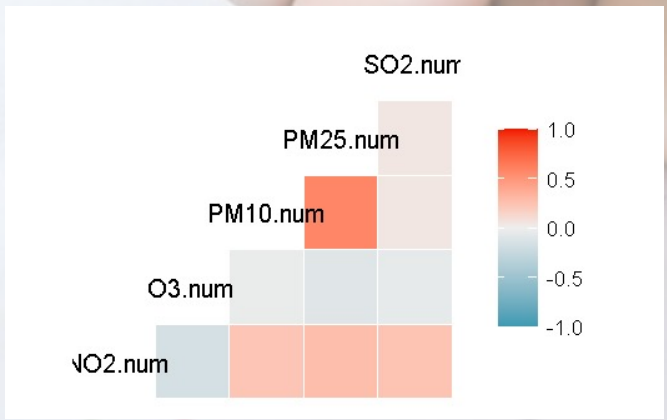


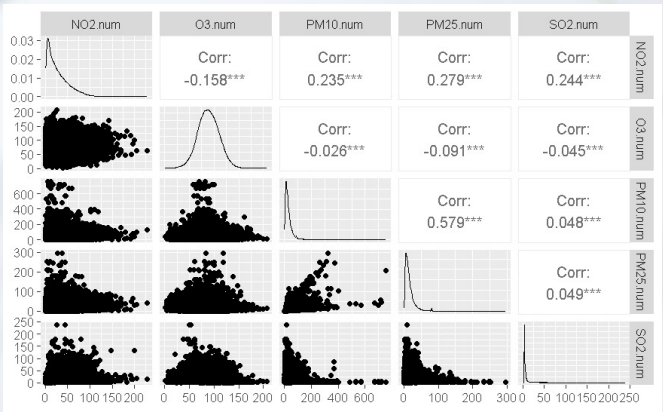
Figura: Compuestos en el tiempo

Análisis Bivariante

Ahora, establecemos relaciones entre las variables para poder analizar que factores afectan a la calidad del aire.







Análisis Bivariante

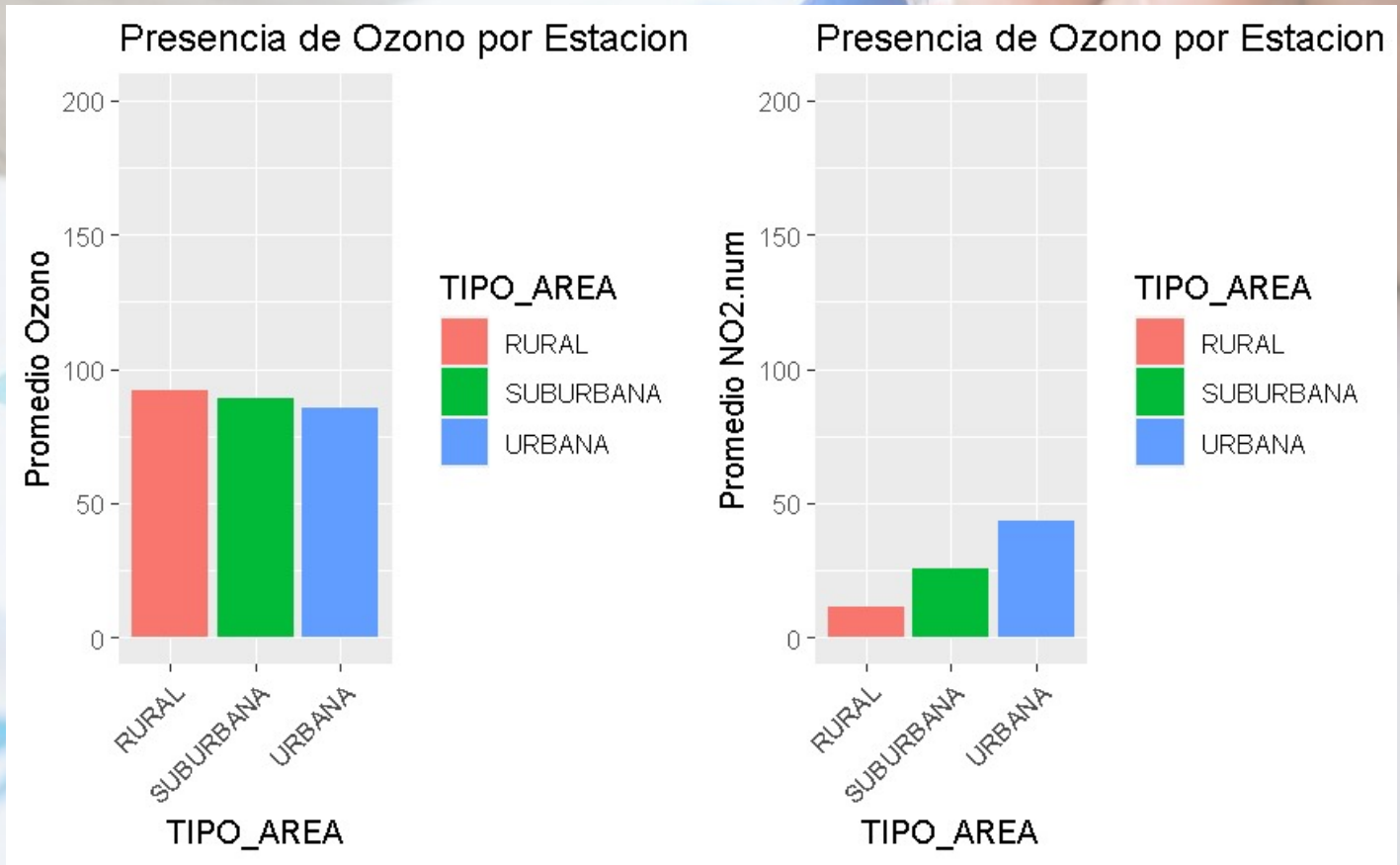


Figura: Presencia de compuestos en determinadas zonas

Análisis Bivariante

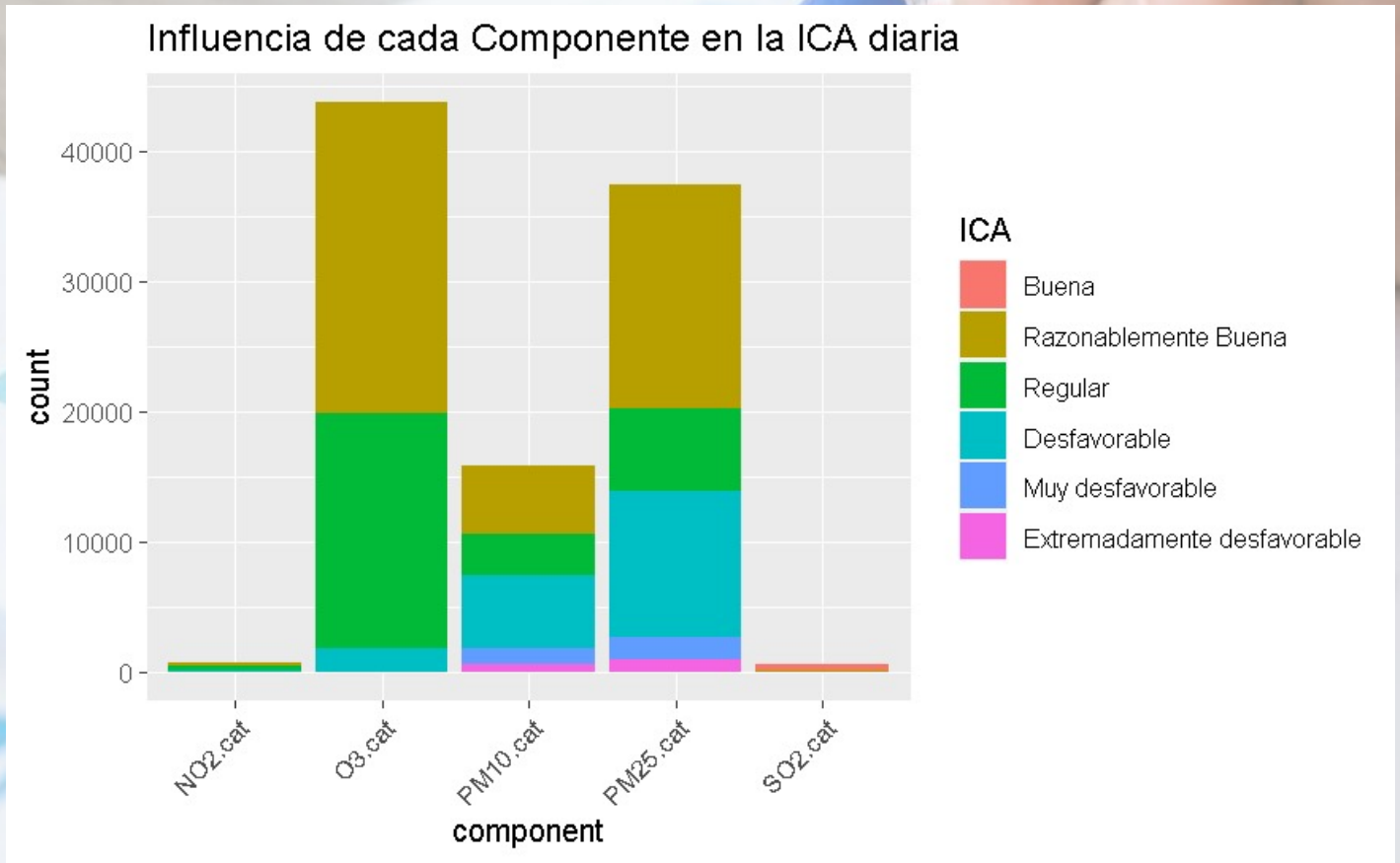


Figura: Influencia del componente en el ICA

Análisis Bivariante

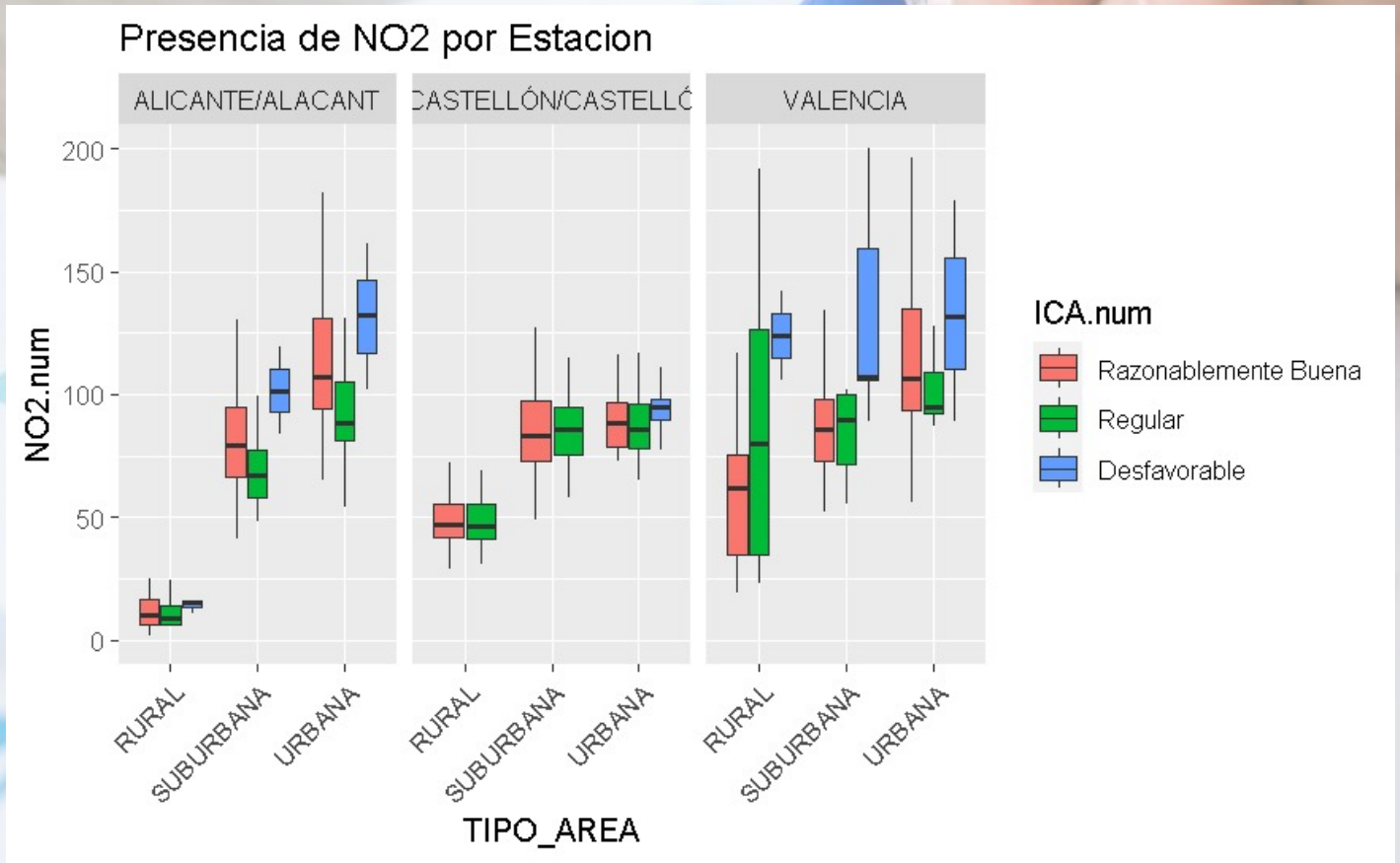


Figura: NO2 por estación

Análisis Bivariante

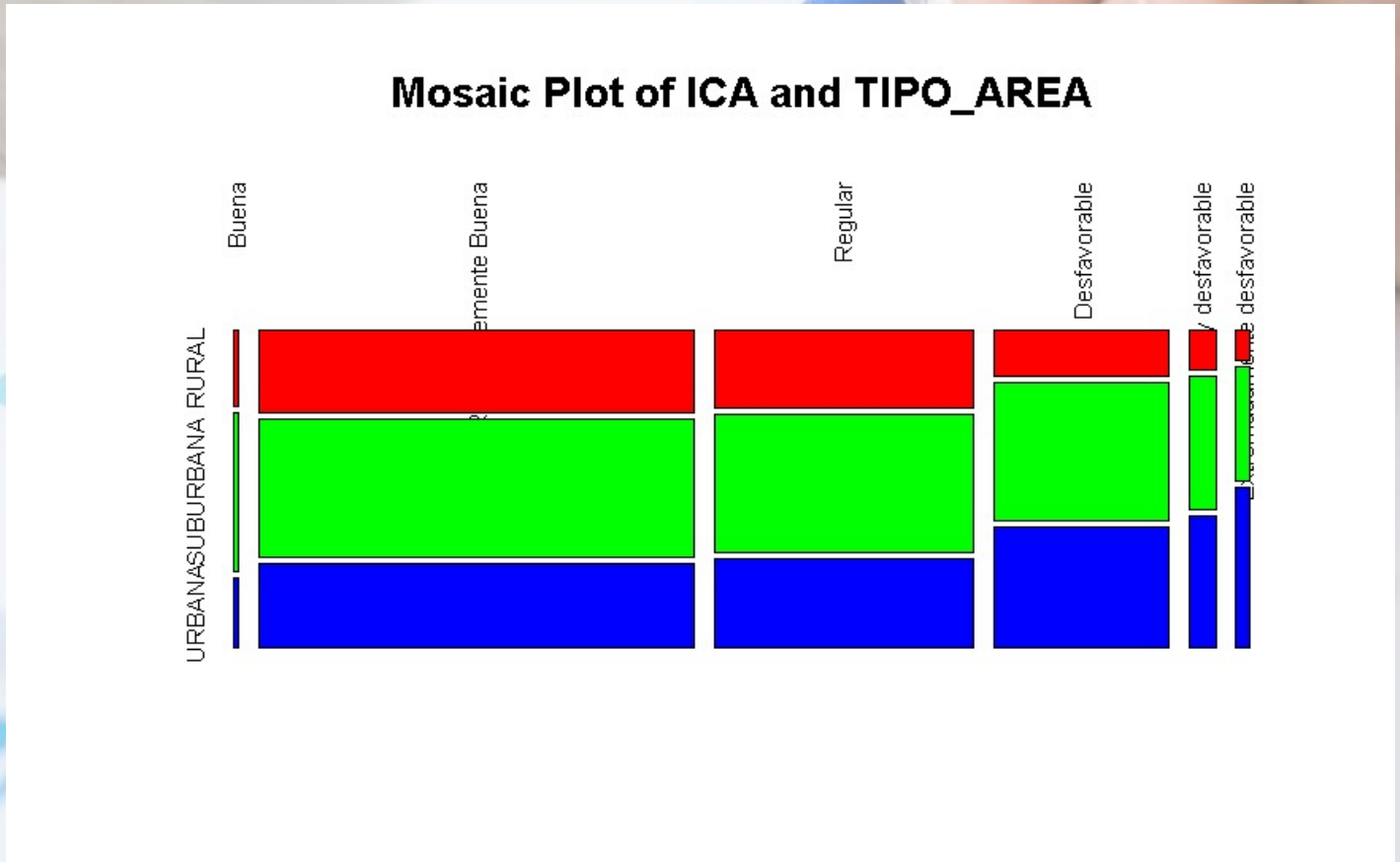


Figura: ICA por Área

Conclusiones

- Existe una fuerte correlación entre el ICA y las partículas en suspensión

- El ICA se ha mantenido estable a lo largo del periodo estudiado, muchos componentes disminuyeron durante la pandemia (COVID 2019) sobretodo los generados por los humanos. No obstante estos valores, se mantenieron en el nivel de "Razonablemente bueno".