Tecnológico de Monterrey

Materia:
Análisis de Métodos Multivariados para Ciencia de Datos

Reto - Etapa 3

Profesores:
Blanca Rosa Ruiz Hernández
Monica Guadalupe Elizondo Amaya

Alumnos:
Gian Marco Innocenti A00834310
Andrés Villarreal González A00833915
Andrea Hernandez A00835225

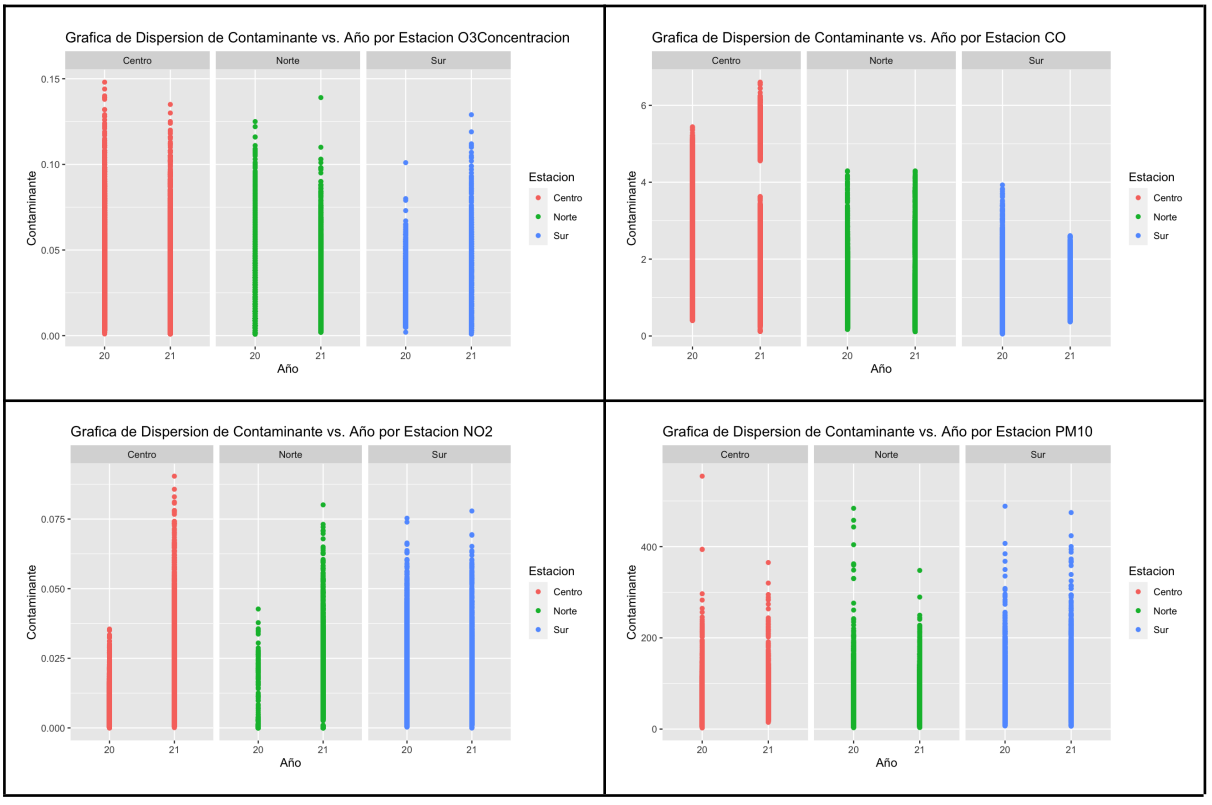
5 de octubre de 2023, Monterrey, N.L.

Objetivo

El objetivo de nuestro análisis es identificar si hubo una diferencia significativa en la calidad del aire en el área metropolitana de Monterrey entre los años 2020 y 2021. Para esto se escogieron los registros de una estación en el Centro,Norte, y Sur de la Ciudad. Este Análisis lo lograremos con una prueba de Intervalo de Confianza de medias entre los registros de los contaminantes para 2020 y 2021. Un modelo de Anova Multifactorial para comparar si hay diferencia entre las medias de las estaciones y año para la media las mediciones de los contaminantes. Finalmente presentaremos un modelo de Análisis de Discriminante para ver qué variables meteorológicas influyen en la clasificación del Índice de Calidad de Aire y Salud. Uno de los cambios más importantes que realizamos fue acotar nuestro análisis a 4 contaminantes O3,CO,NO2, y PM10. Esto se debe a la gran falta de valores faltantes que se tenía en la base de datos de los otros contaminantes.

Análisis Exploratorio de los Contaminantes

Gráfica de Dispersión



En estas gráficas de Dispersión se puede ver como los valores de las concentraciones de los diferentes contaminantes en el año 2020 y 2021 y en las diferentes estaciones. Las conclusiones principales para este analisis preliminar es que para el contaminante O3 la dispersión de los datos fue mayor en el año 2020 para las estaciones Centro y Norte. Para el contaminante CO hubo mayor concentración en 2021 para la estación centro, y hubo mayor concentración en 2020 para la estación Sur.

Análisis de Intervalos de Confianza de Medias por Contaminante

En el siguiente Análisis haremos un Análisis de Intervalos de Confianza de Medias por contaminante. Este análisis se hará comparando las medias de la población de las muestras del año 2020 y 2021 y comparando las medias por estación de sus muestras en 2020 y 2021. La metodología del Análisis es de la siguiente manera.Se calculará un Limite Superior y un Límite inferior y si dentro de este límite se encuentra el cero podremos concluir que no hay diferencias significativas entre la media de los dos poblaciones. En el caso de que el cero no se encuentre dentro del intervalo podremos concluir que si hay diferencias significativas entre las poblaciones. Si el Límite superior es menor a cero la media de la población 2 es mayor a la media de la población 1 y si el Límite superior es mayor a 0 la media de la población 1 es mayor a la media de la población 2.

$$X_1: \textit{Muestras tomadas en } 2020 \quad X_2: \textit{Muestras tomadas en } 2021$$

$$\textit{Pruebas de Hipótesis}$$

$$H_0: \mu_1 - \mu_2 \neq 0 \quad H_1: \mu_1 - \mu_2 \neq 0$$

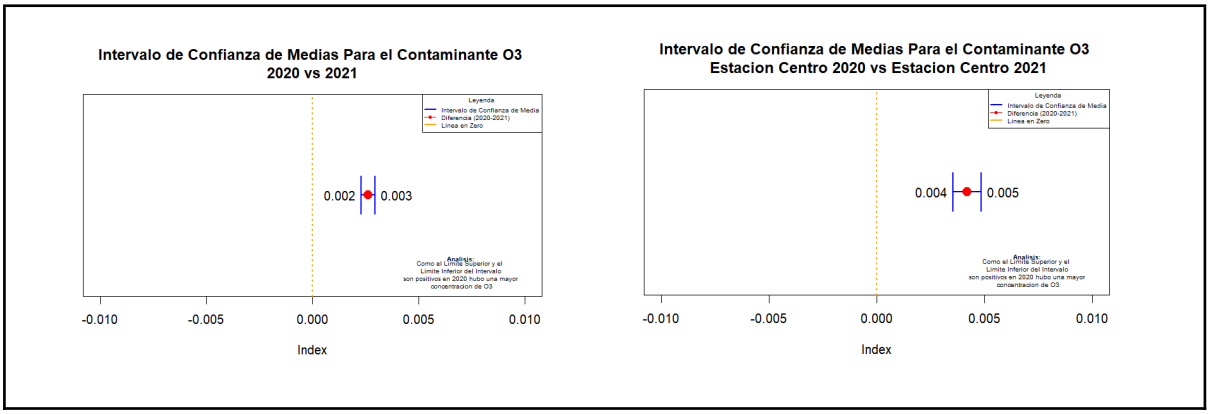
$$\alpha = 0.05 \;\; z = 1.96 \textit{ o el valor z para } (\alpha/2) \textit{ de una dist normal}$$

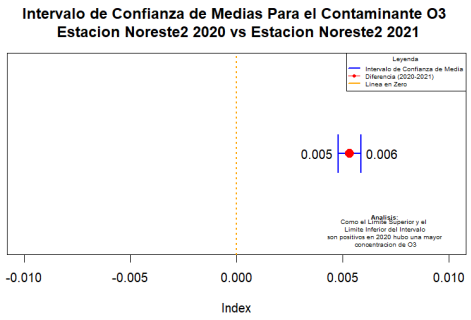
$$S_n = \textit{Desviación Estándar de la Muestra} \;\; n_n: \textit{Tamaño muestra}$$

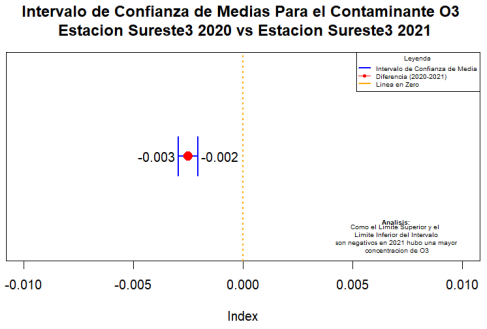
$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$Linf = (\mu_1 - \mu_2) - (z * SE) \; Lsup = (\mu_1 - \mu_2) + (z * SE)$$

O3:



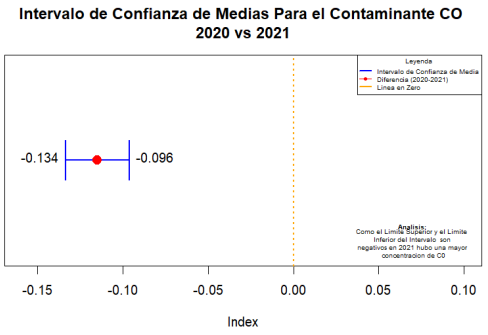


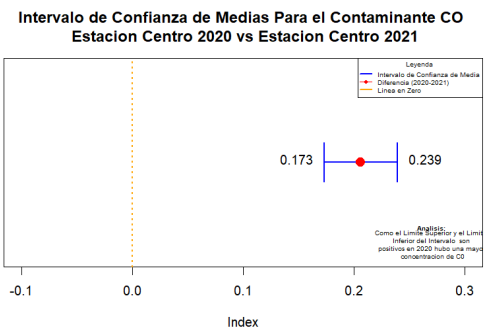


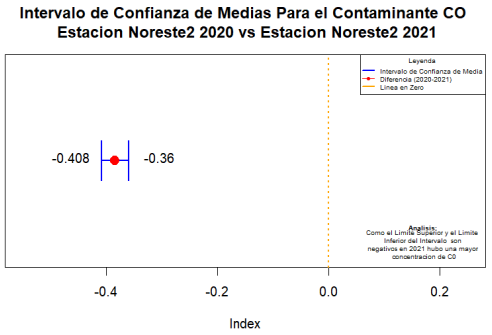
Podemos observar que todas nuestras medias de O3 demuestran que hubo una mayor concentración en 2020 que en 2021, con excepción de la Estación Sureste, la cual presenta una mayor media en 2021

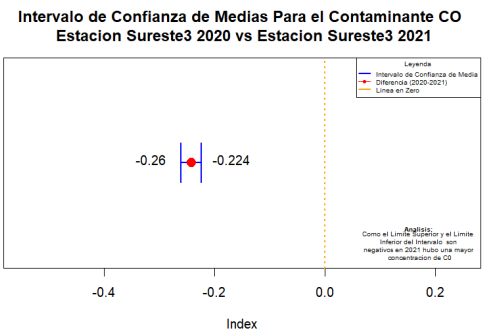
Durante la Pandemia de Covid 19 se recomendó que todos los habitantes desinfectan todos los artefactos y superficies con desinfectantes en aerosol. El ozono se "se forma a partir de reacciones fotoquímicas complejas con intensa luz solar entre contaminantes primarios como son los óxidos de nitrógeno (NO, NO2) y compuestos orgánicos volátiles (COV) " (*Ozono Troposférico • Ecologistas En Acción*, 2013). Se sabe que la mayoría de desinfectantes utilizados para matar los virus y bacterias del covid 19 están hechos en su gran mayoría por compuestos orgánicos volátiles como el etanol, éter y acetona. Por lo que el gran incremento en su uso por toda la población debido a las instrucciones sanitarias para combatir el Covid 19 significó una mayor concentración de O3 durante el año 2020 a comparación de 2021.

CO:



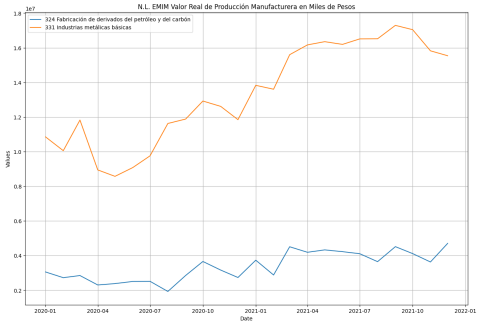




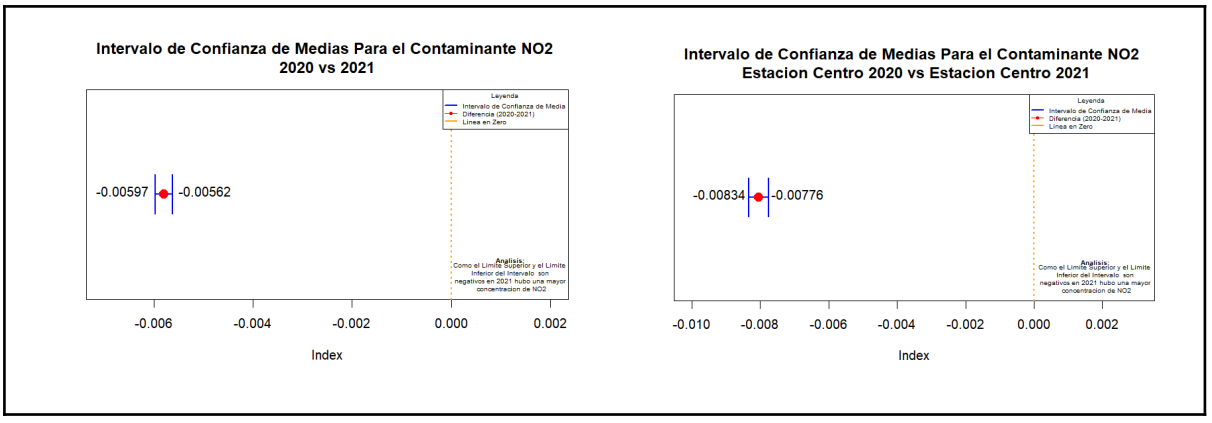


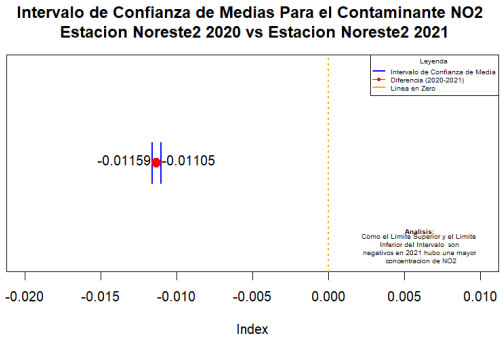
Nuestras gráficas muestran una mayor media de CO en 2020 para la Estación Centro mientras que el resto presenta una mayor media en 2021.

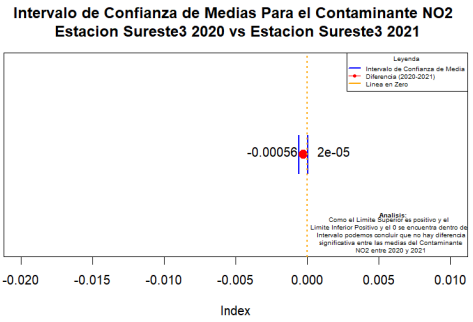
En este análisis podemos ver como si hubo una diferencia significativa para las medias de las concentraciones de CO entre 2020 y 2021. Siendo el 2021 un año con mayor concentración de CO registradas. El CO se produce en el momento de la combustión de combustibles fósiles para la creación de energía. Según la revista WRI México la actividad que más aporta a la creación de CO es el tráfico vehicular. (*Cuatro Gráficos Que Explican Las Emisiones De Gases De Efecto Invernadero Por País Y Por Sector*, 2021). Durante el año 2021 hubo un incremento drástico en el tráfico vehicular en Monterrey ya que la mayoría de personas empezaban a retomar su rutina y utilizar más los medios de transporte por lo que incrementó la creación de CO en la atmósfera. También la revista PRTR comenta que las industrias que más CO generan son las Industrias Petroleras y las Industrias Metálicas (*CO (Monóxido De Carbono)*, n.d.). Según la base de datos del gobierno de Nuevo León en su base de datos N.L. EMIM Valor Real de Producción Manufacturera en Miles de Pesos (*DATA NUEVO LEÓN | N.L. EMIM Valor Real De Producción Manufacturera*, n.d.). En la siguiente gráfica se puede ver el incremento en Producción de ambas industrias en 2021 en comparación a 2020 por lo que tuvo un impacto en su mayor concentración.



NO2:

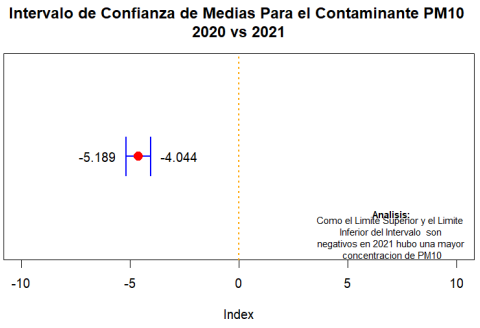


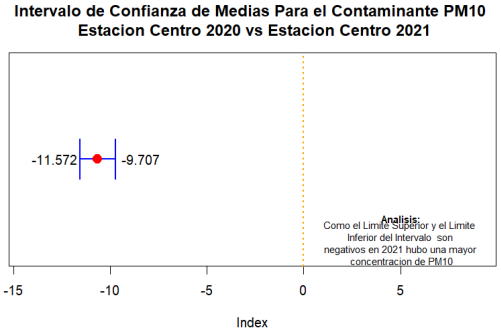


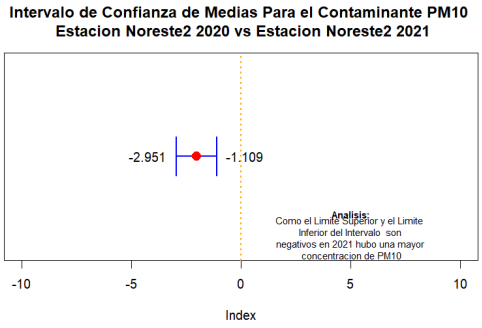


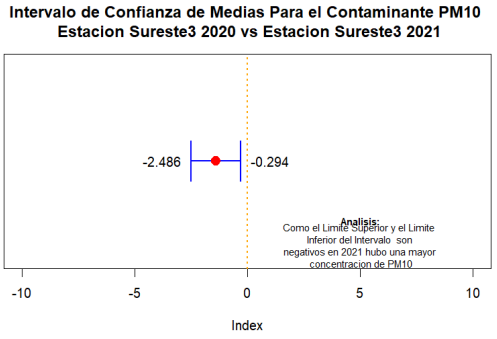
La mayoría de nuestras gráficas indican que en 2021 las medias de NO2 fueron considerablemente mayores, sin embargo en la Estación Sureste las medias de ambos años son bastante similares.

PM10:



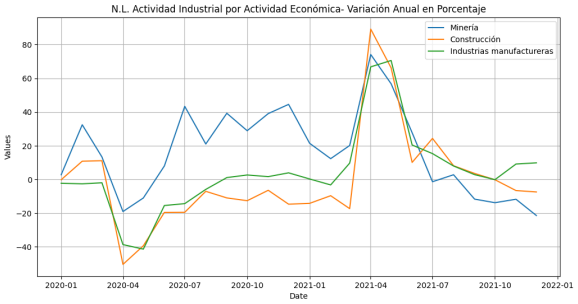






Las medias de PM10 fueron mayores en 2021 en todos los casos

En el análisis de medias el PM10 es el que más diferencia hubo en las concentraciones. En el 2021 hubo una gran diferencia en comparación al 2020. Las partículas de PM10 se generan en gran medida en la industria por partículas generadas por la construcción y por la minería. Según la Base de datos del gobierno de Nuevo León titulada N.L. Actividad Industrial por Actividad Económica- Variación Anual se (*DATA NUEVO LEÓN | N.L. Actividad Industrial Por Actividad Económica- Variación Anual*, n.d.) se podrá ver la siguiente gráfica. Donde se aprecia un incremento porcentual drástico en la industria de Construcción y Minería en 2021 a comparación de 2021 lo cual genera una concentración de PM10 más alta en 2021.



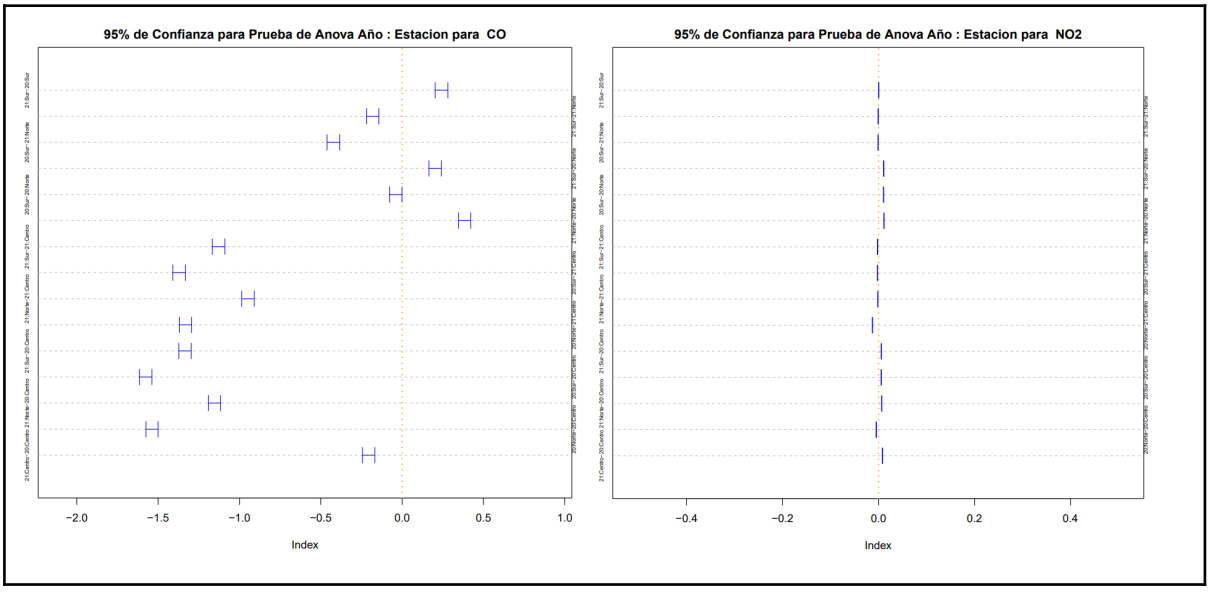
Análisis de Anova Multifactorial

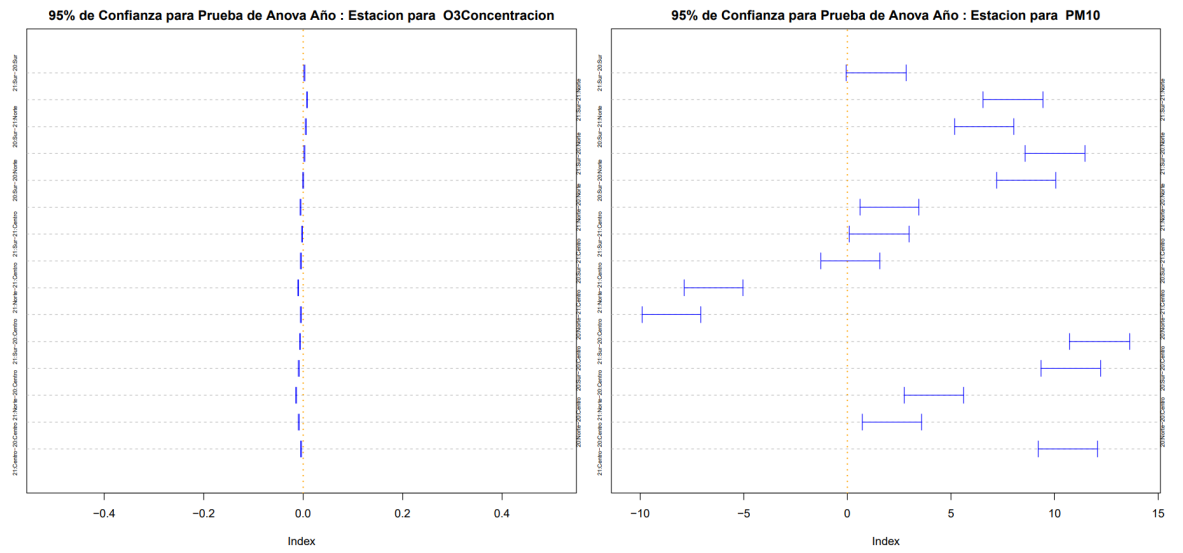
Para nuestro siguiente análisis buscamos un Análisis de Anova Multifactorial. Este con el propósito de ver si la media de los contaminantes es igual entre los factores de año(2020,2021) y estación (Centro, Noreste 2, y Sureste 3). La interpretación de las gráficas se debe hacer de la siguiente manera. Si un límite de un intervalo se encuentra debajo del cero no hay suficiente información para rechazar H0 por lo que se puede concluir que no todas las medias son iguales y existe una diferencia significativa entre las medias de las muestras.

$$H_0: \textit{Todas las medias de las muestras son iguales}$$

$$H_1: \textit{No Todas las medias de las muestras son iguales}$$

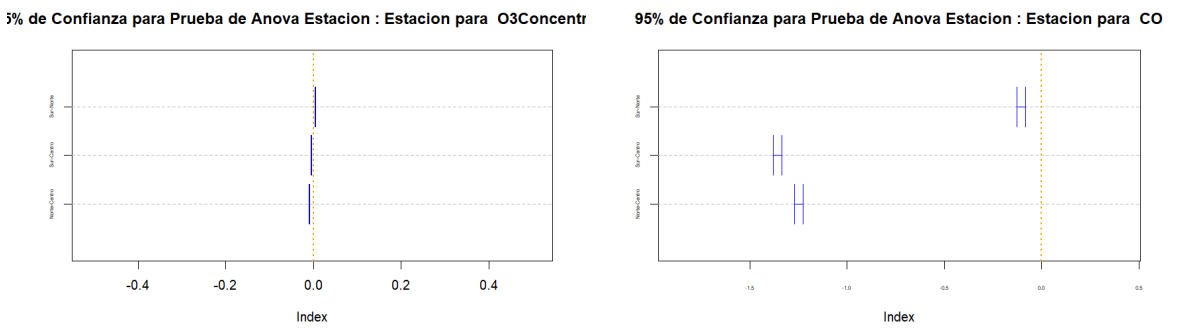
En la siguiente tabla se puede observar el análisis de Anova entre Estación y Año.
Se busca una combinatoria para comparar las medias por ejemplo el contaminante CO entre la estación Centro en el año 2020 y la estación Sureste 3 en el año 2021.

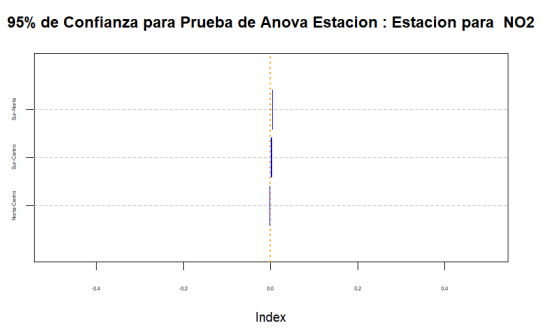


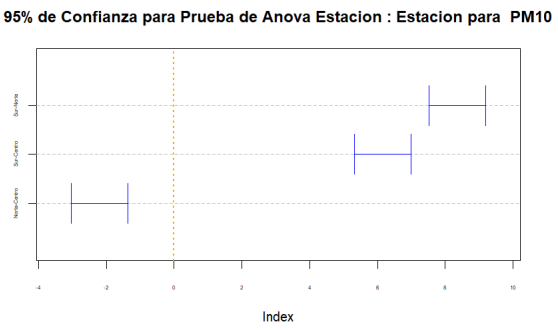


Al analizar los resultados de la prueba de anova multifactorial entre año y estación podemos concluir que para todos los contaminantes se debe rechazar la hipótesis nula y concluir que hay suficiente información para concluir que no todas las medias son iguales. Cabe recalcar que para los contaminantes O3 y NO2 los intervalos son casi iguales y todos se encuentran muy cerca de cero. Esto queda respaldado en el Análisis de Intervalos de Confianza de Medidas para estos contaminantes donde si existió una diferencia significativa y se concluyó que las medias eran diferentes. Pero su escala era muy pequeña y sus intervalos se encontraban muy cerca de cero. Para los contaminantes PM10 y CO sus intervalos eran más grandes por lo que había una mayor variación entre las medidas de las poblaciones, también hubo una mayor diferencia entre las poblaciones dado la ubicación de los intervalos. Los intervalos que se encontraban por debajo del cero tienden a hacer combinaciones con años diferentes y estaciones diferentes. Por lo que podemos concluir que para estos contaminantes si hubo variación entre el año y la zona de la ciudad.

En el siguiente anova se buscará comparar las medias para los cuatro contaminantes entre las diferentes estaciones.







Al observar los resultados de la prueba anova comparando las medias de los contaminantes en las distintas estaciones podemos rechazar la hipótesis nula en todos los casos y podemos decir que hay suficiente evidencia para concluir que no todas las medias son iguales. También en estas gráficas podemos notar que hay algunos contaminantes (O3 y NO2) cuyos intervalos son similares y muy cercanos a cero.

Modelo análisis discriminante:

Para este Modelo de Análisis discriminante generamos una columna categórica de la Clasificación del Índice de Calidad de Aire y Salud para las entradas donde se pudo generar una clasificación a los contaminantes según la normativa. Para este modelo buscamos identificar la asociación entre las variables meteorológicas y su clasificación en el Índice de la calidad de Aire y Salud.

	PRS	RH	SR	TOUT	WSR	WDR
Function 1	-0.001720	0.020090	-0.375485	-0.006393	0.023135	-0.000712
Function 2	0.001835	-0.010446	0.251573	0.020031	-0.011342	-0.000068
Function 3	-0.001914	-0.025640	0.517943	-0.011635	-0.034783	0.001437
Function 4	0.034369	-0.040042	-0.989694	-0.069750	-0.007073	0.003582
Function 5	0.065038	-0.064756	-0.681316	-0.150686	0.077438	0.006033)

El modelo LDA obtuvo una precisión de aproximadamente 50.85% en el conjunto de prueba. Aunque no es una precisión excelente, el objetivo principal aquí es identificar qué variables influyen más en la clasificación del Índice de Calidad de Aire y Salud.

Para determinar las variables más influyentes, podemos examinar los coeficientes de las funciones discriminantes. Cada fila en la tabla de coeficientes representa una función discriminante, y cada columna representa una variable. Los valores más grandes (ya sean positivos o negativos) indican una mayor influencia de esa variable en la clasificación.

Al observar los coeficientes notamos los siguientes resultados:

Estos resultados sugieren que la variable que más influye en la clasificación del Índice de Calidad de Aire y Salud es SR, seguida por TOUT, WSR y RH.

Conclusiones

En conclusión durante este Análisis hemos logrado las siguientes conclusiones. Primero que hubo una diferencia significativa en la Concentración de los cuatro contaminantes en los diferentes años. El único

contaminante donde hubo una mayor concentración en el año 2020 fue O3. Esto pudo ser causado por el incremento en el uso de Compuestos Orgánicos Volátiles por toda la población para mitigar la pandemia del Covid 19 como fue instruido por los diferentes Organismos de Salud. Es un ejemplo de cómo el uso por toda la población de ciertos químicos a gran escala puede impactar en la calidad del aire. Para el caso de CO descubrimos que en 2021 hubo una mayor concentración de ese contaminante. Descubrimos a través de la base de datos del Gobierno de Nuevo León donde se registra el valor de la producción manufacturera se vio un incremento drástico en la producción de Metales y de derivados del petróleo, junto con el incremento en el uso de transportes automovilísticos por toda la población. Todo esto aportó a una mayor concentración de CO en 2021. Finalmente para el contaminante PM10 fue donde se vio una mayor diferencia entre las concentraciones. En el 2021 hubo una mayor concentración por gran diferencia. Este contaminante es generado en su mayoría por la industria de construcción y la minería. También con el uso de la base de datos del gobierno de Nuevo León titulada Actividad Industrial por Actividad Económica vimos un incremento drástico en ambas de estas industrias. Lo cual generó un incremento en la concentración de PM10 en 2021.

Otra conclusión que llegamos durante este estudio y utilizando un análisis de anova multifactorial. Es que no tuvimos suficiente información para concluir que las medias de concentración entre 2020, 2021 y entre las diferentes estaciones fue igual. Por lo que podemos concluir que existió una diferencia significativa entre las muestras. Finalmente implementamos un Modelo de Análisis Discriminante para ver qué variables meteorológicas tienen un mayor impacto a la hora de discriminar entre las diferentes clasificaciones según el índice de calidad de aire y Salud. Dado que la base de datos está muy sesgada solo logramos predecir el 50% de las observaciones correctamente.

Liga a Carpeta con Códigos Implementados durante esta Etapa

https://drive.google.com/drive/folders/1WlKljZvw5Mqf_GbQz1AM7YhXsSEzHa0q?usp=sharing

Referencias

CO (Monóxido de carbono). (n.d.). PRTR España. Retrieved October 12, 2023, from https://prtr-es.es/CO-Monoxido-de-carbono,15589,11,2007.html

Cuatro gráficos que explican las emisiones de gases de efecto invernadero por país y por sector. (2021, September 2). WRI Mexico. Retrieved October 12, 2023, from https://wrimexico.org/bloga/cuatro-gr%C3%A1ficos-que-explican-las-emisiones-de-gases-de-efecto-invernadero-por-pa%C3%ADs-y-por

DATA NUEVO LEÓN | N.L. Actividad Industrial por Actividad Económica- Variación Anual. (n.d.). DATA NUEVO LEÓN. Retrieved October 12, 2023, from http://datos.nl.gob.mx/n-l-actividad-industrial-por-actividad-economica-variacion-anual/

DATA NUEVO LEÓN | N.L. EMIM Valor Real de Producción Manufacturera. (n.d.). DATA NUEVO LEÓN. Retrieved October 12, 2023, from http://datos.nl.gob.mx/1287-2/

Ozono troposférico • Ecologistas en Acción. (2013, December 1). Ecologistas en Acción. Retrieved October 12, 2023, from https://www.ecologistasenaccion.org/27108/ozono-troposferico/