IMPUTACIÓN REPRODUCIBLE DE SERIES TEMPORALES CON TIDYVERSE Y TIDYMODELS

Autores: F. Muñoz[1] y S. Orellana[1].

Afiliaciones:

1. Cienciambiental Consultores S.A.

Palabras clave: R, tidyverse, tidymodels, reproducibilidad, series temporales, imputación, estimación, machine learning

Abstract

En el ámbito de las ciencias ambientales, uno de los desafíos frecuentes es trabajar con series temporales incompletas. Esto ocurre frecuentemente en datos hidrológicos, donde los registros de caudales suelen presentar vacíos debido a problemas operacionales en estaciones o falta de instrumentación. En este trabajo, se aborda un ejemplo de imputación de caudales diarios en múltiples estaciones ubicadas en el Salar de Atacama (norte de Chile), un ecosistema de alta relevancia ecológica y socioambiental. Aunque este caso es de temas hidrológicos, la metodología desarrollada en R es generalizable a múltiples dominios donde se requiera reconstruir series con información faltante.

El desarrollo se implementó exclusivamente en R, utilizando paquetes que permitieron articular un flujo de trabajo coherente y escalable. La lectura y normalización de datos en formatos heterogéneos se resolvió con el ecosistema **tidyverse** ({readr}, {dplyr}, {stringr}, {lubridate}), mientras que la programación funcional con {purrr} hizo posible aplicar procedimientos de forma sistemática a múltiples archivos y estaciones. Una vez unificados los datos, se generaron variables derivadas con el paquete {slider}, que facilita el cálculo de acumulados, retardos y medias móviles, mostrando la flexibilidad de R para la ingeniería de características en series temporales.

El modelado se llevó a cabo con **tidymodels**, que ofrece una gramática unificada para preprocesamiento y algoritmos. Con el paquete {recipes} se aplicaron transformaciones de la variable objetivo para estabilizar la varianza, así como imputación de predictores faltantes mediante medianas. Posteriormente, con el paquete {parsnip} se especificó un modelo Random Forest implementado con {ranger*}*, configurado para priorizar rapidez y capacidad de capturar relaciones no lineales. La evaluación se realizó con el paquete {yardstick}, empleando métricas estándar como RMSE, MAE y el índice de eficiencia Nash–Sutcliffe (NSE). Finalmente, los resultados se comunicaron a través de {ggplot2}, integrando visualizaciones reproducibles que facilitan la interpretación y comparación de series observadas y estimadas.

En conjunto, esta experiencia ilustra cómo R permite resolver las etapas de un análisis de series temporales incompletas: limpieza, generación de variables, modelado predictivo, evaluación y visualización. Aunque aquí se presenta en el contexto de los caudales en el Salar de Atacama, el mismo flujo es aplicable en múltiples áreas de la academia y la

industria, reafirmando el potencial de R como herramienta reproducible, flexible y de amplio alcance para la ciencia de datos en América Latina.