Estadística No Paramétrica: Taller 2

Sebastian Carvalho Salazar, ✉ scarvalhos@eafit.edu.co
Daniel Loaiza Lopez, ✉ dloaizal@eafit.edu.co
Sebastian Ramirez Escobar, ✉ sramireze1@eafit.edu.co
Asmec Duvan Urrea Uribe, ✉ adurreau@eafit.edu.co
Hernan Felipe Sanchez Cardenas, ✉ hfsanchezc@eafit.edu.co

Asesor:

Pablo Andres Saldarriaga Aristizabal

Universidad EAFIT
Maestría en Ciencia de Datos y Analítica
Medellín
2024

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

I. Punto 1

I-A. A

Se realiza el análisis del cálculo de la tabla de frecuencias de las temperaturas en la provincia de Alicante. Aquí, se agrupan las temperaturas en intervalos específicos, y se determina la frecuencia de ocurrencia de cada intervalo. Este proceso proporciona comprensión de la distribución de las temperaturas observadas.

TABLA. I: Temperature Frequency Distribution

Id	Temperature Range	Frequency	Relative Frequency
0	-14 to -13	1	0.000183
1	-12 to -11	0	0.000000
2	-10 to -9	1	0.000183
3	-8 to -7	12	0.002190
4	-6 to -5	54	0.009856
5	-4 to -3	140	0.025552
6	-2 to -1	313	0.057127
7	0 to 1	529	0.096550
8	2 to 3	433	0.079029
9	4 to 5	480	0.087607
10	6 to 7	447	0.081584
11	8 to 9	421	0.076839
12	10 to 11	456	0.083227
13	12 to 13	571	0.104216
14	14 to 15	551	0.100566
15	16 to 17	517	0.094360
16	18 to 19	314	0.057310
17	20 to 21	150	0.027377
18	22 to 23	62	0.011316
19	24 to 25	21	0.003833
20	26 to 27	6	0.001095

La tabla de frecuencias proporciona una visión detallada de la distribución de las temperaturas en la provincia de Alicante. Se observa que las temperaturas más comunes se encuentran en el rango de 12 a 13 grados Celsius, con una frecuencia de 571 ocurrencias y una frecuencia relativa del 10.42 %. Por otro lado, las temperaturas más extremas, tanto en el rango negativo como en el positivo, tienen una frecuencia mucho menor, lo que sugiere que estos eventos climáticos son menos frecuentes en la región. La distribución muestra una tendencia hacia temperaturas moderadas, con una cantidad considerable de observaciones en los rangos entre 0 y 17 grados Celsius. Esto puede reflejar el clima típico de la región de Alicante, caracterizado por inviernos suaves y veranos cálidos.

I-B. B

Se lleva a cabo la estimación de la densidad de probabilidad utilizando el método de Kernel Density Estimation (KDE). Se prueban diferentes kernels y anchos de banda para modelar la distribución de las temperaturas. Este enfoque permite visualizar y comprender mejor la distribución de las temperaturas, así como capturar posibles patrones o tendencias subyacentes en los datos.

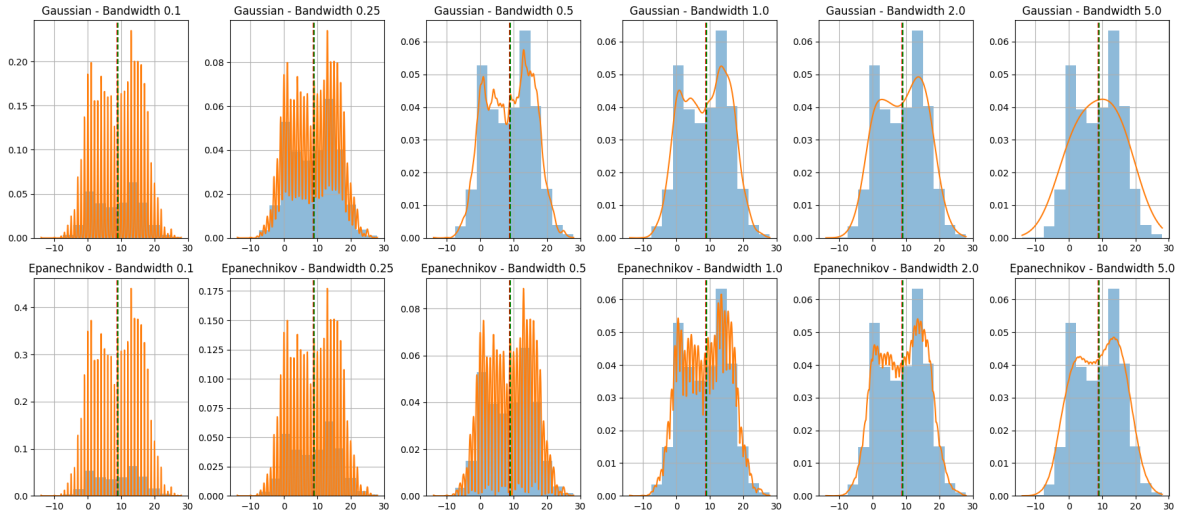

FIG. 1: Kernel Density Estimations

Los resultados de esta exploración revelan cómo la elección del kernel y el ancho de banda afecta la forma y la precisión de la estimación de la densidad de probabilidad de las temperaturas. Mientras que algunos kernels y anchos de banda pueden proporcionar una mejor adaptación a la distribución de datos observados, otros pueden introducir cierto grado de suavizado o sesgo en la estimación.

Al variar el ancho de banda en el método de Estimación de Densidad de Kernel (KDE), se observa un impacto significativo en la suavidad y la precisión de la estimación de la densidad de probabilidad. Un ancho de banda más pequeño resulta en una estimación más detallada y sensible a las fluctuaciones locales de los datos. Por otro lado, un ancho de banda más grande produce una estimación más suavizada y generalizada de la distribución, lo que puede ocultar detalles finos pero también reducir el ruido y la variabilidad aleatoria en los datos.

I-C. C

La estimación de las probabilidades para la media, moda y mediana se realizó utilizando dos enfoques: la tabla de frecuencias y el método de Kernel Density Estimation (KDE). En primer lugar, mediante la tabla de frecuencias, se calcularon las probabilidades acumuladas de que una temperatura sea menor o igual a la media, moda y mediana. Estas probabilidades se calcularon sumando las frecuencias relativas de todas las temperaturas iguales o menores que los valores respectivos de la media, moda y mediana. Para la media, moda y mediana de las temperaturas en la provincia de Alicante (8.77, 9.0 y 13, respectivamente), las probabilidades resultantes fueron del 47.18 %, 51.67 % y 70.41 %, respectivamente.

Por otro lado, se utilizó el método de KDE. Se integró la densidad de probabilidad estimada por KDE desde menos infinito hasta los valores respectivos de la media, moda y mediana para obtener las probabilidades acumuladas. Las probabilidades resultantes utilizando KDE fueron del 48.35 %, 49.42 % y 67.41 %, respectivamente, para la media, moda y mediana.

A continuación, se presenta una tabla resumiendo los valores estadísticos y sus correspondientes probabilidades utilizando ambos enfoques:

TABLA. II: Estadísticos y probabilidades

Estadístico	**Valor**	**Probabilidad (Tabla de frecuencias)**	**Probabilidad (KDE)**
Media	8.77	0.4718	0.4835
Moda	9.0	0.5167	0.4942
Mediana	13	0.7041	0.6741

Dado que las probabilidades estimadas para la media, moda y mediana de las temperaturas de la provincia difieren de 0.5, lo que sería esperado en una distribución normal no sesgada, parece que hay evidencia de sesgo en los datos. Las probabilidades estimadas, obtenidas tanto a través de la tabla de frecuencias como mediante el método de estimación de densidad de kernel (KDE), muestran desviaciones de este valor esperado. Por lo tanto, sugiere que la distribución de las temperaturas de la provincia no sigue una distribución normal sin sesgo y podría estar influenciada por factores que sesgan la distribución.

II. PUNTO 2

II-A. A

se está utilizando el método de Kernel Density Estimation (KDE) para simular datos de temperatura. El KDE es una técnica estadística que estima la función de densidad de probabilidad de una variable aleatoria a partir de un conjunto de datos. En este caso, se ajusta un KDE a los datos reales de temperatura en la provincia de Alicante. Luego, se

generan muestras simuladas utilizando la distribución de probabilidad estimada por el KDE. Esto se logra muestreando aleatoriamente nuevos valores de temperatura de acuerdo con la distribución de densidad de probabilidad obtenida del KDE. La comparación visual entre las temperaturas reales y las temperaturas simuladas permite evaluar cuán bien se ajusta el modelo de KDE a los datos reales y determinar si las temperaturas simuladas reproducen adecuadamente la distribución de las temperaturas reales.

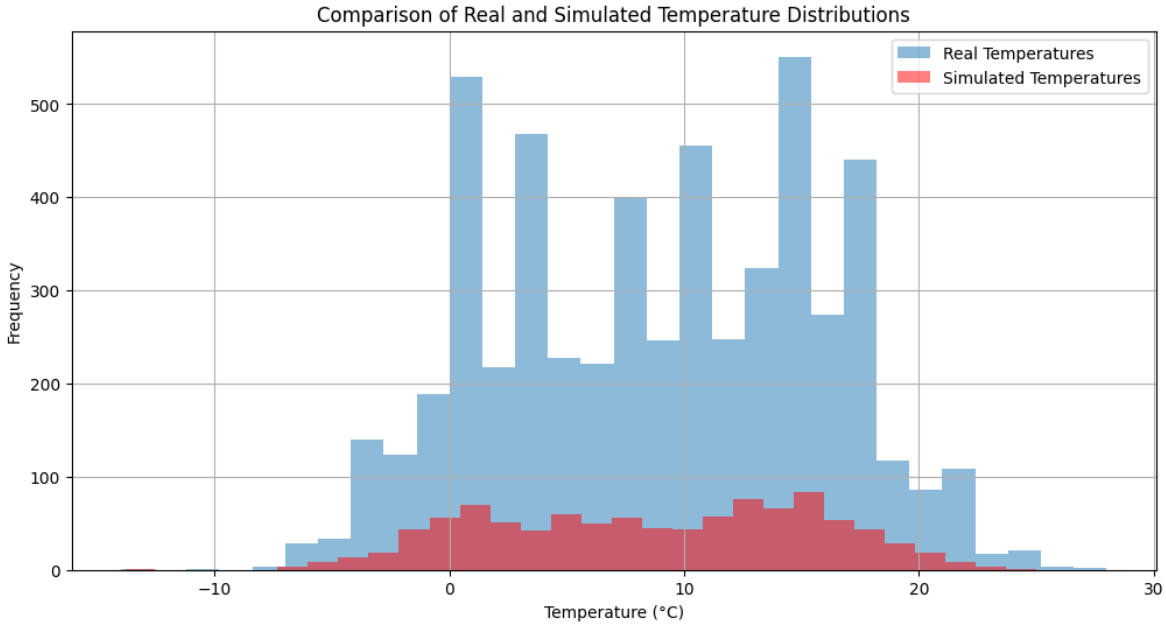

FIG. 2: Comparison of Real and Simulated Temperature Distributions

II-B. B

En este proceso, se realizan 10,000 simulaciones para calcular el promedio de temperaturas en la provincia de Alicante. En cada simulación, se generan muestras aleatorias de temperaturas utilizando el kernel. Luego, se calcula el promedio de temperatura para cada una de estas muestras simuladas. Este proceso se repite 10,000 veces para obtener una amplia distribución de promedios de temperatura simulados.

Una vez que se han calculado los promedios de temperatura para cada simulación, se procede a visualizar la distribución de estos promedios mediante un histograma. En el histograma, el eje x representa los promedios de temperatura, mientras que el eje y representa la frecuencia con la que ocurre cada promedio. Cada barra en el histograma muestra la frecuencia de ocurrencia de un rango específico de promedios de temperatura.

El proceso estadístico detrás de este análisis se basa en el Teorema del Límite Central (TLC). Este teorema establece que, independientemente de la distribución de los datos originales, la distribución de los promedios de muestras grandes tiende a aproximarse a una distribución normal. Por lo tanto, al realizar múltiples simulaciones y calcular los promedios

de temperatura en cada una, se espera que la distribución de estos promedios se aproxime a una distribución normal. El histograma resultante proporciona una representación visual de esta distribución y permite verificar si se cumplen las condiciones del TLC.

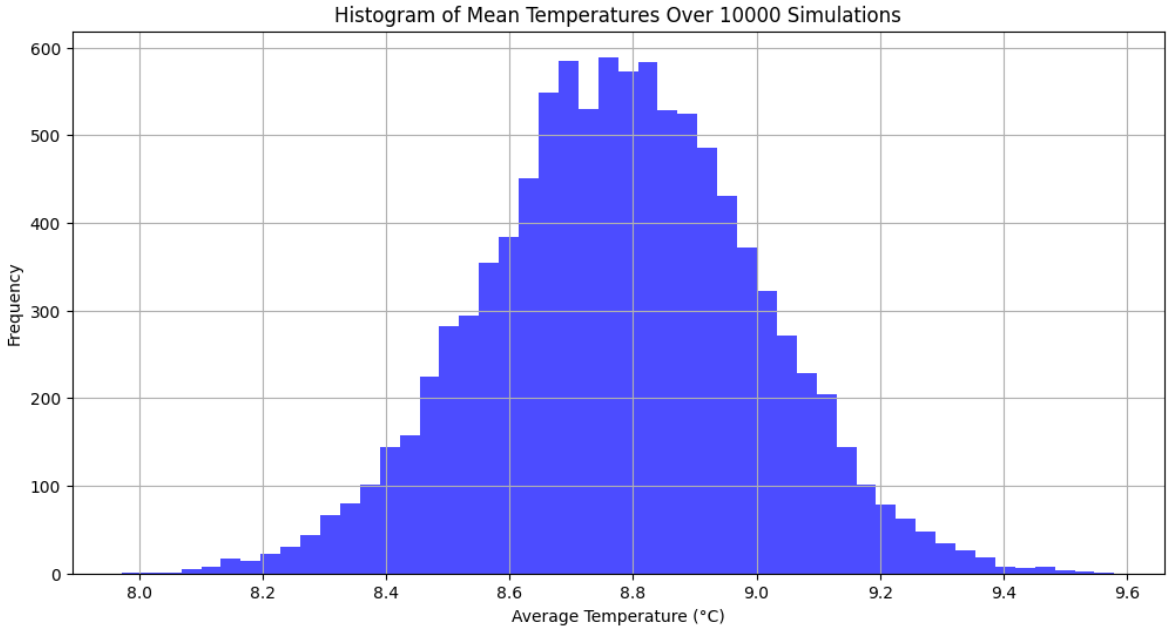

FIG. 3: Histogram of Mean Temperatures Over 10000 Simulations

III. PUNTO 3

Se quiere verificar que el Kernel Gaussiano cumple con las propiedades de un kernel de densidad, cabe recordar que en una distribución normal la varianza es 1 y la media es 0 lo que hace que la función del kernel Gaussiano está dado por la siguiente expresión:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{1}$$

Pero antes de esto, veamos cuales son las propiedades de un kernel de densidad:

$$k(x) \geq 0 \tag{2}$$

$$\int k(x)\, dx = 1 \tag{3}$$

$$\int xk(x)\, dx = 0 \tag{4}$$

$$\int x^2 k(x)\, dx > 0 \tag{5}$$

III-A. ***Primera propiedad*** $k(x) \geq 0$

Para verificar que el Kernel Gaussiano es siempre no negativo, consideramos los dos componentes principales de la función:

III-A1) ***El factor constante:*** El factor constante en el Kernel Gaussiano es $\frac{1}{\sqrt{2\pi}}$, que es positivo dado que $\sqrt{2\pi}$ es un número real positivo.

III-A2) ***El término exponencial:*** El término exponencial, $\exp\left(-\frac{x^2}{2}\right)$, es también siempre positivo. Esto se debe a las siguientes razones:

- La función exponencial e^t es siempre positiva para cualquier valor real t.
- El exponente $-\frac{x^2}{2}$ es negativo, ya que, x^2 (el cuadrado de cualquier número real) es siempre positivo y al multiplicarlo por -1 se convierte en negativo.
- Aunque el exponente es negativo, la exponencial de cualquier número no positivo es aún positiva, puesto que la función exponencial nunca toca ni cruza el eje x.

III-B. ***Segunda propiedad*** $\int k(x)\, dx = 1$

La integral del Kernel Gaussiano tiene la siguiente forma:

$$\int_{-\infty}^{\infty} K(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\, dx \tag{6}$$

III-B1) ***Propiedad de la Distribución Normal:*** La función $\exp\left(-\frac{x^2}{2}\right)$ dentro de la integral es la función de densidad de probabilidad de la distribución normal estándar:

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} I_1.$$

Calculemos el valor de I_1. Para ello, tengamos en cuenta lo siguiente:

$$I_1 I_1 = I_1^2 = \left(\int_0^{\infty} e^{-y^2/2}\, dy\right)\left(\int_0^{\infty} e^{-x^2/2}\, dx\right) = \int_0^{\infty}\int_0^{\infty} e^{-(x^2+y^2)/2}\, dx\, dy \quad (*)$$

Realizando un cambio a polares:

$$x = r\cos\theta \quad dx = \cos\theta\, dr - r\sin\theta\, d\theta$$

$$y = r\sin\theta \quad dy = \sin\theta\, dr + r\cos\theta\, d\theta$$

$$|J| = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

Resulta

$$(*) = \int_0^{\pi/2}\int_0^{\infty} re^{-r^2/2}\, dr\, d\theta = \int_0^{\pi/2}\left(\int_0^{\infty} re^{-r^2/2}\, dr\right) d\theta = \int_0^{\pi/2} 1\, d\theta = \int_0^{\pi/2} d\theta = [\theta]_0^{\pi/2} = \frac{\pi}{2}.$$

Por tanto, $I_1 = \sqrt{\frac{\pi}{2}} = \frac{\sqrt{2\pi}}{2}$, de donde, finalmente, se tiene que la integral buscada, $I = 1$.

III-C. ***Tercera propiedad*** $\int xk(x)\, dx = 0$

III-C1) ***Propiedades de Simetría y Antisimetría:*** El Kernel Gaussiano $K(x)$ es una función simétrica:

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

La función x es antisimétrica:

$$x = -(-x)$$

III-C2) ***Integral del Producto de Funciones Simétrica y Antisimétrica:*** Dado que $K(x)$ es simétrica y x es antisimétrica, el producto $x \cdot K(x)$ es antisimétrico. La integral de una función antisimétrica sobre un intervalo simétrico desde $-\infty$ a ∞ es cero:

$$\int_{-\infty}^{\infty} x \cdot K(x)\, dx = 0$$

Esto se puede ver dividiendo la integral en dos partes y utilizando la propiedad de antisimetría:

$$\int_{-\infty}^{0} x \cdot K(x)\, dx + \int_0^{\infty} x \cdot K(x)\, dx = 0$$

Cada parte es el negativo de la otra, por lo tanto, se cancelan mutuamente.

III-D. ***Cuarta propiedad*** $\int x^2 k(x)\, dx > 0$

La forma de la integral a evaluar es:

$$\int_{-\infty}^{\infty} x^2 \cdot K(x)\, dx = \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{7}$$

III-D1) Evaluación de la Integral: Por la demostración realizada en la propiedad sabemos que:

$$\int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1$$

También sabemos que x^2 es mayor que 0 para todo numero, entonces la ecuación $\int x^2 k(x)$ siempre es mayor que 0, cumpliendo la cuarta propiedad

IV. PUNTO 4

Para analizar las dos afirmaciones, es importante revisar las propiedades que debe cumplir un Kernel:

$$k(x) \geq 0 \tag{8}$$

$$\int k(x)\, dx = 1 \tag{9}$$

$$\int xk(x)\, dx = 0 \tag{10}$$

$$\int x^2 k(x)\, dx > 0 \tag{11}$$

a. La suma de Kernels es un Kernel:

Sean $k_1(x)$ y $k_2(x)$ dos kernels.

Condición (1): Por definición, se cumple que: $k_1(x) \geq 0$ y $k_2(x) \geq 0$. Por lo tanto, $k_1(x) + k_2(x) \geq 0$

Condición (2):

$$\begin{gathered} \int [k_1(x) + k_2(x)]\, dx = 1 \\ \int k_1(x)\, dx + \int k_2(x)\, dx = 1 \\ 1 + 1 = 1 \\ 2 \neq 1 \end{gathered}$$

Dado que esta condición no se cumple, el resultado de la suma de dos kernels no es un kernel, y por lo tanto, la afirmación es falsa. Habría que hacer ajustes de normalización que permitan modificar la integral de los kernels.

b. Cualquier Kernel puede ser utilizado para generar una regresión no paramétrica:

Según Wasserman (2006), para generar una regresión no paramétrica haciendo uso del estimador kernel Nadaraya-Watson, la elección de un kernel K no es tan importante, e incluso, estimaciones usando diferentes kernels suelen dar resultados numéricamente muy similares, siendo mucho más importante seleccionar correctamente el parámetro h que hace referencia al bandwidth. En otras palabras, si una función es, por definición, un kernel, teniendo en cuenta los requisitos mencionados previamente, puede se rutilizado para generar una regresión no paramétrica, y por lo tanto, la afirmación es verdadera.

V. PUNTO 5

Se aplican tres modelos de regresión donde la variable explicativa (Y) es la provincia de Alicante, y las variables regresoras(X) son las demás provincias del conjunto de datos.

V-A. Modelo Robusto: RLM

Se empleó un modelo de regresión lineal robusto (RLM) con la función de pérdida de Hampel para estimar los coeficientes del modelo.

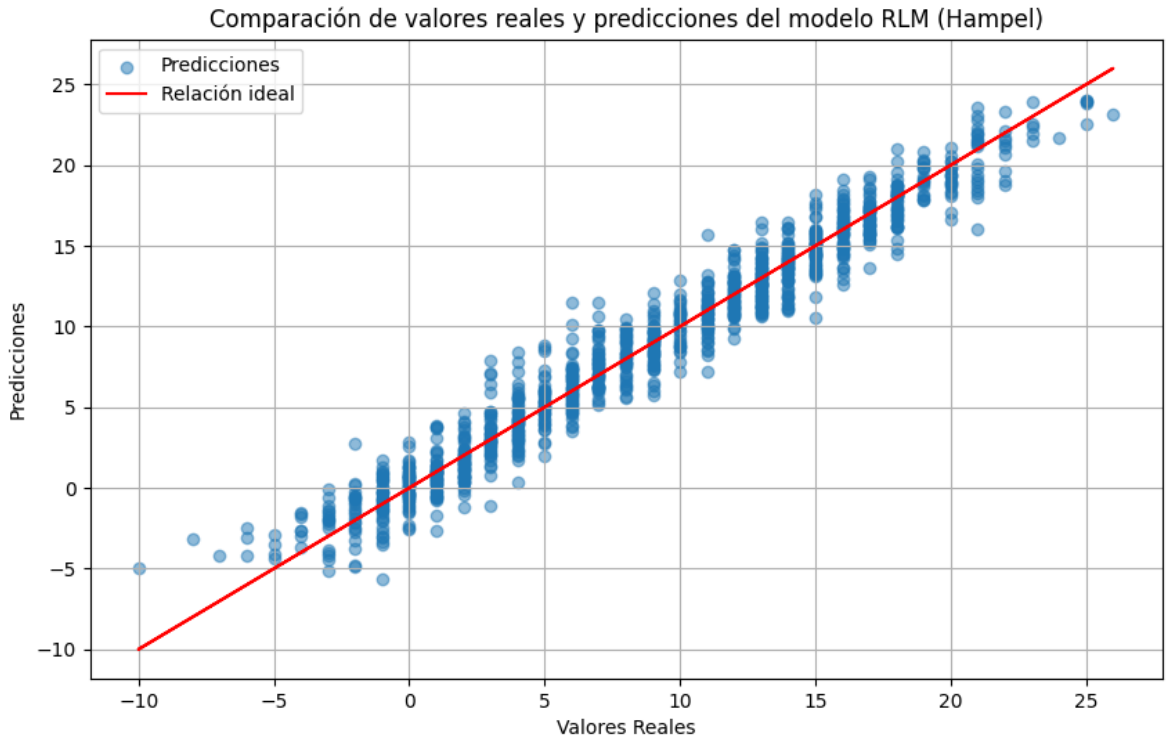

FIG. 4: Resultado modelo Robusto RLM

V-B. Modelo Robusto: RANSAC

Se utilizo el algoritmo RANSAC el cual es robusto a los outliers y es útil para ajustar modelos en presencia de ruido significativo.

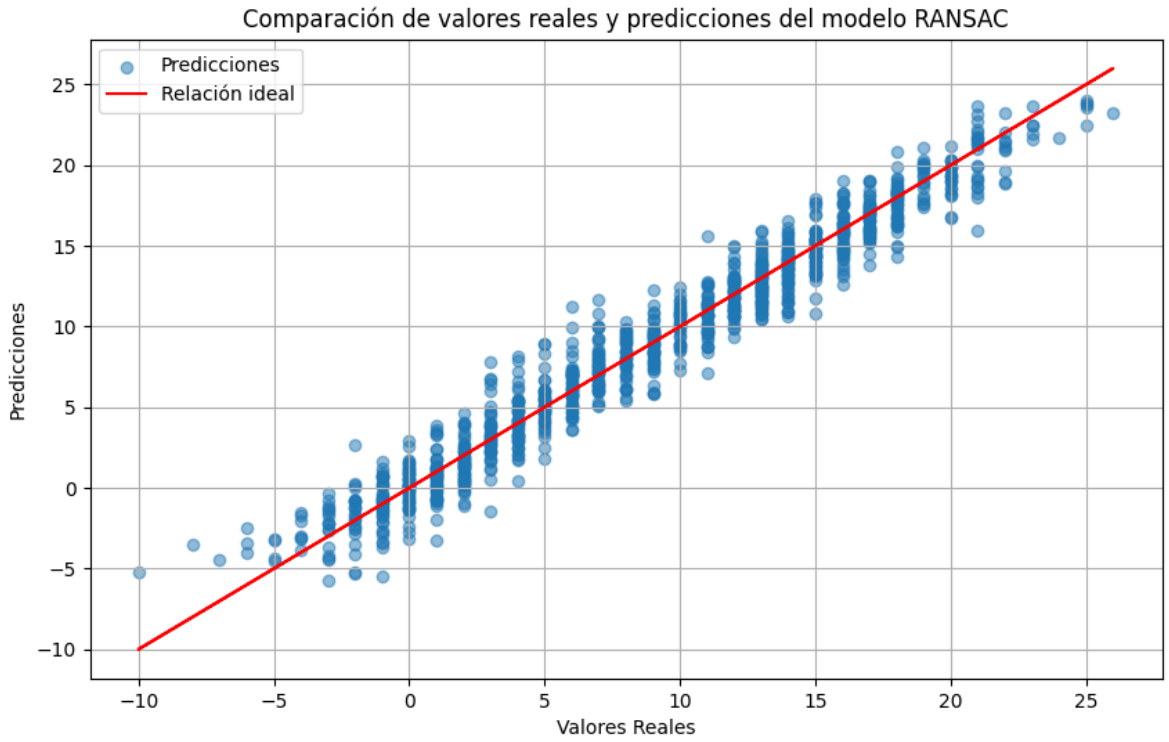

FIG. 5: Resultado modelo RANSAC

V-C. Modelo no Paramétrico: KernelReg

Se implemento el modelo Kernel Regression, esta técnica no paramétrica puede capturar relaciones no lineales entre variables. Utiliza un núcleo (kernel) para ponderar las observaciones en función de su distancia a un punto de interés.

Debido a su alto costo computacional, para este modelo se utilizaron los datos de temperatura de las dos provincias (X) más correlacionadas con Alicante (Y) y así calcular la temperatura de la provincia de Alicante.

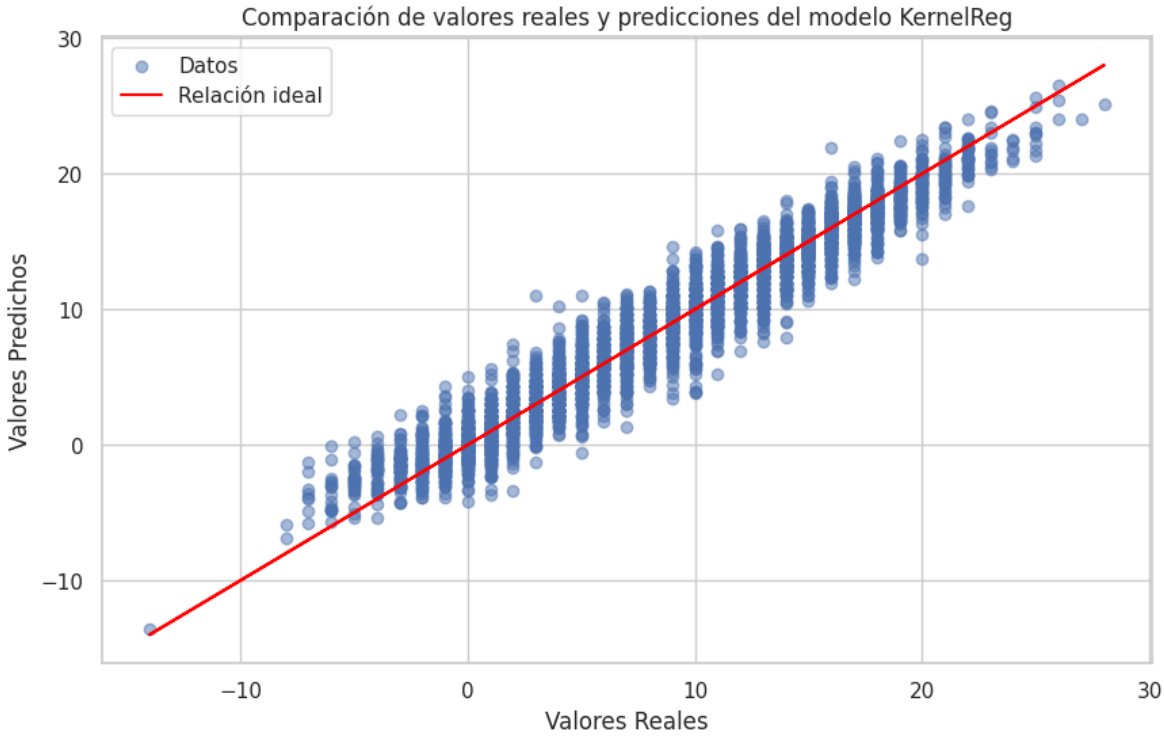

FIG. 6: Resultado modelo KernelReg

V-D. Tabla métricas de los Modelos

Modelo	**R^2**	**MSE**
RLM_Hampel	0.9565649760935544	2.0881983936109894
RANSAC	0.9557862333623158	2.1256375193246346
KernelReg	0.9428325306050978	2.693667214512897

TABLA. III: Resultados de diferentes modelos de regresión

Para predecir la temperatura de Alicate (Y) en función de las demás provincias (X), el modelo RLM con norma Hampel es el mejor en términos de precisión y capacidad de ajuste, seguido de cerca por RANSAC, mientras que Kernel Regression, aunque aún efectivo, muestra un rendimiento ligeramente inferior, teniendo en cuenta que por su costo computacional solo se utilizaron dos provincias(X).

V-E. Provincias más significativas

Se identifican las provincias (X) más relevantes para explicar la provincia de Alicante (Y) teniendo en cuenta la combinación de las provincias más significativas en la Correlación de Spearman y la de Pearson, esto permite capturar las relaciones más fuertes, tanto lineales como las no lineales.

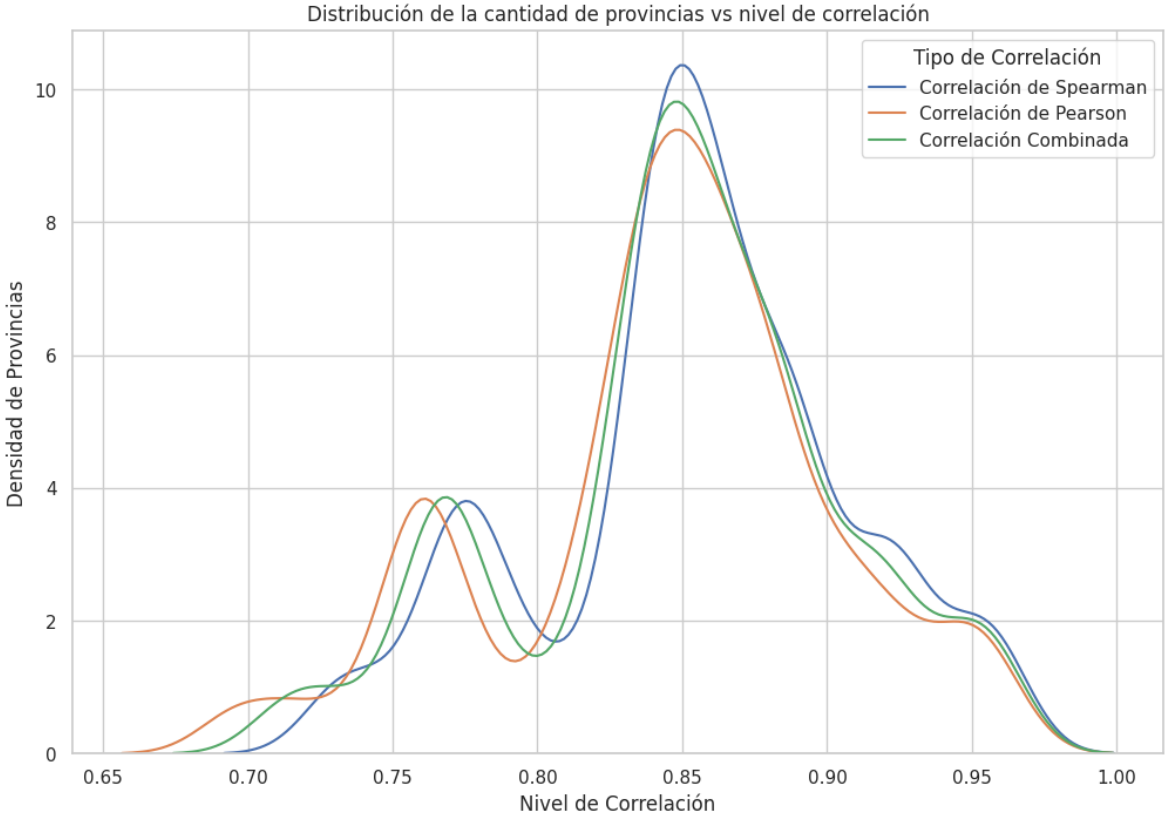

FIG. 7: Resultado Correlaciones

Se determina el umbral de 0.9 para seleccionar las provincias (X) que son más relevantes para explicar la provincia de Alicante (Y).

Provincia	**Correlación**
murcia36	0.960284
valencia53	0.952818
castellon16	0.948760
baleares11	0.933236
barcelona12	0.923415
tarragona49	0.917461
malaga34	0.913420
ceuta17	0.910598

TABLA. IV: Provincias con una correlación significativa superior a 0.9

VI. PUNTO 6

Sea la función definida por la siguiente expresión:

$$f(x,y) = \begin{cases} \frac{3}{16}(x^2+y^2) & \text{si } 0 < x < y < 2, \\ 0 & \text{E.O.C} \end{cases} \tag{12}$$

Para encontrar el modelo de regresión teórico se resuelve la siguientes ecuaciones:

$$f(y|x) = \frac{f(x,y)}{f(x)} \tag{13}$$

$$E[Y|X] = \int_x^2 y f(y|x) dy \tag{14}$$

Primero para resolver la ecuación 13 primero resolvemos la parte $f(x,y)$:

$$f(x) = \int_x^2 f(y|x) dy$$

La cual luego de las respectivas operaciones matemáticas da:

$$f(x) = -0{,}25x^3 + 0{,}375x^2 + 0{,}5 \tag{15}$$

Obteniendo la expresión dada por la ecuación 13, usando la ecuación 15 y la ecuación inicial 12

$$f(y|x) = \frac{0{,}1875x^2 + 0{,}1875y^2}{-0{,}25x^3 + 0{,}375x^2 + 0{,}5}$$

Ahora, la ecuación a calcular dada por la expresión 14

$$E[Y|X] = \int_x^2 y \frac{0{,}1875x^2 + 0{,}1875y^2}{-0{,}25x^3 + 0{,}375x^2 + 0{,}5} dy$$

La cual, luego de resolver la integral se obtiene la siguiente expresión

$$E[Y|X] = \frac{0{,}28125x^3 + 0{,}5625x^2 + 0{,}375x + 0{,}75}{0{,}5x^2 + 0{,}25x + 0{,}5} \tag{16}$$

La cual equivale al modelo de regresión teórico encontrado.

Una vez encontrado el modelo teórico, se hace una simulación de 1000 datos y se agrega un ruido blanco con una amplitud de 0.1. En la siguiente figura vemos la función limpia y la función con el ruido blanco:

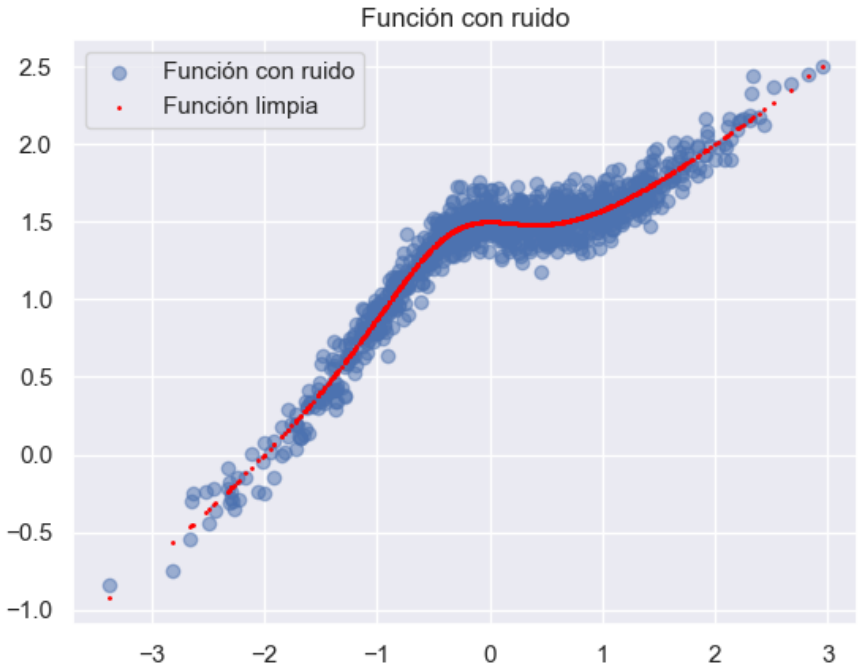

FIG. 8: Simulación de 1000 datos. Limpia y con ruido blanco

El objetivo es encontrar una función que describa mejor los datos. Para esto se usarán 3 modelos diferentes: Regresión Lineal tradicional, Modelado no parametrico y Modelado Robusto.

FIG. 9: Ajuste a los datos usando los diferentes modelos planteados

Se puede observar que la regresión no parametrica se adapta mejor a la naturaleza no lineal de la señal. Mientras que la regresión Robusta y la regresión tradicional no logran obtener

unos buenos resultados. Esto está sustentado por la siguiente tabla, donde están el r2 y RMSE de cada modelo, en el que el ajuste logrado por la regresión no paramétrica es considerablemente mejor:

Modelo	R2	RMSE
Regresion Lineal	0.78	0.04
Regresion No Parametrica	0.95	0.01
Regresión Robusta	0.67	0.07

VII. PUNTO 7

La regresión lineal es una técnica ampliamente utilizada para modelar la relación entre una variable dependiente y una o más variables independientes. El modelo de regresión lineal tradicional busca minimizar la suma de los cuadrados de los residuos, lo que hace que sea sensible a valores atípicos en los datos.

Para abordar este problema, en este trabajo se exploran métodos robustos para la estimación de parámetros en el modelo de regresión lineal. Se implementan dos enfoques robustos:

1. **Método Robusto 1**: Este método utiliza la mediana y la mediana absoluta de las desviaciones para calcular los coeficientes de la regresión lineal, lo que lo hace menos sensible a valores atípicos en los datos.

2. **Método Robusto 2**: En este método, se utiliza una medida de correlación robusta entre las variables predictoras y la variable de respuesta para calcular los coeficientes de la regresión. Esto ayuda a mitigar el efecto de valores atípicos en la estimación de los parámetros.

A continuación, se presentan las ecuaciones que definen estos métodos robustos:

1. **Método Robusto 1**:

$$\beta_1 = \frac{\text{comedian\_matrix}(X, y)}{\text{mad}(X)^2}$$
$$\beta_0 = \text{median}_y - \beta_1 \times \text{median}_x$$

2. **Método Robusto 2**:

$$\beta_1 = Corr_{kendall}(X, y)$$
$$\beta_0 = \text{median}_y - \beta_1 \times \text{median}_x$$

Donde:

- β_0 y β_1 son los coeficientes de la regresión lineal.

- X es la matriz de variables independientes.
- y es el vector de la variable dependiente.
- comedian_matrix(X, y) es la matriz de covarianza utilizando medianas.
- mad(X) es la mediana absoluta de las desviaciones de X.
- $Corr_{kendall}(X, y)$ es la medida de correlación robusta entre las variables X e Y.
- median$_x$ y median$_y$ son las medianas de X e y, respectivamente.

Estos métodos robustos se comparan con el enfoque tradicional de regresión lineal mediante la evaluación del error absoluto mediano y la raíz del error cuadrático medio en un conjunto de datos de temperatura para las provincias de España mas y menos correlacionadas entre si:

Se realizaron análisis de regresión lineal para las variables con mayor correlación entre ellas. Los resultados se presentan a continuación:

TABLA. V: Comparación de Resultados de Regresión No Robusta y Robusta

Método	β_0 (Intercepto)	β_1 (Pendiente)	Error Absoluto Mediano	Raíz del Error Medio Cuadrático
No Robusto	-3.1226	1.0286	0.9222	1.5865
Robusto (Método 1)	-8.4977	1.4641	4.0000	4.9155
Robusto (Método 2)	-1.0445	0.9317	1.3857	2.0613

Como se puede observar en la tabla VI, los resultados varían significativamente entre los diferentes métodos de regresión. En el caso de la regresión no robusta, se obtiene un coeficiente de pendiente de aproximadamente 1.0286, con un error absoluto mediano de 0.9222 y una raíz del error medio cuadrático de 1.5865. Por otro lado, los métodos robustos presentan coeficientes de pendiente y de intercepto diferentes, así como mayores errores absolutos medianos y raíces del error medio cuadrático.

En particular, el Método 1 de regresión robusta muestra un aumento significativo en los errores absolutos medianos y raíces del error medio cuadrático en comparación con la regresión no robusta. Esto sugiere que el Método 1 puede ser menos efectivo para modelar la relación entre las variables en presencia de valores atípicos. Por otro lado, el Método 2 de regresión robusta muestra una mejora en los errores absolutos medianos y raíces del error medio cuadrático en comparación con el Método 1, aunque todavía son mayores que los obtenidos con la regresión no robusta. Esto indica que el Método 2 puede proporcionar una mejor robustez a los valores atípicos en los datos.

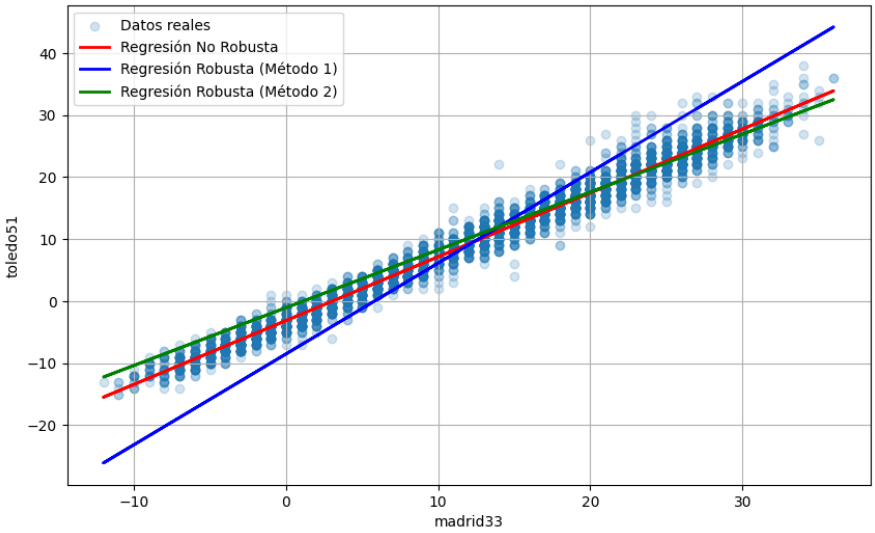

FIG. 10: Simple Regression Vs Robust regressions (Toledo Vs Madrid)

Se realizaron análisis de regresión lineal para las variables con menor correlación entre ellas. Los resultados se presentan a continuación:

TABLA. VI: Comparación de Resultados de Regresión No Robusta y Robusta

Método	β_0 (Intercepto)	β_1 (Pendiente)	Error Absoluto Mediano	Raíz del Error Medio Cuadrático
No Robusto	2.1501	0.3158	1.8349	2.8569
Robusto (Método 1)	-2.4808	0.6165	2.1504	3.5890
Robusto (Método 2)	-0.4319	0.4960	1.9840	3.1649

Como se puede observar en la tabla VI, los resultados varían entre los diferentes métodos de regresión, aunque en menor medida que en el caso de las variables con mayor correlación. En este caso, la regresión no robusta y los métodos robustos muestran coeficientes de pendiente y de intercepto que difieren entre sí, así como diferentes niveles de error absoluto mediano y raíz del error medio cuadrático.

La regresión no robusta presenta un coeficiente de pendiente de aproximadamente 0.3158, con un error absoluto mediano de 1.8349 y una raíz del error medio cuadrático de 2.8569. Los métodos robustos, por otro lado, muestran coeficientes de pendiente y de intercepto diferentes, así como niveles de error absoluto mediano y raíz del error medio cuadrático que también difieren entre sí.

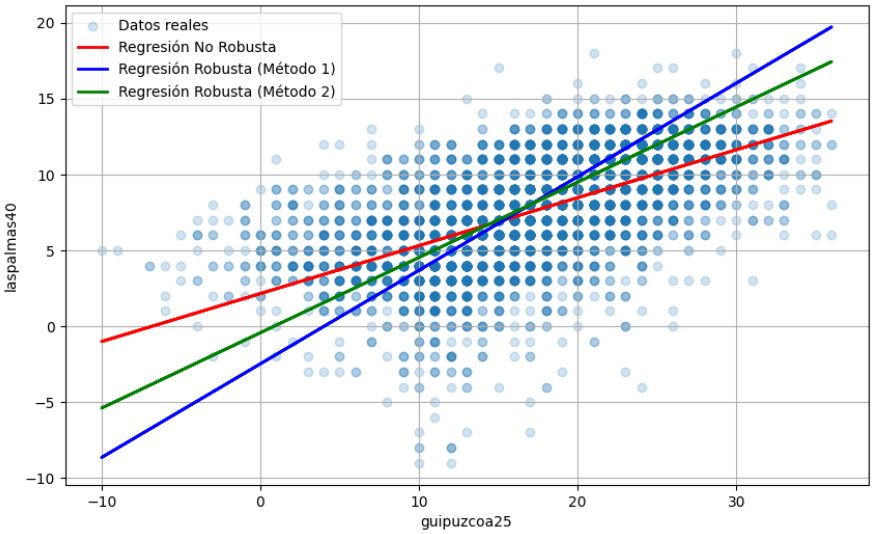

FIG. 11: Simple Regression Vs Robust regressions (Las Palmas Vs Guipuscoa)

VIII. PUNTO 8

Para las provincias de Las Palmas y Guipúzcoa se realizó el test de Mann-Whitney para determinar si hay diferencias significativas entre las distribuciones de dos grupos para cada una de las otras provincias españolas.

VIII-A. Resultados del test Mann-Whitney

Las Palmas:

TABLA. VII: Sin Diferencias Significativas - Las Palmas

	Stat	P-valor
cordoba18	15236684.5	0.169994

Guipúzcoa:

TABLA. VIII: Sin Diferencias Significativas - Guipúzcoa

	Stat	P-valor
avila9	15148837.5	0.400528
burgos13	15330580.0	0.052508
soria48	14842162.0	0.311277

Para las palmas se determina que la mayoría de las provincias muestran una p-valor menor a 0.05, lo que indica que hay diferencias estadísticamente significativas en las distribuciones de temperatura entre Las Palmas y estas provincias. En el caso de Córdoba (con un p-valor de 0.169994), no se encontraron diferencias significativas. Lo que indica que las distribuciones de temperatura entre Las Palmas y Córdoba pueden ser similares.

Con repecto a Guipúzcoa también la mayoría de las provincias muestran p-valores significativos, lo que indica diferencias notables en las distribuciones de temperatura comparadas con Guipúzcoa. Las provincias de Avila, Burgos y Soria presentan p-valores superiores a 0.05, indicando que no hay evidencias suficientes para afirmar que existen diferencias significativas en las temperaturas comparadas con Guipúzcoa. El test sugiere que las condiciones de temperatura en estas provincias son comparables a las de Guipúzcoa.

IX. PUNTO 9

Después de adaptar el dataset como datos funcionales, se obtuvo la siguiente gráfica:

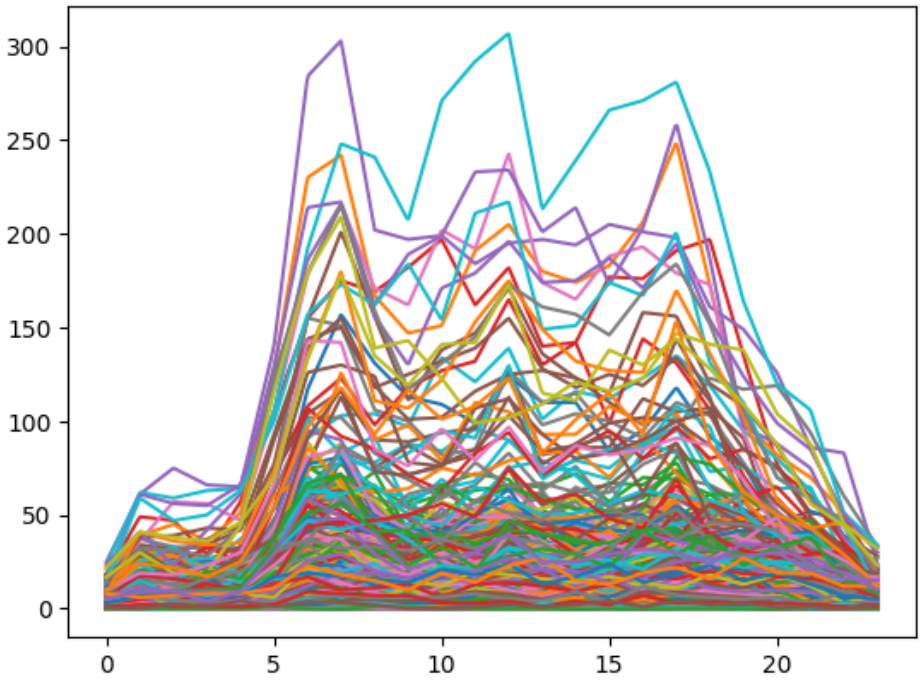

FIG. 12: Curvas de datos de accidentalidad por barrio

Se optó por utilizar la medida de Fraiman Muniz para identificar las profundidad de cada barrio. En la siguiente gráfica, se puede apreciar la curva más profunda con base en esta medida:

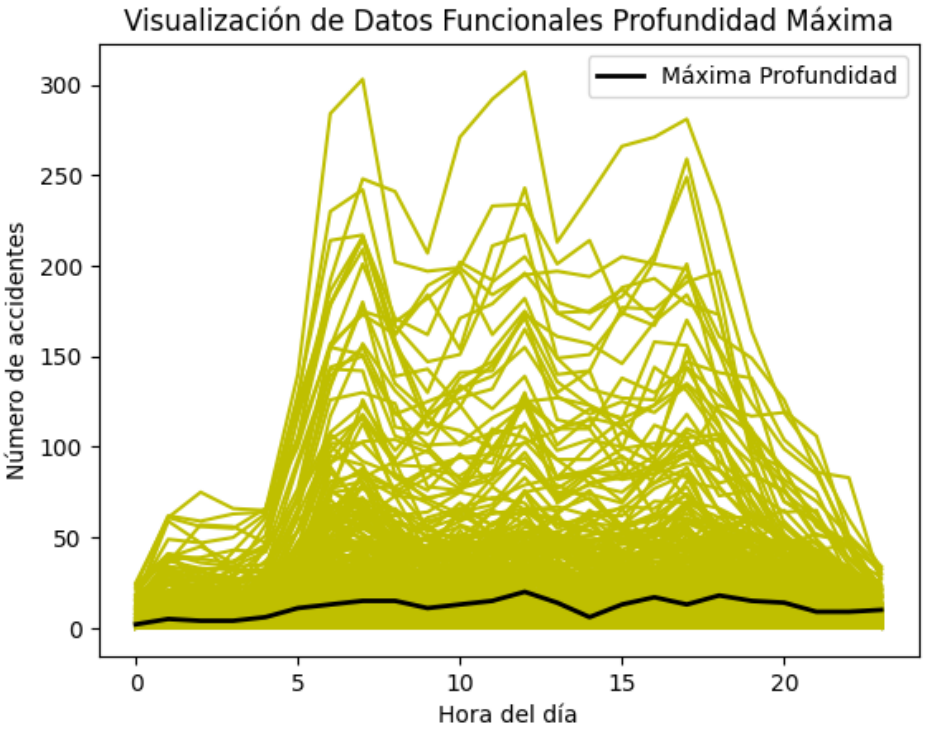

FIG. 13: Máxima profundidad de Fraiman Muniz

Tras organizar en orden descendente los barrios en función de su profundidad, se obtuvo que los barrios más profundos, es decir, con valores más altos, son los siguientes:

TABLA. IX: Top 10 Barrios más profundos

BARRIO	PROFUNDIDAD
MoscuNo1	0.94798
LasMercedes	0.940114
LaPradera	0.938337
Granizal	0.938108
SimonBolivar	0.937087
CalasanzParteAlta	0.936594
SantaMonica	0.936181
VilladelSocorro	0.934886
BosquesdeSanPablo	0.934496
NuevaVilladelaIguana	0.931079

Tiene sentido que sean barrios profundos, pues es de esperar que no tengan mayores particularidades en cuanto a accidentalidad. En otras palabras, son barrios que no destacan en términos de accidentalidad, ni para bien, ni para mal.

X. PUNTO 10

Dado que en el punto anterior se calculó la profundidad con la medida de Fraiman Muniz, a partir de esto se buscó encontrar el 5 % de outliers. Para hacerlo, se organizaron los datos en orden ascendente con base en la profundidad, y luego, se encontró el 5 % de datos con menor profundidad, es decir, los outliers, presentados a continuación:

TABLA. X: Outliers

BARRIO	PROFUNDIDAD
LaCandelaria	0.502477
Caribe	0.505504
LosConquistadores	0.513702
PerpetuoSocorro	0.515100
CampoAmor	0.515146
BarrioColon	0.522209
SanBenito	0.523516
Guayaquil	0.528538
SantaFe	0.530144
VillaNueva	0.532827
CarlosERestrepo	0.533595
SanDiego	0.537929
TerminaldeTransporte	0.539030
Naranjal	0.542653
Castilla	0.545496

Dentro de lo que conocemos de la planificación territorial de Medellín, sí tiene sentido que estos barrios sean outliers a nivel de accidentalidad. Como se puede ver en la figura 14, estos barrios más alejados de la curva más profunda son aquellos con mayor accidentalidad. Esto tiene sentido especialmente en zonas de alto flujo vehicular, como La Candelaria, Caribe, San Diego, Terminal de Transporte, Guayaquil y Santa Fe. Las otras, se pueden explicar por las grandes vías que pasan por estos barrios, como pasa, por ejemplo, con Los Conquistadores, que limita con la calle San Juan y la Carrera 65, Castilla, que también limita con la Carrera 65 y la regional, o Carlos E. Restrepo, que limita con la avenida regional y la calle Colombia, donde es esperable que haya un alto número de accidentes.

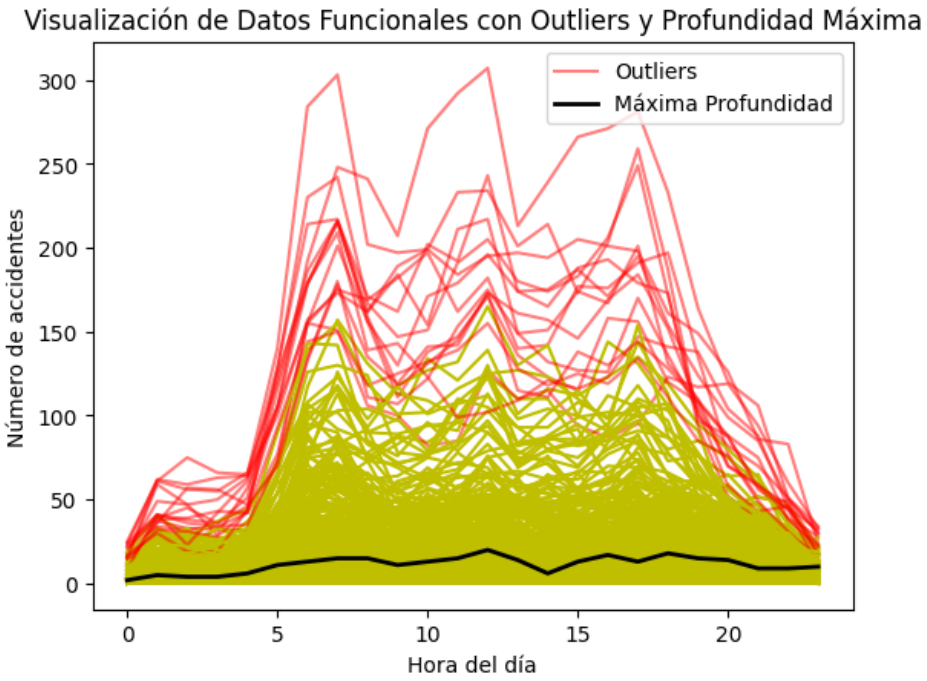

FIG. 14: Comparación entre máxima profundidad de Fraiman Muniz con los outliers

XI. PUNTO 11

Felipe y Daniel

Functional boxplot fue el nombre propuesto por Hyndman and Shan en el 2010 [1] para nombrar un nuevo metodo de graficación para datos funcionales. Este metodo es una variacion del boxplot tradicional. Este está construido usando una estimación de densidad de kernel bivariante $\hat{f}(z)$, el cual está definido por:

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} K_h(z - Z_i) \tag{17}$$

Donde Z_i representa un set de puntos, $K_h(\cdot) = \frac{K(\cdot/h_i)}{h_i}$, donde K es la función kernel y h_i es el ancho de banda para la dimensión i. Usando el estimado del kernel de densidad, una region de alta densidad está definida por:

$$R_\alpha = \{z : \hat{f}(z) \geq f_\alpha\}, \tag{18}$$

Donde f_α es tal que $\int_{R_\alpha} \hat{f}(z)\, dz = 1-\alpha$, o sea, es la region de covertura de la probabilidad $1-\alpha$, donde todos los puntos dentro de la región tienen una estimación de densidad mas alta que cualquiera de los puntos fuera de esta región, esto le da el nombre de region de alta densidad. Para una densidad bivariante, las regiones de alta densidad pueden considerarse

como contornos, con una cobertura cada vez mayor a medida que α disminuye.

En el caso funcional, un boxplot no se obtiene de los boxplots puntuales en cada punto t. Básicamente, se encuentra la mediana funcional, y luego el rango intercuantil del conjunto de datos que representa el 50 % de los mismos (región central). Para los bigotes, usualmente se extiende la región central 1.5 veces su tamaño, y los datos que se encuentren por fuera de estos son considerados outliers. Para el conjunto de datos de accidentes en Medellín, se utilizó la función ”Boxplot”de la librería *scikit-fda*, obteniendo la siguiente gráfica:

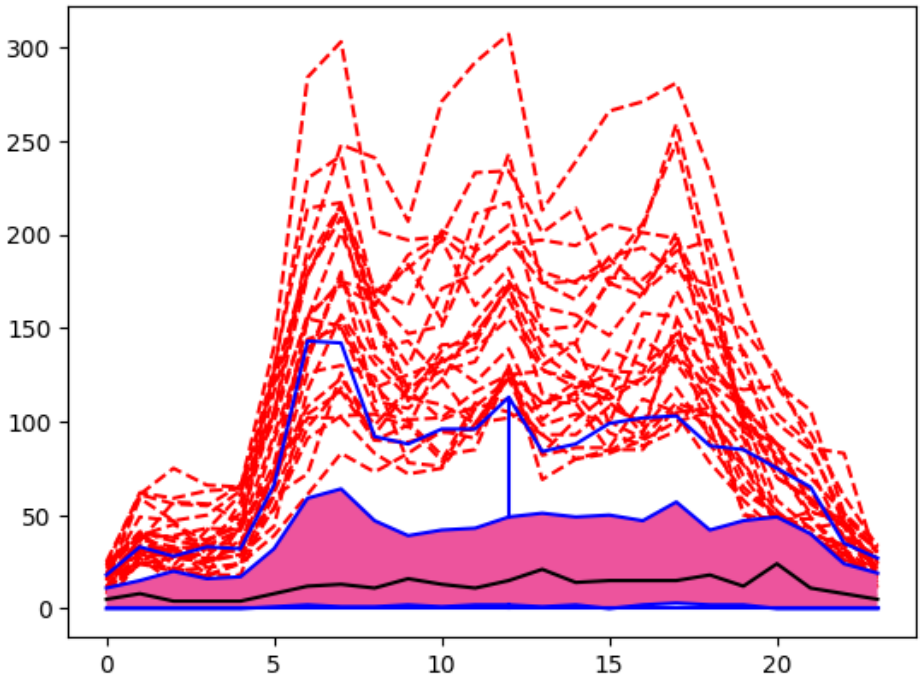

FIG. 15: Boxplot Funcional

Como se puede apreciar, la región central está representada en color morado, delimitada por líneas azules. En ella, se muestra en color negro la mediana funcional. Los bigotes, son representados por otras líneas azules, y los outliers, se muestran en color rojo. Siguiendo la lógica de cómo se calculan los outliers con este método, se obtuvo la siguiente tabla:

TABLA. XI: Outliers según el Boxplot

BARRIO			
BarrioColon	**Castilla**	LaAguacatala	Prado
Belen	CorazondeJesus	**LaCandelaria**	Rosales
Boston	CristoRey	LasAcacias	**SanBenito**
CabeceraSanAntoniodePrado	ElChagualo	LasGranjas	**SanDiego**
CalleNueva	ElEstadio	Laureles	**SantaFe**
CampoAmor	ElPoblado	LosColores	Suramericana
CampoValdesNo1	ElProgreso	**LosConquistadores**	**TerminaldeTransporte**
CampoValdesNo2	Guayabal	Manila	UniversidadNacional
Caribe	**Guayaquil**	**Naranjal**	VillaCarlota
CarlosERestrepo	JesusNazareno	**PerpetuoSocorro**	**VillaNueva**

El método boxplot funcional encontró más outliers que el utilizado con la medida de profundidad de Fraiman Muniz en el punto 10. Comparando con la Tabla X, la Tabla XI también incluye los barrios **LaCandelaria, Caribe, LosConquistadores, PerpetuoSocorro, CampoAmor, BarrioColon, SanBenito, Guayaquil, SantaFe, VillaNueva, CarlosERestrepo, SanDiego, TerminaldeTransporte, Naranjal y Castilla**, es decir, en ambos métodos estos barrios son encontrados como Outliers por su alta accidentalidad.

REFERENCIAS

[1] R. J. Hyndman y H. L. Shang, «Rainbow Plots, Bagplots and Boxplots for Functional Data,» *Journal of Computational and Graphical Statistics*, vol. 18, n.º 2, págs. 335-347, jun. de 2009. DOI: 10.1198/jcgs.2009.07098.