UNIVERSIDAD CATÓLICA DE CUENCA

Comunidad Educativa al servicio del Pueblo

UNIDAD ACADÉMICA DE INGENIERÍA, INDUSTRIA Y CONSTRUCCIÓN

CARRERA DE INGENIERÍA AMBIENTAL

IMPUTACIÓN DE DATOS FALTANTES DE REGISTROS HIDROMETEOROLÓGICOS DE LAS CUENCAS DE LOS RÍOS JUBONES, CAÑAR Y ESMERALDAS MEDIANTE MÉTODOS ESTADÍSTICOS Y MACHINE LEARNIG.

TRABAJO DE INVESTIGACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO AMBIENTAL

AUTOR: PAÚL EDUARDO VÁSQUEZ ÁLVAREZ

DIRECTOR: ING. DIEGO AQUILES HERAS BENAVIDES MsC.

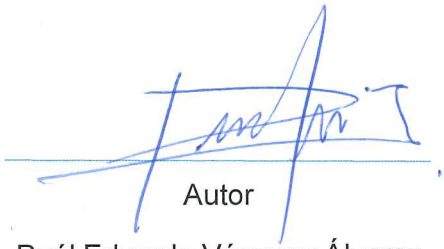
MATRIZ CUENCA

2019

DECLARATORIA

Yo, Paúl Eduardo Vásquez Álvarez, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento; y eximo expresamente a la Universidad Católica de Cuenca y a sus representantes legales de posibles reclamos o acciones legales.

La Universidad Católica de Cuenca puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y la normatividad institucional vigente.



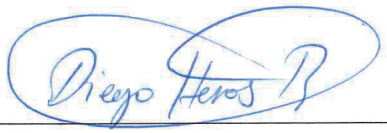
Autor

Paúl Eduardo Vásquez Álvarez

C.I. 0106558679

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Paúl Eduardo Vásquez Álvarez, bajo mi supervisión.



Ing. Diego Aquiles Heras Benavides MsC.

DIRECTOR

AGRADECIMIENTOS

A mis Padres; Eduardo y Mercedes, a mi Hermana Gabriela, por su incansable esfuerzo y paciencia, sus vidas han sido un ejemplo y fuente de enseñanzas; dedicación, honestidad y sacrificio. Todo aquello que este trabajo refleja.

A mi enamorada Od. Katherine Chuchuca y a su querida familia por todo el apoyo a hacia mi persona en este esfuerzo por desarrollar esta investigación.

A mi Director de tesis Ing. Diego Heras B. por guiarme a través del laberinto de la ciencia, y sin lugar a dudas aportar con su intelecto para que esta investigación se llevare a cabo.

A mi amigo Sr. Juan Marcelo Arpi, por ser un gran colaborador de esta investigación y brindarme su ayuda continua y desinteresada.

A todas aquellas personas que me brindaron su ayuda; lograron así iluminar un poquito más el extenso camino del saber.

DEDICATORIA

De manera emotiva a toda mi Familia, a ellos con profundo cariño y admiración porque nunca dejaron de confiar en mí.

A mis catedráticos que me dieron las bases para cumplir este desafío, que de la forma más cordial siempre atendieron a mis más profundas dudas de conocimiento.

"Caminante, no hay camino,
se hace camino al andar"

Antonio Machado

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE CÓDIGOS

RESUMEN

La presente investigación pretende identificar un método que permita imputar los valores faltantes de los registros hidrometeorológicos mediante el análisis de métodos de imputación clásicos y métodos computacionales basados en máquinas de aprendizaje automático denominado *Machine Learning.*

Para la elaboración de esta investigación se utilizaron los registros hidrometeorológicos de las estaciones de monitoreo ubicadas en las cuencas hidrográficas de los ríos Esmeraldas, Cañar y Jubones, en un periodo de 22 años, comprendidos entre 1990 y 2012. Las variables que se imputaron fueron la precipitación y el caudal.

En la investigación se utilizó métodos de regresión lineal simple y múltiple como métodos traiciónales de imputación, además, se empleó máquinas de aprendizaje automático del módulo *Scikit_Learn* de Python, estos módulos integran una amplia gama de algoritmos de aprendizaje automático como *Lineal regressor* y *Random Forest.*

Finalmente, se obtuvieron resultados óptimos que condujeron a un error cuadrático medio efectivo mínimo de 0.01 en los valores imputados, mediante el método de máquinas de aprendizaje automático *Random Forest.*

PALABRAS CLAVES: DATOS HIDROMETEOROLÓGICOS, IMPUTACIÓN, REGRESIÓN LINEAL, *MACHINE LEARNING*.

ABSTRACT

This research seeks to identify a method that allows to impute the missing values from the hydrometeorological records through the analysis of both classic imputation methods and computational ones based on automatic learning machines called Machine Learning.

In order to carry out the present research the hydrometeorological records from the monitoring stations located in the watersheds of the Esmeraldas, Cañar, and Jubones rivers were used over a period of 22 years, comprised between 1990 and 2012.

The imputed variables were precipitation and flow. The research used simple and multiple linear regression methods as betrayal methods of imputation, in addition, automatic learning machines from the Python Scikit_Learn module were used, these modules integrate a wide range of automatic learning algorithms such as Linear Regressor and Random Forest.

Finally, results led to a minimum effective mean square error, this being the imputation method of Random Forest automatic learning machines.

KEYWORDS: HYDROMETEOROLOGICAL DATA, IMPUTATION, LINEAR REGRESSION, MACHINE LEARNING.

INTRODUCCIÓN

Las series hidrometeorológicas pueden tener variables en evolución temporal como: caudales; precipitación; temperatura; evapotranspiración; entre otros. A menudo presentan datos faltantes o no disponibles, debido a diferentes causas que los provocan, que van desde defectos instrumentales, fallas técnicas o incluso errores humanos.

En la actualidad predecir el comportamiento de un fenómeno natural de cualquier índole, a partir de mólelos matemáticos requiere una base de datos fiable. En el campo de la hidrología estos modelos matemáticos requieren de registros hidrometeorológicos confiables, precisos, robustos y completos. Sin estos datos la predicción de comportamientos climáticos no será precisa, además incurrirá en errores potenciales que podrían no reflejar el verdadero comportamiento de un fenómeno.

Luego de un análisis de los datos meteorológicos e hidrológicos de los registros de series temporales del INAMHI se identificaron registros de datos con valores faltantes en sus registros, sin lugar a dudas la carencia de información completa impide un análisis climático preciso, predicciones acertadas sobre el comportamiento de las variables climáticas o un adecuado manejo del recurso hídrico (Campozano, Sánchez, Aviles, & Samaniego, 2014).

En las vertientes de las cuencas hidrográficas del pacifico del Ecuador, el número de estaciones de monitoreo hídrico son en cantidad menores respecto a la zona de la Sierra. Es primordial que la información observada sea en su mayoría completada, además, la integridad de los registros es un componente crucial para darle una correcta utilidad. Incluso brechas muy cortas impiden el cálculo de estadísticas importantes, como los totales mensuales de escorrentía o los flujos mínimos de (n) días, inhiben el análisis y la interpretación de la variabilidad del recurso.

De hecho, los estudios hidrológicos requieren datos completos de series de tiempo, recopilados en un período prolongado de varias estaciones, especialmente en grandes estudios de cuencas, distribuidos en toda la región de interés (Carrera Villacrés et al., 2016).

El objetivo principal de esta investigación es completar los registros de datos con la certeza de que son muy confiables y con la mejor estimación posible de las variables de precipitación mensual y del caudal promedio mensual en estaciones hidrometeorológicas de las Cuencas de los ríos, Jubones, Cañar y Esmeraldas en un periodo de tiempo de 22 años desde 1990 hasta el 2012.

Se propone un algoritmo para la optimización en la imputación de datos que minimice el error medio cuadrático de los valores completados. Para cumplir con este objetivo se evaluaron métodos de imputación determinísticos como regresiones lineales y métodos de inteligencia artificial con máquinas de aprendizaje automático, estos últimos proponen una solución moderna y efectiva de imputación.

CAPITULO I

1. GENERALIDADES

1.1. OBJETIVO

1.1.1. GENERAL

Imputar los datos faltantes en los registros hidrometeorológicos de las estaciones de las cuencas hidrográficas de los ríos Jubones, Cañar y Esmeraldas, mediante métodos estadísticos que permitan obtener datos confiables.

1.1.2. ESPECÍFICOS

- Realizar un estado del arte sobre métodos de imputación de datos aplicado a estaciones meteorológicas e hidrológicas para identificar la metodología estadística de estimación de datos ausente más adecuada para las variables a analizar.
- Realizar un análisis exploratorio y selectivo de registros de datos de los anuarios hidrometeorológicos del Instituto Nacional de Meteorología e Higrología (INAMHI), que correspondan a las estaciones de análisis.
- Implementar los diferentes modelos matemáticos con métodos estadísticos y de *machine learning*; realizar simulaciones para evaluar el comportamiento de los datos.
- Proponer una metodología de trabajo para la estimación de datos faltantes en los registros meteorológicos en los sectores de análisis.

1.2. JUSTIFICACIÓN

La presente investigación pretende ser una guía que plantea alternativas metodológicas en la imputación de datos faltantes en los registros de las zonas analizadas. Además, pretende contribuir con los registros hidrometereológicos de los de las cuencas hidrográficas de los ríos Cañar, Esmeraldas y Jubones, con la finalidad de proporcionar información válida y precisa para futuras investigaciones sobre el modelamiento en el aprovechamiento y gestión del recurso hídrico.

CAPITULO II

2. MARCO TEÓRICO

2.1. VARIABLES CLIMÁTICAS

El estudio de la atmosfera mediante la meteorología, implica enfocarse en variables como: la temperatura; precipitación; presión; humedad; caudal, evapotranspiración, entre otras, todas estas variables están relacionadas con el tiempo (Rodríguez Jimenez, Capa, & Portela Lozano, 2004).

2.1.1. PRECIPITACIÓN

La precipitación es de vital importancia dentro del área de la hidrología; es uno de los principales ingresos en un sistema o una cuenca hidrográfica (Bateman- et al., 2007). La variación en este comportamiento es importante para entender la recarga de los acuíferos, la escorrentía sobre la superficie terrestre y el caudal de los ríos (Urrutia, Palomino, & Salazar, 2010)

Se considera precipitación a toda manifestación de agua que impacta sobre la superficie terrestre, esta agua puede encontrarse en diferentes estados de la materia: solido o líquido.

La precipitación del agua se da a causa de la disminución de la temperatura sobre la humedad contenida en las masas de vapor de aire. De esta manera las diminutas partículas se agrupan hasta tener el tamaño y el peso adecuado para precipitar a la superficie terrestre (Rodríguez Jimenez et al., 2004)

2.1.2. CAUDAL

El caudal es el volumen de agua que pasa por un lugar en un determinado tiempo, el lugar puede ser una tubería, un canal o un río (Bello & Pino, 2000), generalmente las unidades que se representa al volumen respecto al tiempo viene expresadas en litros/segundos (l/s) ó metros cúbicos/horas (m3/h).

Existen diversos métodos para medir los caudales entre los más conocidos se encuentran:

- Método de trayectoria
- Método del flotador
- Método de estructuras de medidas

2.2. SISTEMAS DE OBSERVACIÓN CLIMÁTICA

Los datos climáticos se observan mediante diferentes equipos especializados para este propósito, además de acuerdo a las características de cada equipo se registran diferentes variables climáticas como: viento, humedad, temperatura, presión, altitud, precipitaciones, caudales, evapotranspiración, entre otros (Velázquez, 2014). De acuerdo con el tipo de datos climáticos observados y registrados, contamos con estaciones meteorológicas y estaciones hidrológicas.

Las estaciones registran datos climáticos que mantienen cierta normalización con la finalidad de estandarizar las observaciones. (OMM, 2011), es decir que las variables observadas pueden ser interprendas en cualquier parte del mundo.

2.2.1. ESTACIONES METEOROLÓGICAS

La estaciones meteorológicas convencionales se especializan en el registro de datos climáticos comunes siendo estas: el viento, precipitación, humedad, velocidad del viento, evapotranspiración, entre otras, este tipo de estaciones por lo general se encuentran en un lugar fijo donde sea indispensable la observación de las variables climáticas (Velázquez, 2014).

Para la observación de las variables climáticas las estaciones meteorológicas cuentan con varios instrumentos como son: altímetro, pluviómetro, vela, termómetro de mercurio entre otros más como se indica en la siguiente ilustración:

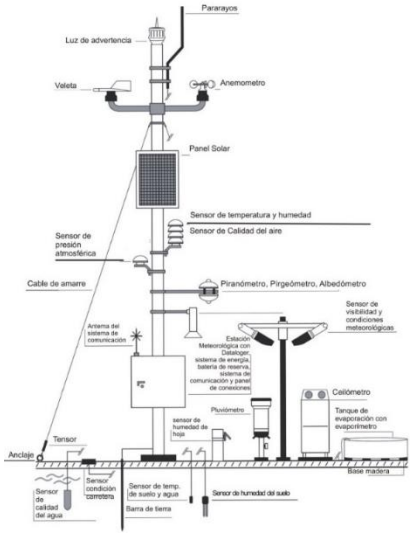


Ilustración 1: Esquema de una estación meteorológica completa

Fuente: (http://www.gisiberica.com, 2018).

2.2.2. ESTACIONES HIDROLÓGICAS

Las estaciones hidrológicas también conocidas como estaciones limnimétricas (ilustración 2) si cuentan con una regla para medir el nivel del agua o limnigráfica (ilustración 3) además de contener en su estructura un limnímetro o regla graduada, posee un limnígrafo, este último instrumento permite registrar de manera automática las variaciones en el nivel del agua de la sección analizada. Las estaciones hidrológicas se encuentran en un lugar específico en la sección de un río o aforo permitiendo registrar el nivel de agua que posee el río en un instante dado (Vera, 2012).

Mediante los registros de las estaciones hidrométricas se pueden calcular datos importantes como son el caudal que mantiene el río en el transcurso del tiempo.

Ilustración 2: Regla limnimétrica usada para determinar el nivel del agua

Fuente: (http://www.gisiberica.com, 2018).

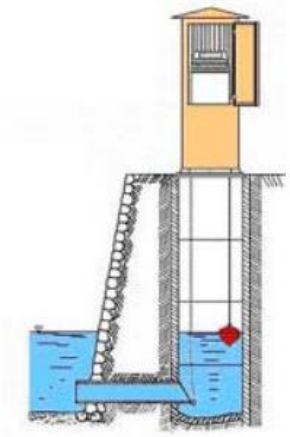


Ilustración 3: Estructura básica de una estación limnigráfica

Fuente: (Vera, 2012)

2.3. ANÁLISIS DE SERIES TEMPORALES

Los datos obtenidos de las observaciones recopiladas secuencialmente a lo largo del tiempo son extremadamente comunes. En meteorología se observa diariamente altas y bajas temperaturas, índices anuales de precipitación y sequía, velocidad del viento por hora, por citar algunas variables.

El propósito del análisis de series de tiempo es generalmente doble: comprender o modelar el mecanismo estocástico, es decir, se hace referencia a procesos, algoritmos y modelos que existen en una secuencia cambiante en el tiempo de eventos no deterministas analizables en términos de probabilidad, que da lugar a una serie observada en el tiempo para poder predecir valores o factores relacionados (D. Cryer & Kung-Sik, 2008).

Para la elección de modelos adecuados de estimación de datos en una serie de tiempo, dependen principalmente de tres pasos siendo los siguientes:

- Especificación del modelo (o identificación): Las clases de modelos de series de tiempo que serán seleccionados debe ser apropiadas para una serie observada dada, además, al elegir un modelo, intentaremos adherirnos al principio de parsimonia; es decir, el modelo utilizado debe requerir el menor número de parámetros que represente adecuadamente la serie temporal.
- Ajuste del modelo: La adaptación del modelo consiste en encontrar las mejores estimaciones posibles de esos parámetros desconocidos dentro de un modelo dado. Se considera criterios como los mínimos cuadrados y la máxima probabilidad de estimación.

- Diagnóstico del modelo: El diagnóstico del modelo se refiere a la evaluación de la calidad del modelo que ha especificado y estimado (D. Cryer & Kung-Sik, 2008).

2.4. DATOS AUSENTES.

Las observaciones de cualquier fenómeno pueden generar datos ausentes en sus registros. La imputación de esta información es el paso principal para efectuar cualquier análisis estadístico.

Para completar estos datos o llenar estos vacíos se requiere identificar qué tipo o mecanismo genera la ausencia de datos, además, comprender si la ausencia de datos es o no aleatoria (Goicoechea, 2002), estos mecanismos se pueden clasificar en 3 grupos.

- MCAR: *Missing completely at Random* se trata de datos perdidos completamente aleatorios, es decir, la ausencia de información no depende de otras variables de los datos observados.
- NMAR: *No missing at Random*, se refiere a datos perdidos cuando estos dependen de la misma variable observable, la ausencia de observación impide registrar su dato en la matriz.
- MAR: *Missing* at Random comprende la aleatoriedad en la ausencia de datos, la variable que no está presente en el registro depende de otras variables observables, es decir que la ausencia de datos está relacionada con otras variables dentro de la matriz de datos.

Comprender el mecanismo que produce que los datos estén ausentes es clave para poder imputarlos, es esencial reconocer si una o varias variables registradas afectan a otras variables dentro de la matriz de datos climatológicos.

2.5. IMPUTACIÓN DE DATOS.

En el análisis de eventos climáticos los datos son su principal herramienta para efectuar estudios estadísticos. Para ello se requiere información con suficientes observaciones de las variables a analizar.

En el afán de completar los vacíos de información en las variables climáticas se han diseñado técnicas o metodologías que puedan estimar el valor del dato ausente. La estimación de datos asume además un papel de control de calidad de la información,

permitiendo comparar con estaciones cercanas donde se registran datos en tiempo y espacio similares (OMM, 2011)

El objetivo principal de un análisis estadístico es la estimación de parámetros o variables de un modelo postulado (Walpole et al., 2012), estos modelos se basan en la probabilidad o recurrencia de eventos, generalmente estos modelos de predicción se caracterizan por ser suposiciones que se efectúan al momento de determinarlos.

La estadística infiere al momento de procesar información que pretende ser predicha, es decir estima un valor faltante o un resultado futuro mediante metodologías que ayudan a acercarse a ese valor.

La inferencia estadística en análisis de datos se puede catalogar dentro de dos grandes grupos: la estimación y pruebas de hipótesis (Walpole, Myers, Myers, & Keying, 2012). Para el caso de análisis de datos meteorológicos la estimación es la aproximación estadística más efectiva, mientras que las pruebas de hipótesis se pueden aplicar al momento de comparar dos o más metodologías de imputación de datos.

2.6. MÉTODOS ESTADÍSTICOS DE IMPUTACIÓN DE DATOS

Existen diferentes métodos de estimación de datos ausentes entre los más complejos de estas técnicas están los de regresión lineal y razón normal, regresión múltiple y krigeado, además de la implementación de redes neuronales que permiten el análisis de datos (OMM, 2011), estas técnicas generalmente permiten analizar los datos con relación a otros tipos de información complementaria que permita mejorar este análisis.

Actualmente la estimación por métodos estadísticos se la realiza mediante sistemas computacionales que permiten trabajar con grandes volúmenes de información en menores tiempos de procesamiento y con menores esfuerzos humanos (O Reilly Medina, 2013). De esta manera y con los softwares especializados para efectuar análisis estadísticos permiten que los diferentes métodos de estimación de datos sean procesados por las computadoras.

2.6.1. MÉTODOS DE ESTIMACIÓN CON ESTACIONES CERCANAS DE REFERENCIA

Existen diferentes métodos de estimación de datos ausentes y muy diversos dependiendo lógicamente de la cantidad de parámetros, similitudes topoclimáticas entre otras, en la siguiente tabla se exponen algunos de estos métodos.

Tabla 1. Métodos clásicos de imputación de datos faltantes

Método	Mediante	Abreviatura
Valor promedio de estaciones cercanas	Distancias	VPECd
Valor promedio de estaciones cercanas	Coeficiente de Person	VPECp
Valores vecinos cercanos	Distancias	VVCd
Valores vecinos cercanos	Coeficiente de Person	VVCp
Inverso de la distancia	Distancias	IDd
Regresión lineal con estaciones cercas	Distancias	RLECd
Regresión lineal con estaciones cercas	Coeficiente de Person	RLECp
Regresión lineal con estaciones cercas	Coeficiente de Person/Distancia	RLECp
Regresión lineal múltiple	Distancias	RLMECd

con estaciones cercas		
Regresión lineal múltiple con estaciones cercas	Coeficiente de Person	RLMECp

Valor promedio de estaciones cercanas mediante distancias:

Los valores de los datos ausente se obtienen mediante una ecuación aritmética en relación con los datos de las estaciones de referencia (Bennett, Newham, Croke, & Jakeman, 2007) como lo indica la ecuación (1):

$$\boldsymbol{V_{est} = \frac{\sum_{i=1}^{n} V_i}{n}} \quad (1)$$

Dónde:

$V_{est} = Valor\ a\ estimar\ del\ dato\ ausente$

$V_i = es\ el\ valor\ del\ mismo\ parametro\ que\ se\ estima\ pero\ de\ la\ estacion\ de\ referencia$

$n = es\ el\ numero\ de\ estaciones\ de\ referencia$

Valor promedio de estaciones cercanas mediante el coeficiente de Person:

El coeficiente de Person denominado comúnmente como R, permite identificar el grado de relación entre dos variables, para este caso los valores de R se transforman en pesos al ponderarlos, para estimar el valor faltante con n estaciones de referencia se suma los valores de las ponderaciones para su respectivo peso (Bennett et al., 2007), esto se refleja en las dos siguientes ecuaciones:

$$\boldsymbol{w_i = \frac{R_i}{\sum_{j=1}^{n} R_j}} \quad (2)$$

$$\boldsymbol{V_{est} = \sum_{j=1}^{n} V_j * w_i} \quad (3)$$

Valores vecinos cercanos mediante las distancias y el coeficiente de Person:

Este método usa estaciones cercanas o vecinas para estimar los datos ausentes mediante el análisis de las distancias geométricas entre las estaciones más cercanas,

además Campozano (2014) explica que este método tiene pobres resultados si las estaciones cercanas tienen una alta variabilidad espacial.

Regresión Lineal:

El análisis de datos mediante la regresión lineal es uno de los métodos más comunes empleadas para imputar datos ausentes (Herrera, Campos, & Carrillo, 2017), este método estadístico estima el valor ausente en el registro climático de la estación de estudio, basándose en estaciones de referencia que mantengan un registro de datos más completo. Además, se fundamenta en la ecuación lineal, esta es una herramienta estadística que usa datos de estaciones cercanas con el objetivo de mantener características topológicas similares a la estación de análisis en la estación de referencia (R. D. Medina, Montoya, & Jaramillo, 2008).

Como característica principal de este método de imputación de datos tenemos la incorporación de estaciones de referencia que nos permite tener una base de datos con caracterizas similares a estación de estudio, permitiendo de esta manera tener datos observados para predecir los datos ausentes.

Para la ejecución de este método se debe establecer datos de referencia, generalmente tienen que contener similitudes, es decir deben presentar características semejantes como zona de interés, elevación, temperaturas similares, generalmente características topoclimáticas (Antelo & Fernández Long, 2014), además la estación de referencia con el que se aplicará el método de regresión lineal tiene que estar completa (Herrera, Campos, & Carrillo, 2017)

El análisis de regresión lineal tiene además un completo para su ejecución denominado correlación lineal, este análisis permite identificar la relación que tienen los datos de la estación de análisis con la estación de referencia. El coeficiente de correlación se encuentra entre -1 y 1, al acercarse a los valores -1 o 1 el coeficiente de correlación lineal indica si los datos presentan una fuerte relación lineal definida por la ecuación (1), generalmente los datos hidrológicos entre estaciones cercanas presentan una correlación de 0.7 que se acepta para elegir la técnica de regresión lineal como adecuada para la predicción de datos ausentes (OMM, 2011).

$$\boldsymbol{r} = \frac{\boldsymbol{S_{xy}}}{\sqrt{\boldsymbol{s_x^2 s_y^2}}} \qquad (4)$$

Dónde:

X: datos de la estación de referencia.

Y: datos de la estación estudiada.

La recta de regresión lineal se representa de la siguiente manera:

$$\boldsymbol{y = b + mx} \quad (5)$$

Dónde:

$y = dato\ ausente\ en\ el\ registro$

$$\boldsymbol{m = \frac{S_{XY}}{S_x^2}} \quad (6)$$

$$\boldsymbol{b = \overline{y} - m\overline{x}} \quad (7)$$

Regresión lineal con estaciones cercanas mediante las distancias y el coeficiente de Person:

Este método selecciona la estación de referencia con menor distancia a la estación de análisis o la que posea el mayor coeficiente de correlación Person de las estaciones vecinas (Campozano et al., 2014). Finalmente se aplica la fórmula (8) de regresión lineal.

$$V_{est} = b * V_i + a \quad \textbf{(8)}$$

Dónde:

$\boldsymbol{V_{est}} = Valor\ a\ estimar\ del\ la\ estacion\ de\ analisis$

$\boldsymbol{V_i} = pendiente\ de\ la\ ecuacion$

$\boldsymbol{b}, a = Parametros\ de\ regresion\ lineal$

Razón normal:

La característica principal de este método estadístico para la imputación de datos ausentes es la utilización de la razón de los valores normales de las estaciones de referencia más cercanas a la del estudio.

La razón normal se recomienda usar cuando existen por lo menos tres estaciones cercanas de referencia (Carrera Villacrés et al., 2016), esto con la finalidad de normalizar los datos con los que se desean estimar.

La siguiente ecuación permite aplicar la razón normal a los registros de datos para la predicción de los mismos:

$$\boldsymbol{Px} = \frac{\mathbf{1}}{\boldsymbol{n}}\left[\left(\frac{\boldsymbol{N_x}}{\boldsymbol{N_1}}\right)\boldsymbol{P1} + \left(\frac{\boldsymbol{N_x}}{\boldsymbol{N_2}}\right)\boldsymbol{P2} + \cdots + \left(\frac{\boldsymbol{N_x}}{\boldsymbol{N_x}}\right)\boldsymbol{Px}\right] \quad (9)$$

Dónde:

Px = Precipitación de la estación con valores ausentes.

N= Número de estaciones con datos completos

P1 a Pn= Precipitación de las estaciones auxiliares para completar los datos.

Nx= Precipitación media de la estación de análisis.

N1 a Nx= Precipitación media de las estaciones de referencia.

Regresión múltiple:

Esta técnica estadística es una modificación de la ecuación de regresión lineal simple (Luna & Lavado, 2015) , la extensión de esta técnica estadística sobre su predecesora la regresión lineal simple se fundamenta en la incorporación de más de una estación de referencia para la estimación de datos ausentes en la estación de control. Se la puede definir con la siguiente ecuación:

$$\boldsymbol{Y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \cdots + \boldsymbol{\beta_i X_{i+}} \boldsymbol{e} \quad (10)$$

Dónde:

$Y_i = la\ i - enésima\ observación\ de\ la\ variable\ dependiente$

$\beta_1 \beta_2 \ldots \beta_i = parámetros\ de\ correlación\ lineal$

$X_1 X_2 \ldots X_i = valores\ conocidos\ de\ las\ variables\ dependientes$

$e_1 = valor\ de\ la\ perturbación\ aleatoria\ o\ residuo$

Regresión lineal múltiple con estaciones cercanas mediante las distancias y el coeficiente de Person.

El método de regresión lineal múltiple combina dos o más variables para predecir una respuesta teniendo en cuenta otras relaciones que la regresión lineal simple no contempla, varios factores intervienen en la ecuación de la regresión lineal múltiple, siendo estos: la distancia geométrica entre estaciones cercas, el coeficiente de Person o la relación entre ambos y el peso de k (Campozano et al., 2014).

$$V_{est} = \frac{\sum_{i=1}^{n} w_i^k * (b_i * V_i + a_i)}{\sum_{i=1}^{n} w_{mi}^k} \quad \textbf{(11)}$$

2.6.2. MACHINE LEARNING

El aprendizaje autónomo es una de las áreas de más rápido crecimiento de la informática, con aplicaciones de gran alcance.(Shalev-Shwartz & Ben-David, 2014). Este aprendizaje de maquina es una rama de la inteligencia artificial que basa en crear sistemas computacionales que puedan aprender automáticamente. Aprender para la maquina supone encontrar complejos patrones en un gran volumen de datos (J. Smola & S.V.N., 2008).

La máquina de aprendizaje basa su estructura en comprender un algoritmo que le permita revisar los datos, procesarlos e identificar patrones con los cuales podrá predecir comportamientos futuros, importante en este tipo de algoritmos computacionales de aprendizaje automático es que a medida que pasa el tiempo la máquina de aprendizaje mejora en el entendimiento de los patrones en la base de datos.

Existen en este contexto dos grandes ramas de aprendizaje automático, aprendizaje supervisado y aprendizaje no supervisado, estas se pueden definir de acuerdo a la manera en que el algoritmo de aprendizaje es entrenado (Shalev-Shwartz & Ben-David, 2014), es decir que la máquina de aprendizaje no supervisado no se le presenta la respuesta, con ello la maquina trabajara generalmente con datos con los cuales se les pueda clasificar, por otra parte para la máquina de aprendizaje supervisado es aquella a la cual se le presenta la respuesta y de esta forma buscara un mecanismo para satisfacer esta solución.

Generalmente las máquinas de aprendizaje se basan es modelos de clasificación y modelos de regresión, Ilustración *4*,esto dependerá del tipo de datos, así como también del tipo de solución a la que se quiera llegar, para ello existen múltiples máquinas de

aprendizaje con la se puede optar para encontrar una solución a un determinado problema (Shalev-Shwartz & Ben-David, 2014)

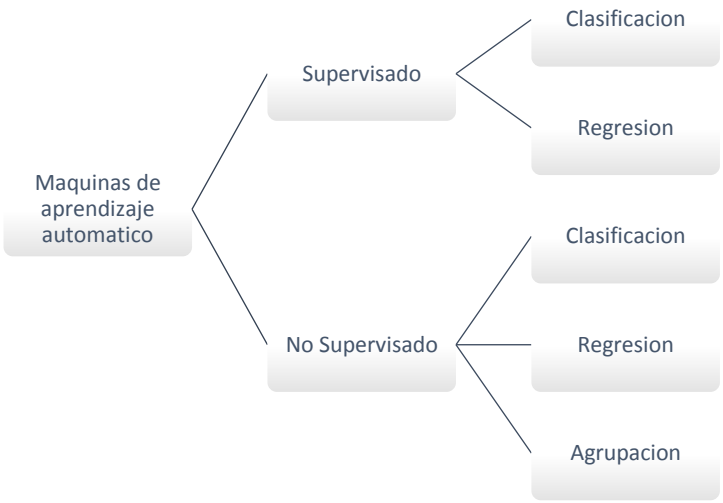


Ilustración 4: Esquema de simplificado de Machine Learning basado en modelos de clasificación y regresión.

Hay dos tipos de algoritmos de aprendizaje automático supervisado: Regresión y clasificación. El primero predice salidas de valor continuo mientras que el segundo predice salidas discretas. Este tipo de aprendizaje se acopla mejor a datos continuos como lo son datos hidrometeorológicos y en contexto general su algoritmo mantiene la siguiente estructura:

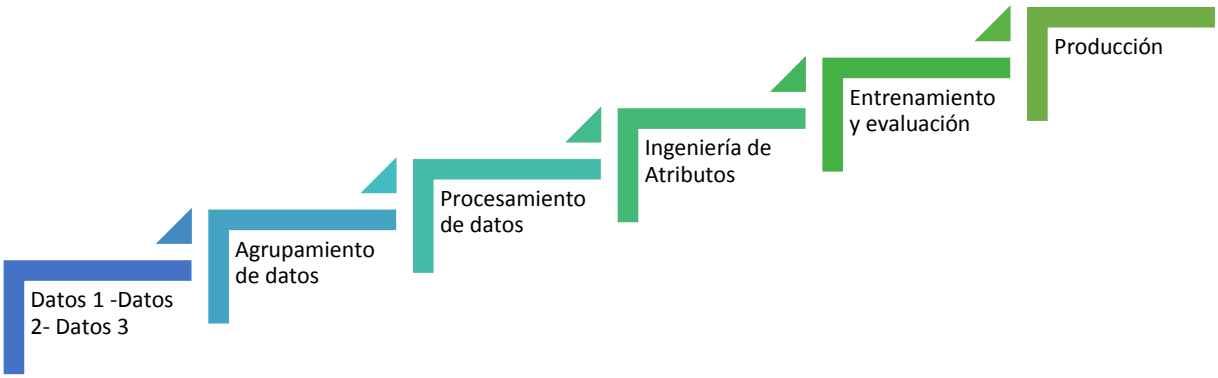


Ilustración 5: Estructura de algoritmo basado en máquinas de aprendizaje automático.

REGRESIÓN LINEAL:

Basado en regresiones lineales simples y múltiples, donde lo existen datos de entrenamiento y prueba con los cuales la maquina realiza múltiples regresiones con la finalidad de entrar una solución que satisfaga la respuesta (J. Smola & S.V.N., 2008).

Ampliamente usado en datos no categóricos donde existan dos o más variables de explicativas que actúen sobre la variable de respuesta.

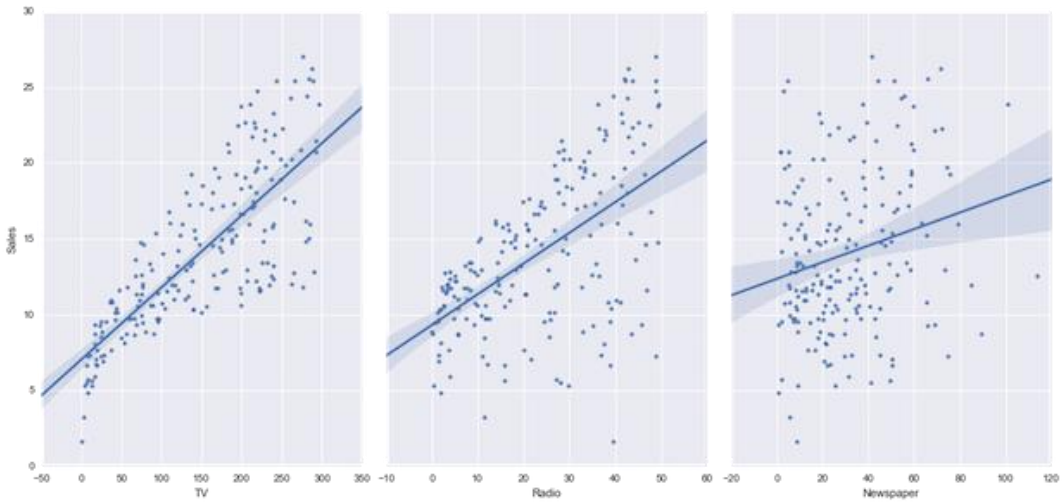


Ilustración 6: Ejemplo de regresión lineal basado en máquinas de aprendizaje automático, en busca del mejor modelo

RANDOM FOREST:

Random Forest o un bosque aleatorio es un meta estimador que se ajusta a una serie de árboles de decisión de clasificación en varias sub muestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y el ajuste excesivo del control (Shalev-Shwartz & Ben-David, 2014), es decir, Random Forest es un algoritmo de aprendizaje supervisado; crea un bosque y lo hace de manera aleatoria. El "bosque" que construye, es un conjunto de árboles de decisión, la mayoría de las veces entrenados con el método de "embolsado". La idea general del método de embolsado es que una combinación de modelos de aprendizaje aumenta el resultado general, ya que el tamaño de la submuestra es siempre el mismo que el tamaño de la muestra de entrada original.

Los bosques aleatorios son una combinación de factores predictivos de árboles, de modo que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles en el bosque (Breiman, 2001). El error de generalización para los bosques converge en cuanto a un límite a medida que aumenta la cantidad de árboles en el bosque.

CAPÍTULO III

3. MATERIALES Y MÉTODOS

3.1. TIPO DE ANÁLISIS

Esta investigación es de tipo analítico, abarca datos de estaciones hidrometeorológicas y se basa en el análisis estadístico para estimar datos ausentes de las variables de precipitación mensual acumulada y caudales medios mensuales en un período de tiempo de 22 años.

3.2. UBICACIÓN.

Las estaciones de análisis se encuentran ubicadas en tres cuencas hidrográficas principales de la zona costanera del Ecuador, siendo estas las siguientes cuencas: Río Cañar, Río Jubones y Río Esmeraldas.

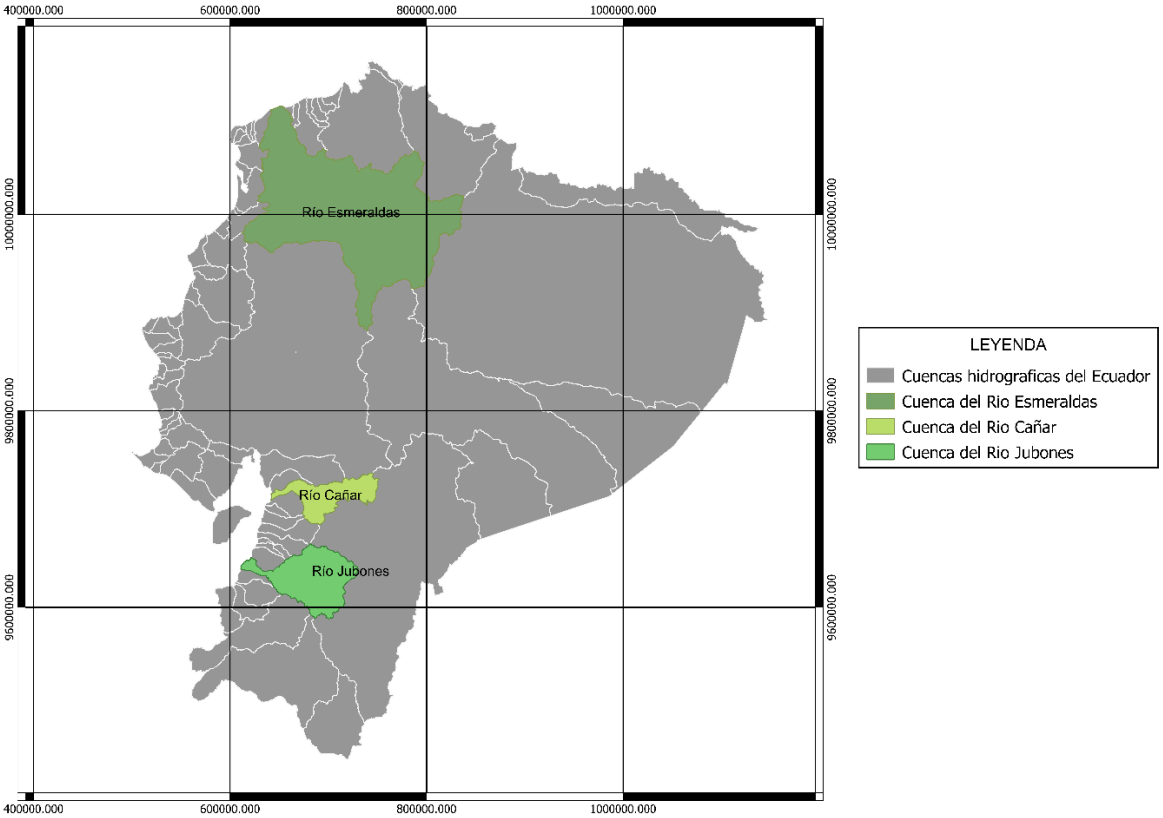


Ilustración 7: Ubicación de las cuencas hidrográficas de estudio

3.3. POBLACIÓN

Las cuencas hidrográficas de análisis tienen al alrededor de 318 estaciones meteorológicas y 106 estaciones hidrológicas, 334 estaciones monitorean la cuenca hidrográfica del río Esmeraldas, 55 estaciones la cuenca hidrográfica del río Jubones, 35 estaciones monitorean la cuenca del río Cañar.

Las estaciones meteorológicas encargadas esencialmente de recolectar datos de precipitación en las cuencas de análisis mantienen registros en las publicaciones oficiales de INAMHI desde 1970 hasta el 2016. Mientras que las estaciones de análisis hidrológicas mantienen registros desde 1990 hasta el 2012.

Tabla 2. Número total de estaciones hidrometeorológicas por cuenta de análisis

Distribución de estaciones hidrometeorológicas por cuenca hidrográfica		
Cuenca hidrográfica	# de estaciones meteorológicas	# de estaciones hidrológicas
Esmeraldas	257	77
Cañar	21	14
Jubones	40	15

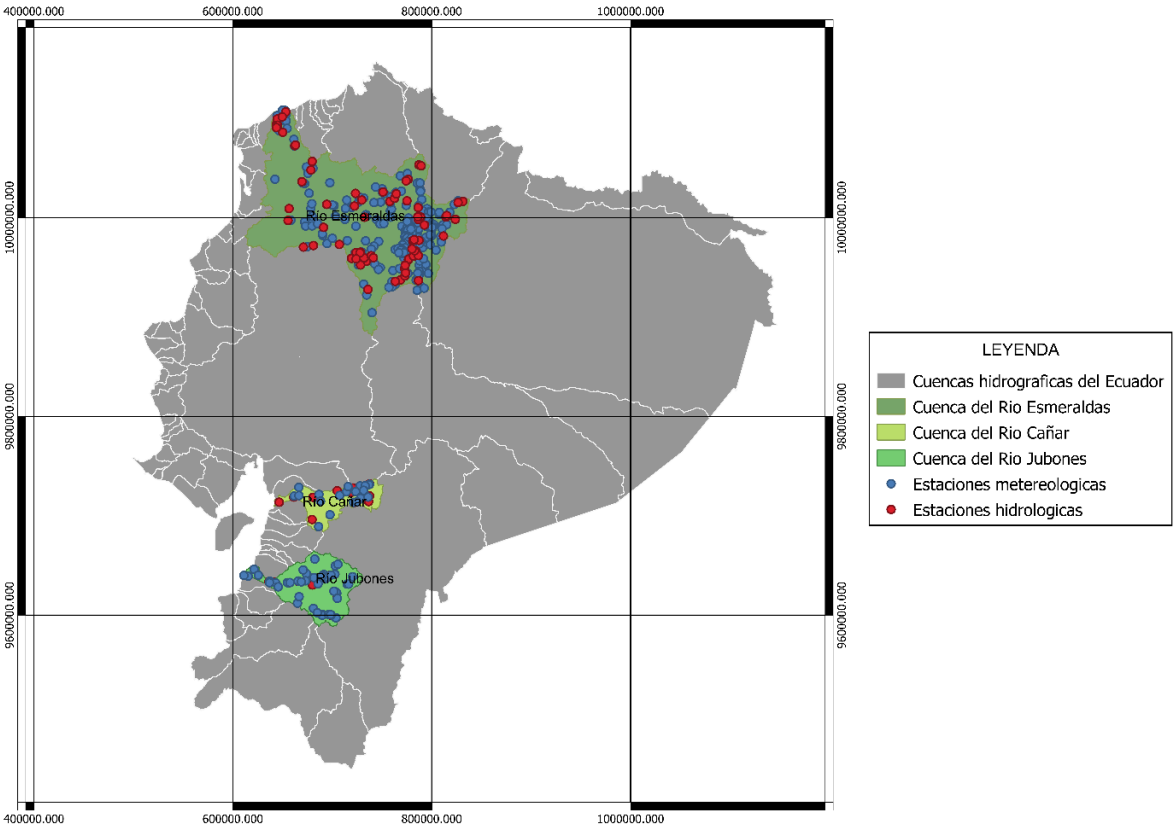


Ilustración 8: Ubicación de las estaciones hidrometeorológicas en las cuencas hidrográficas de estudio

3.4. MUESTRA

Las estaciones a analizar se obtuvieron mediante muestro aleatorio simple; se selecciono al azar por cada cuenca hidrográfica una estación meteorológica de un total de 318 estaciones y una estación hidrológica de un total de 106, adicionalmente, dos estaciones de referencia cercanas para el caso de las estaciones meteorológicas y una estación de referencia para el caso de las estaciones hidrológicas, estas estaciones cercanas de referencia fueron seleccionadas como predictoras en el momento de análisis de datos. En definitiva, se seleccionaron tres estaciones meteorológicas y dos estaciones hidrológicas para cada cuenca hidrográfica de análisis.

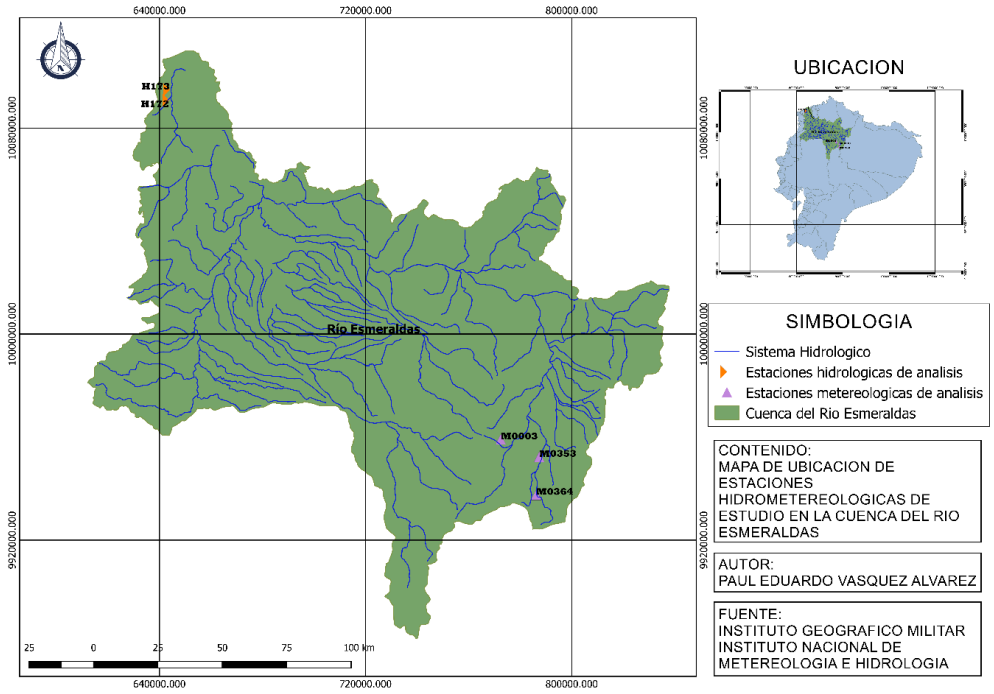


Ilustración 9: Ubicación de las estaciones hidrometeorológicas en la cuenca del río Esmeraldas

En la cuenca del río Esmeraldas las estaciones meteorológicas a estudiar son las estaciones M003, M353 y M364, a su vez las estaciones hidrológicas a analizar son las estaciones H172 y H173. Estos códigos de nombres están normalizados por el Banco Nacional de Datos Hidrometereológicos del INAMHI.

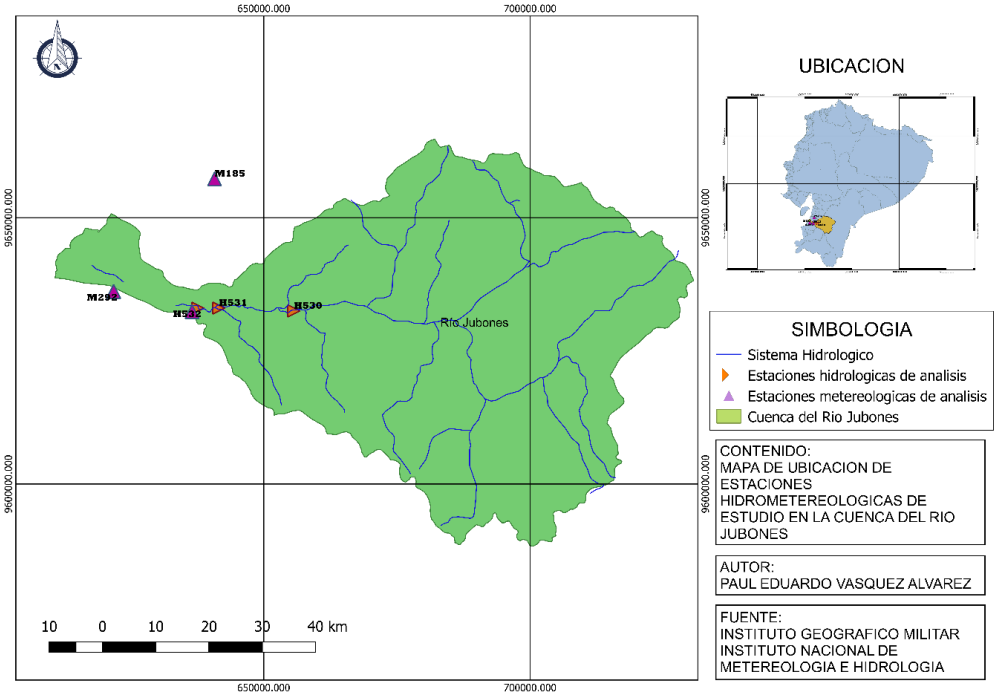


Ilustración 10: Ubicación de las estaciones hidrometeorológicas en la cuenca del río Jubones

En el caso de la cuenca del río Jubones se estudian las estaciones meteorológicas M185, M292 y M040, para las estaciones hidrológicas se analizan las estaciones H531 y H530,

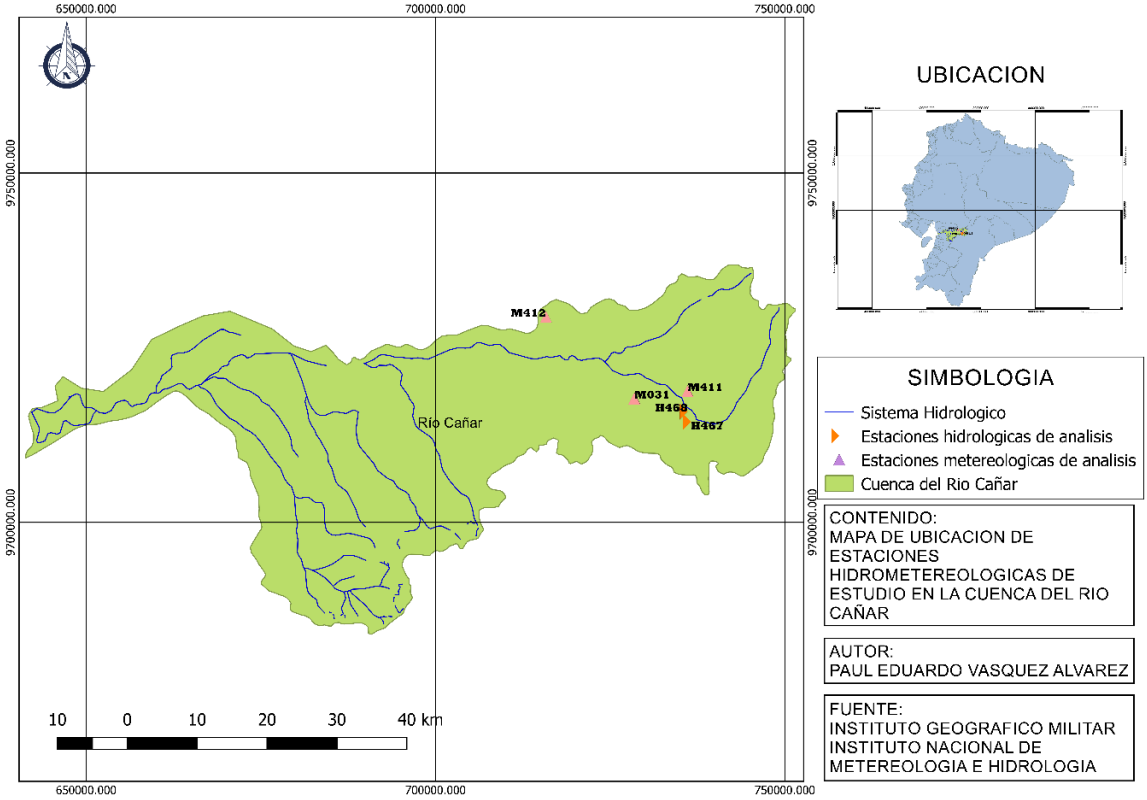


Ilustración 11: Ubicación de las estaciones hidrometeorológicas en la cuenca del río Cañar

Las estaciones meteorológicas en la cuenca del río Cañar son las estaciones M031, M411 y M412, mientras que las estaciones hidrológicas a analizar son las estaciones H467 y H468.

El período de análisis se seleccionó de acuerdo a la intercesión de temporalidad de las estaciones meteorológicas e hidrológicas, es decir un período de tiempo donde los registros de datos de las estaciones meteorológicas correspondan con los datos de las estaciones hidrológicas, de esta forma se estableció un período de 22 años comprendido entre 1990 y el 2012.

3.5. EQUIPOS Y MATERIALES

Para la elaboración de esta investigación se empleó un computador alta capacidad de procesamiento para el manejo de datos. El software científico utilizado Phyton es de libre acceso, con el cual se analizaron los datos hidrometereológicos en las estaciones de análisis.

Los materiales de análisis para esta investigación son los registros oficiales de los anuarios hidrometeorológicos de libre acceso que se obtuvieron de la página oficial del INAMHI.

3.6. METODOLOGÍA

Para la selección del método más adecuado para la imputación de datos ausentes en series temporales de precipitación y caudales se elaboró una recopilación de diferentes fuentes bibliográficas de autores reconocidos en el mundo de meteorología, esto como resultado del cumplimiento del primer objetivo específico, así tenemos las siguientes conclusiones de acuerdo con sus experiencias en el análisis de datos ausentes:

- La mejor metodología para el relleno o estimación de datos ausentes en series temporales en la región costanera del Ecuador es el método de regresión lineal simple debido a la gran cantidad de datos faltantes que provee el INAMHI (Carrera Villacrés et al., 2016).
- Campozano (2014) En su investigación para completar datos de estaciones meteorológicas en el caso de estudio de la Cuenca del río paute concluye:
  "Los resultados revelan que, para rellenar series temporales de precipitación diaria, el método de regresión lineal múltiple ponderada es el mejor, debido a la consideración de la razón entre el coeficiente de correlación de Pearson y la distancia con respecto a otras estaciones como factor de ponderación, dando mayor importancia a las estaciones más cercanas altamente correlacionadas".
- Herrera (2017) Efectúa un análisis de métodos completo, además un estado del arte sobre métodos de imputación de datos meteorológicos de diversos autores y finalmente decide realizar su investigación mediante regresión lineal justificando de la siguiente manera:
  "El método de regresión simple es superior entre los tradicionales para las variables temperatura mínima, máxima y precipitación en diferentes condiciones climáticas".

Finalmente se empleó una adaptación de la metodología usada por Patt (2012) en la que reúne características claves de lo antes expuesto y que sigue el siguiente esquema:

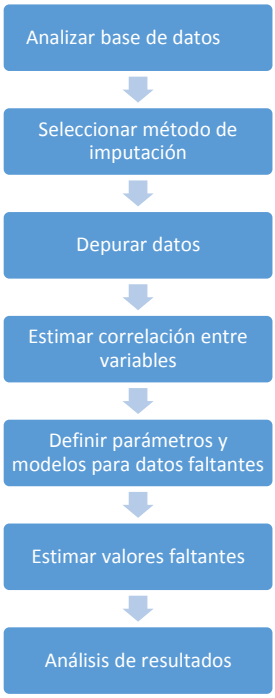


Ilustración 12: Diagrama de procesos para aplicar el método de regresión lineal simple con la finalidad de imputar datos ausentes en series temporales

Fuente: (Patt, 2012).

El método seleccionado se puede resumir de la siguiente manera:

1- Para el análisis de datos de las estaciones meteorológicas seleccionadas se debe evaluar el grado de datos ausentes en toda la temporalidad del registro.
2- Seleccionar el método de imputación permite evaluar las variables que disponemos del conjunto de datos y si estas son suficientes para garantizar una correcta imputación de acuerdo al método seleccionado.
3- Seleccionado el método estadístico en los objetivos anteriores permite depurar y estandarizar la data, para que las matrices tengan el mismo tamaño tanto la estación de análisis como la o las estaciones de referencia (Patt, 2012),
4- Se seleccionan la estación de referencia que tenga una mejor correlación con la de análisis es decir que tenga una correlación superior a 0.7 como lo recomienda (Urrutia et al., 2010).
5- Caso seguido realizar el análisis con el método seleccionado y aplicar pruebas de confiablidad para verificar la bondad y el ajuste de los datos imputados, siguiendo los lineamientos establecidos en la Guía de prácticas meteorológicas (OMM, 2011).
6- Analizar los resultados nos permite identificar la calidad de nuestros modelos de imputación, es indispensable para poder garantizar los datos imputados.

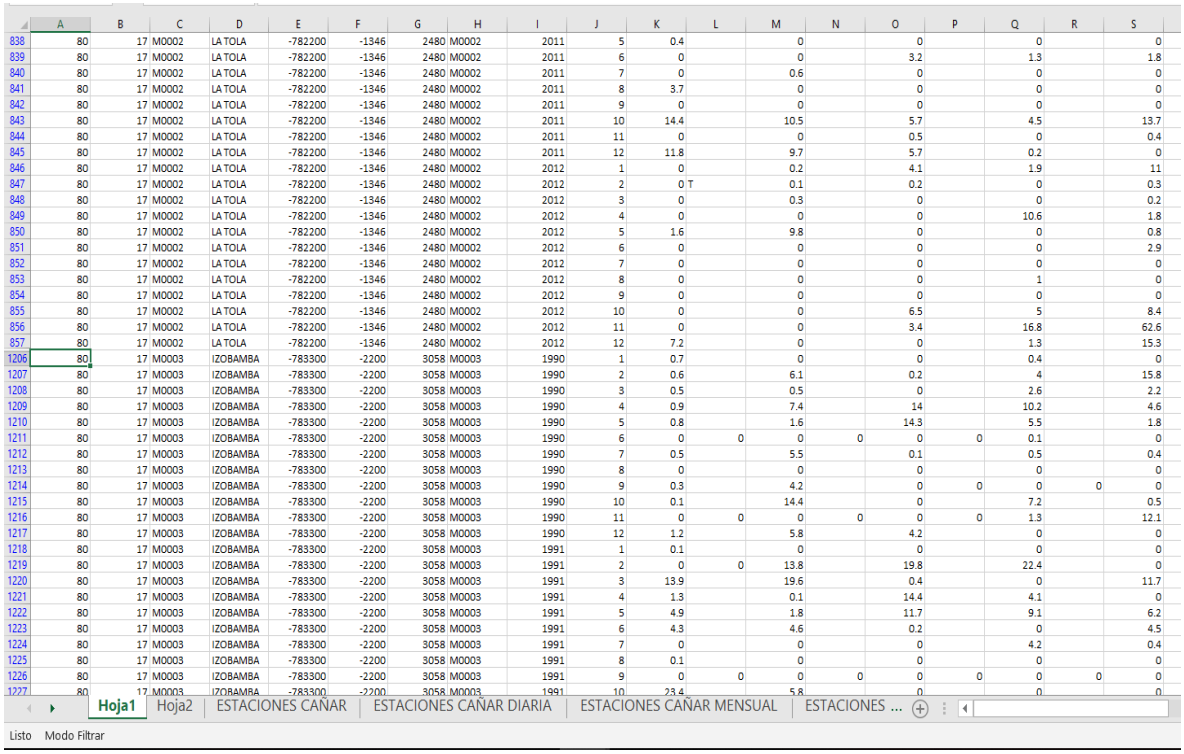
CAPÍTULO IV

4. RESULTADOS Y DISCUSIONES

4.1. PROCESAMIENTO DE INFORMACIÓN

Para el procesamiento de la información se empleó el Software Excel, en este proceso se seleccionó la información presentada en los registros del INAMHI donde el archivo de origen contenía la información de las estaciones por filas y era necesario modificar ese formato para el consiguiente procesamiento de datos en el Software estadístico y de programación, Minitab y Python sucesivamente. Dichos programas informáticos utilizan formatos (*csv)[1], donde la información se la tiene que presentar por columnas respecto a cada estación hidrometeorológica.

Además, se seleccionaron las estaciones por cuenca hidrográfica donde estas se encuentran, como se indica una muestra en la siguiente gráfica:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
838	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	5	0.4		0		0		0		0
839	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	6	0		0		3.2		1.3		1.8
840	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	7	0		0.6		0		0		0
841	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	8	3.7		0		0		0		0
842	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	9	0		0		0		0		0
843	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	10	14.4		10.5		5.7		4.5		13.7
844	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	11	0		0		0.5		0		0.4
845	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2011	12	11.8		9.7		5.7		0.2		0
846	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	1	0		0.2		4.1		1.9		11
847	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	2	0	T	0.1		0.2		0		0.3
848	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	3	0		0.3		0		0		0.2
849	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	4	0		0		0		10.6		1.8
850	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	5	1.6		9.8		0		0		0.8
851	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	6	0		0		0		0		2.9
852	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	7	0		0		0		0		0
853	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	8	0		0		0		1		0
854	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	9	0		0		0		0		0
855	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	10	0		0		6.5		5		8.4
856	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	11	0		0		3.4		16.8		62.6
857	80	17	M0002	LA TOLA	-782200	-1346	2480	M0002	2012	12	7.2		0		0		1.3		15.3
1206	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	1	0.7		0		0		0.4		0
1207	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	2	0.6		6.1		0.2		4		15.8
1208	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	3	0.5		0.5		0		2.6		2.2
1209	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	4	0.9		7.4		14		10.2		4.6
1210	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	5	0.8		1.6		14.3		5.5		1.8
1211	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	6	0	0	0	0	0	0	0.1		0
1212	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	7	0.5		5.5		0.1		0.5		0.4
1213	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	8	0		0		0		0		0
1214	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	9	0.3		4.2		0	0	0	0	0
1215	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	10	0.1		14.4		0		7.2		0.5
1216	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	11	0	0	0	0	0	0	1.3		12.1
1217	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1990	12	1.2		5.8		4.2		0		0
1218	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	1	0.1		0		0		0		0
1219	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	2	0	0	13.8		19.8		22.4		0
1220	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	3	13.9		19.6		0.4		0		11.7
1221	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	4	1.3		0.1		14.4		4.1		0
1222	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	5	4.9		1.8		11.7		9.1		6.2
1223	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	6	4.3		4.6		0.2		0		4.5
1224	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	7	0		0		0		4.2		0.4
1225	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	8	0.1		0		0		0		0
1226	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	9	0	0	0	0	0	0	0	0	0
1227	80	17	M0003	IZOBAMBA	-783300	-2200	3058	M0003	1991	10	23.4		5.8		0		0		0

Ilustración 13: Presentación de los registros hidrometeorológicos sin procesamiento

[1] Valores separados por comas, es un tipo de formato que permite mantener los datos en formas de tablas.

Este procesamiento se lo efectuó mediante la fórmula (12) y aplicando Macros dentro del Excel, todo esto con la finalidad de presentar la información de una forma que pueda ser asimilada por los programas con los cuales se analizara la información.

= DESREF (rango1,Residuo(FILA()-FILA($x#x), FILAS(rango1)), TRUNCAR ((FILA()-FILA(xx))/FILAS (rango1)),1,1) (12)

Esta fórmula (12) permite transponer la matriz, es decir, intercambiar filas por columnas manteniendo siempre sus dimensiones. Una vez procesada la información se procede a efectuar las estimaciones de las estaciones hidrometeorológicas.

	A	B	C	D	E	F	G
1	fecha	M0411	M0412	M031			
2	01-01-90	27	87.5	19.9			
3	01-02-90		254.3	48.6			
4	01-03-90		109.5	20.5			
5	01-04-90		148.3	84.4			
6	01-05-90		76	46.7			
7	01-06-90		38.4	16.1			
8	01-07-90		21	12.5			
9	01-08-90		8	14.6			
10	01-09-90		18.2	22.4			
11	01-10-90		62	71.4			
12	01-11-90		41.8	26.9			
13	01-12-90	56.6	61.8	28.3			
14	01-01-91	25.4	64.4	26.3			
15	01-02-91	37.5	162.7	26.2			
16	01-03-91	72.8	189.9	64.7			
17	01-04-91	47.2	108.4	60.8			
18	01-05-91	58.8	125.6	73.8			
19	01-06-91	29.4	19.3	28.7			
20	01-07-91	24.5	12.8	22.5			
21	01-08-91	23.7	31.1	20.3			
22	01-09-91	18	9	29.2			
23	01-10-91	44.5	44.9	32.4			
24	01-11-91	61.1	25.4	45.6			
25	01-12-91	26.8	96.7	34.2			
26	01-01-92	10.8	67	9.5			
27	01-02-92	47.5	152.5	38.6			
28	01-03-92	54	167.5	57			
29	01-04-92	69.1	217.5	60.8			

Hoja1 | Hoja2 | ESTACIONES CAÑAR | ESTACIONES CAÑAR DIARIA | ES

Ilustración 14: Procesamiento de los registros hidrometeorológicos por cuenca hidrográfica

4.2. IMPUTACIÓN DE DATOS FALTANTES EN ESTACIONES HIDROMETEOROLÓGICAS MEDIANTE REGRESIONES LINEALES ITERATIVAS

Para la imputación de los datos ausentes de las estaciones meteorológicas, se empleó el método de regresiones lineales múltiples consecutivas, este método tiene la particularidad de aproximarse a un mejor resultado luego de efectuarse regresiones lineales sucesivas, de esta forma se evalúan las regresiones lineales hasta que los coeficientes de la ecuación de la recta de regresión lineal se estabilizan, permitiendo

así el máximo ajuste de los datos analizados (Gómez García, Palarea Albaladejo, & Martín Fernández, 2006).

4.2.1. ANÁLISIS DE CORRELACIONES DE ESTACIONES METEOROLÓGICAS

En el caso del Río Jubones todas las estaciones de análisis de esta cuenca se pueden imputar con las estaciones cercanas de la misma cuenca, porque mantienen un coeficiente de determinación alto entre ellas como se puede apreciar en la siguiente tabla.

Tabla 3. Análisis de correlaciones entre las estaciones meteorológicas M0040, M0185 y M0292 del Río Jubones

	M040	M185
M185	0.826	
	0.000	
M292	0.850	0.911
	0.000	0.000

Contenido de la celda
Correlación de Pearson
Valor p

De acuerdo a R. Medina (2008) y Urrutia (2010) si los coeficientes de correlación son mayores a 0.75 los modelos cumplen con los siguientes supuestos: linealidad de modelo, varianza constante y sobre todo la normalidad de los datos e independencia. Esto permite que modelo estadístico sea confiable y que los datos imputados sean significativos.

Para el caso de la cuenca del Río Cañar el único modelo estadístico que sería válido de acuerdo a los coeficientes de correlación determinados en la Tabla 3. Son las estaciones meteorológicas M411 y M31.

De acuerdo a Medina (2008) no es recomendable la creación de modelos estadísticos con correlaciones menores a 0.75, esto ocurre en el caso la estación M412 debido a que no cumple con los supuestos estadísticos que validen el modelo.

Tabla 4. Análisis de correlaciones entre las estaciones meteorológicas M411, M412 y M031 del Río Cañar

	M411	M412
M412	0.604	
	0.000	
M031	0.853	0.551

	0.000	0.000

Contenido de la celda
Correlación de Pearson
Valor p

Para el caso de las estaciones meteorológicas de la cuenca del río Esmeraldas el único modelo estadístico que sería válido de acuerdo a los coeficientes de correlación y que cumplen con las condiciones, son entre las estaciones meteorológicas M364 y M003, por lo que la creación de modelos estadísticos para la estación meteorológica M353 no es válido debido que la correlación entre las demás estaciones no es fuerte.

Tabla 5. Análisis de correlaciones entre las estaciones meteorológicas M003, M353 y M364 del Esmeraldas

	M003	M353
M353	0.643	
	0.000	
M364	0.826	0.624
	0.000	0.000

Contenido de la celda
Correlación de Pearson
Valor p

En este estudio se analizaron solo estaciones que mantengan un coeficiente de correlación mayor a 0.75 entre las estaciones de análisis respecto a las estaciones de referencia, tanto para el caso de estaciones meteorológicas como para las hidrológicas.

4.2.1.1. ANÁLISIS DE VARIANZA

La varianza mide qué tan dispersos están los datos alrededor de su media. La varianza es igual a la desviación estándar elevada al cuadrado.

El análisis de varianza efectuado a la combinación de estaciones meteorológicas de las diferentes cuencas hidrográficas indica que el valor p es significativamente representativo[2], además la hipótesis nula, ecuación (13), es rechazada ya que los modelos estadísticos analizados mantienen un valor p menor a 0.05, es decir que al rechazar la hipótesis nula afirmamos que existe una relación entre las variables analizadas y que se puede construir un modelo matemático lineal.

[2] Implica rechazar la hipótesis nula H_0, la cual indicaría que el modelo sería una línea horizontal

$$H_0 = \beta_1 = 0 \quad (13)$$

$$H_1 = \beta_1 \neq 0 \quad (14)$$

β_1= *pendiente del modelo matemático lineal.*

(12) Hipótesis nula

(14) Hipótesis alternativa

Rechazar la hipótesis nula permite identificar que el coeficiente de la variable predictora será diferente de cero, con lo cual se asume que la recta de regresión lineal analizada no será una recta horizontal.

No rechazar la hipótesis nula, ecuación (13), significaría que el modelo lineal no sería capaz de explicar la variación en la respuesta en la variable predictora. Por lo general de manera estándar con respecto al error tipo I se declara por defecto un valor del nivel de significancia alfa de 0.05 que indica un riesgo del 5% de concluir que el modelo explica la variación en la respuesta cuando no es así.

De esta manera se efectúan los análisis de varianza en las estaciones meteorológicas de estudio de las 3 cuencas representadas en las siguientes tablas de datos.

Tabla 6. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Jubones, entre las estaciones meteorológicas M040, M185 y M292

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	2	1218533	609267	272.35	0.000
M185	1	35820	35820	16.01	0.000
M292	1	75940	75940	33.95	0.000
Error	198	442945	2237		
Total	200	1661479			

GL: es el número total de grados de libertad de los datos, en este análisis se utiliza para determinar el número de variables con las que se puede estimar los datos ausentes.

SC: La suma de los cuadrados (SC), que son las sumas ajustadas de los cuadrados, son medidas de la variación para los diferentes componentes del modelo.

MC: Los cuadramos medios ajustados permiten medir cuanta variación explica el termino o el modelo, considerando el grado de libertad.

Valor F: Este valor es estadístico de prueba que determina si el modelo tiene una asociación con la respuesta, además permite calcular p, que se usa para determinar la significancia del modelo

Valor p: El valor p es una probabilidad que mide la evidencia en contra de la hipótesis nula. Las probabilidades más bajas proporcionan una evidencia más fuerte en contra de la hipótesis nula. La hipótesis nula para la regresión general es que el modelo no explica ninguna variación en la respuesta.

Falta de ajuste: es una medida que indica que falta variables para poder ajustar adecuadamente el modelo.

Error puro: son las réplicas en los valores observados, es decir que existen múltiples observaciones con valores idénticos (Montgomery,Runger, 2016).

Tabla 7. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Jubones, entre las estaciones meteorológicas M185, M292 y M040

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	2	4428515	2214258	731.95	0.000
M040	1	64915	64915	21.46	0.000
M292	1	764912	764912	252.85	0.000
Error	238	719986	3025		
Total	240	5148501			

Tabla 8. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Jubones, entre las estaciones meteorológicas M292, M185 y M040

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	2	1524374	762187	786.89	0.000
M040	1	35175	35175	36.31	0.000
M185	1	247485	247485	255.51	0.000
Error	240	232466	969		
Total	242	1756839			

Tabla 9. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Cañar, entre las estaciones meteorológicas M411 y M031

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	223865	223865	569.35	0.000
M031	1	223865	223865	569.35	0.000
Error	214	84143	393		
Falta de ajuste	202	80710	400	1.40	0.265
Error puro	12	3433	286		
Total	215	308008			

Tabla 10. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Cañar, entre las estaciones meteorológicas M031 y M411

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	170814	170814	851.88	0.000

M411	1	170814	170814	851.88	0.000
Error	267	53537	201		
Falta de ajuste	245	46947	192	0.64	0.944
Error puro	22	6590	300		
Total	268	224351			

Tabla 11. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Esmeraldas, entre las estaciones meteorológicas M003 y M364

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	1112428	1112428	575.17	0.000
M364	1	1112428	1112428	575.17	0.000
Error	268	518337	1934		
Falta de ajuste	250	498284	1993	1.79	0.074
Error puro	18	20053	1114		
Total	269	1630766			

Tabla 12. Análisis de varianza de las estaciones meteorológicas de la cuenca hidrográfica del Río Esmeraldas, entre las estaciones meteorológicas M364 y M003

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	1142475	1142475	575.17	0.000
M003	1	1142475	1142475	575.17	0.000
Error	268	532338	1986		
Falta de ajuste	263	530346	2017	5.06	0.037
Error puro	5	1992	398		
Total	269	1674813			

En definitiva, al rechazar la hipótesis nula se asume que el modelo de regresión lineal no será una recta horizontal y que los valores predictivos que se obtengan del modelo son satisfactorios. Todas las estaciones analizadas cumplieron con el rechazo de la hipótesis nula, esto permite que todos los modelos matemáticos serán válidos.

Además, este análisis permite reconocer de forma numérica si los modelos de regresión serán satisfactorios o no, de esta manera podemos observar que los análisis de varianza para las estaciones seleccionadas luego de del análisis de correlación cumple con los criterios para que se pueda establecer adecuadamente modelos estadísticos que permitan en el consiguiente completar los registros meteorológicos con datos ausentes.

4.2.1.2. MODELOS DE IMPUTACIÓN MEDIANTE REGRESIONES ITERATIVAS PARA ESTACIONES METEOROLÓGICAS.

Los modelos de estimación son representados mediante las ecuaciones de regresión lineal, además, son la representación matemática que permite predecir los valores ausentes en los registros hidrometeorológicos.

En este estudio se aplica regresiones lineales con el método de mínimos cuadrados, y uso del algoritmo de regresiones iterativas. El método de mínimos cuadros permite ajustar la recta de regresión a una serie de datos presentados o preestablecidos, mientras que las regresiones lineales iterativas permiten obtener un mejor ajuste en la recta de regresión. Generalmente se la emplea el método de mínimos cuadrados en casos donde se requiera estudiar la variación de una magnitud X respecto a otra magnitud Y.

Todo este procesamiento de la información y la aplicación de los métodos señalados para la determinación de los modelos de regresión lineal se empleó el programa MiniTab.

Prosiguiendo con los análisis antes señalados tenemos para el caso de las estaciones meteorológicas de la cuenca del río Jubones para la estimación de datos en la estación M040 se efectuaron dos regresiones lineales en un proceso iterativo usando como referencia las estaciones M185 Y M292, para obtener como resultado la siguiente ecuación matemática con la cual se estimarán en lo posterior los valores ausentes dentro de su registro.

$$M040 = 24.75 + 0.2145\ M185 + 0.5375\ M292$$

Para estimar los datos de la estación meteorológica M185 de la cuenca del río Jubones se efectuaron 3 regresiones lineales múltiples mediante un proceso iterativo con las estaciones de referencia M040 y M292 con el objetivo de estabilizar los coeficientes de la regresión, de esta manera la ecuación para estimar los datos ausentes de la estación meteorológica M185 se representa en la siguiente manera.

$$M185 = 23.25 + 0.3488\ M040 + 1.2632\ M292$$

Para concluir con las estimaciones de las estaciones meteorológicas de la cuenca del río Jubones se realizaron 2 regresiones lineales múltiples consecutivas en un proceso iterativo a la estación M292 con sus estaciones de referencia cercanas M040 y M185, obteniendo como ecuación de predicción de datos la siguiente ecuación lineal.

$$M292 = -14.28 + 0.2496\ M040 + 0.4078\ M185$$

De esta manera se crearon tres ecuaciones lineales para la imputación de datos de las estaciones meteorológicas de la cuenca del río Jubones con las cuales se pudo completar la información faltante del registro de precipitaciones mensuales desde 1990 hasta el 2012.

Tabla 13. Resumen de los modelos lineales de las estaciones meteorológicas del rio Jubones

Modelos lineales para estimación de datos de precipitaciones de la cuenca hidrográfica del río Jubones

Estaciones meteorológicas	Modelo lineal
M040	24.75 + 0.2145 M185 + 0.5375 M292
M185	23.25 + 0.3488 M040 + 1.2632 M292
M292	-14.28 + 0.2496 M040 + 0.4078 M185

En el caso de análisis de la cuenca del río Cañar, para la imputación de datos de la estación meteorológica M0411 se utilizó como predictora la estación M0031 que presentaba una mayor correlación que las demás estaciones cercanas, a su vez, se efectuaron 2 regresiones lineales consecutivas donde los coeficientes de la ecuación se estabilizaron permitiendo en lo posterior una adecuada imputación de datos meteorológicos.

$$M0411 = 1.59 + 1.0939\ M031$$

La ecuación lineal que permite la imputación de los datos meteorológicos de la estación M0031 se generó usando la estación cercana de referencia M0411, para la obtención de la ecuación se realizó una regresión lineal puesto que los datos vacíos en esta estación de análisis son tan solo 7 valores.

$$M0031 = 10.52 + 0.6960\ M0411$$

A continuación, se resumen las ecuaciones para la estimación de las estaciones meteorológicas de análisis de la cuenca del río Cañar que permiten la imputación de los datos ausentes, con las cuales luego se completaran las series meteorológicas.

Tabla 14. Resumen de los modelos lineales de las estaciones meteorológicas del rio Cañar

Modelos lineales para estimación de datos de precipitaciones de la cuenca hidrográfica del río Cañar	
Estaciones meteorológicas	**Modelo lineal**
M411	1.59 + 1.0939 M31
M031	10.52 + 0.6960 M411

Para la cuenca del río Esmeradas la obtención de la ecuación de estimación de datos de la estación meteorológica M003, se utilizó la estación de referencia cercana M364, con la cual mantiene una correlación muy fuerte de 0.83, de la misma manera se realizó dos regresiones lineales consecutivas con lo cual se obtuvo la siguiente ecuación que permite predecir los datos ausentes en su registro meteorológico.

$$M003 = 20.91 + 0.8150\ M364$$

De la misma forma se obtuvo la ecuación lineal que permite imputar los datos ausentes en la estación M364 ubicada en la misma cuenca hidrográfica, usando como estación de referencia para todo el período de tiempo analizado la estación meteorológica cercana M003 y realizando dos regresiones lineal consecutivas.

$$M364 = 22.68 + 0.8370\ M003$$

Tabla 15. Resumen de los modelos lineales de las estaciones meteorológicas del rio Esmeraldas

Modelos lineales para estimación de datos de precipitaciones de la cuenca hidrográfica del río Esmeraldas	
Estaciones meteorológicas	**Modelo lineal**
M003	20.91 + 0.8150 M364
M364	22.68 + 0.8370 M003

La aplicación de estas ecuaciones permitirá en lo posterior completar el registro meteorológico para las estaciones de estudio. Además, se tiene un nivel de confianza del 95 % sobre todos los modelos matemáticos realizados en todas las estaciones meteorológicas donde se realizaron las imputaciones.

A continuación, se presenta un resumen de todos los modelos que permiten predecir los datos ausentes en los registros de precipitación desde 1990 hasta el 2012 de las cuencas hidrográficas de los ríos Jubones, Cañar y Esmeraldas.

Tabla 16. Resumen de los modelos lineales de las estaciones meteorológicas para las estaciones de análisis

RESUMEN DE MODELOS LINEALES DE ESTIMACIÓN DE VALORES AUSENTES	
Estaciones meteorológicas de la cuenca del río Esmeraldas	**Modelo lineal**
M003	20.91 + 0.8150 M364
M364	22.68 + 0.8370 M003
Estaciones meteorológicas de la cuenca del río Cañar	**Modelo lineal**
M411	1.59 + 1.0939 M31
M031	10.52 + 0.6960 M411
Estaciones meteorológicas de la cuenca del río Jubones	**Modelo lineal**
M040	24.75 + 0.2145 M185 + 0.5375 M292
M185	23.25 + 0.3488 M040 + 1.2632 M292
M292	-14.28 + 0.2496 M040 + 0.4078 M185

4.2.1.3. ANÁLISIS DE SUPUESTOS

Los modelos para la imputación de datos tienen que ser validados mediante los supuestos estadísticos, es decir que los modelos mediante regresión lineal tienen que cumplir algunos criterios para que el mismo no represente incorrectamente los datos.

Para que el modelo de regresión sea validado tiene que cumplir con los siguientes supuestos:

- Linealidad: se dice que si los datos no guardan una relación lineal se tiene un error de especificación, este supuesto se analizó al principio con el análisis de correlación y los diagramas de dispersión, este es uno de los supuestos más importantes para garantizar el modelo.
- Independencia: de la variable aleatoria, este análisis es importante si los datos tienen una secuencia temporal.
- Normalidad: sobre los residuos, es decir si los datos siguen en patrón de una distribución normal.
- Homocedasticidad: El supuesto de homocedasticidad implica que la variación de los residuos sea uniforme en todo el rango de valores de los pronósticos.

Procesando con el Software estadístico Minitab, se debe tener en cuenta la siguiente información para validar los supuestos en este programa informático.

Tabla 17. Características de un modelo de regresión adecuado, interpretación de graficas de resultados

Características de un modelo de regresión adecuado	Verificar usando	Soluciones posibles
La forma funcional modela adecuadamente cualquier curvatura que esté presente.	Prueba de falta de ajuste Gráfica de residuos vs. variables	Agregar término de orden superior al modelo Transformar las variables Regresión no lineal
Los residuos tienen una varianza constante.	Gráfica de residuos vs. ajustes	Transformar las variables Mínimos cuadrados

		ponderados
Los residuos son independientes (no están correlacionados) entre sí.	Estadístico de Durbin-Watson Gráfica de residuos versus orden	Agregar nuevo predictor Usar análisis de series de tiempo Agregar una variable de desfase
Los residuos están distribuidos normalmente.	Histograma de residuos Gráfica normal de residuos Gráfica de residuos vs. ajuste Prueba de normalidad	Transformar las variables Verificar si hay valores atípicos
Sin observaciones poco comunes ni valores atípicos.	Gráficas de residuos Apalancamientos Distancia de Cook DFITS	Transformar las variables Eliminar la observación atípica
Los datos no están mal condicionados.	Factor de inflación de la varianza (FIV) Matriz de correlación de los predictores	Eliminar el predictor Regresión de mínimos cuadrados parciales Transformar las variables

De esta manera tenemos las siguientes graficas donde se comprobará los supuestos para la validación de modelos antes generados para las estaciones meteorológicas de las cuencas hidrográficas de los ríos Esmeraldas, Cañar y Jubones.

Guía para identificación de los supuestos en los modelos:

1- La grafica de porcentaje vs residuo nos permite analizar los supuestos de normalidad para los modelos de las estaciones hidrometeorológicas, si la mayoría de los datos se encuentran sobre la recta o muy cerca, es decir que el modelo cumple con el supuesto de normalidad, teniendo en cuenta que puedan existir a su vez también valores en los extremos que se alejan de la línea estos se denominan valores atípicos. Esta grafica se encuentra ubicada en la parte superior izquierda de las ilustraciones.
2- La gráfica de residuos vs. ajustes permite verificar el supuesto donde los residuos están distribuidos aleatoriamente y tienen una varianza constante. Es decir, se identifica la homocedasticidad de las variables, los datos deben distribuirse aleatoriamente alrededor de la línea recta y una variabilidad aleatoria uniforme para que se cumpla el supuesto de homocedasticidad. La gráfica esta ubicada en la parte superior derecha de las ilustraciones del análisis de los supuestos.
3- La independencia del modelo se observa en la gráfica de residuos vs orden, se debería cumplir que los residuos independientes no muestran tendencias ni patrones cuando se muestran en orden cronológico. Los patrones en los puntos podrían indicar que los residuos que están cercanos entre sí podrían estar correlacionados y, por lo tanto, podrían no ser independientes. Lo ideal es que los residuos que se muestran en la gráfica se ubiquen aleatoriamente alrededor de la línea central. Gráfica ubicada en la parte inferior derecha de las ilustraciones.

Graficas de análisis de supuestos para las estaciones meteorológicas M031 y M0411 de la cuenca hidrográfica del río Cañar:

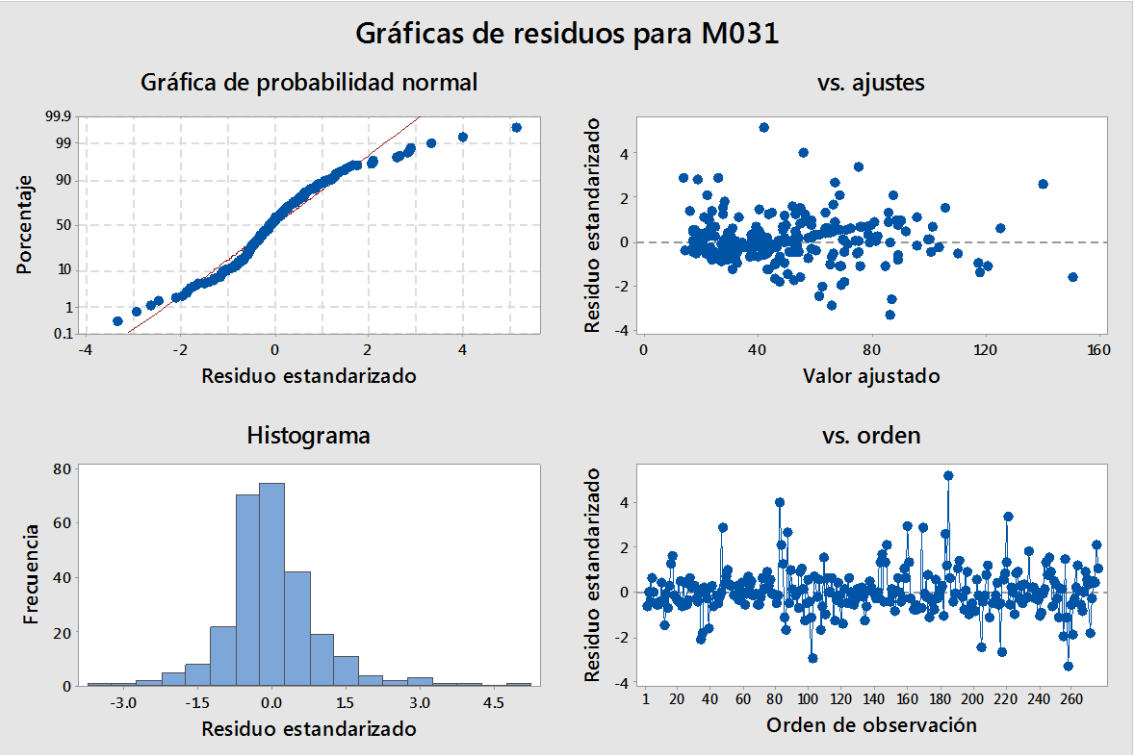


Ilustración 15: Análisis de supuestos de la estación meteorológica M031

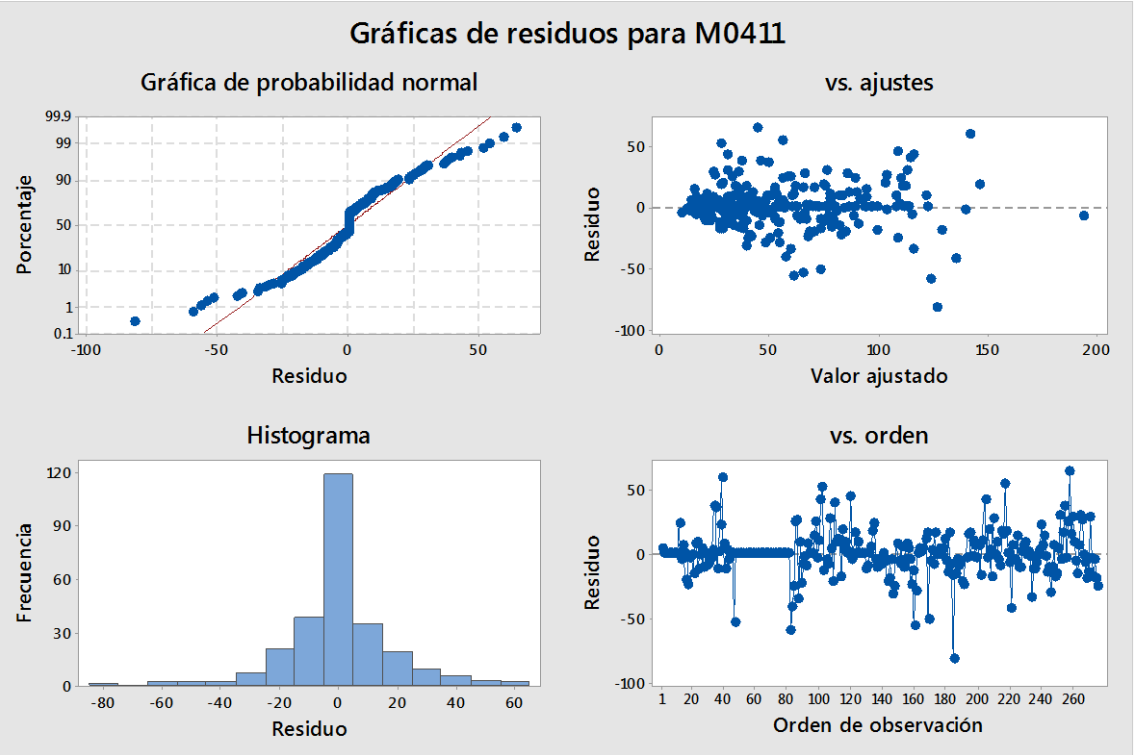


Ilustración 16: Análisis de supuestos de la estación meteorológica M0411

Las gráficas para las estaciones meteorológicas del M031 y M0411 presentan características similares aun que se puede observar una mejor adaptación de los

datos en la gráfica de la normalidad de la estación meteorológica M031 siguen una mejor distribución normal en relación al modelo de su estación vecina M0411.

El supuesto de homocedasticidad en las gráficas de errores vs ajuste no es completamente aleatorio pero se puede utilizar los modelos sin alteraciones graves

Respecto a los supuestos de linealidad e independencia los modelos para ambas estaciones cumplen sin ninguna observación importante.

Graficas de análisis de supuestos para las estaciones meteorológicas M003 y M0364 de la cuenca hidrográfica del río Esmeraldas:

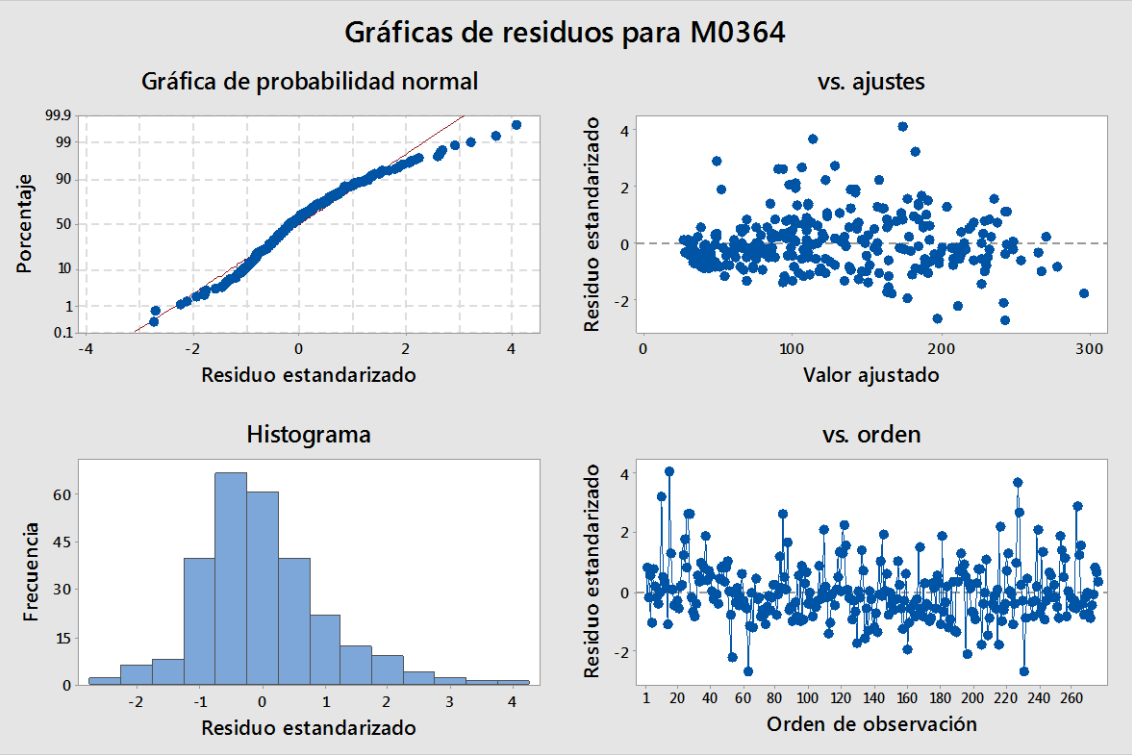


Ilustración 17: Análisis de supuestos de la estación meteorológica M0364

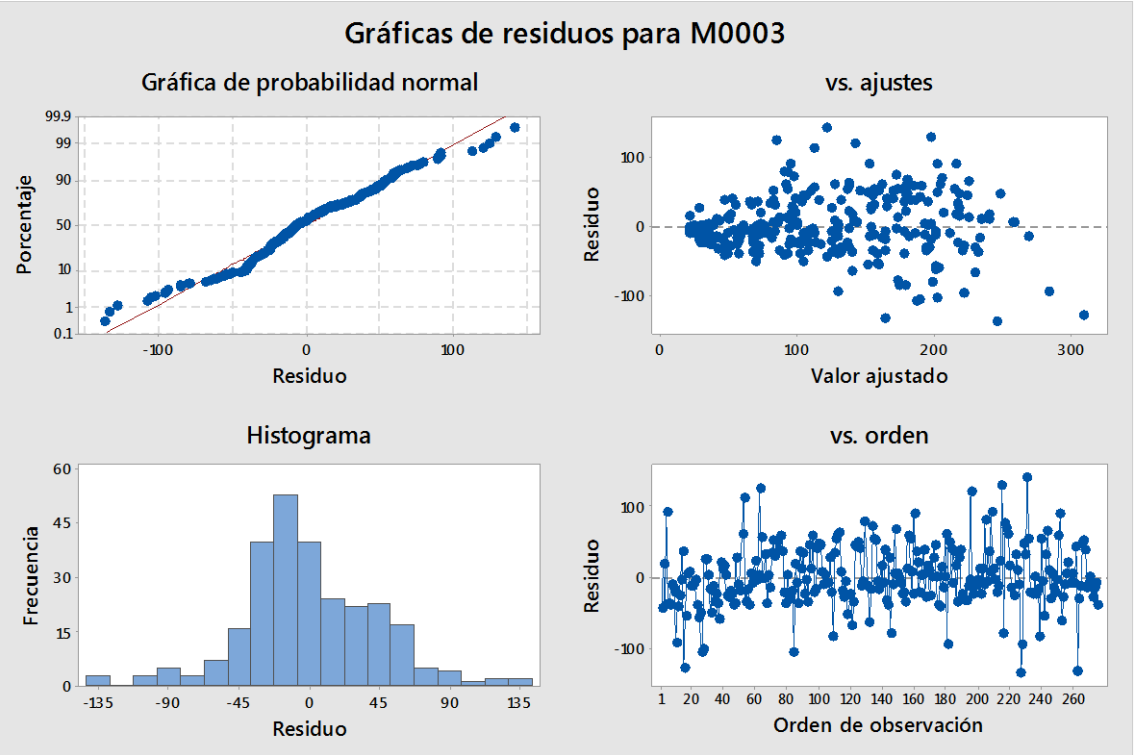


Ilustración 18: Análisis de supuestos de la estación meteorológica M0003

Las gráficas para las estaciones meteorológicas del M003 y M0364 presentan una distribución normal en sus graficas de porcentaje vs residuo se puede apreciar esta característica.

La homocedasticidad de estos modelos presenta una distribución aleatoria lo que nos permite concluir con que cumplen con este supuesto y el modelo es adecuado.

Respecto a los supuestos de linealidad e independencia los modelos para ambas estaciones cumplen del todo.

Graficas de análisis de supuestos para las estaciones meteorológicas M031, M0185 y M0292 de la cuenca hidrográfica del río Esmeraldas:

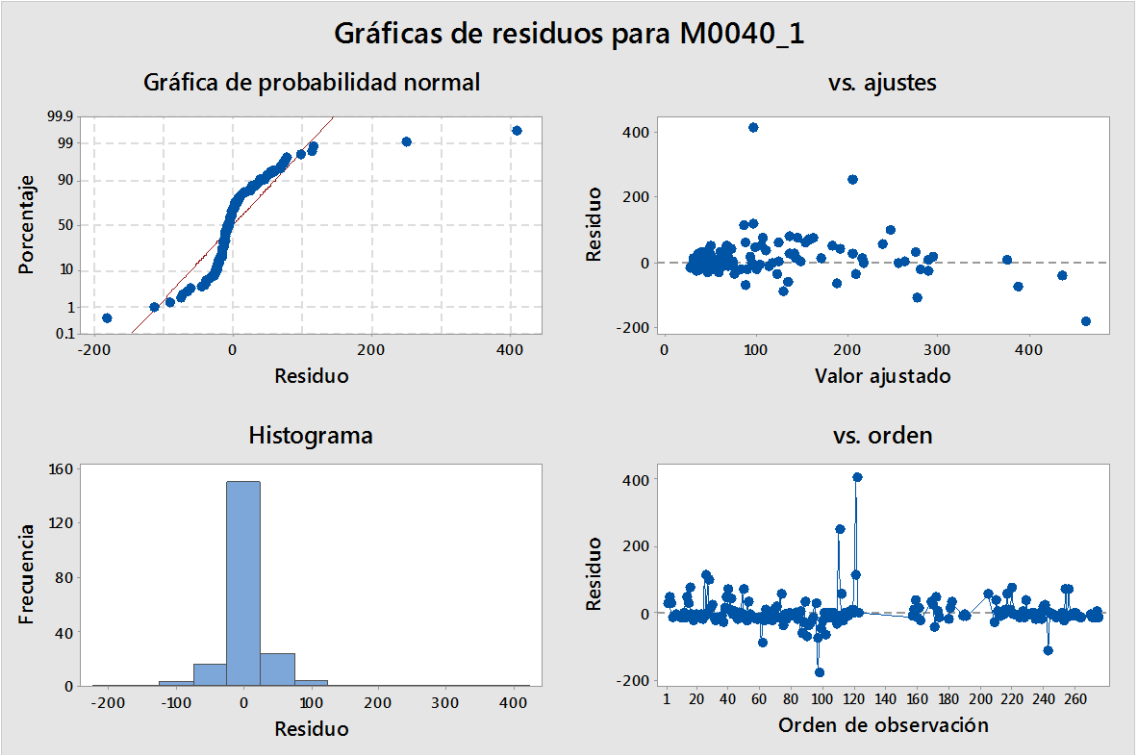


Ilustración 19: Análisis de supuestos de la estación meteorológica M0040

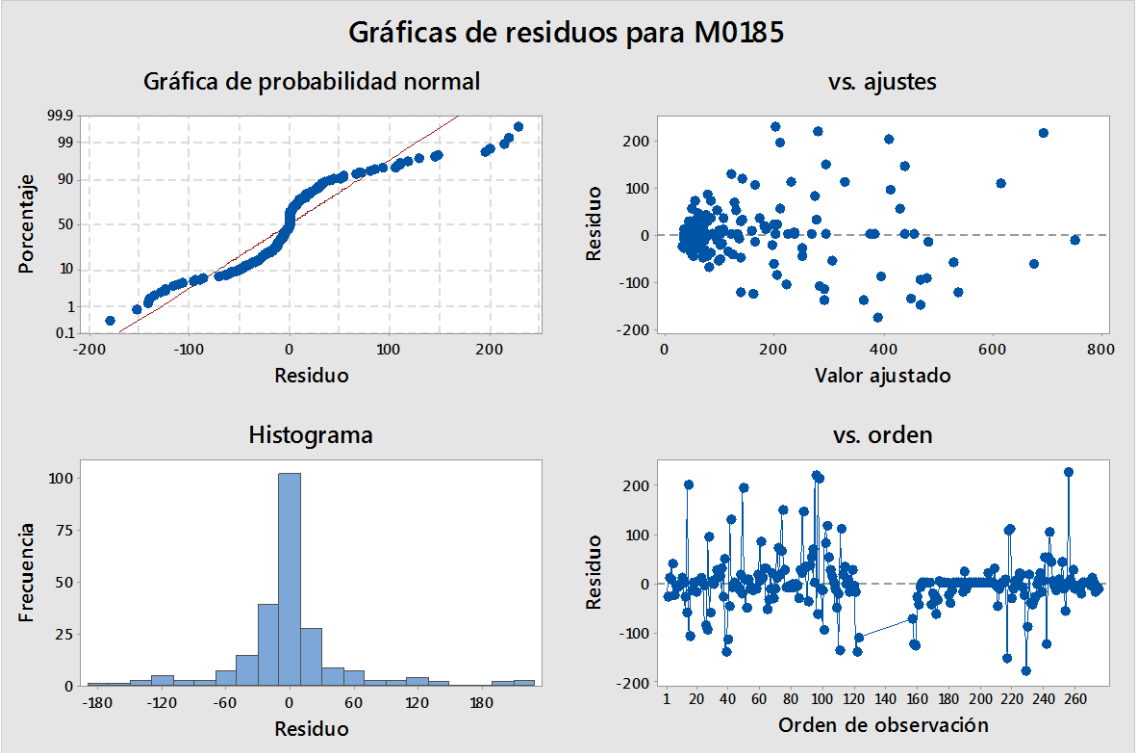


Ilustración 20: Análisis de supuestos de la estación meteorológica M0185

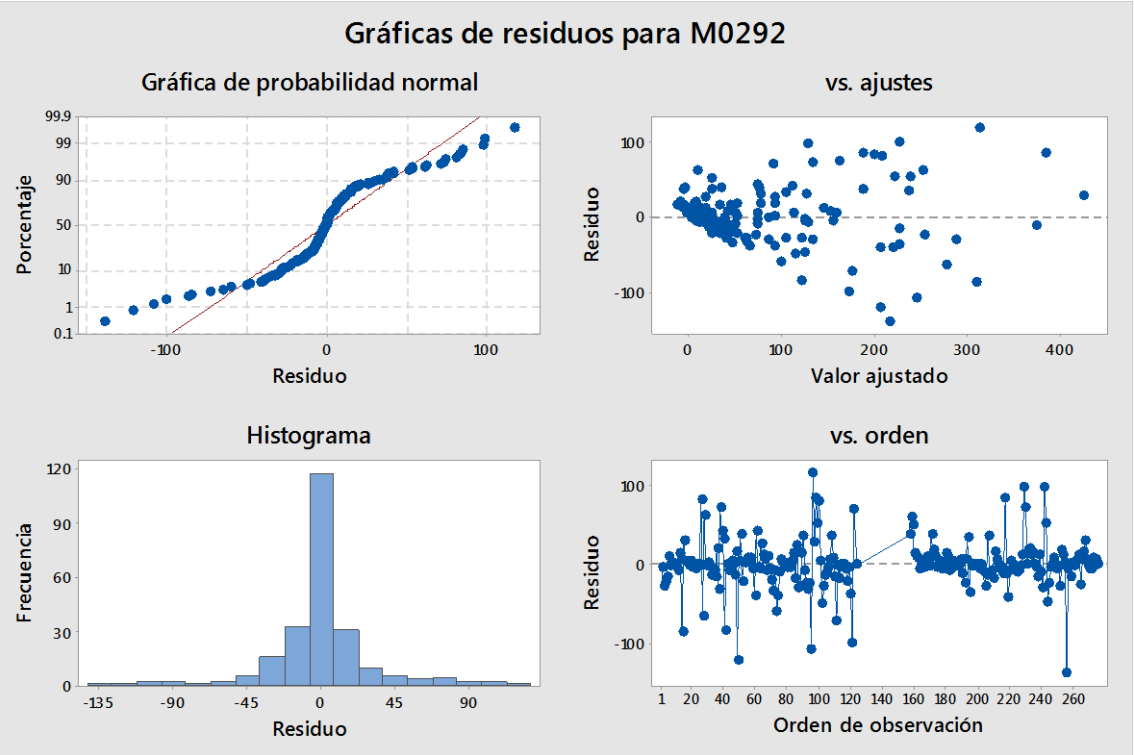


Ilustración 21: Análisis de supuestos de la estación meteorológica M0292

Las gráficas de residuos para las estaciones meteorológicas M031, M0185 y M0292 de la cuenca hidrográfica del río Esmeraldas presentan una distribución no normal en sus graficas de porcentaje vs residuo, pero no validar este supuesto no necesariamente se tiene que dudar de los modelos puesto que para imputación de datos hidrológicos un supuesto importante y en él se basan los modelos de regresión es el supuesto de linealidad y este parámetro es muy alto, esto se ve reflejado en el coeficiente de correlación de los análisis preliminares.

La homocedasticidad de estos modelos presenta una distribución aleatoria en forma de abanico lo que nos permite concluir con que existe una variable que está produciendo este efecto y debe ser tomado en cuenta para la generación de modelos más exactos.

En el supuesto independencia los modelos para las estaciones analizadas cumplen los criterios al estar de forma aleatoria respecto a la línea central en la gráfica de residuo vs orden, además la linealidad de los modelos es muy alta y cumple satisfactoriamente los supuestos de los siguientes modelos.

4.2.2. ANÁLISIS DE CORRELACIONES DE LAS ESTACIONES HIDROLÓGICAS

En el análisis de las correlaciones de datos con las estaciones hidrológicas cercanas se identifica si el modelo cumplirá con una correlación fuerte para poder continuar con los demás análisis de datos, es decir correlaciones superiores a 0.75. De esta manera se seleccionó la estación cercana que mayor grado de correlación mantenga en sus registros con la estación de análisis.

Para el caso de las estaciones de la cuenca del río Jubones el análisis de correlaciones entre las estaciones hidrológicas H530 y H531 no existe una correlación significativa, es decir que la correlación entre las estaciones es muy débil como se indica en la siguiente tabla.

Tabla 18. Análisis de correlaciones entre la estación hidrológica H530 y H531

Correlación de Pearson	-0.063
Valor p	0.459

Se estudió la posibilidad de imputar los datos con las demás estaciones hidrológicas de la cuenca, pero no existen correlaciones fuertes entre ninguna de ellas.

Las correlaciones en las estaciones hidrológicas H467 y H468 de la cuenca del río Cañar presentan una correlación media, pero no es lo suficientemente fuerte como para elaborar un modelo de predicción entre ellas, al igual que con las estaciones hidrológicas de la cuenca del río Jubones, se efectuó un análisis de las demás estaciones hidrológicas de la cuenca del río Cañar, obteniendo resultados similares que en el análisis de la anterior cuenca.

Tabla 19. Análisis de correlaciones entre la estación hidrológica H467 y H468

Correlación de Pearson	0.520
Valor p	0.000

En el caso de análisis de las estaciones hidrológicas de la cuenca del río Esmeraldas las estaciones hidrológicas de estudio H172 y H173, presentan una correlación positiva fuerte como se refleja en la siguiente tabla. Con un valor de correlación de Pearson alto se pude generar un modelo estadístico que permita predecir los datos de las series hidrológicas faltantes.

Tabla 20. Análisis de correlaciones entre la estación hidrológica H172 y H173

Correlación de Pearson	0.818
Valor p	0.000

4.2.2.1. ANÁLISIS DE VARIANZA

Una vez realizado del análisis de correlación se procede a seleccionar las estaciones que presentan correlaciones fuertes de la cuenca del río Esmeraldas. Se analizan los modelos de regresión lineal particionando de la variabilidad de los mismos mediante la técnica del análisis de varianza para probar la significación de la regresión en regresiones múltiples.

Tabla 21. Análisis de varianza entre las estaciones hidrológicas H172 y H173

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	5990	5990.18	315.17	0.000
H173	1	5990	5990.18	315.17	0.000
Error	156	2965	19.01		
Total	157	8955			

El valor p al ser menor a 0.05 en el análisis de varianza permite rechazar la hipótesis nula, ecuación (13) y asumir que los datos de modelo son significativos o que los modelos de imputación serán válidos.

Rechazar la hipótesis nula permite identificar que el coeficiente de la variable predictora será diferente de cero, con lo cual se asume que la recta de regresión lineal analizada no será una recta horizontal.

La hipótesis nula para la regresión general significa que el modelo no explica ninguna variación en la respuesta. Por lo general, un nivel de significancia (denotado como α o alfa) de 0.05 funciona adecuadamente. Un nivel de significancia de 0.05 indica un riesgo de 5% de concluir que el modelo explica la variación en la respuesta cuando no es así.

Tabla 22. Análisis de valores de R entre las estaciones hidrológicas de la cuenca del rio Esmeraldas

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
4.35960	66.89%	66.68%	62.10%

La tabla (22) permite identificar que la estación hidrológica H173 puede predecir el 66.89% de los datos contenidos en la estación hidrológica H173 de la cuenca del río Esmeraldas.

El análisis de varianza de la estación hidrológica como respuesta H173 respecto a su estación predictora H172 permite observar que al igual que el anterior análisis de varianza el valor p es menor que 0.05 y de esta forma rechazar la hipótesis nula para el modelo de predicción de datos y asumir que los valores son significativos, como se puede observar en la siguiente tabla:

Tabla 23. Análisis de varianza entre las estaciones hidrológicas H173 y H172

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	16275.6	16275.6	315.17	0.000
H172	1	16275.6	16275.6	315.17	0.000
Error	156	8055.9	51.6		
Falta de ajuste	155	8054.9	52.0	48.49	0.114
Error puro	1	1.1	1.1		
Total	157	24331.5			

Tabla 24. Análisis de varianza entre las estaciones hidrológicas H173 y H172

	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
7.18614	66.89%	66.68%	65.26%

De acuerdo modelo que permite la predicción de los datos de la estación H173 el 66.89% de sus datos pueden ser predichos por la estación H172.

4.2.2.2. MODELOS DE IMPUTACIÓN MEDIANTE REGRESIONES ITERATIVAS PARA ESTACIONES HIDROLÓGICAS

La siguiente ecuación representa el modelo de imputación de datos de la estación hidrológica H172 que permite predecir el registro hidrológico para las estaciones de estudio. Además, se efectuó el modelo con un nivel de confianza del 95%.

$$H172 = 1.923 + 0.4962\ H173$$

La ecuación del modelo que permite predecir los datos de la estación H173 respecto a la estación H172 se obtiene luego de pasar por los diversos filtros de análisis de datos, como resultado de dichos análisis y se obtiene un modelo de regresión lineal representado por la siguiente ecuación.

$$H173 = 0.346 + 1.3481\ H172$$

Tabla 25. Resumen de modelos lineales de imputación para las estaciones hidrológicas de la cuenca del río Esmeraldas

CUENCA HIDROGRÁFICA DEL RÍO ESMERALDAS	
Estaciones hidrológicas	**Ecuaciones**
H172	1.923 + 0.4962 H173
H173	0.346 + 1.3481 H172

Dichos modelos representados por las ecuaciones de regresión lineal permitieron en lo posterior completar los registros de caudales medios mensuales en las estaciones hidrológicas de la cuenca del río Esmeraldas en un período de 22 años, comprendidos entre 1990 y el 2012.

4.2.2.3. ANÁLISIS DE SUPUESTOS

Gráficas para análisis de supuestos de las estaciones hidrométricas H173 y H172 ubicadas en la cuenca hidrográfica del río Esmeraldas:

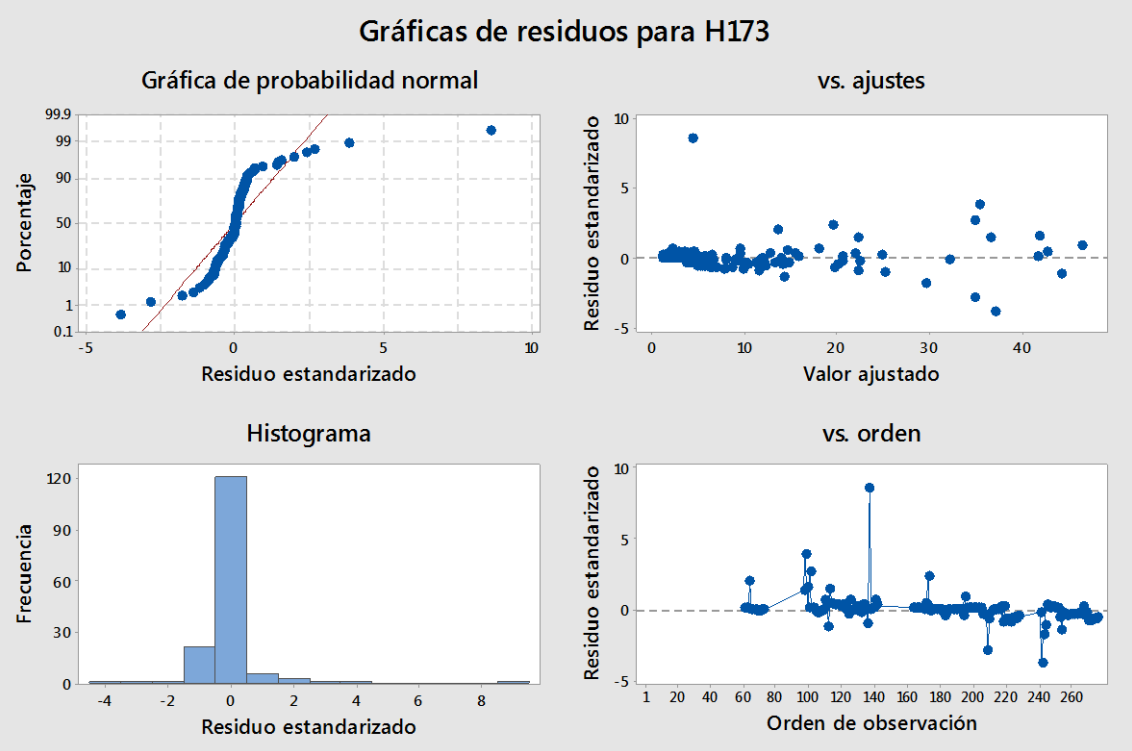


Ilustración 22: Análisis de supuestos de la estación hidrológica H173

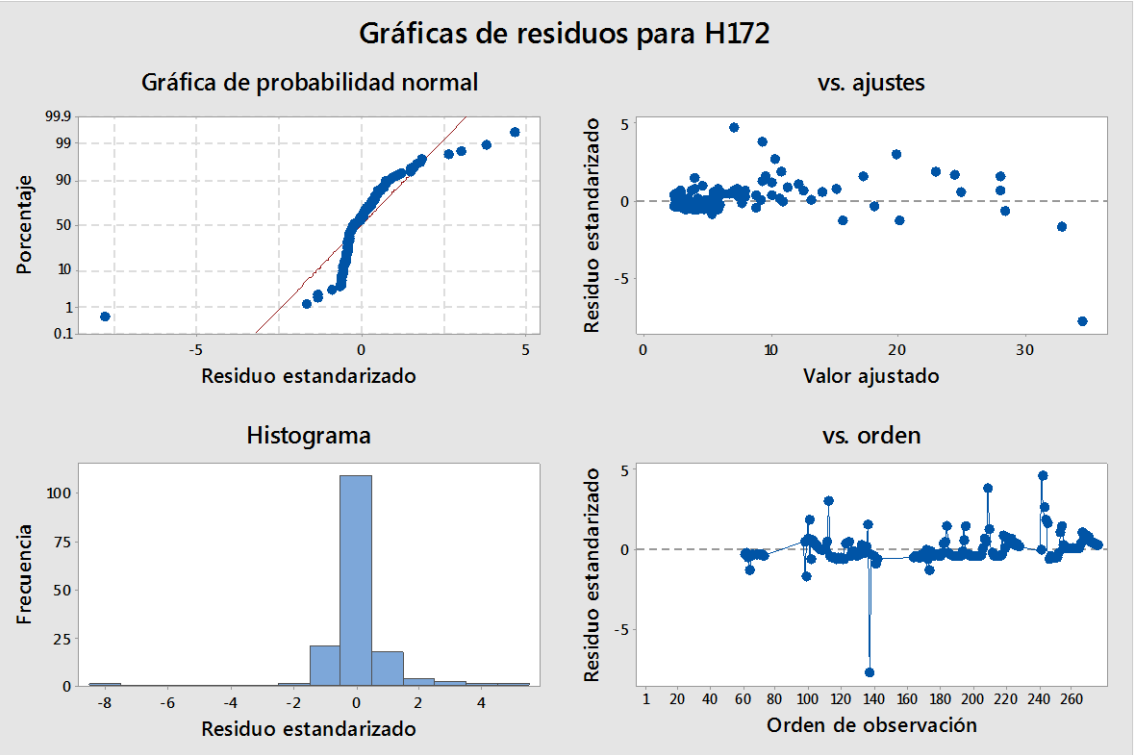


Ilustración 23: Análisis de supuestos de la estación hidrológica H172

En los gráficos de las ilustraciones (22-23) se observa que los datos no siguen una distribución normal significativa para los residuos de la estación H173 y H172, además, existen valores anormales en los extremos de la recta y una distribución en forma de S.

Lo más relevante de los gráficos de supuestos para la validación del modelo es grafica de homocedasticidad que permite observar un nivel de varianza alto entre sus datos.

Manteniendo el criterio del supuesto de linealidad como correlación fuerte positiva se asume que el modelo es válido para las dos estaciones hidrológicas de la cuenca del río Esmeraldas.

4.3. IMPUTACIÓN DE DATOS FALTANTES EN ESTACIONES HIDROMETEOROLÓGICAS MEDIANTE MACHINE LEARNING

Esta metodología se basa en inteligencia automatizada basada en lenguaje de programación Python. Este método de imputación se basa en modelos de aprendizaje supervisado, es decir a la maquina se le presenta la información de respuesta al

mismo tiempo que la información de entrada, con lo cual la maquina aprenderá a llegar a la respuesta mediante un proceso iterativo.

En definitiva, una máquina de aprendizaje autónomo divide los datos en tres secciones, una para datos de prueba, una segunda parte para datos de entrenamiento y finamente una tercera parte de los datos para comprobaciones, de esta manera efectúa simultáneamente miles de procesamientos con la finalidad de encontrar una respuesta o un modelo que pueda satisfacer al problema.

Python permite acceder a librerías que contienen algoritmos computacionales que facilitan la implementación de las máquinas de aprendizaje. Las librerías usadas en este análisis fueron las siguientes:

- Pandas: es una librería de código abierto para lenguaje de programación Python, esta librería permite hacer análisis matemáticos y de datos de alto rendimiento.
- Numpy: permite la manipulación de los datos, generar matrices que puedan ser analizadas entre una de sus funciones principales.
- Scikit-learn: esta herramienta es una librería de libre acceso permite la minería de datos, además es una potente máquina de aprendizaje autónomo, dentro de sus funciones tiene la de predicción, agrupación y clasificación.
- Seaborn: Seaborn es una biblioteca de visualización de datos de Python basada en matplotlib. Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos.
- Statsmodels: es un módulo de Python que proporciona clases y funciones para la estimación de muchos modelos estadísticos diferentes, así como para realizar pruebas estadísticas y la exploración de datos estadísticos. Una lista extensa de estadísticas de resultados está disponible para cada estimador.

4.3.1. ALGORITMO DE IMPUTACIÓN DE DATOS MEDIANTE MACHINE LEARNING.

El algoritmo computacional se puede definir como una secuencia de pasos ordenados que en conjunto solucionan un problema o hallan una solución específica.

Para la imputación de los registros hidrometeorológicos se creó un algoritmo en Python, ilustración (24) que permite incorporar en sus líneas una máquina de aprendizaje “Sckitlearn_linear_model” basado en un módulo de regresiones lineales que permite estimar un modelo matemático con el cual se puede imputar los datos faltantes de los registros hidrometeorológicos. Esta máquina tiene como función

efectuar regresiones en un proceso iterativo que procesa la información proporcionada.

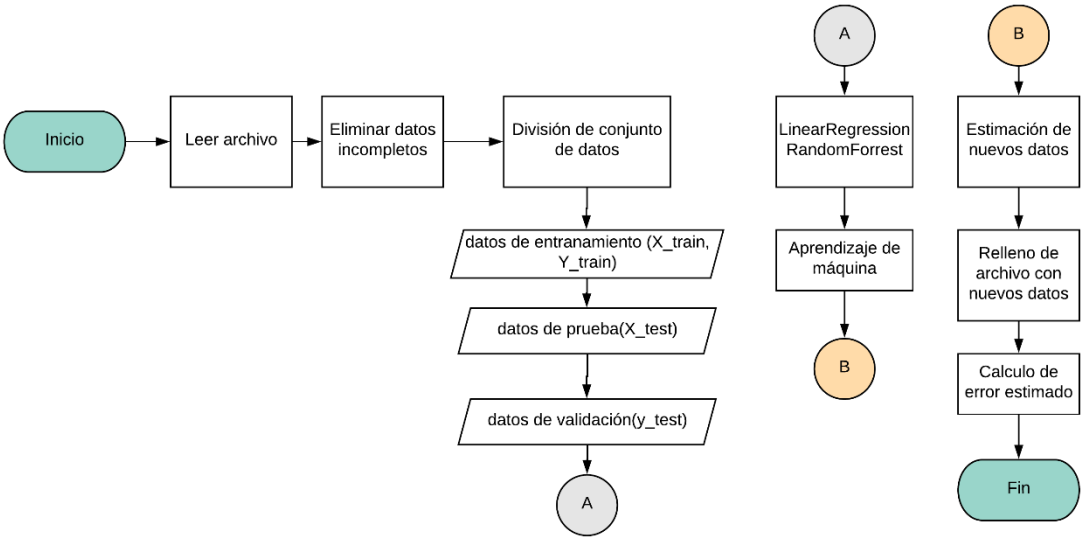


Ilustración 24: Algoritmo computacional para la imputación de datos faltantes mediante Machine Learning

Algoritmo de Machine Learning explicado:

1- Inicio: cargar las librerías adecuadas en lenguaje de programación Python utilizando el entorno Spyder.
2- Lee el archivo csv que contiene la estación hidrometeorológica de análisis y las estaciones de referencia que tendrán la función de ser las variables predictoras.
3- Se indica la estación que será la variable de respuesta y las estaciones que serán las variables predictoras, es decir, se estable las variables X e Y que configurarán el modelo de respuesta.
4- Se elimina los datos faltantes para crear un matriz de dimensiones similares entre las variables predictoras y la variable de respuesta.
5- Mediante validación cruzada se divide el conjunto de datos en datos de entrenamiento, prueba y validación, esta división permite un mejor ajuste en el modelo, puesto que la validación cruzada permite interactuar a estos conjuntos para que el modelo se ajuste a todo el conjunto de datos, esto se puede entender mejor en la ilustración (25).
6- Ingeniería de aprendizaje autónomo, en este paso se establece los parámetros adecuados para el correcto funcionamiento de la máquina de aprendizaje, es decir, se establece el tamaño de los datos de entrenamiento y de prueba. Cabe

recalcar que la labor del *machine learning* es ajustar automáticamente los valores aprendiendo de las entradas.

7- Con el modelo creado por la máquina de aprendizaje se completarán los datos faltantes que en al principio se eliminaron.
8- Se calcula el error cuadrático medio entre los datos observados y los datos predichos para evaluar el desempeño de la máquina de aprendizaje automático.
9- Para finalizar se repite desde el paso 4 donde se corrigen los parámetros de la máquina de aprendizaje automático con la finalidad de determinar el error cuadrático medio más bajo, el cual determinará el mejor modelo de imputación de datos.

A continuación, se presenta las líneas de programación que contiene al algoritmo computacional para poder determinar un modelo que permita completar los registros hidrometeorológicos cuando existan más de una estación de referencia para la imputación de datos hidrometeorológicos, también se permite entender el funcionamiento de este proceso computacional.

Importar librerías necesarias con estas líneas de código permite como primer paso importar las librerías de libre acceso que permitirán la ejecución de todo el algoritmo computacional.

```
1.  from pandas import DataFrame
2.  import pandas as pd
3.  from sklearn.linear_model import LinearRegression
4.  import statsmodels.api as sm
5.  import numpy as np
6.  from sklearn import metrics
7.  from sklearn.cross_validation import train_test_split
8.  from sklearn.model_selection import cross_val_score
```

Código 1: Importación de librerías

Paso seguido es necesario procesar la información adecuadamente para que la máquina de aprendizaje automatizado identifique las variables analizadas y pueda trabajar adecuadamente.

Se ingresa las variables a analizar, en este caso se seleccionan dos variables predictoras (variable_x1, variable_x2) y una variable de respuesta (variable_y), se identifica el archivo (*.csv) en el cual se está trabajando y se creara un nuevo archivo csv, el cual contendrá la nueva información para la variable_y de respuesta que en resumen será el resultado del modelo de Machine Learning.

De esta manera quedan establecidas cuales estaciones asumirán las variables predictoras y que variable será la variable de respuesta, también se puede asignar más de 2 variables predictoras modificando las líneas de comando para que pueda satisfacer las necesidades de los datos.

```
9.  variable_x1 = "M0292"
10. variable_x2 = "M0185"
11. variable_y = "M0040"
12. nombre_archivo = "jubones.csv"
13. nombre_copia = "copia_multiple.csv"
```

Código 2: Definición de las variables a analizar.

A continuación, se convierte la lista de datos en un matriz, esta lista de datos procede de la lectura del archivo (*.csv) que contiene las estaciones hidrometereológicas de análisis.

```
14. def convertir_lista(array):
15.     n = len(array)
16.     l = []
17.     for i in range(n):
18.         l.append(float(array[i]))
19.     return l
```

Código 3: Procesamiento de los registros

Se continua con la conformación de funciones para la lectura de los archivos y se establece además como se creará el nuevo archivo que tendrá la información imputada por el modelo lineal múltiple que la máquina de aprendizaje lo obtendrá al final del proceso.

Para crear el modelo lineal al final del análisis de imputación de datos se establece en la línea 20 del código 4 la estructura del modelo de imputación de datos que mantiene la forma de una ecuación lineal múltiple, siendo b_0 y b_1 los coeficientes de la recta de regresión y $const$ que equivale a la pendiente del modelo.

```
20. def fun(b0, b1, const, x1, x2):
21.     return (b0*x1 + b1*x2 + const)
22.
23. def agregar_linea(fecha, v1, v2, v3, nombre_destino):
24.     f = open(nombre_destino,"a")
25.     cadena = str(fecha)+","+str(v1)+","+str(v2)+ "," +str(v3) +"\n"
26.     f.write(cadena)
27.     f.close()
28.
29. def cargar_informacion(nombre):
30.     data = pd.read_csv(nombre)
31.     data.head()
32.     return data
33.
34. def crear_archivo_csv(nombre_destino, v_x, v_y, v_z):
35.     cadena = "fecha"+","+v_x+","+ v_y + "," + v_z +"\n"
36.     f = open(nombre destino,"w")
37.     f.write(cadena)
38.     f.close()
```

Código 4: Procesamiento de la información, cargar y crear nuevos archivos

En el siguiente paso la data se selecciona; se divide en partes: datos de entrenamiento, prueba y datos de validación, esta división se establece mediante validación cruzada que permite una distribución adecuada de los datos, entre datos de prueba y validación para que el modelo no se sobre ajuste en los datos entrenados y tenga deficiencias en los datos de validación, ilustración (25). Seguido la máquina de aprendizaje supervisado usando el módulo scikit learn Linear Regression, procesa esta información y establezca un modelo lineal basado en el método de mínimo cuadrados. Las múltiples iteraciones que efectúa la máquina de aprendizaje con los datos de entrenamiento y prueba permiten identificar el mejor modelo lineal, que permitirá en lo consiguiente la imputación de los datos hidrometereológicos.

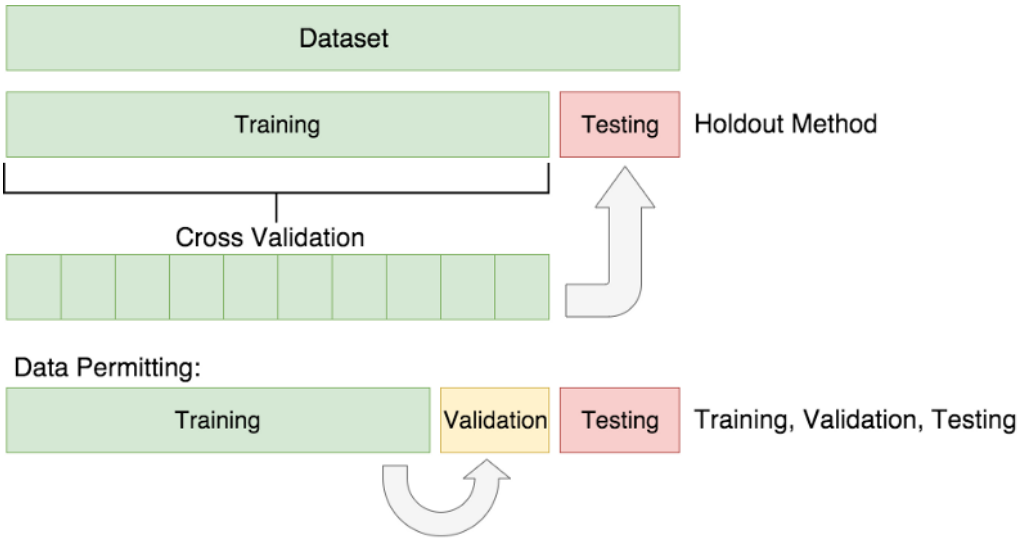


Ilustración 25: Funcionamiento de validación cruzada, entre los datos de entrenamiento, prueba y validación.

Fuente: www.towardsdatascience.com

Por otra parte, en el código 5, en la línea de programación 53 se establece el porcentaje de datos que se tomara como muestra para definir el modelo adecuadamente, en otras palabras, es aquí donde se ajusta los datos que se analizaran para que el modelo predictor mantenga en sus resultados finales de imputación con un error cuadrático medio bajo.

```
39. def funcion_minimos_cuadrados(nombre, v_x1, v_x2, v_x3):
40.     slr_df = pd.read_csv(nombre, sep=",")
41.     slr_df.head()
42.     slr_df = slr_df.dropna()
43.     x1 = slr_df[v_x1].values.reshape(-1,1)
44.     x2 = slr_df[v_x2].values.reshape(-1,1)
45.     x3 = slr_df[v_x3].values.reshape(-1,1)
46.     X1 = convertir_lista(x1)
```

```
47.     X2 = convertir_lista(x2)
48.     X3 = convertir_lista(x3)
49.     diccionario = {v x1:X1, v x2:X2, v x3:X3}
50.     df = DataFrame(diccionario,columns=[v x1,v x2,v x3])
51.     X_sol = df[[v_x1,v_x2]]
52.     Y_sol = df[v_x3]
53.     X_sol_train, X_sol_test, Y_sol_train, Y_sol_test = train_test_split(X_sol, Y_sol
    , test_size=0.16, random_state=5)
54.     print ('varaibles de entremamiento: ', X_sol_train.shape)
55.     print ('caracteristicas de entrenamiento:', Y_sol_train.shape)
56.     print ('varaibles de prueba: ', X_sol_test.shape)
57.     print ('caracteristicas de prueba: ', Y_sol_test.shape)
58.     regr = LinearRegression()
59.     regr.fit(X sol train, Y sol train)
60.     predictions = regr.predict(X sol test)
61.     print('MAE: ', metrics.mean_absolute_error(Y_sol_test, predictions))
62.     print ('r2_score', metrics.r2_score(Y_sol_test, predictions))
63.     print('Intercept: \n', regr.intercept_)
64.     print('Coefficients: \n', regr.coef_)
65.     print ("validacion
    cruzada:",cross val score(regr,X sol train,Y sol train,cv=10))
66.
67.     # with statsmodels
68.     X_sol = sm.add_constant(X_sol) # adding a constant
69.     model = sm.OLS(Y sol, X sol).fit()
70.     predictions = model.predict(X sol)
71.     print model = model.summary()
72.     print(print_model)
73.     constante = regr.intercept_
74.     b0  = regr.coef_[0]
75.     b1  = regr.coef_[1]
76.     l =[ b0, b1, constante]
77.     return l
```

Código 5: Implementación de máquina de aprendizaje autónomo basado en módulos de regresiones lineales múltiples.

En la selección de la máquina de aprendizaje autónomo se puede también seleccionar la máquina de aprendizaje basado en bosques aleatorios de decisión conocida como Random Forest, código 6, esta máquina basa su aprendizaje en un conjunto de datos de referencia y a su vez en la selección aleatoria de decisiones que le permite predecir la variable analizada, para ello selecciona un porcentaje de datos que le servirán como datos de entrenamiento y datos de prueba, empleando de la misma manera la validación cruzada en los datos de entrenamiento, prueba y validación, como se explica en la ilustración (25).

```
1.  def forrest(nombre_archivo, nombre_destino, variable_x, variable_y):
2.      data = pd.read_csv(nombre_archivo)
3.      x_ori = data[variable_x].values
4.      y_ori = data[variable_y].values
5.      dataset = pd.read_csv(nombre_archivo)
6.      dataset.dropna
7.      dataset = dataset.dropna()
8.      X = dataset[variable_x].values.reshape(-1,1)
9.      y = dataset[variable_y].values.reshape(-1,1)
10.     # Splitting the dataset into the Training set and Test set (cross validation)
11.     X train, X test, y train, y test = train test split(X, y, test size = 0.28, rand
    om state = 5)
12.     regressor = RandomForestRegressor(n_estimators = 100, random_state = 0)
13.     regressor.fit(X_train, y_train.ravel())
```

Código 6: Implementación de máquina de aprendizaje autónomo Random Forest.

Finalmente, el programa procederá al cálculo del error cuadrático medio para poder estimar el error entre los datos observados y los datos predichos.

Para esta función la maquina utiliza el archivo con la información hidrometereológica original y el archivo nuevo con los datos imputados por Machine Learning; efectúa el cálculo aritmético entre estos dos archivos con la finalidad de determinar el error cuadrático medio.

```
78. def calcular_error(nombre,nombre_destino, variable_y):
79.     prom_orig = calcular_promedio_col(variable_y,nombre)
80.     print("promedio original: ", prom_orig)
81.     prom copy = calcular promedio col(variable y, nombre destino)
82.     print("promedio copia: ", prom_copy)
83.     total = ((prom orig - prom copy)**2)
84.     print("error: ",total)
```

Código 7: Cálculo del error cuadrático medio entre valores observados y los valores imputados.

De la misma manera que la programación anterior se procedió a crear una variación de lo expuesto con la finalidad que el programa pueda reconocer solamente dos estaciones; una de referencia y una estación de análisis, es decir que el programa identificará una variable predictora y una variable de respuesta, X e Y respectivamente, para determinar un modelo basado en regresiones lineales simples, esto será útil en el caso de tener dos estaciones, con un valor de correlación lineal fuerte entre ellas, código (8).

```
1.  nombre = "ESMERALDAS.csv"
2.  nombre_destino= "copia_2.csv"
3.  v_x = "M0003"
4.  v y = "M0364"
5.  forrest(nombre, nombre destino, v x, v y)
6.  calcular_error(nombre, nombre_destino, v_y)
```

Código 8: Definición de dos estaciones a ser analizadas.

Con dos variables la maquina a de aprendizaje autónomo predictora efectúa el mismo procesamiento de los datos que cuando existe más de dos variables. A continuación, se presenta un extracto del código de programación donde la máquina de aprendizaje utiliza únicamente dos estaciones hidrometeorológicas.

```
1.  def machine_learning(nombre, variable_x, variable_y):
2.      slr_df = pd.read_csv(nombre, sep=",")
3.      slr_df.head()
4.      slr_df = slr_df.dropna()
5.      x = slr df[variable x].values.reshape(-1,1)
6.      y = slr df[variable y].values.reshape(-1,1)
7.      X_train, X_test, y_train, y_test = train_test_split(x, y , test_size=0.30, rando
    m_stat  e=10)
8.      lm = LinearRegression()
9.      lm.fit(X train, y train)
10.     slope = lm.coef
11.     predictions = lm.predict(X_test)
12.     sns.distplot((y_test - predictions), bins = 1)
```

```
13.     print ('La recta de regresión es: y = %f + %f * x'%(lm.intercept_, slope))
14.     func_x=[lm.intercept_, slope]
15.     valores y = fun x(x, lm.intercept , slope)
16.     #plt.plot(x,valores y )
17.     plt.scatter(y_test, predictions,  c='r', edgecolors=(0, 0, 0), alpha=0.5)
18.     plt.plot(x,valores_y, c = 'b')
19.     plt.title(' vs test values', fontsize=10)
20.     plt.xlabel('test values')
21.     plt.ylabel('predicted values')
22.     plt.show()
23.     print ("validacion cruzada:",cross_val_score(lm,X_train,y_train,cv=5))
24.     print('Evaluacion de metodo',lm.score(X_test,y_test))
25.     return func_x
```

Código 9: máquina de aprendizaje autónomo basado en módulos de regresiones lineales simples.

Como ejemplo se observa el extracto del resultado de las imputaciones realizadas en las estaciones hidrológicas de la cuenca del río Esmeraldas, donde no había información alguna en la estación H172 para el periodo de 18 meses consecutivos; luego de aplicar Machine Learning podemos tener datos confiables en los registros.

Tabla 26. Ejemplo de la imputación de datos faltantes mediante Machine Learning en la estación hidrológica H172 mediante Linear Regression.

Datos originales de estaciones hidrológicas		
fecha	H173	H172
01/01/1990	4.392	
01/02/1990	12.683	
01/03/1990	11.522	
01/04/1990	11.248	
01/05/1990	3.757	
01/06/1990		
01/07/1990		
01/08/1990		
01/09/1990	2.243	
01/10/1990	2.017	
01/11/1990	1.936	
01/12/1990	2.123	
01/01/1991		
01/02/1991	15.92	
01/03/1991	10.919	
01/04/1991	8.432	
01/05/1991	10.127	
01/06/1991	3.494	
01/07/1991	2.964	
01/08/1991	2.686	
01/09/1991	2.376	
01/10/1991	2.275	
01/11/1991	2.216	
01/12/1991	2.445	

Datos imputados de estaciones hidrológicas		
fecha	H173	H172
01/01/1990	4.392	4.073
01/02/1990	12.683	8.055
01/03/1990	11.522	7.497
01/04/1990	11.248	7.366
01/05/1990	3.757	3.768
01/06/1990		
01/07/1990		
01/08/1990		
01/09/1990	2.243	3.041
01/10/1990	2.017	2.932
01/11/1990	1.936	2.893
01/12/1990	2.123	2.983
01/01/1991		
01/02/1991	15.92	9.610
01/03/1991	10.919	7.208
01/04/1991	8.432	6.013
01/05/1991	10.127	6.827
01/06/1991	3.494	3.642
01/07/1991	2.964	3.387
01/08/1991	2.686	3.254
01/09/1991	2.376	3.105
01/10/1991	2.275	3.056
01/11/1991	2.216	3.028
01/12/1991	2.445	3.138

4.3.2. MODELOS RESULTANTES DE LA IMPUTACIÓN MEDIANTE MACHINE LEARNING

La máquina de aprendizaje autónomo basado en regresiones lineales y en bosques aleatorios de decisión, obtuvieron modelos que permitieron la imputación de datos faltantes en los registros hidrometeorológicos de las estaciones ubicadas en las cuencas de estudio, siendo estas las estaciones de las cuencas de los ríos Esmeraldas, Cañar y Jubones.

Los siguientes modelos están calibrados para satisfacer la imputación de datos faltantes de cada estación dentro del período de tiempo de registros que comprenden 22 años, desde 1990 hasta el 2012, tanto para las estaciones meteorológicas como las estaciones hidrológicas. Se debe también tener en cuenta que se trabajó únicamente con las estaciones hidrometeorológicas cercanas por cuenca hidrográfica y que en el análisis de correlaciones mantuvieron entre ellas un valor de correlación mayor o igual a 0.75.

A continuación, se representan la recta de regresión que mejor resultados obtuvo para cada estación hidrometeorológica de análisis de las diferentes cuencas hidrográficas, estos modelos se establecieron con la máquina de aprendizaje Linear Regression de la biblioteca Sklearn de Python y a sus conjuntos de datos se les aplico validación cruzada, como un pilar fundamental para la validación de los resultados. Los gráficos tienen una relación entre los valores de prueba y los valores predichos, como consiguiente a partir de estos resultados los modelos lineales permitieron imputación de datos faltantes en los registros hidrometeorológicos.

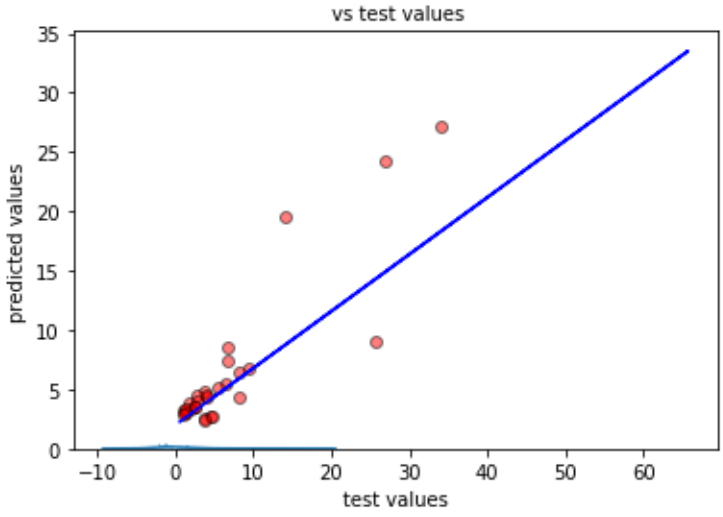


Ilustración 26: Recta de regresión lineal obtenida mediante el algoritmo de Machine Learning Linear Regression entre valores de prueba y los valores predichos de las estaciones H172 y H173

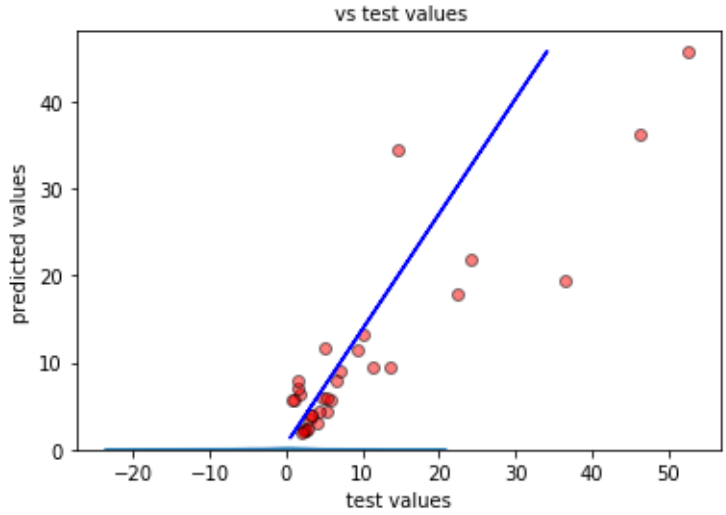


Ilustración 27: Recta de regresión lineal obtenida mediante el algoritmo de Machine Learning Linear Regression entre valores de prueba y los valores predichos de las estaciones H173 y H172

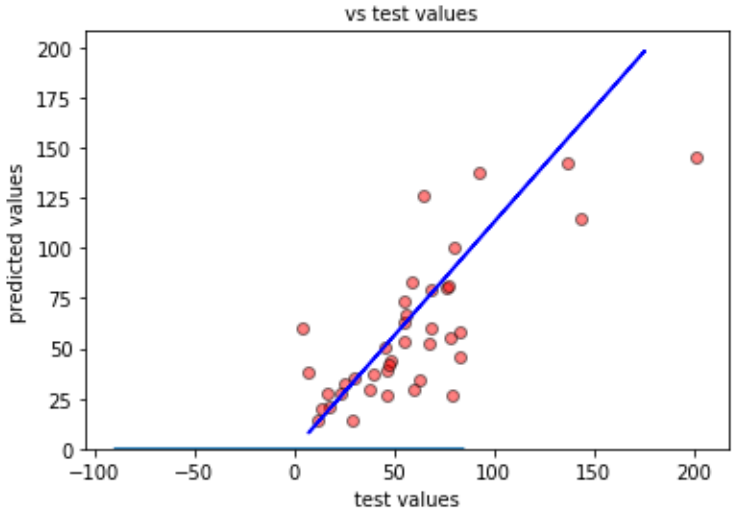


Ilustración 28: Recta de regresión lineal obtenida mediante el algoritmo de Machine Learning Linear Regression entre valores de prueba y los valores predichos de las estaciones M0411 y M031

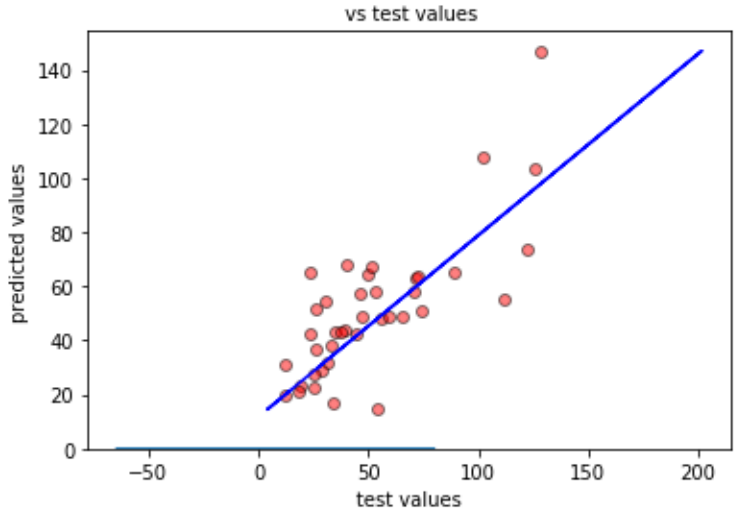


Ilustración 29: Recta de regresión lineal obtenida mediante el algoritmo de Machine Learning Linear Regression entre valores de prueba y los valores predichos de las estaciones M031 y M0411

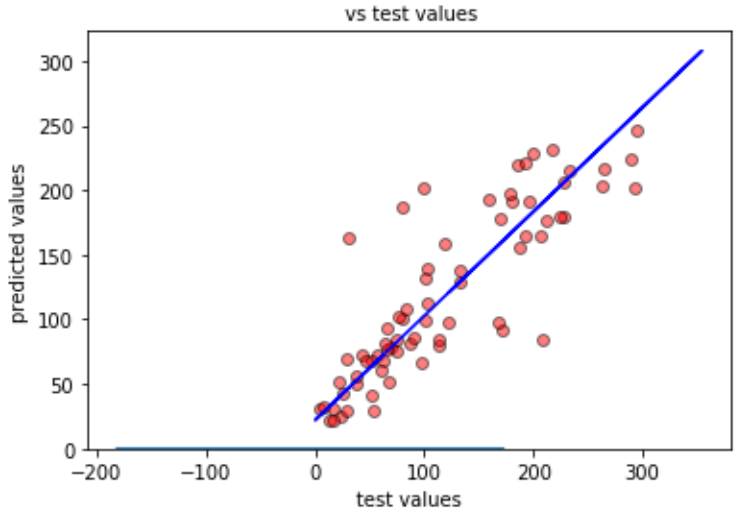


Ilustración 30: Recta de regresión lineal obtenida mediante el algoritmo de Machine Learning Linear Regression entre valores de prueba y los valores predichos de las estaciones M0003 y M0364

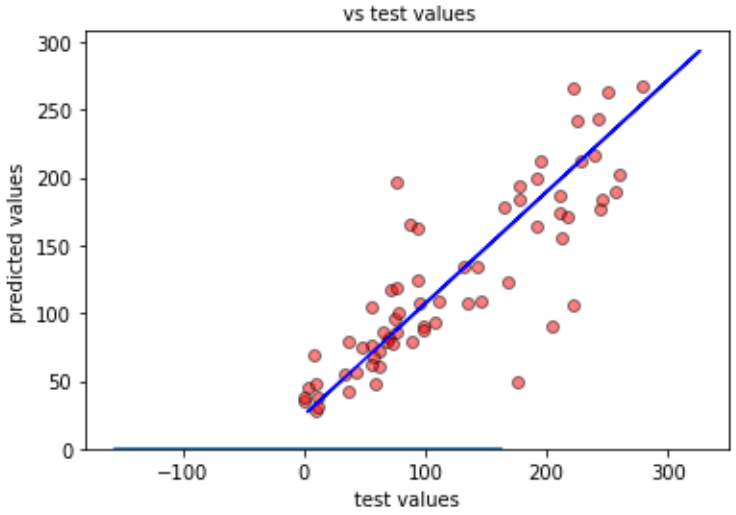


Ilustración 31: Recta de regresión lineal obtenida mediante el algoritmo de Machine Learning Linear Regression entre valores de prueba y los valores predichos de las estaciones M0364 y M0003

Tabla 27. Resumen de modelos lineales de imputación mediante machine Learning basado en regresiones lineales para las estaciones hidrometeorológicas de las cuencas analizadas.

RESUMEN DE MODELOS LINEALES DE IMPUTACIÓN DE VALORES FALTANTES MEDIANTE MACHINE LEARNING	
Estaciones meteorológicas de la cuenca del río Esmeraldas	Modelo lineal
M003	30.509 + 0.775 M364
M364	21.339 + 0.8121 M003
H172	1.894+ 0.488 H173
H173	0.596+ 1.323 H173
Estaciones meteorológicas de la cuenca del río Cañar	Modelo lineal
M411	1.59 + 1.0939 M31

M031	11.768 + 0.671 M411

Estaciones meteorológicas de la cuenca del río Jubones	Modelo lineal
M040	26.927 + 0.189 M185 + 0.652 M292
M185	27.846 + 0.271 M040 + 1.306 M292
M292	-15.282 + 0.249 M040 + 0.4152 M185

De la misma manera se presentan los gráficos para obtención de los modelos de imputación basados en máquina de aprendizaje Random Forest, los gráficos presentan la relación entre los valores de los datos de la estación de referencia y los datos de la estación de análisis, para los gráficos siguientes sus parámetros están calibrados para que el resultado en el error cuadrático medio sea el menor posible.

Ilustración 32: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predictores de las estaciones M0364 y M0003

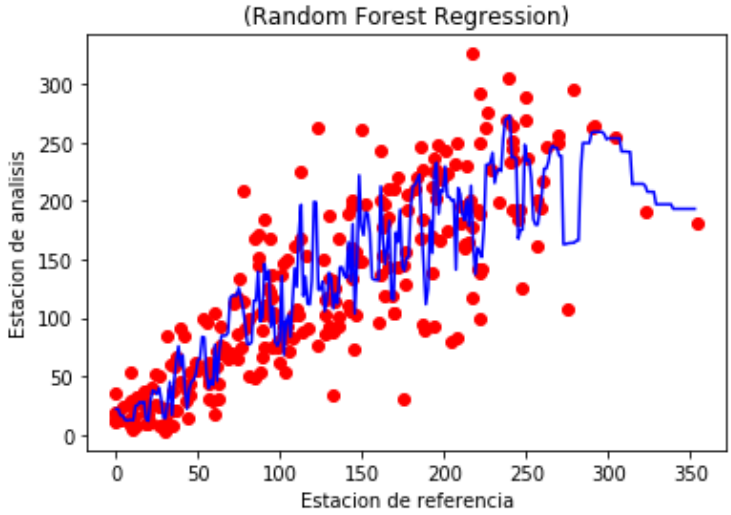


Ilustración 33: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones M0003 y M0364

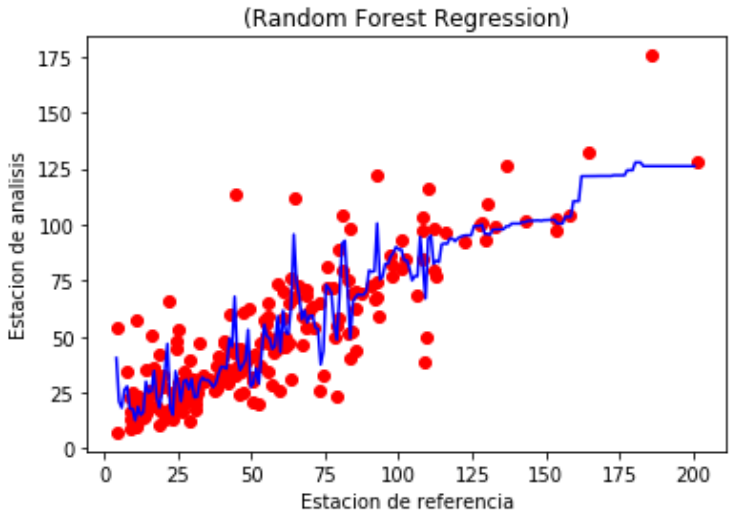


Ilustración 34: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones M0411 y M031

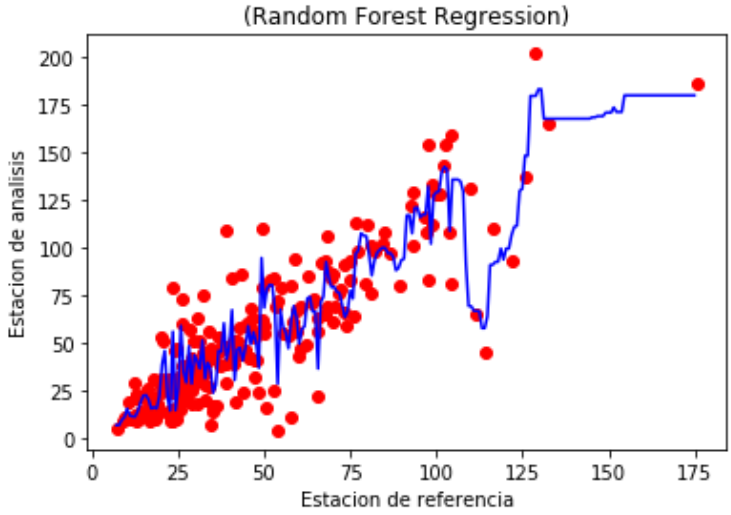


Ilustración 35: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predicho de las estaciones M031 y M0411

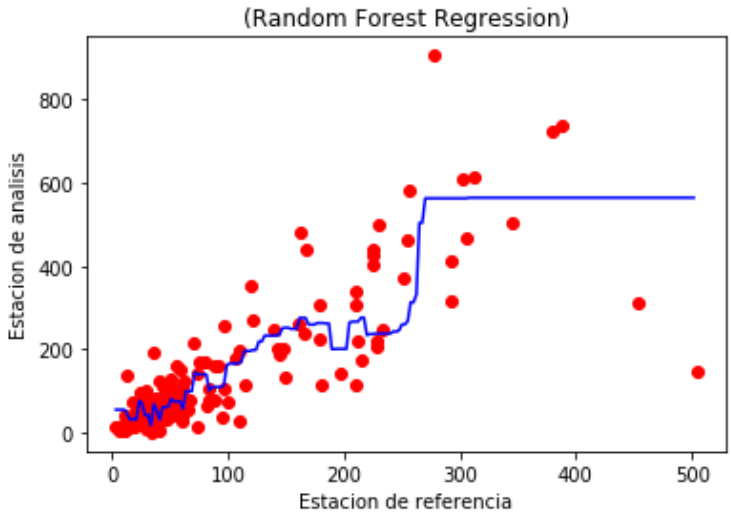


Ilustración 36: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones M0040 y M0185

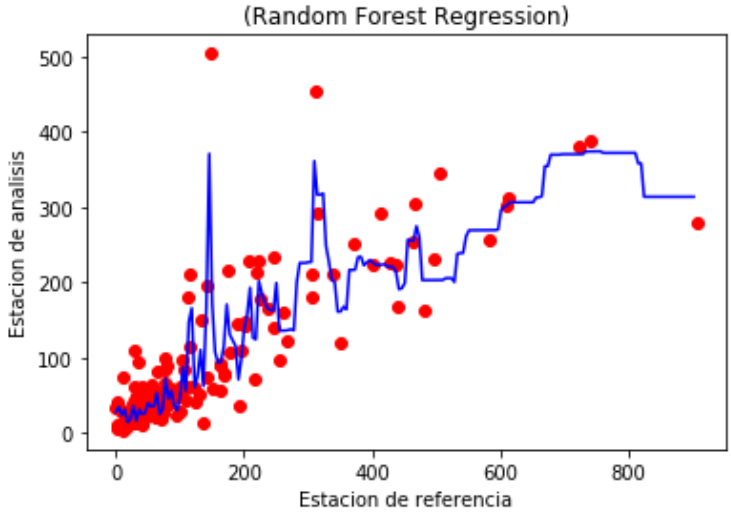


Ilustración 37: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones M0185 y M0040

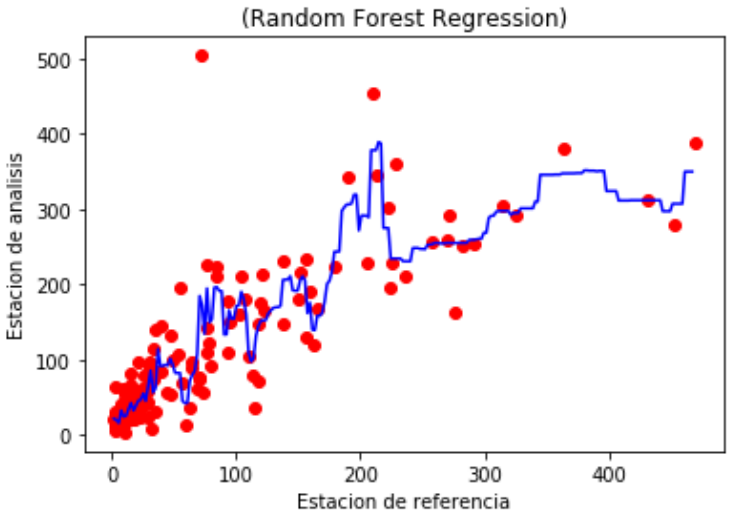


Ilustración 38: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones M0292 y M0040

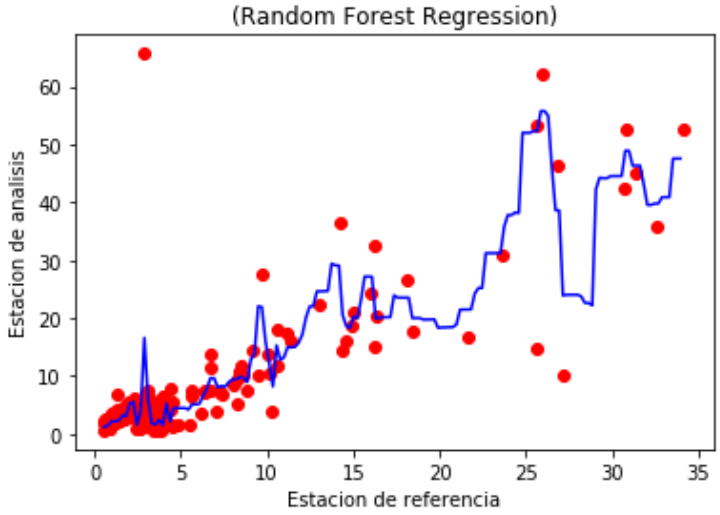


Ilustración 39: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones H172 y H173

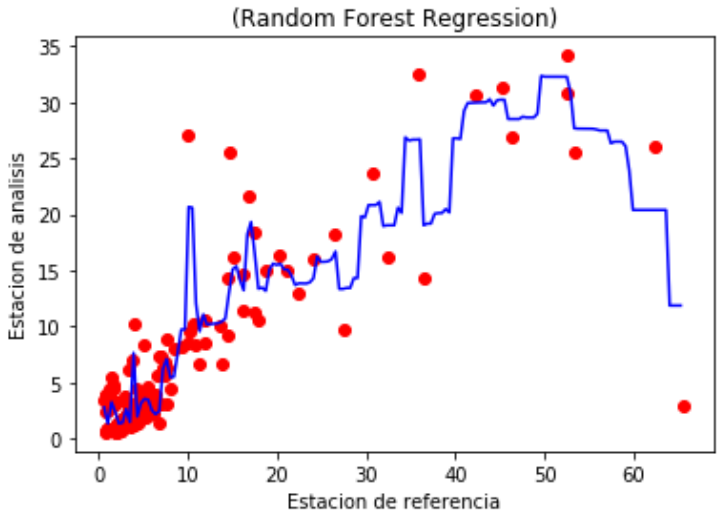


Ilustración 40: Modelo predictivo obtenido mediante el algoritmo de Machine Learning Random Forest entre valores de prueba y los valores predichos de las estaciones H173 y H172

4.3.3. ERROR CUADRÁTICO MEDIO DE LOS MODELOS DE IMPUTACIÓN MEDIANTE MACHINE LEARNING LINEAL REGRESSION Y RANDOM FOREST

El cálculo del error cuadrático medio de los modelos de imputación mediante Machine Learning nos permite tener una métrica con la cual podemos deducir la diferencia que existe entre los datos observados en relación a los datos imputados por los diferentes modelos lineales.

El error cuadrático medio también nos indica con facilidad, cuan diferentes son los datos observados respecto a los datos predichos para cada modelo en este caso observamos el error cuadrático medio al usar Machine Learning con el módulo Linear Regression.

Tabla 28. Resumen de error cuadrático medio calculado mediante machine Learning Linear Regression para las estaciones hidrometeorológicas de las cuencas analizadas.

Cuadro de error cuadrático medio por estación de análisis	
Estación meteorológica	Error cuadrático medio
M003	0.004
M364	0.000

M411	0.021
M031	0.1
M040	0.000
M185	0.32
M292	0.08

Estación hidrológica	Error cuadrático medio
H172	0.05
H173	0.004

A continuación, se presenta la tabla del error cuadrático medio obtenido mediante la máquina de aprendizaje supervisado Random Forest, de los modelos de imputación para las estaciones hidrometeorológicas de análisis.

Tabla 29. Resumen de error cuadrático medio calculado mediante machine Learning Random Forest para las estaciones hidrometeorológicas de las cuencas analizadas.

Cuadro de error cuadrático medio por estación de análisis	
Estación meteorológica	Error cuadrático medio
M003	0.004
M364	0.009
M411	0.001
M031	0.000
M040	0.06

M185	0.07
M292	0.025

Estación hidrológica	Error cuadrático medio
H172	0.001
H173	0.001

4.3.1. RESUMEN DE ERROR CUADRÁTICO MEDIO DE LOS MODELOS DE IMPUTACIÓN.

Para determinar el mejor método de imputación de datos de las estaciones de análisis se elaboró un resumen de los errores cuadráticos medios obtenidos por los diferentes modelos de imputación analizados, los mismos que fueron regresión lineal iterativo como método clásico, y las máquinas de aprendizaje autónomo, como lo fueron Linear Regression y Random Forest ejecutadas mediante lenguaje de programación Python.

Tabla 30. Resumen de error cuadrático medio calculado mediante machine Learning Random Forest, Linear Regression y regresión lineal iterativa como método clásico para las estaciones hidrometeorológicas de las cuencas analizadas.

Error cuadrático medio		M003	M364	M411	M031	M040	M185	M292	H172	H173	Promedio
Clásico	Regresión lineal	0.06	0.04	0.39	0.19	0.06	2.45	21.82	0.02	0.1	2.79
Machine Learning	Linear Regression	0	0	0.02	0.1	0	0.32	0.08	0.05	0	0.06
	Random Forest	0	0	0	0	0.06	0.07	0.02	0	0	0.01

Como es pude observar en la tabla (30) el mejor método predictivo es la máquina de aprendizaje autónomo basado en el módulo de aprendizaje supervisado Random Forest, seguida de la máquina de aprendizaje autónomo Linear Regression, ambas maquinas pertenecen a la librería Sklearn de machine learning.

4.3.1. METODOLOGÍA RECOMENDADA PARA LA IMPUTACIÓN DE DATOS HIDROMETEOROLÓGICOS.

Luego del análisis de diferentes métodos de imputación dentro de esta investigación para los datos hidrometeorológicos y con los resultados de la evaluación de los mismos, podemos inferir que el mejor método de imputación es el método computacional machine learning con el módulo Randon Forest, el cual permite tener resultados confiables con el menor error cuadrático respecto a los otros métodos evaluados en esta investigación.

A continuación se propone una metodología para llevar a cabo adecuadamente la imputación de datos hidrometeorológicos mediante machine learning usando el módulo Random Forest de la librería Scikit Learn:

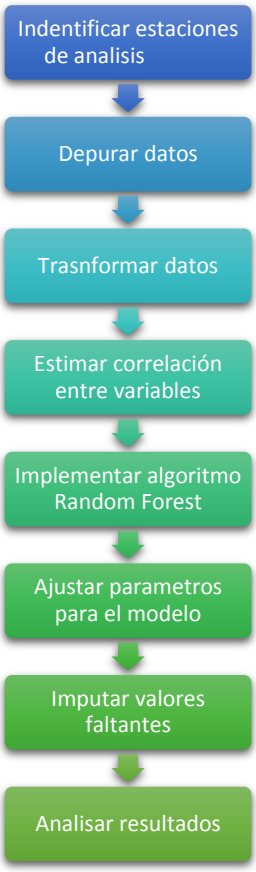


Ilustración 41: Diagrama de la metodología propuesta para la imputación de datos hidrometeorológicos mediante Machine Learning con el módulo Random Forest.

Explicación detallada de la metodología propuesta para la imputación de datos cantantes en los registros hidrometeorológicos:

1- Identificar las estaciones meteorológicas o hidrológicas muy cercanas a las estaciones de interés o a las estaciones a las que se les desea imputar los datos faltantes.
2- Depurar los datos hidrometeorológicos de la estación de la estación de análisis y la estación cercada de referencia, con la finalidad de tener la información de ambas estaciones en el mismo periodo de tiempo.
3- Transformar la información depurada en una matriz de datos, los mismos que deben contener una cabecera con el nombre de la estación y una columna identificando claramente las fechas de los registros hidrometeorológicos como se observa en la ilustración (41).
4- Estimar las correlaciones de las variables a analizar entre las estaciones para imputar y las estaciones cercanas de referencia.
5- Implementar el método de imputación de los datos faltantes, en este caso será el algoritmo computacional donde se utiliza el módulo Random Forest de la librería Scikit Learn para Python.
6- Ajustar los parámetros del modelo para encontrar los mejores resultados en cuanto al error cuadrático medio del conjunto de datos observados respecto a al conjunto de datos imputados.
7- Imputar valores faltantes de los registros hidrometeorológicos.
8- Analizar resultados.

4.4. DISCUSIÓN

Para lograr una imputación correcta de datos en cualquier ámbito, se debe aplicar un método de acuerdo a la característica de los datos (Patt, 2012), siguiendo este criterio se debe plantear que tipo de datos se tiene de entrada y con qué datos se los puede contrarrestar. Caso seguido escoger un método de imputación adecuada para estos los datos en cuestión.

En los métodos de imputación mediante procedimiento estadísticos clásicos unos de los métodos ampliamente usados y recomendados para la imputación de los datos hidrometereológicos es el método de regresión lineal simple como le describe Carrera Villacrés (2016) en su estudio “Relleno de series anuales de datos meteorológicos

mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media", esto también lo corrobora Luna & Lavado (2015) en su artículo "Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenca Jetepeque, Perú". En nuestro análisis se puede observar que el método de regresiones lineales convencional también permite tener resultados satisfactorios, tanto para las estaciones meteorológicas como para las estaciones hidrológicas, siempre y cuando las imputaciones se den entre estaciones que mantengan un coeficiente de correlación mayor a 0.75, permitiendo así tener validar el supuesto mayor de la linealidad de los datos como también lo reconoce Herrera (2017) en su estudio "Estimación de datos faltantes de precipitación por el método de regresión lineal: Caso de estudio Cuenca Guadalupe, Baja California, México"

Los métodos de imputación de datos de grandes volúmenes de información solo pueden ser tratados mediante procesos computacionales (O Reilly Medina, 2013), para ello se han creado algoritmos o máquinas de aprendizaje que permiten procesar la información y predecir la misma de forma autónoma (Michael, Andrianto, & Trafalis, 2015), es evidente que el procesamiento de la información por métodos computaciones permite trabajar con grandes volúmenes de información en menor tiempo que por métodos de evaluación distintos, lo cual se puedo evidenciar en el análisis de los datos para esta investigación.

Los métodos de imputación con máquinas de aprendizaje autónomo supervisado por medio de regresiones o clasificación, además, la inteligencia artificial permite soluciones eficaces a problemas donde plantear una solución adecuada es complicado. En estos casos como en los de esta investigación hallar la mejor solución al problema de imputación de datos hidrometeorológicos, solo nos brindará las máquinas de aprendizaje autónomo supervisado.

CAPÍTULO V

5. CONCLUSIONES

Para este estudio un punto de partida fue la selección de estaciones de análisis en las cuencas de los ríos Cañar, Jubones y Esmeraldas, con la particularidad de que las estaciones seleccionadas estén cercanas entre si y que la correlación entre ellas mantenga un índice superior a 0.75 con lo cual garantiza el supuesto de linealidad esto permite que los modelos de imputación sean significativos. Toldas las estaciones analizadas e imputadas mediante dos diversos métodos de esta investigación mantienen una correlación lineal entre ellas superior al índice indicado.

El método de imputación determinístico como la regresión lineal en este estudio obtuvo un promedio en el error cuadrático medio de 2.79 en los datos imputados respecto a los datos observados en las estaciones de análisis. Aunque, es evidente que en la mayoría de las estaciones el error cuadrático medio es menor a 1, esto es un reflejo de que los modelos lineales se ajustan muy bien para la imputación de datos faltantes. Por otra parte, existe un valor demasiado alto en el error cuadrático medio de 21.82 en la estación a análisis meteorológica M0292, esto también se evidencia en las gráficas de la validación de los supuestos para esta estación, donde los datos no siguen un patrón normal, lo cual afecta directamente al promedio general de error cuadrático medio con este método de imputación.

Los métodos computacionales basados en máquinas de aprendizaje autónomo respecto a los métodos determinísticos de imputación de datos hidrometeorológicos reflejan en el análisis del promedio general del error cuadrático medio que son superiores al momento de imputar los datos. Es decir, el conjunto de datos imputados respecto a al conjunto de datos observados mantiene una menor variabilidad en los modelos de imputación basados en máquinas de aprendizaje autónomo supervisado. En definitiva, el mejor método es aquel minimiza el error medio cuadrático en las predicciones de prueba, esto garantiza que los datos a imputar sean más confiables.

A su vez, la evaluación de las máquinas de aprendizaje autónomo supervisado, *Linear Regressor* y *Random Forest* nos permiten concluir que la mejor máquina de aprendizaje autónomo supervisado para la imputación de datos hidrometeorológicos

en todas las estaciones de análisis es la maquina *Random Forest* la cual obtuvo un promedio general en el error cuadrático medio de 0.01, esto en términos generales permite resumir que prácticamente no existe variabilidad entre el conjunto de datos imputados respecto al conjunto los datos observados.

Finalmente, se concluye que obtener registros completos hidrometeorológicos a través de la implementación de un algoritmo computacional, permite tener una la información meteorológica e hidrológica valida. Esto otorgará a futuros a estudios sobre el medio ambiente, cambio climático y estudios de relevancia sobre el recurso hídrico tener los insumos necesarios para crear modelos que reflejen con mayor precisión la realidad de nuestro territorio y de sus recursos naturales.

CAPÍTULO VI

6. RECOMENDACIONES

Para futuros estudios de imputación en las estaciones hidrológicas se recomienda efectuar un análisis morfométrico de las microcuencas donde se encuentre estaciones de monitoreo hídrico, con la finalidad de determinar las estaciones que compartan condiciones topo climáticas y morfométricas similares, permitiendo así identificar estaciones que mantengan correlaciones lineales fuertes en los registros de caudales. Por otra parte, no se recomienda la selección aleatoria de estaciones hidrológicas, aunque cercanas estas estaciones generalmente no presentan correlación alguna.

Para la selección de estaciones meteorológicas se recomienda, identificar las estaciones más cercanas posibles y que la altura en la que se encuentran ubicadas sea similar, esto garantizará un coeficiente de correlación lineal alto que permitirá en lo posterior obtener modelos matemáticos de imputación significativos.

Estas metodologías de imputación de datos hidrometeorológicos sirven exclusivamente para interpolar datos y no se han puesto a prueba para predicciones futuras para lo cual se recomienda en estos casos fusionar estas técnicas con teorías de series temporales.

Luego de la evaluación de los métodos de imputación de esta investigación, se recomienda emplear el algoritmo computacional donde se incluya la máquina de aprendizaje autónomo supervisado *Random Forest* para imputar los datos con un el menor error cuadrático medio por estación.

REFERENCIAS BIBLIOGRÁFICAS

Bateman-, A., Camacho, A., Toro, M., Elosegi, A., Sabater, S., Ollero, A., & GAMA. (2007). *Hidrologia basica y aplicada*. *Grupo de Investigación en Transporte de Sedimetos*. Retrieved from http://www.floodup.ub.edu/hidro/%5Cnhttp://www.magrama.gob.es/es/biodiversidad/temas/espacios-protegidos/red-natura-2000/rn_fichas_be_agua_dulce.aspx

Bello, M., & Pino, M. (2000). *Medición de Presion y caudal*. *Inia*. Punta Arenas. https://doi.org/ISSN 0717-4829

Bennett, N. D., Newham, L. T. H., Croke, B. F. W., & Jakeman, A. J. (2007). Patching and Disaccumulation of Rainfall Data for Hydrological Modelling. *Int. Congress on Modelling and Simulation (MODSIM 2007)*, 2520–2526.

Breiman, L. E. O. (2001). Random Forest（LeoBreiman）.pdf, 5–32. https://doi.org/10.1023/A:1010933404324

Campozano, L., Sánchez, E., Aviles, A., & Samaniego, E. (2014). Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes. *Maskana*, *5*(1), 99–115. Retrieved from http://dspace.ucuenca.edu.ec:8080/handle/123456789/5586

Carrera Villacrés, D. V., Guevara García, P. V., Tamayo Bacacela, L. C., Balarezo Aguilar, A. L., Narváez Rivera, C. A., & 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)*, *34*(3), 81–90. https://doi.org/10.4067/S0718-34292016000300010

Goicoechea, A. P. (2002). Imputación basada en árboles de clasificación.

Gómez García, J., Palarea Albaladejo, J., & Martín Fernández, J. (2006). Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. *Estadistica Española*, *48*, 241–270.

Herrera, C. S., Campos, J., & Carrillo, F. (2017). Estimación de datos faltantes de precipitación por el método de regresión lineal: Caso de estudio Cuenca Guadalupe, Baja California, México. *Investigacion Y Ciencia*, *28*, 34–44. Retrieved from http://www.redalyc.org/pdf/674/67452917005.pdf

J. Smola, A., & S.V.N., V. (2008). *Introduction to machine learning*. *Methods in*

molecular biology (Clifton, N.J.) (Vol. 1). https://doi.org/10.1007/978-1-62703-748-8_7

Luna, E., & Lavado, W. (2015). Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenca Jetepeque, Perú. *Revista Tecnológica ESPOL – RTE*, *28*(3), 42–52. https://doi.org/10.1089/ees.2013.0409

Medina, R. (2008). ESTIMACIÓN ESTADÍSTICA DE VALORES FALTANTES EN SERIES HISTÓRICAS DE LLUVIA. Pereira.

Medina, R. D., Montoya, E. C., & Jaramillo, Á. (2008). Estimación estadística de valores faltantes en series históricas de lluvia. *Cenicafé*, *59*(3), 260–273. https://doi.org/10.1017/CBO9781107415324.004

Michael, R., Andrianto, I., & Trafalis, T. B. (2015). Missing data imputation through machine learning algorithms. *ResearchGate*, *1*(January). https://doi.org/10.1007/978-1-4020-9119-3

O Reilly Medina, I. (2013). *Big data now*. (J. Webb & T. O Brien, Eds.).

OMM. (2011). *Guía de prácticas climatológicas-N° 100*. https://doi.org/OMM-N^{o} 168

Patt, G. (2012). *Imputación de datos faltantes : una aplicación del método de regresión lineal en las estaciones agrometeorológicas del Valle del Mayo*. Universidad de Sonora.

Rodríguez Jimenez, R. M., Capa, Á. B., & Portela Lozano, A. (2004). *Meteorología y Climatología*. https://doi.org/M-XXXXX-2004

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambrige University Press. https://doi.org/10.1017/CBO9781107298019

Urrutia, J. A., Palomino, R., & Salazar, H. D. (2010). Metodología para la imputación de datos faltantes en metereología. *Sciencia et Technica*, *17*(46), 44–49.

Velázquez, J. E. (2014). *Calibración de un modelo de estimación de lluvia con imágenes de satélite, utilizando datos de estaciones climatológicas, para la región hidrológica número 30 de México*.

Vera, L. E. (2012). Análisis de aforo de la Estación Hidrométrica periodo 2000-2001 ., 34. https://doi.org/10.1017/CBO9781107415324.004

Walpole, R., Myers, R., Myers, S., & Keying, Y. (2012). *Probabilidad y estadística para ingeniería y ciencias*. *Journal of Chemical Information and Modeling* (Novena edi, Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

ANEXOS

ANEXO 1

Código para imputación de datos hidrometeorológicos mediante machine Learning empleando Linear Regression Multiple.

```
from pandas import DataFrame
import pandas as pd
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
import numpy as np
from sklearn import metrics
from sklearn.cross_validation import train_test_split
from sklearn.model_selection import cross_val_score

def convertir_lista(array):
    n = len(array)
    l = []
    for i in range(n):
        l.append(float(array[i]))
    return l

    def fun(b0, b1, const, x1, x2):
        return (b0*x1 + b1*x2 + const)

    def agregar_linea(fecha, v1, v2, v3, nombre_destino):
        f = open(nombre_destino,"a")
        cadena = str(fecha)+","+str(v1)+","+str(v2)+ "," +str(v3) +"\n"
        f.write(cadena)
        f.close()

    def cargar_informacion(nombre):
        data = pd.read_csv(nombre)
        data.head()
        return data

    def crear_archivo_csv(nombre_destino, v_x, v_y, v_z):
        cadena = "fecha"+","+v_x+","+ v_y + "," + v_z +"\n"
        f = open(nombre_destino,"w")
        f.write(cadena)
        f.close()

    def rellenar_datos_ausentes(b0, b1, const,nombre, nombre_destino,
                                variable_x1, variable_x2, variable_x3):
        data = cargar_informacion(nombre)
        fecha = data['fecha'].values
        X1 = data[variable x1].values
        X2 = data[variable_x2].values
        X3 = data[variable_x3].values
        n = len(X1)
        crear_archivo_csv(nombre_destino, variable_x1, variable_x2, variable_x3)
        for i in range(n):
            if (np.isnan(X1[i]) == False) and (np.isnan(X2[i]) == False):
                z = fun(b0, b1, const, X1[i], X2[i])
            elif(np.isnan(X1[i]) == False) and (np.isnan(X2[i]) == True) and (np.isna
    n(X3[i]) == False):
                z = X3[i]
            elif (np.isnan(X2[i]) == False):
                z = X3[i]
            else:
                z = np.NaN
            agregar_linea(fecha[i], X1[i], X2[i], z, nombre_destino)

```

```
141.    def calcular_promedio_col(columna, nombre_archivo):
142.        data = pd.read_csv(nombre_archivo)
143.        data.head()
144.        y = data[columna].values
145.        n = len(y)
146.        suma = 0
147.        c = 0
148.        for i in range(n):
149.            if not(np.isnan(y[i]) == False):
150.                suma += 0
151.            else:
152.                c = c + 1
153.                suma += float(y[i].item(0))
154.        promedio = suma / c
155.        print("datos tomados:",c)
156.        return promedio
157.
158.
159.
160.
161.    def funcion_minimos_cuadrados(nombre, v_x1, v_x2, v_x3):
162.        slr_df = pd.read_csv(nombre, sep=",")
163.        slr_df.head()
164.        slr_df = slr_df.dropna()
165.        x1 = slr df[v x1].values.reshape(-1,1)
166.        x2 = slr df[v x2].values.reshape(-1,1)
167.        x3 = slr df[v x3].values.reshape(-1,1)
168.        X1 = convertir_lista(x1)
169.        X2 = convertir_lista(x2)
170.        X3 = convertir_lista(x3)
171.        diccionario = {v_x1:X1, v_x2:X2, v_x3:X3}
172.        df = DataFrame(diccionario,columns=[v x1,v x2,v x3])
173.        X_sol = df[[v_x1,v_x2]] # here we have 2 variables for multiple regression.
    If you just want to use one variable for simple linear regression, then use X =
    df['Interest_Rate'] for example.Alternatively, you may add additional variables
    within the brackets
174.        Y sol = df[v x3]
175.        # with sklearn
176.        X_sol_train, X_sol_test, Y_sol_train, Y_sol_test = train_test_split(X_sol, Y_
    sol, test_size=0.16, random_state=5)
177.        print ('varaibles de entremamiento: ', X_sol_train.shape)
178.        print ('caracteristicas de entrenamiento:', Y_sol_train.shape)
179.        print ('varaibles de prueba: ', X_sol_test.shape)
180.        print ('caracteristicas de prueba: ', Y_sol_test.shape)
181.        regr = LinearRegression()
182.        regr.fit(X_sol_train, Y_sol_train)
183.        predictions = regr.predict(X_sol_test)
184.        print('MAE: ', metrics.mean_absolute_error(Y_sol_test, predictions))
185.        print ('r2_score', metrics.r2_score(Y_sol_test, predictions))
186.        print('Intercept: \n', regr.intercept_)
187.        print('Coefficients: \n', regr.coef_)
188.        print ("validacion
    cruzada:",cross_val_score(regr,X_sol_train,Y_sol_train,cv=10))
189.
190.        # with statsmodels
191.        X_sol = sm.add_constant(X_sol) # adding a constant
192.        model = sm.OLS(Y_sol, X_sol).fit()
193.        predictions = model.predict(X_sol)
194.        print model = model.summary()
195.        print(print_model)
196.        constante = regr.intercept_
197.        b0  = regr.coef_[0]
198.        b1  = regr.coef_[1]
199.        l =[ b0, b1, constante]
200.        return l
201.
202.
203.    def calcular_error(nombre,nombre_destino, variable_y):
204.        prom_orig = calcular_promedio_col(variable_y,nombre)
205.        print("promedio original: ", prom_orig)
206.        prom copy = calcular promedio col(variable y, nombre destino)
207.        print("promedio copia: ", prom_copy)
208.        total = ((prom_orig - prom_copy)**2)
209.        print("error: ",total)
210.
211.
212.    variable_x1 = "M0292"
```

```
    variable_x2 = "M0185"
    variable_y = "M0040"
    nombre archivo = "jubones.csv"
    nombre copia = "copia multiple.csv"

    lista = funcion_minimos_cuadrados(nombre_archivo,variable_x1, variable_x2, variable_y )
    rellenar_datos_ausentes(lista[0], lista[1], lista[2], nombre_archivo, nombre_copia, variable x1,
                            variable x2, variable y)
    calcular_error(nombre_archivo, nombre_copia, variable_y)
```

ANEXO 2

Código para imputación de datos hidrometeorológicos mediante machine Learning empleando Linear Regression simple.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.cross_validation import train_test_split
from sklearn.model_selection import cross_val_score


def fun(b0, b1, x):
    return (b0 + b1*x)

def fun_x(lista,b0, b1):
    l = []
    for x in lista:
        n = float(fun(b0,b1,x))
        l.append(n)
    return l

def calcular_promedio_col(columna, nombre_archivo):
    data = pd.read_csv(nombre_archivo)
    data.head()
    y = data[columna].values
    n = len(y)
    suma = 0
    c = 0
    for i in range(n):
        if not(np.isnan(y[i]) == False):
            suma += 0
        else:
            c = c + 1
            suma += float(y[i].item(0))
    promedio = suma / c
    print("datos tomados:",c)
    return promedio



def agregar_linea(fecha, v1, v2, nombre_destino):
    f = open(nombre_destino,"a")
    cadena = str(fecha)+","+str(v1)+","+str(v2)+ "\n"
    f.write(cadena)
    f.close()

def crear_archivo_csv(nombre,nombre_destino, variable_x, variable_y):
    data = cargar informacion(nombre)
    l = list(data)
    cadena = l[0]+","+variable x+","+ variable y +"\n"
    f = open(nombre_destino,"w")
    f.write(cadena)
    f.close()


def cargar_informacion_y(nombre):
    data = pd.read_csv(nombre)
```

```
    data.head()
    data.drop(pd.isnull(data).any(1).nonzero()[0],inplace = True)
    return data['M031'].values

def cargar_informacion(nombre):
    data = pd.read_csv(nombre)
    data.head()
    return data

def fun(b0, b1, x):
    return (b0 + b1*x)

def rellenar_datos_ausentes(b0, b1, nombre,nombre_destino,variable_x, variable_y):
    data = cargar informacion(nombre)
    fecha = data['fecha'].values
    X = data[variable_x].values
    Y = data[variable_y].values
    n = len(X)
    crear_archivo_csv(nombre,nombre_destino, variable_x, variable_y)
        for i in range(n):
            if (np.isnan(X[i]) == False) and (np.isnan(Y[i]) == False):
                y = fun(b0,b1,X[i])
            elif (np.isnan(Y[i]) == True) and not(np.isnan(X[i]==False)):
                y = fun(b0,b1,X[i])
            else:
                y = Y[i]
            agregar linea(fecha[i], X[i], y, nombre destino)

    def machine_learning(nombre, variable_x, variable_y):
        slr_df = pd.read_csv(nombre, sep=",")
        slr_df.head()
        slr df = slr df.dropna()
        x = slr_df[variable_x].values.reshape(-1,1)
        y = slr_df[variable_y].values.reshape(-1,1)
        X_train, X_test, y_train, y_test = train_test_split(x, y , test_size=0.30, ra
    ndom_state=10)
        lm = LinearRegression()
        lm.fit(X train, y train)
        slope = lm.coef_
        predictions = lm.predict(X_test)
        sns.distplot((y_test - predictions), bins = 1)
        print ('La recta de regresión es: y = %f + %f * x'%(lm.intercept_, slope))
        func x=[lm.intercept , slope]
        valores_y = fun_x(x, lm.intercept_, slope)
        #plt.plot(x,valores_y )
        plt.scatter(y_test, predictions,  c='r', edgecolors=(0, 0, 0), alpha=0.5)
        plt.plot(x,valores_y, c = 'b')
        plt.title(' vs test values', fontsize=10)
        plt.xlabel('test values')
        plt.ylabel('predicted values')
        plt.show()
        print ("validacion cruzada:",cross_val_score(lm,X_train,y_train,cv=5))
        print('Evaluacion de metodo',lm.score(X_test,y_test))
        return func_x

    def calcular_error(nombre,nombre_destino, variable_y):
        prom_orig = calcular_promedio_col(variable_y,nombre)
        print("promedio original: ", prom_orig)
        prom copy = calcular promedio col(variable y, nombre destino)
        print("promedio copia: ", prom_copy)
        total = ((prom_orig - prom_copy)**2)
        print("error cuadratico medio: ",total)


    nombre = "ESMERALDAS.csv"
    nombre copia = "copia1.csv"
    variable_x ="M0364"
    variable_y = "M0003"
    #nombre = str(input("ingrese el nombre del archivo: "))
    #variable x = str(input("Ingrese el parametro de la variable x:"))
    #variable y = str(input("Ingrese el parametro de la variable y:"))
    #ignorado = str(input("Ingrese la variable a ignorar:"))
    l=machine_learning(nombre, variable_x, variable_y)
    rellenar_datos_ausentes(float(l[0]),float(l[1]), nombre, nombre_copia,
                            variable_x, variable_y)
    #calcular error("archivo.csv", "copia1.csv")
```

```
157.
158.    calcular_error(nombre,nombre_copia, variable_y)
```

ANEXO 3

Código para imputación de datos hidrometeorológicos mediante machine Learning empleando Random Forest

```
7.  # Random Forest Regression
8.
9.  # Importing the libraries
10. import numpy as np
11. import matplotlib.pyplot as plt
12. import pandas as pd
13. from sklearn.ensemble import RandomForestRegressor
14. from sklearn.cross_validation import train_test_split
15. from sklearn.model_selection import cross_val_score
16.
17. def eliminar_vacios(lista):
18.     lista nueva=[]
19.     for i in lista:
20.         if not(np.isnan(i)):
21.             lista_nueva.append(i)
22.     return lista_nueva
23.
24.
25. def calcular_promedio_col(columna, nombre_archivo):
26.     data = pd.read_csv(nombre_archivo)
27.     data.head()
28.     y = data[columna].values
29.     n = len(y)
30.     suma = 0
31.     c = 0
32.     for i in range(n):
33.         if (y[i] == "[nan]"):
34.             suma += 0
35.         elif (str(y[i]) == "nan"):
36.             suma += 0
37.         else:
38.             c = c + 1
39.
40.             suma += float(str(y[i]))
41.     promedio = suma / c
42.     return promedio
43.
44.
45. def cargar_informacion(nombre):
46.     data = pd.read csv(nombre)
47.     data.head()
48.     return data
49.
50. def agregar_linea(fecha, v1, v2, nombre_destino):
51.     f = open(nombre destino,"a")
52.     cadena = str(fecha)+","+str(v1)+","+str(v2)+ "\n"
53.     f.write(cadena)
54.     f.close()
55.
56. def crear_archivo_csv(nombre_archivo,nombre_destino, variable_x, variable_y):
57.     data = cargar informacion(nombre archivo)
58.     l = list(data)
59.     cadena = l[0]+","+variable_x+","+ variable_y +"\n"
60.     f = open(nombre_destino,"w")
61.     f.write(cadena)
62.     f.close()
63.
64. def calcular_error(nombre,nombre_destino, variable_y):
65.     prom_orig = calcular_promedio_col(variable_y,nombre)
66.     print("promedio original: ", prom_orig)
67.     prom_copy = calcular_promedio_col(variable_y, nombre_destino)
68.     print("promedio copia: ", prom_copy)
69.     total = ((prom orig - prom copy)**2)
70.     print("error: ",total)
```

```
71.
72.
73. def rellenar_datos_ausentes(nombre_archivo, nombre_destino,variable_x, variable_y):
74.     data = cargar informacion(nombre archivo)
75.     fecha = data['fecha'].values
76.     X = data[variable_x].values
77.     Y = data[variable_y].values
78.     n = len(X)
79.     crear archivo csv(nombre archivo, nombre destino, variable x, variable y)
80.     for i in range(n):
81.         agregar_linea(fecha[i], X[i], Y, nombre_destino)
82.
83. def forrest(nombre_archivo, nombre_destino, variable_x, variable_y):
84.     # Importing the dataset
85.     data = pd.read csv(nombre archivo)
86.     x_ori = data[variable_x].values
87.     y_ori = data[variable_y].values
88.     dataset = pd.read_csv(nombre_archivo)
89.     dataset.dropna
90.     dataset = dataset.dropna()
91.     X = dataset[variable x].values.reshape(-1,1)
92.     y = dataset[variable_y].values.reshape(-1,1)
93.     # Splitting the dataset into the Training set and Test set
94.     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.28, rand
    om state = 5)
95.     # Fitting Random Forest Regression to the dataset
96.     regressor = RandomForestRegressor(n estimators = 100, random state = 0)
97.     regressor.fit(X_train, y_train.ravel())
98.
99.     # Visualising the Random Forest Regression results (higher resolution)
100.
101.        min x = min(X)
102.        max_x = max(X)
103.        intervalo = (max_x - min_x)/len(X)
104.
105.        X_grid = np.arange(min(X), max(X), intervalo)
106.        X grid = X grid.reshape((len(X grid), 1))
107.        plt.scatter(X, y, color = 'red')
108.        plt.plot(X_grid, regressor.predict(X_grid), color = 'blue')
109.        plt.title('(Random Forest Regression)')
110.        plt.xlabel('Estacion de referencia')
111.        plt.ylabel('Estacion de analisis')
112.        plt.show()
113.        #aqui predice con los nuevos datos
114.        x_sin_nan = np.asarray(eliminar_vacios(x_ori)).reshape(-1,1)
115.        print("los nuevos valores sin nan son:",len(x_sin_nan))
116.        lr =regressor.predict(x_sin_nan)
117.        print("los valores predecido:",len(lr))
118.        print("dataos predichos:",len(lr))
119.        print("datos para entrenar: ",len(x_sin_nan))
120.        print("Datos existentes",len(x_ori))
121.        y_train = np.array(y_train.ravel())
122.        print ("validacion
    cruzada:",cross val score(regressor,X train,y train,cv=5,))
123.        print('Evaluacion de metodo',regressor.score(X_test,y_test))
124.        crear_archivo_csv(nombre_archivo, nombre_destino, variable_x, variable_y)
125.        #para agregar la informacion predecida
126.        fecha = data['fecha'].values
127.        cont = 0
128.        for i in range(0, len(x_ori)):
129.
130.            #rellena si los dos estan llenos
131.            if (np.isnan(x_ori[i]) == False) and (np.isnan(y_ori[i]) == False):
132.
133.                y = lr[cont]
134.                cont += 1
135.            #rellena si x exite  y 'y' no
136.            elif(np.isnan(x_ori[i])==False) and (np.isnan(y_ori[i])==True):
137.
138.                y = lr[cont]
139.                cont += 1
140.             #cuando x y 'y' no existen
141.            elif ((np.isnan(x_ori[i]) == True) and (np.isnan(y_ori[i]) == True)):
142.
143.                y = y_ori[i]
144.                #cuando x no existe y "y" si
145.            elif((np.isnan(x_ori[i]) == True) and (np.isnan(y_ori[i]) == False)):
```

```
146.
147.                 y = y_ori[i]
148.             agregar linea(fecha[i], x ori[i], y, nombre destino)
149.
150.     nombre = "ESMERALDAS.csv"
151.     nombre_destino= "copia_2.csv"
152.     v_x = "M0003"
153.     v_y = "M0364"
154.     forrest(nombre, nombre destino, v x, v y)
155.     calcular_error(nombre, nombre_destino, v_y)
```

PERMISO DEL AUTOR DE TESIS PARA SUBIR AL REPOSITORIO INSTITUCIONAL

Yo, **Paúl Eduardo Vásquez Álvarez** portador(a) de la cédula de ciudadanía Nº 0106558679. En calidad de autor/a y titular de los derechos patrimoniales del trabajo de titulación **"IMPUTACION DE DATOS FALTANTES DE REGISTROS HIDROMETEOROLOGICOS DE LAS CUENCAS DE LOS RIOS JUBONES, CAÑAR Y ESMERALDAS MEDIANTE METODOS ESTADISTICOS Y MACHINE LEARNING"** de conformidad a lo establecido en el artículo 114 Código Orgánico de la Economía Social de los Conocimientos, Creatividad e Innovación, reconozco a favor de la Universidad Católica de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos, Así mismo; autorizo a la Universidad para que realice la publicación de éste trabajo de titulación en el Repositorio Institucional de conformidad a lo dispuesto en el artículo 144 de la Ley Orgánica de Educación Superior.

Cuenca, 11 de febrero de 2019

Autor
Paúl Eduardo Vásquez Álvarez
C.I. 0106558679