Estrategias de imputación de datos faltantes y outliers en estudios de metabolómica, ¿cambian realmente los resultados?

E.L. Reinoso-Peláez[1], J.L. López-Martínez[1], M.J. Carabaño[1], M. Ramón[1], C. Meneses[1], C. González[1], A. Hernández-Pumar, and C. Díaz[1]

[1] Centro Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Ctra. de La Coruña, km 7,5, 28040, Madrid, España
* Corresponding author: edgar.reinoso@inia.csic.es

Resumen

La metabolómica captura la actividad bioquímica celular en tiempo real por lo que nos permite estudiar la respuesta fisiológica de los individuos a distintos tipos de estrés, entre ellos el estrés térmico. No obstante, su potencial suele verse limitado por la alta dimensionalidad, los datos faltantes (*NAs*) y los valores atípicos (*outliers*), lo que hace indispensable un riguroso preprocesamiento de los datos. En este contexto, este estudio plantea un doble objetivo: primero, evaluar la precisión de 15 estrategias de imputación para reconstruir los NAs y el tratamiento de outliers, y segundo, evaluar el impacto de éstas sobre la elección de metabolitos como biomarcadores asociados al estrés calórico en dos razas bovinas.

Se analizó el perfil metabolómico en el plasma de 81 vacas Holstein (FR) y 67 terneros Avileña Negra Ibérica (ANI), clasificados como tolerantes (TOL) y susceptibles (SUS) según su ritmo respiratorio bajo estrés por calor. Estos perfiles se cuantificaron mediante UPLC-MS/MS, excluyendo xenobióticos y metabolitos con >20% NAs. Se evaluaron tres escenarios de filtrado de outliers basados en el rango intercuartílico (*Ninguno*, *5IQR* y *3IQR*) con cinco métodos de imputación: dos variantes de MICE (Multivariate Imputation by Chained Equations) diferenciadas por su preselección de predictores (basadas en *top-correlación* y *selección dinámica*), Random Forest (*rf*), k-Nearest Neighbors (*knn*) y regresión lineal de componentes principales (*lm-pcr*). La evaluación se estructuró en dos fases. Primero, se determinó la precisión de la imputación mediante validación cruzada (20 k-folds), empleando la raíz del error cuadrático medio (RMSE) y la correlación de Pearson (*r*). Segundo, las 15 matrices imputadas se normalizaron (Box-Cox), se corrigieron por covariables (rebaño y lactancia en FR; peso en ANI), y se sometieron a un análisis discriminante PLS-DA para clasificar los fenotipos SUS y TOL. La importancia de los metabolitos se evaluó mediante su índice VIP (Variable Importance in Projection), que cuantifica la contribución individual de cada metabolito a dicha discriminación.

La validación cruzada mostró que la precisión en la imputación depende fundamentalmente del método empleado en ambas razas. El método *lm-pcr* mostró la mayor precisión (*r* media ~ 0,73; RMSE medio ~ 0,12), mientras que *knn* la peor (*r* media ~ 0,50; RMSE medio~ 0,17). El análisis PLS-DA reveló que la importancia de los metabolitos (valores VIP) depende fundamentalmente del escenario de outliers. El escenario *3IQR* maximizó sistemáticamente la capacidad discriminante de los métodos de imputación, impactando el número y tipo de metabolitos más importantes (VIP > 1) a la hora de discriminar entre TOL y SUS, aunque un alto porcentaje son comunes a las 15 estrategias estudiadas (196 para ANI y 130 para FR).

En conclusión, la elección de un método robusto (destacando lm-pcr) es vital para minimizar el error de imputación. Por otro lado, el tratamiento de outliers es un factor que puede tener un impacto relevante en los resultados biológicos, máxime cuando los tamaños muestrales son reducidos. El análisis de estos datos requiere evaluar distintas estrategias para identificar un conjunto de metabolitos de consenso, con el fin de aportar mayor fiabilidad y robustez a los resultados.

Keywords: Metabolómica, datos faltantes, outliers, Imputación, PLS-DA, Estrés por calor.

Agradecimientos: Este trabajo ha sido financiado por el proyecto Re-Livestock (EU Horizon Europe, Grant Agreement No. 101059609). Las Asociación de Avileña-Negra Ibérica (AEANI) y de Frisona de Castilla-La Mancha (AFRICAMA) han contribuido en la obtención de muestras. Se agradece a Samuele Bovo, Matteo Bolner y Luca Fontanesi por su apoyo.