



Re-Livestock
RESILIENT FARMING SYSTEMS

Reunión Nacional de Mejora Genética Animal 2026

National Meeting on Animal Breeding · 11-12 June 2026,
Orihuela, Spain

Re-Livestock Abstracts

IMPUTATION STRATEGIES FOR MISSING DATA AND OUTLIERS IN METABOLOMIC STUDIES: DO THE RESULTS REALLY CHANGE?.....2

E.L. REINOSO-PELÁEZ^{1*}, J.L. LÓPEZ-MARTÍNEZ¹, M.J. CARABAÑO¹, M. RAMÓN¹, C. MENESES¹, C. GONZÁLEZ¹, A. HERNÁNDEZ-PUMAR, C. DÍAZ¹



**Funded by
the European Union**

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.





Re-Livestock
RESILIENT FARMING SYSTEMS

Imputation strategies for missing data and outliers in metabolomic studies: do the results really change?

E.L. Reinoso-Peláez^{1*}, J.L. López-Martínez¹, M.J. Carabaño¹, M. Ramón¹, C. Meneses¹, C. González¹, A. Hernández-Pumar, C. Díaz¹

¹Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Ctra. de La Coruña, km 7,5, 28040 Madrid, Spain

Abstract

Metabolomics captures cellular biochemical activity in real time, allowing us to study the physiological response of individuals to different types of stress, including heat stress. However, its potential is often limited by high dimensionality, missing data (NAs) and outliers, which makes rigorous data pre-processing essential. In this context, this study has a twofold objective: first, to evaluate the accuracy of 15 imputation strategies for reconstructing NAs and handling outliers, and second, to assess their impact on the choice of metabolites as biomarkers associated with heat stress in two cattle breeds.

The plasma metabolomic profile of 81 Holstein cows (FR) and 67 Avileña Negra Ibérica calves (ANI), classified as tolerant (TOL) and susceptible (SUS) according to their respiration rate under heat stress, was analysed. Profiles were quantified by UPLC-MS/MS, excluding xenobiotics and metabolites with >20% NAs. Three outlier-filtering scenarios based on the interquartile range (None, 5IQR and 3IQR) were evaluated with five imputation methods: two variants of MICE (Multivariate Imputation by Chained Equations) differing in their pre-selection of predictors (based on top correlation and dynamic selection), Random Forest (rf), k-Nearest Neighbors (knn) and principal component linear regression (lm-pcr). The evaluation was carried out in two phases. First, imputation accuracy was determined by cross-validation (20 k-folds), using the root mean square error (RMSE) and Pearson correlation (r). Second, the 15 imputed matrices were normalised (Box-Cox), corrected for covariates (herd and lactation in FR; weight in ANI), and subjected to a PLS-DA discriminant analysis to classify the SUS and TOL phenotypes. Metabolite importance was assessed using the VIP (Variable Importance in Projection) index, which quantifies the individual contribution of each metabolite to this discrimination.

Cross-validation showed that imputation accuracy depends mainly on the method used in both breeds. The lm-pcr method showed the highest accuracy (mean r ~ 0.73; mean RMSE ~ 0.12), while knn showed the lowest (mean r ~ 0.50; mean RMSE ~ 0.17). PLS-DA analysis revealed that metabolite importance (VIP values) depends mainly on the outlier scenario. The 3IQR scenario systematically maximised the discriminant capacity of the imputation methods, affecting the number and type of the most important metabolites (VIP > 1) for discriminating between TOL and SUS, although a high percentage are common to the 15 strategies studied (196 for ANI and 130 for FR).

In conclusion, choosing a robust method (notably lm-pcr) is vital to minimise imputation error. On the other hand, outlier treatment can have a relevant impact on biological results, especially when sample sizes are small. Analysing these data requires evaluating different strategies to identify a consensus set of metabolites, in order to provide greater reliability and robustness to the results.

Keywords: Metabolomics, missing data, outliers, imputation, PLS-DA, heat stress.



Funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.





Re-Livestock
RESILIENT FARMING SYSTEMS

Original abstract (Spanish)

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

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.

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



**Funded by
the European Union**

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

