CORRELACIÓN ESPACIO TEMPORAL DE LOS CONTAMINANTES DE LA CIUDAD DE MÉXICO Y ENFERMEDADES CRÓNICAS NO TRANSMISIBLES

INTRODUCCIÓN

La contaminación en la Ciudad de México es considerada como una gran amenaza para la salud pública. Aunque el informe de calidad del aire en tiempo real se puede utilizar para actualizar nuestro conocimiento sobre la calidad del aire, las preguntas sobre cómo evolucionan los contaminantes a lo largo del tiempo y cómo se correlacionan espacialmente siguen siendo un enigma. En vista de este punto, adoptamos el método KDE o Estimación de Densidad de Kernel (núcleo) para analizar los datos por hora de nueve contaminantes en la Ciudad de México en un intento de averiguar cómo estos contaminantes se correlacionan temporal y espacialmente.

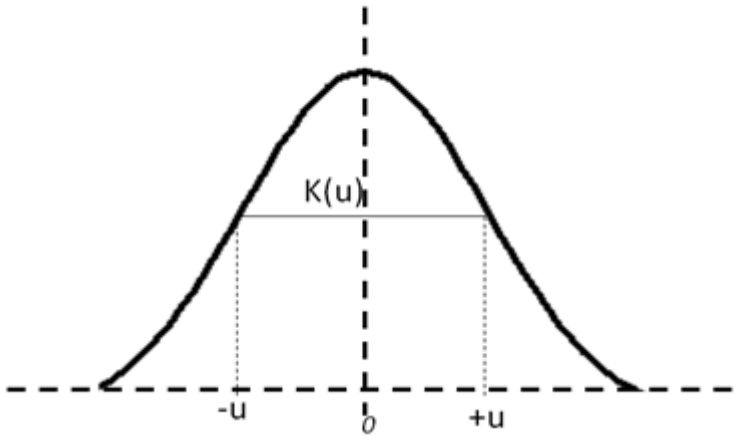
METODOLOGÍA

La metodología utilizada para este estudio fue la utilización de la Estimación de Densidad de Kernel KDE por sus siglas en inglés la cual nos permite interpolar datos y poder estimar información en espacios vacíos de información. A continuación, se explican los conceptos de esta metodología.

KERNEL

Las funciones Kernel se utilizan para estimar la densidad de variables aleatorias y como función de ponderación en la regresión no paramétrica. Esta función también se utiliza en el aprendizaje automático como método kernel para realizar la clasificación y la agrupación.

La primera propiedad de una función kernel es que debe ser simétrica. Esto significa que los valores (u) de la función kernel son los mismos tanto para +u como para –u, como se muestra en la siguiente gráfica. Esto se puede expresar matemáticamente como K (-u) = K (+u). La propiedad simétrica de la función kernel permite que el valor máximo de la función (max(K(u)) se encuentre en medio de la curva.



El área bajo la curva de la función debe ser igual a uno. Matemáticamente, esta propiedad se expresa como:

$$\int_{-\infty}^{+\infty} K(u)du = 1$$

La función de densidad gaussiana se usa como función kernel porque el área bajo la curva de densidad gaussiana es uno y también es simétrica.

El valor de la función kernel, que es la densidad, no puede ser negativo, $K(u) \geq 0$ para todo $-\infty < u < \infty$.

La ecuación para el Kernel gaussiano es:

$$K(x) = \frac{1}{h\sqrt{2\pi}} e^{-0.5\left(\frac{x-xi}{h}\right)^2}$$

Donde xi es el punto de datos observado. x es el valor donde se calcula la función kernel y h se denomina ancho de banda o desviación estándar.

Ejemplo:

Digamos que tenemos las calificaciones obtenidas por seis alumnos en una materia. Se construye el núcleo en cada punto de datos utilizando la función del núcleo gaussiano.

xi = {65, 75, 67, 79, 81, 91}

x1 = 65, x2 = 75 … x6 = 91.

Se requieren tres entradas para desarrollar una curva kernel alrededor de un punto de datos. Estas son:

a) El punto de datos de observación que es xi.

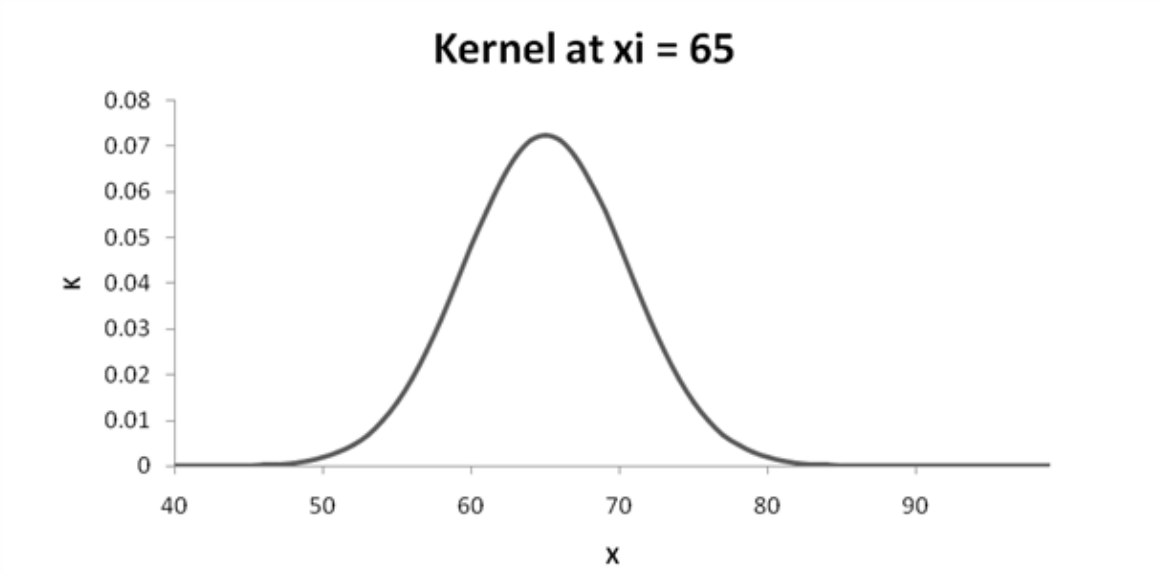
b) El valor de h.

c) Una serie de puntos de datos espaciados linealmente que alberga los puntos de datos observados donde se estiman los valores de K. Xj = {50,51,52…. 99}

El cálculo de los valores de K para todos los valores de Xj para valores dados de xi y h se muestra en la siguiente tabla; donde xi = 65 y h = 5,5.

X_j	x_i	h	$A = \frac{1}{h\sqrt{2\pi}}$	$B = -0.5\left(\frac{X_j - x_i}{h}\right)^2$	$K = Ae^B$
50	65	5.5	0.072536	-3.71901	0.00175958
51	**65**	**5.5**	**0.072536**	**-3.23967**	**0.002841733**
52	65	5.5	0.072536	-2.79339	0.00444018
–	–	–	–	–	–
–	–	–	–	–	–
–	–	–	–	–	–
–	–	–	–	–	–
99	**65**	**5.5**	**0.072536**	**-19.1074**	**0.000000000365**
Sum					1.000

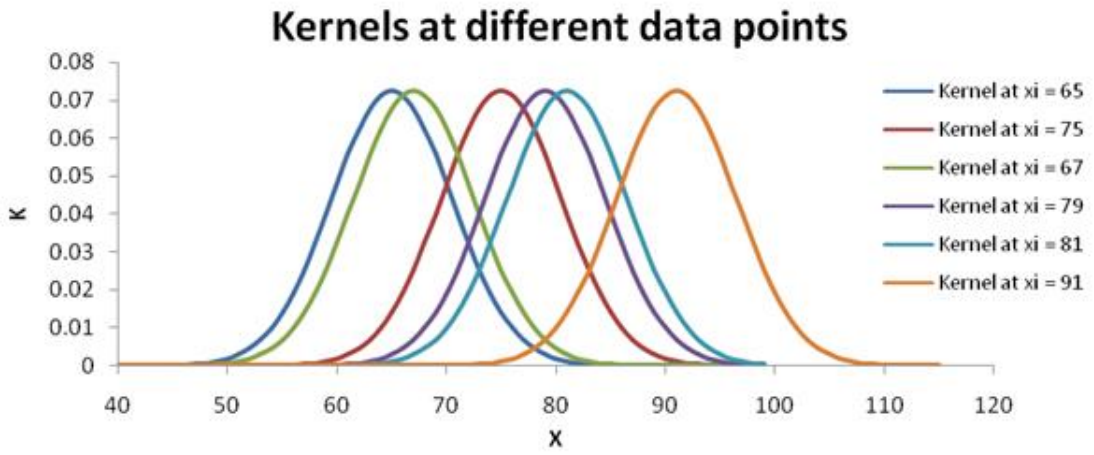
Xj y K se trazan a continuación para visualizar el núcleo (Kernel).



De manera similar, en los seis puntos de datos observados, los valores del kernel se estiman como se muestra en la tabla y se trazan a continuación.

	K(x)					
X	Xi = 65	Xi = 75	Xi=67	Xi=79	Xi=81	Xi=91
50	0.00175958	0.00000237	0.00061093	0.00000007	0.00000001	0.00000000
51	0.00284173	0.00000532	0.00105409	0.00000017	0.00000003	0.00000000
52	0.00444018	0.00001157	0.00175958	0.00000042	0.00000007	0.00000000
53	0.00671214	0.00002433	0.00284173	0.00000102	0.00000017	0.00000000
-	-	-	-	-	-	-
.	.	.	.	.	.	.
-	-	-	-	-	-	-
78	0.00444	0.06251	0.009817	0.071347	0.06251	0.00444
79	0.002842	0.05568	0.006712	0.072536	0.067895	0.006712
80	0.00176	0.047984	0.00444	0.071347	0.071347	0.009817
81	0.001054	0.040007	0.002842	0.067895	0.072536	0.01389
.	.	.	.	.	.	.
99	0.00000	0.0000000	0.00000	0.000000	0.000342567	0.025184586

Se observa en la tabla que el valor de la función kernel es casi 0 para valores Xj que están bastante lejos de xi. Por ejemplo, el valor de la densidad del kernel en Xj = 99 es cero cuando xi = 65.

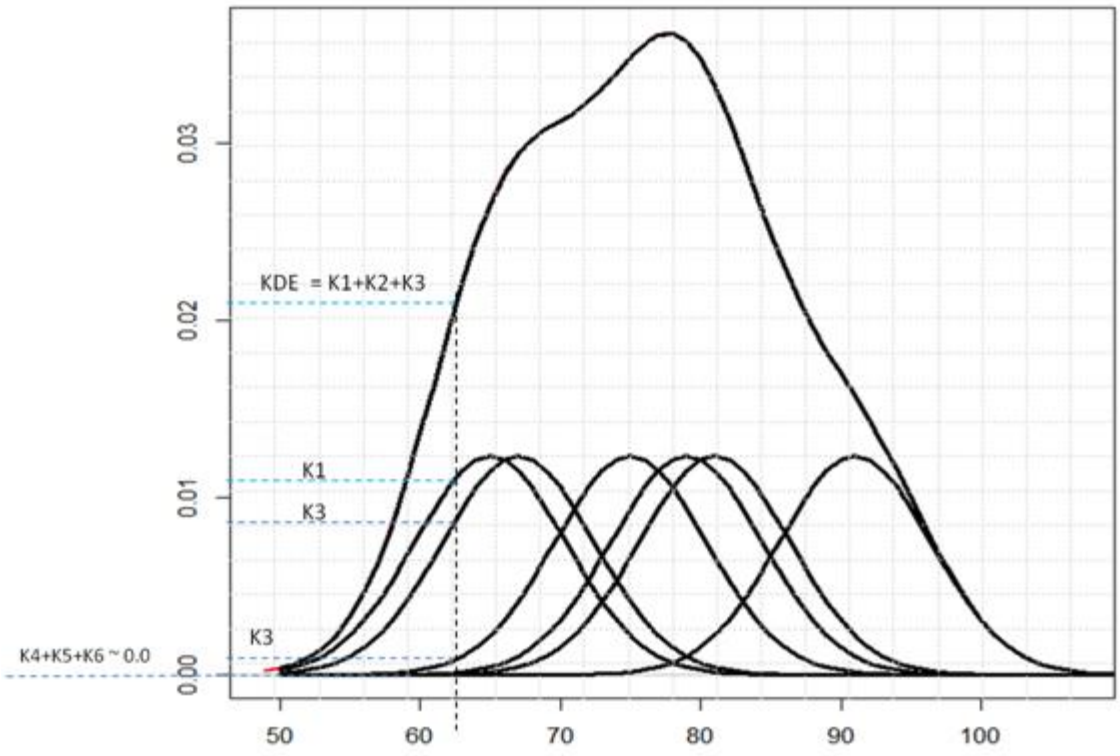


Estimación de la densidad del kernel (KDE)

Hasta ahora se han calculado kernels individuales sobre puntos de datos. Ahora, los valores de densidad compuesta se calculan para todo el conjunto de datos. Se estima simplemente sumando los valores kernel (K) de todos los Xj. Con referencia a la tabla anterior, el KDE para el conjunto de datos completo se obtiene sumando todos los valores de las filas. Luego, la suma se normaliza dividiendo entre el número de puntos de datos, que es seis en este ejemplo. La normalización se realiza para llevar el área bajo la curva KDE a uno. Por lo tanto, la ecuación para calcular KDE para cada Xj se expresa como:

$$KDE_j = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{h\sqrt{2\pi}} e^{-0.5\left(\frac{x_j - x_i}{h}\right)^2}$$

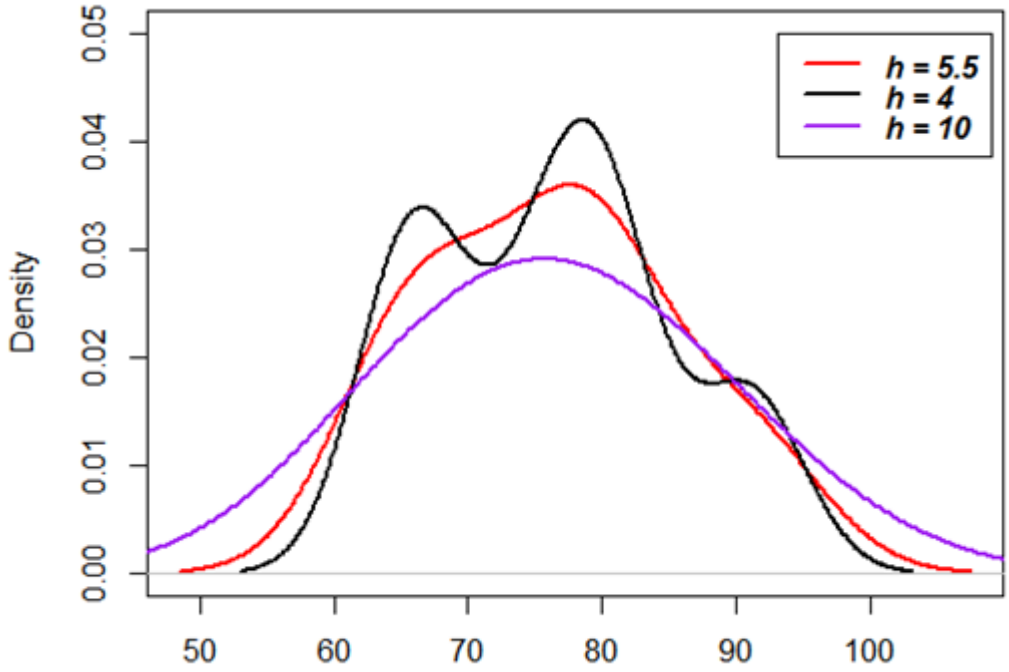
Donde n es el número de puntos de datos. El KDE después de agregar los seis núcleos normalizados se muestra a continuación en la gráfica.



OPTIMIZACIÓN DE ANCHO DE BANDA

El ancho de banda (h) de una función del núcleo juega un papel importante para ajustar los datos de manera adecuada. Un valor bajo de ancho de banda estima la densidad con mucha variación, mientras que un valor alto de h produce un gran sesgo. Por lo tanto, la estimación de un valor óptimo

de h es muy importante para construir la densidad más significativa y precisa. Como se muestra en la siguiente gráfica, tres valores diferentes de anchos de banda producen tres curvas diferentes. El negro ofrece mucha variación en los valores de densidad que no parece realista, mientras que el morado no explica la densidad real al ocultar información.



Hay varios métodos propuestos por los investigadores para optimizar el valor del ancho de banda en la estimación de la densidad del kernel. Uno de ellos es el método de validación cruzada de máxima probabilidad.

Validación cruzada de máxima verosimilitud (MLCV)

Este método fue propuesto por Hobbema, Hermans y Van den Broeck (1971) y por Duin (1976). En este método, la función kernel se estima en un subconjunto de Xj basado en un enfoque de validación cruzada de exclusión. La función objetivo que maximiza MLCV se expresa como:

$$MLCV_{maximize} = \frac{1}{n}\sum_{i=1}^{n} log\left[\sum_{j} K\left(\frac{x_j - x_i}{h}\right)\right] - log[(n-1)h]$$

Primer término.

$$log\left[\sum_{j} K\left(\frac{x_j - x_i}{h}\right)\right]$$

Se calculan los valores de la función kernel como se mencionó anteriormente. Aquí, en una sola observación de x1, los valores de K se calculan para cierta h en el rango de Xj que excluye a x1 de Xj. Se suman los valores de K y finalmente se calcula el logaritmo de la suma.

Ejemplo

Consideremos h = 3 (h debe optimizarse. Se ha tomado h = 3 sólo para explicar cómo se calcula la función de optimización).

xi = {65, 75, 67,79, 75,63,71,83,91, 95}

	K at difference Xi					
Xj	63	65	67	-	-	95
60	**0.2419707245**	**0.0994771388**	**0.0262218891**	-	-	**0.0000000000**
61	0.3194480055	0.1640100747	0.0539909665	-	-	0.0000000000
62	**0.3773832277**	**0.2419707245**	**0.0994771388**	-	-	**0.0000000000**
63	0.0000000000	0.3194480055	0.1640100747	-	-	0.0000000000
64	**0.3773832277**	**0.3773832277**	**0.2419707245**	-	-	**0.0000000000**
65	0.3194480055	0.0000000000	0.3194480055	-	-	0.0000000000
66	**0.2419707245**	**0.3773832277**	**0.3773832277**	-	-	**0.0000000000**
67	0.1640100747	0.3194480055	0.0000000000	-	-	0.0000000000
68	**0.0994771388**	**0.2419707245**	**0.3773832277**	-	-	**0.0000000000**
69	0.0539909665	0.1640100747	0.3194480055	-	-	0.0000000000
70	**0.0262218891**	**0.0994771388**	**0.2419707245**	-	-	**0.0000000000**
71	0.0113959860	0.0539909665	0.1640100747	-	-	0.0000000000
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	
95	-	-	-			0.0000000000
sum	**2.601048**	**2.601057**	**2.601058**	-	-	-
log(sum)	0.9559146	0.955918	0.955918			

Los valores K para Xj = Xi se establecen en cero para asegurarse de que se excluyen al realizar la suma.

Segundo término.

$$log[(n-1)h]$$

En este ejemplo, n = 10 y h = 3. Podemos estimar fácilmente el término como:

$$log[(10-1)*3] = 3.295837$$

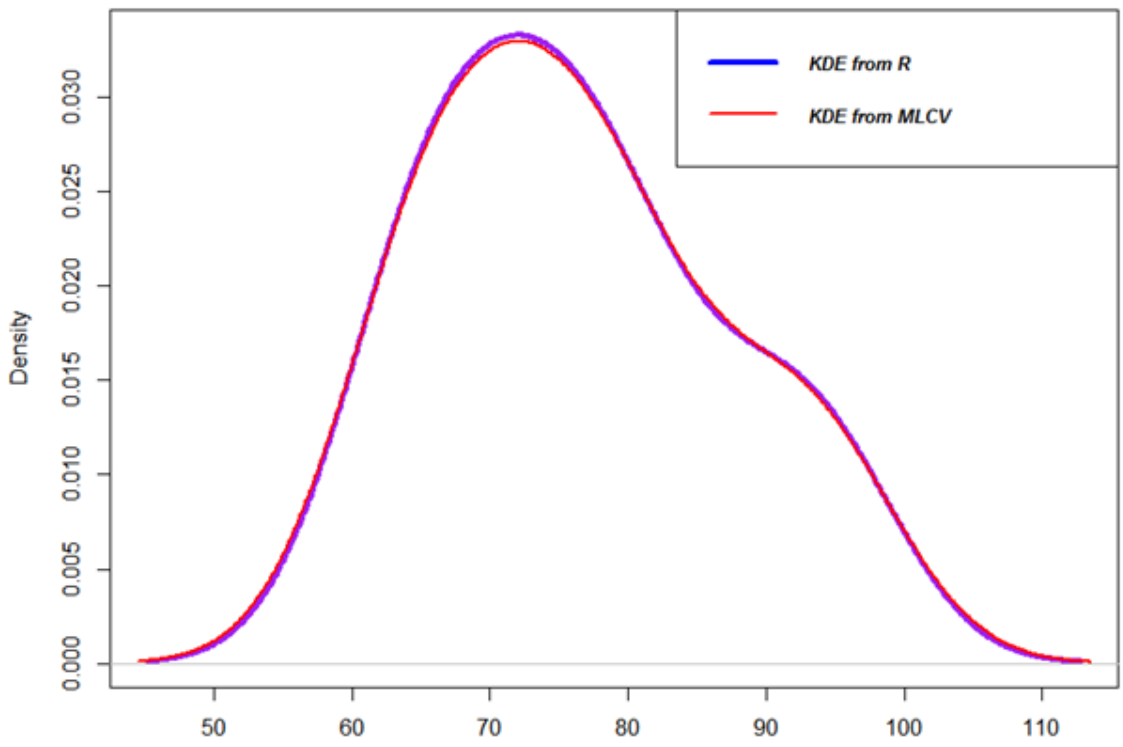
El valor final de la función objetivo (MLCV) se calcula tomando la media de las diferencias obtenidas al restar el Término 2 del Término 1 como se muestra en la siguiente tabla.

Xi	Term1	Term2	Difference
63	0.95591	3.29584	-2.339922
65	0.9559181	3.29584	-2.339919
67	0.9559182	3.29584	-2.339919
71	0.9559182	3.29584	-2.339919
75	0.9559182	3.29584	-2.339919
75	0.9559182	3.29584	-2.339919
79	0.9559182	3.29584	-2.339919
83	0.9559182	3.29584	-2.339919
91	0.9550724	3.29584	-2.340764
95	0.9174173	3.29584	-2.37842
Mean			-2.3438539

El valor de la función objetivo, es decir, MLCV, es -2.3438539 para el ancho de banda, h = 3. Se repite el mismo proceso seleccionando diferentes valores de h para que el valor de MLCV se acerque a un máximo finito para optimizar h. El algoritmo de optimización de búsqueda de la sección dorada en la función "optimizar" en R se utiliza para maximizar la función MLCV.

El valor optimizado de h en este ejemplo es 6.16 y el valor de la función objetivo es -2.295783.

El KDE se estima y se traza utilizando un ancho de banda optimizado (= 6,16) y se compara con el KDE obtenido mediante la función de densidad en R. Como se muestra en el siguiente gráfico, El KDE con h optimizada es bastante similar al KDE trazado mediante la función de densidad R.



Referencias

Kernel Estimator and Bandwidth Selection for Density and Its Derivatives by Arsalane Chouaib Guidoum (2015)

Density Estimation: Histogram and Kernel Density Estimator by Yen-Chi Chen (2018)