EVALUACIÓN DE MÉTODOS PARA LA IMPUTACIÓN DE REGISTROS FALTANTES

ASSESSMENT OF METHODS TO IMPUTATE MISSING WEATHER RECORDS

Federico Schmidt[1], Federico Bert[2], Guillermo Podestá[3], Hernán Veiga[4], Natalia Herrera[4], María de los Milagros Skansi[4], Federico Claus[5], Adriana Basualdo[5]

schmidt.federico@hotmail.com

[1] Universidad Tecnológica Nacional, Facultad Regional Buenos Aires

[2] Facultad de Agronomía, Universidad de Buenos Aires

[3] Universidad de Miami, Escuela Rosenstiel de Ciencias Marinas y Atmosféricas, USA

[4] Servicio Meteorológico Nacional

[5] Oficina de Riesgo Agropecuario, Ministerio de Agricultura, Ganadería y Pesca de la Nación

RESUMEN

Algunos modelos hidrológicos y agronómicos requieren series meteorológicas completas. Sin embargo, es frecuente encontrar valores faltantes en registros meteorológicos que deben ser completados. Este trabajo tiene como objetivo implementar y evaluar métodos de imputación (rellenado) de temperatura máxima, temperatura mínima y precipitación diaria.

Se evaluaron dos grupos de métodos: autónomos y cooperativos. Los métodos autónomos utilizan en la imputación únicamente datos de la estación meteorológica a completar. Los cooperativos añaden datos de estaciones vecinas. Los métodos autónomos evaluados fueron: (i) interpolación lineal de valores usando días contiguos, (ii) MICE y (iii) MissForest. Se evaluaron los métodos cooperativos: (i) reemplazo por valor en estación más cercana, (ii) Inverse Distance Weighting (IDW), (iii) Kriging simple, (iv) Kriging universal, (v) MICE y (vi) MissForest. La implementación de los métodos se realizó en el lenguaje R.

La evaluación de las imputaciones se realizó para la estación meteorológica Junín del Servicio Meteorológico Nacional, ya que tiene datos completos entre 1994 y 2013. Además se dispuso de registros climáticos para las seis estaciones más cercanas a Junín (distancia máxima 170km).

Los métodos se evaluaron para diferentes escenarios. Cada escenario se definió en base a: (a) distintas proporciones de datos faltantes (de 5 a 50%) y, (b) distintas longitudes de secuencias faltantes (de 1 a 90 días). Para crear los escenarios se generaron faltantes artificiales aleatoriamente en los registros completos de Junín.

Para temperaturas, todos los métodos cooperativos mostraron muy buen desempeño, independientemente del largo o proporción de faltantes. El ajuste (R^2) entre los datos observados e imputados fue mayor a 0.9 en todos los escenarios, con un RMSE cercano a 1.3°C y un MAPE < 4%. Los métodos autónomos mostraron un rendimiento inferior y más afectado por la longitud de faltantes, con un R^2 cercano a 0.75, un RMSE mínimo de 3°C y un MAPE cercano al 14%.

Las imputaciones de precipitaciones resultaron más imprecisas. Con los métodos cooperativos se obtuvieron valores de R^2 cercanos a 0.3; el R^2 para métodos autónomos estuvo entre 0.01 y 0.11. Una alternativa es usar dos etapas: imputar primero la ocurrencia de lluvia (precipitación > 1mm), lo que permite asignar 0 a días secos y en una segunda etapa imputar montos para días lluviosos.

Usando MissForest con estaciones vecinas la imputación de día seco/lluvioso es correcta en el 86-90% de los casos, según escenarios. Cuando se imputa el monto de lluvia en la segunda etapa el error no mejora mucho con la imputación directa: el R^2 estuvo entre 0.26 y 0.3.

Se concluye que la imputación de temperaturas tiene buen desempeño, especialmente si se tiene datos de estaciones vecinas. Si no hay datos de vecinos, se obtienen buenos resultados usando interpolación lineal o MissForest para secuencias de faltantes cortas (1-3 días). En cambio, la imputación de precipitación tiene resultados mucho más inciertos, y los valores imputados generalmente son mucho menores que los observados. La imputación separada de ocurrencia y monto de lluvia no mejoró significativamente el error final.

ABSTRACT

Some hydrological and agronomic models require complete meteorological data series. However, meteorological series usually have missing values that must be imputed (filled). This work aims to implement and assess different imputation methods for maximum and minimum temperatures and daily rainfall.

Two groups of imputation methods were evaluated: autonomous and cooperatives. The autonomous methods use data only from the meteorological station to fill. Conversely, the cooperative methods also use data from neighboring stations. The autonomous methods tested were: (i) linear interpolation using contiguous days, (ii) MICE and (iii) MissForest. The cooperative methods tested were: (i) imputation of the value from the nearest neighbor, (ii) Inverse Distance Weighting (IDW), (iii) Simple Kriging, (iv) Universal Kriging, (v) MICE and (vi) MissForest.

Imputation tests were performed for the weather station of the Argentinean National Meteorological Service at Junín, since it has complete records from 1994 to 2013. Additionally, we used meteorological records from Junín's six nearest neighbors (max distance: 170 km).

Imputation methods were tested in different scenarios. Each scenario was defined by: (a) different proportions of missing data (from 5 to 50%) and, (b) the length of consecutive missing values (between 1 and 90 days). The scenarios were created by randomly generating artificial missing records into the complete series for Junín.

All of the cooperative methods performed really well when imputating temperatures, regardless of the length or proportion of missing values. The fit (R^2) between observed and imputated values was greater than 0.9 for all scenarios, with a RMSE close to 1.3°C and a MAPE < 4%. Autonomous methods showed worse results in general but the performace varied with the scenarios. These methods showed an R^2 close to 0.75, a RMSE > 3°C and a MAPE close to 14%.

Rainfall imputations turned out to be less precise. Cooperative methods showed R^2 values close to 0.3 while autonomous methods' values ranged between 0.01 and 0.11. Alternatively, a two stages approach could be used: first to imputate the occurrence or not of rain (dry days are imputated with 0 mm) and then to imputate the rainfall amount to rainy days. Using MissForest and neighbors' data we were able to predict rainy days correctly in 86 to 90% of the cases, depending on the scenario. However, this approach doesn't improve significantly the results when compared to the original approach: the R^2 between observed and imputated rainfall values stayed between 0.26 and 0.3.

We concluded that temperature imputation performs quite well, especially if neighbors' data is

available. Otherwise, acceptable results can be obtained by using linear interpolation or MissForest to fill short missing values series (1-3 days). Conversely, rainfall imputation shows less precise results and imputated values tend to be lower than the observed ones. Finally, imputating rainfall in two stages did not improve the results.

Palabras clave: imputación, variables meteorológicas.