IMPUTACIÓN MEDIANTE MÉTODO DE ESPERANZA-MAXIMIZACIÓN

UN ESTUDIO DE SIMULACIÓN SOBRE DATOS DE ENSAYOS MULTIAMBIENTE

Autoras: Angelini, Julia; Armendáriz, Aldana; Macat, Paula | **Licenciatura en estadística** | Tutora: Quaglino, Marta

INTRODUCCIÓN

Entre las herramientas más importantes para la evaluación de genotipos superiores durante el fitomejoramiento se encuentran los ensayos multiambiente (MET). Consisten en evaluar un conjunto de genotipos en múltiples ambientes. Originalmente, están pensados para evaluar todos los genotipos en todos los ambientes, pero ya sea por el azar o por la acción planificada de los investigadores, tienen la particularidad de presentar valores faltantes con frecuencia.

Uno de los modelos más utilizados en el análisis de datos MET es el Additive Main-effects and Multiplicative Interaction (AMMI), que combina el ANOVA de efectos principales genotipo y ambiente con un análisis de componentes principales para la interacción entre ellos (IGA). Sin embargo, el AMMI no permite analizar datos incompletos.

Una de las numerosas metodologías de imputación desarrolladas para afrontar este inconveniente es el EM-AMMI, un método que combina el modelo AMMI con el algoritmo de esperanza-maximización.

OBJETIVO

El objetivo de este estudio es comparar la eficiencia del método EM-AMMI en la reproducibilidad de los datos faltantes considerando sólo los efectos aditivos del modelo (EM-AMMI0), y una única componente multiplicativa en el modelo (EM-AMMI1).

MATERIALES Y MÉTODOS

MODELOS AMMI

La estructura de un modelo AMMI es:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{K} \theta_k a_{ki} b_{kj} + \varepsilon_{ij} \quad i = 1,\dots,I \; j = 1,\dots,J \; k = 1,\dots,K$$

Donde μ es la media general, α_i y β_j son los efectos principales asociados al i-ésimo genotipo y al j-ésimo ambiente respectivamente, y ε_{ij} es el error aleatorio asociado al i-ésimo genotipo y al j-ésimo ambiente. $\sum_{k=1}^{K} \theta_k a_{ki} b_{kj}$ es la sumatoria de componentes multiplicativas que modela las interacciones Genotipo-Ambiente, donde a_{ki} es elemento del autovector asociado al i-ésimo genotipo y la k-ésima componente, b_{kj} es el elemento del autovector asociado al j-ésimo ambiente y la k-ésima componente y θ_k es el autovalor asociado a esa misma componente.

MÉTODO DE IMPUTACIÓN EM-AMMI

1. Las celdas faltantes se inicializan con:
Y_{ij} = Media Genotipo + Media Ambiente$_i$ – Media General$_j$ obtenidas de los datos observados.
2. Los parámetros aditivos se inicializan con la media general, la media genotípica y la media ambiental de la matriz completa actualizada.
3. Se obtiene la matriz de residuos R, donde cada elemento se calcula como: r_{ij} = Observación – Media Genotipo – Media Ambiente + Media general.
4. Los estimadores mínimo cuadráticos de $\theta_k a_{ki} b_{kj}$ (k = 0,…,K K=0 para AMMI0 y K=1 para AMMI1) se obtienen aplicando la SVD a la matriz de residuos R del paso 3.
5. Se actualizan las celdas faltantes usando el modelo AMMI con K componentes multiplicativas.
6. Se calcula la distancia de Tchebychev entre los valores imputados de dos iteraciones sucesivas. Si ésta es < 0.01 el proceso termina. De lo contrario, se vuelve al paso 2.

ESTUDIO POR SIMULACIÓN

Utilizando una estructura AMMI con dos componentes multiplicativas, se genera un conjunto de datos de 100 genotipos y 8 ambientes.
Se elimina al azar un 20% de los datos y se imputan utilizando los métodos EM-AMMI0 y EM-AMMI1. Para evaluar la eficiencia de los métodos en reproducir el dato faltante, se calculan las estadísticas M^2 de Procrustes, NRMSE y el Coef. de Correlación S de Spearman. Este proceso se repite 1000 veces.
A su vez, se lleva a cabo un Test de Wilcoxon para cada estadística a fin de determinar si existen diferencias significativas entre los métodos a través de las 1000 simulaciones.

RESULTADOS

Según la estadística M^2, el método EM-AMMI es levemente superior respecto del AMMI0 (Figura 1). La NRMSE muestra una notoria diferencia de medianas, sugiriendo de nuevo que el EM-AMMI1 es superior, aunque presenta más variabilidad y una distribución asimétrica (Figura 2). El coeficiente S indica una fuerte correlación entre los valores verdaderos y los imputados por EM-AMMI1, y una débil correlación con los del EM-AMMI0 (Figura 3).

De acuerdo al Test de Wilcoxon, la evidencia muestral es suficiente para concluir que cada una de las estadística a través de las 1000 simulaciones difiere entre el EM-AMMI0 y el EM-AMMI1 ($p < 0.01$, para las tres estadísticas).

CONCLUSIÓN

En el presente trabajo se observa que el método EM-AMMI1 es más eficiente que el EM-AMMI0 en la reproducibilidad del dato perdido, cuando se considera una proporción de valores faltantes del 20%.

Figura 1. "Estadística M^2 según método de imputación"

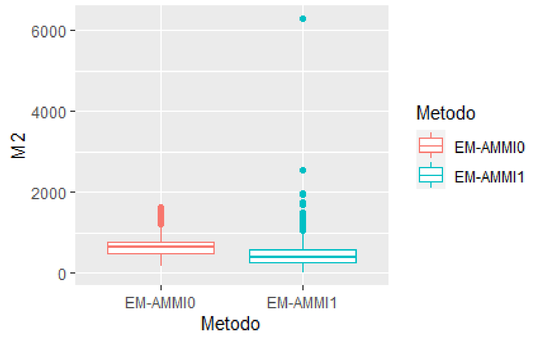

Figura 2. "Estadística NRMSE según método de imputación"

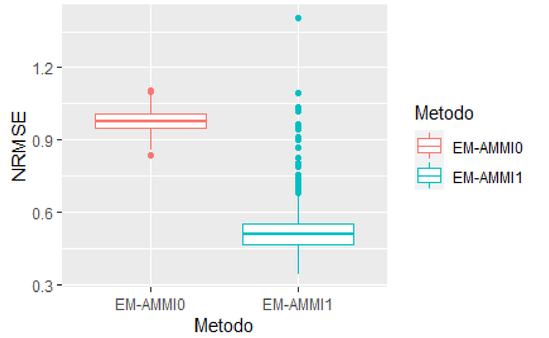

Figura 3. "Coeficiente S según método de imputación"