ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

COMPARACIÓN DE TÉCNICAS DE RELLENO DE DATOS FALTANTES DE LA VARIABLE VELOCIDAD DE VIENTO DE LOS AÑOS 2014 AL 2021

Trabajo de Integración Curricular

Tipo: Proyecto de Investigación

Presentado para optar el grado académico de:

INGENIERA/O EN ESTADÍSTICA INFORMÁTICA

AUTORES:

VELASTEGUI CUJILEMA EVELYN MISHELLE

HORNA ZHININ ERICK ADRIAN

DIRECTORA: ING. NATALIA ALEXANDRA PÉREZ LONDO, MGS.

Riobamba – Ecuador

2023

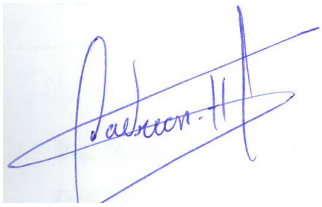
Nosotros, Evelyn Mishelle Velastegui Cujilema y Erick Adrian Horna Zhinin, declaramos que el presente trabajo de integración curricular es de nuestra autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autores asumimos la responsabilidad legal y académica de los contenidos de este trabajo de integración curricular; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 13 de diciembre de 2023

Evelyn Mishelle Velastegui Cujilema

065003320-2



Erick Adrian Horna Zhinin

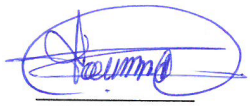
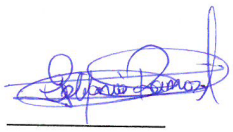
060577961-0

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

El Tribunal del Trabajo de Integración Curricular certifica que: el Trabajo de Integración Curricular; Tipo: Proyecto de Investigación. **COMPARACIÓN DE TÉCNICAS DE RELLENO DE DATOS FALTANTES DE LA VARIABLE VELOCIDAD DE VIENTO DE LOS AÑOS 2014 AL 2021**, realizado por la señorita: **EVELYN MISHELLE VELASTEGUI CUJILEMA** y el señor: **ERICK ADRIAN HORNA ZHININ**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Integración Curricular, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	FIRMA	**FECHA**
Ing. Johanna Enith Aguilar Reyes, Mgs. **PRESIDENTE DEL TRIBUNAL**		2023-12-13
Ing. Natalia Alexandra Pérez Londo, Mgs. **DIRECTORA DEL TRABAJO DE INTEGRACIÓN CURRICULAR**		2023-12-13
Ing. Cristina Estefanía Ramos Araujo, Mgs. **ASESORA DEL TRABAJO DE INTEGRACIÓN CURRICULAR**		2023-12-13

DEDICATORIA

La fortaleza que he tenido para culminar este trabajo va dedicado primeramente a Dios por regalarme vida, salud, sabiduría y poder alcanzar mi meta de ser Ingeniera Estadística, a mis padres Jorge e Isabel que día a día han estado conmigo apoyándome motivándome y no dejándome caer hasta lograrlo, a mi hijo Gabriel Sánchez que es quien me ha dado la suficiente motivación de querer superarme profesionalmente para poder brindarle un mejor futuro, a mi hermano Dennis que con su ejemplo de sacrificio y dedicación se logra alcanzar los propósitos que anhelamos.

Evelyn

La fe y la esperanza son cosas que han hecho que tenga mucha gratitud y amor donde este trabajo va dedicado a Dios, por darme fuerza, valentía y persistencia en todo momento, además de contar con mis padres que han sido mi eje fundamental para poder seguir en este largo camino a mi madre Mónica Zhinin y a mi padre Juan Horna que con su sacrificio. A cada uno de las personas que han confiado en mí en todo mi transcurso de formación académica quienes me han acompañado en este transcurso académico con sus ánimos y consejos de superación, enseñándome valores y afrontar la vida, convirtiéndome en la persona que soy hoy en día.

Adrian

AGRADECIMIENTO

En primera instancia queremos agradecer a la Ing. Natalia Pérez y a la Ing. Cristina Ramos por guiarnos en el transcurso del desarrollo de nuestro proyecto de investigación, a su vez a toda la planta docente de la carrera de Estadística que con el pasar de los semestres nos fueron forjando como profesionales. A todos nuestros compañeros/as con los/as que hemos compartido espacio de clase y vivencias dentro y fuera de la universidad. En especial a todos los docentes que nos inculcaron ese amor por la carrera y supieron plantearnos retos en los distintitos semestres.

Evelyn

Adrian

ÍNDICE DE CONTENIDO

CAPÍTULO II

CAPÍTULO III

CAPÍTULO IV

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

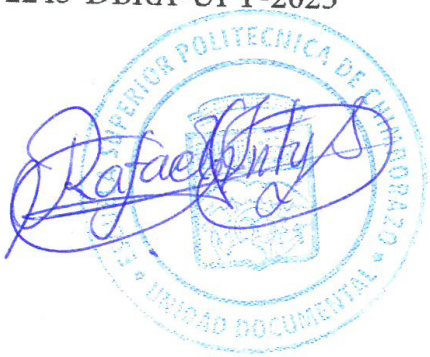
ÍNDICE DE ILUSTRACIONES

ÍNDICE DE ANEXOS

RESUMEN

En base al estudio meteorológico de una de las variables importantes que es la velocidad de viento que permitió la generación de fuentes eólicas siendo la energía que fue puesto a prueba para poder conocer su consistencia, a esto se agregó que los estudios en meteorología son muy cambiantes debido a distintas variables que se produjeron con el pasar del tiempo. El presente trabajo tuvo como objetivo principal comparar las técnicas de relleno de datos faltantes de la variable Velocidad de Viento de los años 2014 al 2021, en la provincia de Chimborazo. La matriz de datos de este estudio se obtuvo del Grupo de Energías Alternativas y Ambiente (GEAA), con un total de 703248 datos, donde a través de una depuración de información se tomó una muestra de 10964 datos faltantes en las 11 estaciones. El análisis estadístico descriptivo mostró una velocidad de viento mínima (1.000 m/s) en la estación Cumandá y una velocidad máxima (20.703 m/s) en la estación Tixán, mediante el Test de Rosner's Outlierse evidenció (677) datos atípicos. A partir de la función (*VIM.impute.pmm()*) se identificó datos faltantes, donde la mayor representatividad de los datos se encontraron en las estaciones Atillo, Multitud y Tunshi en los años de estudio. Con los métodos Random Forest, Media (MICE), Hot Deck e Iterative PCA Imputation se logró imputar y se realizaron los ajustes necesarios en cada estación, gráficamente se observó que Alao tuvo un buen ajuste entre los datos reales y los imputados, mediante las métricas EMP, EMC y DAM la estación Quimiag presentó valores bajos 0.0021, 0.0002 y 0.0021 respectivamente para las técnicas de la media (MICE) y Hot Deck siendo las que mejor se ajustan a los datos reales.

Palabras clave: <ESTADÍSTICA>, <IMPUTACÍON>, <RELLENO DE DATOS>, <METEREOLÓGICOS>, <MÉTRICAS>, <ESTACIONES>.



ABSTRACT

Based on the meteorological study of the wind speed, which is one of the important variables that allowed the generation of energy from wind sources, being the latter that was tested to be able to know its consistency, to this was added that the studies in meteorology are very changeable due to different variables that occurred with the passing of time. The main objective of this work was to compare the techniques for filling missing data of the variable Wind Speed from 2014 to 2021, in the province of Chimborazo. The data matrix of this study was obtained from the Alternative Energies and Environment Group (GEAA), with a total of 703248 data, where through a purification of information a sample of 10964 missing data was taken from the 11 stations. The descriptive statistical analysis showed a minimum wind speed (1,000 m/s) at the Cumandá station and a maximum speed (20,703 m/s) at the Tixán station; the Rosner's test for outliers revealed (677) outlier data. From the function ($VIM.impute.pmm()$) missing data were identified, where the most representative data were found at the Atillo, Multitud and Tunshi stations in the study years. With the Random Forest, Mean (MICE), Hot Deck and Iterative PCA Imputation methods, it was possible to impute and make the necessary adjustments in each station, graphically it was observed that Alao had a good adjustment between the real data and the imputed data, by means of the metrics EMP, EMC and DAM the Quimiag station presented low values 0.0021, 0.0002 and 0.0021 respectively for the mean (MICE) and Hot Deck techniques being those that best adjusted to the real data.

Keywords: <STATISTICS>, <IMPUTATION>, <DATA FILLING>, <METEOROLOGICAL>, <METRICS>, <STATIONS>.

Edgar Mesias Jaramillo Moyano
0603497397

INTRODUCCIÓN

La meteorología se presenta como una ciencia interdisciplinaria que aborda distintas definiciones y fenómenos que se amplía en las ciencias básicas (Física, Química, Matemáticas), se han desarrollado una diversidad de estudios que ayudan a integrar información y ampliar el conocimiento acerca de los fenómenos atmosféricos que son tan antiguos como la propia atmósfera, la creación de distintas organizaciones que se encargan de monitorear esta información es de suma necesidad dado que los cambios climáticos han sido muy recurrentes en las últimas décadas, lo que ha provocado que los pronósticos meteorológicos sean muy importantes para entender el comportamientos de determinados fenómenos (calor extremo, huracanes, inundaciones, incendios forestales, etc.) (Galán, et al., 2015).

En la ciudad de Mazatlán, México se han realizados estudios que explica como la Velocidad del Viento presentes durante el mes de Mayo en dicha localidad se pueda relacionar con un análisis numérico estadístico que entregue y ayude a determinar la densidad de potencia y potencia eólica de la velocidad, a su vez que trabaje con información faltante, en otro caso se ha estudiado a la misma variable como un generador de fuente eólica en Santa Cruz una urbe ubicado en el interior de Argentina que busca obtener una correlación entre los datos que arrojaban la torre del control y los de la superficie dada esta información se pudo conocer que en dicho lugar existe mucha dificultad para realizar cualquier tipo de estudio de variabilidad provocado por el tipo y calidad de instrumentos usados, cambios de ubicación y altura de los instrumentos, a esto se agrega que los estudios en meteorología son muy cambiantes dado a distintas variables que se pueden producir con el pasar del tiempo (Otero , et al., 2016).

En base al estudio meteorológico de una de las variables importantes la velocidad de viento permite la generación de fuentes eólicas siendo la energía que se está poniendo a prueba para conocer su utilización y consistencia sobre otras fuentes de energía tradicionales (carbón, petróleo, gas natural, energía hidráulica y nuclear), que pondera principalmente su comportamiento que se origina a través del movimiento del aire superior e inferior a la presión atmosférica sin embargo tiene mayores resultados en lugares que están a grandes alturas sobre el nivel del mar. Por problemas de índole natural, falta de señal o mantenimiento de los dispositivos utilizados de los estudios meteorológicos, esta variable presenta datos faltantes (Castillo, et al., 2017).

Por esta razón el análisis de los datos faltantes de la variable velocidad del viento de los años 2014-2021 contribuirán al proyecto medite comparación de las técnicas de relleno, los métodos

aplicados para este estudio son Random Forest Imputation, Imputación por la Media (MICE), Hot Deck Imputation e Iterative PCA Imputation, utilizando el software R Studio el programa permitirá llegar a los resultados esperados.

CAPÍTULO I

1. PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento del problema

Los datos suelen perderse debido a problemas con la estación meteorológica, como: calibración y daños en el equipo o falta de mantenimiento. Los datos se recopilan y almacenan desde las estaciones meteorológicas actuales mediante la descarga de datos, que el equipo GEEA actualiza a través de un sistema automatizado debido a una falla del anemómetro debido a una configuración de grupo no válida. No importa dónde se encuentren los datos, las anomalías ocurren en diferentes intervalos de tiempo y los datos faltantes son problemáticos y reproducibles porque la mayoría de los métodos estadísticos no pueden aplicar el análisis requerido.

1.2. Limitaciones y delimitaciones

Una de las posibles limitaciones de los trabajos de comparación de llenado de datos es que los grandes sistemas de información se tratan con agregación histórica, lo que resulta en una toma de decisiones basada en la información proporcionada y, por lo tanto, los datos analizados de las estaciones meteorológicas de 2014-2021 tendrán que ser vistos sistemáticamente para rellenar los datos con mayor precisión.

1.3. Problema general de investigación

¿Cuál es la mejor técnica de relleno de datos faltantes de la variable Velocidad de Viento de los años de estudio?

1.4. Problemas específicos de investigación

- ¿Cuáles son los datos que presentan mayor problema al momento de rellenar dicha información?
- ¿Cómo ejecutar el relleno de datos dada la comparación de técnicas?
- ¿Cuáles fueron los años que más influyeron en el estudio y de que estaciones fue entrega dicha información?

1.5. Objetivos

1.5.1. *Objetivo general*

Comparar las técnicas de relleno de datos faltantes de la variable Velocidad de Viento de los años 2014 al 2021, en la provincia de Chimborazo.

1.5.2. *Objetivos específicos*

- Realizar un análisis estadístico descriptivo mediante gráficas de comportamiento de la variable Velocidad de Viento.
- Determinar las mejores técnicas estadísticas a través de revisión bibliográfica para completar datos faltantes de Velocidad de Viento.
- Comparar las técnicas de completación de datos por medio de mediciones métricas.

1.6. Justificación

1.6.1. *Justificación teórica*

El objetivo del estudio, fue la búsqueda de teorías y conceptos subyacentes que fundamenten los métodos y técnicas de relleno de datos, estos previamente han sido comprobados con una regresión las distintas metodologías como Hot-Deck es una técnica de relleno de datos el cual identifica, evalúa y analiza la variable Velocidad de Viento en el período 2014 al 2021, en la provincia de Chimborazo.

1.6.2. *Justificación metodológica*

El estudio se desarrolló con el fin de solucionar un problema de datos faltantes en las bases meteorológicos, los valores faltantes en el área de meteorología pueden causar sesgos y problemas de falta de información.

Sin embargo, la investigación se centró en encontrar procedimientos de comparación de relleno de datos que mejoren el ajuste de la variable velocidad de viento. El primer punto es validar la información mediante una depuración de datos, revisando la información faltante y errónea. Cada observación realizada después de tomar una de las lecturas, ayudó a avanzar en las características climáticas requeridas y poder establecer diferentes escalas de tiempo dependiendo de la situación. Por tanto, la información obtenida nos permitió captar la información necesaria para contribuir a

futuras investigaciones en este campo. Cabe señalar claramente que en los diversos campos del conocimiento, la necesidad de recabar información sobre los fenómenos objeto de estudio se está convirtiendo en una forma cada vez más eficaz de comprobar su realidad. Finalmente, se evaluó el efecto del proceso de adaptación de la información sobre la efectividad del método utilizado, así como los cambios en la estructura de los componentes y el tamaño de la variable representada, donde se explicó mejor su característica más importante.

1.6.3. *Justificación práctica*

De acuerdo con los objetivos del estudio, los resultados permitieron determinar el mejor método para el relleno de datos. Con base en los probables efectos, es posible analizar qué procedimiento es el más adecuado para futuras investigaciones que puedan completar la información faltante.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. Antecedentes de investigación

Ecuador cuenta con 4 regiones naturales costa, sierra, oriente y la región insular mostrando una infinidad de especies y una producción agrícola grande, la provincia de Chimborazo es uno de los lugares que presenta dichos recursos esto se caracteriza por tener climas para la proliferación de productos como banano, cacao, caña de azúcar y ganadería, desde lo tropical (cantón Cumandá) hasta el páramo (cantones Colta y Riobamba).

Las condiciones climáticas en Chimborazo, por ejemplo, son influenciadas por dos variables o factores principales la temperatura y la precipitación, que dan lugar a marcados cambios temporales y espaciales en los distintos cantones de la provincia. A diferencia de otros lugares, en Chimborazo se observan dos épocas bien diferenciadas por la distribución temporal de las precipitaciones, una época lluviosa y otra seca, al igual que en el resto del planeta, las observaciones de fuertes vientos muestran un leve cambio que se ha ido registrando de manera paulatina a través del tiempo.

En la última época se presenta diferentes investigaciones realizadas por organizaciones nacionales e internacionales relacionadas, a las técnicas de relleno de datos faltantes en diferentes variables de estaciones meteorológicas.

Ayala aplico los datos faltantes de 6 estaciones meteorológicas de la variable precipitación en Chone que sirvió para el relleno de dicha información Ayala et al. (2018, pp. 298-313), Carrera en su investigación “Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la velocidad media de los vientos” en el cual su objetivo fue comprobar el método estadístico para el relleno de datos en dos zonas y obtener el modelo más adecuado Carrera et al. (2016, pp. 81-90), Ferreira el enfoque de la completación de valores faltantes en series de velocidad de viento realizo el promedio a la imputación de los datos basado en la metodología Hot-Deck Ferreira (2004, pp. 163-170) , Araya concluyó que mediante la metodología Hot-Deck su ventaja principal fue la imputación de valores para rellenar datos faltantes, evaluando su eficiencia en distintas situaciones con el propósito de comparar el comportamiento de los parámetros como MAE (Error Absoluto Medio), RMSE (Error Cuadrático Medio) Araya (2014, pp.70-79) .

2.1.1. *Antecedentes históricos*

Urrutia investigó una metodología para la imputación de datos faltantes en meteorología de series de precipitación y/o temperatura, el procedimiento consiste en hacer uso de correlaciones parciales, modelos de regresión, ajustes de los datos por medio del método de doble masa y verificación de la tendencia a través del test de Kendal, finalmente evaluó el 20 % de datos faltantes dada la metodología resulta apropiada, sin embargo, los datos faltantes que sobrepasan el 20 % el modelo adecuado es AMI bivariado Urrutia et al. (2010, pp. 44-49) . Araya manifestó que para ciertas experiencias en la aplicación operativa de un método multivariado de imputación meteorológica, se han propuesto diferentes técnicas estadísticas para poder lidiar con el problema de los datos ausentes, este estudio se analizó la aplicación de un modelo de imputación usando componentes principales, su fin fue discutir las posibilidades de este procedimiento para poder completar la información en resolución horaria generados por la red de estaciones del Instituto Meteorológico Nacional, realizaron algunos experimentos de prueba, a los que aleatoriamente se les eliminaron datos para su posterior estimación (Araya, 2014, pp.70-79) ; Viada realizó la revisión sistemática de los métodos de imputación de datos faltantes, existieron muchas maneras de llevar a cabo la evaluación y validación de una o más técnicas de imputación que hayan sido aplicadas a una situación particular como el método MAR, donde los valores faltantes dependen de los observados y no de las variables faltantes en sí, la pérdida de información sobre la variable no depende de la variable en sí, mientras que en el modelo MCAR, los valores faltantes son independientes de otros valores observados y datos faltantes, la más común de ellas es el uso de simulación, que consistió en aplicar pérdidas artificiales de información de una base original con una de datos disponibles, posteriormente se suprimió todos los registros que tengan al menos un elemento faltante, obteniendo finalmente un registro más pequeño pero completo Viada et al. (2016, pp. 113-130) .

Chica investigó una herramienta para la simulación de datos faltantes en series climáticas diarias de zonas ecuatoriales y con base en el análisis de las tres series originales (escala diaria), se simularon tres series climáticas diarias con las variables (lluvia, brillo solar, temperatura máxima, temperatura mínima y humedad relativa) que tienen la misma longitud, generando un promedio en esa escala de tiempo tanto para los datos simulados como para los observados y utilizó la recta de pendiente 1 e intercepto 0 para evaluar la calidad de la información simulada Chica et al. (2014, pp. 7365-7373) ; Salgado investigó la “Imputación de datos faltantes de temperatura mediante técnicas geoestadísticas en estaciones climáticas del Valle del Cauca en el periodo de 1985 a 2015” dónde pretendió ratificar el método de imputación mediante validación cruzada, retirando los listados completos de una estación, luego imputó dichos registros apartados y comparó los datos reales con sus respectivos procedimientos, respecto a su ECM, por lo cual usaron simulaciones en las que se

retiraron el 20 %, 40 % o 60 % del total de información de cada estación, y así se pudo decir qué tan bueno resultó ser el método de imputación Salgado (2016).

Quishpe realizó el "Relleno de datos de velocidades de viento mediante la aplicación de método de Hot Deck para la estimación de producción de energía eléctrica en base al recurso eólico", la simulación para cada uno de los escenarios se conforma por diferentes porcentajes de datos utilizables en series de viento comparando el comportamiento de 90 %, 60 % y 30 % de las sucesiones presentes, esto a su vez podrá generar valores para la estimación de producción de energía eléctrica en base al recurso eólico, además se observó como la secuencia de información con valores faltantes han sido imputadas con el método Hot Deck, las cantidades de la media con distintos porcentajes de datos disponibles, se evalúa su eficacia mediante las cantidades desiguales de parámetros como MAE (Error Absoluto Medio), RMSE (Error Cuadrático Medio) Quishpe (2020); Cárdenas y Urgilés analizaron el espacio-temporal meteorológico en una cuenca andina tropical del sur de Ecuador tuvo el objetivo analizar espacial y temporalmente evaluando imputando y realizando un análisis exploratorio de los registros de las estaciones meteorológicas en la cuenca alta del Río Paute, en la serie existen faltantes para lo cual se aplicó métodos de imputación múltiple por ecuaciones encadenadas (MICE) para cada variable, obtuvieron resultados aceptables a escala horaria para temperatura del aire, presión atmosférica, humedad relativa y radiación solar global, y a escala diaria para precipitación, velocidad y dirección del viento Cárdenas y Urgilés (2020, p. 162) ; Mariño realizó una predicción de la temperatura ambiental mediante modelos estadísticos funcionales de las estaciones monitoreadas por el GEAA en la provincia de Chimborazo (2014-2019), donde las imputaciones fueron mediante Random Forest, las ventajas de este método es que es certero para un conjunto de datos extenso, imputa grandes cantidades de información perdida, mantienen la similitud con la cantidad real, realiza calificación mediante la relación entre las variables y permite localizar valores, como resultado eliminó los valores erróneos, que no estaban dentro de los límites establecidos, también se completaron 53 bases datos faltantes de las 66 proporcionadas, con un máximo de imputación del 30 % atípicos Mariño (2021).

2.2. Referencias teóricas

2.2.1. *Variable cuantitativa*

Se definen por la existencia de una unidad de medición, que puede ser contable (unidades enteras), medible o ponderada por algún atributo físico con algún instrumento (Rendón et al., 2016, pp. 397-407).

2.2.2. *Datos faltantes*

Los valores faltantes (también conocidos como valores perdidos) son importante cuando se realiza un análisis de datos porque, a pesar de que los valores no son muy abundantes, la mayoría de los métodos estadísticos asumen una matriz de datos completa. Lo que tiene como objetivo de completar los datos mediante la sustitución de los valores perdidos por valores válidos estimados, preferentemente utilizando un camino sin sesgo e informáticamente eficiente (Andrades et al., 2018, pp. 199-212).

2.2.3. *Estadística descriptiva*

La rama de la estadística que formula recomendaciones sobre cómo resumir la información en cuadros o tablas, gráficas o figuras (Rendón et al., 2016, pp. 397-407) y también (Diggle y Chetwynd, 2011, pp. 36-56).

2.2.4. *Diagrama de caja y bigotes*

Resumen visual de la distribución (comportamiento) de una variable que provee detalles acerca de si uno o ambos extremos de la distribución contienen valores inusualmente grandes o pequeños (Parra Olivares, 2002, pp. 115-124).

2.2.5. *Datos atípicos*

Son aquellos excesivamente grandes o pequeños tal que, tras su comprobación, no pueden considerarse como equivocaciones o errores groseros del proceso (Atkinson et al., 2007, pp. 171-187).

2.2.6. *Diagrama de densidad*

Este visualiza la distribución de datos a través de un período de intervalo o de tiempo continuo (IngenioVirtual, 2015).

2.2.7. *Correlación*

Para el estadístico que cuantifica la correlación. Sus valores están comprendidos entre -1 y 1 (Ortega et al., 2009, p. 6).

2.2.8. *Estadística inferencial*

Se enfoca en la toma de decisiones o realización de generalizaciones acerca de las características de todas las observaciones bajo consideración con base en información parcial o incompleta (Velázquez, 2017).

2.2.9. *Prueba de rosner*

La prueba de uso común para "valores atípicosçuando está dispuesto a asumir que los datos sin valores atípicos siguen una distribución normal (gaussiana), diseñado para evitar el enmascaramiento, que ocurre cuando un valor atípico no se detecta porque tiene un valor cercano a otro (R, s.f).

2.2.10. *Hipótesis estadística*

Es una afirmación o conjetura acerca del valor de un parámetro o parámetros de una población según (Velázquez, 2017).

2.2.11. *Hipótesis nula*

Se plantea de manera que no hay diferencia o cambio en el parámetro de la población, pues el objetivo de la prueba es rechazarla (Velázquez, 2017).

2.2.12. *Serie de tiempo*

Se define como una colección de datos obtenidos por mediciones de algún evento natural o inducido, los cuales son reunidos sobre la misma variable, bajo iguales condiciones a lo largo del tiempo y con intervalos de semejante medida (Barón Orozco, 2018, p. 103).

2.2.13. *Comportamiento temporal*

El comportamiento temporal de las precipitaciones, sus tendencias, su variabilidad pasada y futura, así como las diferencias espaciales de estos cambios, representan algunos de los procesos más estudiados por las ciencias del clima. Las principales razones son su inherente naturaleza caótica, su importancia para la vida y el desarrollo de las sociedades, el efecto que sobre ellas puede tener el calentamiento atmosférico o su indudable relación con el ciclo hidrológico (Ceballos et al., 2013, pp. 235-260).

“La atmósfera de las primeras épocas de la historia de la Tierra estaría formada por vapor de agua

(H2O), dióxido de carbono (CO2) y nitrógeno (N2), junto a muy pequeñas cantidades de hidrógeno (H2) y monóxido de carbono (CO) pero con ausencia de oxígeno", según (Ceballos et al., 2013, pp. 235-260).

2.2.14. *La troposfera*

Abarca hasta un límite superior llamado tropopausa que se encuentra a los 9 Km en los polos y los 18 km en el ecuador. En ella se producen importantes movimientos verticales y horizontales de las masas de aire (vientos) y hay relativa abundancia de agua, por su cercanía a la hidrosfera. Por todo esto es la zona de las nubes y los fenómenos climáticos: lluvias, vientos, cambios de temperatura, etc. Es la capa de más interés para la meteorología. En la troposfera la temperatura va disminuyendo conforme se va subiendo, hasta llegar a -70ºC en su límite superior (Ceballos et al., 2013, pp. 235-260).

2.2.15. *La estratosfera*

Sigue a la tropopausa y llega hasta un límite superior llamado estratopausa que se sitúa a los 50 kilómetros de altitud. En esta capa la temperatura va aumentando hasta llegar a ser de alrededor de 0ºC en la estratopausa. Casi no hay movimiento en dirección vertical del aire, pero los vientos horizontales llegan a alcanzar frecuentemente los 200 km/h. En esta parte de la atmósfera, entre los 30 y los 50 kilómetros, se encuentra el ozono que tan importante papel cumple en la absorción de las dañinas radiaciones de onda corta (Ceballos et al., 2013, pp. 235-260) .

2.2.16. *Velocidad del viento*

"La velocidad del viento mide la componente horizontal del desplazamiento del aire en un punto y en un instante determinados". Se mide mediante un anemómetro, y la unidad de medida es habitualmente metros por segundo (m/s). Las ausencias de viento se denominan calmas, según (Sarochar, 2009).

Tiende a incrementarse a medida que se asciende y que la superficie terrestre ejerce una acción de fricción o de retardo sobre la velocidad el viento (Guevara, 2013, pp. 81-101).

2.2.17. *Dirección del viento*

La dirección mide la componente horizontal de la velocidad del viento. En meteorología es importante tener en cuenta que la dirección nos indica de dónde viene el viento, no hacia dónde va. Por ejemplo, el viento norte es el que sopla desde el norte. Se mide en grados, desde 0º (excluido)

hasta 360º (incluido), girando en el sentido de las agujas del reloj en el plano horizontal visto desde arriba. Valores cercanos a 1º y 360º indican viento del norte, cercanos a 90º viento del este, 180º del sur y 270º del oeste. "Entre estos valores tendremos el resto de direcciones: nordeste, sureste, suroeste y noroeste" según (Burzykowski et al., 2010, pp. 288-297).

2.2.18. *Método de imputación*

La imputación es llenar los espacios en blanco de una base de datos incompleta con valores confiables y así obtener un archivo completo para su análisis esta usado como tratamiento para la falta de respuesta parcial.

"El proceso de imputación está incluido dentro de la etapa de preparación de datos, que es el paso que se realiza luego de llevar a cabo el entendimiento de los mismos. Aquí, los valores faltantes son reemplazados por valores estimados conocidos, esto con el fin de obtener un conjunto de datos completo, al cual se le pueda aplicar diversas técnicas estadísticas. La importancia de encontrar un correcto método de imputación es esencial, ya que trabajar con datos equivocados, implicaría tener alteraciones en el resultado final de las estimaciones. De acuerdo con algunos autores, las técnicas simples de imputación pueden tener algunas ventajas sobre las múltiples, ya que se dice que hay menos riesgo de pérdida de eficiencia en comparación con las técnicas múltiples.

2.2.18.1. *Imputación con la media*

Se estima la media absoluta de los registros presentes en la base de datos completa para la variable a imputar, por lo tanto, el valor resultante (media absoluta) será el valor donante para los registros con datos faltantes de esta variable. De esta misma forma se aplica para cada una de las variables que presenten al menos un registro ausente (Márquez et al, 2017, pp. 9-40). Según se expresa de la siguiente manera: sea yij el valor de Y la unidad i en la variable j, $i = 1, 2, \ldots, N$, $j = 1, \ldots, J$. La media imputada sustituye la media $\bar{y}jR$ de la mj unidad respondiente en la celda j para unidades que son muestreadas, pero que no responden (Little and Rubin 2014). Para diseños igualmente ponderados, la media poblacional $\bar{Y}$ podría ser estimada por la media de las unidades observadas e imputadas, a saber (Márquez et al, 2017, pp. 9-40):

$$\frac{\sum_{j=1}^{J} n_j \widehat{\bar{y}}_j}{\sum_{j=1}^{J} n_j} \tag{1–2}$$

donde $\widehat{\bar{y}}_j$ es la media de las unidades observadas e imputadas en la celda j (Márquez et al, 2017, pp. 9-40). Ahora,

$$\widehat{\bar{y}}_j = \frac{m_j \bar{y}_j R + n_j m_j \bar{y}_j R}{n_j} \tag{2–2}$$

2.2.18.2. *Random forest*

Es un algoritmo de aprendizaje automático basado en ensamble que combina múltiples árboles de decisión para mejorar la precisión y reducir la varianza en las predicciones (Breiman, 2001, pp. 5-32). Antes de construir un modelo de Random Forest, es esencial preparar los datos adecuadamente. Esto implica limpiar y preprocesar los datos, lo que puede incluir el manejo de datos faltantes, la codificación de variables categóricas y la normalización de variables numéricas (Kuhn y Johnson, 2013).

Los árboles de decisión son fáciles de interpretar todos los árboles comienzan en la raíz o la raíz y terminan en diferentes hojas o nudos de hojas. En cada el árbol de decisión se divide en dos nodos según el regresor más importante o la variable categórica. Depende de si la variable probada es categórica o no continúo, el tipo de árbol será diferente de los primeros árboles de clasificación segunda regresión, para analizar conjuntos de datos de supervivencia, pero no tendrá que seguirlos este estudio porque los datos utilizados no entran en esta categoría. Una vez construido el árbol completo obtiene una lista de reglas de decisión para usar categorizar o predecir una nueva observación (Bou-Hamad y col., 2011).

2.2.19. *Error medio de pronóstico (EMP)*

Se utiliza para medir la precisión de un pronóstico. Encontrando el error en cada periodo, dividiendo esto entre el valor real de ese período y promediando después estos porcentajes de error (Hanke y Reitsh, 1996, p. 605). La fórmula es:

$$EMP = (1/n) * \Sigma(Yi - \widehat{Yi}) \tag{3–2}$$

Donde:

Y_i : Es el valor real.

$\widehat{Yi}$: Es el valor pronosticado.

n : Es el número de observaciones.

2.2.20. *Error medio cuadrático (EMC)*

Se utiliza para medir la precisión de un pronóstico. Cada error o residual se eleva al cuadrado; luego, estos valores se suman y se divide entre el número de observaciones (Hanke y Reitsh, 1996, p.

605). La fórmula es:

$$EMC = (1/n) * \Sigma(Yi - \widehat{Yi})^2 \tag{4–2}$$

2.2.21. *Diferencia absoluta media (DAM)*

Consiste en obtener la suma de los errores absolutos y dividir para el número de observaciones (Hanke y Reitsh, 1996, p. 605). La fórmula es:

$$DAM = (1/n) * \Sigma|Yi - \widehat{Yi}| \tag{5–2}$$

2.2.22. *RStudio*

El entorno en el que se han implementado técnicas estadísticas, tanto clásicas como modernas, está enmarcado dentro de la plataforma GNU y se distribuye con licencia GNU GPL, dispone versiones de R para Windows de Microsoft, Unix, Linux y MacOS (Mirabal et al., 2010, pp.302-308).

2.2.23. *Paquete (MICE)*

Hace imputación múltiple utilizando Fully Conditionally Specification (FCS) implementado por el algoritmo MICE (Multiple Imputation by Chained Equations). Cada variable tiene su propio modelo de imputación. Se proporcionan modelos de imputación incorporados para datos continuos (pmm), datos binarios (regresión logística), datos categóricos no ordenados (regresión logística politómica) y datos categóricos ordenados (odds proporcional). Se puede utilizar imputación pasiva para mantener consistencia entre las variables. Se dispone de varios gráficos de diagnóstico para examinar la calidad de las imputaciones (Buuren y Groothuis, 2011, pp. 1-67).

2.2.24. *Paquete (VIM)*

Presenta nuevas herramientas para la visualización de valores perdidos, que pueden utilizarse para explorar los datos y la estructura de los valores imputados, se puede explorar visualmente utilizando varios métodos de gráficos univariados, bivariados, múltiples y multivariados (Kowarik y Templ, 2016).

2.2.25. *Función t.test*

Se encargada de los procedimientos de inferencia sobre la media en poblaciones normales (Santana y Hernández, 2016).

CAPÍTULO III

3. MARCO METODOLÓGICO

3.1. Enfoque de investigación

La presente investigación aplicó un enfoque cuantitativo, para el análisis de la velocidad de viento máxima en el período 2014 al 2021 fue tomado por horas de las estaciones meteorológicas de la provincia de Chimborazo, esta investigación se centró en la solución de imputación o relleno de datos faltantes presentes en cada matriz de información de dicha variable mediante el uso de cuatro métodos de imputación: Random Forest, Imputación por la Media (MICE), Hot Deck e Iterative PCA imputation.

3.2. Localización de estudio

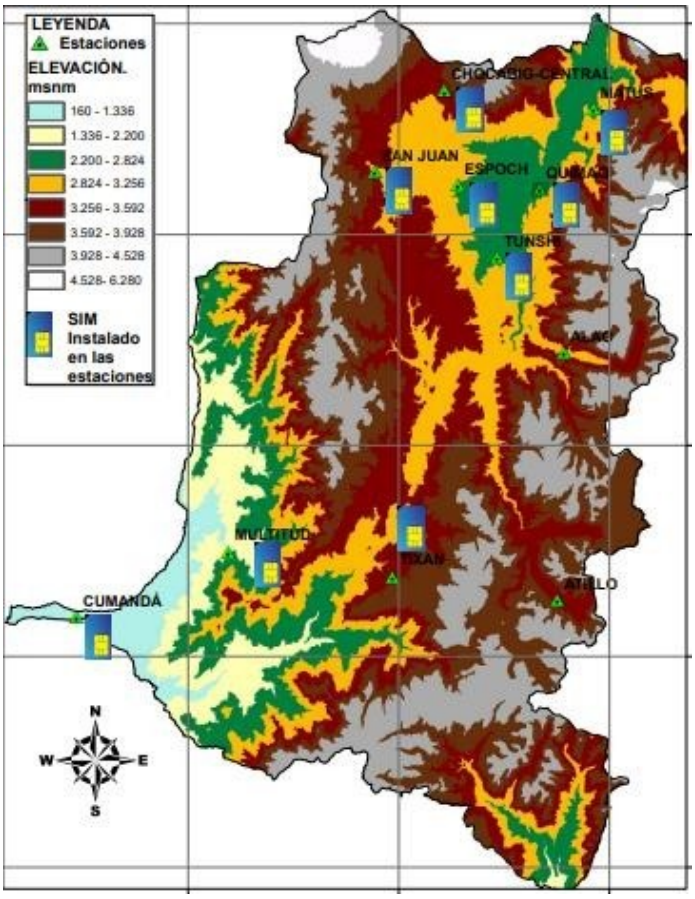


Ilustración 1–3: Estaciones Meteorológicas establecidas

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2023.

Tabla 1–3: Tabla de las coordenadas por estación

Estación	Latitud	Longitud
Alao	9793173	773499
Atillo	9758048	772610
Cumandá	9755579.92	706262.4
Espoch	9816965	758398
Matus	9827878	777759
Multitud	9711374,13	722699,63
Quimiag	9816392.86	770083.61
San Juan	9818849	746596
Tixan	9761332	749103
Tunshi	9806678	764087
Urbina	9835326	754533

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

3.3. Nivel de investigación

El nivel de investigación fue de manera exploratoria e imputación de datos faltantes, donde se buscó determinar técnicas adecuadas para imputación de datos faltantes, haciendo uso de metodos como: Random Forest e Imputación por la Media (MICE), Hot Deck e Iterative PCA imputation y seleccionar el mejor método para la imputación de la variable velocidad de viento.

3.4. Diseño de investigación

3.4.1. *Según la manipulación de las variables*

La investigación es no experimental, sin embargo, es de tipo inferencial e inductiva debido a las técnicas influyentes en el relleno de datos a través de la información recolectada en el GEAA; el periodo de tiempo es transversal, se analizó la Velocidad del Viento de los años 2014 al 2021.

3.5. Población

Para el estudio de la variable velocidad de viento la población de datos fue de 703248 registros tomados de las matrices recoletadas por el grupo GEAA.

3.6. Tamaño de muestra

En cada estación meteorológica se hizo la comparación de datos faltantes por año teniendo un total de 10964 en las 11 estaciones en el período 2014 al 2021.

3.7. Métodos de investigación aplicados

3.7.1. *Método inductivo*

El método aplicado es inductivo puesto que se contó con la base de datos y se procedió ha aplicar a la investigación el relleno de información.

3.7.2. *Método analítico*

Se recopiló la información de la variable velocidad de viento en la cual se observó datos faltantes por tanto se procedió a aplicar la imputación de las matrices.
Se realizó un análisis descriptivo, analítico y gráfico del comportamiento de la velocidad de viento, además de hizo el tratamiento de datos atípicos mediante la aplicación de la prueba Rosner's Outlier donde se detectaron los mismos, los cuales fueron separados de las matrices.
Mediante el test Kolmogórov-Smirnov con la corrección de Lilliefors se realizó el contraste de normalidad.
Obteniendo las matrices idóneas y realizado el tratamiento adecuado se procedió a la aplicación de 4 métodos de imputación se hizo uso de las distintas librerías del software RStudio de cada uno de los métodos: $library(randomForest)$, $library(mice)$, $library(VIM)$.
En la validación de los métodos de imputación aplicados consideramos las metricas del error cuadrático medio (EMC), error medio de pronóstico (EMP) y el desvió absoluto medio (DAM).

3.7.3. *Instrumentos de investigación*

Para la investigación fue tomada la información de los registros meteorológicos que posee el GEAA desde el 2014 al 2021 en las 11 estaciones situadas en la provincia de Chimborazo, el software utilizado para la imputación fue RStudio con sus librerías para cada método: Random Forest Imputation, Imputación por la Media (MICE), Hot Deck Imputation y Iterative PCA Imputation.

3.7.4. *Revisión bibliográfica*

En base a la revisión bibliográfica que se realizó en las distintas plataformas digitales se verificó que las mejores técnicas para el relleno de datos faltantes fueron los métodos de Random Forest, Hot Deck, Imputación por la media e Iterative PCA imputation.

Una de las plataformas que ayudo en la búsqueda de los mejores métodos fue DSpace ESPOCH, los resultados de diferentes tesis admiten que para rellenar datos faltantes en el área meteorológica de forma correcta es Random Forest, Imputación por la media e Iterative PCA imputation.

Otro espacio que concurrió de manera similar en que uno de los métodos que mejor se ajustan a los datos originales para datos meteorológicos es el de Hot Deck e Iterative PCA imputation ya que los resultados obtenidos en diferentes publicaciones de artículos en la plataforma SciELO así lo ratifican. Similarmente en la plataforma RepositorioTEC se halló una tesis que argumento que con los resultados encontrados para un correcto relleno de datos meteorológicos es el método Iterative PCA imputation, de manera semejante en la plataforma DSpaceUPS se averiguó en dos tesis distintas con los resultados obtenidos que los mejores métodos para la imputación de datos meteorológicos fueron el de Hot Deck y de Random Forest.

Con los resultados obtenidos en los distintos artículos de las diferentes plataformas manejadas permitieron elegir los métodos para realizar el relleno de datos en las estaciones meteorológicas.

Random Forest

- ATKINSON, A.D.J., ARIZA, F.J. y GARCÍA BALBOA, J.L., 2007. Estimadores robustos: una solución en la utilización de valores atípicos para el control de la calidad posicional.
- CHECA GAMARRA, M.C., 2020. Análisis geoestadístico de datos funcionales de temperatura del aire en la provincia de Chimborazo.
- CEBALLOS BARBANCHO, A., MORÁN TEJEDA, E. y LÓPEZ MORENO, J.I., 2013. Análisis de la variabilidad espacio-temporal de las precipitaciones en el sector español de la cuenca del Duero (1961-2005).

Hot Deck

- ARAYA LÓPEZ, J.L., 2014. 2014. Experiencias en la aplicación operativa de un método multivariado de imputación de datos meteorológicos.
- AYALA, M.F., CARRERA VILLACRÉS, D. y TIERRA, A., 2018. Relación espacio-temporal entre estaciones utilizadas para el relleno de datos de precipitación en Chone, Ecuador.
- CHICA RAMÍREZ, H.A., PEÑA QUIÑONES, A.J., GIRALDO JIMÉNEZ, J.F., OBANDO

BONILLA, D. y RIAÑO HERRERA, N.M., 2014. SueMulador: Herramienta para la Simulación de Datos Faltantes en Series Climáticas Diarias de Zonas Ecuatoriales.

Imputación por la media

- BARÓN OROZCO, A.F., 2018. Análisis espacio temporal de la precipitación mensual, en la depresión momposina para los años 2012 a 2015.
- CÁRDENAS CAMPOVERDE, H.P. y URGILÉS ÁVILA, C.C., 2020. Análisis espacio-temporal meteorológico en una cuenca andina tropical del sur de Ecuador.
- CARRERA VILLACRÉS, D.V., GUEVARA GARCÍA, P.V., TAMAYO BACACELA, L.C., BALAREZO AGUILAR, A.L., NARVÁEZ RIVERA, C.A. y MOROCHO LÓPEZ, D.R., 2016. Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media.
- SALGADO, A., 2016. "Imputación de datos faltantes de temperatura mediante técnicas geoestadísticas en estaciones climáticas del Valle del Cauca en el periodo de 1985 a 2015".

Iterative PCA imputation

- ARIAS MUÑOZ, A.C., 2022. 2014. ARIAS MUÑOZ, A.C., 2022. Propuesta y evaluación de una estrategia para la imputación múltiple y multivariada de valores faltantes en series de tiempo del campo meteorológico utilizando aprendizaje automático.
- ANDRADES GRASSI, J.E., TORRES MANTILLA, H.A., LÓPEZ HERNÁNDEZ, J.Y., GOITÍA ACOSTA, A. y MEJÍAS DELGADO, J.E., 2018. Exploración espacio temporal de la distribución de datos faltantes de precipitación mensual en el centro occidente de Venezuela, con fines de selección de estaciones.
- HARO RIVERA, S., ZÚÑIGA LEMA, L., MENESES FREIRE, A. y ESCUDERO VILLA, A., 2020. Determinación del comportamiento meteorológico del viento en la provincia de Chimborazo, Ecuador.

CAPÍTULO IV

4. MARCO DE ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

4.1. Análisis descriptivo

Tabla 1–4: Resumen estadístico de las estaciones

Descripción\Estación	Alao	Atillo	Cumandá	Espoch	Matus	Multitud
Media	5.573	4.667	2.937	4.789	4.925	1.486
Moda	0.000	0.000	0.000	2.500	2.102	0.000
Mediana	5.102	4.102	2.867	3.602	4.102	1.297
Desviación	3.087	4.397	1.215	3.066	2.893	2.176
Coef. Variación	0.126	0.054	0.059	0.033	0.030	0.023
Asimetría	0.561	0.444	0.511	1.110	0.625	4.629
Curtosis	-0.473	-0.962	1.246	0.323	-0.639	40.834
Mínimo	0.000	0.000	0.000	0.000	0.000	0.000
Máximo	14.344	22.297	8.047	18.203	16.500	37.398
Descripción\Estación	**Quimiag**	**San Juan**	**Tixan**	**Tunshi**	**Urbina**	
Media	0.090	4.431	5.519	2.522	7.229	
Moda	0.086	0.000	0.000	0.000	0.000	
Mediana	0.086	3.898	4.602	2.000	7.297	
Desviación	0.025	3.187	4.524	2.573	2.997	
Coef. Variación	0.0006	0.036	0.049	0.035	0.034	
Asimetría	-0.603	0.729	0.752	1.234	-0.182	
Curtosis	2.016	0.451	-0.121	1.927	-0.213	
Mínimo	0.000	0.000	0.000	0.000	0.000	
Máximo	0.125	19.500	20.898	24.500	18.898	

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

En la tabla 1–4 la velocidad máxima fue de 0.125 m/s y mínima de 0.000 m/s con una desviación de 0.025 m/s y una mediana de 0.086 m/s de la estación Quimiag. La estación de Multitud presentó un promedio de 1.486 m/s de la velocidad de viento con un máximo de 37.398 m/s, un mínimo de 0.000 m/s y una desviación de 2.176 m/s con una mediana de 1.297 m/s.

4.1.1. *Detección de datos atípicos en cada una de las estaciones*

La detección de datos atípicos se realizó de forma gráfica y mediante el test de Rosner's Outlier.

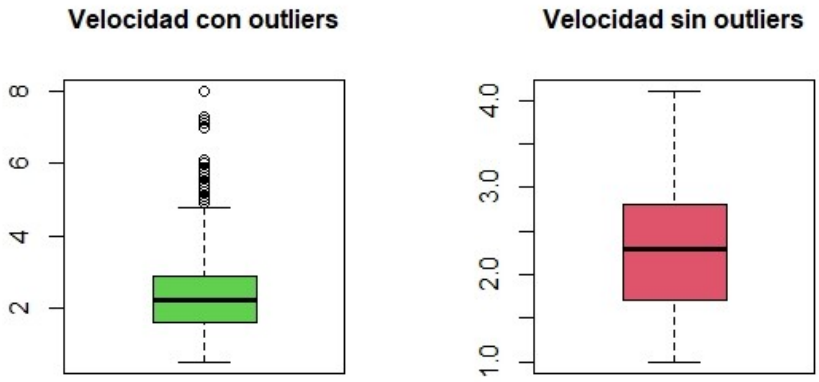


Ilustración 1–4: Estación meteorológica gráfica estación Cumandá

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El número de los datos atípicos que se procesaron fueron de 10 de los años 2014 al 2021, presentó una dispersión menor en la velocidad de viento en la estación Cumandá a su vez entregó asimetría, la posición de su mediana es de 2.867 m/s como mostró la ilustración 1–4.

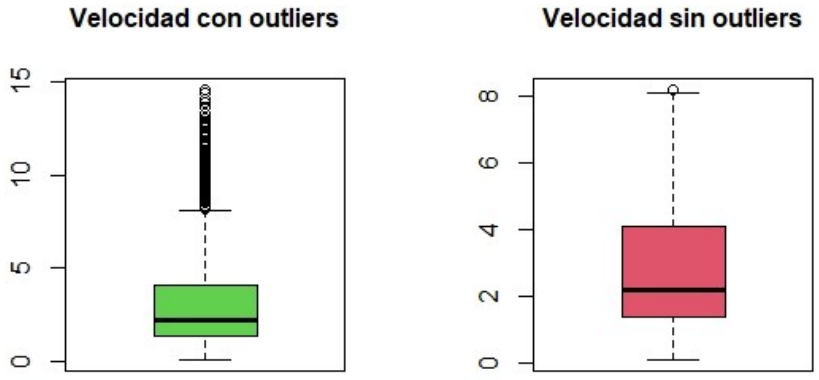


Ilustración 2–4: Estación meteorológica gráfica estación ESPOCH

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación ESPOCH mostró una menor dispersión de la variable velocidad del viento, el número de datos atípicos procesados fueron de 117 dentro de los ocho años, a su vez se observó asimetría negativa por la posición de su mediana 3.602 m/s según la gráfica 2–4.

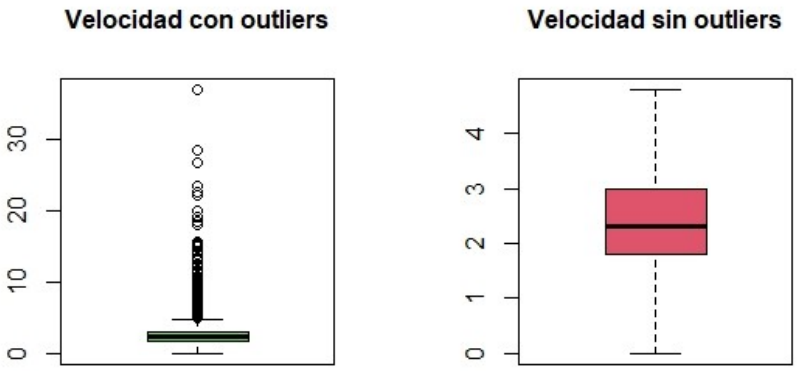


Ilustración 3–4: Estación meteorológica gráfica estación Multitud

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación Multitud se observó una dispersión muy baja de la variable velocidad del viento, el número de datos atípicos procesados fueron de 163 de los años 2014 al 2021, por otra parte se halló simetría positiva por la posición de su mediana 1.297 m/s como mostró la ilustración 3–4.

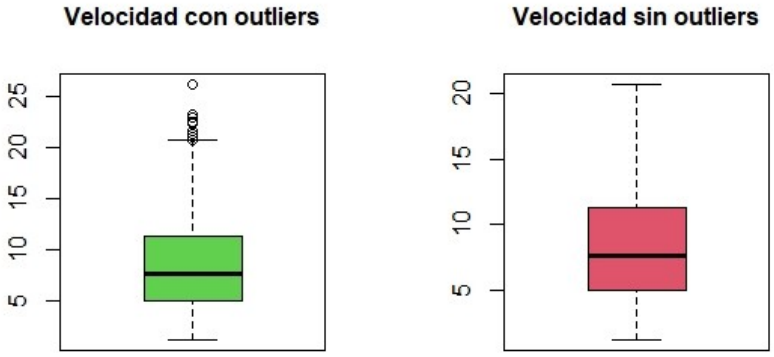


Ilustración 4–4: Estación meteorológica gráfica estación Tixan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 4–4 mostró una menor dispersión de la variable velocidad del viento en la estación Tixan, el número de datos atípicos fueron 64, procesados dentro de los ocho años, por lo tanto presentó asimetría negativa por la posición de su mediana 4.602 m/s.

4.1.2. *Resumen de atípicos y prueba de normalidad*

Tabla 2–4: Resumen estadístico de las estaciones

Estación	Rosner's test (atípicos)	Lilliefors (p-value)
Alao	2	0.0647
Atillo	3	0.0456
Cumandá	10	0.0512
Espoch	117	0.0268
Matus	3	0.0624
Multitud	163	0.0381
Quimiag	30	0.0496
San Juan	179	0.0592
Tixan	64	0.0398
Tunshi	98	0.0484
Urbina	8	0.0516

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

$$p - value > \alpha : \text{No se rechaza H0}$$
$$p - value < \alpha : \text{Se rechaza H0}$$

En la estación San Juan la cantidad de atípicos fue 179 datos, seguido de la estación de Multitud con 163 mientras que la estación de Alao y Atillo se observo 2 y 3 datos atípicos respectivamente (tabla 2–4).

Mediante la prueba de Kolmogorov-Smirnov con una corrección de Lilliefors se determinó la normalidad para las estaciones Atillo (0.0456), Tixan (0.0398), Tunshi (0.0484), ESPOCH (0.0268), Multitud (0.0381) y Quimiag (0.0496) (tabla 2–4).

4.1.3. *Análisis descriptivo de cada una de las estaciones sin presencia de outliers*

Tabla 3–4: Resumen estadístico sin atípicos de las estaciones

Descripción\Estación	Alao	Atillo	Cumandá	Espoch	Matus	Multitud
Media	5.168	6.486	2.303	3.028	4.609	2.494
Moda	5.102	6.636	2.297	8.203	1.898	4.797
Mediana	5.168	6.486	2.303	3.028	4.609	2.494
Desviación	2.772	2.772	0.737	2.178	2.831	0.974
Coef. Variación	0.445	0.427	0.320	0.719	0.614	0.390
Asimetría	0.623	0.071	0.341	1.076	0.748	0.797
Curtosis	-0.246	-0.969	-0.486	0.086	-0.582	-0.015
Mínimo	1.602	1.797	1.000	0.102	0.000	0.000
Máximo	11.000	12.297	4.102	8.203	13.898	4.797
Descripción\Estación	**Quimiag**	**San Juan**	**Tixan**	**Tunshi**	**Urbina**	
Media	5.228	4.086	8.312	6.811	6.340	
Moda	2.203	9.297	5.5	4.602	6.398	
Mediana	5.228	4.086	8.312	6.811	6.340	
Desviación	3.199	2.391	4.038	3.607	2.731	
Coef. Variación	0.612	0.585	0.486	0.530	0.431	
Asimetría	0.705	0.748	0.511	0.595	0.103	
Curtosis	-0.578	-0.268	-0.577	-0.400	-0.464	
Mínimo	0.703	0.000	1.203	0.000	0.000	
Máximo	14.898	10.297	20.703	17.000	14.102	

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

En la tabla 3–4 la velocidad máxima fue de 4.102 m/s y mínima de 1.000 m/s con una desviación de 0.737 m/s y una mediana de 2.303 m/s de la estación Cumandá. La estación de Tixán presentó un promedio de 8.31 m/s de la velocidad de viento con un máximo de 20.703 m/s, un mínimo de 1.203 m/s y una desviación de 4.038 m/s con una mediana de 8.312 m/s.

4.1.4. *Estadística descriptiva mediante gráficas de comportamiento de la variable Velocidad de viento (máxima)*

Estación meteorológica Alao

Comportamiento de la velocidad de viento en el año 2017

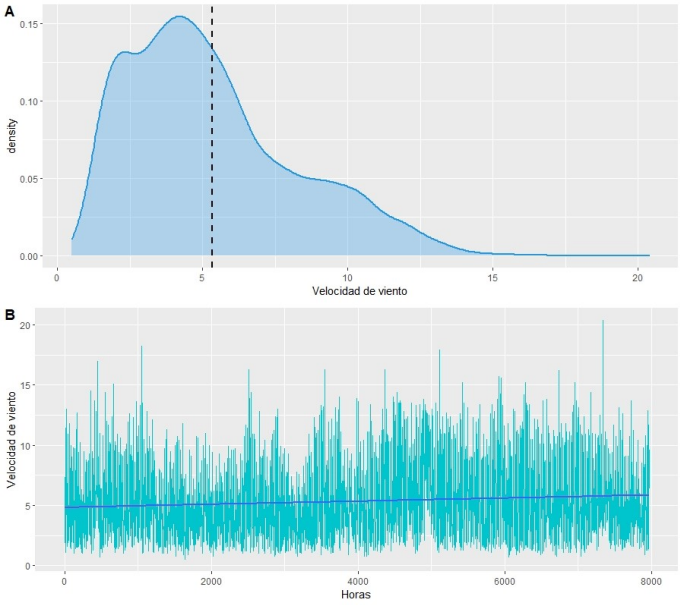


Ilustración 5–4: Estación meteorológica Alao del 2017

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación Alao el año con mayor representatividad de datos de la velocidad de viento fue en el 2017. La ilustración 5–4 "A", mostró asimetría positiva, donde la velocidad promedio de viento fue de 5.1 m/s. En la gráfica 5–4 "B" se observó el comportamiento de la velocidad de viento, por hora en el transcurso del año, evidenciando picos máximos de 21 m/s y mínimos de 0.1 m/s.

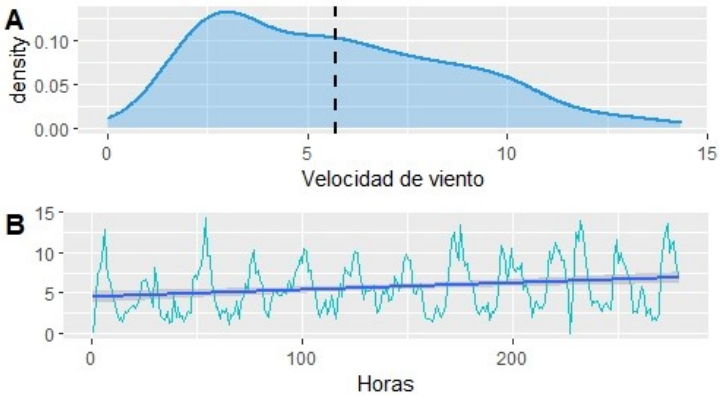


Ilustración 6–4: Estación meteorológica de Alao del 2014

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación Alao el año menos relevante de la variable velocidad de viento fue en el 2014. La ilustración 6–4 "A", presentó asimetría positiva, donde la velocidad promedio de viento fue de 5.1 m/s. En la gráfica 6–4 "B", se verificó el comportamiento de los datos por hora en el transcurso del año, constatando picos máximos de 14 m/s y mínimos de 0.15 m/s.

Estación meteorológica Quimiag

Comportamiento de la velocidad de viento en el año 2017

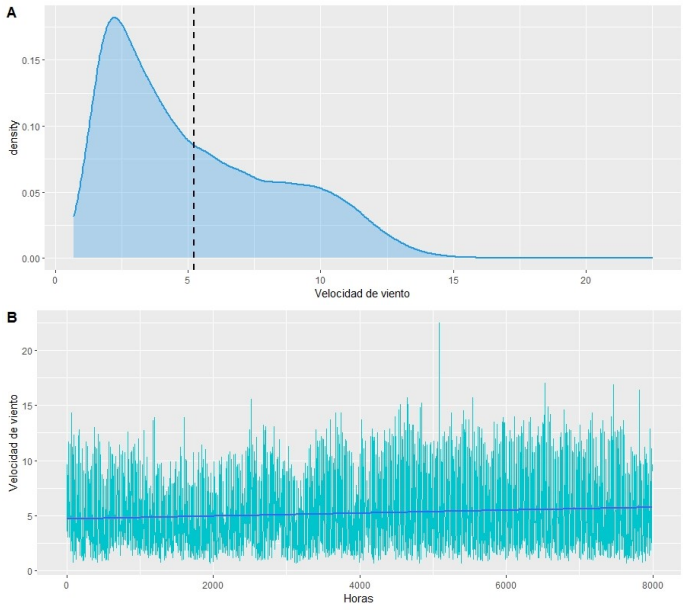


Ilustración 7–4: Estación meteorológica Quimiag 2017

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación Quimiag el año con mayor importancia de la variable velocidad de viento fue en el 2017. La ilustración 7–4 "A", tuvó asimetría positiva, donde la velocidad promedio de viento fue de 5.23 m/s. En la gráfica 7–4 "B" se representó el comportamiento de la variable por hora en el transcurso del año, evidenciando picos máximos de 22.5 m/s y mínimos de 0.70 m/s.

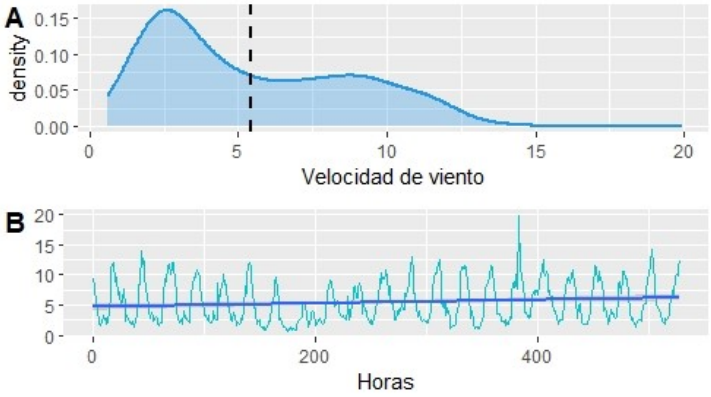


Ilustración 8–4: Estación meteorológica de Quimiag 2014

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación Quimiag el año con menor repercusión de los datos de la velocidad de viento fue en el 2014. La ilustración 8–4 "A", se observó asimetría positiva, donde la velocidad promedio de

viento fue de 5.23 m/s. En la gráfica 8–4 "B" mostró el comportamiento de la variable por hora en el transcurso del año, evidenciando picos máximos de 20 m/s y mínimos de 0.11 m/s.

Comparación de las técnicas para imputación de datos

Cálculo del porcentaje de datos faltantes para cada estación

Para el cálculo del porcentaje de datos faltantes se creó una función que permitió identificar datos nulos o ceros y reemplazar con "NA" debido al reconocimiento "NA" en los métodos de imputación seleccionados.

Porcentaje de datos faltantes en cada una de las estaciones

Tabla 4–4: Porcentaje de datos faltantes

Estación Metereológica	Variable	Año	Datos faltantes	Porcentaje
Alao	SpdMax	2014	7	1,17 %
Atillo	SpdMax	2018	2263	34,94 %
Cumandá	SpdMax	2014	12	2,86 %
Espoch	SpdMax	2017	33	0,39 %
Matus	SpdMax	2015	30	0,34 %
Multitud	SpdMax	2016	4169	47,46 %
Quimiag	SpdMax	2018	30	2,25 %
San Juan	SpdMax	2014	1132	14,73 %
Tixan	SpdMax	2021	1369	16,70 %
Tunshi	SpdMax	2019	1698	31,71 %
Urbina	SpdMax	2014	221	2,88 %
		TOTAL	**10964**	

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E. y Velastegui M, 2022.

La estación Atillo presentó en el año 2018 un total de 2263 datos faltantes que representó al 34.94 %, en el año 2016 la estación Multitud tuvó un porcentaje del 47.76 % con 4169 datos incompletos, la estación Tunshi en el año 2019 dio un total de 1698 faltantes esto índico el 31.71 %, en la estación Tixan con un porcentaje del 16.70 % existió 1369 datos ausentes en el año 2021 como se mostró en la tabla 4–4.

4.1.5. *Gráficas de imputación por método Random Forest*

Estación Alao

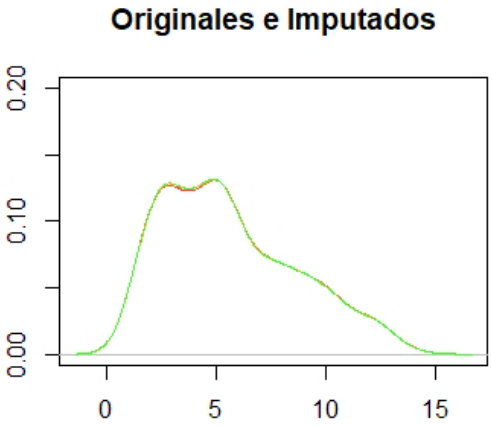


Ilustración 9–4: Random Forest Alao

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 9–4 mostró la gráfica de densidad de los datos reales (color rojo), mientras que la línea verde a la información imputada, el método Random Forest para la estación Alao presentó un buen ajuste.

Estación Atillo

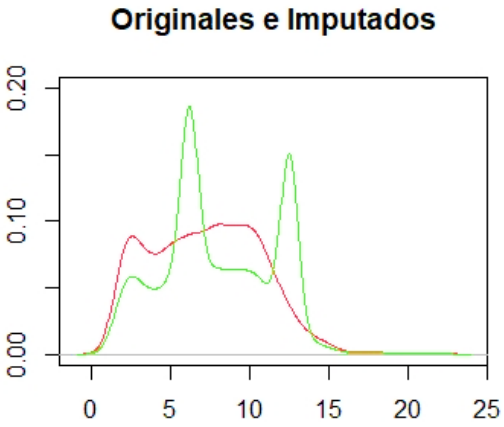


Ilustración 10–4: Random Forest Atillo

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método Random Forest no obtuvo un buen ajuste entre los datos reales y los imputados debido al porcentaje alto de datos faltantes en la estación Atillo como se observó en la ilustración 10–4.

Estación Cumandá

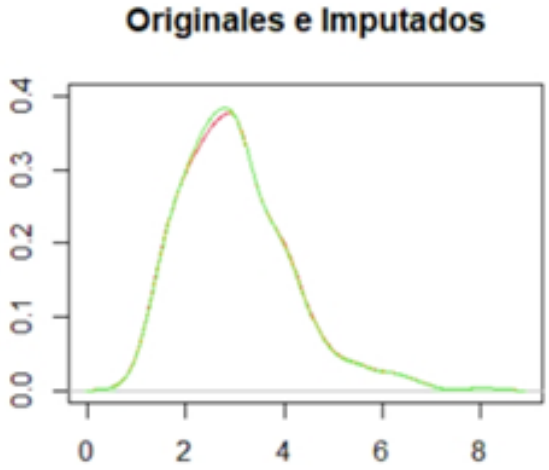


Ilustración 11–4: Random Forest Cumandá

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Cumandá el método Random Forest presentó un buen ajuste en la velocidad de viento como se puede ver en la ilustración 11–4.

Estación ESPOCH

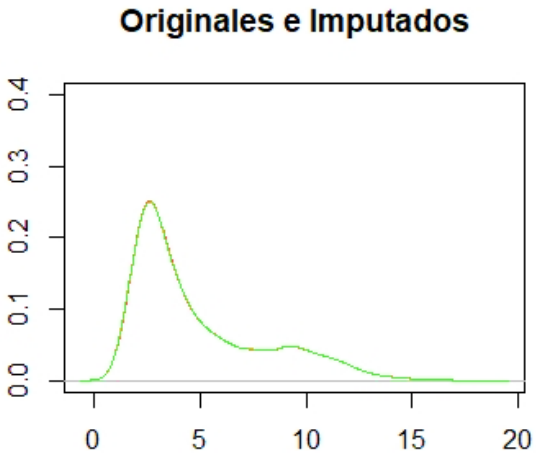


Ilustración 12–4: Random Forest ESPOCH

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 12–4 se mostró que el método Random Forest para la estación ESPOCH presentó un buen ajuste.

Estación Matus

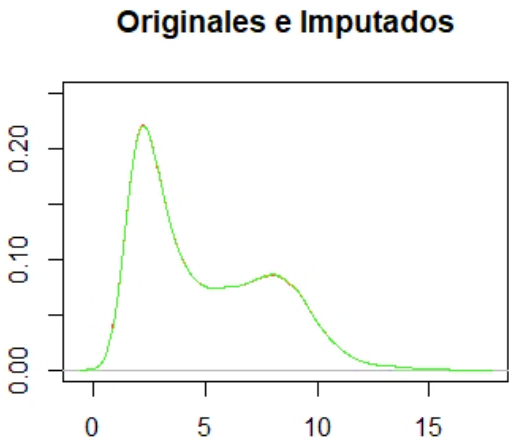


Ilustración 13–4: Random Forest Matus

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método Random Forest obtuvo un buen ajuste entre los datos reales y los imputados debido a que no índico un porcentaje alto de datos faltantes en la estación Matus.

Estación Multitud

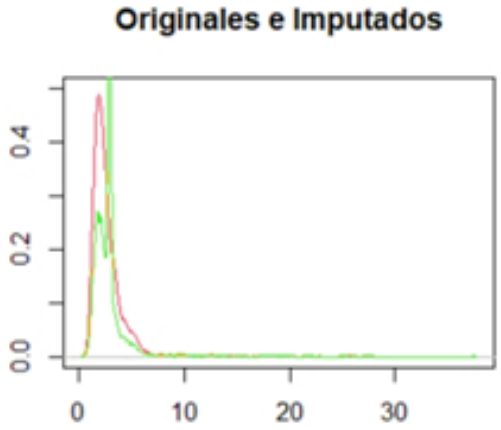


Ilustración 14–4: Random forest Multitud

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 14–4 se observó un alto porcentaje de valores faltantes que representó el 47,46 %. Debido a esto, el método Random Forest no logró imputar correctamente dicho porcentaje en la estación Multitud.

Estación Quimiag

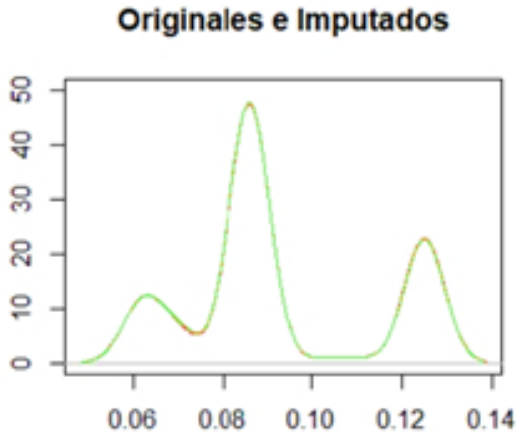


Ilustración 15–4: Random Forest Quimiag

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método Random Forest obtuvo un buen ajuste entre los datos reales y los imputados debido a que no presentó un porcentaje alto de datos faltantes en la estación Quimiag como se observó en la ilustración 15–4.

Estación San Juan

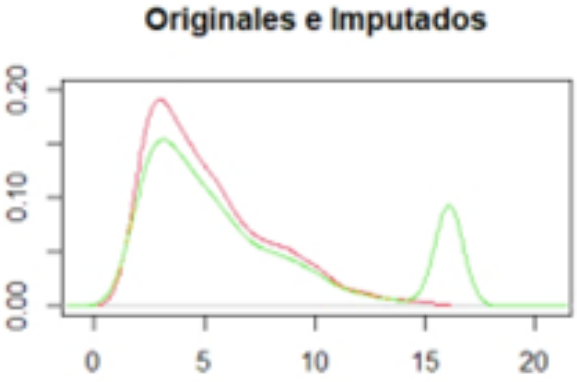


Ilustración 16–4: Random Forest de San Juan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 16–4 mostró un porcentaje alto de valores faltantes, debido a esto el método Random Forest no logró imputar correctamente la información en la estación San Juan.

Estación Tixan

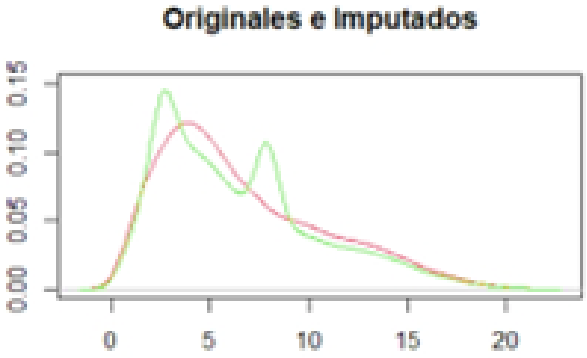


Ilustración 17–4: Random Forest estación Tixan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método Random Forest evidenció un buen ajuste entre los datos reales y los imputados debido al porcentaje bajo de datos faltantes en la estación Tixan como se observó en la ilustración 17–4.

Estación Tunshi

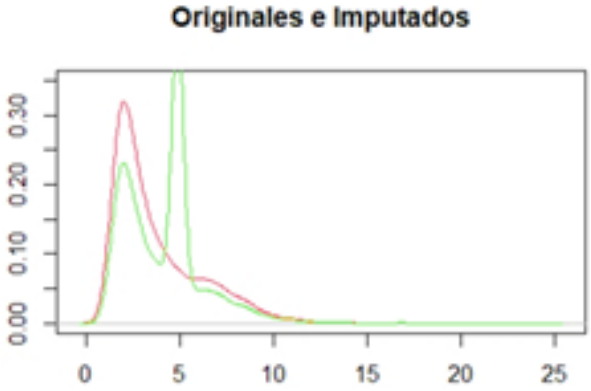


Ilustración 18–4: Random Forest gráfica Tunshi

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método Random Forest no obtuvo un buen ajuste entre los datos reales y los imputados debido al porcentaje alto de datos faltantes en la estación Tunshi como se observó en la ilustración 18–4.

Estación Urbina

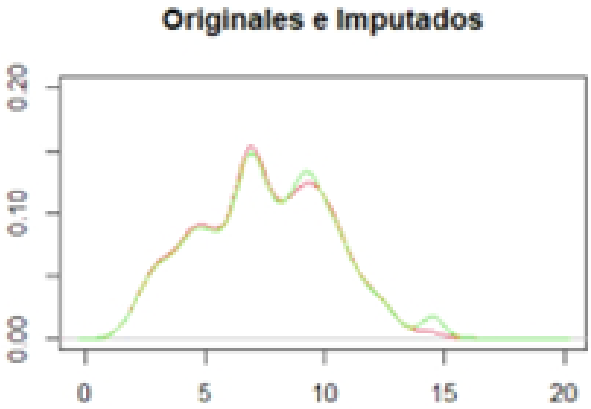


Ilustración 19–4: Random Forest gráfica Urbina

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 19–4 se mostró un porcentaje bajo de valores faltantes que representó 2,88 %. Debido a esto, el método Random Forest logró imputar correctamente dicho porcentaje en la estación Urbina.

4.1.6. *Gráficas de imputación por la Media (MICE)*

Haciendo uso de la librería *library(mice)* se realizó la imputación de los datos como se presenta a continuación:

Estación Alao

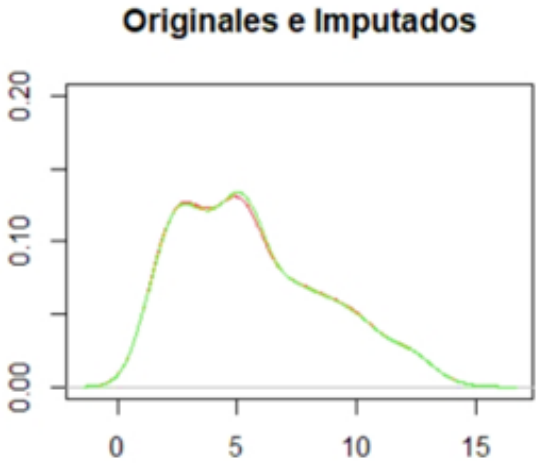


Ilustración 20–4: Relleno de la media Alao

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método de la media (MICE) obtuvo un buen ajuste entre los datos reales y los imputados debido al porcentaje bajo de datos faltantes (1,17 %) en la estación Alao como se observó en la ilustración

20–4.

Estación Atillo

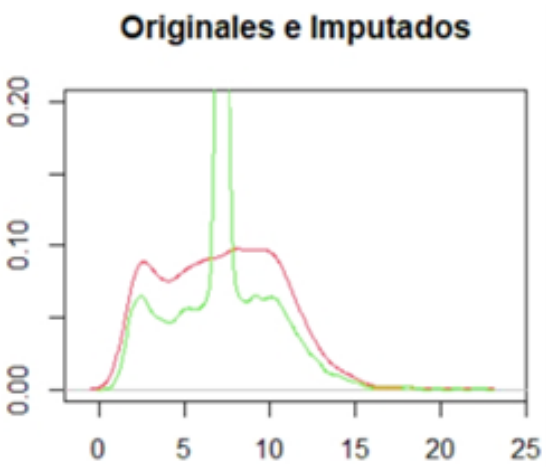


Ilustración 21–4: Relleno de la media Atillo

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 21–4 se evidenció un porcentaje alto de valores faltantes que representó 34,94 %. Debido a esto, el método de la media (MICE) logró imputar correctamente dicho porcentaje en la estación Atillo.

Estación Cumandá

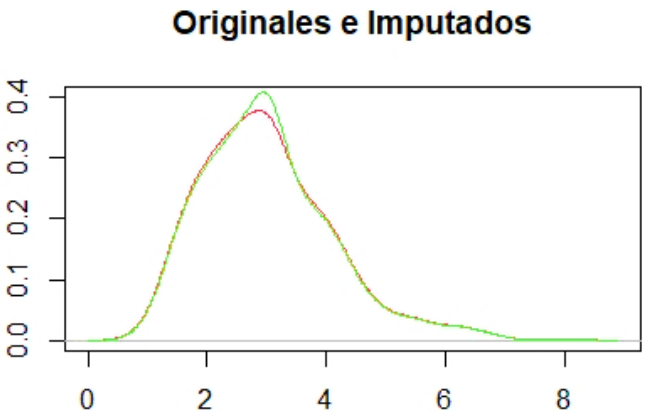


Ilustración 22–4: Imputación por la media Cumandá

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Cumandá el método de la media (MICE) presentó un buen ajuste en la velocidad

de viento como se puede ver en la ilustración 22–4.

Estación ESPOCH

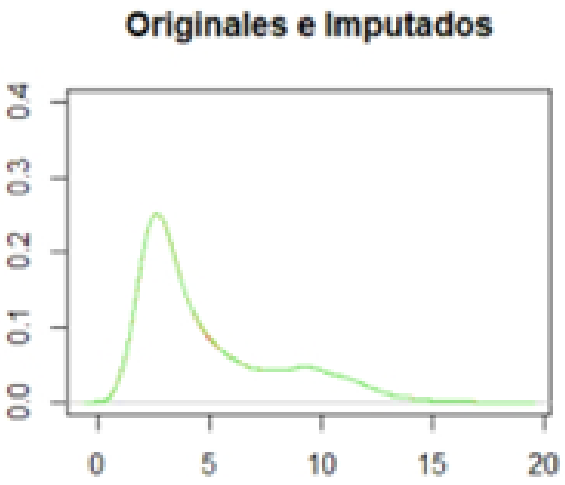


Ilustración 23–4: Relleno de media ESPOCH

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 23–4 se mostró un porcentaje bajo de valores faltantes que representó 0,39 %. Debido a esto, el método de la media (MICE) logró imputar correctamente dicho porcentaje en la estación ESPOCH.

Estación Matus

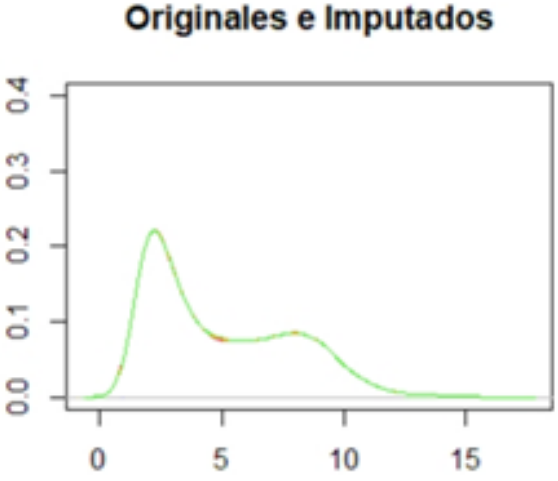


Ilustración 24–4: Relleno por la media Matus

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Matus el método de la media (MICE) presentó un buen ajuste en la velocidad de viento por un bajo porcentaje de datos faltantes (0,34 %) como se puede ver en la ilustración 24–4.

Estación Multitud

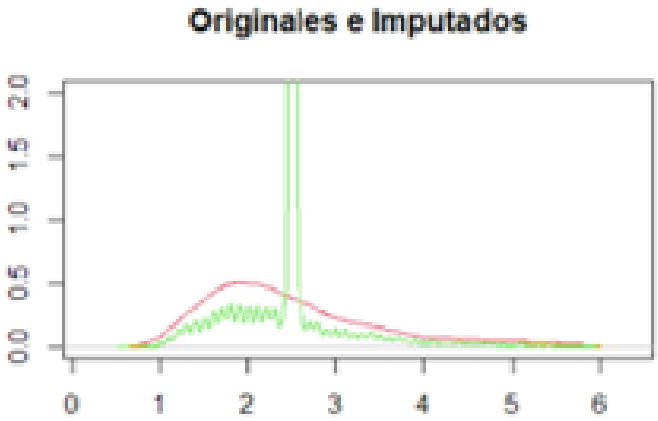


Ilustración 25–4: Imputación por la media de Multitud

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La gráfica de densidad 25–4 para la estación Multitud, se observó que el método de la media (MICE) no logró imputar correctamente los datos.

Estación Quimiag

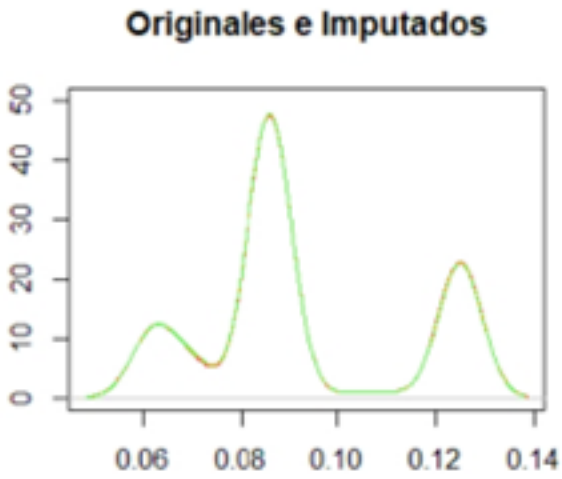


Ilustración 26–4: Relleno de la media Quimiag

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Quimiag el método de la media (MICE) mostró un buen ajuste en la velocidad de viento por un porcentaje bajo de datos faltantes 2,25 % según la figura 26–4.

Estación San Juan

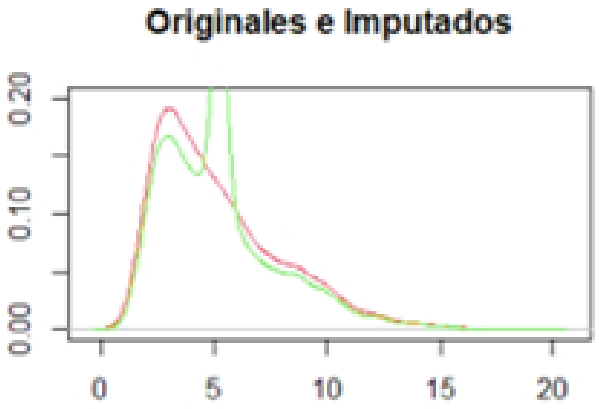


Ilustración 27–4: Imputación de media San Juan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 27–4 se evidenció un porcentaje alto de valores faltantes que representó el 14,73 %. Por tanto, el método de la media (MICE) no logró imputar correctamente dicho porcentaje en la estación San Juan.

Estación Tixan

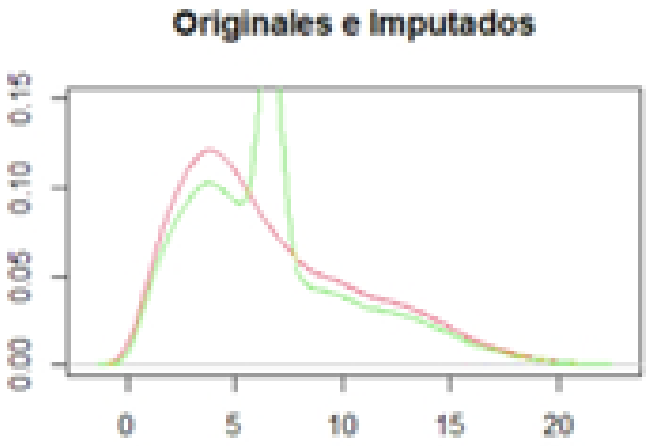


Ilustración 28–4: Imputación de media estación Tixan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método de la media (MICE) presentó que no se ajustan los datos reales y los imputados debido al porcentaje alto de información faltante, en la estación Tixan como mostró la figura 28–4.

Estación Tunshi

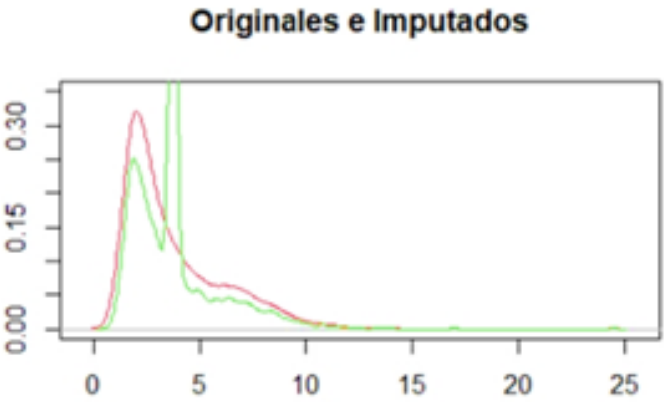


Ilustración 29–4: Relleno por la media estación Tunshi

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Como se observó en la gráfica 29–4 el método de la media (MICE) no mostró un buen ajuste en la información imputada por un alto porcentaje de datos faltantes en la estación Tunshi.

Estación Urbina

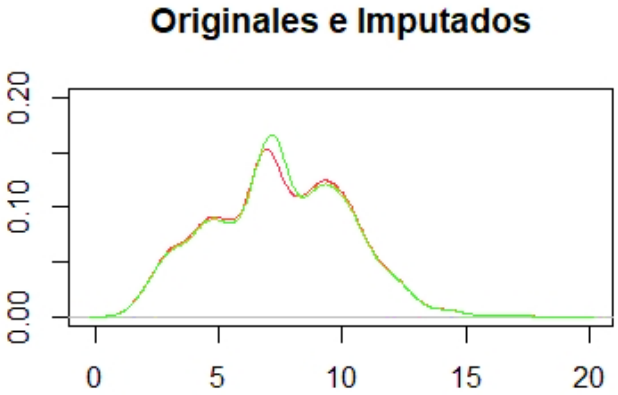


Ilustración 30–4: Imputación por la media Urbina

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 30–4 de la estación Tunshi evidenció un porcentaje alto de valores faltantes que representó el 31,71 % de la información. Por ende, el método de la media (MICE) logró imputar correctamente dicho porcentaje esto probó que es un buen método.

4.1.7. *Gráficas de imputación por método Hot Deck*

Para este método de asignación de "mejor ajuste"se hizo uso de la librería *library(VIM)* con la función *hotdeck()* en el software estadístico R.

Estación Alao

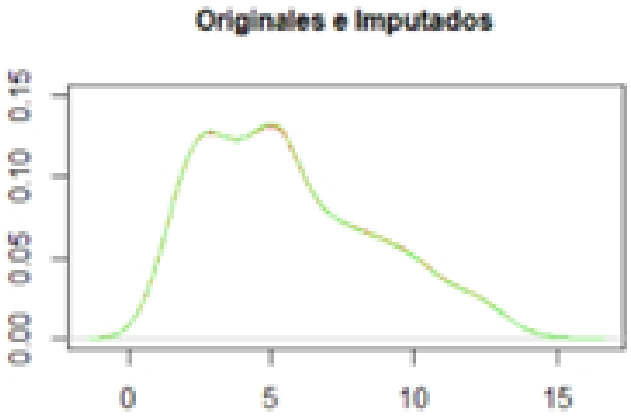


Ilustración 31–4: Imputación por Hot Deck de Alao

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Alao el método de Hot Deck presentó un buen ajuste en la velocidad de viento por un bajo porcentaje de datos faltantes (1,17 %) como se puede ver en la ilustración 31–4.

Estación Atillo

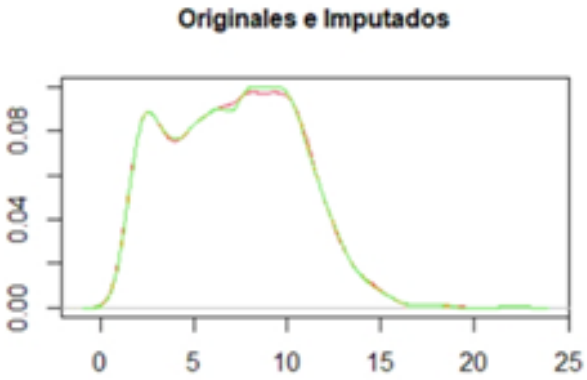


Ilustración 32–4: Imputación por Hot Deck Atillo

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 32–4 de la estación Atillo mostró un porcentaje alto de valores faltantes que representó el 34,94 % de la información. Por ende, el método de Hot Deck logró imputar correctamente dicho porcentaje esto probó que es un buen método.

Estación Cumandá

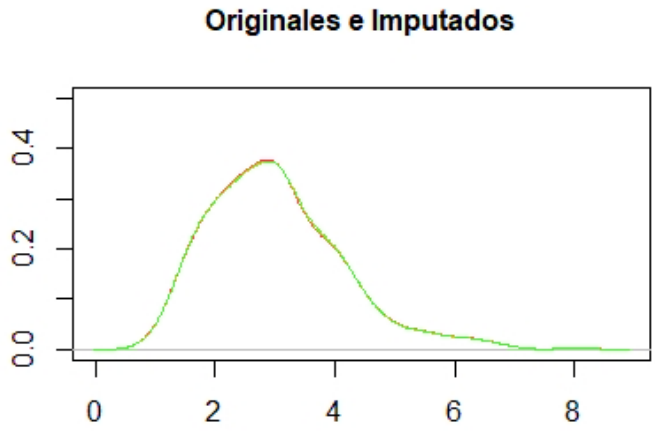


Ilustración 33–4: Relleno por Hot Deck de Cumandá

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 33–4 se observó un porcentaje bajo de valores faltantes que representó el 2,86 %. Por tanto, el método Hot Deck logró imputar correctamente dicho porcentaje en la estación Cumandá esto probó que es un buen método.

Estación ESPOCH

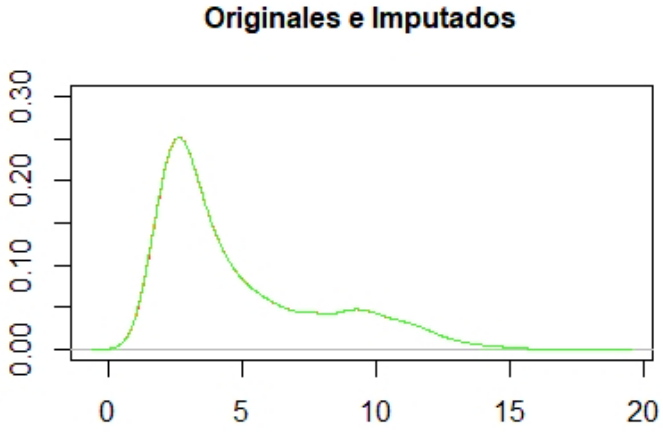


Ilustración 34–4: Imputación por Hot Deck ESPOCH

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método de Hot Deck obtuvo un buen ajuste entre los datos reales y los imputados por un

porcentaje bajo de información faltante en la estación ESPOCH como se presentó en la ilustración 34–4.

Estación Matus

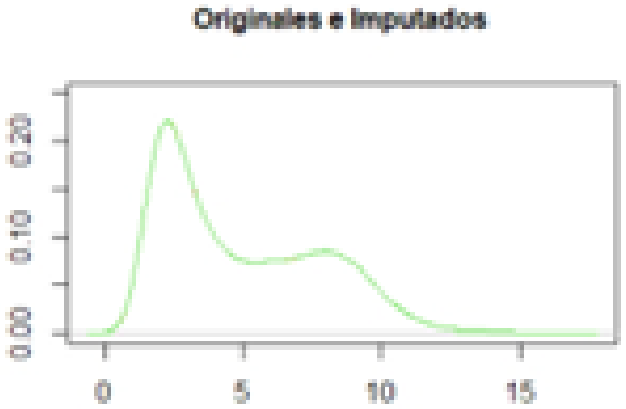


Ilustración 35–4: Imputación por Hot Deck Matus

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 35–4 de la estación Matus evidenció un porcentaje bajo de valores faltantes que representó el 0,34 % de la información. Por ende, el método de Hot Deck logró imputar correctamente dicho porcentaje.

Estación Multitud

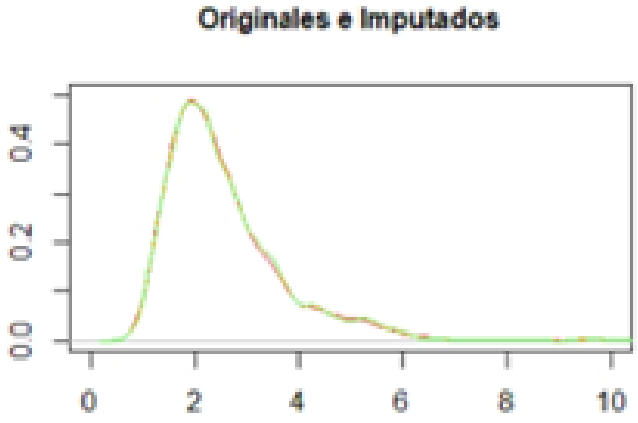


Ilustración 36–4: Imputación por Hot Deck Multitud

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Multitud el método de Hot Deck se observó un buen ajuste en la velocidad de viento además teniendo un alto porcentaje de datos faltantes del 47,46 % evidenciando que es un buen método para dicha estación como mostró la figura 36–4.

Estación Quimiag

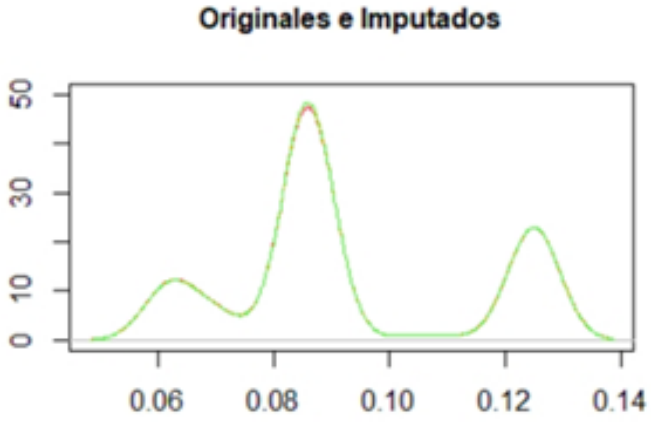


Ilustración 37–4: Imputación por Hot Deck Quimiag

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación Quimiag presentó un buen ajuste entre los datos reales y los imputados dado por un porcentaje bajo de información faltante del 2,25 % aplicando el metodo de Hot Deck como se observó en la ilustración 37–4.

Estación San Juan

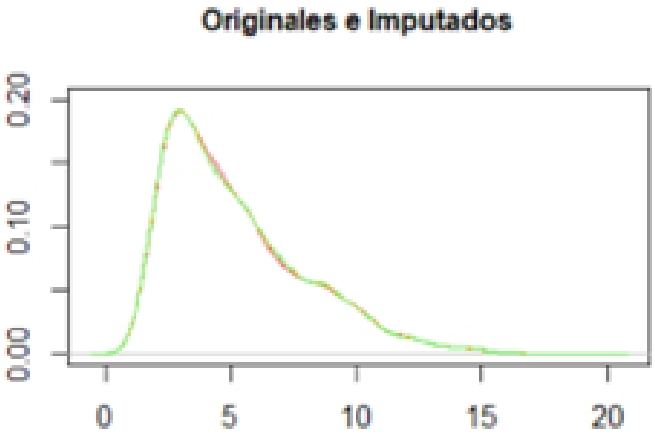


Ilustración 38–4: Imputación por Hot Deck San Juan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 38–4 de la estación San Juan se mostró un porcentaje bajo de valores faltantes que representó el 14,73 % de la información. Como resultado, el método de Hot Deck logró imputar correctamente dichos valores.

Estación Tixán

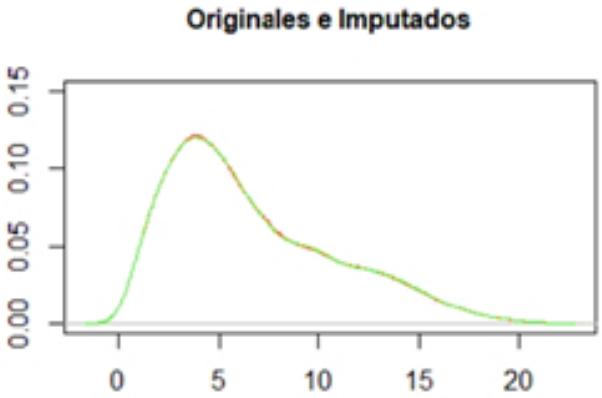


Ilustración 39–4: Imputación por Hot Deck Tixán

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método de Hot Deck obtuvo un buen ajuste entre los datos reales y los imputados dado por un bajo porcentaje de valores faltantes en la estación Tixan como mostró la ilustración 39–4.

Estación Tunshi

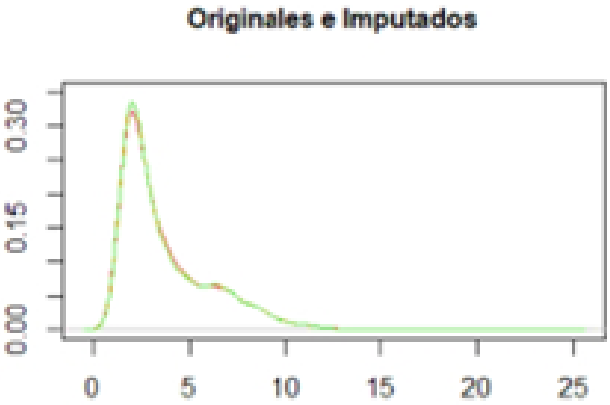


Ilustración 40–4: Imputación Hot Deck de Tunshi

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El gráfico 40–4 evidenció un buen ajuste entre los datos reales y los imputados además que tuvo un porcentaje alto de valores faltantes del 31,71 %. Como resultado, el método Hot Deck es un procesó muy adecuado para la imputación de la estación Tunshi.

Estación Urbina

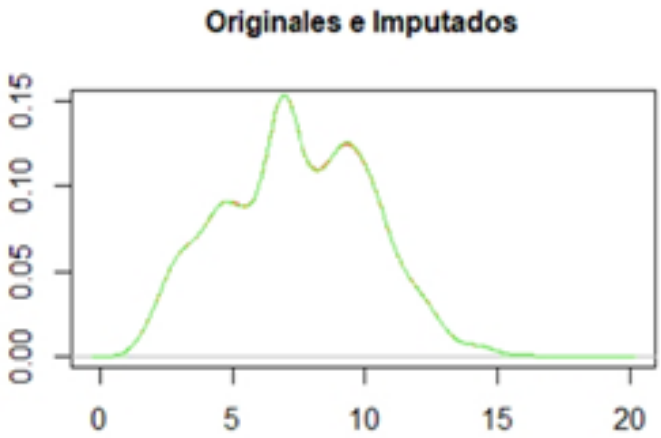


Ilustración 41–4: Imputación por Hot Deck de Urbina

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método de Hot Deck presentó un buen ajuste entre los datos reales y los imputados con un porcentaje bajo de valores faltantes del 2,88 % en la estación Urbina como mostró la figura 41–4.

4.1.8. *Gráficas de imputación por método PCA*

El algoritmo de PCA (Análisis de Componentes Principales) es una técnica de imputación de datos donde la idea básica es que los valores faltantes se pueden estimar como una composición de las variables restantes en el dataset, se hizo el uso de la función *impPCA()* de la librería *library(VIM)*.

Estación Alao

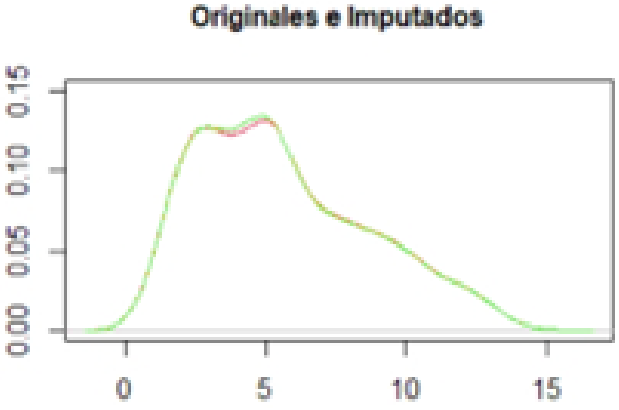


Ilustración 42–4: Imputación PCA gráfica de Alao

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Alao el método PCA índico un buen ajuste en la velocidad de viento por un bajo porcentaje de datos faltantes (1,17 %) como se puede ver en la ilustración 42–4.

Estación Atillo

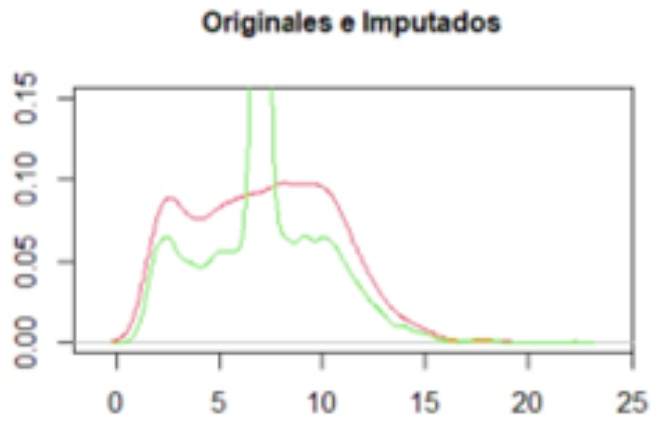


Ilustración 43–4: Imputación por PCA gráfica Atillo

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 43–4 de la estación Atillo mostró un porcentaje alto de valores faltantes que representó el 34,94 % de la información. Por ende, el método PCA no logró imputar correctamente dicho porcentaje.

Estación Cumandá

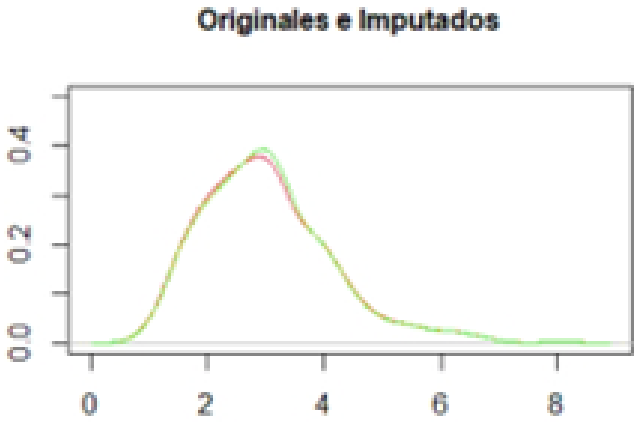


Ilustración 44–4: Imputación PCA gráfica Cumandá

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 44–4 se evidenció un porcentaje bajo de valores faltantes que representó el 2,86 %. Por tanto, el método PCA logró imputar correctamente dicho porcentaje en la estación Cumandá esto probó que es un buen método.

Estación ESPOCH

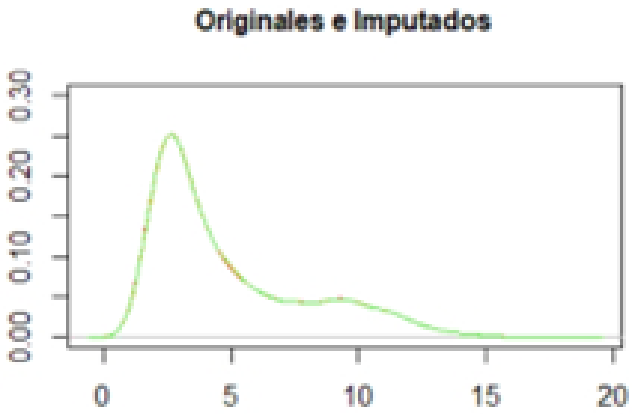


Ilustración 45–4: Imputación PCA gráfica ESPOCH

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método PCA obtuvo un buen ajuste entre los datos reales y los imputados por un porcentaje bajo de información faltante en la estación ESPOCH como se observó en la ilustración 45–4.

Estación Matus

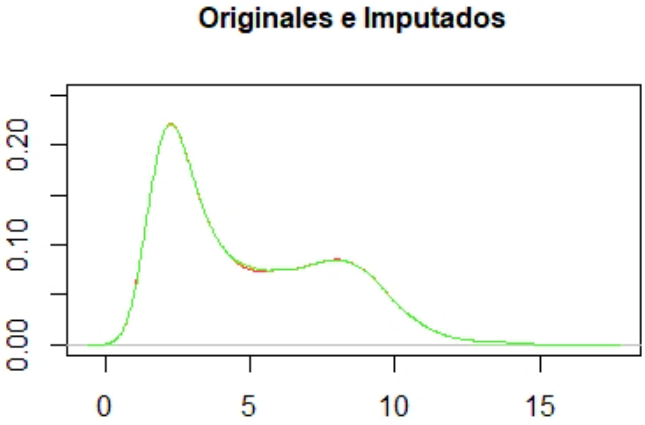


Ilustración 46–4: Imputación PCA gráfica de Matus

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 46–4 de la estación Matus presentó un porcentaje bajo de valores faltantes que representó el 0,34 % de la información. Por ende, el método PCA logró imputar correctamente dicho porcentaje.

Estación Multitud

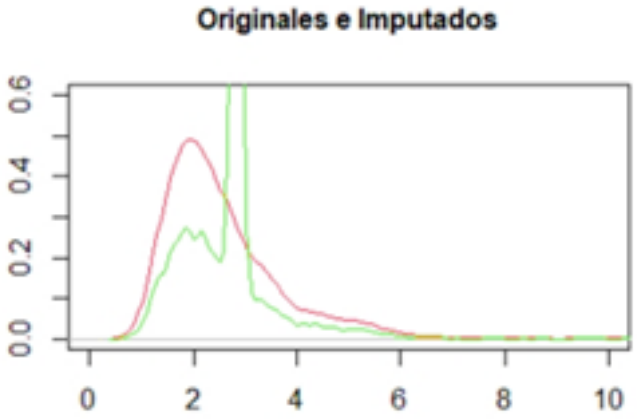


Ilustración 47–4: Imputación PCA gráfica Multitud

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Para la estación Multitud en el método PCA se observó un mal ajuste en la velocidad de viento por un alto porcentaje de datos faltantes del 47,46 % evidenciando que no es un buen método para dicha estación como mostró la figura 47–4. .

Estación Quimiag

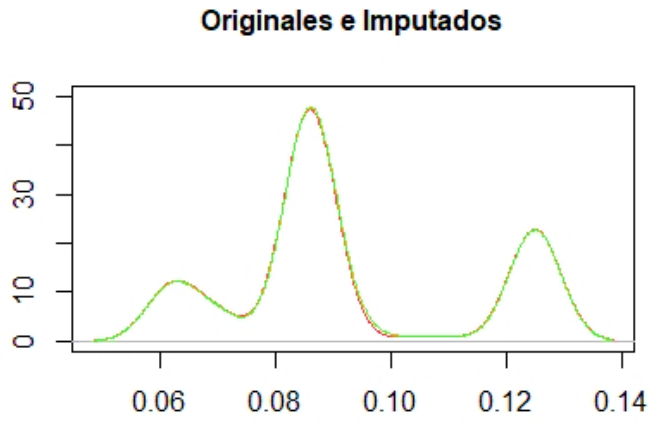


Ilustración 48–4: Imputación PCA gráfica Quimiag

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación Quimiag se evidenció un buen ajuste entre los datos reales y los imputados dado por un porcentaje bajo de información faltante del 2,25 % aplicando el metodo PCA como se observó en la ilustración 48–4.

Estación San Juan

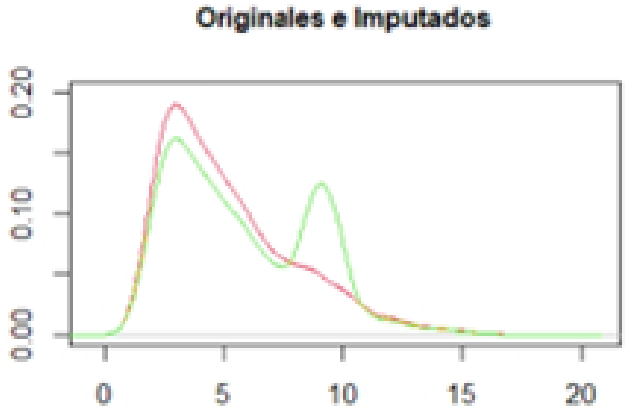


Ilustración 49–4: Imputación PCA gráfica San Juan

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 49–4 de la estación San Juan se mostró un porcentaje bajo de valores faltantes que representó el 14,73 % de la información. Como resultado, el método PCA no logró imputar correctamente dichos valores.

Estación Tixán

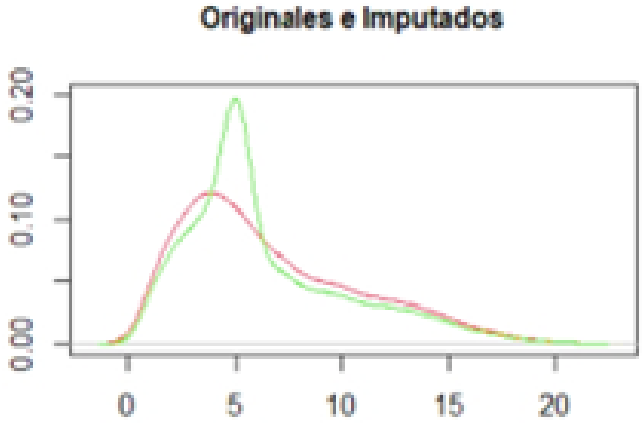


Ilustración 50–4: Imputación por PCA gráfica Tixán

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método PCA no obtuvo un buen ajuste entre los datos reales y los imputados aunque tuvo un bajo porcentaje de valores faltantes en la estación Tixan como mostró la ilustración 50–4.

Estación Tunshi

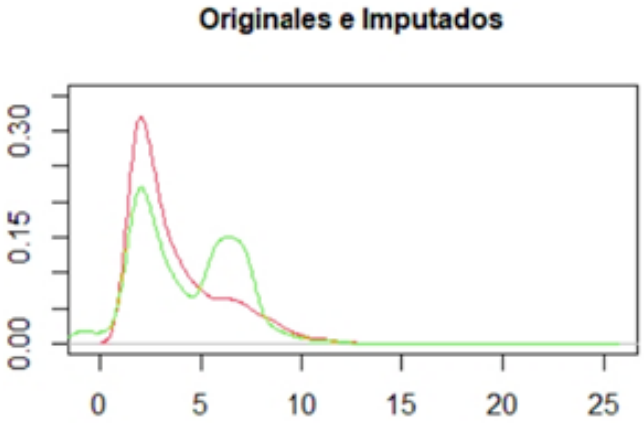


Ilustración 51–4: Imputación PCA gráfica de Tunshi

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El gráfico 51–4 índico un mal ajuste entre los datos reales y los imputados dado por un porcentaje alto de valores faltantes del 31,71 %. Como resultado, el método PCA es un procesó nada adecuado para la imputación de la estación Tunshi.

Estación Urbina

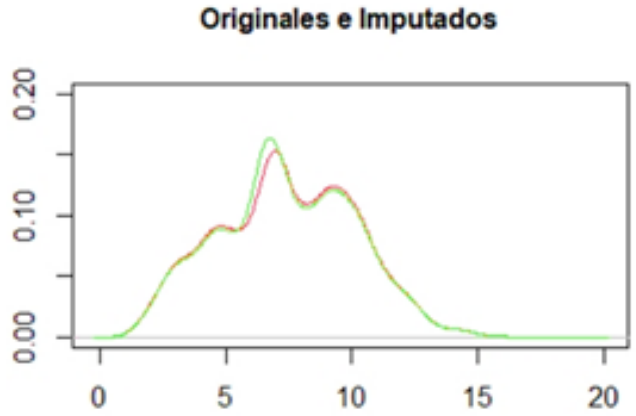


Ilustración 52–4: Imputación PCA gráfica de Urbina

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método PCA presentó un buen ajuste entre los datos reales y los imputados por un porcentaje bajo de valores faltantes del 2,88 % en la estación Urbina como mostró la figura 52–4.

4.1.9. *T-test: Comparación de medias poblacionales independientes*

Se hizo uso del test de comparación de medias poblacionales para determinar si existe diferencia estadística entre los grupos de muestras, el método utilizado fue el test mencionado entre los datos originales e imputados.

```
        Welch Two Sample t-test

data:  ori and d3$Media
t = -0.37263, df = 1197.5, p-value = 0.7095
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4121350  0.2805716
sample estimates:
mean of x mean of y
 5.572652  5.638433
```

Ilustración 53–4: T de student método de la Media (MICE) establecidos

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación Cumandá el *p-value* para el método de la Media (MICE) fue de 0.7095 mayor al nivel de significancia esto índico que las medias son iguales estadisticamente, es decir no existió diferencia entre los datos originales y los imputados como mostró la figura 53–4.

```
        Welch Two Sample t-test

data:  ori and d3$Forest
t = -0.28408, df = 1197.7, p-value = 0.7764
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3971551  0.2966909
sample estimates:
mean of x mean of y
 5.572652  5.622884
```

Ilustración 54–4: T de student método de Random Forest de árboles predictores

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El *p-value* mostrado para el método Random Forest fue de 0.7764 mayor al nivel de significancia esto índico que las medias son iguales estadisticamente, por lo tanto no existió diferencia entre los datos originales y los imputados en la estación ESPOCH según la ilustración 54–4.

```
        Welch Two Sample t-test

data:  ori and HD$x
t = -0.32122, df = 1197.6, p-value = 0.7481
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4035183  0.2899750
sample estimates:
mean of x mean of y
 5.572652  5.629423
```

Ilustración 55–4: T de student método de Hot Deck para las variables predictoras

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

El método Hot Deck aplicando el test mostró un *p-value* de 0.7481 mayor al nivel de significancia esto índico que las medias son iguales estadisticamente, como resultado no existió diferencia entre

los datos originales y los imputados en la estación Matus según la ilustración 55–4.

```
        Welch Two Sample t-test

data:  ori and PCA$x
t = -0.28152, df = 1197.6, p-value = 0.7784
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3962973  0.2968385
sample estimates:
mean of x mean of y
 5.572652  5.622381
```

Ilustración 56–4: T de student método PCA o análisis de componentes principales

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 56–4 presentó un *p-value* de 0.7784 para el método PCA que fue mayor al nivel de significancia esto índico que las medias son iguales estadisticamente, por tanto no existió diferencia entre los datos originales y los imputados en la estación Quimiag.

```
        Welch Two Sample t-test

data:  ori and PCA$x
t = -37.748, df = 10907, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.562566 -2.309564
sample estimates:
mean of x mean of y
 4.665990  7.102056
```

Ilustración 57–4: T de student método PCA, análisis de componentes principales de las variables

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En las ilustraciones (A: Multitud, B: San Juan, C: Tixan, D: Tunshi, E: Urbina) se verificó el test de T de student a los métodos de la media (MICE), Random Forest, Hot Deck y PCA en el que su *p-value* fue menor *2.2e-16* por tanto se rechazó la hipótesis nula, es decir existió diferencia significativa entre las medias poblacionales con un porcentaje alto de datos faltantes.

4.1.10. *Comparación gráfica de los métodos aplicados*

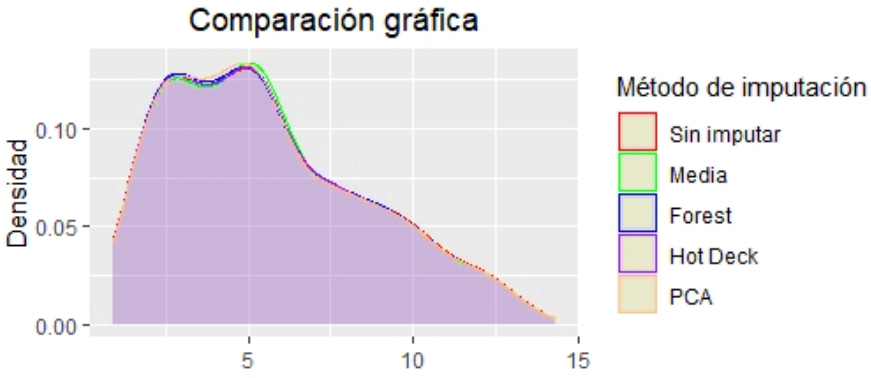


Ilustración 58–4: Comparación de los métodos "Alao" la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 58–4, presentó que la gráfica de densidad con los distintos métodos aplicados, tuvieron un buen ajuste a los datos originales en la estación Alao.

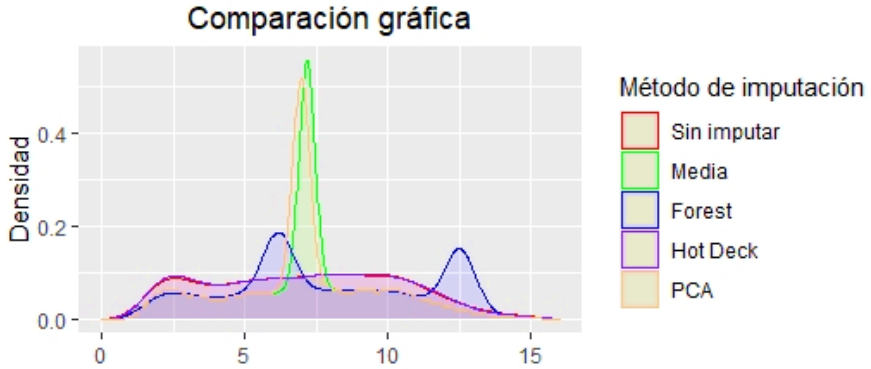


Ilustración 59–4: Comparación de métodos "Atillo" la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la ilustración 59–4 se mostró que no existío una buena aproximación a la gráfica de densidad original con los métodos de Random Forest, Media y PCA en comparación con Hot Deck, que presentó una buena aproximación a la gráfica original de la estación Atillo.

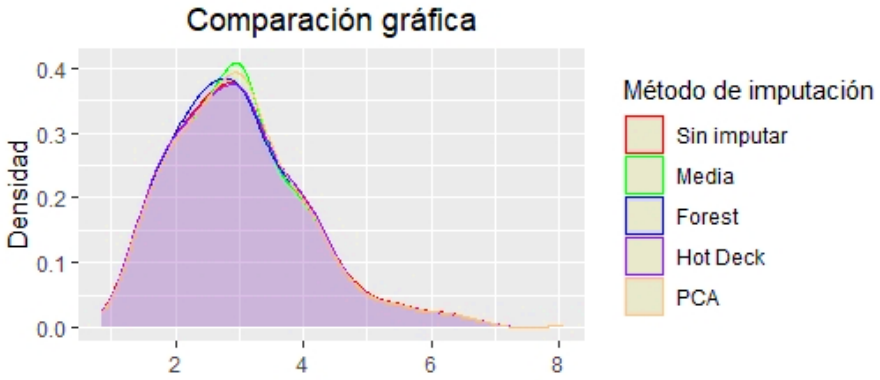


Ilustración 60–4: Comparación de los métodos desde la gráfica "Cumandá"

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La ilustración 60–4 evidenció que el método de imputación *Random Forest* y *Hot Deck* tuvo una buena aproximación a los datos originales en comparación con los métodos por la media y PCA, que presentaron picos en la gráfica de densidad de la estación Cumandá.

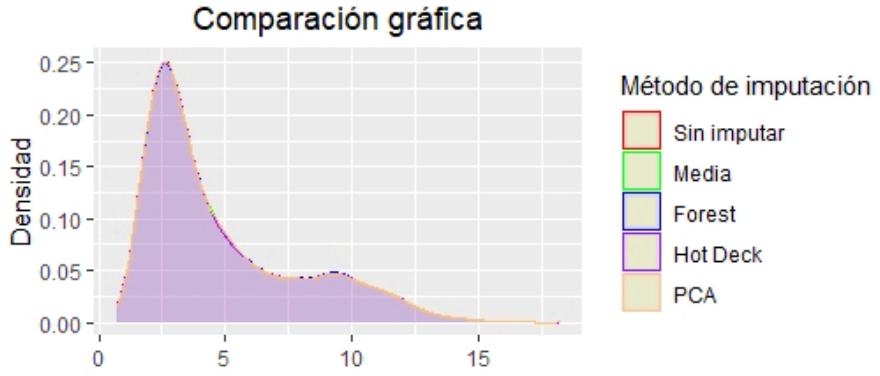


Ilustración 61–4: Comparación de los métodos desde la gráfica "ESPOCH"

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Se apreció en la ilustración 61–4 que los métodos de imputación aplicados generaron un buen acercamiento a la distribución de datos original de la estación meteorológica ESPOCH, donde el porcentaje de datos faltantes en la locación no era elevado.

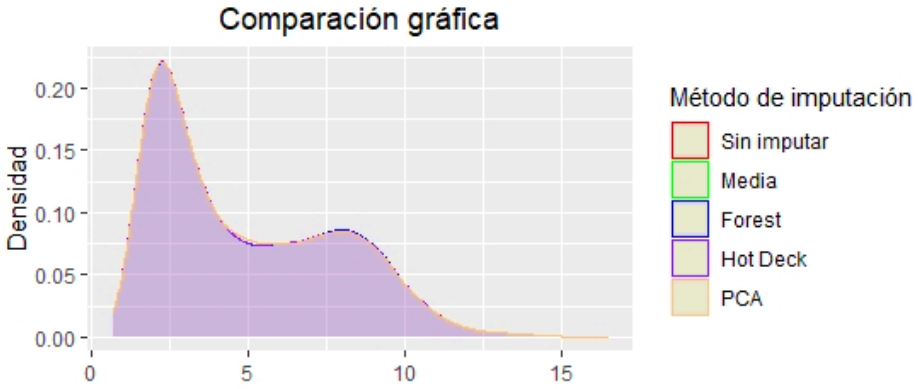


Ilustración 62–4: Comparación los métodos "Matus" en la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

En la estación meteorológica Matus se observó que los métodos de imputación aplicados generaron un buen acercamiento a la distribución de datos original, donde el porcentaje de datos faltantes en la locación no era elevado como presentó la ilustración 62–4.

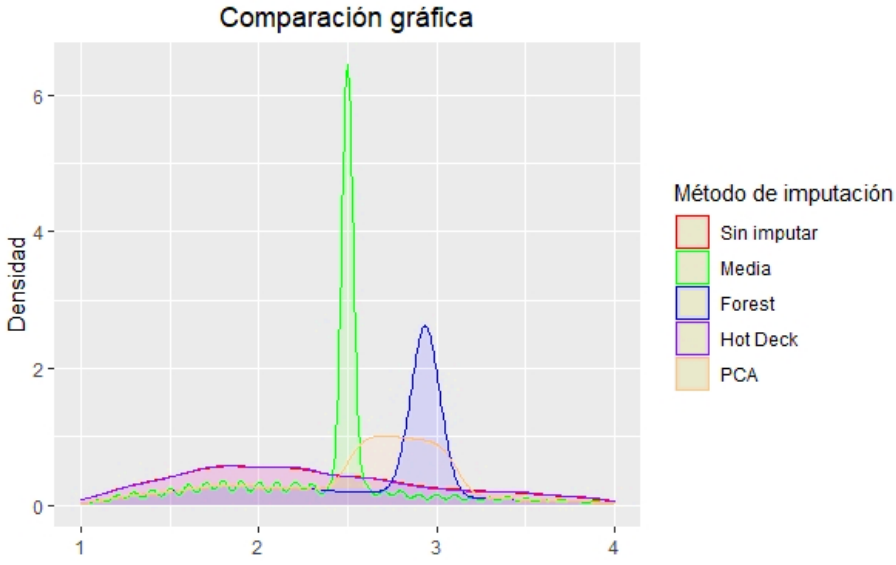


Ilustración 63–4: Comparación de métodos "Multitud" gráfica de dispersión

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación meteorológica Multitud mostró que los métodos de imputación PCA, Forest y Media tenían problemas en la imputación debido a la presencia de una gran cantidad de datos faltantes, en comparación con el método Hot Deck, que presentó una buena aproximación en comparación con los demás métodos como presentó la ilustración 63–4.

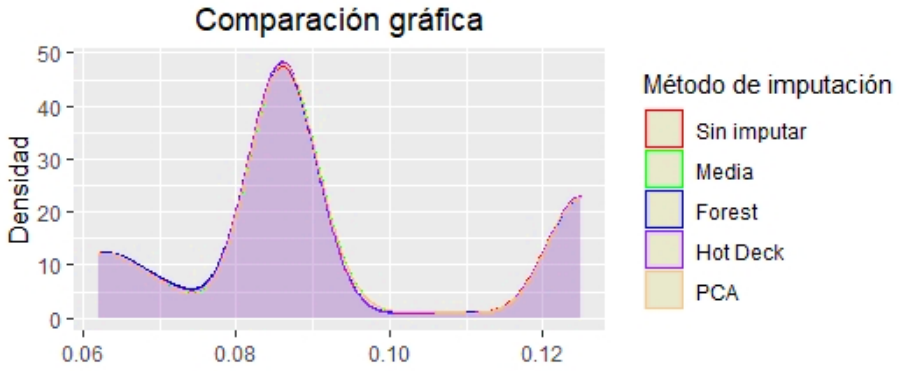


Ilustración 64–4: Comparación de métodos "Quimiag" la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación meteorológica Quimiag presentó que los métodos de imputación aplicados generaron un buen acercamiento a la distribución de datos original, donde el porcentaje de datos faltantes en la locación no era elevado como mostró la ilustración 64–4.

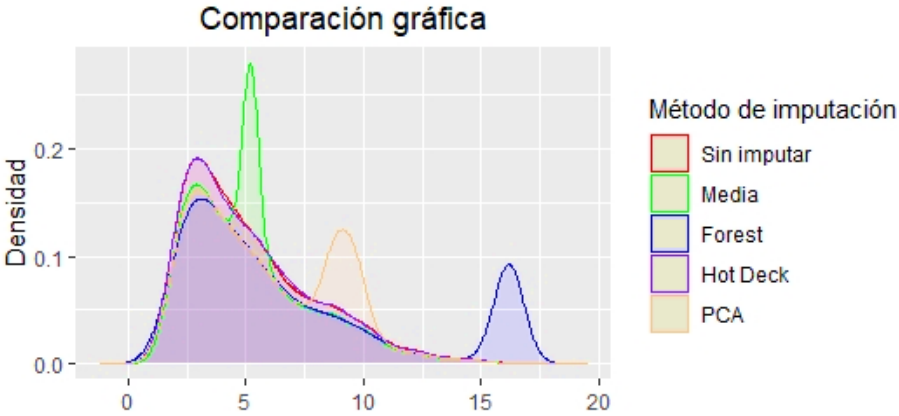


Ilustración 65–4: Comparación de métodos "San Juan" la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación meteorológica San Juan evidenció que los métodos de imputación PCA, Forest y Media no lograron una buena aproximación a los datos originales cuando existió un alto porcentaje de información faltante. En contraste, el método Hot Deck presentó una mejor adaptación a situaciones de este tipo como presentó la ilustración 65–4.

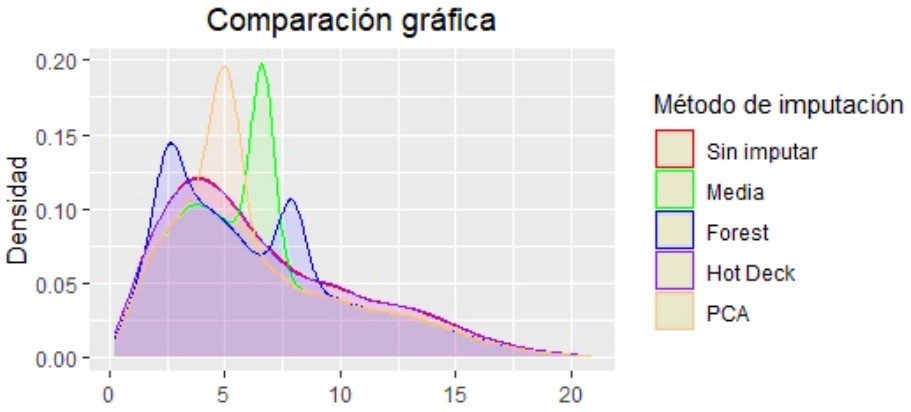


Ilustración 66–4: Comparación de los métodos "Tixán" la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación meteorológica Tixan mostró que los métodos de imputación PCA, Random Forest y Media no lograron una buena aproximación a los datos originales cuando existió un alto porcentaje de información faltante. En contraste, en el método Hot Deck se observó una mejor adaptación a situaciones de este tipo según la ilustración 66–4.

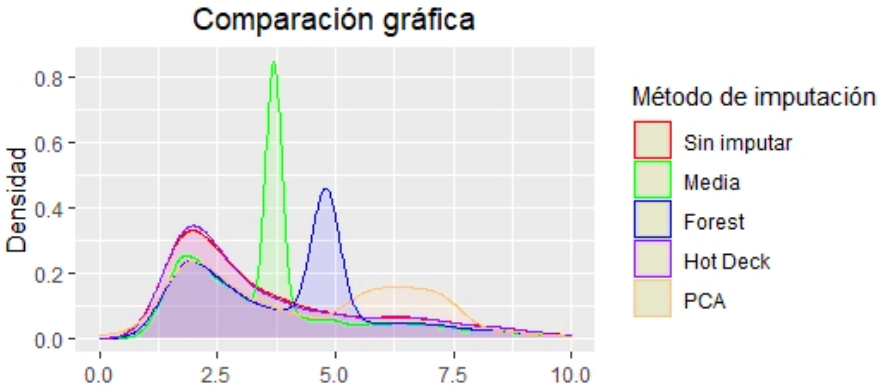


Ilustración 67–4: Comparación de métodos "Tunshi" la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

La estación meteorológica Tunshi índico que los métodos PCA, Forest y Media no mostraron una buena aproximación a la distribución de los datos originales, incluso se presenció un gran porcentaje de información faltante. Por otro lado, el método Hot Deck fue el que presentó una adecuada aproximación como se detalló en la ilustración 67–4.

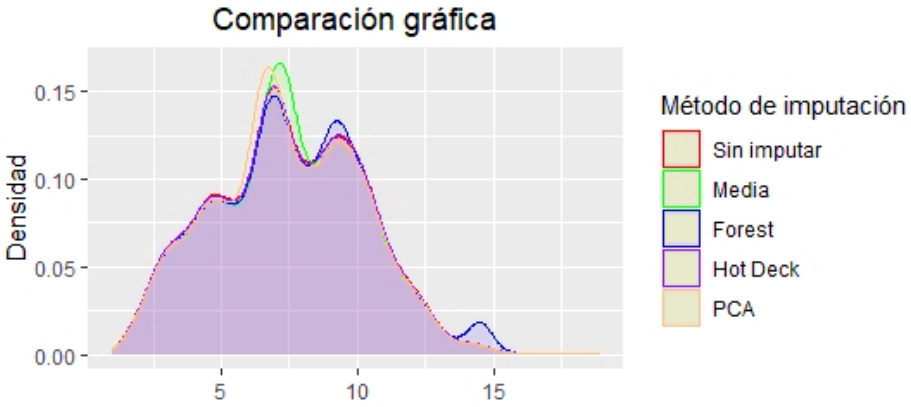


Ilustración 68–4: Comparación de métodos "Urbina" de la gráfica de densidad

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Velastegui, Evelyn y Horna, Adrian, 2022.

Se apreció en la ilustración 68–4 los métodos de imputación por media y PCA no lograron una buena aproximación, presentando diferencias significativas con los datos originales. Esto sugirió que los procesos Random Forest y Hot Deck son más adecuados en la imputación de información. Sin embargo, esto dependió del caso específico y pudo haber situaciones en las que la media o PCA brindaron mejores resultados.

4.1.11. *Comparación de los métodos mediante los errores*

En base a los análisis anteriores se cálculo los errores para cada estación como el Error medio de pronóstico (EMP), el Error medio cuadrático (EMC) y el Desvío absoluto medio (DAM).

Tabla 5–4: Errores método random Forest

Error\Estación	Alao	Atillo	Cumandá	Espoch	Matus	Multitud
EMP	-0.0509	-3.2561	-0.0761	-0.0267	-0.026	-1.4005
EMC	0.2621	33.8399	0.2093	0.2048	0.2076	4.1328
DAM	0.0509	3.2561	0.0761	0.0267	0.026	1.4005
Error\Estación	**Quimiag**	**San Juan**	**Tixan**	**Tunshi**	**Urbina**	
EMP	-0.0018	-2.3703	-0.8742	-1.5565	-0.3382	
EMC	0.0002	38.1596	5.754	7.7574	4.1929	
DAM	0.0018	2.3703	0.8742	1.5565	0.3382	

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

Tabla 6–4: Errores método de la Media

Error\Estación	Alao	Atillo	Cumandá	Espoch	Matus	Multitud
EMP	-0.0658	-2.5057	-0.0864	-0.0190	-0.0169	-1.3421
EMC	0.3709	17.9704	0.2605	0.0912	0.0836	3.7953
DAM	0.0658	2.5057	0.0864	0.0190	0.0169	1.3421
Error\Estación	**Quimiag**	**San Juan**	**Tixan**	**Tunshi**	**Urbina**	
EMP	-0.0021	-0.7656	-1.1066	-1.1717	-0.2142	
EMC	0.0002	3.9784	7.3321	4.3290	1.5944	
DAM	0.0021	0.7656	1.1066	1.1717	0.2142	

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

Tabla 7–4: Errores método Hot Deck

Error\Estación	Alao	Atillo	Cumandá	Espoch	Matus	Multitud
EMP	-0.0568	-2.4635	-0.0846	-0.0200	-0.0163	-1.3309
EMC	0.3118	21.3972	0.2726	0.1366	0.1006	5.6807
DAM	0.0568	2.4635	0.0846	0.0200	0.0163	1.3309
Error\Estación	**Quimiag**	**San Juan**	**Tixan**	**Tunshi**	**Urbina**	
EMP	-0.0021	-0.7773	-1.1323	-1.1526	-0.2169	
EMC	0.0002	5.2707	10.7009	5.8465	1.8466	
DAM	0.0021	0.7773	1.1323	1.1526	0.2169	

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

Tabla 8–4: Errores método PCA

Error\Estación	Alao	Atillo	Cumandá	Espoch	Matus	Multitud
EMP	-0.0497	-2.4361	-0.0908	-0.0193	-0.0169	-1.3425
EMC	0.2133	16.9897	0.2897	0.0946	0.0838	3.7976
DAM	0.0497	2.4361	0.0908	0.0193	0.0169	1.3425
Error\Estación	**Quimiag**	**San Juan**	**Tixan**	**Tunshi**	**Urbina**	
EMP	-0.0021	-1.3542	-0.8397	-1.6914	-0.1842	
EMC	0.0002	12.4932	4.2450	11.0224	1.1863	

Continúa en la siguiente página.

Tabla 8–4 : *Continuación de la página anterior.*

Error\Estación	Quimiag	San Juan	Tixan	Tunshi	Urbina
DAM	0.0021	1.3545	0.8397	1.7390	0.1842

Fuente: Grupo De Energías Alternativas y Ambiente.

Realizado por: Horna E y Velastegui M, 2022.

Considerando todas las estaciones se han evaluado las métricas de los métodos de imputación: Error medio de pronóstico (EMP), Error medio cuadrático (EMC) y Desvío absoluto medio (DAM), con el objetivo de identificar cuál de los métodos obtuvieron estimaciones bajas, de las comparaciones gráficas anteriores los métodos Media (MICE) y Hot Deck presentaron errores bajos en el objeto de análisis. En efecto, la estación de Quimiag presentó valores bajos en EMP, EMC y DAM indicando un mejor ajuste en la imputación de datos faltantes. Sin embargo, en la estación de Atillo se evidenció que no hubo un buen ajuste en el relleno por los valores 2.4635, 21.3972, 2.4635 para el EMP, EMC y DAM respectivamente en el método de Hot Deck y de manera similar, para el MICE (EMP = 2.5057, DAM = 2.5057, EMC = 17.974).

4.1.12. *Discusión de resultados*

Para rellenar los datos faltantes de la velocidad de viento se utilizó la metodología de Hot Deck, donde se aplicó en la estación de Quimiag en el año 2018 estimando 30 datos lo cual representó el 2.25 % del total de la información. Este resultado es semejante al estudio de Ferreira (2004, pp. 163-170) que establece la imputación del 3.25 % de datos ausentes, con respecto a la máxima similitud por causas provocadas o naturales, de igual forma Quishpe (2020) estableció el 2.28 % y 10 % de faltantes para la imputación. Cárdenas y Urgilés (2020, p. 162) trabajaron con la metodología de la Media (MICE) donde mostraron un máximo del 10 % de datos faltantes, Checa (2020) realizó el estudio de la estación meteorológica Quimiag de la provincia de Chimborazo del año 2017 mediante el método de la Media (MICE) con un total de 17 valores perdidos que representa el 2.01 % determinando una imputación eficaz.

Cárdenas y Urgilés (2020, p. 162) aplicaron el método de la media (MICE) en la investigación efectuada a las estaciones metereológicas donde la que mejor se ajustó fue Alao por su bajo porcentaje de valores perdidos, en cambio las estaciones que no se adecuaron a los datos originales fueron Quimiag, Multitud, Atillo y Tixan verificando de forma gráfica las imputaciones; el estudio de Checa (2020) coincide en que la estación Alao se ajustó a sus datos originales de forma analítica y gráfica, en tanto Quishpe (2020) en las estaciones Tunshi y Multitud donde se aplicó el método Hot Deck se visualiza un buen ajuste de los datos reales con los imputados, Arias (2022) realizó el método

Random Forest con distintos porcentajes de datos ausentes, las estaciones evaluadas en el estudio concordaron que Alao, ESPOCH y Quimiag tiene un buen ajuste, en tanto las demás estaciones no imputaron correctamente sus datos. Haro et al. (2020) determinó que la estación Alao junto a Quimiag son las que más se adecuaron a los datos originales, al contrario que Cumandá y Multitud no se ajustaron de la forma necesaria utilizando el método PCA.

Cárdenas y Urgilés (2020, p. 162) determinaron un RMS de 0.29, EMP 0.00 y DAM 0.21 para el método de la media (MICE), Haro et al. (2020) utilizaron la imputación por PCA determinando un RMS de 0.85, DAM 0.58 y EMP 0.47; Arias (2022) haciendo uso de Random Forest da a conocer un RMSE de 0.54 y EMP de 0.51; Quishpe (2020) dio a conocer un RMS de 0.0051 y DAM de 0.009 aplicando Hot Deck. Los estudios realizados por distintos autores coinciden en que los métodos de la media (MICE) y Hot Deck son lo más confiables dado el tamaño del porcentaje de datos faltantes y la comparación de sus errores preciso que son los modelos más adecuados para la imputación de este estudio.

CONCLUSIONES

- Mediante el análisis exploratorio se evidenció un total de 703248 registros tomados de las matrices meteorológicas, con 677 datos atípicos obtenidos mediante el test de Rosner's Outlier y un total de 10964 datos faltantes, a partir de esto se pudo procesar la información y verificar los estadísticos de cada una de las estaciones, con un promedio de 2.303 m/s, una velocidad de viento máxima de 4.102 m/s y mínima de 1.000 m/s en la estación Cumandá, mientras que en Tixán presentó un promedio de 8.31 m/s con un máximo de 20.703 m/s, un mínimo de 1.203 m/s respectivos a la variable mencionada.

- En base a la revisión bilbiografica las técnicas estadísticas para la completación de datos faltantes en la variable velocidad de viento fueron los métodos Random Forest, Hot Deck, Imputación por la media (MICE) e Iterative PCA imputation. Por tanto, dichos métodos se aplicaron en la imputación de datos en las estaciones meteorologicas, donde el 1,17 % de faltantes se encontró en la estación Alao y el 47,46 % en la estación de Multitud.

- Mediante los errores EMP, EMC y DAM se compararon los métodos mencionados anteriormente, verificando que el método Hot Deck y la media (MICE) presentaron errores más bajos en las estaciones metereologicas. Sin embargo, la estación de Quimiag presentó errores cercanos a cero estó índica un mejor ajuste en la imputación de datos faltantes en dicha estación.

RECOMENDACIONES

- Efectuar un estudio investigativo más afondo de otras variables que contemplen las bases de datos proporcionadas por el GEAA.
- Aplicar nuevas técnicas de relleno de datos para poder darle continuidad a este estudio, explorando afondo la variable velocidad de viento de suma importancia para obtener resultados relevantes para la problemática, mismo que pretenden reducir la cantidad de información faltante en matrices muy grandes.
- Se recomienda que los grupos de investigación encargados de monitorear la información a partir de los instrumentos de medición de datos puedan darles mantenimiento a dichas herramientas, dado que a partir de ellos se puede realizar cualquier tipo de investigación y estudio necesario por la información que procesan.

BIBLIOGRAFÍA

ARAYA LÓPEZ, J.L., 2014. 2014. Experiencias en la aplicación operativa de un método multivariado de imputación de datos meteorológicos. En: Accepted: 2015-03-23T21:03:49Z, Tecnología en Marcha [en línea], vol. 27, no. 3, [consulta: 3 julio 2023]. Disponible en: https://repositoriotec.tec.ac.cr/handle/2238/4304.

ARIAS MUÑOZ, A.C., 2022. 2014. ARIAS MUÑOZ, A.C., 2022. Propuesta y evaluación de una estrategia para la imputación múltiple y multivariada de valores faltantes en series de tiempo del campo meteorológico utilizando aprendizaje automático = Proposal and evaluation of a strategy for multiple and multivariate imputación of missing values in time series of the meteorological field using machine learning. En: Accepted: 2022-12-08T16:05:47Z [en línea], [consulta: 24 agosto 2023]. Disponible en: https://repositoriotec.tec.ac.cr/handle/2238/14060.

AYALA, M.F., CARRERA VILLACRÉS, D. y TIERRA, A., 2018. Relación espacio-temporal entre estaciones utilizadas para el relleno de datos de precipitación en Chone, Ecuador. Revista Geográfica Venezolana [en línea], vol. 59, no. 2, [consulta: 3 julio 2023]. ISSN 1012-1617, 2244-8853. Disponible en: https://www.redalyc.org/articulo.oa?id=347760473005.

ANDRADES GRASSI, J.E., TORRES MANTILLA, H.A., LÓPEZ HERNÁNDEZ, J.Y., GOITÍA ACOSTA, A. y MEJÍAS DELGADO, J.E., 2018. Exploración espacio temporal de la distribución de datos faltantes de precipitación mensual en el centro occidente de Venezuela, con fines de selección de estaciones. Ciencia e Ingeniería [en línea], vol. 39, no. 2, [consulta: 28 junio 2023]. Disponible en: https://www.redalyc.org/journal/5075/507557606011/html/.

ATKINSON, A.D.J., ARIZA, F.J. y GARCÍA BALBOA, J.L., 2007. Estimadores robustos: una solución en la utilización de valores atípicos para el control de la calidad posicional. GeoFocus. International Review of Geographical Information Science and Technology [en línea], no. 7, [consulta: 28 junio 2023]. ISSN 1578-5157. Disponible en: https://www.geofocus.org/index.php/geofocus/article/view/116.

BARÓN OROZCO, A.F., 2018. Análisis espacio temporal de la precipitación mensual, en la depresión momposina para los años 2012 a 2015. Caso de estudio: municipio de Mompox. [en línea]. Proyecto de grado. Bogotá, Colombia: Universidad Distrital Francisco José de Caldas. [consulta: 19 julio 2021]. Disponible en: http://repository.udistrital.edu.co/handle/11349/13816.

BUUREN, S. van y GROOTHUIS OUDSHOORN, K., 2011. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software [en línea], vol. 45, [consulta: 28 junio 2023]. ISSN 1548-7660. DOI 10.18637/jss.v045.i03. Disponible en: https://doi.org/10.18637/jss.v045.i03.

BREIMAN, L., 2001. Random Forests. Machine Learning [en línea], vol. 45, no. 1, [consulta: 7 agosto 2023]. ISSN 1573-0565. DOI 10.1023/A:1010933404324. Disponible en: https://doi.org/10.1023/A:1010933404324.

CÁRDENAS CAMPOVERDE, H.P. y URGILÉS ÁVILA, C.C., 2020. Análisis espacio-temporal meteorológico en una cuenca andina tropical del sur de Ecuador [en línea]. bachelorThesis. S.l.: Universidad de Cuenca. [consulta: 4 julio 2023]. Disponible en: http://dspace.ucuenca.edu.ec/handle/123456789/35041.

CARRERA VILLACRÉS, D.V., GUEVARA GARCÍA, P.V., TAMAYO BACACELA, L.C., BALAREZO AGUILAR, A.L., NARVÁEZ RIVERA, C.A. y MOROCHO LÓPEZ, D.R., 2016. 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 (Arica) [en línea], vol. 34, no. 3, [consulta: 21 enero 2021]. ISSN 0718-3429. DOI 10.4067/S0718-34292016000300010. Disponible en:http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1025-028X2014000100006&lng=es&nrm=iso

CEBALLOS BARBANCHO, A., MORÁN TEJEDA, E. y LÓPEZ MORENO, J.I., 2013. Análisis de la variabilidad espacio-temporal de las precipitaciones en el sector español de la cuenca del Duero (1961-2005). Boletín de la Asociación de Geógrafos Españoles [en línea], no. 61, [consulta: 28 junio 2023]. ISSN 0212-9426, 2605-3322. Disponible en:

https://dialnet.unirioja.es/servlet/articulo?codigo=4157738.

CHECA GAMARRA, M.C., 2020. Análisis geoestadístico de datos funcionales de temperatura del aire en la provincia de Chimborazo. En: Accepted: 2021-01-15T17:19:55Z [en línea], [consulta: 23 agosto 2023]. Disponible en: http://dspace.espoch.edu.ec/handle/123456789/14280.

CHICA RAMÍREZ, H.A., PEÑA QUIÑONES, A.J., GIRALDO JIMÉNEZ, J.F., OBANDO BONILLA, D. y RIAÑO HERRERA, N.M., 2014. SueMulador: Herramienta para la Simulación de Datos Faltantes en Series Climáticas Diarias de Zonas Ecuatoriales. Revista Facultad Nacional de Agronomía Medellín [en línea], vol. 67, no. 2, [consulta: 3 julio 2023]. ISSN 0304-2847. DOI 10.15446/rfnam.v67n2.44179. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0304-28472014000200012&lng=en&nrm=iso&tlng=es.

DIGGLE, P.J. y CHETWYND, A.G., 2011. 1 Introduction. Statistics and Scientific Method: An Introduction for Students and Researchers [en línea]. S.l.: Oxford University Press, pp. 36-56. [consulta: 28 junio 2023]. ISBN 978-0-19-954318-2. Disponible en: https://doi.org/10.1093/acprof:oso/9780199543182.003.0001.

FERREIRA, A.M., 2004. Metodologías de análisis y imputación de datos faltantes en series de velocidad del viento. VI Congreso Galego de Estatística e Investigación de Operacións: Vigo, 5,6 e 7 de novembro de 2003, 2004, ISBN 8468819492, págs. 163-170 [en línea]. S.l.: Universidade de Vigo, pp. 163-170. [consulta: 3 julio 2023]. ISBN 978-84-688-1949-5. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=6453266.

GUEVARA DÍAZ, J.M., 2013. Cuantificación del perfil del viento hasta 100 m de altura desde la superficie y su incidencia en la climatología eólica. Terra. Nueva Etapa [en línea], vol. XXIX, no. 46, [consulta: 7 agosto 2023]. ISSN 1012-7089, 2542-3266. Disponible en: https://www.redalyc.org/articulo.oa?id=72130181006.

HANKE, J. y REITSH, A., 1996. Pronosticos en los negocios [en línea]. Quinta edición. México. España: Prentice-Hall Hispanoamericana. [consulta: 7 agosto 2023]. ISBN 978-968-880-681-4.

Disponible en: http://www.marcialpons.es/libros/pronosticos-en-los-negocios/9789688806814/.

HARO RIVERA, S., ZÚÑIGA LEMA, L., MENESES FREIRE, A. y ESCUDERO VILLA, A., 2020. Determinación del comportamiento meteorológico del viento en la provincia de Chimborazo, Ecuador. En: Accepted: 2021-08-26T17:54:33Z [en línea], [consulta: 24 agosto 2023]. Disponible en: http://dspace.espoch.edu.ec/handle/123456789/14577.

INGENIOVIRTUAL, 2015. Tipos de gráficos y diagramas para la visualización de datos. ingeniovirtual.com [en línea]. [consulta: 28 junio 2023]. Disponible en: https://www.ingeniovirtual.com/tipos-de-graficos-y-diagramas-para-la-visualizacion-de-datos/.

KOWARIK, A. y TEMPL, M., 2016. Imputation with the R package VIM. Journal of Statistical Software, vol. 74, DOI 10.18637/jss.v074.i07.

KUHN, M. y JOHNSON, K., 2013. Applied predictive modeling. S.l.: s.n. ISBN 978-1-4614-6848-6.

MÁRQUEZ PÉREZ, V.E., USECHE CASTRO, L.M., MESA AVILA, D.M. y CHACON CONTRERAS, A.I., 2017. Estrategia de imputación con la media bajo el uso de árboles de regresión. Comunicaciones en Estadística [en línea], vol. 10, no. 1, [consulta: 7 agosto 2023]. ISSN 2027-3355, 2339-3076. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=6765744.

MIRABAL SOSA, M., ROBAINA GARCÍA, M. y URANGA PIÑA, R., 2010. R: una herramienta poco difundida y muy útil para la investigación clínica. Revista Cubana de Investigaciones Biomédicas [en línea], vol. 29, no. 2, [consulta: 13 febrero 2021]. ISSN 0864-0300. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0864-03002010000200012&lng=es&nrm=iso&tlng=es.

ORTEGA, R.M.M., PENDÁS, L.C.T., ORTEGA, M.M., ABREU, A.P. y CÁNOVAS, A.M., 2009. El coeficiente de correlación de los rangos de spearman caracterización. Revista Habanera

de Ciencias Médicas [en línea], vol. 8, no. 2, [consulta: 28 junio 2023]. ISSN , 1729-519X. Disponible en: https://www.redalyc.org/articulo.oa?id=180414044017.

R. RPubs by RStudio [en línea], s.f. [consulta: 28 junio 2023]. Disponible en: https://search.r-project.org/CRAN/refmans/EnvStats/html/rosnerTest.html.

RENDÓN MACÍAS, M.E., VILLASÍS KEEVE, M.Á. y MIRANDA NOVALES, M.G., 2016. Estadística descriptiva. Revista Alergia México [en línea], vol. 63, no. 4, [consulta: 28 junio 2023]. ISSN 0002-5151, 2448-9190. Disponible en: https://www.redalyc.org/articulo.oa?id=486755026009.

SALGADO, A., 2016. "Imputación de datos faltantes de temperatura mediante técnicas geoestadísticas en estaciones climáticas del Valle del Cauca en el periodo de 1985 a 2015". Anteproyecto [en línea], [consulta: 3 julio 2023]. Disponible en: https://www.academia.edu/26958191/Imputaci%C3%B3ndedatosfaltantesdetemperaturamediante t%C3%A9cnicasgeoesta%C3%ADsticasenestacionesclim%C3%A1ticasdelValledelCaucaenelpe riodode1985a2015.

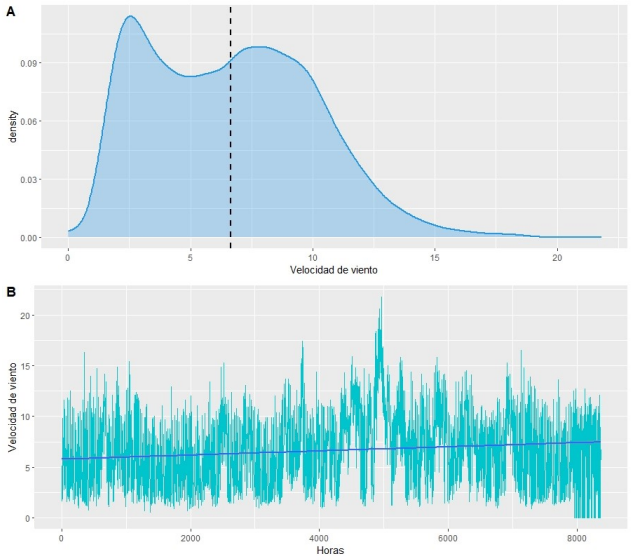
SANTANA, A. y HERNÁNDEZ, C., 2016. Inferencia estadística con R. Introducción a R [en línea]. [consulta: 28 junio 2023]. Disponible en: https://estadistica-dma.ulpgc.es/cursoR4ULPGC/11-inferencia-MediaVar.html.

VELÁZQUEZ, A.P., 2017. Estadística inferencial. Documento de trabajo [en línea], [consulta: 28 junio 2023]. Disponible en: https://www.repositorionacionalcti.mx/recurso/oai:centrogeo.repositorioinstitucional.mx:1012/159.

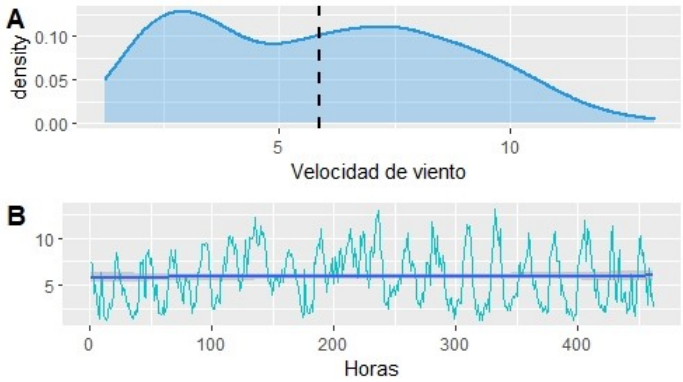
VIADA, C., BOUZA, C., BALLESTEROS, J., FORS, M., ROBAINA, M. y ROLANDO, U., 2016. Revisión sistemática de los métodos de imputación de datos faltantes. . S.l.: s.n., pp. 113-130. ISBN 978-84-608-4246-0.

ANEXOS

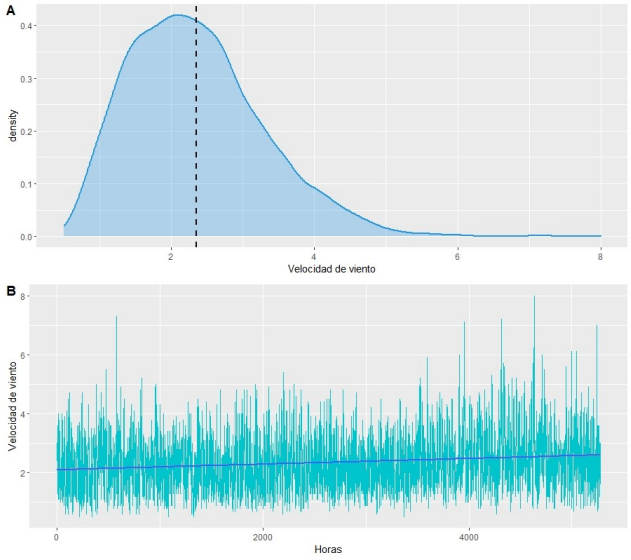
ANEXO A: ESTACIÓN METEREOLÓGICA ATILLO 2017



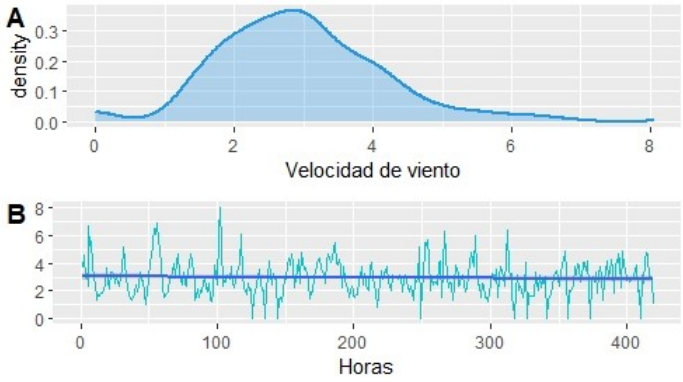
ANEXO B: ESTACIÓN METEREOLÓGICA ATILLO 2014



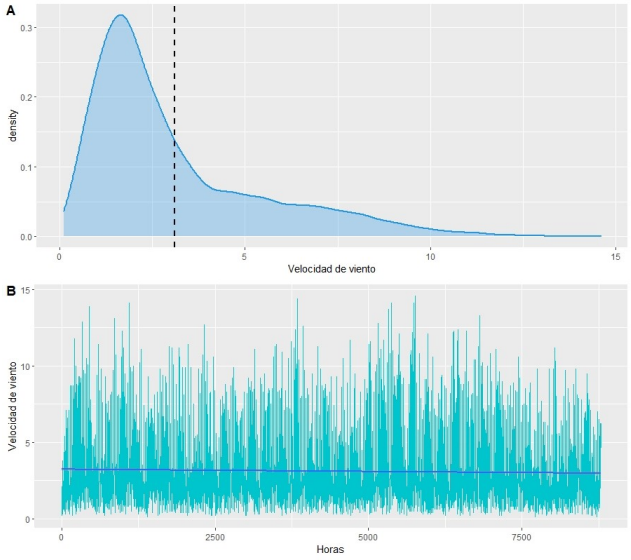
ANEXO C: ESTACIÓN METEREOLÓGICA CUMANDÁ 2016



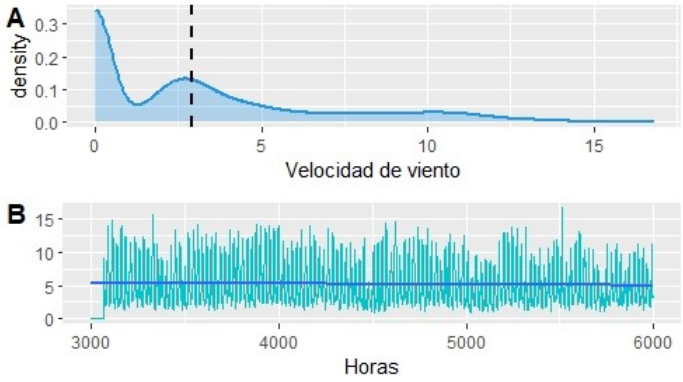
ANEXO D: ESTACIÓN METEREOLÓGICA CUMANDÁ 2014



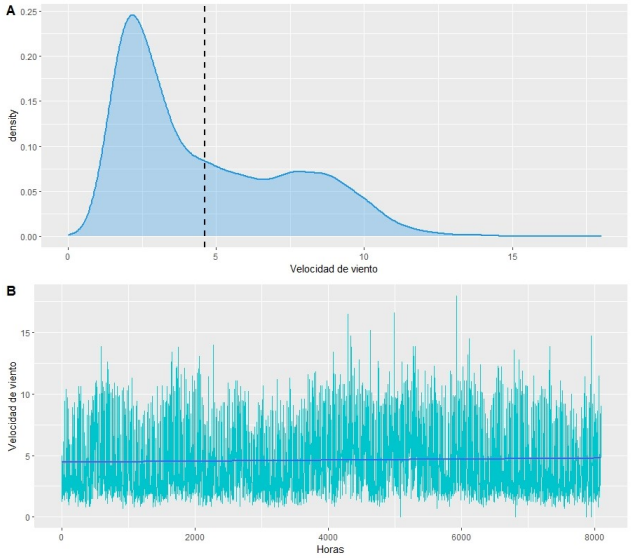
ANEXO E: ESTACIÓN METEREOLÓGICA ESPOCH 2020



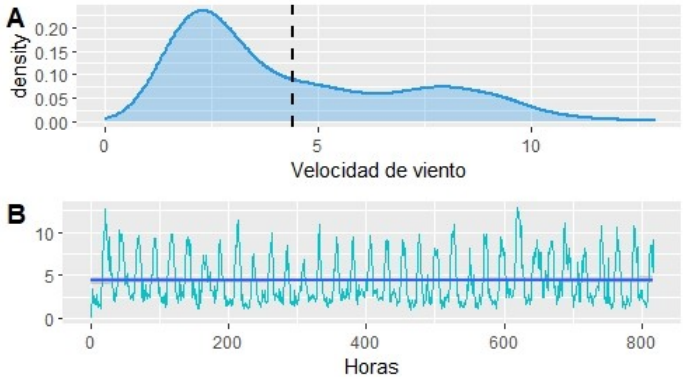
ANEXO F: ESTACIÓN METEREOLÓGICA ESPOCH 2014



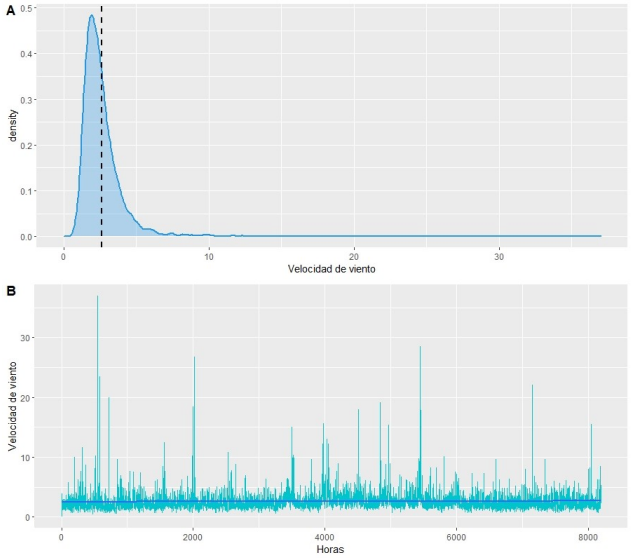
ANEXO G: ESTACIÓN METEREOLÓGICA MATUS 2018



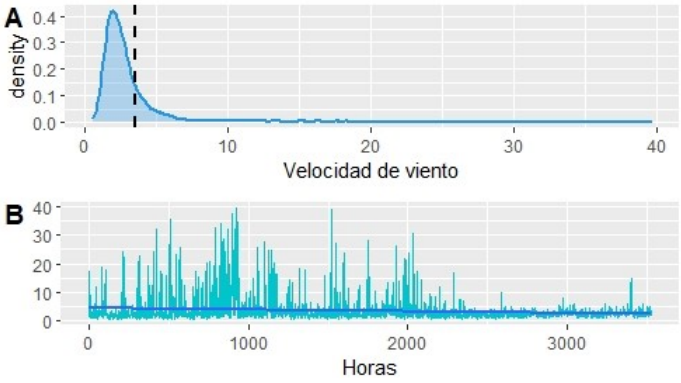
ANEXO H: ESTACIÓN METEREOLÓGICA MATUS 2014



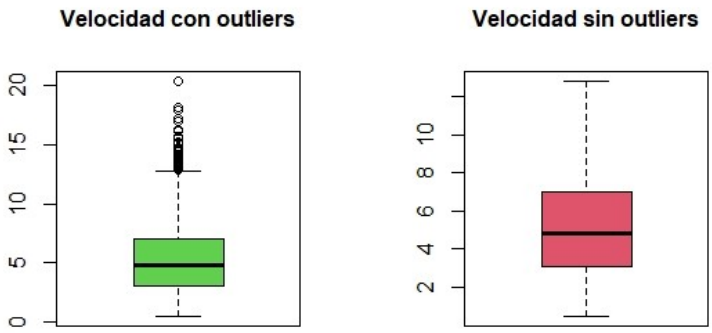
ANEXO I: ESTACIÓN METEREOLÓGICA MULTITUD 2017



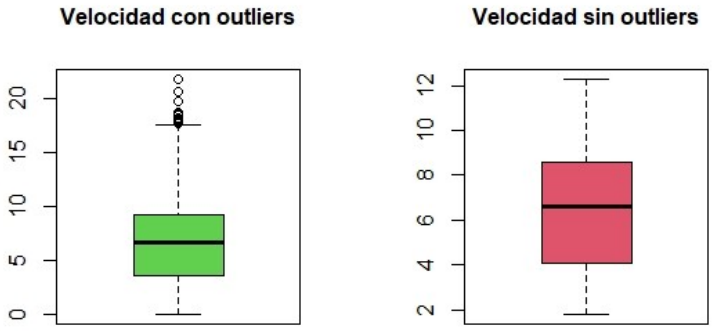
ANEXO J: ESTACIÓN METEREOLÓGICA MULTITUD 2017



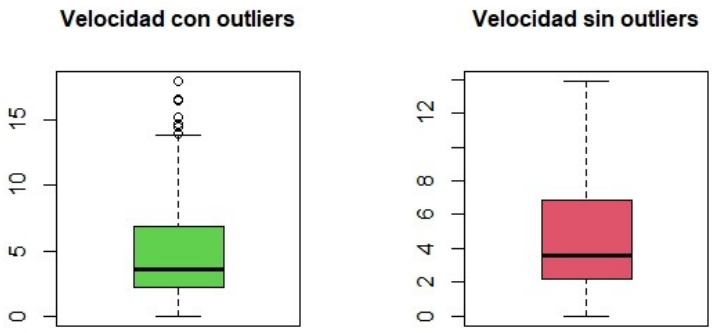
ANEXO K: BOX-PLOT ESTACIÓN METEREOLÓGICA ALAO



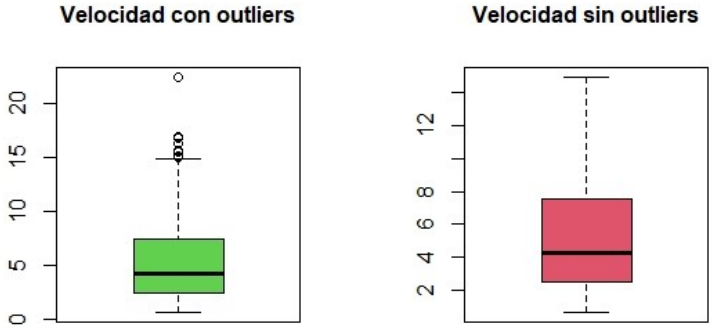
ANEXO L: BOX-PLOT ESTACIÓN METEREOLÓGICA ATILLO



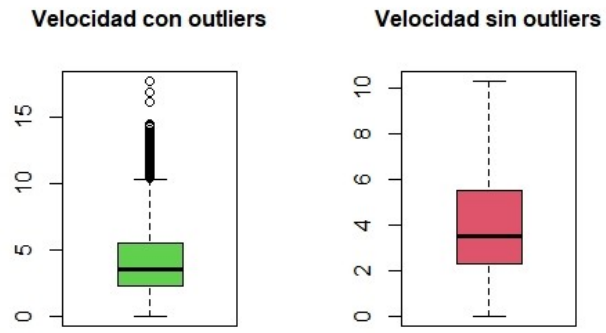
ANEXO M: BOX-PLOT ESTACIÓN METEREOLÓGICA MATUS



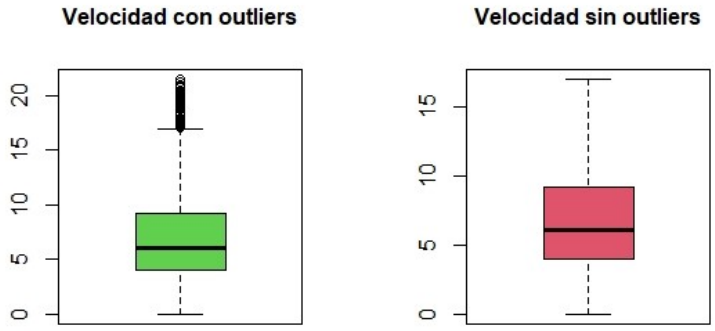
ANEXO N: BOX-PLOT ESTACIÓN METEREOLÓGICA QUIMIAG



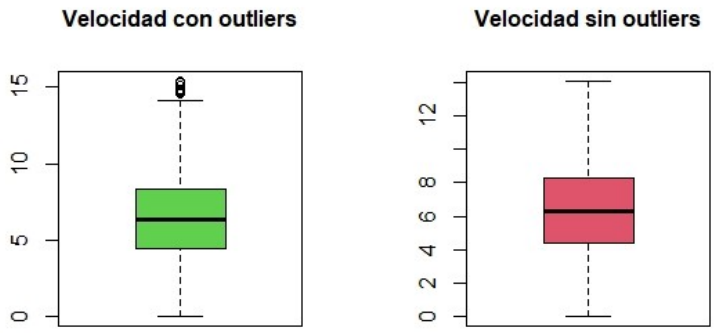
ANEXO Ñ: BOX-PLOT ESTACIÓN METEREOLÓGICA SAN JUAN



ANEXO O: BOX-PLOT ESTACIÓN METEREOLÓGICA TUNSHI



ANEXO P: BOX-PLOT ESTACIÓN METEREOLÓGICA URBINA



ANEXO Q: CÓDIGO R UTILIZADO PARA LA INVESTIGACIÓN

```
#####################################################
############## Analisis descriptivo ##############
############## Metodo de imputacion ##############
#####################################################
#####################################################

library(randomForest)
library(openxlsx)
library(readxl)
library(ggplot2)
library(reshape)
library(dplyr)
library(ggpubr)
library(RcmdrMisc)
library(tidyverse)
library(missForest)

#Of Monsters and Men - The Cabin Sessions
#GenWind_1h_SDI12.8
#GenWind_1h.8
dat<-read.csv("E:/TESIS VELASTEGUI/IMPUTACION/ANUAL/URBINA/
              URBINA-2021-L2-ANUAL/L2210101 - L2211015.csv"
              ,sep = ",")

```

```r
numSummary(df2$Velocidad, statistics=c("mean", "sd", "IQR", "quantiles"
                ,"cv", "skewness", "kurtosis"), type=c("2", "1", "3"),
                     quantiles=c(0, .25, .5, .75, 1))
summary(df2$Velocidad)
#boxplot(df2$Velocidad, col = "deepskyblue")
{
Velocidad<-dat$GenWind_1h_SDI12.8[2:length(dat$GenWind_1h_SDI12.8)]
Fecha<-dat$X[2:length(dat$GenWind_1h_SDI12.8)]

df1<-cbind(Fecha,Velocidad)
df2<-data.frame(Velocidad=as.numeric(Velocidad))
#write.csv(df1,"prue.csv")

p1<-ggplot(df2, aes(x=Velocidad))+geom_density(color=4,
lwd = 1,fill=4,alpha=0.30)+labs(x="Velocidad de viento")
+geom_vline(aes(xintercept=mean(Velocidad)),
                        linetype="dashed", size=1)

df <- data.frame(x = seq_along(df1[, 1]),
                 df1)

###p2<-ggplot(df,aes(x,as.numeric(Velocidad),color=Fecha))
###+geom_line(lwd = 1)+geom_point()+
###(y="Velocidad de viento",x="Horas")

p2<-ggplot(df,aes(x,as.numeric(Velocidad)))
+geom_line(color="turquoise3")
  +#geom_point(color="turquoise3")+
  labs(y="Velocidad de viento",x="Horas")+geom_smooth(method = lm)

ggarrange(p1,p2,ncol =1,nrow =2,labels = c("A","B"))

}



####IMPUTACION#######

#GenWind_1h_SDI12.8
```

```r
#GenWind_1h.8

Velocidad<-dat$GenWind_1h_SDI12.8[2:length(dat$GenWind_1h_SDI12.8)]
Fecha<-dat$X[2:length(dat$GenWind_1h_SDI12.8)]


Velocidad<-replace(Velocidad,Velocidad=="0.0","NA")

df2<-data.frame(Velocidad=as.numeric(Velocidad))

#Porcentaje de faltantes
NAS<- function(x)
{
  Velocidad<-replace(Velocidad,Velocidad=="0.0",NA)
  Fal<-sum(is.na(df2$Velocidad))
  Porcen<-round((Fal*100)/length(Velocidad),2)
  data.frame(Faltantes=Fal, Porcentaje=Porcen)
}
NAS(Velocidad)




p1<-ggplot(df2, aes(x=Velocidad))+geom_density(color=4,lwd = 1)
+labs(x="Velocidad de viento")


df <- data.frame(x = seq_along(df1[, 1]),
                 df1)


p2<-ggplot(df,aes(x,as.numeric(Velocidad),color=Fecha))
+geom_line(lwd = 1)+geom_point()+
  labs(y="Velocidad de viento",x="Horas")

ggarrange(p1,p2,ncol =2,nrow =1,labels = c("A","B"))

muestra
```

```r
prov


### Seleccion de t-student
outt <- vector("list", length(muestra))

for(i in 1:length(muestra)){

  if((levels(factor(dt$ESTACION,
                    levels = prov)))[i] == prov[i]){
    outt[[i]] <- dt[sample(which(dt$ESTACION
                                 == prov[i]), muestra[i]) ,]

  }
}

DATA_MUESTRAS <- do.call(rbind.data.frame, outt)
### Analisis de los errores
#######
dt_MCA <- MCA(DATA_MUESTRAS[, -c(5, 6)],
              method = 'Burt', graph = F)

##########
fviz_mca_var(dt_MCA, repel = T)

##########
fviz_mca_biplot(dt_MCA, repel = T)

#######
fviz_mca_ind(dt_MCA)

#######
factoextra::fviz_screeplot(dt_MCA, addlabels = T)

### Analisis
tabla_pro_c1 <- table(DATA_MUESTRAS[, c('Velocidad de viento',
                                        'HORA')])
COP_1 <- FactoMineR::CA(tabla_pro_c1, graph = F)
summary(COP_1)
```

```
142
143 factoextra::fviz_screeplot(COP_1, addlabels = T)
144 factoextra::fviz_ca_biplot(COP_1, repel = T)
145
146 #######ERRORES
147 tabla_pro_c2 <- table(DATA_MUESTRAS[, c('DIA', 'HORA')])
148 COP_2 <- FactoMineR::CA(tabla_pro_c2, graph = F)
149 summary(COP_2)
150
151 factoextra::fviz_screeplot(COP_2, addlabels = T)
152 factoextra::fviz_ca_biplot(COP_2, repel = T)
153
154 #######ERRORES
155 tabla_pro_c3 <- table(DATA_MUESTRAS[, c('DIA', 'HORA')])
156 COP_3 <- FactoMineR::CA(tabla_pro_c3, graph = F)
157 factoextra::fviz_screeplot(COP_3, addlabels = T)
158 table(DATA_MUESTRAS[, c('DIA', 'HORA')])
159 par(cex=1.5)
160 barplot(table(DATA_MUESTRAS[, c('DIA', 'HORA')]),
161         cex.names = 0.5, legend = rownames(table(DATA_MUESTRAS
162                     [, c('HORA', 'DIA')])), beside = F,
163                      col = c('lightblue',"pink"))
164
165 tabla_pro_c4 <- table(DATA_MUESTRAS[, c('DIA', 'HORA')])
166 COP_4 <- FactoMineR::CA(tabla_pro_c4, graph = F)
167 summary(COP_4)
168
169 factoextra::fviz_screeplot(COP_4, addlabels = T)
170 factoextra::fviz_ca_biplot(COP_4, repel = T)
171
172 # precision del RELLENO DE DATOS con
173 # los datos originales
174 sum(diag(con_table))/sum(con_table) * 100
175
176 roc_multi <- multiclass.roc(test$Y, pre_train)
177
178 ###IMPMIRIR LINEA anterior
```

espoch | Dirección de Bibliotecas y Recursos del Aprendizaje

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 09 / 01 / 2024

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: Evelyn Mishelle Velastegui Cujilema Erick Adrian Horna Zhinin
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Estadística
Título a optar: Ingeniero/a en Estadística Informática
f. Analista de Biblioteca responsable: Ing. Rafael Inty Salto Hidalgo

2243-DBRA-UPT-2023