Programación Matemática y Software (2021)13(3): 75-90. ISSN:2007-3283

Simulación de los niveles máximos de ozono en la ciudad de Puebla por medio del método kernel

Simulation of maximum ozone levels in Puebla City through the kernel method

Juan Antonio Vazquez-Morales, Hortensia Josefina, Reyes-Cervantes, Bulmaro, Juárez-Hernández

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas,
Av. San Claudio y 18 Sur, Col. San Manuel, C.P. 72570, Puebla, Puebla

218470565@alumnos.fcfm.buap.mx, hreyes@fcfm.buap.mx, bjuarez@fcfm.buap.mx

PALABRAS CLAVE:

Medio ambiente, simulación, método kernel, datos faltantes.

RESUMEN

En este artículo se presenta una propuesta para reemplazar los datos faltantes de una base de datos, basado en la simulación de una variable aleatoria teórica que tendría los niveles máximos de ozono en la ciudad de Puebla de la República Mexicana, ciudad con problemas ambientales y deficiencia en el monitoreo de la calidad del aire. La metodología planteada está basada en el método kernel, que se utiliza para la aproximación de una función de densidad a partir de una muestra independiente. Se presenta una propuesta de algoritmo para el reemplazo de datos faltantes, así como una prueba que valida su utilidad y por último su aplicación a los datos de niveles de ozono de la estación de monitoreo automática Agua Santa ubicada en la ciudad de Puebla.

KEYWORDS:

Environment, simulation, kernel method, missing data

ABSTRACT

This article presents a proposal to replace missing data from a database, established on the simulation of a theoretical random variable that has the maximum ozone levels in Puebla City, a Mexican city with environmental problems and deficiencies on air quality monitoring. The proposed approach is based on the kernel method, which is used to approximate a density function from an independent sample. An algorithm proposal is presented to replace missing data, as well as a test that validates its usefulness and finally its application to the ozone level data from Agua Santa automatic monitoring station located in the Puebla City.

•**Recibido:** 17 de junio de 2021 • **Aceptado:** 2 de septiembre de 2021 • **Publicado en línea:** 1 de octubre de 2021

1. Introducción

En los últimos años, la contaminación ambiental provocada por mano del ser humano ha sido un tema de interés, en especial los efectos en la salud. En la contaminación fotoquímica1, el ozono es considerado principalmente el componente más toxico de la mezcla. Los estudios han revelado que las altas concentraciones de ozono, tienen efecto relacionados con el sistema respiratorio, como la disminución de la función pulmonar o agravamiento del asma como se menciona en [1]. En el caso del monitoreo de los niveles de ozono, tanto como en otras áreas del conocimiento, existe perdida de datos por muy diversas situaciones que el investigador no puede controlar, como son, falta de apoyo económico, los individuos han muerto o se cambiaron de sitio geográfico, desperfectos en los dispositivos de medición, etc.

Rodríguez S. et. al., trabajó los niveles máximos de ozono en la ciudad de la México a partir del supuesto de que los datos tienen una distribución generalizada de valores extremos [2]. En este trabajo se estimó los parámetros de la distribución teórica para el manejo de los datos faltantes en las bases de datos.

En este artículo se propone un método no paramétrico para el reemplazamiento de los niveles máximos de ozono en la ciudad de Puebla, en este sitio Cruz- Juárez et al. han estimado los datos faltantes por otros métodos y han realizado estimaciones de los niveles de ozono [3]. Por la experiencia en estudios anteriores, se asume que los datos tienen una función de densidad continua y se precede a encontrar la función de densidad por el método kernel.

2. Método kernel

Se supone una función de densidad $f(x)$. Si no coincide con un modelo conocido, entonces el problema radica que a partir de muestra aleatoria $x_1, x_2, \dots, x_n$ se debe encontrar la función de densidad de la variable X. Para tal objetivo, se utiliza el método de kernel, el cual utiliza el conjunto de datos $x_1, \dots, x_n$ para aproximar a la función de densidad.

2.1. Construcción de los kernel's

Definición 2.1 (Kernel) Un kernel es una función $K: \mathbb{R} \to \mathbb{R}$ que cumple:

$$K(x) \in [0, \infty), x \in [-1, 1],$$

$$K(x) = 0, x \notin [-1,1],$$

$$K(x) = K(-x),$$

$$\int_1^{-1} K(x)dx = 1,$$

$\int_1^{-1} xK(x)dx = 0$ y

$$\int_1^{-1} x^2 K(x)dx > 0.$$

Parametrización de kernel's

En la definición anterior, el intervalo donde K es no negativa es $[-1, 1]$, pero esto puede modificarse mediante un parámetro.

[1] Contaminación principalmente procedente de las reacciones de los hidrocarburos y los óxidos de nitrógeno, estimuladas por la luz solar intensa y el incremento de la temperatura.

Sea $h > 0$, el kernel parametrizado en h es

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right), x \in [-h, h].$$

Esta modificación mantiene las propiedades, pero ahora sobre el intervalo $[-h, h]$. A h se le conoce como el ancho de banda de K.

Traslación de kernel's

En la Definición 2.1, el kernel está centrado en 0, pero se puede centrar en cualquier otro punto $x_i \in \mathbb{R}$. El kernel parametrizado en h y centrado en x_i está dado por

$$K_h(x) = K\left(\frac{x - x_i}{h}\right), \qquad x \in [x_i - h, x_i + h].$$

Las condiciones de la Definición 2.1 para un kernel parametrizado en h y centrado en x_i son las siguientes:

$$K_h(x) \in [0, \infty), x \in [x_i - h, x_i + h],$$
$$K_h(x) = 0, x \notin [x_i - h, x_i + h],$$
$$K_h(x + x_i) = K(-x + x_i),$$
$$\int_{x_i - h}^{x_i + h} K_h(x)dx = 1,$$

$\int_{x_i - h}^{x_i + h} xK_h(x)dx = 0$ y

$$\int_{x_i + h}^{x_i + h} x^2 K_h(x)dx > 0.$$

2.2. Construcción de funciones de densidad de probabilidad

Sea X una variable aleatoria continua, con densidad $f(x)$ desconocida. Supóngase que se dispone una muestra aleatoria independiente $x_1, x_2, \ldots, x_n$. El objetivo es obtener un estimador $\hat{f}(x)$ de la función $f(x)$ a partir de dicha muestra. Para tal objetivo se explicará el método de kernel.

El método kernel

Un kernel no es más que una función de densidad. Si se coloca un kernel en cada uno de los datos de la muestra, la suma ponderada de estas funciones será una función de densidad. Esta suma es una función continua, que capta la influencia de los datos cercanos.

Definición 2.2 (Estimación por kernel's) Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria, entonces la estimación por kernel es

$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}K\left(\frac{x - x_i}{h}\right), \ x \in \mathbb{R}.$

El ancho de banda h es el parámetro de suavizado de $f(x)$. Si h es pequeño, más concentrada está la construcción del kernel en cada punto x_i. Si h es muy grande, habrá mayor la influencia e interacción del kernel hacia los puntos vecinos. Cuando $h \to 0$, la contribución de cada kernel estará concentrada en cada punto x_i, así $\hat{f}(x)$ tendrá una distribución puntual. Si $h \to \infty$, $\hat{f}(x)$ se aplana en un solo cúmulo y con mayor dispersión [4]. De lo anterior, se expresa la necesidad de buscar un ancho de banda adecuado para construir $\hat{f}(x)$.

2.3. Eficiencia del estimador $\hat{\boldsymbol{f}}(\boldsymbol{x})$ con respecto a $\boldsymbol{f}(\boldsymbol{x})$

Ahora es importante saber cómo medir que tan eficiente es la estimación $\hat{f}(x)$, para ello se da la definición siguiente.

Definición 2.3 Se definen las medidas siguientes: Error cuadrático medio como

$$ECM\left[\hat{f}(x)\right] = E\left[\hat{f}(x) - f(x)\right]^2.$$

El error cuadrático medio integrado como

$$ECMI\left[\hat{f}(x)\right] = \int_{\mathbb{R}} ECM\,[f\hat{}(x)]. \quad (1)$$

Teorema 2.4 Sea $x_1, \ldots, x_n$ una muestra aleatoria e independiente de una variable aleatoria con densidad $f(x)$, la cual sea continua o acotada, y sea f̂(x) la aproximación por método kernel, entonces

$$\lim_{n\to\infty} ECM\left[\hat{f}(x)\right] \to 0,$$

para toda $x \in \mathbb{R}$.

Demostración. Véase [5].

Actualmente, el método kernel es aceptado y la investigación sobre este, están basadas en la elección del ancho de banda h. El selector Plug-In se basa en la ecuación (1), el enfoque habitual es encontrar una gran aproximación de muestra del ECMI utilizando la técnica de expansión de la serie Taylor, explicado en [6] y [7].

3. Simulación de variables aleatorias

Actualmente la simulación de una variable aleatoria con cierta distribución está programada en varios softwares como R. Por otro lado, existen variables aleatorias no tan comunes que no están programadas o bien, un investigador propone una nueva variable, por ello, se necesita métodos para simular variables aleatorias. Todos los software o paquete estadístico ya tiene programada la simulación de una variable aleatoria uniforme, hecho que se utiliza para simular otras variables, como se mostrará a continuación.

[2] Vector uniforme en $\Omega = \{(x;u) : 0 < u < f(x)\}$, para más información véase [8].

Definición 3.1 Para una función no decreciente F en $\mathbb{R}$, el inverso generalizado de F, F^-, es la función definida por

$$F^{-1}(u) = inf\{x: F(x) \geq u\}.$$

Lema 3.2 (Transformación Integral de Probabilidad) Si $U \sim Unif[0,1]$ y F una función de distribución, entonces la variable aleatoria $F^{-1}(U)$ tienen la distribución F.

Demostración. Véase [8].

Por lo tanto, usando el Lema anterior, para generar una variable aleatoria X que tenga función de distribución F , es suficiente generar $U \sim Unif[0,1]$ y luego hacer la transformación $x = F^{-1}(u)$.

Teorema 3.4 (Teorema fundamental de simulación) Simular X con función de densidad $f(x)$ es equivalente a simular

$$(X,U) \sim U\,nif\{(x,u): 0 < u < f(x)\}.\,2$$

Demostración. Véase [8].

4. Simulación de los datos faltantes

Una forma de simular datos usando kernel's, es la de asignar un cierto peso relacionado con la distancia de los datos disponibles más cercanos. Este método puede ser observado en [9]. El propósito de este artículo es proponer un método para asignar a partir de la aproximación por método kernel reemplazando a los datos faltantes por un número aleatorio que esté acorde con los datos observados.

4.1. Prueba de la metodología

Con ayuda de la prueba Kolmogorov-Smirnov se hace un experimento, con el cual se busca verificar que el método planteado es efectivo. Para esto se usará la prueba de hipótesis al, $\alpha = 0.05$:

H_0: La muestra proviene de la función de densidad f(x).

vs

H_1: La muestra no proviene de la función de densidad f(x).

4.2 Discreteando la aproximación por método kernel

El experimento se desarrolla de la manera siguiente:

1. Se simulan 1500 datos de una función de densidad $f(x)$.
2. A partir de los datos generados, se crea la estimación por kernel $\hat{f}(x)$, tomando el ancho de banda h como el dado por el selector directo Plug-In, con los kernel's rectangular[3], epanechnikov[4], cuártico[5] y triweight[6], con los códigos en [10].
3. Se discretica f^(x), usando intervalos de $M = 100, 500$ y 1000.
4. Después de haber obtenido la discretización, se simulan 1500 datos con dicha variable.
5. Se calcula el p-valor de la prueba.

En el Tabla 1 se utiliza $f(x)$ como la función de densidad de una $Beta(3,2)$, cuyo rango es finito, por lo que se espera que $\hat{f}(x)$ sea una buena aproximación a $f(x)$. Se observa que en su mayoría la prueba de hipótesis es aceptada a un nivel de significancia $\alpha = 0.05$. Como se esperaba, al tomar más valores de M, la hipótesis es aceptada por todos los kernel, aunque en 3 casos la hipótesis es rechazada.

Tabla 1 Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra Beta(3,2).

	KERNEL			
	Rectangular	Epanechnikov	Cuártico	Triweight
h	0.08096936	0.103014	0.1220371	0.138579
p-valor para M = 100	0.008589	0.3811	0.009352	0.6535
p-valor para M = 500	1	0.0407	0.3061	0.6287
p-valor para M = 1000	0.6474	0.1086	0.09922	0.3946

[3] El kernel rectangular se define por

$$K(x) = \frac{1}{2}.$$

[4] El kernel de Epanechnikov se define por

$$K(x) = \frac{3}{4}(1 - x^2).$$

[5] El kernel cuartico se define por

$$K(x) = \frac{15}{16}(1 - x^2)^2.$$

[6] El kernel triweight se define por

$$K(x) = \frac{35}{32}(1 - x^2)^3.$$

En el caso de la densidad $Gamma(3, 0.5)$, cuyo rango son los números positivos, la experimentación con $M = 100$, las pruebas fueron rechazadas, como se observa en el Tabla 2. Sin embargo, con los otros parámetros la hipótesis es aceptada, lo que nos permite formular que, si el rango de la discretización aumenta, los resultados son mejores.

Tabla 2Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra Gamma(3,0.5).

	KERNEL			
	Rectangular	Epanechnikov	Cuártico	Triweight
h	1.130752	1.43861	1.70427	1.935281
p-valor para $M = 100$	0.002666	0.002353	0.01432	0.03368
p-valor para $M = 500$	0.442	0.4958	0.11	0.8366
p-valor para $M = 1000$	0.5356	0.1149	0.1549	0.7328

En el Tabla 3 se muestran los $p-$valores obtenidos del experimento con una densidad $Gumbel(3, 4)$. Se observa que la hipótesis nula se rechaza en los casos donde se discretiza la variable en un rango de $M = 100$, la mitad en donde $M = 500$. En este caso es fácil pensar que si el número del rango en que se discretiza la variable aumenta se obtiene una mejor aproximación a $f(x)$.

Tabla 3.Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra Gumbell(3,4).

	KERNEL			
	Rectangular	Epanechnikov	Cuártico	Triweight
	1.750271	2.2268	2.63801	2.995588
p-valor para $M = 100$	0.05419	0.002279	0.04846	0.06553
p-valor para $M = 500$	0.2569	0.07	0.001438	0.04282
p-valor para $M = 1000$	0.34	0.5499	0.05375	0.3512

En el último caso, se trabaja con la densidad $Weibull(1, 2/3)$. En este, el caso se tiene información negativa, ya que, en todo momento, la prueba de hipótesis es rechazada, como se observa en el Tabla 4.

Tabla 4Prueba de Kolmogorov-Smirnov para el método propuesto, para una muestra Weibull(1,2/3).

	KERNEL			
	Rectangular	Epanechnikov	Cuártico	Triweight
h	0.1065738	0.1355896	0.1606282	0.182401
p-valor para $M = 100$	7.876×10^{-5}	0.0007375	0.005991	8.056×10^{-8}
p-valor para $M = 500$	0.05436	0.01336	0.01059	0.09265
p-valor para $M = 1000$	0.02518	0.02696	0.07134	9.579×10^{-5}

Analizando la información en conjunto, se formula lo siguiente:

Para densidades de cola pesada como la Gumbell y Weibull, con pocos datos, parecen no dar buenos resultados, ya que recordemos que la teoría de kernel, entre más datos, mejor la aproximación, entonces para ellas se necesitan más datos.

Entre mayor sea la discretización, el método funciona mejor.

Aunque el kernel Epanechnikov es el kernel más estudiado, los kernel más útiles son el kernel Rectangular y Triweight.

Los resultados permiten pensar que la discretización es una buena opción excepto para las distribuciones de cola pesada, las cuales son la Gumbel y Weibull, esto se debe a los pocos datos utilizados.

4.3. Algoritmo para la simulación de datos faltantes

Con los datos obtenidos, se tiene evidencia de que la propuesta es viable para ser incluido en los métodos para trabajar con datos faltantes, por lo que ya se puede formular el algoritmo para simulación de datos faltantes.

Para el algoritmo, se requiere de los datos con los que se quiere trabajar que tienen datos faltantes, donde se utiliza la aproximación $\hat{f}(x)$. Las ideas utilizadas hasta el momento se resumen en el Algoritmo 4.1.

Algoritmo 4.1 (Reemplazamiento de datos faltantes)

Entrada: Datos en vector $DATOS$, el kernel K y el entero M.

Salida: Datos en vector $DATOS$.

1. Se crea una copia de $DATOS$, la cual es llamada D;
2. Se eliminan las entradas de D donde se tiene datos faltantes;
3. Se obtiene el ancho de banda h utilizando el selector DPI, usando el kernel K y D;
4. Se obtiene la aproximación por método kernel $\hat{f}(x)$ con ancho de banda h, datos D y kernel K;
5. Se crea matriz V de tamaño $2 \times M$;
6. Se obtiene $maxi = max\ D + h$;
7. Se obtiene $mini = \min D - h$;
8. Se calcula $Paso = (maxi - mini)/(M +$

1);

9. Para i = 1 hasta M hacer
10. Se calcula $V[1,i] = mini + i \times PASO - PASO/2$, es decir, un punto medio;
11. Se calcula $V\,[2,i] = P\,(mini + (i - 1) \times P\,ASO < X < mini + i \times P\,ASO)$, es decir, la probabilidad;
12. Fin Para
13. Se eliminan las columnas de V donde la segunda entrada de la columna es 0.
14. Para cada entrada de $DATOS$ que sea un dato faltante, reemplazar por una simulación de una variable aleatoria V con función de densidad $\hat{f}(x)$.

En el Algoritmo 4.1, los pasos 9-12 son para obtener una discretización de la variable aleatoria asociada a $\hat{f}(x)$, siendo la primera fila el rango de la variable y la segunda fila las probabilidades. En el paso 13, se hace para la mejor implementación del Lema 3.2.

5. Aplicación. Niveles máximos de ozono en la ciudad de puebla

En esta sección, se utiliza el Algoritmo 4.1 para simular los niveles máximos de ozono, de la estación Agua Santa en la ciudad de Puebla, obtenidos en la página de la SINAICA obtenidos en [11]. Los datos corresponden del 1 de enero de 2001 al 31 de diciembre de 2014, tomando como datos las partes por millón de $O3$.

5.1. Estimación de datos dudosos

Se puede pensar que los máximos tomados cada 2 semanas son independientes, con esta idea, se toman los máximos cada 2 semanas, sin omitir los datos faltantes, con lo que se obtienen 365 datos. Los datos omitidos son aquellos que, en las 2 semanas consideradas, los datos completos son menores del $100p\%$, donde $p = 0.5,\ 0.25$ y 0.1, los datos restantes son considerados datos completos u originales. Para el cálculo del ancho de banda h, se consideran solo los datos completos, y con los códigos en [10], dichos anchos de banda se encuentran en la Tabla 5. Se observa que, para cada kernel, al haber menor número de datos, el ancho de banda aumenta, esto se debe que, al tener una menor cantidad de datos, se necesita tener más información de los puntos vecinos.

Tabla 5Estimación del ancho de banda h para distintos p.

Kernel	Rectangular	Epanechnikov	Cuártico	Triweight
h para $p = 0.5$	0.01802722	0.02293531	0.0271065	0.03085358
h para $p = 0.25$	0.01732555	0.02204259	0.02611308	0.02965266
h para $p = 0.1$	0.01717444	0.02185035	0.02588533	0.02939404

Una vez calculados, se realiza la simulación de los datos faltantes, esto se hace discretizando la aproximación por método de $\hat{f}(x)$, con un número esperado de intervalos $M = 1000$, y se sustituyen los datos faltantes por una simulación

de dicha variable, y como el valor simulado puede no ser entero, se trunca este valor. Lo anterior se realiza con los kernel's rectangular, epanechnikov, cuártico y triweight.

5.2. Resultados

En la Tabla 6, se muestran los resultados obtenidos cuando la proporción de datos faltantes, es mayor a $p = 0.5$, por lo que fueron omitidos 144 datos. Todas las estadísticas sufren cambios, en especial con el kernel cuártico, como por ejemplo el mínimo y máximo, lo cual puede ser por la forma del kernel. Por otro lado, las estadísticas exhiben el hecho que, a una cantidad grande de datos incompletos, las estadísticas muestran una mayor dispersión.

Tabla 6Cálculo de las estadísticas para los datos originales y los datos obtenidos con la simulación de datos, con un p = 0.5.

	Min	1er Cuar.	Mediana	Media	3er Cuar.	Max
Original	0.021	0.059	0.084	0.08171	0.101	0.178
K.	0.01021	0.05674	0.082	0.08052	0.101	0.17836
K.	0.006687	0.06	0.085198	0.082378	0.102	0.178
K. Cuártico	6.392	5.839×10^{-2}	8.4×10^{-2}	8.25×10^{-2}	1.026×10^{-1}	1.86
K. Triweight	0.006659	0.06	0.084081	0.082926	0.103	0.178

En la Figura 1(OzonoP0.5.png) se observa el histograma de los datos originales, así como de los datos simulados con la metodología planteada, todas las simulaciones tienen mías barras que de los datos originales. En la Figura 2(ComparacionP0.5.png), se colocan las curvas de densidad aproximadas dadas por el paquete R, con lo que se observa una mínima diferencia entre ellas, pues sus graficas muestran cambios por unas unidades y cambios en el rango.

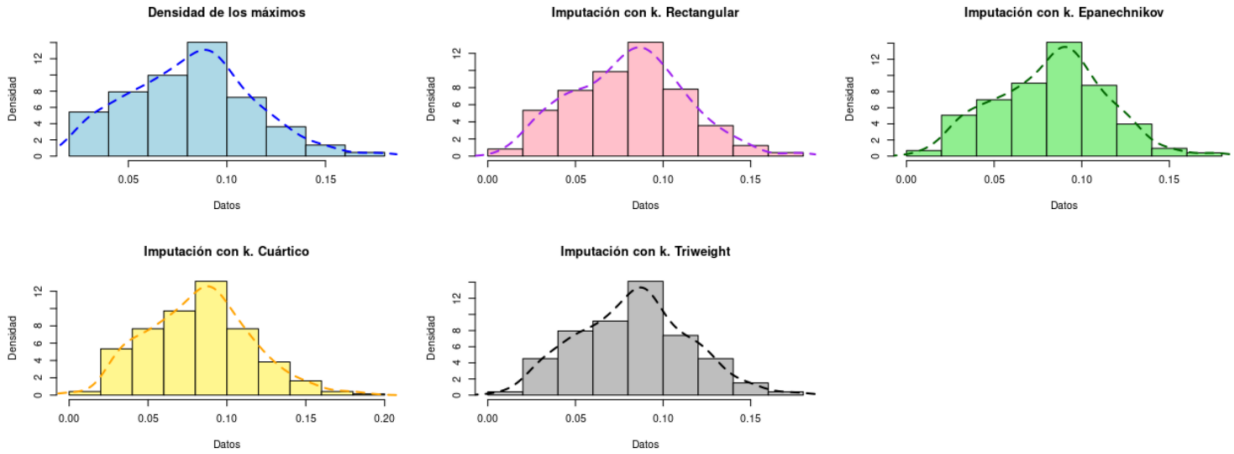


Ilustración 1Histogramas y densidades para p = 0.5.

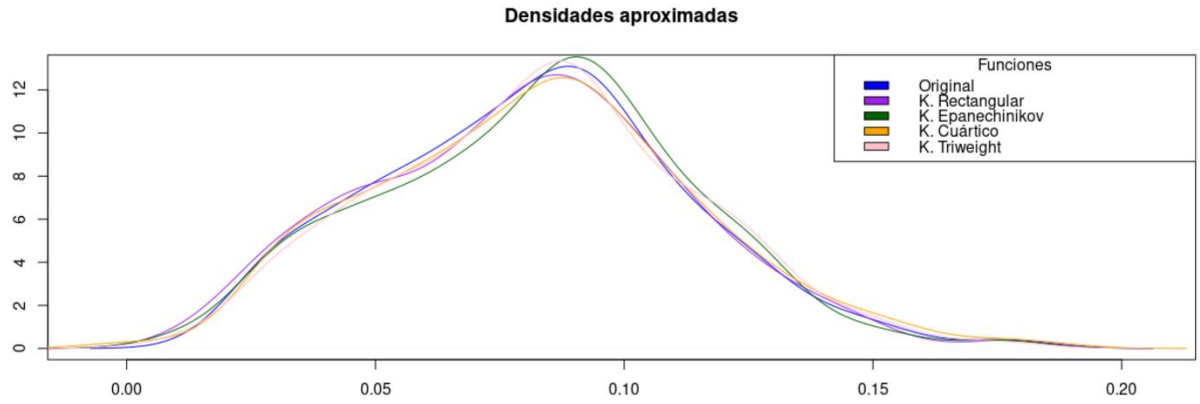


Ilustración 2Comparación de las densidades obtenida para p = 0.5.

Por lo anterior se podría decir que el kernel con resultados que más variaciones tiene, es el cuártico, y el kernel que da mejor resultado es el kernel rectangular, cuyas estadísticas con respecto a la original están cercanas a los originales.

Para el caso donde $p = 0.25$, se omitieron un total de 138 datos faltantes, solo 6 menos que en caso anterior. Como se observa en la Tabla 7, los máximos no cambian en general. Se empieza a observar una mayor cercanía en las otras estadísticas, que se esperaba al aumentar la cantidad de datos faltantes, la estadística que parece ser menos influenciada por la simulación de los datos faltantes es la mediana.

Tabla 7Cálculo de las estadísticas para los datos originales y los datos obtenidos con la simulación, con un p=0.25.

	Mín	1er Cuar.	Mediana	Media	3er Cuar.	Máx
Original	0.021	0.058	0.084	0.08151	0.1005	0.178
K. Rectangular	0.01642	0.057	0.083	0.08141	0.101	0.178
K. Epanechnikov	0.009514	0.06	0.082	0.081256	0.101	0.178
K. Cuártico	0.01508	0.05985	0.084	0.08221	0.104	0.178
K. Triweight	0.00465	0.06	0.084	0.08245	0.103	0.18981

Por otro lado, en cuanto a los histogramas, haciendo una inspección a simple vista a la Figura 3(OzonoP0.25.png), se puede observar el cambio en el número de barras, pero ninguna tan parecida a la original. En cuanto a las funciones de densidad aproximadas en la Figura 4(ComparacionP0.25), todas las gráficas de las simulaciones empiezan a ver una mayor cercanía a la gráfica de los datos originales, a excepción

del kernel cuártico.

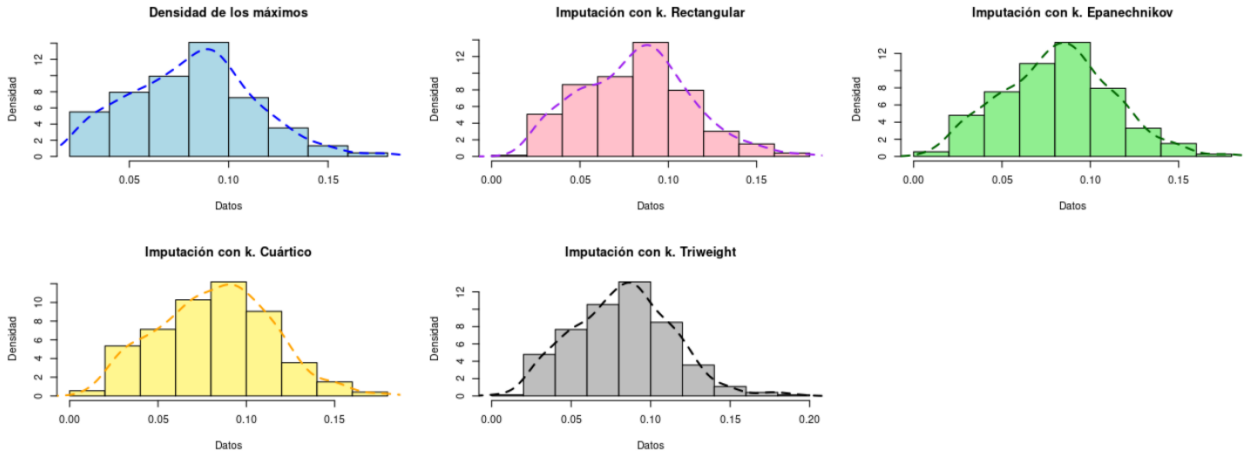


Ilustración 3Histogramas y densidades para p=0.25.

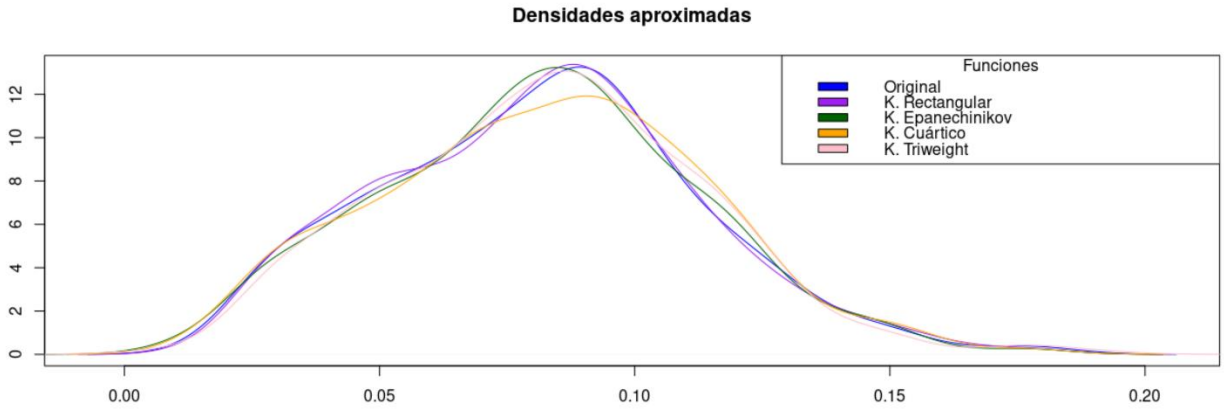


Ilustración 4Comparación de las densidades obtenida para p=0.25.

Se puede decir que el kernel que da mejor resultado es el rectangular, ya que son los que más cercanas son sus estadísticas con respecto a de los originales, a excepción de la media. A diferencia del caso anterior, el que da peores resultados, pero no necesariamente malos, parece ser el kernel triweight, que es el que tiene mayor diferencia en las estadísticas y el cuártico que tiene un ajuste de función de densidad con mucha diferencia.

En la Tabla 8, se observan las estadísticas para el caso donde $p = 0.1$, la cual se omiten 135 datos, no tanta diferencia con el caso anterior. La diferencia entre las estadísticas se mantiene, ahora si se nota diferencia en el máximo.

Tabla 8: Calculo de las estadísticas para los datos originales y los datos obtenidos con la simulación de datos, con un p = 0.1.

	Mín	1er Cuar.	Mediana	Media	3er Cuar.	Máx
Original	0.021	0.0575	0.084	0.08152	0.10075	0.178
K. Rectangular	0.006217	0.056	0.082948	0.081337	0.101	0.195079
K. Epanechnikov	0.0129	0.05906	0.083	0.0809	0.1004	0.178
K. Cuártico	0.01067	0.057	0.08207	0.08108	0.10065	0.19773
K. Triweight	0.01588	0.05861	0.082	0.08073	0.1	0.

En la observación de los histogramas de la Figura 5(OzonoP0.5.png) se ven parecidos al caso anterior, excepto con el kernel triweight, que ahora tiene un número igual en el número de barras que de los datos originales. En la observación de las densidades aproximadas, en la Figura 6(ComparacionP0.5), no se observa la diferencia entre las colas de las densidades, pero si al rededor del punto 0.05, pero general todos son buenas funciones de densidades aproximadas a la de los casos originales.

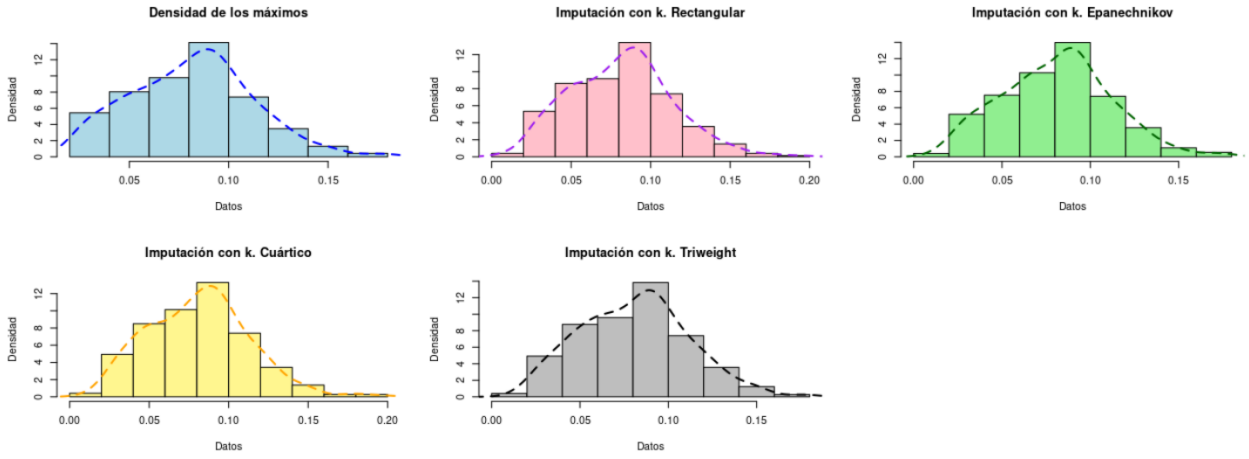


Ilustración 5Histogramas y densidades para p=0.1.

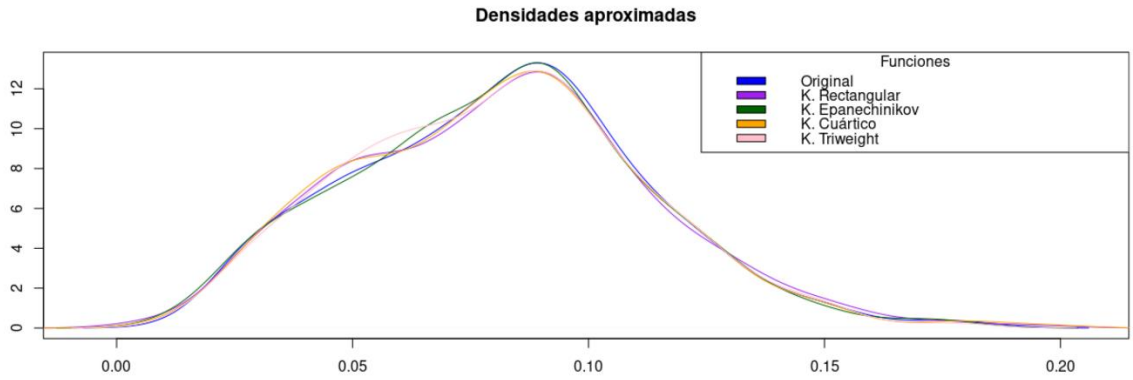


Ilustración 6Comparación de las densidades obtenida para p=0.1.

Por lo anterior se podría decir que el kernel con mejores resultados resulta ser el triweight, y solo por el hecho que su rango es más aproximado a la de la función de densidad original.

En general, se puede decir, que, de acuerdo a las gráficas y las estadísticas, cualquier kernel proporciona una buena simulación de datos, ya que las estadísticas solo difieren por unas cuantas centésimas, por ejemplo, en la Tabla 8, en el caso de usar un kernel rectangular, el cual la diferencia entre el 3er cuartil de los datos originales con respecto al obtenido con los datos simulados, al truncar los datos a centésimas, todas son iguales.

Por otro lado, al usar los datos como se muestra, y no hacerles una escala para que los datos fueran enteros, es más fácil que los máximos y mínimos no cambien, y permitiendo que todas las funciones de densidad obtenidas tengan el mismo rango, pero al dejarlos como datos como se encontraron, al tener una cantidad reducida de datos, se encuentra un mejor ajuste si hay una gran cantidad de datos, pues exactamente 134 máximos, provienen de días en que no se tomó muestra. Por tanto, la simulación de datos faltantes por medio del método kernel, es un buen método para trabajar cuando existen datos faltantes.

6. Conclusiones

En este trabajo, se desarrolló la idea de utilizar las estimaciones por método kernel, para simular los datos faltantes en bases de datos, por lo que se planteó una metodología. Se diseñan experimentos donde se ve la viabilidad del método propuesto, por lo que la metodología expuesta es aceptada como método para trabajar con datos faltantes. Por todo esto, las ideas se resumen en el Algoritmo 4.1.

Las bases en las que se recomienda aplicar los algoritmos, son aquellas en las que existe independencia entre los datos, ya que es un requerimiento para el método kernel. Se debe considerar que las observaciones provienen de una variable aleatoria, o bien, hacer un tratamiento a los datos, que permitan estudiar el fenómeno de interés, y así aplicar el Algoritmo

4.1.

Es claro que al experimentar, $\hat{f}(x)$, tenga un rango mayor a la $f(x)$ si esta es de rango finito, esto por el ancho de banda obtenido, pero esto no fue especialmente un problema, ya que la probabilidad de que se obtuviera un número fuera del rango de la densidad $f(x)$, es relativamente pequeña, contrastando por el hecho de que nunca se obtuvo un dato simulado fuera del rango en los experimentos que se llevaron a cabo.

En cuanto a los códigos referentes al Algoritmos 4.1 son útiles, ya que permitieron hacer este trabajo, aunque hay inconvenientes. Para el Algoritmo 4.1, el código es rápido en su ejecución y el problema es que, si no hay suficientes datos, en especial, para variables aleatorias con rango grande, no se tendrá una buena simulación de la variable, por lo que no se puede asegurar una buena simulación.

La aplicación, sobre los niveles máximo de ozono, se ve que es una buena opción para tratar con los datos faltantes, pues en las gráficas y estadísticas dan buenos resultados. Como conclusión, los datos que se simularon resultan ser buenos datos para sustituir la información que falta, por lo que, si se desea calcular cualquier estadística que requiera una gran cantidad de datos, el Algoritmo 4.1 es una herramienta útil para tal objetivo, ya que a pesar que la base de datos tenía un rango de 2 años sin datos, su simulación no afecto lo observado en los máximos.

REFERENCIAS

[1] Ballester, F. Contaminación atmosférica, cambio climático y salud. Revista española de salud pública, 2005,79(2), 159-175.

[2] Rodríguez, S., Huerta, G., Reyes, H. A study of trends for Mexico City ozone extremes: 2001-2014. Atmósfera. 2016, 29(2), 107-120.

[3] Cruz, J., Reyes H., Rodrigues, E. Analysis of Ozone Behaviour in the City of Puebla-Mexico Using Non-Homogeneous Poisson Models with Multiple Change-Points. Journal of Environmental Protection. 2016, 7, 1886-1903.

[4] Rodríguez, L. Construcción de kernels y funciones de densidad de probabilidad. Matemáticas. 2013, 11(2), 27-40.

[5] Miñarro, A. Estimación no paramétrica de la función de densidad. Documento de Trabajo. Universidad de Barcelona, 1998.

[6] Gramacki, A. Nonparametric kernel density estimation and its computational aspects. Springer, 2018.

[7] Wand, M. P., M. Jones M. C. Kernel smoothing. Chapman and Hall/CRC, 1995.

[8] Robert, C., and Casella, G. Monte Carlo Statistical Methods, 2da ed. Springer Texts in Statistics. Springer, 2004.

[9] Lee, H., Kang, K. Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. Advances in Meteorology. 2015, 2015, 1-12.

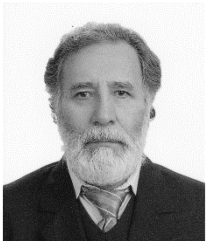
[10] Wand, M., Ripley, B. Functions for kernel smoothing supporting Wand & Jones (1995) (R package version 2.23-15). http://CRAN.R-project.org/package=KernSmooth, 1999.

[11] SINAICA. PUE_Puebla-Máximo 24 horas - O3. Base de datos. Recuperado de https://sinaica.inecc.gob.mx/, 2017

Acerca de los autores

Juan Antonio Vázquez Morales es estudiante de maestría en Ciencias en Matemáticas de la Benemérita Universidad Autónoma de Puebla BUAP. Egresado de la Licenciatura en Matemáticas Aplicadas de la FCFM BUAP y Técnico en Informática del CBTis 257 Ricardo Flores Magón. Pertenece al grupo de divulgación matemática Pitagóricos BUAP, participando en varios eventos como creador de talleres y expositor. Actualmente se encuentra laborando como docente en el Instituto del Bosque, en la ciudad de Puebla y se encuentra terminando sus estudios de maestría.

Dra. Hortensia Josefina Reyes Cervantes profesor investigador de tiempo completo, perteneciente SNI 1, con 33 años trabajando en la BUAP, las áreas de interés son la estadística aplicada al medio ambiente y las pruebas de no inferioridad para datos discretos.



Bulmaro Juárez Hernández. Doctorado en Ciencias con especialidad en Estadística. Instituto de Socioeconomía, Estadística e Informática del Colegio de Postgraduados. México. Profesor de la Facultad de Ciencias Físico Matemáticas de la Benemérita Universidad Autónoma de Puebla de 1987 a la fecha. Áreas de Interés: Series de Tiempo, Análisis de Supervivencia, Diseños Experimentales, Análisis de Regresión y Teoría de la Inferencia Estadística.