XLVIII Coloquio Argentino de Estadística

VI JORNADA DE EDUCACIÓN ESTADÍSTICA "MARTHA DE ALIAGA"

27 al 30 oct 2020

Poster:

Comparación de métodos de imputación en datos climáticos

Gabriela S. Faviere, Julia Angelini, Eugenia B. Bortolotto, Gabriel Valentini, Gerardo Cervigni

FACULTAD DE CIENCIAS ECONÓMICAS

UNC Universidad Nacional de Córdoba

SAE

COMPARACIÓN DE MÉTODOS DE IMPUTACIÓN EN DATOS CLIMÁTICOS

FAVIERE GABRIELA[1], ANGELINI JULIA[1], BORTOLOTTO EUGENIA[1], VALENTINI GABRIEL[2], CERVIGNI GERARDO[1]

[1]Centro de Estudios Fotosintéticos y Bioquímicos (CEFOBI), UNR—CONICET ; [2]Estación Experimental Agropecuaria INTA San Pedro, Argentina

E-mail: *faviere@cefobi-conicet.gov.ar*

INTRODUCCIÓN

El duraznero requiere una determinada cantidad de frío invernal para atravesar el período de dormancia, maximizando el rendimiento y la calidad de la fruta en cada región. A causa del cambio climático global las futuras variedades podrían requerir menos frío que las actuales. Para realizar estudios de proyección climática se consideran registros climáticos, los cuales presentan frecuentemente valores faltantes. Es necesario tener un registro completo de datos climáticos, por lo que se deben considerar metodologías estadísticas apropiadas para imputar los datos perdidos.

OBJETIVO

El objetivo de este trabajo fue evaluar la eficiencia de diferentes métodos de imputación para datos climáticos.

MATERIALES Y MÉTODOS

Se cuenta con datos diarios de **radiación**, **temperatura máxima** y **mínima** en 7 estaciones meteorológicas: San Pedro (Bs As), Concordia (Entre Rios), Cerro Azul (Misiones), El Colorado (Formosa), Salta, La Consulta (Mendoza) y Alto Valle (Río Negro).

Se comparó la reproducibilidad de las imputaciones de los métodos **Predictive Mean Matching** (PMM), **K-Nearest Neighbors** (KNN) y **Random Forest Imputation** (RF) mediante la raíz del error cuadrático médio (*RMSE*), el coeficiente de correlación (*r*) y un índice de acuerdo (d_2).

SIMULACIÓN

Se generaron distintos porcentajes de pérdidas aleatorias en cada una de las variables: 5, 10, 20 y 40%. Luego, se imputaron los valores faltantes con los tres métodos, los cuales fueron evaluados comparando los datos observados con los imputados. Este proceso se repitió mil veces para cada estación meteorológica y para cada porcentaje de pérdida.

RESULTADOS

En la Figura 1 se representan las medianas del *r* para la temperatura máxima. Se omiten los resultados restantes por ser similares a los presentados.

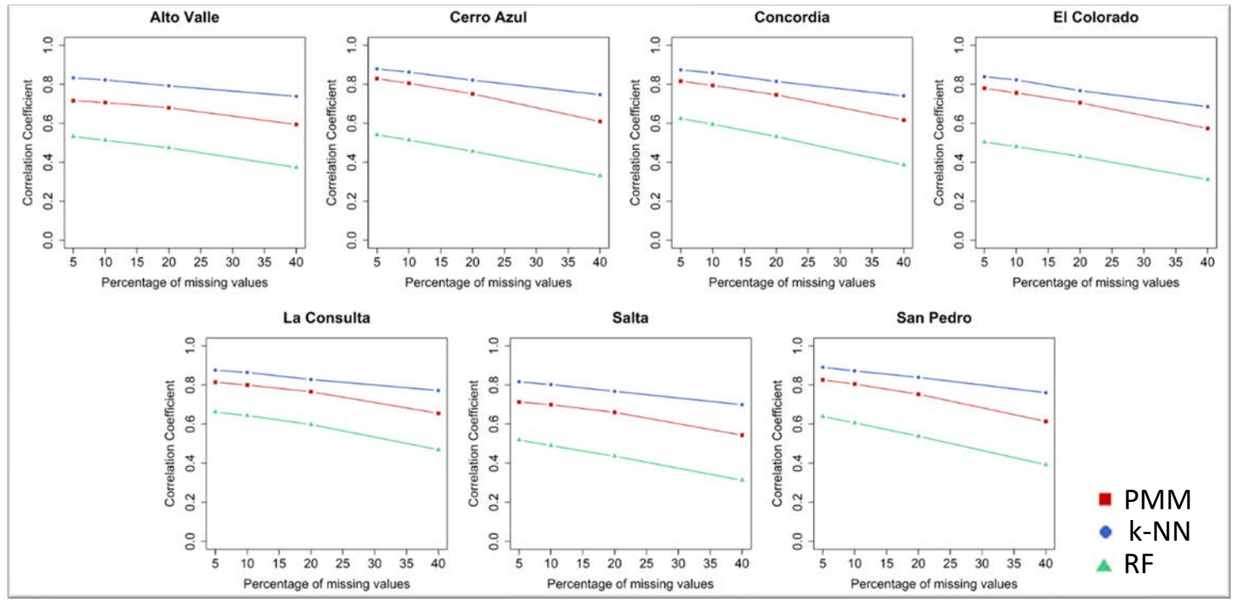

Figura 1: Coeficiente de correlación para la temperatura máxima a través de los distintos porcentajes de pérdida

CONCLUSIÓN

- En todas las estaciones meteorológicas, el método k-NN presentó el mejor rendimiento de acuerdo a las tres medidas de calidad, seguido por PMM y RF.
- RF tuvo el peor desempeño en la imputación de datos climáticos en base a las tres medidas.
- A medida que aumentó el porcentaje de pérdida disminuyó el rendimiento de los métodos de imputación.