UNIVERSIDAD NACIONAL AGRARIA LA MOLINA - ESCUELA DE POSGRADO

DOCTORADO EN INGENIERÍA Y CIENCIAS AMBIENTALES



DISEÑO Y ANÁLISIS DE EXPERIMENTOS EN INGENIERÍA Y CIENCIAS AMBIENTALES

**Actividad:** Trabajo final encargado de teoría

Docente:

Ph.D. Christian René Encina Zelada

Presenta:

José Augusto Zevallos Ruiz

Lima – Perú

10 de diciembre del 2024

# INTRODUCCIÓN

La región de Piura, en la costa norte de Perú, enfrenta desafíos significativos asociados a eventos de precipitaciones extremas vinculados al fenómeno El Niño. Estudios previos han demostrado una relación consistente entre las precipitaciones intensas y las anomalías positivas en la temperatura de la superficie del mar (SST), que favorecen la formación de sistemas convectivos al modificar las condiciones en la capa límite planetaria (Tapley & Waylen, 1990; Waylen & Caviedes, 1986). Durante estos eventos, las precipitaciones acumuladas en Piura pueden ser hasta treinta veces superiores a lo normal, resultado de incursiones anómalas hacia el polo de la Corriente del Niño que generan inestabilidad atmosférica extendida y condiciones propicias para el ascenso del aire, especialmente sobre pendientes montañosas cercanas (Takahashi, 2004). Estos fenómenos meteorológicos incrementan significativamente el riesgo de inundaciones, lo que subraya la necesidad de herramientas predictivas robustas para la gestión del riesgo.

En el contexto del cambio climático, el aumento en la frecuencia e intensidad de eventos extremos refuerza la urgencia de desarrollar modelos predictivos eficaces que apoyen la planificación y la mitigación de riesgos. Tradicionalmente, se han empleado modelos hidráulicos basados en las ecuaciones de aguas poco profundas, los cuales han sido aplicados para simular inundaciones históricas en Piura, como las asociadas al evento El Niño de 2017, y proyectar escenarios futuros de períodos de retorno hasta 500 años (Álvarez et al., 2021; Muñoz et al., 2022). Aunque útiles, estas metodologías tienden a enfocarse en evaluaciones posteriores a los eventos, dejando aspectos como el llenado de datos faltantes sin explorar completamente.

Por otro lado, los modelos hidrológicos han mostrado avances en la predicción a corto plazo gracias a la integración de datos satelitales, mejorando la precisión en eventos de inundaciones en regiones montañosas como los Andes tropicales (Llauca et al., 2023). Asimismo, sistemas globales como el GloFAS han demostrado ser efectivos para pronosticar inundaciones en cuencas con grandes áreas de captación, alcanzando una precisión del 65% en ríos peruanos (Bischiniotis et al., 2019). Sin embargo, estas herramientas suelen depender de conjuntos de datos completos, lo que representa un desafío en áreas donde las estaciones de monitoreo presentan discontinuidades significativas en sus registros.

En este contexto, las técnicas de regresión múltiple ofrecen una alternativa prometedora para abordar el problema del llenado de datos faltantes, particularmente en zonas con baja densidad de estaciones meteorológicas. La regresión múltiple permite relacionar las precipitaciones en estaciones cercanas mediante la integración de variables predictoras, como la elevación, latitud y longitud, logrando estimaciones precisas a nivel local (Naoum & Tsanis, 2004). Por ejemplo, estudios realizados en la isla de Creta han demostrado cómo la incorporación de estas variables mejora significativamente las predicciones de precipitación en terrenos complejos y con limitadas estaciones de monitoreo.

Además, la aplicación de modelos de regresión múltiple ha sido exitosa en diferentes regiones, como en la isla de Creta, donde se lograron estimaciones robustas de la precipitación media anual al incluir parámetros orográficos y geográficos (Naoum & Tsanis, 2004). Estas metodologías no solo capturan la variabilidad espacial de las precipitaciones, sino que también son aplicables en diferentes escalas, desde cuencas hasta áreas más extensas, lo que resalta su versatilidad.

El presente estudio tiene como objetivo principal utilizar modelos de regresión múltiple para llenar vacíos en datos de precipitación diaria en la cuenca del río Piura. Para ello, se utilizarán datos de cinco estaciones meteorológicas: Chusis, Chalaco, Huamarca, Huancabamba y Miraflores, aprovechando la información de las estaciones vecinas para realizar las estimaciones. Este enfoque busca mejorar la disponibilidad de datos en la región, lo que es esencial para la modelación hidrológica y la planificación de recursos hídricos.

La metodología planteada permitirá no solo abordar la problemática de datos faltantes, sino también generar insumos que puedan integrarse a modelos hidrológicos e hidráulicos para la gestión del riesgo de inundaciones en Piura. Así, se busca contribuir al fortalecimiento de herramientas de predicción, esenciales para mitigar los impactos de los eventos extremos que afectan recurrentemente a esta región.

# MARCO TEÓRICO

**Regresión lineal múltiple**

La regresión múltiple es una técnica estadística utilizada para modelar la relación entre una variable dependiente y varias variables independientes . Su objetivo es predecir o explicar su variabilidad basándose en las variables predictoras. La ecuación general de un modelo de regresión múltiple es:

Donde:

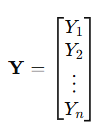
* : Variable dependiente.
* ​: Variables independientes o predictoras.
* ​: Intercepto del modelo.
* Coeficientes de regresión que representan el cambio esperado en por unidad de cambio en cada ​, manteniendo las demás constantes.
* : Término de error o residual, que captura la variabilidad no explicada por el modelo.

**Forma Matricial del Modelo**

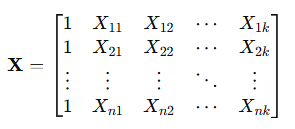
El modelo de regresión múltiple puede expresarse en términos matriciales como:

Donde:

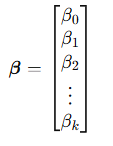
* es el vector de observaciones de la variable independiente .



* es la matriz de diseño , que incluye una columna de unos para el termino de intercepto y las observaciones de las variables independientes



* es el vector de coeficientes :



* es el vector de errores :

**Estimación de los Coeficientes**

Los coeficientes del modelo se estiman minimizando la suma de los cuadrados de los errores , siguiendo el método de los mínimos cuadrados ordinarios (OLS). La estimación matricial de se obtiene como:

Donde:

* es la transpuesta de la matriz de diseño.
* es la matriz inversa del producto.
* es el producto de la transpuesta de con .

El vector contiene los coeficientes estimados del modelo

**Prueba Kolmogórov-Smirnov**

La prueba de Kolmogórov-Smirnov es una prueba no paramétrica que evalúa si una muestra sigue una distribución específica. Es utilizada para verificar la normalidad de los datos en un análisis ANOVA. La estadística de prueba DDD para esta prueba es:

Donde:

* es la función de distribución empírica de la muestra,
* es la función de distribución acumulada teórica.

Si el valor de D es suficientemente grande, se rechaza la hipótesis nula de que los datos siguen la distribución especificada.

**Prueba de Durbin-Watson**

La prueba de Durbin-Watson es una prueba estadística utilizada para detectar la autocorrelación en los residuos de un modelo de regresión. Su estadística de prueba es:

Donde:

* son los residuos,
* es el número de observaciones

Un valor cercano a 2 indica que no hay autocorrelación en los residuos, mientras que valores alejados de 2 sugieren la presencia de autocorrelación.

**Error cuadrático medio**

Es un estimador que mide el promedio de los errores al cuadrado. El error es la diferencia entre el valor estimado por un modelo y el valor real medido de la variable. Esta diferencia se debe a que el modelo generalmente no captura toda la información necesaria para reproducir la realidad o a la existencia de errores de medición aleatoria (ver Ecuación).

# METODOLOGÍA

**4.1. Metodología empleada**

El flujograma representa el proceso de análisis y modelado de datos faltantes mediante regresión múltiple, comenzando con la carga de datos y configuración del entorno de trabajo en herramientas como R o Python. A continuación, se preparan y verifican los datos, identificando columnas con valores faltantes y filtrando estaciones según un umbral predefinido de datos incompletos. Luego, el flujo pasa a iterar sobre los datos faltantes, entrenar modelos de regresión múltiple y realizar predicciones para llenar los vacíos, calculando el RMSE como métrica de desempeño. En cada iteración, se evalúa si hubo mejora; si no, el proceso finaliza y retorna los datos completados, las exclusiones y los errores calculados. Finalmente, se generan gráficos para comparar series originales y completadas, así como un análisis de correlación entre estaciones mediante un mapa de calor.

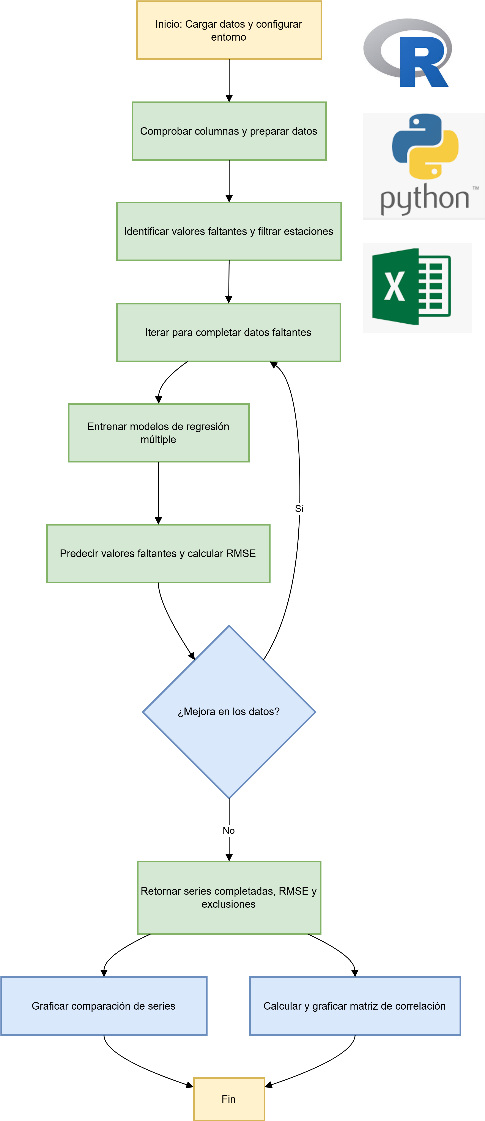


Figura 1 Flujograma metodológico empleado en el presente trabajo

**4.3. Datos**

El conjunto de datos registra las precipitaciones mensuales acumuladas en cinco estaciones meteorológicas (Chusis, Chalaco, Huamarca, Huancabamba y Miraflores) desde enero de 1980 hasta diciembre de 2017. Cada fila representa el total mensual de precipitaciones en milímetros (mm) para cada estación, con una columna adicional que indica la fecha correspondiente al último día del mes.

Los datos muestran variaciones significativas entre estaciones y períodos, con valores mínimos de 0 mm durante meses secos y máximos superiores a 700 mm en eventos extremos, como durante fenómenos de El Niño. También se observan valores faltantes (NA), que son relevantes para análisis de imputación o modelado. Este conjunto de datos es clave para estudiar la estacionalidad y los patrones de lluvia en la región, así como para desarrollar modelos predictivos relacionados con eventos climáticos.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fecha** | **chusis** | **chalaco** | **huamarca** | **huancabamba** | **miraflores** |
| 1980-01-31 | 0 | 70.4 | 38.7 | 23.5 | 0.2 |
| 1980-02-29 | 0 | 108.7 | 139 | 35.9 | 2.5 |
| 1980-03-31 | 6.4 | 121.6 | 116.9 | 51.3 | 13.7 |
| 1980-04-30 | 7.6 | 178 | 129.1 |  | 35.2 |
| 1980-05-31 | 0 | 42.5 | 17.9 | 22.1 | 0.3 |
| 1980-06-30 | 0 | 0 | 2.3 | 8.7 | 0 |
| 1980-07-31 | 0 | 0 | 0.1 | 2.9 | 0 |
| 1980-08-31 | 0 | 0 | 1.1 | 2.8 | 0 |
| 1980-09-30 | 0 | 0 | 0 | 11.2 | 0 |
| 1980-10-31 | 5.2 | 36.7 | 62.2 | 49.9 | 0.4 |
| 1980-11-30 | 4.1 | 40 | 33.3 | 65.6 | 5.6 |

# RESULTADOS Y DISCUSIONES

**Prueba Kolmogórov-Smirnov**

Para el test de normalidad se tiene = distribución normal muestral, = distribución teórica Weibull.

De la siguiente expresión se obtuvo que:

El valor de :

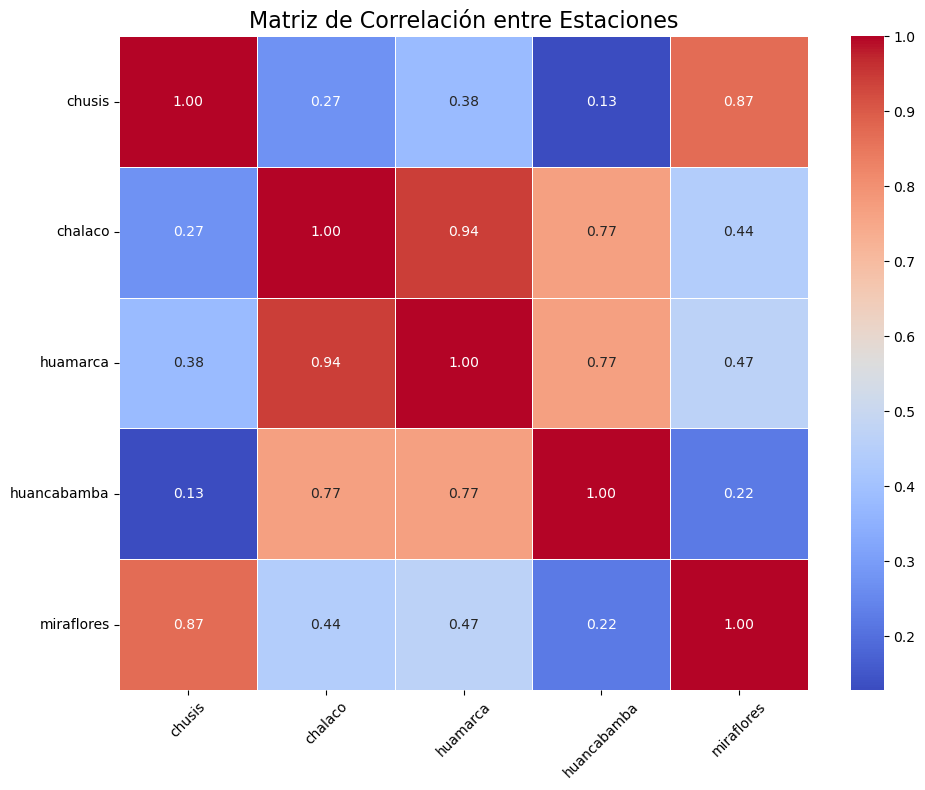
Como el valor de no supera al , se concluye que los residuos si cumplen el criterio de normalidad.

**Prueba de Durbin-Watson**

La prueba de Durbin-Watson es una prueba estadística utilizada para detectar la autocorrelación en los residuos de un modelo de regresión. En este caso, el valor de Durbin-Watson es 1.0659, con un p-valor de 0.0001056. Dado que el p-valor es muy bajo (menor que el umbral común de significancia de 0.05), se rechaza la hipótesis nula de que no hay autocorrelación. El valor de 1.0659 está cerca de 1, lo que sugiere una autocorrelación positiva de primer orden en los residuos, es decir, existe una relación entre los errores residuales consecutivos en el modelo. Esto indica que los residuos no son independientes, lo cual puede ser un problema para la validez de los resultados del análisis ANOVA, ya que una de las suposiciones clave es la independencia de los errores.

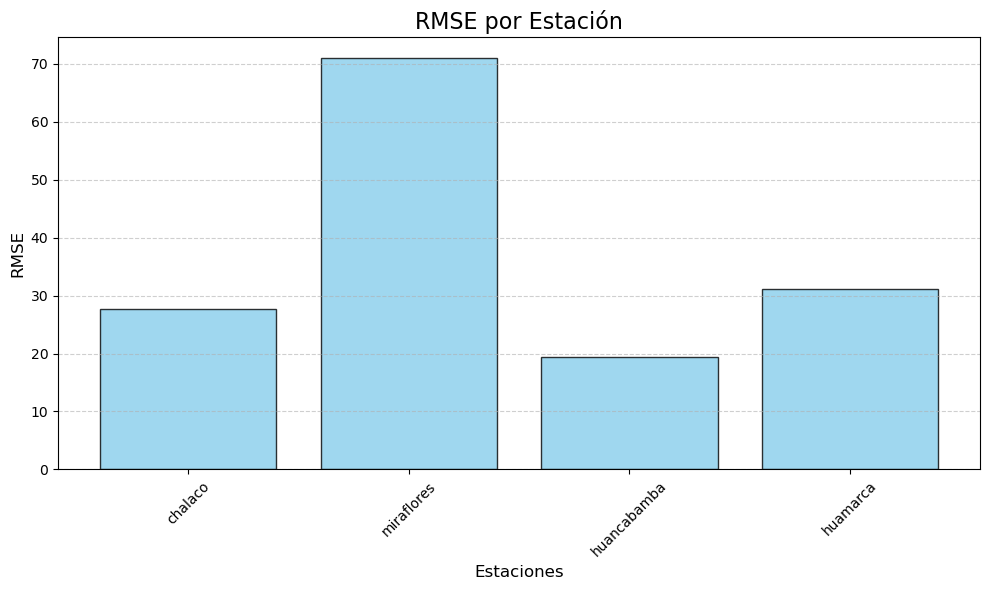
**Matriz de correlaciones**

La imagen muestra una matriz de correlación entre las estaciones meteorológicas Chusis, Chalaco, Huamarca, Huancabamba y Miraflores, con valores que oscilan entre -1 (correlación negativa perfecta) y 1 (correlación positiva perfecta). Las correlaciones más altas se observan entre Chalaco y Huamarca (0.94) y entre Miraflores y Chusis (0.87), indicando que estas estaciones tienen patrones similares de precipitación. Por otro lado, las correlaciones más bajas se encuentran entre Huancabamba y Chusis (0.13) y entre Huancabamba y Miraflores (0.22), lo que sugiere diferencias significativas en los patrones de lluvia entre estas estaciones. El mapa utiliza una escala de colores donde el rojo indica correlación positiva fuerte, el azul muestra correlación baja o negativa, y el blanco representa valores intermedios.

****

**Errores cuadráticos medios**

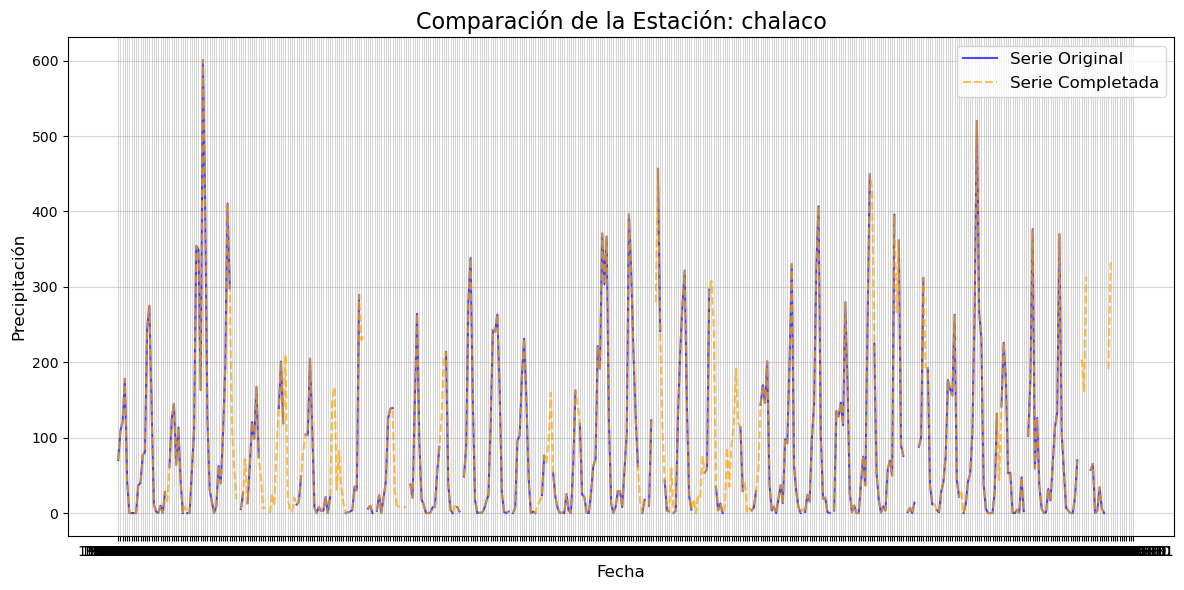
La imagen muestra un gráfico de barras que representa el error cuadrático medio (RMSE) para diferentes estaciones meteorológicas: Chalaco, Miraflores, Huancabamba y Huamarca. El RMSE es una métrica que mide la diferencia promedio entre los valores observados y los valores predichos, con valores más bajos indicando mejor precisión en el modelo. La estación Miraflores tiene el mayor RMSE, superando los 70, lo que sugiere mayores discrepancias en las predicciones. Por otro lado, Huancabamba presenta el RMSE más bajo, cercano a 20, indicando una mayor precisión en los cálculos para esta estación. Chalaco y Huamarca tienen valores intermedios, alrededor de 30. Este gráfico es útil para evaluar la calidad del modelo de imputación de datos en cada estación.

****

**Series completadas**

El gráfico muestra una comparación entre la serie de datos originales y la serie completada para la estación meteorológica "Chalaco". En el eje vertical se representan los valores de precipitación acumulada (en milímetros), mientras que en el eje horizontal se encuentran las fechas de observación. La serie original está representada con una línea azul continua, mientras que la serie completada se muestra con una línea naranja discontinua.

Se observa que ambas series coinciden en gran medida, especialmente en los períodos sin datos faltantes. En los intervalos con valores faltantes en la serie original, la serie completada ofrece estimaciones que siguen la tendencia general de los datos históricos. Esto sugiere que el modelo de imputación utilizado logró reconstruir los valores faltantes de manera consistente con el comportamiento de las precipitaciones en esta estación.

****

# CONCLUSIONES Y RECOMENDACIONES

El presente estudio aplicó modelos de regresión múltiple para completar datos faltantes en registros de precipitación mensual en la cuenca del río Piura, utilizando información de cinco estaciones meteorológicas: Chusis, Chalaco, Huamarca, Huancabamba y Miraflores. Los resultados muestran que esta metodología es efectiva para estimar valores faltantes, especialmente cuando existe una correlación significativa entre las estaciones. La implementación permitió mejorar la disponibilidad de datos, lo cual es esencial para estudios hidrológicos y la gestión de recursos hídricos en la región.

Sin embargo, es importante reconocer que la regresión múltiple es solo una de varias técnicas disponibles para el llenado de datos faltantes. Para robustecer los resultados y reducir posibles sesgos, es recomendable explorar métodos adicionales como la imputación múltiple, modelos basados en series temporales, aprendizaje automático y técnicas geoestadísticas. Además, se requiere un control de calidad más riguroso de los datos, incluyendo la detección y corrección de outliers, análisis de consistencia interna y validación cruzada de los modelos utilizados.

Finalmente, futuros trabajos deberían enfocarse en el análisis de tendencias a largo plazo y la evaluación de la consistencia de los datos completados. Integrar estos datos en modelos hidrológicos e hidráulicos permitirá mejorar las predicciones de eventos extremos y apoyar la planificación estratégica para mitigar los impactos asociados al fenómeno El Niño. Esto contribuirá significativamente a la resiliencia de la región frente a eventos climáticos adversos y al desarrollo sostenible de las comunidades locales.

# REFERENCIAS BIBLIOGRÁFICAS

# ANEXOS