UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Facultad de Ciencias - Física
Temas Selectos de Astrofísica I

Concentración del NO_2 en la CDMX en el año 2017.

Saksevul Arias & Alan Gutierrez

Resumen

En este trabajo realizamos un análisis estadístico, hecho mediante el programa R, de las concentraciones de NO_2 en 5 zonas de la Ciudad de México: noreste, noroeste, centro, suroeste y sureste. La base de datos utilizada es la proporcionada por la SEDEMA en la sección de Índice de la Calidad del Aire.[1] Los resultados más relevantes fueron que no se cumplió la normalidad entre las variables y que para poder hacer una mejor modelo es necesario considerar mas variables en el problema. Además de que en nuestro análisis las concentraciones de NO_2 en el centro son menores a la de los alrededores.

Introducción

Bióxido de Nitrógeno

Los óxidos de nitrógeno son gases formados por átomos de oxígeno y de nitrógeno. De todos estos óxidos que se encuentran en la atmósfera, únicamente el bióxido de nitrógeno (NO_2) resulta ser nocivo para la salud. Esta es una de las principales razones para conocer y , a ser posible, controlar la concentración de NO_2 en el ambiente.

Producción: En general, este contaminante se libera a la atmósfera a través de combustión a altas temperaturas. También se libera en la producción de energía eléctrica. Dado que la mayoría de los vehículos son propulsados por motores de combustión, en la Cuidad de México estos representan la principal fuente de NO_2.

Efectos en el ambiente: En la atmósfera los óxidos de nitrógeno reaccionan con otros compuestos para generar el ácido nítrico (HNO_3). Esto desemboca en la generación de lluvia ácida al combinarse con el agua contenida en las nubes.

Efectos en la salud: La exposición a altas concentraciones puede ocasionar daño en la membrana celular del tejido pulmonar y a bajas concentraciones puede ocasionar irritación en las vías respiratorias o agravar los síntomas de enfermedades respiratorias como bronquitis y pulmonía. La Norma Oficial Mexicana (NOM-023-SSA1-1993) establece un límite para el dióxido de nitrógeno (NO2) de 210 partes por billón (ppb) para el promedio de una hora, el cual no debe excederse más de una vez al año.[2]

Problema a estudiar

Es bien sabido que el centro de la cuidad de México es una de las zonas que mas fuentes de empleo genera, además de que es también una de las zonas con un sin fin de atracciones turísticas, sin embargo, la gente que vive en zonas aledañas tiene que trasladarse de su hogar hasta dicha zona, donde generalmente lo hacen mediante vehículos de combustión ya sea por transporte público o privado. Es por eso que nos interesa estudiar si los contaminantes que se producen en las diferentes zonas de la cuidad, en particular las emisiones de NO_2 que podrían deberse a los carros guardan alguna relación con los producidos en el centro de la ciudad.

Obtención de datos

Los datos fueron obtenidos del índice de la calidad del aire en la Ciudad de México correspondiente al año 2017.[1] Posteriormente se procedió a eliminar manualmente la primer columna de los datos, así como la información preliminar de los datos. Lo anterior se hizo para no tener problemas con R. Como adicional, se adjunta la tabla modificada de datos: *indice_*2017*.csv*.

Variables a usar

En este trabajo nos concentraremos únicamente en los datos, contenidos en $indice_2017.csv$, referentes al bióxido de nitrógeno. Por lo tanto, tendremos la variable cualitativa, de escala nominal, *zona* con las categorías: noroeste, noreste, centro, suroeste y sureste. Y la variable cuantitativa, con escala de razón, *concentración*,

medida en partes por billón (ppb). De esta manera trabajaremos con 5 *grupos*: concentración de NO_2 en el noroeste, concentración de NO_2 en el noreste, concentración de NO_2 en el centro, concentración de NO_2 en el suroeste y concentración de NO_2 en el sureste.

Manejo de datos

Análisis exploratorio

Comenzamos con crear un *data frame* (NO2) de nuestros datos de interés, en este caso, serán 5 columnas de datos correspondientes a cada uno de nuestros grupos. Después eliminamos los datos faltantes.

Ahora, nos interesa saber el comportamiento de nuestros grupos, por lo cual procedemos a hacer análisis gráficos como por ejemplo las gráficas de barras o diagramas de caja y brazos. Ver figura 1 y Apéndice A.

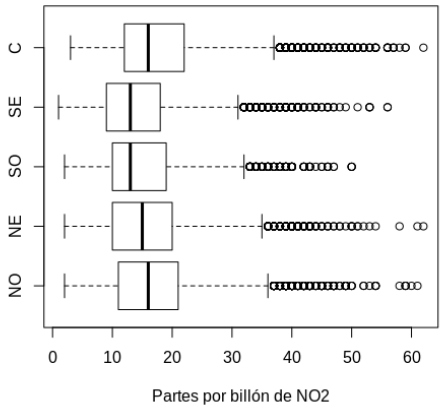

Figura 1: Diagramas de caja y brazo correspondientes a los 5 grupos que tenemos. NO: noroeste, NE: noreste, C: centro SO: suroeste, SE: sureste.

De la figura anterior notamos que no es posible considerar la distribución de nuestras poblaciones como normal. Esto debido a la asimetría de los diagramas y a la gran cantidad de puntos atípicos registrados.

De manera complementaria, es posible hacer un *summary* respecto a nuestro data frame (NO2). Con esto obtenemos, para todos los grupos, los valores mínimos y máximos, valores de los cuantiles y los valores de la mediana y media. Ver tabla 1.

	NO	NE	C	SO	SE
Min.	2.0	2.0	2.0	1.0	3.0
1er Cuan.	11.0	10.0	10.0	9.0	12.0
Mediana	16.0	15.0	13.0	13.0	16.0
Media	16.0	16.0	14.7	13.7	17.7
3er Cuan.	21.0	20.0	19.0	18.0	22.0
Max.	61.0	62.0	50.0	56.0	62.0

Tabla 1: Resumen de nuestros 5 grupos. NO: noroeste, NE: noreste, C: centro SO: suroeste, SE: sureste.

Observemos que tanto las medianas y las medias se encuentran en el mismo orden de magnitud, lo cual sugiere que las distribuciones no son tan distintas a las normales. Además, a primera impresión, puede conllevar a que las concentraciones de NO_2 sean similares en todas las regiones.

Por otro lado podemos calcular la varianza y la desviación estándar de cada uno de los grupos.

	NO	NE	C	SO	SE
Var.	61.6	58.4	63.0	45.2	46.6
Desv. Std.	7.9	7.6	7.9	6.7	6.8

Tabla 2: Varianza y desviación estándar de cada uno de los grupos. NO: noroeste, NE: noreste, C: centro SO: suroeste, SE: sureste.

Notemos que las varianzas no son muy distintas las unas con las otras. Por otro lado, las desviaciones estándar son muy parecidas. Podemos pensar que en las 5 zonas se tiene una distribución similar.

Estimación de Intervalos de Confianza

Recordemos que uno de los supuestos básicos para la inferencia clásica paramétrica es el de normalidad, por lo tanto es necesario verificar si esta se cumple para cada variable. Del análisis exploratorio hecho previamente vimos que este supuesto parecia no cumplirse.
En la figura 2 se puede ver la gráfica QQ (cuantil-cuantil)[1] de nuestros datos originales. Para que se cumpla normalidad es necesario que los puntos caigan en la recta roja, sin embargo, es evidente que esto no es así. Para verificar nuestro supuesto de una forma adecuada realizamos un prueba estadística de normalidad Lilliefors (KS). Para ello usamos la función lilleTest de R en donde cabe resaltar que:

H0=Hay normalidad VS HA≠Hay normalidad
Al realizar dicha prueba tenemos:

$$p - value < 2 \times 10^{16}$$

para cada variable.

[1]Método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación.

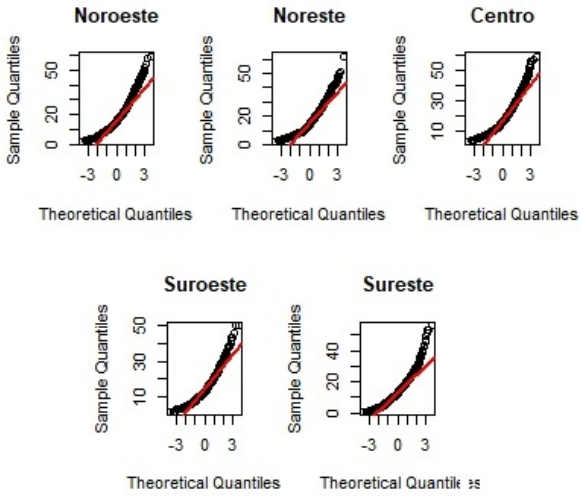

Figura 2: Gráficas QQ del NO_2 para todas las zonas de la ciudad.

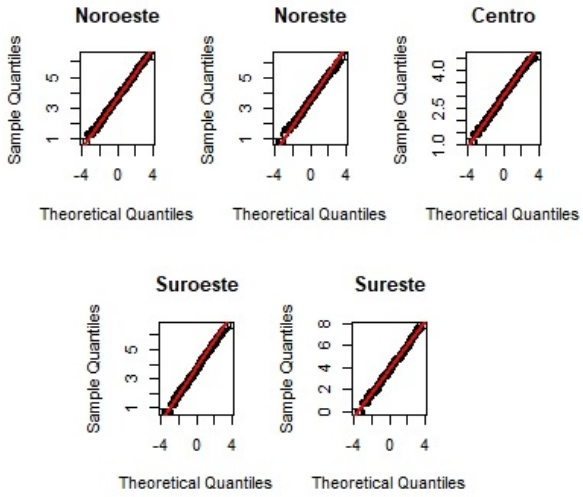

Figura 3: Gráficas QQ del NO_2 de los datos transformados para todas las zonas de la ciudad.

Por lo tanto rechazamos H0, es decir, no hay normalidad en ninguna de nuestras variables.
Con este problema enfrente transformamos nuestras variables mediante la técnica $Box - Cox$[2] con el fin de corregir la no-normalidad de nuestras variables originales, tal como se muestra en la figura 3 para la gráfica QQ de nuestros datos. Vemos que los puntos ahora si caen en la recta roja, pero aún así es necesario realizar la prueba estadística de normalidad. De igual forma que en el caso anterior verificamos por prueba estadística si esta se cumple.

Al realizar la prueba Lilliefors (KS) tenemos que el valor p$<$ 0.05 para todas las variables por lo tanto seguimos sin cumplir normalidad.

Por lo tanto procedemos a verificamos con diferentes pruebas de normalidad como Anderson-Darling, Cramer-von Mises, Lilliefors (Kolmogorov-Smirnov) y Pearson chi-square, sin embargo, en todas ellas seguimos sin cumplir normalidad (puede deberse a que seguimos teniendo datos atípicos en estas variables transfomadas), por lo que nos basaremos en el hipotético caso que si la cumple, esto con el fin didáctico de ver la importancia de la normalidad al estimar los intervalos de confianza. Esto lo podemos hacer debido a que conocemos la media poblacional y la desviación estándar poblacional.

[2]En honor a los estadísticos George E. P. Box y David Cox.

Usamos nuestros datos transformados con promedio poblacional de 3.5291 y desviacion estandar de 0.8934, nos tomamos muestras aleatorias de 3000 elementos para cada variable. Calculamos el intervalo de confianza para la media y desviación conocida. En la tabla 3, vemos que la media poblacional no se encuentra en los intervalos de confianza, donde esperabamos un resultado de esa forma debido a que nuestras variables no cumplen con el criterio de normalidad.

Variable	Intervalo de confianza	¿Contiene a la media POBLACIONAL (3.529)?
NO	(3.618,3.675)	no
NE	(3.505, 3.562)	si
C	(2.947, 3.004)	no
SO	(3.631, 3.688)	no
SE	(3.739, 3.796)	no

Tabla 3: Intervalo de confianza al 92 para una muestra aleatoria de nuestros datos transformados.

Ahora como nos interesa saber si las concentraciones de NO2 en las zonas de alrededor afectan los niveles de concentración en el centro de la ciudad nos tomamos las muestras transfomadas del centro y las muestras transformadas de los alrededores, y procedemos a calcular su intervalo de confianza para las medias, es por eso que solo comparamos la del centro con la de los alrededores. Utilizaremos una confianza del 92 % al calcular el intervalo de las variables comparadas suponiendo muestras independientes (suponemos normalidad y varianzas iguales). En la tabla 4 vemos que todos los intervalos que nos dan son negativos,

Variables comparadas	Intervalo
C y NO	(-0.699, -0.637)
C y NE	(-0.610, -0.547)
C y SO	(-0.727, -0.660)
C y SE	(-0.830, -0.753)

Tabla 4: Comparación de la media muestral en el centro (C) con las del noroeste (NO), noreste (NE), suroeste (SO) y sureste (SE).

Interpretando los resultados anteriores, tenemos que en todas las zonas alrededor del centro, la concentración de NO_2 es mayor. Muy curioso pues esperabamos ver un resultado contrario.

Prueba de hipótesis paramétricas

Hagamos la prueba de hipotesis, ahora deseamos contrastar a un nivel de significancia de 0.08 la hipótesis nula de que la media de las concentraciones del dióxido de nitrógeno transformadas es de 3.529, es decir, tenemos: $H0 : \mu = 3.529 vs Ha : \mu \neq 3.529$

Esto lo hacemos para las diferentes variables como se ve en la tabla 5, donde vemos en todas rechazamos la hipótesis nula de que la media Poblacional es de 3.52 para los datos transformados, excepto para la zona NE, donde si queremos usar a los datos transformados nos dice que deberíamos de hacerlo con estos datos pues con la prueba Ztest, el valor zcalc entra en la zona de aceptación y su valor p es bastante alto.

Variable	Intervalo de aceptacion de H0	Zcalculado	Valor p
NO	(-1.7506,1.7506)	7.258	3.91e-13
NE	(-1.7506,1.7506)	0.283	0.77
C	(-1.7506,1.7506)	-33.9	1.17e-252
SO	(-1.7506,1.7506)	8.022	1.03e-15
SE	(-1.7506,1.7506)	14.644	1.4e-48

Tabla 5: Prueba de hipótesis para cada variable con nivel de confianza del 92 %.

Para dos poblaciones las hipótesis a contrastar son: H0:$\mu_1 >= \mu_2$ vs Ha: $\mu_1 < \mu_2$

Donde $\mu_1 = \mu_C$ y μ_2 son las medias de las zonas alrededor. Como de los intervalos obtuvimos que las concentraciones en el centro son menores esto nos dice que las medias son diferentes, en particular las del centro eran menores a la de los alrededores. Además de que el valor p en todas las pruebas es <2e-16, procedemos con rechazar hipótesis nula, por lo que $\mu_1 < \mu_2$, es decir podemos estar seguros al 92 % que las concentraciones del centro son menores a la de los alrededores.

Hipótesis no paramétricas y análisis categóricos

1. a) Procedemos ahora a comparar las 5 poblaciones, entre si, para obtener información sobre que tan parecidas son las medianas de cada una de las poblaciones. Con esto sabremos si es posible considerar que son iguales o no.

En este caso, nuestra hipótesis nula es H_0: mediana$_1$ = mediana$_2$. Por lo que la hipótesis alternativa resulta H_a: mediana$_1$ $\neq$ mediana$_2$.

Nuestra herramienta será la prueba *wilcox.test* tomando todas las posibles combinaciones de parejas entre las poblaciones sin transformar. Además usaremos un nivel de significancia de 0.08, o lo que es igual, un nivel de confianza de 0.92. En la tabla 6 se muestran los valores obtenidos para el $valor - p$ en cada caso.

	NO	NE	C	SO	SE
NO	1	$< 10^{-12}$	$< 10^{-10}$	$< 10^{-15}$	$< 10^{-15}$
NE	$< 10^{-12}$	1	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$
C	$< 10^{-10}$	$< 10^{-15}$	1	$< 10^{-15}$	$< 10^{-15}$
SO	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	1	$< 10^{-15}$
SE	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	1

Tabla 6: Valor-p respecto a cada pareja del grupo. NO: noroeste, NE: noreste, C: centro SO: suroeste, SE: sureste.

Como podemos ver, en todos los casos relevantes, el valor-p es prácticamente 0. Por lo tanto, rechazamos hipótesis nula H_0. Por lo tanto, concluimos que la concentración de NO_2 son diferentes en todas las zonas. Este resultado es igual al encontrado mediante el enfoque paramétrico.

2. a) Ahora toca construir una tabla de categoría utilizando dos variables. Consideraremos las variables noroeste (NO) y noreste (NE) y usaremos las siguientes tablas de contingencia para construir la tabla categórica de dos variables. $cont < -table(NO2.muestreado\$NO)$; $cont1 < -table(NO2.muestreado\$NE); cont; cont1$ Con lo que obtenemos todos los valores que pueden tomar y la cantidad de veces que lo hacen. Así, construimos, a mano, la siguiente tabla de contingencia mediante un arreglo matricial.

	2	3	4	5	6	7	8	9	...
NO	1	8	27	50	53	90	106	126	...
NE	3	4	19	64	97	122	170	131	...

Tabla 7: Matriz A que representa la tabla de contingencia considerando las variables noroeste (NO) y noreste (NE).

En total se tienen 53 entradas por cada variable. A esta matriz, que llamaremos A, le aplicaremos la prueba de chisq y la de fisher. Al hacerlo, obtenemos un error en la salida de R. Consideramos que es debido a que el arreglo es demasiado grande así que tomamos 9 datos para cada una de las variables (3 del inici, 3 del centro y 3 finales), obteniendo la tabla siguiente, más simplificada:

NO	1	8	27	161	147	179	0	0	1
NE	3	4	19	129	183	157	3	1	0

Tabla 8: Matriz A' que representa la tabla de contingencia reducida considerando las variables noroeste (NO) y noreste (NE).

Esta última tabla está representada por una matriz A' la cual tiene únicamente 9 datos por cada variable.

b) Nuestra hipótesis nula (H_0) es: independencia. La hipótesis alternativa (H_a) es: no independencia. Finalmente aplicamos el test de fisher con una confianza del 0.98 (o significancia=0.02) y el test de chisq. De esta manera obtenemos un valor-p = 0.03. Por lo tanto rechazamos H_0 y concluimos que no existe independencia.

Análisis de Regresión

En esta parte realizaremos el análisis de regresión para nuestras variables, usamos como variable respuesta a la concentración en la zona centro y como covariables a las concentraciones al rededor.

En la figura 4 se nota cierta tendencia entre las variables, a nosotros solo nos interesa poder realizar un modelo para la última fila de la figura.

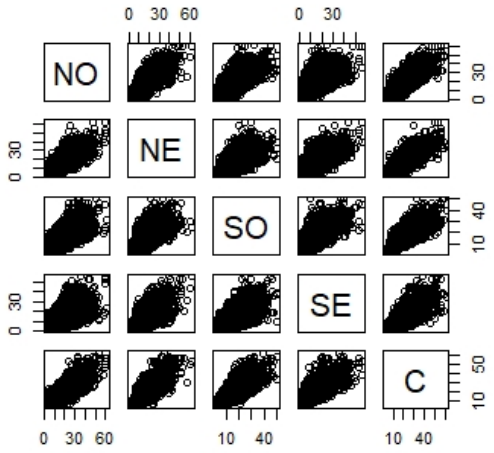

Figura 4: Gráfica de todas las variables vs todas.

Es decir queremos modelar la ecuación
$C = a_0 + a_1NO + a_2NE + a_3SO + a_4SE$
Con C los datos en la zona centro, NO los de la zona Noroeste, así respectivamente.
Al determinar los coeficientes en el programa r obtenemos:

$$C = 0.313 + 0.327NO + 0.259NE + 0.322SO + 0.213SE$$

Sin embargo es necesario hacer una prueba general de los supuestos, tal y como se ve en en la tabla 9.

Prueba	Resultado
Global Stat	No se satisface!
Skewness	No se satisface!
Kurtosis	No se satisface!
Link Function	No se satisface!
Heteroscedasticity	No se satisface!

Tabla 9: Prueba de los supuestos para el modelo con todas las covariables.

Esa tabla nos dice que el modelo no es muy bueno. Al revisar el supuesto de independencia de residuos el valor p indica que NO hay autocorrelación, es decir, se cumple la independencia de residuos. Obtenemos que por lo menos nuestros datos satisfacen el supuesto más importante.

Ahora para estimar un mejor modelo y dado que transformando las variables veiamos que no se satisfacía normalidad, para mejorar el modelo lo que hacemos es eliminar regresores.

Tomamos el de norte con centro ya que al realizar la matriz de correlación resulta haber más entre estas variables con correlación de 0.853. Entonces:
$C = b + mNO$
Haciendo el ajuste resulta que:

$$C = 3.11(0.106) + 0.862(0.005)NO$$

lo que esta entre parentesis es el error en ambos coeficientes tenemos un valor de t de 2e<-16, por lo tanto nos dice que ambos coefientes son distintos de cero, tenemos que la rcuadrada vale 0.727.

Veamos que pasa si quitamos el intercepto. Entonces la ecuación resulta:
$y = ax$
Obteniendo
$C = 1.013(0.002)NO$
vemos que el error en la pendiente disminuye (Figura 5), tenemos valor estadístico t de 2e<-16, por lo tanto nos dice que es correcto tomar el coeficiente distinto de cero. Obtenemos que el valor de rcuadrada aumenta a 0.949, por lo que este parece ser un mejor modelo.

Para ambos modelos verificamos independencia en residuos (recordando que H0: autocorrelación es cero), donde resulta que $p = 2.26e - 16$ por lo tanto rechazamos hipótesis nula, es decir, si hay dependencia en los residuos. Esto en ambos modelos, por lo tanto, usamos el criterio de la rcuadrada para escoger el modelo que

no tiene al intercepto.

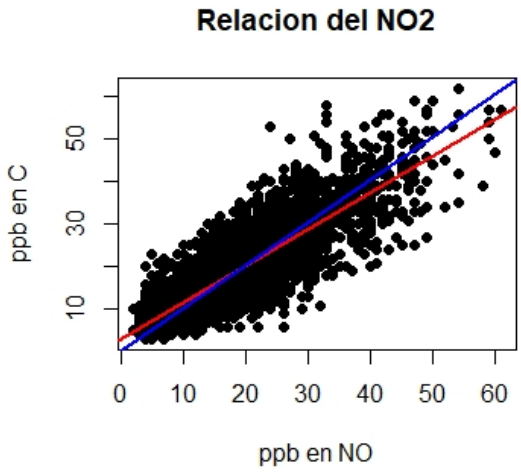

Figura 5: Gráfica de las concentraciones de C vs NO, en rojo el modelo con intercepto y en azul la gráfica con modelo sin intercepto

Al realziar la estimacion de los bandas de confianza y las bandas de predicción del modelo sin intercepto, resulta que las bandas de confianza son muy pequeñas, esto nos dice que es la unica zona en la que estamos seguros que nuestro modelo funciona, como vemos es muy pequeña esta zona.
Al realizar la estimación de las bandas de predicción estas nos dicen la zona en la que podriamos tener nuevos posibles valores, donde vemos que parecen predecir bien en donde esperaríamos tener posibles valores futuros figura 6, esto hasta el valor 150. Para ello hacemos una prueba anova para saber si el modelo es correcto, donde esta prueba nos dice que si lo es.

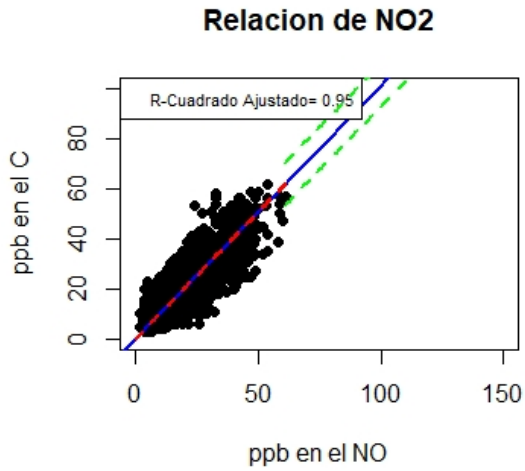

Figura 6: Gráfica de las concentraciones de C vs NO, en rojo las bandas de confianza y en verde la banda de predicción.

Conclusiones

Vemos que al no tener normalidad en nuestros datos, los resultados que nos arrojan difieren de lo que nosotros esperabamos, tal como que las concentraciones en el centro resultan menores que las de al rededor. Además observamos que la concentración de NO2 en el noroeste y en el noreste son dependientes entre si para el enfoque no paramétrico. Para el caso de la regresión obtenemos que más variables parecen explicar mejor al modelo pues al menos se pudo satisfacer el supuesto de independencia en los residuos y los otros supuestos se pueden arreglar transformando las variables a normales, que en nuestro caso veiamos que al transformarlas seguian sin ser normales, es por eso que se debe de utilizar alguna otra técnica, que almenos para el caso de la técnica BOX-COX vemos que no es sufuciente. Al hacer la regresión para una variable, la correlación que se obtiene en el modelo resulta mas alta cuando se elimina el intercepto, las bandas de confianza resultan bajas, pero al menos las bandas de predicción visualmente se nota que es muy probable que tengamos datos dentro de esa banda.
Podemos decir que hay relación entre las emisiones en el centro con las de alrededor, pero hace falta mas variables para poder tener un mejor modelo, como tomar en cuenta la hora y la dirección en la que sopla el viento.

Referencias

[1] `http://www.aire.cdmx.gob.mx/default.php?opc=%27aKBhnmI=%27&opcion=aw==`, consulta; junio, 2018.

[2] `http://www.aire.cdmx.gob.mx/default.php?opc=%27Y6BhnmKkZA==%27`, consulta: junio 2018.

Apéndice A

Complemento análisis exploratorio

Presentamos los diagramas de barras generados mediante las tablas de frecuencias de nuestros 5 grupos. Esos diagramas representan las mediciones de NO_2 en diferentes zonas de la Ciudad de México y sus respectivas frecuencias de incidencia.

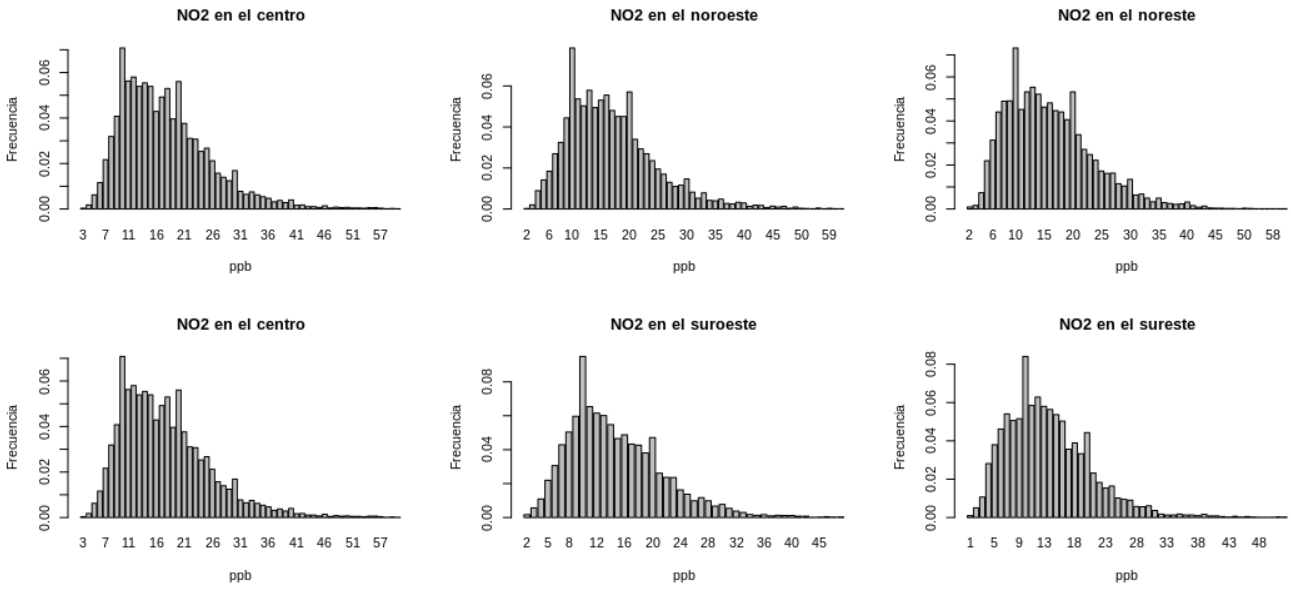

Figura 7: Diagramas de barras correspondientes a la concentración de NO_2 y sus respectivas frecuencias de incidencia en el año 2017. La división se hace en 5 grupos. NO: noroeste, NE: noreste, C: centro SO: suroeste, SE: sureste. Nótese que se repite dos veces la zona centro. La concentración se mide en partes por billón (ppb).

A simple vista podemos apreciar que las concentraciones de NO_2 son similares en la ciudad. Siendo muy atrevidos, podrías pensar que son iguales. Sin embargo es necesario hacer un análisis estadístico para estar seguros.