Cómo citar este artículo:
A. M. Bello, J. A. Cuta, E. K. García, "Técnicas de imputación para datos de precipitación máxima mensual en la zona central de Boyacá". Rev. Revista Ingeniería, Investigación y Desarrollo, vol. 19 (1), pp. 64-79, junio. 2019.

TÉCNICAS DE IMPUTACIÓN PARA DATOS DE PRECIPITACIÓN MÁXIMA MENSUAL EN LA ZONA CENTRAL DE BOYACÁ

Imputation techniques applied in a maximum monthly precipitation data in the central zone of Boyacá

Angie Milena Bello[1], Julián Andrés Cuta[2], Ehidy Karime García[3]
[1] Universidad Pedagógica y Tecnológica de Colombia / Escuela de ingeniería geológica, Grupo de investigación INGEOLOG, Colombia. Email: ingeolog@uptc.edu.co
[2]Universidad Pedagógica y Tecnológica de Colombia / Escuela de ingeniería industrial, Grupo de investigación OBSERVATORIO, Colombia. Email: grupo.observatorio@uptc.edu.co
[3] Universidad Pedagógica y Tecnológica de Colombia / Grupo de investigación GAMMA, Colombia. Email: gamma.estadistica@uptc.edu.co

(Recibido Diciembre 05 de 2018 y aceptado Junio 22 de 2019)

Resumen

La precipitación se encuentra relacionada directamente con el suministro de agua de las cuencas fluviales, convirtiéndose su predicción en un objetivo de estudio en diferentes investigaciones. Sin embargo, los registros históricos a menudo muestran datos faltantes debido a fallas instrumentales, técnicos o humanos. Esta limitación impacta directamente los resultados de los análisis estadísticos que puedan ser realizados posteriormente. Esta investigación aborda este problema para un conjunto de datos con características similares, recopilados en la parte central del departamento de Boyacá- Colombia para el período 1974-2013. Se evaluó el desempeño de los mecanismos de imputación de pérdida MCAR, MAR o MNAR, cada uno de estos se implementó usando una imputación múltiple con un enfoque aleatorio, una asignación por el método de K-Nearest Neighbors con enfoque espacial y una imputación por el método de suavizado de Kalman con enfoque temporal. Se midió la convergencia de los estadísticos descriptivos del valor imputado y el valor original y se realizó la comparación de los ajustes gráficos y sus distribuciones de probabilidad, sugiriendo un mejor ajuste usando la imputación múltiple Amelia en conjunto con un ajuste a una distribución gamma para los datos faltantes en el conjunto de datos de referencia.
Palabras clave: *Imputación múltiple, Precipitación, R, series temporales, datos faltantes, Boyacá.*

Abstract

Precipitation directly affects the water supply of river basins and its prediction becomes the main objective in different investigations. However, historical records often show missing data due to instrumental, technical or human drawbacks. This limitation must be solved to avoid errors in subsequent Analysis. This proposal deal with a similar problem for a data set about precipitation collected in the central part of Boyacá along the years 1974-2013. The performance of the imputation mechanisms of loss MCAR, MAR and MNAR was evaluated. All of them were implemented each one under either a multiple imputation with a random approach based on an allocation by the K-Nearest Neighbors method with spatial focus and an imputation by the Kalman smoothing method time focused approach. We measured the convergence of the descriptive statistics of the imputed value and the original value, and additionally, we compared the graphical adjustments and their probability distributions. Amelia was suggested as a better performance of imputation technique jointly with a gamma distribution associated to the missing data.
Key words: *Multiple imputation, precipitation, R-software, temporal series, Missing data, Boyacá.*

1. INTRODUCCIÓN

La precipitación es el fenómeno meteorológico por el cual el vapor de agua se condensa y desciende de la atmosfera a la superficie terrestre [1]. Este fenómeno es importante para desarrollar estudios hidrológicos necesarios en el diseño de obras civiles, planeamiento del territorio y demás proyectos implicados en el crecimiento económico y social [2]. La medición de las precipitaciones en Colombia se hace mediante pluviómetros, la información meteorológica disponible se ve limitada debido a que las estaciones, su instalación y su mantenimiento, se constituyen como un trabajo arduo y costoso; a esto se suma la complejidad geográfica y climática que afecta la densidad de la red de monitoreo [3]. Estos registros no siempre cuentan con una cobertura y longitud convenientes, pues hay periodos en los cuales no se registran datos debido a fallas humanas o en la instrumentación. En caso de que los estudios requieran esta información, es necesario completar los datos faltantes por medio de diferentes métodos que consideren las particularidades de estas series temporales y logren modelar el fenómeno con cierto grado de validez [4].

En la metodología de imputación de registros para series temporales, debe considerarse el mecanismo de pérdida de los datos, para así elegir el método que se adapte mejor al conjunto de datos; estas clasificaciones son cruciales, ya que las propiedades de estos métodos dependen en gran medida de la naturaleza y las dependencias en estos mecanismos. Dichos mecanismos se clasifican considerando la relación de una distribución condicional y el conjunto de datos faltantes; de acuerdo con lo anterior se definen tres mecanismos de pérdida: Missing completely at random (MCAR), Missing at random (MAR), y Missing not at random (MNAR) [5]. El enfoque para los métodos de imputación ha de ser definido considerando las características que son inherentes a los conjuntos de datos estudiados, siendo de mayor prevalencia aquellas características espaciales, temporales o aleatorias. Actualmente, algunos de estos estudios no consideran este factor al momento de elegir un método para completar los datos faltantes puesto que basan sus estudios en sugerencias de organismos internacionales [6, 7]. El estudio de los métodos para completar datos faltantes ha tenido un gran desarrollo en las últimas décadas, dadas las diferentes herramientas computacionales que facilitan el cálculo de los valores faltantes [8]. Estudios iniciales con valores promedios [9]; estudios internacionales que se basan en un enfoque determinístico [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]; también se han empleado técnicas estocásticas [20, 21, 22, 23]. Esta temática también ha sido tratada en investigaciones colombianas [24, 25, 26, 27, 28, 29], algunas de ellas desarrolladas en Boyacá [30, 31].

En estos estudios, no se consideran los mecanismos de pérdida de los datos para la selección de las técnicas de imputación. No existe un enfoque estocástico para la imputación de los datos de precipitación, tampoco se tienen referencia similar que realice una comparación que permita evaluar el rendimiento de llenado para diferentes técnicas de imputación.

El presente estudio comprende un análisis exploratorio de los datos de precipitación máxima mensual para el lapso entre 1974-2013, recibidos del Instituto de Hidrología, Meteorología y Estudios Ambientales, en adelante IDEAM; una prueba estadística para la identificación del mecanismo de pérdida de los datos; seguido de la selección y aplicación de las técnicas de imputación propuestas con enfoques diferentes, uno aleatorio (imputación múltiple), uno espacial (K-Nearest Neighbors) y uno temporal (Suavizado de Kalman); finalmente se realiza la evaluación de dichos métodos. Todo el estudio se desarrolla en R- Statisfical Software [32, 33]

2. MATERIALES Y MÉTODOS

*2.1 Área de estud*io

La zona elegida para el estudio se encuentra en el área central del departamento de Boyacá y una pequeña área del departamento de Casanare. La Figura 1 representa la ubicación geográfica, en la cual los puntos simbolizan las estaciones meteorológicas del IDEAM.

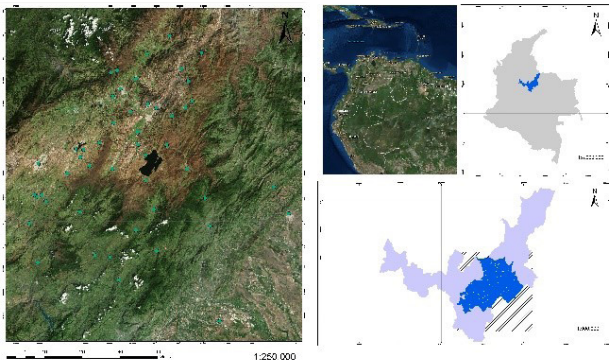

Figura 1: Localización geográfica del área del estudio. Fuente: Autores.

2.2. Datos utilizados

Se recopilaron los registros históricos proporcionados por el IDEAM. Series temporales de precipitación máxima mensual y anual acumulada de 49 estaciones meteorológicas, con periodos de medición de diferente longitud, por lo cual se hizo análisis en un periodo de cuarenta años (1974-2013) y se dividieron los registros en cuatro décadas para analizar la totalidad de los datos con las estaciones que tuvieran el mismo periodo registrado. (Ver tabla 1).

Tabla 1. Estaciones meteorológicas IDEAM.

	Municipio	Estación	Este	Norte	Periodo
1	Gámeza	Nimicia	1145233	1131022	1974-2013
2	Aquitania	Cazadero	1121296	1076091	1984-2013
3	Cerinza	Cerinza	1126120	1151193	1974-1993
4	Chameza	Chameza	1130231	1067376	1974-2013
5	Corinto	Corinto	1150770	1089859	1984-2013
6	Duitama	Duitama	1114779	1136238	1974-2013
7	Cuítiva	Túnel	1125570	1108035	1974-1993
8	Firavitoba	Firavitoba	1122162	1117940	1974-2013
9	Iza	Iza	1122394	1112518	1974-2013
10	Monguí	Monguí	1136599	1124609	1974-2013
11	Pajarito	Pajarito	1152874	1078008	1974-2013
12	Pesca	Pesca	1111267	1102271	1974-2013
13	Aquitania	Potrerito	1125489	1097269	1974-2013
14	Rondón	Rondón	1097245	1084042	1974-2013
15	Sogamoso	Sena	1129941	1128079	1984-2013
16	Toquilla	Toquilla	1142951	1102396	1974-2013
17	Aquitania	Guamos	1129053	1085008	1984-2013
18	Las Cintas	Las Cintas	1134430	1112391	1974-2013
19	Samacá	Teatinos	1078152	1091141	1994-2013
20	Ramiriquí	Villa Luisa	1081074	1091078	1984-2013
21	Tibaná	Tibaná	1075930	1079244	1974-2013
22	Yopal	El Morro	1180075	1094680	1974-2013
23	Chinavita	Chinavita	1079462	1062613	1974-2013
24	Miraflores	Vivero	1103794	1065708	1984-2013
25	Yopal	Yopal	1186343	1083296	1974-1993
26	Mongua	Mongua	1142503	1128364	1974-2013
27	Tasco	Tasco	1143553	1139646	1974-2013
28	Ramiriquí	Ramiriquí	1082899	1088570	1974-2013
29	Siachoque	Siachoque	1091963	1100368	1974-2013
30	Toca	CasaAmar.	1101662	1103447	1974-2013
31	Toca	El Garrocho	1105768	1113370	1974-2013
32	Tibasosa	Tibasosa	1119194	1126706	1974-2013
33	StaRosa V	StaRosa V	1121288	1140524	1974-2013
34	Socha	La Chapa	1149225	1151419	1974-1993
35	Tunja	Uptc	1079795	1104442	1974-2013
36	Beteitiva	Beteitiva	1144829	1079795	1994-2003
37	Páez	Páez	1113963	1055084	1974-2013
38	Berbeo	Buenavista	1110233	1064741	1974-2013
39	Jenesano	Jenesano	1079162	1090461	1984-2003
40	Toca	San Pedro	1097611	1107075	1974-1993
41	Toca	Hotel El	1095759	1110759	1974-1993
42	Toca	S. Cristóbal	1098967	1110423	1974-2013
43	Duitama	SurbataBon.	1111477	1133175	1974-2013
44	Tauramena	Pradera La	1156961	1037577	1974-2013
45	Duitama	An. Rusia	1110582	1143039	1984-2013
46	Nobsa	Nobsa	1126352	1129992	1974-2013
47	Duitama	Andalucía	1113245	1144094	1994-2013
48	Paipa	Cerezo	1111798	1121768	1974-2013
49	Tutazá	Tutazá	1135666	1158792	1974-2013

Fuente: IDEAM-Autores

2.3 Software

El software R, es un lenguaje de programación estadístico que provee infinitas posibilidades para solucionar problemas científicos en casi todas las áreas de investigación debido a la versatilidad que brinda al usuario para solucionar analíticamente los problemas estadísticos, además, su funcionamiento es flexible, pues permite la adaptación de sus librerías a necesidades particulares [34].

Para la realización del estudio se utilizaron los siguientes paquetes, y en paréntesis están las funciones necesarias: Normtest *(sf.test*), BaylorEdPsych (*LittleMCAR*), función genérica (*kruskal.test*), función gráfica (*boxplot*), Trend (*ww.test*), riskDistributions (*fit.cont*), función grafica (*hist*), función genérica(*lines*), MASS (*fit.distr*), bbmle (*mle2*), Flexsurv, goftest (*ad.test*)-(*cvm.test*), función genérica (*ks.test)*, función grafica (*plot*), PerformanceAnalytic (*chart.QQplot*), Amelia (*amelia*), DMuR (*knnImputation*), Impute.TS (*na.kalman*) [33].

2.4 Mecanismo de pérdida de los datos

Los mecanismos de pérdida de datos son consideraciones matemáticas que describen las leyes que rigen la aparición de datos perdidos y que captan las posibles relaciones con los datos no observados en sí mismos.

Se tiene un vector aleatorio **x** con dimensión *k* que genera los datos y un vector **v** también de dimensión *k*, formado por variables aleatorias binarias tomando valores de 0 o 1 para indicar el valor observado o no observado. Se conoce al mecanismo de no respuesta como la distribución de probabilidad de **v**. Asociada a una muestra del vector x se tendrá una muestra de **v**, cuya forma dependerá de la complejidad del patrón de no respuesta. Si se denota mediante $\tilde{\mathbf{X}}$ a una muestra multidimensional de **x** se puede hacer una partición de forma que $\mathbf{x}=(\mathbf{x}_{obs},\mathbf{x}_{per})$, donde $\mathbf{x}_{obs}$ es la parte observada y $\mathbf{x}_{per}$ representa la parte pérdida. Se dice que los datos faltantes son de tipo MCAR si $P[\mathbf{v}|\mathbf{x}_{obs},\mathbf{x}_{per},\xi]=P[R|\xi]$, donde ξ es un vector de parámetros desconocidos del mecanismo de no respuesta. Los datos ausentes del tipo MAR consideran $P[\mathbf{v}|\mathbf{x}_{obs},\mathbf{x}_{per},\xi]=P[\mathbf{v}|\mathbf{x}_{obs},\xi]$.

Por último, se presentó el mecanismo MNAR, en el cual la no respuesta depende del valor verdadero del dato perdido (depende de $\mathbf{x}_{per}$), o de variables no observables. La hipótesis del mecanismo MNAR es la más general, pero al mismo tiempo es la más difícil de modelar ya que exige la especificación de un modelo para **v**, por lo que es frecuente hablar de mecanismo de no repuesta no ignorable [35].

La identificación del mecanismo de pérdida de los datos en el software R Project se hizo por medio de *BaylorEdPsych* (*LittleMCAR*), que proporciona un test que comprueba la hipótesis nula de que los datos tienen un patrón MCAR, un valor *p* de menos de 0.05 generalmente se interpreta como que los datos faltantes no son MCAR, es decir, no son ignorables. No existen procedimientos generales para contrastar algunos de los otros dos mecanismos sobre un conjunto de datos incompletos [36]

2.5 Análisis Exploratorio de datos

El análisis exploratorio de datos (EDA) [37] se inició con la obtención de los estadísticos descriptivos que serán la base para la comparación de las técnicas. Se prosiguió con la realización de la prueba de normalidad de Shapiro-Francia [38]. Para el cálculo del estadístico se utiliza la siguiente formula:.

$$W' = \frac{\sum_{i=1}^{n}\left(x_{(i)}-\bar{x}\right)(m_i-\bar{m})}{\sqrt{\left(\sum_{i=1}^{n}\left(x_{(i)}-\bar{x}\right)^2\right)\left(\sum_{i=1}^{n}(m_i-\bar{m})^2\right)}} \quad (1)$$

Donde $x_{(i)}$ es el valor de muestra ordenado, *n* el tamaño del conjunto de datos y m_i los cuantiles ordenados esperados. El valor *p* se calcula a partir de la fórmula dada por [39]. Se aplicó el test por medio de normtest (sf.test) que aplica el test de normalidad, si el valor *p* es menor a 0,05 se rechazara la hipótesis nula.

Para la prueba de homogeneidad e independencia, se hace uso del test Kruskall-Wallis (kruskal.test) el cual constituye una alternativa no paramétrica al análisis de varianza usual y se considera como una extensión del procedimiento de suma de rangos de Wilcoxon. La hipótesis nula es que no existen diferencias entre en el conjunto de datos (homogéneos), mientras que la hipótesis alternativa es que exista diferencia entre el conjunto de datos (no homogéneos). Considere que se dispone de *k* muestras de tamaños n, donde $n=n_1+n_2+...+n_k$. Si llamamos R_i a la suma de rangos de las observaciones de la i-isima muestra, el estadístico *H* será [40]:

$$H = \left(\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(n+1) \quad (2)$$

Para la prueba de homogeneidad gráfica, se usaron los diagramas de cajas por medio de la función grafica (*boxplot*). La prueba de independencia y estacionalidad, se realizó por medio del test Wald-Wolfowitz, el cual se usa para probar la hipótesis nula de que los conjuntos de datos se han extraído de una muestra idéntica [41]. Bajo la hipótesis nula, el número de corridas en una secuencia de *N* elementos es una variable aleatoria cuya distribución condicional dada la observación de N_+ valores positivos y N_- valores negativos ($N=N_+ + N_-$) es aproximadamente normal, con:

$$\mu = \frac{2N_+N_-}{N} + 1 \quad (3) \qquad \sigma^2 = \frac{2N_+N_-(2N_+N_- - N)}{N^2(N-1)} \quad (4)$$

Para el cálculo del estadístico se usa:

$$Z = N_+ - \frac{\mu(N_-)}{\sqrt{\sigma^2(N_+)}} \quad (5)$$

Para la aplicación en el software se usó trend (ww.test) y se

logró obtener los estadísticos para la evaluación de las hipótesis, se rechaza la hipótesis nula con un valor p menor a 0,05.

2.6 Ajuste a una distribución teórica de probabilidad

Para ajustar los datos experimentales a una distribución teórica de probabilidad es necesario que los datos cumplan con las hipótesis anteriores de modelación, que permitan suponer que los datos tienen un comportamiento aleatorio e independiente. Comprobados estos supuestos se construyeron histogramas, representando la distribución de frecuencias de una variable continua, también puede ser usada para representar la función empírica de densidad de los datos. El histograma estará cercano al valor medio de la función de densidad sobre la clase [42]:

$$\frac{1}{a_h - a_{h-1}} \int_{a_{h-1}}^{a_h} f(x)dx \quad (6)$$

Se puede reconocer a simple vista el ajuste del histograma con la función de densidad teórica que se proponga. Al haber seleccionado un conjunto de posibles distribuciones, se procedió a calcular los valores de los parámetros. Cada función fue definida por varios parámetros específicos, generalmente entre uno o tres dependiendo de la función de distribución. Existen diferentes métodos numéricos para el cálculo de los parámetros (Mínimos cuadrados, Máxima verosimilitud, Método de momentos, entre otros).

Se utilizó el método numérico de Máxima verosimilitud [43]. En cada distribución propuesta es necesario formular la función de máxima verosimilitud que permita encontrar el valor de los parámetros correspondientes. Además por cada función de distribución se calculó un estimador máximo-verosímil.

La aplicación del método de Máxima verosimilitud tuvo dos enfoques diferentes dado que se debe analizar el comportamiento independiente de cada estación y del conjunto total de datos por década. Para el primer caso se hace uso de rriskDistributions (*fit.cont*), que proporciono una interfaz gráfica para la elección de una distribución continua. La función devuelve la distribución continua elegida, los parámetros de la distribución, los estadísticos de la prueba de bondad de ajuste y las siguientes graficas: Histograma de la función de densidad empírica y teórica, el grafico Q-Q, la función de distribución acumulada empírica y teórica. En la aplicación del método de máxima verosimilitud al conjunto total de datos por década, se utilizó MASS (*fit.distr*).

Para el primer enfoque se obtuvieron los parámetros, la función de probabilidad seleccionada y la prueba de bondad de ajuste con sus graficas correspondientes. Pero en el segundo enfoque solo se obtuvieron los parámetros para la familia de distribuciones seleccionadas, es por esto que es necesario aplicar una prueba de bondad de ajuste para seleccionar la distribución que mejor se ajusta a los datos y de igual manera construir los gráficos correspondientes.

Para la prueba de bondad de ajuste se consideró el estimador de máxima verosímil, el criterio de información de Akaike (AIC) y criterio de información bayesiano (BIC) [44], los estadísticos y valor p de las pruebas: Anderson-Darling, Cramer Von Misses y Kolmogorov-Smirnov [45]. Una vez aplicada la prueba se procede a seleccionar la distribución que mejor representa al conjunto de datos totales por década y al conjunto de datos de cada una de las estaciones, proporcionando un punto de partida para la comparación y evaluación de las imputaciones.

2.7 Métodos de imputación para valores faltantes

2.7.1 Imputación múltiple "Amelia":

En la imputación múltiple, los valores perdidos para cualquier variable se predicen utilizando valores existentes de otras variables o de la misma variable. Los valores predichos se sustituyen por los valores faltantes, lo que da como resultado un conjunto completo de datos. Este proceso se realiza varias veces, produciendo múltiples conjuntos de datos imputados. El análisis estadístico estándar se lleva a cabo en cada conjunto de datos, produciendo múltiples resultados de análisis. Estos resultados del análisis se combinan para producir un análisis general [46]. La imputación múltiple completa los datos faltantes restaurando no solo la variabilidad natural de los datos, sino también incorporando la incertidumbre causada por

la estimación de los datos faltantes, para mantener la variabilidad original los datos imputados se basan en variables correlacionadas con los datos faltantes y las causas de la falta. La incertidumbre se considera creando diferentes versiones de los datos faltantes y observando la variabilidad entre los conjuntos de datos imputados [47].

El método de imputación múltiple se puede resumir en tres etapas [48]: (1) cada valor perdido se reemplaza por un conjunto de *m*>1 valores generados por simulación, con lo que se crean m conjuntos de datos completos; (2) se aplica a cada conjunto, los métodos de análisis deseados; (3) los resultados obtenidos se combinan mediante reglas simples para producir una estimación global.

Se considera que el número óptimo de bases de datos *m* depende del porcentaje de información faltante. Cada una de las *m* estimaciones anteriores se pueden crear con una gran variedad de métodos (imputación por media, modelos Monte Carlo con cadenas de Markov) [49]. Para combinar las *m* estimaciones obtenidas se calcula la media de todas las combinaciones, [50, 51, 52]. Sean $\hat{\theta}_i$ las estimaciones realizadas y W_i las varianzas respectivas a cada estimación para un parámetro θ, con conjunto de datos, *i*=1,…,m; la estimación combinada es:

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \quad (7)$$

La variabilidad asociada a esta estimación tiene dos componentes:

- La varianza dentro de cada imputación,

$$\bar{W}_m = \frac{1}{m}\sum_{i=1}^{m} W_i \quad (8)$$

- La varianza entre las imputaciones,

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \bar{\theta}_m\right)^2 \quad (9)$$

Por lo tanto la variabilidad total asociada a la estimación θ_m es:

$$T_m = \bar{W}_m + \frac{m+1}{m}B_m \quad (10)$$

Donde $\frac{m+1}{m}$ es el factor de corrección por ser *m* un número finito. Por lo tanto,

$$\hat{\gamma}_m = \frac{m+1}{m}\frac{B_m}{T_m} \quad (11)$$

Donde $\hat{\gamma}_m$ es una estimación de la fracción de información sobre θ que se pierde por falta de respuesta. Si el parámetro θ es escalar, las estimaciones por intervalo y las pruebas de significación siguen una distribución t de Student:

$$(\theta - \hat{\theta}_m)T_m^{-1/2} \sim t_v \quad (12)$$

Donde los grados de libertad *v*:

$$v = (m-1)\left(1 + \frac{\bar{W}_m}{B_m(m+1)}\right)^2 \quad (13)$$

En el caso contrario, cuando θ tiene *K* componentes, las pruebas de significancia para contrastar la hipótesis de nulidad del parámetro estimado θ deben ser realizadas a partir de las m estimaciones realizadas, y no a partir de la estimación combinada.

Se utilizó Amelia (*amelia*) que ejecutó una imputación múltiple por medio de algoritmo *EM bootstrap* en datos incompletos y creo conjuntos de datos imputados. El algoritmo primero crea una versión inicial *bootstrap* de los datos originales, estima los estadísticos suficientes (con niveles previos, si se especifica) por *EM* en una muestra posterior *bootstrapped*, y luego imputa los valores faltantes de los datos originales utilizando las estadísticas suficientes estimadas. Repite este proceso m veces para valores no observados que se obtienen de sus distribuciones posteriores [51]

2.7.2 Imputación K-Nearest Neighbors (KNN):

En el enfoque espacial para el método de imputación de los datos se consideró el método K-Nearest Neighbors en adelante KNN, el cual es un algoritmo de clasificación simple, que almacena todas las observaciones disponibles y clasifica nuevos casos basado en una medida semejante. Así, al usar KNN una nueva observación es clasificada por votación entre sus *K* vecinos más cercanos [53].

Algunas de las medidas de semejanza son:

- Distancia Euclídea: $d(x,y) = \sqrt{\sum i|x_i - y_i|^2}$
- D. de Manhattan: $d(x,y) = \sum i|x_i - y_i|$
- D. de Minkowski: $d(x,y) = (\sum i|x_i - y_i|^p)^{\wedge}(1/p)$

Es necesario normalizar las variables en una escala común antes de la aplicación del método, de lo contrario podría verse afectado por la escala en que se miden los atributos. Por otra parte, el parámetro *K* debe ser definido para el ajuste del modelo, una forma de hacerlo es por inspección en los datos. Generalmente un *K* más grande es más preciso ya que reduce el ruido, aunque no hay garantía de ello por lo que debe ser evaluado caso a caso. Otra forma de estimar el *K* óptimo es mediante validación cruzada, donde se observan los resultados para distintos valores *K* y se elige aquel que tenga un menor error promedio. Usualmente el *K* óptimo se encuentra entre 3 y 10 en la mayoría de conjuntos de datos [54].

Para aplicarlo se utilizó *DMwR* (*knnImputation*), que consideró la distancia Euclidiana para encontrar los vecinos más semejantes a los valores faltantes en un conjunto de datos. Para cada caso con cualquier valor de NA, buscará sus *K* casos más similares y usará los valores de estos casos para completar las incógnitas.

2.7.3 Imputación por suavizado de Kalman

Este es un algoritmo de procesamiento utilizado para modelar sistemas dinámicos y procesos estocásticos. Considera el estado del proceso en un instante *t* y los factores para modelar probabilísticamente. Todos estos valores se ponderan mediante unos parámetros, que habrá que optimizar de manera que el error sea minimizado [55]. Los modelos de estado pueden representarse por medio de estructuras básicas, tal como la descomposición de una serie de tiempo o por modelos ARIMA los cuales consideran las series temporales en tres componentes principales [56]: El componente auto regresivo, el cual utiliza valores pasados para realizar la regresión de los valores de la serie temporal a modelar; el componente integrador, el cual denotara el grado de diferenciabilidad *d* del modelo, es decir calcular $\sum_{t-d+1}^{t}(y_i - y_{i-1})$, siendo y_t el valor de la serie en el instante *t*. Esta transformación elimina el efecto de componentes como la tendencia y la estacionalidad, lo que convierte la serie temporal en estacionaria. Por último, se tiene el componente media móvil, el cual modela el error del sistema mediante la combinación de términos de error anteriores. Al combinar los tres componentes se obtiene un modelo ARIMA. La aplicación del suavizado de Kalman se realizó por medio de *imputeTS* (*na.kalman*), que imputo valores perdidos por el Suavizado de Kalman y modelos de espacio estado obtenidos a partir de un modelo ARIMA [55].

2.7.4 Evaluación de los métodos de imputación:

Para la evaluación de los métodos, se utilizaron las series de datos completas de cada uno de los métodos de imputación propuestos. Se realizó un análisis exploratorio de estos datos, con el fin de visualizar y comparar la convergencia con respecto a los datos originales. Se hace una comparación de las similitudes en los estadísticos de tendencia central (media, mediana, moda), estadístico de dispersión (desviación estándar y varianza) y estadísticos de forma (coeficiente de asimetría y curtosis); y se evalúa la semejanza de las gráficas obtenidas en cada uno de los conjuntos de datos imputados. Esto con el fin de encontrar tanto visual como numéricamente diferencias que puedan comprometer el ajuste de los datos imputados. Además, se realizó una modelación de las funciones de distribuciones para cada conjunto de datos imputados, con el fin de comparar los parámetros obtenidos en los datos originales y los datos imputados, clarificando si existe una variabilidad en las posibles ocurrencias que puedan suceder. También se comparan las gráficas obtenidas en las pruebas de bondad de ajuste para así tener una referencia de los cambios aleatorios que pueden suceder en las imputaciones propuestas [57]

3. RESULTADOS

De la información recibida del IDEAM se identificaron 49 estaciones con registros de precipitación máxima mensual, de las cuales 33 contienen datos para el periodo 1974-2013; una para el periodo 1974-1993; una para el periodo 1984-1993; cuatro para el periodo 1974-1993; siete para el periodo de 1984-2013; una para el periodo 1994-2003 y dos para el periodo 1994-2013. A partir de esto se realizó. una división en cuatro décadas, clasificando cada estación

dependiendo del registro que contenga. Las cuatro décadas se dividen así: 1974-1983, 1984-1993, 1994-2003 y 2004-2013. Esta clasificación permite agrupar registros semejantes para dichos periodos, facilitando los análisis pues al no realizar dicha clasificación podría conllevar problemas al momento de aplicar las metodologías de imputación propuestas, creando sesgo en los resultados.

Las estaciones con mayor porcentaje de datos perdidos son, Anda Lucia (22,5%) y Beteitiva (18,3%), mientras que las estaciones que menor porcentaje presentan son Nobsa (0,6%), Firavitoba (1,046%) y Santa Rosa (1,064%). El conjunto total suma 20.400 datos de precipitación máxima mensual, se identifican 1.210 datos faltantes, que corresponde a un 5,93% del total de registros. En cuanto a la identificación del mecanismo de pérdida de los datos se inicia con la aplicación del método propuesto para cada una de las décadas, el estadístico que se considero fue el valor p de la prueba, con hipótesis nula de que los datos tienen el mecanismo de pérdida MCAR. Al realizar la prueba "*LittleMCAR*" se obtienen los siguientes resultados:

Tabla 2. Estadísticos para los mecanismos de pérdida

Estadístico	1974-1983	1984-1993	1994-2003	2004-2013
p-value	$1{,}48E^{-06}$	$3{,}63E^{-09}$	$4{,}22E^{-06}$	$3{,}87^{E-05}$
chi-square	2454,38	3187,94	3235,95	8081,19
df	1880	2738	2885	3285

Fuente: Autores

Ninguna década presenta un mecanismo de pérdida MCAR (Ver tabla 2). Ya que no se encontró un procedimiento para comprobar algunos de los otros dos mecanismos de pérdida de datos, se realizó un análisis teórico que permitió hacer la elección por descarte.

Para excluir la hipótesis MNAR, se hace uso del conocimiento empírico sobre el fenómeno de lluvia en la zona de estudio. Se asumió que la pérdida de datos se debía a fallas en el medidor causadas por eventos de lluvia particularmente intensos. Si esta hipótesis fuera correcta y admitiendo que el medidor comienza a funcionar correctamente de manera inmediata en un periodo de lluvia intensa, entonces las altas tasas de precipitación deberían perderse con mayor frecuencia que los valores bajos; pero al revisar los datos de estaciones para diferentes periodos se observa que la pérdida de registros no solo ocurre en los periodos más intensos de lluvia. Por ejemplo: el periodo de lluvia para la década 1984-1993, en la estación El Garrocho, registra valores mayores con respecto a la década 1974-1983 en la cual se presentó menor cantidad de datos faltantes, lo que nos permite descartar dicha hipótesis. En consecuencia, se asume que los datos tienen un mecanismo de pérdida MAR en las cuatro décadas planteadas. Al revisar la literatura se observó que los métodos de imputación propuestos, se basan en el supuesto de que los datos tienen un mecanismo de pérdida MAR. En cuanto al análisis exploratorio de datos se obtienen los siguientes resultados:

Tabla 3. Estadístico análisis exploratorio

Test	74-83	84-93	94-003	004-013
Shapiro-Francia	w=0,685 p=2,21e-16	w=0,760 p=2,2e-16	w=0,1294 p=4,3e-15	w=0,131 p=2,22e-16
Krukal-Wallis	chi2=937,6 df=37 p=2,2e-16	chi2=1140,7 df=45 p<2,2e-16	chi2=1188 df=43 p<1,45e-16	chi2=1187,8 df=41 p=2,2e-16
Wald-Wolfowitz	z=32,67 p=1,1e-16	z=28,407 p<2,2e-16	z=34,804 p<2,2e-16	z=18,501 p=1,5e-16

Fuente: autores

Para la prueba de normalidad Shapiro-Francia, se obtienen estadísticos de baja magnitud lo que sugiere que los datos se desvían fuertemente de una distribución normal, además se obtiene un valor p menor al nivel de significancia propuesto. A partir de esto se aprecia que los datos no son extraídos de una distribución normal. En cuanto a la prueba de homogeneidad se utilizaron las tablas de la distribución chi-cuadrado para evaluar el nivel de significancia (0,05) de la prueba Kruskal-Wallis. Al comparar cada valor respecto a los grados de libertad se observa que rechazan la hipótesis nula, en cada una de las décadas, lo que permite inferir que los datos tienen un comportamiento no homogéneo.

Además, el valor p confirma este planteamiento. La prueba de estacionalidad e independencia de Wald-Wolfowitz muestra un gran número de rachas, en cada década analizada, lo que indica que la hipótesis nula se debería rechazar, suponiendo así que los datos no provienen de la misma muestra.

Como se aprecia es factible realizar el ajuste de la función de probabilidad al conjunto de datos, ya que cumple con los supuestos necesarios para la realización del procedimiento. Cabe resaltar que se obtienen los estadísticos descriptivos para cada una de las estaciones, como también, para cada una de las décadas. Estos valores se pueden apreciar en las tablas 4 y 5.

Al modelar las funciones de probabilidad de los datos el ajuste realizado en las estaciones meteorológicas permite identificar a la distribución Gamma. Se realizó una prueba de bondad de ajuste para cada estación considerada y de igual manera se construyeron las gráficas que permiten evaluar el ajuste obtenido. Sin embargo, al considerar colocar cada una de las pruebas de bondad de ajuste como sus graficas correspondientes, resulta dispendioso mostrar cada uno de estos análisis. Es por esto que se construye la tabla 4, para poder visualizar de manera sencilla los resultados obtenidos. Para interpretar la tabla 4 es necesario entender que se encuentra divida en dos secciones. En la primera parte encontramos la información relacionada con los datos originales (sin imputar), se observa la numeración asignada a cada estación, seguido por los estadísticos de tendencia central, estadísticos de dispersión y estadísticos de forma.

A continuación, se expone la función de probabilidad que mejor describe a los datos y sus respectivos parámetros. En la segunda parte se encuentran los mismos criterios pero de los datos imputados seleccionados. En cuanto al análisis por década, resulta más fácil representar los datos obtenidos, este análisis se llevará a la par con la evaluación de los métodos de imputación ya que uno de los criterios a evaluar fue considerar cambios posibles en las funciones de probabilidad.

Para la realización de las imputaciones fue necesario considerar aspectos precisos para cada método de imputación. En cuanto a la imputación múltiple se consideró realizar m=5 estimaciones para completar el conjunto de datos faltantes. Se obtienen 5 conjuntos de datos completos, los cuales se unen calculado la media de los valores imputados por grupo de conjuntos. Al final se obtiene un solo conjunto de datos completos, conformado por la media de los valores imputados.

Así mismo, para el método de imputación KNN, fue necesario especificar el valor de k, para la primera década k=10, en la segunda década k=12, en la tercera década k=11 y en la última década k=8. Finalmente, en la imputación por suavizado de Kalman es necesario especificar la utilización de dicho método, para ello se utilizó, dentro de la función na.kalman la opción "*smooth=TRUE*".

Después de realizar el proceso de imputación se procedió a evaluar cada uno de los métodos. Inicialmente, se confrontan los resultados obtenidos por estación, calculando la convergencia de los estadísticos descriptivos respecto a los datos originales y los datos imputados. Se comparan los estadísticos de las pruebas de bondad de ajuste y los parámetros de las funciones de probabilidad, con el fin de identificar los conjuntos más semejantes. En la tabla 4 se presenta el conjunto de datos imputados que obtuvo la mayor convergencia, siendo este, el método de imputación múltiple (Amelia).

La tabla 5 representa los resultados de los métodos de imputación para el análisis decadal de los datos. Al interpretar su contenido es posible determinar que método tienen mejor desempeño para el paquete de datos estudiados. Las celdas sombreadas resaltan aquellos valores que obtuvieron mayor convergencia respecto a los valores originales, lo cual ayuda a identificar que método proporciona un valor más cercano al real. En la parte inferior, se encuentra sombreada, la selección del método que mejor ajuste obtuvo para la década analizada. Contigua a esta tabla de estadísticos descriptivos, se

encuentran la numeración de los diferentes estadísticos considerados en las pruebas de bondad de ajuste para el modelamiento de las funciones de probabilidad. De igual manera, se encuentran sombreados aquellos valores que obtuvieron un mejor ajuste respecto a la data original. En algunos casos existe igualdad entre los métodos de imputación, es por esto que se observa en las zonas inferiores la selección de dos métodos de imputación. Los criterios considerados para escoger el "mejor" modelo fueron observar los estadísticos por método, y considerar aquellos que tuvieron un mejor ajuste, respecto al valor original, entre más estadísticos obtuvieron, mejor se consideró el método. Es por esto que puede existir la posibilidad de que dos métodos tuvieran igual número de estadísticos ajustados, considerando que los dos métodos son adecuados, en cuanto a la evaluación se refiere. Partiendo de esto, para la década 1974-1983, el método imputación que mejor se adapta al conjunto de datos fue el de imputación múltiple, pero en la evaluación del test de bondad de ajuste se observa una igualdad con el modelo de suavizado de Kalman. Para la década 1984-1993 se obtuvo un equilibrio en los criterios descriptivos y en el test de bondad de ajuste de los métodos Amelia y Kalman. En cuanto a la década 1994-2003 el método que mejor convergencia tuvo fue Amelia, tanto para los criterios descriptivos como para el test de bondad de ajuste.

Tabla 4. Análisis exploratorio de los registros: original e imputado por medio de Amelia

s	Registro original con datos faltantes									Registro con datos imputados por medio de Amelia								
	μ	x	σ	σ^2	g_2	A_5	dt	PAR. s	r	μ	X	σ	σ2	g2	A_5	dt	PAR. s	R
1	23,0	20,0	17,6	310,3	3,3	46	G	56,3	5,8	22,7	19,0	18,0	292,3	3,5	1,6	G	54,2	5,6
2	20,1	18,1	12,3	151,2	2,5	31	G	67,1	6,2	20,2	19,0	11,7	136,5	2,8	1,2	G	146,1	13,8
3	19,2	17,6	9,9	97,2	4,2	22	G	47,2	3,2	18,8	17,0	9,5	89,8	4,2	1,4	G	60,8	4,1
4	16,8	15,5	10,3	106,4	5,2	10	G	52,6	5,7	16,8	15,5	10,2	104,6	5,3	1,5	G	95,3	10,7
5	19,1	17,8	10,1	101,3	0,5	16	G	66,9	4,8	19,2	18,0	9,9	9,8	0,6	0,7	G	261,6	17,9
6	21,2	19,0	13,7	187,0	5,9	5	G	67,2	6,8	21,2	19,0	13,6	185,2	6,0	1,7	G	79,9	8,1
7	21,3	20,0	10,5	110,5	0,4	10	G	79,6	4,0	21,3	20,1	10,4	108,0	0,4	0,6	G	52,6	2,7
8	22,0	20,6	11,1	122,2	7,3	54	G	46,9	2,3	22,3	21,1	10,0	99,9	8,6	1,8	G	76,4	3,7
9	19,3	17,0	12,1	146,6	14,2	29	G	92,0	6,5	19,3	17,0	11,8	139,2	14,9	2,3	G	119,8	8,5
10	18,4	16,8	10,6	112,8	3,7	6	G	69,9	4,6	18,4	16,8	10,6	111,6	3,8	1,4	G	81,9	5,6
11	18,9	17,6	11,4	130,8	2,2	3	G	91,7	6,4	18,9	17,8	11,4	130,2	2,2	1,1	G	167,4	11,4
12	18,2	19,1	10,4	108,8	10,3	7	G	87,8	4,7	19,2	18,5	10,4	108,8	10,1	1,6	G	115,2	6,0
13	18,2	15,9	12,4	152,6	5,3	42	G	84,1	5,5	18,6	15,9	11,8	141,4	5,9	1,8	G	93,0	6,1
14	18,1	15,8	12,7	160,6	10,2	7	G	207,6	20,0	18,2	16,0	12,6	159,2	10,2	2,3	G	201,1	19,4
15	16,4	14,5	9,9	97,1	2,2	25	G	52,3	3,9	16,5	14,8	9,6	92,1	2,3	1,1	G	61,7	4,3
16	17,3	15,7	10,7	114,4	1,9	5	G	114,0	11,7	17,3	16,0	10,7	113,5	1,9	1,0	G	54,9	5,6
17	18,1	16,3	10,5	110,8	0,5	37	G	50,8	3,5	18,3	17,0	10,0	99,0	0,7	0,7	G	70,3	4,5
18	20,9	20,0	11,9	141,7	2,8	7	G	38,1	2,4	20,9	20,0	11,8	140,1	2,8	1,1	G	60,8	3,3
19	17,1	14,5	13,5	181,3	12,4	26	G	197,5	20,0	17,1	15,0	13,2	173,0	13,0	2,8	G	185,7	18,8
20	15,3	13,1	10,4	109,0	0,6	34	G	315,1	30,9	15,3	13,4	10,2	103,2	0,7	0,9	G	82,8	8,0
21	16,8	15,7	12,0	144,3	48,0	64	G	68,0	4,3	17,9	16,5	12,1	146,6	39,0	3,8	G	71,5	4,5
22	20,8	20,5	13,1	171,4	1,5	35	G	49,4	4,6	20,9	20,8	12,5	156,9	1,8	0,8	G	345,6	15,1
23	44,3	43,5	23,4	545,8	-0,1	45	G	52,0	1,2	44,5	44,3	22,2	492,8	0,1	0,4	G	57,8	1,3
24	20,2	19,0	11,1	122,2	1,4	16	G	114,4	7,1	20,2	19,2	10,9	118,6	1,5	0,8	G	60,9	3,5
25	19,9	18,5	14,1	198,0	5,1	20	G	52,0	2,7	19,9	18,5	13,8	191,0	5,4	1,6	G	163,9	8,2
26	60,0	60,0	38,3	1464,3	-0,6	38	G	30,5	0,5	60,5	59,7	36,6	1.340,8	-0,4	0,2	G	34,6	0,6
27	61,0	58,0	40,2	1614,1	-0,4	22	G	26,7	3,7	61,1	59,0	38,6	1.486,8	-0,2	0,3	G	0,0	0,0
28	63,0	58,9	42,2	1784,7	2,4	53	G	2305,2	576,3	64,1	60,0	40,6	1.645,1	2,5	0,9	G	4585,4	1146,3
29	63,5	59,0	42,9	1843,6	2,1	37	G	104,5	3,0	63,3	58,5	41,6	1.729,6	2,4	1,2	G	35,4	1,0
30	16,3	15,7	9,5	90,0	1,5	48	G	76,9	4,5	16,3	15,9	9,1	82,4	1,9	0,9	G	81,2	4,7
31	14,8	14,5	8,5	72,0	0,6	32	G	97,6	6,4	14,8	14,5	8,3	69,0	0,7	0,6	G	0,1	0,0
32	35,7	35,5	17,8	315,3	0,1	18	G	48,7	1,7	35,7	35,6	16,3	265,6	0,2	0,4	G	39,8	1,4
33	18,0	17,0	11,5	131,4	4,6	19	G	191,6	17,3	18,1	17,0	11,3	127,3	4,7	1,4	G	127,4	11,6
34	17,0	16,0	10,8	117,6	2,6	4	G	66,3	5,4	17,1	16,0	10,8	116,0	2,7	1,1	G	71,6	5,8
35	17,0	15,6	9,6	92,5	1,6	2	G	45,8	5,4	17,1	15,7	9,6	91,6	1,6	1,0	G	85,9	7,4
36	16,2	14,5	9,9	98,4	1,2	9	G	73,3	5,6	16,2	14,7	9,8	96,9	1,2	1,0	G	69,7	5,3
37	18,4	16,3	11,5	133,2	8,8	36	G	30,1	2,5	18,7	16,8	11,1	123,1	9,2	2,0	G	79,4	5,0
38	15,7	14,3	9,9	97,9	1,5	3	G	36,3	2,7	15,7	14,5	9,8	97,0	1,5	0,8	G	53,4	3,9
39	13,3	12,2	7,9	62,8	2,1	16	G	145,0	11,8	13,4	12,2	7,9	61,7	2,1	1,0	G	174,2	14,3
40	15,2	13,5	9,0	80,2	2,9	0	G	123,5	10,5	18,6	17,1	9,7	94,2	1,0	0,8	G	90,4	6,2
41	18,6	17,0	9,8	96,3	0,9	14	G	146,7	10,0	31,2	29,6	15,5	240,4	1,8	0,8	G	118,0	4,2
42	31,2	29,6	15,9	253,4	1,6	31	G	62,3	2,2	18,9	16,9	12,5	155,8	8,1	2,0	G	65,1	4,5
43	18,7	16,5	12,7	160,9	8,2	29	G	119,7	9,2	23,3	21,0	15,2	232,2	32,5	3,6	G	66,7	3,4
44	23,5	21,0	15,6	242,3	31,5	20	G	96,2	4,8	32,8	30,0	21,4	457,9	9,4	1,9	G	213,6	7,0
45	32,8	30,0	21,5	462,2	9,3	6	G	80,4	2,6	26,4	25,0	14,3	203,2	1,8	0,9	G	77,4	3,2
46	26,6	25,5	14,8	218,8	1,5	38	G	48,2	2,0	22,6	20,4	11,9	142,0	2,2	1,3	G	45,1	2,6
47	22,5	20,0	12,1	147,0	2,1	11	G	43,7	2,5	46,3	44,4	23,3	542,1	0,2	0,5	G	40,4	1,0
48	46,2	43,7	24,7	609,2	0,0	65	G	42,3	1,2	52,4	51,0	34,4	1.183,3	-0,2	0,4	G	32,9	1,0
49	51,6	50,0	36,1	1300,3	-0,3	57	G	47,4	2,1	22,7	19,0	18,0	292,3	3,5	1,6	G	54,2	5,6

E= Estaciones IDEAM μ=Media x=mediana σ= Desviación estándar σ=varianza g2=Curtosis A5= coeficiente de asimetría Par=parámetros s= shape r = rate dt=distribución teórica G = Gamma L = Logistic C =Cauchy

1.Tutazá 2.La Chapa 3.Beteitiva 4.Tasco 5.Cerinza 6.SantaR. 7.La Rusia 8.Andalucia 9.Duitama 10.Surbata. 11.Nobsa 12.Nimicia 13.Mongua 14.Mongui 15.Sena 16.Firavitoba 17.Tibasosa 18.Cerezo 19.Garrocho 20.Iza 21.Tunel 22.Guamos 23.Corinto 24.Cintas 25.Toquilla 26.El morro 27.Yopal 28.Pajarito 29.Chameza 30.Potrerito 31.Pesca 32.Cazadero 33.CasaAma. 34.El hotel 35.Jenesano 36.S.Cristobal 37.VillaLuisa 38.S.Pedro 39.Siachoqu 40.Uptc 41.Ramiriquí 42.Rondon 43.Tibana 44.Chinavita 45.Buenavista 46.El Vivero 47.Teatinos 48.Paez 49.La pradera

Fuente: autores

En la década 2004-2013 el método de suavizado de Kalman obtuvo un ajuste deseable para ambas consideraciones. Se escoge el método de Amelia ya que presento los mejores ajustes gráficos, además los parámetros en las distribuciones tuvieron los ajustes más cercanos, respecto a los datos originales. En general, se presenta un empate entre los métodos Amelia y suavizado de Kalman, se selecciona el método de imputación múltiple, debido a las consideraciones matemáticas que desarrolla, así también, cuantifica la incertidumbre del modelo imputado.

Tabla 5. Análisis de la imputación de los datos por década

Década 1974-1983										**Década 1984-1993**									
ME	Orig	Amel	KNN	Kal		Distribución de Probabilidad				ME	Orig	Amel	KNN	Kal.		Distribución de Probabilidad			
					Eb	Distribución = Gamma									Eb	Distribución = Gamma			
μ	25,7	26,4	28,5	26,0		Orig	Amelia	KNN	kal	μ	23,29	23,08	24,10	23,92		Orig	Amelia	KNN	kal
E	0,390	0,391	0,392	0,386	1	-155,1	-171,82	-181,7	162,4	E	0,284	0,287	0,289	0,258	1	-188,7	-195,9	-198,3	-197.3
X	19,0	20,0	20,4	20,6	2	-6,08	-6,29	-6,40	-6,17	X	17,7	18,3	18,5	17,0	2	-6,48	-6,55	-6,58	-6,57
x̄	20,0	20,0	20,0	20,0	3	6,65	3,64	3,53	3,75	x̄	20,0	20,0	20,0	20,0	3	6,63	6,67	6,65	6,661
σ	26,39	26,45	26,48	26,11	4	483,14	1086,7	1210	474,2	σ	21,13	21,38	21,49	21,22	4	394,34	485,3	473,3	403,4
σ	696,9	699,9	701,5	681,6	5	0,116	0,223	0,236	0,114	σ	446,6	457,4	462,0	450,4	5	0,093	0,117	0,114	0,096
g2	15,2	14,0	14,0	15,1	6	inf	inf	inf	inf	g2	8,08	7,06	6,96	7,30	6	inf	inf	inf	inf
AS	3,18	3,04	3,04	3,15	7	0	0	1,32E^{-07}	1,32E^{-07}	AS	2,47	2,31	2,30	2,34	7	1,8E^{-07}	1,8E^{-07}	1,8E^{-07}	1,8E^{-07}
m	0,0	0,0	0,0	0,0	8	0,99	0,99	0,99	0,99	Mn	0,0	0,0	0,0	0,0	8	0,99	0,999	0.99	0,99
M	315,0	315,0	315,0	315,0	9	2,20E^{-16}	2,20E^{-16}	2,20E^{-16}	2,20E^{-16}	Mx	181,3	181,3	181,3	181,3	9	2,2E^{-16}	2,2E^{-16}	2,2E^{-16}	2,2E^{-16}
Amelia				Kalman						Kalman				Amelia -Kalman					
Década 1994-2003										**Década 2004-2013**									
ME	Orig	Amel	KNN	Kal		Distribución de Probabilidad				ME	Orig	Amel	KNN	Kal		Distribución de Probabilidad			
					Eb	Distribución = Gamma									Eb	Distribución = Gamma			
μ	24,93	25,01	24,95	25,23		Orig	Amelia	KNN	kal	μ	25,13	25,02	24,83	24,89		Orig	Amelia	KNN	kal
E	0,291	0,288	0,287	0,296	1	-181,26	-187,78	-195,1	-201,1	E	0,289	0,283	0,283	0,281	1	-168,1	-166,9	-189,9	-183,3
X	19,6	19,601	19,603	19,608	2	-6,39	-6,471	-6,547	-6,607	X	20	20	19,7	19,8	2	-6,25	-6,24	-6,49	-6,42
x̄	20	20	20	20	3	6,618	6,67	6,596	6,536	x̄	10	10	10	10	3	6,667	6,814	6,557	6,627
σ	21,20	20,927	20,922	21,28	4	1056,8	442,18	382,5	1546,4	σ	20,53	20,12	20,156	20,17	4	506,7	341,9	694,1	555,4
σ	449,7	437,9	438	452,8	5	0,217	0,106	0,090	0,270	σ	421,72	404,7	406,29	406,8	5	0,12	0,08	0,16	0,13
g2	7,352	7,379	7,373	6,889	6	inf	inf	inf	inf	g2	6,188	6,450	6,423	6,431	6	inf	inf	inf	inf
AS	2,311	2,308	2,306	2,254	7	1,8E^{-07}	1,8E^{-07}	1,8E^{-07}	1,8E^{-07}	AS	2,06	2,09	2,09	2,1	7	1,19E^{-7}	1,19E^{-7}	1,19E^{-7}	1,19E^{-7}
m	0	0	0	0	8	0,99	0,99	0,9	0,99	Mn	0	0	0	0	8	0,999	0,999	0,999	0,999
M	207	207	207	207	9	2,2E^{-16}	2,2E^{-16}	2,2E^{-16}	2,2E^{-16}	Mx	190,8	191	191	191	9	2,2e^{-16}	2,2e^{-16}	2,2e^{-16}	2,2e^{-16}
Amelia				Amelia						Amelia-Kalman				Kalman					

ME= Medidas estadísticas μ= media E =Error típico X = mediana x̄= moda σ = desviación estándar σ2 = varianza de la muestra
g2 = Curtosis As = coeficiente de asimetría m = Mínimo M = Máximo Orig= Datos originales Amelia=Datos imputados con Amelia
KNN= Datos imputados con KNN kal= Datos imputados con Kalman Eb=Estadísticos necesarios para la prueba de bondad de ajuste
1=Log-L 2=AIC 3=BIC 4=CVM(value) 5=CVM(p-value 6= AD(value) 7= AD(p-value) 8=KS(value)9 9=KS(p-value)

Fuente: autores

En cuanto al análisis que se realizó al conjunto de datos por década se obtienen los siguientes resultados: en la Figura 2 (década 1974-1983), inicialmente, se exponen las gráficas de las pruebas de bondad de ajuste realizada a los datos originales. En la Figura 2-a) se observa el histograma de distribución de densidad empírica y teórica. Puede apreciarse que su forma empírica (barras) se acopla sin ningún problema a la forma teórica (línea roja), lo mismo sucede con la gráfica 2-e), la cual representa el histograma de distribución de densidad de los datos imputados por el método Amelia.

En cuanto a las figuras 2-b) y 2-f) representan la comparación de la distribución empírica respecto a la distribución Gamma. Puede observarse que no existe una diferencia considerable en las comparaciones, para ninguno de los dos conjuntos de datos. En las figuras 2-c) y 2-g) se presentan las funciones de distribución acumuladas, las cuales no difieren mucho en su representación para datos originales y datos imputados.

Por último, las figuras 2-d) y 2-h) representan las gráficas boxplot, las cuales permiten observar diferencias mínimas en algunos valores máximos de los valores imputados respecto a los originales. En la Figura 3 (década 1984-1993), nos muestra las gráficas obtenidas en la prueba de bondad de ajuste, tanto para los datos originales como para los datos imputados. Se logra observar diferencias mínimas en las gráficas 3-b) y 3-f). La grafica 3-h) respecto a la 3-d), tiene un leve cambio en la estación Pajarito, lo que permite intuir que existe mayor variabilidad en los datos imputados. Esto se debe a que en esta estación se presentan grandes periodos de lluvias intensas, al estar ubicada en zona de selva tropical, lo mismo sucede en la estación La Pradera.

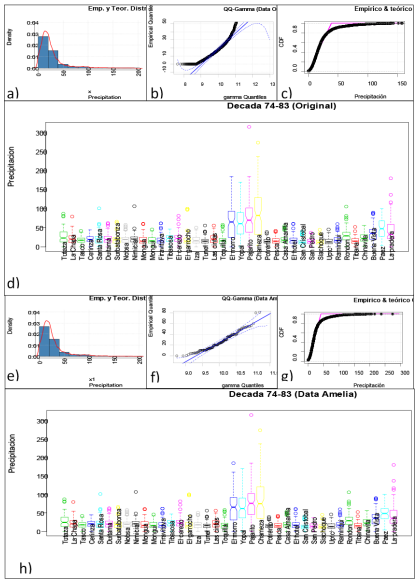

Figura 1. Análisis de la imputación para la década de 1974-1983. Fuente: Autores

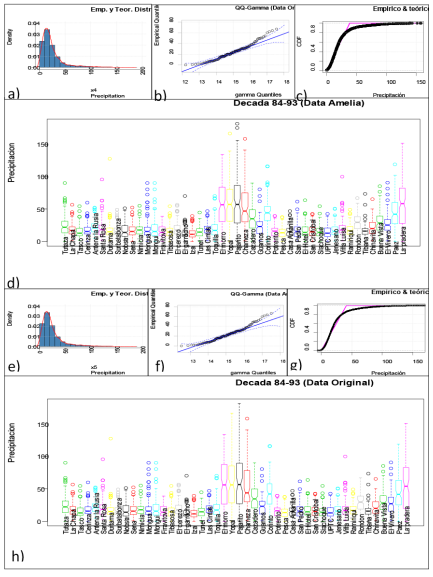

Figura 3. Análisis de la imputación para la década de 1984-1993. Fuente: Autores

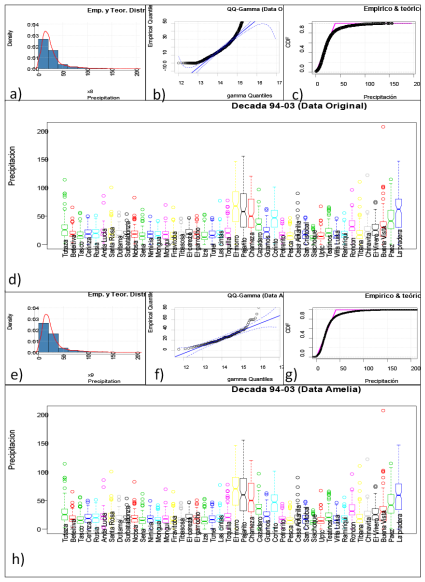

Figura 4. Análisis de la imputación para la década de 1994-2003. Fuente: Autores

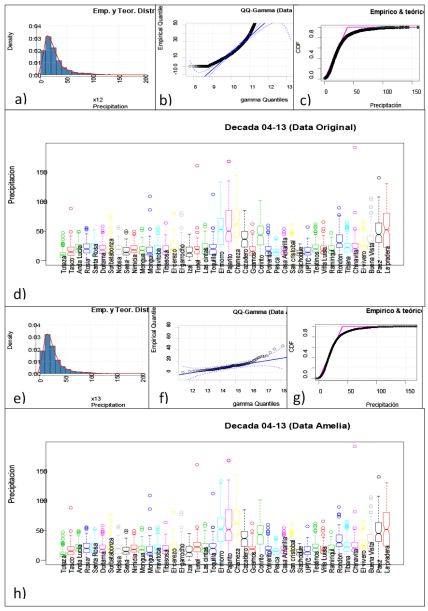

Figura 5. Análisis de la imputación para la década de 2004-2013. Fuente: Autores

En la Figura 4 (década 1994-2003), se observa un cambio sustancial en la gráfica 4-f) respecto a la gráfica 4-b), debido a la gran cantidad de valores máximos que se presentaron en el periodo analizado. Puede verse esto en las gráficas 4-d) y 4-h) en donde existen estaciones con más de dos valores atípicos. Estos valores atípicos resultan en un inconveniente al momento de calcular probabilidades en un conjunto datos. En la Figura 5 (década 2004-2013), se puede apreciar el aumento de datos atípicos en zonas de alta montaña, lo que genera el mismo efecto anterior, en las gráficas 5-b) y 5-f). Puede observarse que para las zonas cercanas al piedemonte pese a tener precipitaciones altas en las cuatro décadas no se visualiza discrepancia entre las gráficas de los valores originales y los valores imputados. En cuanto a las gráficas 5-a) y 5-e) puede notarse un leve cambio en la forma de la campana de la función de densidad teórica, debido a los registros de los datos que fueron imputados. Estos valores debieron tener un valor menor a la media del conjunto de datos.

DISCUSIÓN Y CONCLUSIÓN

Se logra identificar el mecanismo de pérdida de los datos (MAR), a partir de este, se escogen tres métodos de imputación que consideren el mecanismo de pérdida encontrado. Para cada método de imputación (imputación múltiple, KNN, Suavizado de Kalman) se elige un enfoque (aleatorio, espacial, temporal). Las pruebas aplicadas permiten inferir que los datos no provienen de una distribución normal, además no se consideran homogéneos. Por otro lado las muestras son independientes y no muestran un factor estacional. Se caracterizó la función de distribución de probabilidad del conjunto de datos, siendo esta la distribución Gamma. Se elige el método de imputación múltiple (Amelia) para completar las series de precipitación máxima en la zona central de Boyacá. Al comparar los resultados obtenidos con los estudios realizados en la zona, se observa que ningún estudio había caracterizado el mecanismo de pérdida de los datos, ni la función de distribución de probabilidad de los registros de precipitación. Tampoco se había planteado una metodología para la evaluación de la imputación.

Cabe mencionar que existen falencias al momento de elegir los mecanismos de pérdida, ya que no existen procedimientos para evaluar las hipótesis de los mecanismos MAR y MNAR. Por esta razón, es importante conocer bien el área de estudio, para así poder plantear hipótesis que permitan rechazar alguno de los mecanismos mencionados. Se recomienda aplicar el estudio en zonas con registros homogéneos, con el fin de identificar posibles fallas que puedan ocurrir al momento de evaluar los métodos de imputación. Se concluye que el mecanismo de pérdida de los datos es MAR, la mejor técnica que se adapta a la naturaleza de los datos de precipitación máxima mensual es la imputación múltiple, pues tiene los mejores índices estadísticos y el mejor ajuste gráfico. Se identificó la distribución de los datos, siendo esta la distribución Gamma. Se recomienda ampliar el conjunto de datos para comprobar la confiabilidad del método propuesto, además de aplicar la metodología propuesta en otras zonas geográficas.

REFERENCIAS

[1] C. Segerer y R. Villodas, HIDROLOGIA I,Unidad 5: Las Precipitaciones, Mendoza, Argentina: Universidad Nacional de Cuyo, Facultad de Ingeniería. Ingeniería Civil, 2006.

[2] O. M. M. OMM, «Hidrología – De la medición a la información hidrológica.,» *Guía de prácticas hidrológicas. Ginebra: Organización Meteorológica*, vol. Volumen I., nº OMN-Nº 168, 6ta. ed. , 2011.

[3] A. Hurtado y Ó. Mesa, «*Reanalysis of monthly precipitation fields in Colombian territory*,» *DYNA, 2014. ISSNelectrónico2346-2183.ISSNimpreso0012-7353.* , pp. Volumen 81, Número 186, p. 251-258, 2014.

[4] D. Carrera, P. Guevara, L. Tamayo, A. Balarezo, C. Narváez y D. Morocho, «Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media,» IDESIA Volumen 34, Nº 3. Páginas 81-90, Chile, 2016.

[5] R. J. Little A. y D. Rubin B., Statistical Analysis With Missing Data, Hoboken, New Jersey: John Wiley & Sons, 1987.

[6] WMO, Some Methods of Climatological Analysis, Ginebra, Zuisa: Secretariat of the World Meteorological Organization, 1966.

[7] CEPAL, Estudios estadísticos y prospectivos. Imputación de datos: Teoría y práctica, Santiago de Chile: Publicación de las Naciones Unidas, 2007.

[8] J. L. Schafer, Analysis of Incomplete Multivariate Data, Boca Raton, Florida: Chapman & Hall/CRC, 1997.

[9] D. R. Dawdy y R. W. Lichty, «*Methodology of hydrologic model building*,» *Proceedings, use of analog and digital computers in hydrology,* vol. 2, pp. 347-355, 1968.

[10] R. P. Rosario A., «Aplicación de algunos métodos de relleno a series anuales de lluvia de diferentes regiones de Costa Rica,» *Tópicos Meteorológicos y Oceanográficos*, vol. 7, nº 1, pp. 1-20, 2000.

[11] R. Lo Presti, E. Barca y G. Passarella, «*A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy),*» *Environ Monit Assess*, vol. 160, pp. 1-22, 2010.

[12] F. Merlos. V, S. T. Sánchez Q., J. A. Almanza C. y C. Domínguez S., «Evaluación de la gestión de datos para estudios hidrológicos,» *III Congreso Naciononal de Manejo de Cuencas Hidrograficas*, pp. 368-379, 2013.

[13] M. E. Fernández L. y M. R. Antelo, «Estimación de datos faltantes de precipitación diaria para las distintas ecorregiones de la República Argentina,» 2do Encuentro de Investigadores en Formación en Recursos Hídricos, Ezeiza, 2014.

[14] P. V. Guevara G., D. V. Carrera V., L. C. Tamayo B., A. L. Balarezo A., C. A. Narváez R. y D. R. Morocho L., «Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media,» *IDESIA*, vol. 34, nº 3, pp. 81-90, 2016.

[15] C. M. Ilbay Y., K. Fonseca L., A. Quichimbo M., R. Lara L. y J. Tiche T., «Estimación de datos faltantes de precipitación en la subcuenca del Río Patate,» Revista Bases de la Ciencia , vol. 2, nº 3, pp. 37-48, 2017.

[16] C. S. Herrera O., J. R. Campos G. y F. M. Carrillo G., «Estimación de datos faltantes de precipitación por el metodo de regresión lineal: Caso de estudio Cuenca Guadalupe, Baja California, México,» *Redalyc*, vol. 25, nº 71, pp. 34-44, 2017.

[17] L. Useche y D. Mesa, «Una introducción a la imputación de valores perdidos,» *Terra Nueva Etapa*, vol. 12, nº 31, pp. 127-151, 2006.

[18] S. Infante , J. Ortega y F. Cedeño, «Estimación de datos faltantes en estaciones meteorológicas de Venezuela vía un modelo de redes neuronales,» *Revista de Climatología*, vol. 8, pp. 51-70, 2008. V. Jimenez,

[19] A. Will, S. Rodríguez y C. Lamelas, «Imputación de datos climáticos utilizando algoritmos genéticos niching,» *Acta de la XXXVII Reunión de Trabajo de la Asociación Argentina de Energías Renovables y Medio Ambiente ,* vol. 2, pp. 11139-11148, 2014.

[20] M. Benítez G. y M. Álvarez C., «Reconstrucción de series temporales en ciencias ambientales,» *Revista Latinoamericana de Recursos Naturales*, vol. 4, nº 3, pp. 326-335, 2008.

[21] C. Guevara O., N. Briceño, E. Zimmermann, L. Vives, M. Blanco, G. Cazenave y G. Ares, «Relleno de series de precipitación diaria para largos periodos de tiempo en zonas de llanura. Caso de estudio cuenca superior del arroyo del Azul,» Geoacta, vol. 42, nº 1, pp. 38-62, 2017.

[22] A. J. Peña Q., H. A. Chica R., J. F. Giraldo J., D. Obando B. y N. M. Riaño H., «SueMulador: Herramienta para la simulación de datos faltantes en series climáticas diarias de zonas ecuatoriales,» *Revista Facultad Nacional de Agronomía Medellín*, vol. 67, nº 2, pp. 7365-7373, 2014.

[23] L. Ingsrisawang y D. Potawee, «Multiple imputation for missing data in repeated measurements using MCMC and Copulas,» *Proceedings of the internacional multiconference of engineers and computer scientists*, vol. II, pp. 1-5, 2012.

[24] S. T. Escobar C., H. O. González P., H. F. Aristizabal R. y Y. Carvajal E., Aplicación de técnicas estadísticas en las series climatológicas mensuales totales de precipitación, evaporación y brillo solar, con el fin de corregir, completar y verificar la calidad de la información, Primera ed., Santiago de Cali: Corporación Autónoma regional del Valle del Cauca, 2005.

[25] R. D. Medina R., E. C. Montoya R. y Á. Jaramillo R.,

«Estimación estadística de valores faltantes en series históricas de lluvia,» *Cenicafé*, vol. 59, nº 3, pp. 260-273, 2008.

[26] J. A. Urrutia, R. Palomino y H. D. Salazar, «Metodología para la imputación de datos faltantes en meteorología,» *Scientia Et Technica*, vol. XVII, nº 46, pp. 44-49, 2010.

[27] D. A. Castro Ll. y Y. Carvajal E., «Análisis de tendencia en la precipitación pluvial anual y mensual en el departamento del Valle del Cauca,» *Memorias*, vol. 11, nº 20, pp. 9-17, 2013.

[28] P. L. García R., «Imputación de datos en series de precipitación diaria caso de estudio cuenca del Río Quindío,» *Ingeniare*, vol. 10, nº 18, pp. 73-86, 2015.

[29] J. Leal R. y M. E. Rivera , «Estimación de datos faltantes de precipitación de la estación meteorológica ISER Pamplona, Colombia,» *Revista Ingenieros Militares*, nº 11, pp. 83-89, 2016.

[30] E. M. Caicedo, Water Quality Assessment of Lake Tota using a 3D modelling approach, Primera ed., Delft: UNESCO-IHE, 2016.

[31] C. Gonzáles M., Impactos de la variabilidad climática y las actividades humanas en la dinámica hidrológica del Lago de Tota, Primera ed., Medellín: Universidad de Antioquia, Facultad de Ingeniería , 2016.

[32] R Foundation, «The R Project for Statistical Computing,» 1993. [En línea]. Available: https://www.r-project.org/about.html. [Último acceso: 19 Enero 2019].

[33] R Foundation , «Cran.r-project,» 2004. [En línea]. Available: https://cran.r-project.org/. [Último acceso: 19 Enero 2019].

[34] «R Core Team,» [En línea]. Available: http://www.R-project.org.

[35] D. B. Rubin, «Inference and missing data,» Biometrika, vol. 63, pp. 581-592, 1976.

[36] J. Gómez G., J. Palarea A. y J. Matín F., «Métodos de inferencia estadística con datos faltantes. Estudio de simulacion sobre los efectos en las estimaciones,» *Estadística Española*, vol. 48, nº 162, pp. 241-270, 2006.

[37] M. J. Rodríguez J. y R. Mora C., Estadística Informática, Ilustrada ed., Alicante: Universidad de Alicante. Servicio de publicaciones, 2001.

[38] S. S. Shapiro y R. S. Francia, «An Approximate Analysis of Variance Test for Normality,» *Journal of the American Statistical Association*, vol. 67, pp. 215-216, 1972.

[39] P. Royston, «A pocket-calculator algorithm for the shapiro-francia test for non-normality: An application to medicine,» *Statistics in Medicine*, vol. 12, pp. 181-184, 1993.

[40] W. H. Kruskal y A. W. Wallis, «Use of Ranks in One-Criterion Variance Analysis,» *Journal of the American Statistical Association*, vol. 47, pp. 583-621, 1952.

[41] A. Wald y J. Wolfowitz, «On the test whether two samples are from the same population,» *The annals of Mathematical Statistics*, vol. 11, pp. 147-162, 1940.

[42] B. Ycart y C. Robert, «Statistique Médicale En ligne,» Université Paris Descartes, Paris, Francia, 2018.

[43] R. Fisher, «On the Mathematical Foundations of Theoretical Statistics,» *Philosophical Transactions of the Royal Society of London*, vol. 222, nº Series A, pp. 309-368, 1922.

[44] F. F. Caballero D., Selección de modelos mediante criterios de información en análisis factorial. Aspectos teóricos y computacionales, Granada, España: Universidad de Granada, 2011.

[45] D. Evans, J. Drew y L. Leemis, «The Distibution of the Kolmogorov-Smirnov, Cramer-Von Misses, and Anderson-Darling Test Statistics for Exponential Populations with Estimated Parameters,» *Taylor & Francis Group*, vol. 37, pp. 1396-1421, 2008.

[46] J. Wayman, «Multiple Imputation For Missing Data: What Is It And How Can I Use It?,» Annual Meeting of the American Educational Research, Chicago, 2003.

[47] D. B. Rubin, «Multiple imputation for non-response in surveys.,» Jhon Wiley $ Sons, New York, 1987.

[48] D. Otero G., Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo, Santiago de Compostela: Universidade de Santiago de Compostela, Universidade da Coruña, Universidade de Vigo, 2011.

[49] D. B. Rubin, «Multiple imputations in sample surveys. A phenomenological bayesian approach to non-response.,» *American Statistical Association*, pp. 20-34, 1978.

[50] D. B. Rubin, «Mutiple imputation after 18+ years.,» *Journal of the American Statistical Association.*, vol. 91, pp. 473-489, 1996.
[51] J. Honaker, G. King y M. Blackwell, «AMELIA II: A program for Missing Data,» Harvard, 2018.
[52] A. D. García U., Análisis de datos y búsqueda de patrones de aplicaciones médicas, Santiago de Chile: Universidad de Chile. Facultad de Ciencias Fisicas y Matemáticas., 2015.
[53] R. Aler M., Clasificadores KNN, Madrid: Universidad Carlos III de Madrid, 2015.
[54] J. Vadillo J., Procesado y análisis de datos procedentes de una máquina de extrusión de pólimeros, País Vasco: Euskal Herriko Unibertsitatea, 2018.
[55] G. Welch y G. Bishop, «An Introduction to the Kalman Filter,» SIGGRAPH, Berlin, 2001.
[56] J. Durbin y S. J. Koopman, Time Series Analysis by State Space Methods, Oxford, England: Oxford Statistical Science Series, 2012.
[57] A. Goicoechea P., «Imputación basada en árboles de decision de clasificación,» Eustat, Bilbao, 2002.