UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Enero 2020

Estimación e imputación de datos faltantes mediante métodos de interpolación espacial para precipitación mensual acumulada en el departamento de Antioquia durante el periodo 2014-2018.

Estimation and imputation of missing values using spatial interpolation methods for accumulated monthly precipitation in the department of Antioquia during the 2014-2018 period.

Autor: Leandro Sánchez Quiroga[a]
leandro.sanchez@usantotomas.edu.co

Director: Wilmer Pineda Rios[b]
wilmwepineda@usantotomas.edu.co

Resumen

Los valores faltantes son comunes en las bases de datos que trabajamos a diario, el saber que hacer con esos datos faltantes es fundamental, en algunas ocasiones la solución inmediata es quitar los registros y perder información que puede ser de gran valor, el propósito es aprovechar la información que se tiene disponible en la mayor posibilidad, para ello se disponen de varias técnicas que permiten imputar de manera eficiente los valores faltantes. El presente trabajo busca contrastar métodos de interpolación espacial en la imputación de valores faltantes de precipitación mensual acumulada en el departamento de Antioquia durante el periodo 2014-2018, con datos suministrados por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), para ello se utiliza la distancia inversa ponderada(IDW, por sus siglas en inglés), Spline de placa delgada (TPS, por sus siglas en inglés) y kriging ordinario, como métodos evaluados a través de la raíz del error cuadrático medio(RMSE, por sus siglas en ingles). El Kriging ordinario fue efectivo cunado se tiene mas del 10 % de los valores faltantes, el método de la distancia inversa ponderada fue el que arrojó mejores resultados cuando se tienen 5 % de los valores faltantes. Se aplicaron los resultados obtenidos a los datos correspondientes al año 2018..

***Palabras clave*:** Valores faltantes, interpolación espacial, precipitación, IDW, TPS, Kriging.

Abstract

Missing values are common in the databases we work on a daily basis, knowing what to do with those missing data is essential, sometimes the immediate solution is to remove the records and lose information that can be of great value, the purpose is to take advantage of the information that is available in the greatest way, for this objective, there are several techniques that allow the efficient imputation of the missing values. This paper seeks to contrast methods of spatial interpolation in the imputation of missing values of accumulated monthly precipitation in the department of Antioquia during the 2014-2018 period, with data provided by the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM), to achieve this goal, weighted inverse distance (IDW), thin plate spline (SPT) and ordinary kriging are used as evaluated methods through the root of the mean square error (RMSE). Ordinary Kriging was effective when you have more than 10 % of the missing values, the weighted inverse distance method was the one that yielded the best results when you have 5 % of the missing values. The results obtained were applied to the data corresponding to the year 2018.

***Keywords*:** Missing values, spatial interpolation, precipitation, IDW, TPS, Kriging.

[a]Estudiante. Facultad de Estadística, Universidad Santo Tomás
[b]Docente. Facultad de Estadística, Universidad Santo Tomás

1. Introducción

Durante las últimas décadas se han presentado una gran variedad de métodos que permiten realizar el procedimiento de identificación, depuración e imputación de valores faltantes en datos georreferenciados; es de suma importancia la completitud de las observaciones para asegurar buenas predicciones y elegir el modelo que mejor representa los datos. En la mayoría de los casos al ser datos tratados y obtenidos de manera automatizada y remota, es común encontrar valores faltantes en estos, que no se pueden recuperar de alguna manera, adicional existen fallas en la recolección de datos que pueden ser errados o salidos de contexto para el estudio realizado y que por ende deben ser imputados, de dicho tratamiento depende que los pronósticos sean acertados. Castaño (2007) afirma "Los métodos de imputación de valores faltantes suponen conocimiento del modelo o que los datos sean tales que un subconjunto de ellos permita identificar el modelo" (p.260).

En particular para la imputación de valores faltantes en datos georreferenciados es posible emplear varias metodologías que se aplican en campos ambientales, climatológicos y de las ciencias naturales, entre ellos valores de precipitación, obteniendo diversos resultados en cuanto precisión y procesamiento computacional. Existen diferentes perspectivas que permiten abordar la imputación de valores faltantes en datos georreferenciados, desde una metodología simple como los promedios aritméticos, por interpolación espacial, hasta modelos estructurados de orden estocástico, bayesiano y satelital que permiten evaluar que tan acertadas pueden ser las predicciones en una estructura de datos.

En Londoño (2015) se usó la interpolación espacial IDW (distancia inversa ponderada, por sus siglas en ingles) y se comparó con el método e interpolación Spline a través de sistemas de información geográfica para la imputación de datos faltantes en partículas contaminantes del aire en la ciudad de Medellín, Colombia, se llegó a la conclusión que la interpolación Spline arrojó mejores resultados en pruebas de validación cruzada.

En Salgado y Largo (2018) se usaron métodos geoestadísticos para la imputación de datos faltantes en la temperatura media del departamento del Valle del Cauca, Colombia, en el cual se compararon varias metodologías y se llegó a la conclusión que para estos datos, modelos de regresión lineal con estructura de correlación espacial en los residuales, aumentan la calidad de la predicción y por ende la imputación que se realizó, se recomendó usar Kriging como alternativa de imputación.

En Cruz y Barrios (2018) se realizó la estimación de datos faltantes para la lluvia mensual en Colombia, con métodos geoestadísticos para dicha imputación, se usó Krigeado Ordinario y Cokrigeado Ordinario, se llegó a la conclusión que el Cokrigeado estima con mejores aproximaciones los datos faltantes, esto con el uso de variables auxiliares como la altitud y datos de lluvia satelital.

Al abordar una estructura de datos georreferenciados en un escenario ideal, se asume que los datos están completos, pero difícilmente sucede, lo cual despierta un interés por evaluar y comparar diferentes metodologías de interpolación espacial para la imputación de valores faltantes en datos georreferenciados y así mejorar las predicciones futuras. En la presente investigación se quiere abordar y contrastar metodologías propuestas para la imputación de valores faltantes en datos georreferenciados, en específico; métodos de interpolación espacial que se puedan aplicar a la precipitación mensual acumulada de Antioquia, evaluar su eficiencia a través de simulaciones computacionales mediante la raíz del error cuadrático medio y sugerir el método de mejor comportamiento para luego aplicar los resultados obtenidos.

Este trabajo es de carácter cuantitativo experimental y el diseño de la investigación es de visualización, estimativo y predictivo. La población a abordar son los datos de precipitación mensual acumulada en el departamento de Antioquia para los periodos 2014-2018 y el sujeto son los valores faltantes en dicha estructura de datos, se quiere aplicar métodos de interpolación espacial para la imputación de datos faltantes, evaluar su eficiencia a través de simulaciones computacionales cuando existen 5 %, 10 %, 15 % y 20 % de valores perdidos para los periodos 2014-2017, como criterio de evaluación y validación entre las metodologías de imputación usadas, se propone la raíz del error cuadrático medio(RMSE, por sus siglas en inglés), los resultados obtenidos se usaran para imputar los valores faltantes correspondientes al año 2018. Para lograr lo mencionado es necesario usar el software *R*, se utilizará para realizar simulaciones

con datos faltantes y de allí realizar predicciones que se ajusten a los datos estudiados.

En la primera parte del trabajo se muestra la definición de precipitación y el área de interés de estudio, la siguiente parte inicia con la definición de valores faltantes, una pequeña reseña y la evolución de la imputación de valores faltantes en general, luego se profundiza con definiciones y generalidades de la geoestadística y los datos georreferenciados, para dar paso a las metodologías que se usaran para la validación de la efectividad en la imputación de los valores faltantes, por último se muestran los resultados obtenidos, la aplicación de dichos resultados y las conclusiones finales.

2. Área de interés y estaciones meteorológicas.

2.1. Precipitación.

Según el IDEAM la precipitación "es la caída de partículas de agua líquida o sólida que se originan en una nube, atraviesan la atmósfera y llegan al suelo. La cantidad de precipitación es el volumen de agua lluvia que pasa a través de una superficie en un tiempo determinado", el instrumento utilizado para recolectar la información es el pluviómetro y la medida es en milímetros.

2.2. Ubicación.

Antioquia está ubicada al noroeste de Colombia, sus puntos extremos son: al sur Cerro de Caramanta (5° 25' 30"), al norte Punta de Arboletes(8° 55' 00"), al este Frente a Barrancabermeja (0° 11' 30") y al oeste límite norte con Chocó (Puerto López) (3° 09' 00"), la altura máxima es de 4080 metros sobre el nivel del mar según el Departamento Administrativo de Planeación de la gobernación de Antioquia. El análisis se realizó en 151 estaciones meteorológicas activas del departamento de Antioquia administradas por el IDEAM que tienen información de precipitación mensual acumulada en el periodo 2014-2018 (Figura 1.) y que se distribuyen como se muestra en la tabla 1:

Tabla 1: Latitud y longitud máximas y mínimas estaciones meteorológicas.Fuente: Elaboración propia

Antioquia	Mínima	Máxima
Latitud	5.547833	8.846944
Longitud	-76.85527	-73.94416

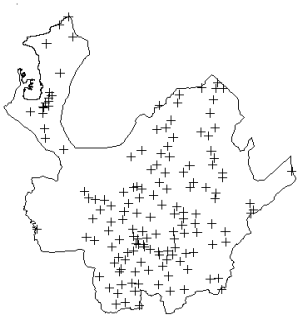

Figura 1: Distribución estaciones meteorológicas. Fuente: Elaboración propia

3. Datos Faltantes.

En Medina (2007) se define un dato faltante como una observación que simplemente está ausente en los datos recolectados o también como observaciones aberrantes o poco probables según el contexto de los datos.

La imputación de datos se aborda desde diferentes metodologías, dependiendo de la estructura de los datos pueden adoptar un modelo que asigne estimaciones a valores faltantes, en la década de los sesenta los modelos Hot-Deck y Cold-Deck se utilizaban de manera significativa para suplir información en encuestas y censos, Rubin en 1976 aplicó una metodología de inferencia estadística donde los datos faltantes se asumían como variables aleatorias y los datos eran imputados sin ajustar modelos, más tarde Rubin en 1987 introdujo el termino de imputación múltiple, planteando que cada dato faltante debe imputarse luego de una serie de simulaciones (como se cita en Medina, 2007).

Se dispone de una clasificación para el reconocimiento de los tipos de datos faltantes:

- MCAR (Missing Completely At Random): La probabilidad de que una respuesta sea un dato faltante es independiente a los valores de las variables observadas y no observadas del conjunto de datos.
- MAR (Missing At Random): La probabilidad de que una respuesta sea un dato faltante es independiente al valor de la variable observada, pero es dependiente de otras variables no observadas del conjunto de datos.
- NMAR (Not Missing At Random): La probabilidad de que una respuesta sea un dato faltante es dependiente a los valores de las variables del conjunto de datos.

3.1. Imputación simple:

Consiste en asignar un único valor a cada dato faltante y así generar una base de datos completa, dentro de estos métodos resaltan los métodos de interpolación espacial y los que se muestran a continuación:

3.1.1. Imputación por media no condicional:

Consiste en asignar el promedio de la totalidad de los datos a los valores faltantes, este método no afecta el promedio, pero si afecta la variabilidad, el sesgo y los percentiles.

3.1.2. Imputación Cold Deck:

Este método consiste en utilizar datos históricos, es decir; se toman de observaciones con características similares al dato faltante pero provenientes de un estudio o medición anterior, el éxito de este método depende de la calidad de información que se pueda obtener.

3.1.3. Imputación Hot Deck:

Este método consiste en duplicar un valor ya existente en las observaciones por el dato faltante, es un proceso de duplicación, su propósito es reducir el sesgo.

3.1.4. Imputación por regresión:

Para este método se ajusta un modelo de regresión para predecir el valor de los valores faltantes a través de covariables que se encuentren correlacionadas con la variable de interés.

Sea Y la variable interés y $X = (X_1, \ldots .., X_k)$ las covariables, el modelo general es el siguiente:

$$g[E(Y)] = X\beta, \qquad Y \sim F \tag{1}$$

Donde g es la función y F la distribución de Y, Si Y es una variable continua:

$$E(Y) = X\beta, \qquad Y \sim Normal \tag{2}$$

3.2. Imputación múltiple Markov Chain Monte Carlo (MCMC):

Consiste en asignar múltiples valores a cada dato faltante, generando la misma cantidad de conjuntos de datos completos para luego combinar los resultados obtenidos. Es uno de los métodos más usados para estimación de datos faltantes, se basa en una simulación mediante cadenas de Markov y utilizando inferencia bayesiana se realizan estimaciones paramétricas, asumiendo que las observaciones provienen de una distribución normal multivariada realizando iteraciones de la siguiente manera:

- Se realiza el proceso de imputación para cada una de las observaciones independientemente, usando la matriz de covarianzas y el vector de medias
- Con los datos obtenidos en el paso anterior se realizan de nuevo las estimaciones con el vector de medias y la matriz de covarianzas de los datos completos.

4. Estadística espacial y datos georreferenciados.

En Giraldo (2002) se define la estadística espacial como una reunión de metodologías que se usan para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios de una región, a su vez los datos georreferenciados son mediciones que están asociadas a las coordenadas de los sitios en que fueron tomadas y una variable regionalizada es una medición en el espacio que muestra una estructura de correlación. El principal objetivo de la geoestadística es la simulación, estimación y predicción de fenómenos espaciales.

4.1. Variograma y Semivariograma:

En un análisis geoestadístico el primer paso es establecer si los datos asociados a una variable aleatoria presentan una dependencia espacial, el variograma es una función en la que se determina si la varianza de la variable regionalizada es finita en una estructura de datos, dicha función está dada por: Giraldo (2002)

$$2\gamma(h) = V(Z(x+h) - Z(x)) = E((Z(x+h) - Z(x))^2) \tag{3}$$

La función de semivarianza muestra la dependencia espacial, dicha función es estimada a través del semivariograma experimental y se obtiene mediante:

$$\bar{\gamma}(h) = \frac{\sum (Z(x+h) - Z(x))^2}{2n} \tag{4}$$

h es la distancia que separa un valor de la variable de interés con otro valor en un sitio x, y n es el número de puntos emparejados que están separados por la distancia h. El semivariograma experimental parte del principio de que existe correlación espacial a menor distancia entre los sitios y por ende existe mayor semejanza.

4.2. Covariograma y Correlograma:

Siguiendo el mismo principio del semivariograma las funciones de Covariograma y Correlograma se definen como:

$$C(h) = \frac{\sum_{i=1}^{n}(Z(x+h)\cdot Z(x))}{n} - m^2 \tag{5}$$

$$r(h) = \frac{C(h)}{S_x^2} = \frac{C(h)}{C(0)} \tag{6}$$

m es el promedio de todos los puntos en la estructura de datos. En Giraldo(2002) se afirma que tanto el semivariograma como el covariograma y correlograma son empleados para establecer la dependencia espacial, pero en la practica la más usada es la función de semivarianza ya que no necesita de la estimación de parámetros.

4.3. Modelos de semivarianza.

Al establecer una dependencia espacial en una estructura de datos georreferenciados o de variables regionalizadas mediante el semivariograma, es necesario ajustar un modelo teórico que generalice lo observado en cuanto estructura de correlación en los datos, para ello, en Giraldo(2002) se definen dos tipos de modelos: modelos acotados (gaussiano, esférico, exponencial) y modelos no acotados (logarítmico, lineal, potencial) , en los dos grupos se tienen tres parámetros como se muestra en la figura 2:

- Pepita (C_0): Se define como una interrupción del semivariograma en el origen.
- Meseta ($C_0 + C_1$): Representa el límite del semivariograma cuando la distancia h tiende a infinito, se pude decir que es la cota superior.
- Rango (a): Es la zona de influencia, representa la distancia en la cual las observaciones son independientes.

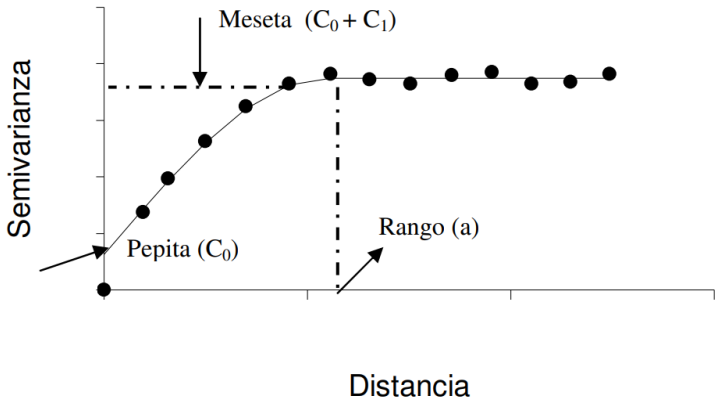

Figura 2: Parámetros semivariograma: Fuente (Giraldo 2002, p.25)

5. Distancia inversa ponderada (IDW, por sus siglas en inglés).

La distancia inversa ponderada es un método de interpolación determinista que actúa como un promedio ponderado, allí la influencia de un punto con respecto a otros, se ve afectada por su distancia y a partir de la asignación de pesos se acepta que el comportamiento de un punto se asemeja a los puntos más próximos. (Roshan & Lulu, 2011)

El predictor IDW está dado por:

$$\hat{y}(x) = \sum_{k=1}^{n} v_k(x) y_k, \tag{7}$$

Donde

$$v_k(x) = \frac{w_k(x)}{\sum_{i=1}^{n} w_i(x)} \tag{8}$$

para $x \notin \{x_1, ..., x_n\}$, $v_k(x_i) = 1$ si $i = k$ y 0 en otro caso

Ahora

$$w_i(x) = 1/d^2(x, x_i), \tag{9}$$

Donde $d(x, x_i) = \sqrt{\sum_{j=1}^{p}(x_j - x_{ij})^2}$ es la distancia Euclidiana de x a x_i

6. Spline de placa delgada (TPS, por sus siglas en inglés).

La interpolación Spline de palca delgada al igual que la interpolación IDW es un método determinista, lo que se busca es ajustar una superficie que pase por todos los puntos conocidos y que a su vez suavice la curvatura de dicha superficie, se usan funciones polinómicas por tramos, cada una de ellas acorde al intervalo que representa, en lugar de usar una función general para todo el intervalo de puntos. (Hancock Hutchinson, 2005)

Sea: $(z_i, x_{1i}, x_{2i}, \ldots, x_{1di})$ con z una variable dependiente y d un conjunto de predictores de variables $x_1, \ldots, x_d$.

$$z_i = g(x_{1i}, x_{2i}, ..., x_{1di}) + \epsilon_i, \tag{10}$$

$i = 1, ..., n$ donde n es el número de observaciones en los datos, g es una función continua y ϵ_i son los errores los cuales se suponen independientes con media cero y varianza σ^2.

El método TPS tiene como objetivo estimar el proceso g mediante una función continua f, dicha función se puede estimar minimizando:

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - f_i)^2 + \lambda J_m^d(f) \tag{11}$$

Donde f_i son valores de la función ajustada en el $i - ésimo$ punto de los datos, λ es un parámetro de suavizado y $J_m^d(f)$ es una medida de la rugosidad de la función f en términos de derivadas parciales de orden m, por ejemplo si $m = 2$:

$$J_2^2(f) = \int_{-\infty}^{\infty} f_{x1x1}^2 + 2f_{x1x2}^2 + f_{x2x2}^2 dx_1 dx_2 \tag{12}$$

La estimación puede optimizarse minimizando la validación cruzada generalizada (GCV, por sus siglas en inglés) dada por:

$$\frac{1}{n}\frac{(z-\hat{z})^T(z-\hat{z})}{(Tr(I-A(\lambda))/n)^2} \tag{13}$$

$$\hat{z} = A(\lambda)z,$$

Donde $\hat{z}$ es el vector de valores predichos, $A(\lambda)$ se denomina la matiz de influencia, a partir de esto se puede obtener la estimación de la varianza de los errores ϵ_i :

$$\hat{\sigma^2} = \frac{(z-\hat{z})^T(z-\hat{z})}{Tr(I-A(\lambda))} \tag{14}$$

7. Kriging Ordinario

La interpolación por Kriging a diferencia del IDW y el TPS es un método geoestadístico, esto quiere decir que se incluyen modelos estadísticos que permiten observar la correlación espacial entre puntos y a través de dicha correlación se puede explicar la variación de la superficie estudiada. (Giraldo, 2002)

El Kriging ordinario plantea que el valor de un punto no medido puede predecirse a través de una combinación lineal de los puntos originales:

Sea x_0 un punto sin medición, Z una variable de interés en los puntos x_i, $i = 1, \ldots, n$

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{15}$$

Donde λ_i son los pesos de los valores en los puntos conocidos, la suma de estos pesos debe ser igual a uno para que el valor esperado del predictor sea igual al valor esperado de la varibale, esto se conoce como el requisito de insesgamiento: (Giraldo, 2002)

$$E(Z^*(x_0)) = E(Z(x_0))$$

La estimación de los pesos se obtiene minimizando:

$$V[Z^*(x_0) - Z(x_0)] \text{ sujeto a } \textstyle\sum_{i=1}^{n} \lambda_i = 1$$

Por ende:

$$V[Z^*(x_0) - Z(x_0)] = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C_{ij} - 2\sum_{i=1}^{n} \lambda_i C_{i0} + \sigma^2 \tag{16}$$

Donde $C_{ij} = COV[Z(x_i), Z(x_j)]$ y $\sigma^2 = V[Z(x_0)]$

Mediante multiplicadores de LaGrange:

$$\sigma_k^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C_{ij} - 2\sum_{i=1}^{n} \lambda_i C_{i0} + \sigma^2 + 2\mu[\sum_{i=1}^{n} \lambda_i - 1] \tag{17}$$

Donde 2μ es el multiplicador de LaGrange, ahora los pesos que minimizan el error de predicción se establecen con la función de covariograma $\lambda = C_{ij} \bullet C_{i0}$ y la varianza de predicción está dada por:

$$\sigma_k^2 = \sigma^2 - \sum_{i=1}^{n} \lambda_i C_{i0} + \sigma^2 - \mu \tag{18}$$

8. Resultados

Se simularon escenarios con 5 %, 10 %, 15 % y 20 % de valores faltantes de precipitación en Antioquia para el periodo 2014-2017 en un total de 151 estaciones climatológicas, esto con el fin de validar la metodología que mejor se ajuste a los escenarios propuestos, para aplicar los resultados a los datos de precipitación correspondientes al año 2018. Se usó la imputación de datos por promedio como metodología base de comparación, adicional se realizó la simulación con los métodos de la distancia inversa ponderada (IDW, por sus siglas en ingles), Spline de placa delgada (TPS, por sus siglas en inglés) y Kriging ordinario, para validar la efectividad de las metodología propuestas se usó validación cruzada con la raíz del error cuadrático medio (RMSE, por sus siglas en inglés) como medida de evaluación, a continuación se muestra los resultados obtenidos a nivel general (figura 3).

Los parámetros a seleccionar en la simulación para el IDW se eligieron a través de algoritmos de validación cruzada con el error cuadrático medio como medida de evaluación, estos parámetros corresponden al peso y el número de observaciones más cercanas a un punto o valor determinado. Para el Kriging ordinario en cada una de las simulaciones se ajustó un modelo de semivarianza teórico el cual es dedujo del semivariograma experimental, teniendo en cuenta la pepita, la meseta y el rango.

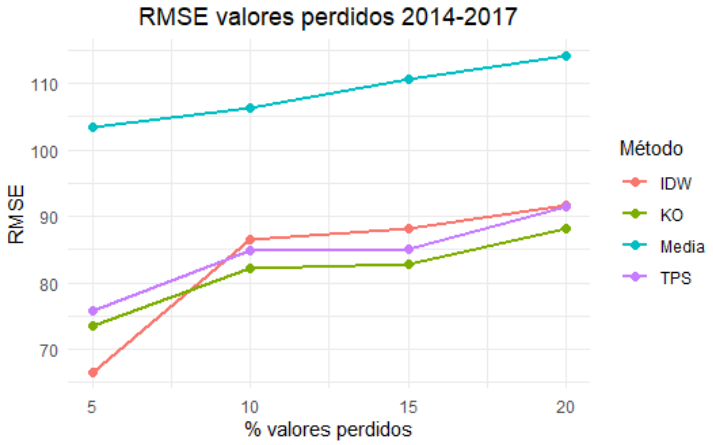

Figura 3: Comportamiento RMSE simulación valores perdidos 2014-2017. Fuente: Elaboración propia

En primera medida se puede observar como la imputación por la media resulta menos efectiva en todos los casos con respecto a los otros métodos utilizados, la imputación por TPS no resulta ser la más efectiva en ninguno de los escenarios, para un 5 % de los valores faltantes el método que mejor comportamiento presenta, fue el IDW con un RMSE de 66.44, el Krigin ordinario muestra mejores resultados en los escenarios de 10 %, 15 % y 20 % de los valores faltantes para precipitación con RMSE de 82.26, 82.79 y 88.14 respectivamente. Todos los métodos de imputación usados en general arrojan mejores resultados a medida que el porcentaje de valores perdidos es menor.

Tabla 2: Resultados smulación RMSE. Fuente: Elaboración propia

% Valores faltantes	Método	2014	2015	2016	2017
		RMSE			
5 %	Media	101,456	111,856	100,733	99,0112
	IDW	74,5196	68,4035	61,2725	60,2948
	TPS	81,5103	77,0077	75,5508	68,5295
	Kriging	80,1443	74,9366	71,7977	66,0831
10 %	Media	107,916	104,739	95,2861	116,739
	IDW	92,6084	78,5002	78,7436	94,5196
	TPS	87,0443	78,7574	77,0159	95,4305
	Kriging	85,5507	74,5935	74,3804	93,0622
15 %	Media	111,621	108,901	100,661	120,761
	IDW	92,4273	87,5949	79,8204	91,9271
	TPS	84,0309	84,3238	82,8662	88,8935
	Kriging	87,1369	78,8907	76,2082	88,1917
20 %	Media	110,778	110,700	106,790	127,598
	IDW	88,7517	86,2246	91,3844	99,9499
	TPS	83,5070	85,1816	87,2722	107,9491
	Kriging	86,0148	79,5659	86,5436	99,3870

Al observar los resultados discriminados por año (Figura 4), para el periodo 2014 el método TPS logra ser más efectivo cuando existen 15 % y 20 % de los valores faltantes, el RMSE del Kriging ordinario sigue mostrando mejor comportamiento cuando existen 10 % de los valores perdidos para todos los años, al igual que el IDW cuando se presenta un 5 % de los valores como faltantes, en ninguno de los casos el Kriging muestra el peor comportamiento si no se tuviese en cuenta el método por la media. Los valores del RMSE se pueden observar en la tabla 2.

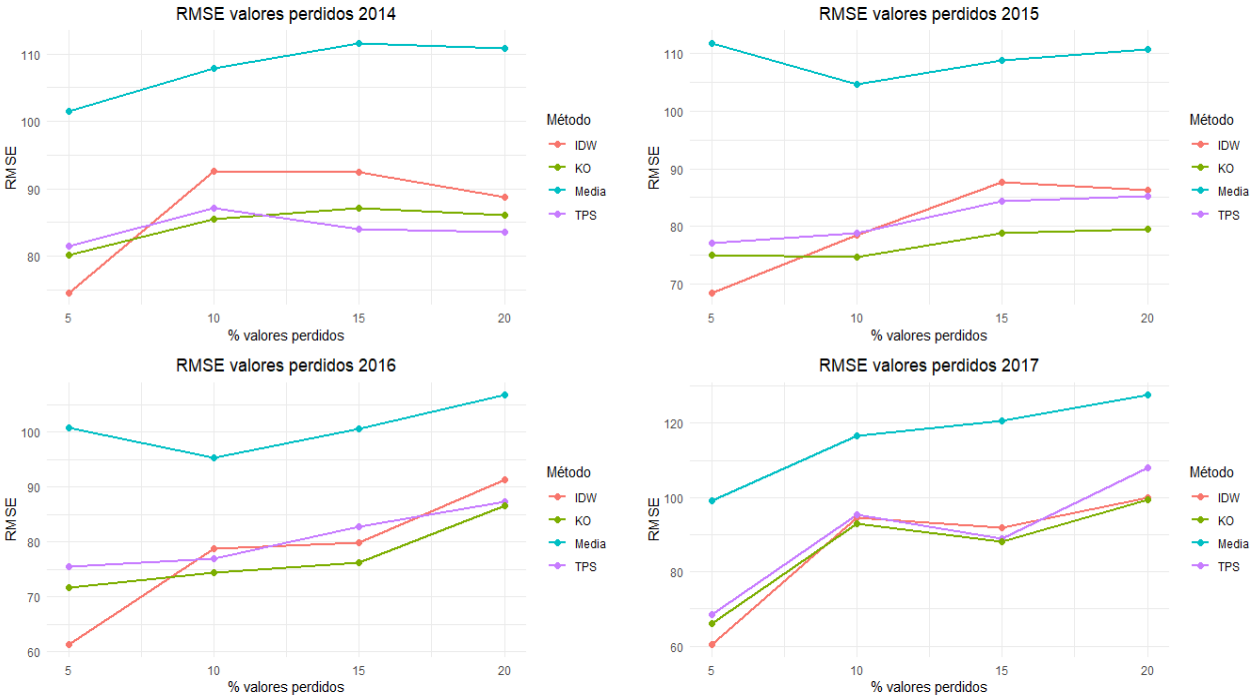

Figura 4: Comportamiento RMSE simulación valores perdidos 2014-2017 por años. Fuente: Elaboración propia

9. Aplicación de los resultados.

Se seleccionó el 2018 como el año donde se aplican los resultados obtenidos en la simulación, para este año se dispone de una base de datos de 151 estaciones climatológicas en las que se tienen medidas de precipitación mensual acumuladas en mm para el departamento de Antioquia para los doce meses del año, dependiendo del mes existen diferentes estaciones climatologícas con valores perdidos como se muestra en la tabla 3, se parte del principio de que cuando existen más del 10 % de los valores perdidos se utiliza el kriging ordinario como método de imputación, si el porcentaje es menor al mencionado se usa el IDW como método de imputación.

Tabla 3: Porcentaje De valores faltantes para el año 2018 con una base de 151 estaciones climatológicas. Fuente: Elaboración propia

Mes	Valores perdidos	Porcentaje valores perdidos
Enero	10	7 %
Febrero	8	5 %
Marzo	40	26 %
Abril	9	6 %
Mayo	6	4 %
Junio	7	5 %
Julio	8	5 %
Agosto	9	6 %
Septiembre	11	7 %
Octubre	8	5 %
Noviembre	10	7 %
Diciembre	9	6 %

Se observa que para todos los casos existe un porcentaje de valores faltantes inferiores al 10 %, con excepción del mes de marzo el cual tiene un 26 % de valores faltantes, para este mes se utiliza el método de interpolación por kriging ordinario, para los meses restantes se utiliza el método por IDW. Se usa la validación cruzada para seleccionar los parámetros del IDW y se ajusta un semivariograma teórico basado en la semivarianza experimental para el Kriging ordinario.

Tabla 4: Porcentaje De valores faltantes para el año 2018 con una base de 151 estaciones climatológicas. Fuente: Elaboración propia

Mes	Pesos	Número de observaciones mas cercanas
Enero	0,500	1
Febrero	0,599	5
Abril	2,346	3
Mayo	10,73	3
Junio	0,880	6
Julio	0,001	5
Agosto	1,853	4
Septiembre	44,13	2
Octubre	0,001	4
Noviembre	0,001	4
Diciembre	0,001	4

En la tabla 4 la columna pesos hace referencia al parámetro de pesos del método IDW y el número de observaciones más cercanas indican la preferencia al número de puntos vecinos en la ponderación. Para el mes de marzo se realizó la imputación con un modelo de semivarianza esférico de parámetros: Pepita (0), Meseta (3290,7), y rango (30,8).

Tabla 5: Parámetros imputación de datos por IDW. Fuente: Elaboración propia

Mes	Modelo	Pepita	Meseta	Rango
Enero	Exponencial	3914,75	8395,21	56,97
Febrero	Gaussiano	6111,56	8021,61	97,38
Marzo	Esferico	0,00	3085,37	36,17
Abril	Exponencial	0,00	12949,00	30,70
Mayo	Exponencial	2224,52	18604,93	25,93
Junio	Esferico	16076,24	20624,08	57,03
Julio	Gaussiano	11519,84	33409,87	152,11
Agosto	Esferico	4519,99	176201,20	1873,76
Septiembre	Exponencial	2990,72	23388,94	94,25
Octubre	Exponencial	9214,27	41318,51	219,85
Noviembre	Matern	5003,74	18138,06	35,35
Diciembre	Esferico	8431,93	8281,37	40,99

Con la totalidad de las mediciones correspondientes a precipitación mensual acumulada para el departamento de Antioquia en las 151 estaciones climatológicas, se ajustan modelos de semivarianza teórica basados en la observación del semivariograma experimental.

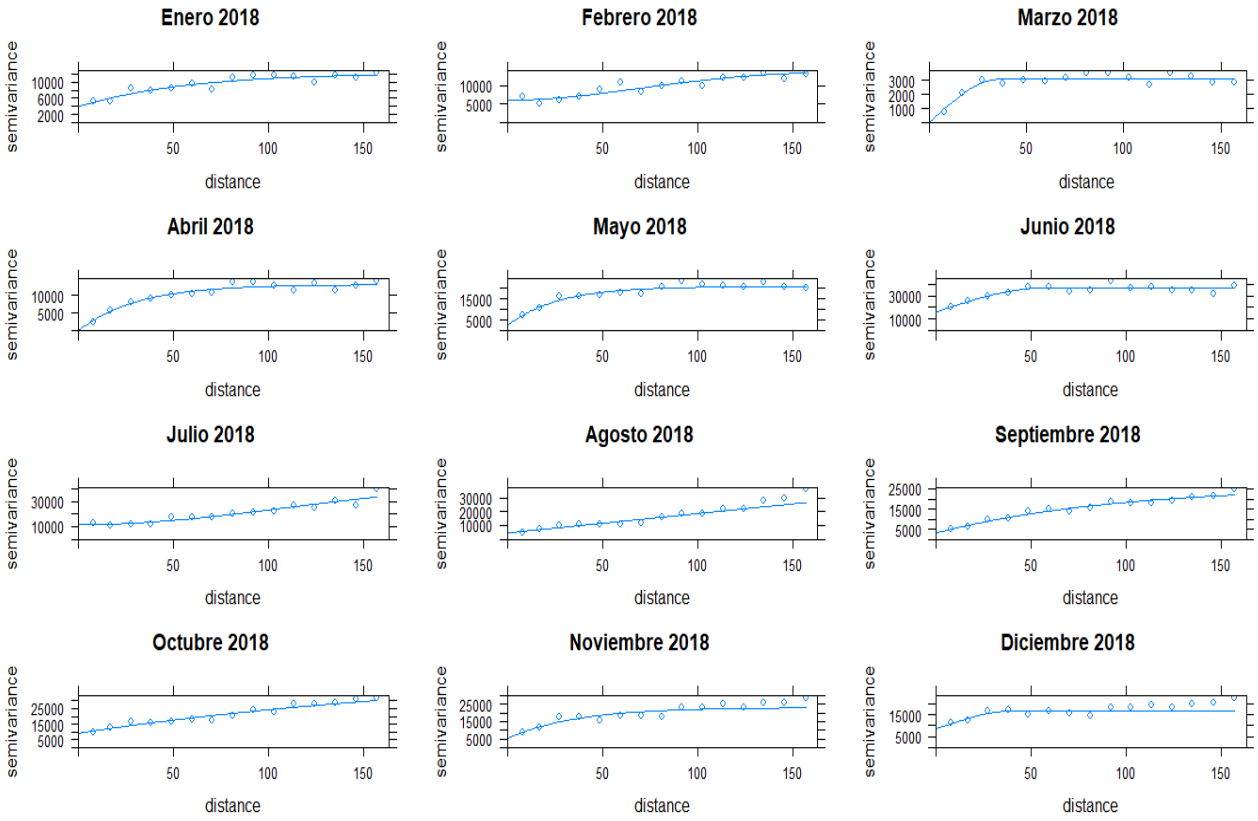

Figura 5: Comportamiento Semivariograma por meses para el año 2018. Fuente: Elaboración propia

En total se ajustaron 12 modelos de semivarianza teóricos, además se pueden observar los parámetros con los que fueron ajustados para cada uno de los meses y gráficamente el ajuste con respecto a los semivaroiogramas experimentales (figura 5).

Se realiza la interpolación espacial para la precipitación mensual acumulada en el departamento de Antioquia para el año 2018 por Kriging ordinario con la base de datos completa, es decir con la totalidad de las mediciones correspondientes a las 151 estaciones climatológicas y las imputaciones ya antes mencionadas.

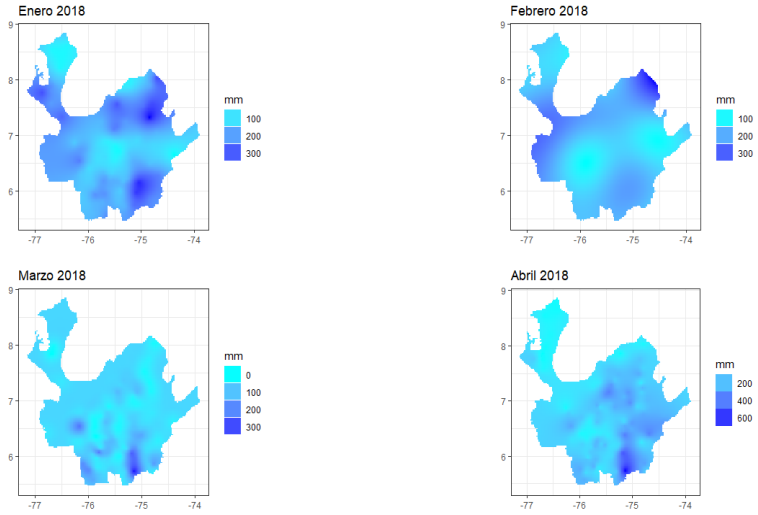

Figura 6: Interpolación espacial por Kriging ordinario primer cuatrimestre 2018.

Para el primer cuatrimestre del 2018 se logra evidenciar que el mes con menor precipitación en Antioquia fue el de marzo y el enero de menor precipitación, en el noroccidente del departamento se observa baja precipitación y en la parte suroriental es donde se logra observar la mayor concentración de lluvia (figura 6).

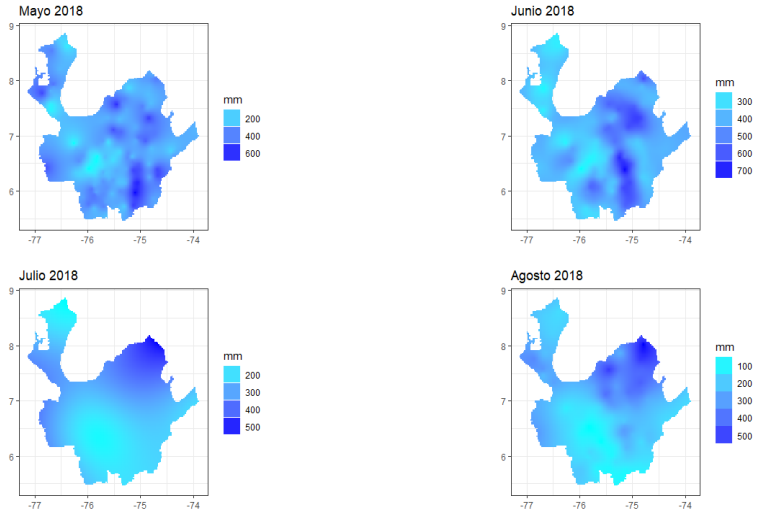

Figura 7: Interpolación espacial por Kriging ordinario segundo cuatrimestre 2018. Fuente: Elaboración propia

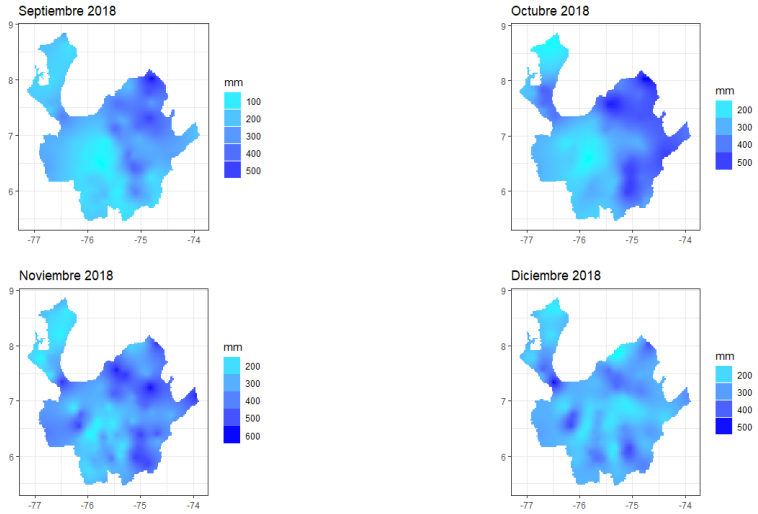

Figura 8: Interpolación espacial por Kriging ordinario tercer cuatrimestre 2018. Fuente: Elaboración propia

En la figura 7 se observa que para el segundo cuatrimestre del 2018 en el departamento de Antioquia una mayor precipitación con respecto al primer cuatrimestre, para este periodo el mes con mayor lluvia fue el de mayo seguido de junio, principalmente al oriente del departamento. Para el tercer cuatrimestre se observa una transición de septiembre a noviembre donde la precipitación va aumentando, para el mes de diciembre se observa una menor precipitación.

10. Conclusiones.

En el diario vivir es común encontrar valores faltantes, en muchas ocasiones se incurre en la mala práctica de eliminar las observaciones que no están completas o simplemente se imputan con el promedio o la moda de los registros, el objetivo es dar a conocer que no siempre la solución más inmediata es la mejor, existen técnicas que al usarlas de una forma adecuada pueden mostrar muy buenos resultados en la imputación de datos faltantes.

La imputación por el promedio arroja una baja efectividad comparada con los otros métodos, se quiso usar esta imputación como metodología base de comparación para mostrar la importancia de usar técnicas de imputación acordes a la estructura de datos estudiada, para este caso datos georreferenciados en los cuales su medición depende en gran medida de su ubicación.

En el proceso de imputación se logra identificar en general que a medida que aumenta el porcentaje de valores perdidos, la efectividad de los métodos usados es menor, lo anterior lo puede explicar que a menor cantidad de mediciones, la distancia entre los puntos observados será mayor y las predicciones se verán afectadas por los puntos de proximidad.

El método determinista por IDW es más efectivo cuando se tienen 5 % de los valores perdidos en la precipitación mensual acumulada de Antioquia, a medida que el porcentaje fue más alto se hizo evidente el uso de la geoestadística en cuanto a identificar y ajustar una dependencia y correlación espacial, el Kriging en ninguno de los escenarios de imputación sin tener en cuenta el promedio fue el de peor resultado, lo cual confirma la importancia de la usar estructuras de correlación espacial.

En la imputación de valores faltantes para precipitación acumulada en Antioquia el método determinista TPS aunque tuvo un comportamiento similar al Kriging en general no resulto siendo más efectivo, con excepción del año 2014 cuando se tienen 15 % y 20 % de los valore perdidos, este método muestra un buen comportamiento cuando las mediciones asociadas a una ubicación son de poca variabilidad y de medida constante, esto nos indica que para dichos escenarios la precipitación no tuvo grandes variaciones.

Como trabajos futuros se propone expandir la validación de las técnicas de imputación para la precipitación mensual acumulada, a nivel país, a su vez incluir metodologías que utilicen covariables tales como la altitud, la temperatura, el brillo solar, entre otras y validar su efectividad respecto a las estudiadas en el presente trabajo.

Referencias

[1] Bihrmann, E. (2015). *Estimación del rango de influencia en caso de que falten datos espaciales: un estudio de simulación sobre datos binarios.*International Journal of Health Geographics, 1-14.

[2] Castano, E. (2007). *Reconstrucción de datos de series de tiempo: una aplicación a la demanda horaria de la electricidad.*Revista Colombiana de Estadística, 247-263

[3] Castro, M. (2015). *Imputación de datos faltantes en un modelo de tiempo de fallo acelerado.*(tesis Maestria). Universidad Santiago De Compostela, La Coruña, España.

[4] Cruz, A., Barrios,M. (2018). *Estimación de datos faltantes de lluvia mensual a través de la asimilación de información satelital y pluviométrica en una cuenca andina tropical.* (tesis Pregrado). Idesia, 1-7.

[5] Departamento administrativo de planeación de Antioquia. (2019). *Obtenido de http://www.antioquiadatos.gov.co/.*

[6] Giraldo, R. (2002). *Introducción a la geoestadística: Teoría y aplicación.*Bogotá: Universidad Nacional de Colombia.

[7] Hancock , P., Hutchinson , M (2005). *Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines.* Environmental Modelling Software 21 (2005) 1684-1694.

[8] Instituto de Hidrología, Meteorología y Estudios Ambientales. (2019). *Obtenido de http://www.ideam.gov.co/.*

[9] Londono, L. (2015). *Imputation of spatial air quality data using gis-spline and the index of agreement in sparse urban monitoring networks.* Revista Facultad de Ingeniería, Universidad de Antioquia, 73-81.

[10] Medina, F. (2007). *CEPAL.* Recuperado el 22 de Febrero de 2018, de http://repositorio.cepal.org/bitstream/handle/11362/4755/S0700590*es.pdf*.

[11] Roshan , J., Lulu, K (2011). *Regression-Based Inverse Distance Weighting With Applications to Computer Experiments.* Technometrics, Vol. 53, No. 3, pp. 254-265.

[12] Salgado, C., Largo, J. (2018). *Imputación de datos faltantes de temperatura máxima media mensual mediante métodos geoestadísticos en estaciones climáticas del valle del cauca en el periodo 2013-2014* . (tesis Pregrado). Universidad Del Valle, Santiago De Cali, Colombia.

[13] Tovar, L. (2017). *Elaboración de tesis estructura y metodología.* Ciudad De México: Trillas.