Universidad Nacional de Córdoba
Maestría en Estadística Aplicada

Estrategias de modelación interacción especie ambiente en función del rendimiento de materia seca en cultivos de cobertura

Tesis
para optar al grado académico de
Magíster en Estadística Aplicada
Ing. Agr. Américo Nicolás Contreras Valdovinos

2019

Universidad Nacional de Córdoba
Maestría en Estadística Aplicada

Director:
Dr. Julio Alejandro Di Rienzo
Facultad de Ciencias Agropecuarias -UNC

Tribunal Evaluador:
Dra. Susana Beatriz Ferrero
Facultad de Ciencias Exactas, Físico-Químicas y Naturales- UNRC

Dr. Mariano Augusto Córdoba
Facultad de Ciencias Agropecuarias -UNC

Mgter. Elena Margot Tablada
Facultad de Ciencias Agropecuarias -UNC

FECHA DE APROBACIÓN DE TESIS: 29 de agosto de 2019.

Agradecimientos

En primer lugar, quisiera agradecer el trabajo de mi profesor guía, Dr. Julio Di Rienzo en el desarrollo de esta tesis. Sus acertados consejos me ayudaron de manera importante en el planteamiento y desarrollo de este maravilloso tema.

A mis padres y hermanos, a quienes llevo en mi memoria siempre.

A mis amigos cordobeses, Norma, Ana, Laura, Germán y Juan, con quienes pasamos momentos increíbles durante mi maestría.

A Erika, quien me respaldó y cubrió mis espaldas en la Universidad de Chile mientras estaba en Argentina.

Y de manera especial a Karen, quién nunca ha dudado de mis capacidades. A ella dedico de manera especial este trabajo.

Índice

Resumen

Ensayos agrícolas realizados durante varias temporadas, localidades y utilizando diferentes líneas de un único genotipo son comunes en experimentación agrícola. Diversas técnicas son utilizadas para evaluar este tipo de ensayos multiambientales, principalmente modelos AMMI o Modelos SREG, los cuales frecuentemente se enfrentan a desbalances, ya sea realizados deliberadamente, o de manera involuntaria, lo que tiene consecuencias a la hora de realizar el modelamiento estadístico. Este trabajo consistió en adaptar algunas de las técnicas utilizadas en ensayos multiambientales con una única especie y diversas líneas a ensayos en los cuales se analizaron diferentes cultivos de cobertura en función de su rendimiento de materia seca, para distintos ambientes y temporadas, y que presentaron desbalance de datos. Se propusieron diferentes escenarios de modelación matemática mediante modelos lineales mixtos, imputación de datos faltantes y presentación de resultados tales como modelos AMMI y modelos SREG a través de sus respectivos gráficos biplot. Los resultados sugieren que tanto la utilización de modelos lineales mixtos, con su ventaja a la hora de modelar datos heteroscedásticos y la utilización de técnicas de imputación de datos ausentes son una alternativa válida a la hora de considerar desbalances implícitos. A pesar de lo anterior debe considerarse, al momento de presentar los resultados, el porcentaje de datos ausentes y la gran variabilidad de rendimientos en especies sin un cercano parentesco, lo que se reflejará en el momento de analizar los gráficos, principalmente los gráficos GGE producto de modelos SREG.

Palabras claves: AMMI, GGE, Modelos lineales mixtos, SREG, imputación de datos.

Abstract

Agricultural trials conducted over several seasons, different environments and using different varieties of a genotype are common in agricultural research. Various techniques can be used to evaluate this type of multi-environmental analysis, mainly AMMI models or SREG models, which frequently face the disadvantage of missing data, either deliberately or involuntarily, which has consequences when performing statistical modelling. In this work some of the techniques used in multi-environmental trials with a single species and several varieties were applied to trials in which different cover crops were analysed based on dry matter yield, in different environments and seasons, with unbalanced dataset. Different scenarios of mathematical modelling were proposed through mixed linear models, imputation of missing data, AMMI and SREG models through their respective biplot graphics. The results suggest that both, the use of mixed linear models, with their advantage when modelling heteroscedastic data, and the use of missing data imputation techniques are a valid alternative when considering implicit unbalances. In spite of the above, the percentage of missing data and the great variability of yield in different species should be considered when of presenting the results, which will be reflected in the graph´s analysis, mainly the GGE graphs from SREG models.

Keywords: AMMI, GEE, Linear mixed models, SREG, data imputation.

1. Introducción

1.1 Cultivos de cobertura

La Región de Atacama, República de Chile, está ubicada entre los paralelos 26° y 29°20' latitud sur. El clima se caracteriza por ser desértico árido (BW[1]) con una gran amplitud térmica y precipitaciones que varían entre 12 mm anuales en la ciudad de Copiapó hasta los 32 mm anuales en la ciudad de Vallenar. La Región se caracteriza por enfocar su producción frutícola bajo riego especialmente en uva de mesa, la cual representa aproximadamente 6.835 ha de superficie, establecidas en dos valles transversales, conocidos como el valle de Copiapó y el valle del Huasco (Odepa, 2018).

La producción de uva de mesa está enfrentada a importantes dificultades edafoclimáticas en la Región de Atacama, lo cual obliga a realizar todos los esfuerzos técnicos posibles que permitan evaluar y controlar las condiciones de crecimiento y desarrollo de la fruta. En general, las condiciones de los suelos en la Región son: pH entre 7,8 y 8,5, conductividad eléctrica (CE) entre 1,2 a 4 $dS \cdot m^{-1}$, altos niveles de sodio, cloro y boro y bajos niveles de materia orgánica (MO) (Callejas et al., 2013). Lo anterior lleva a obtener rendimientos de uva de mesa por hectárea y porcentajes de fruta exportada inferiores a los potenciales que presenta la zona. Una de las razones más discutidas como responsable de los bajos rendimiento en la región son las características específicas de los suelos: poco estructurados, con baja fertilidad natural y bajo contenido de materia orgánica (Baginsky et al., 2010). En forma natural, los suelos de la Región de Atacama poseen bajos niveles de materia orgánica, lo que provoca que ésta sea una limitante en los sistemas productivos de la Región. Cabe recordar que los efectos positivos que tiene la materia orgánica sobre las propiedades físicas del suelo son: favorecer una mayor disponibilidad de agua (Rawls et al., 2003), mejorar la estabilidad estructural, influyendo directamente en el transporte de agua y gases (Horn and Baumgartl, 1999) y entregar una mejor resistencia mecánica cuando estos suelos se encuentran en condiciones de humedad. En relación a los efectos positivos sobre las condiciones químicas, la materia orgánica aumenta la capacidad de intercambio catiónico y reduce las pérdidas de potasio, calcio y magnesio, aportando a su vez altos porcentajes de nitrógeno, fosfatos y azufre (Varnero, 1992), además de ser un quelante en la absorción de micronutrientes como hierro, cobre y zinc (Bohn et al., 1993).

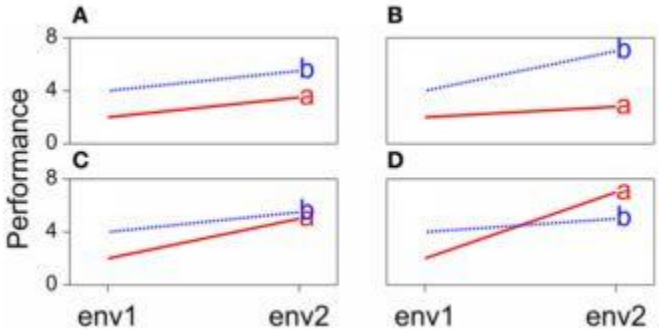
[1] Clasificación climática de Köppen

El manejo agronómico habitual de los agricultores de Atacama, con el fin de subsanar las dificultades asociadas a los bajos niveles de materia orgánica, consiste en aplicar enmiendas en los predios productivos, siendo la aplicación de compost y guano de cabra las labores más recurrentes. Una segunda opción, evaluada con éxito en Chile y otras regiones del mundo, consiste en la utilización de cultivos de cobertura asociados al cultivo principal. El cultivo de cobertura, que se desarrolla en conjunto con el cultivo principal, aporta, a través de la descomposición de sus raíces y la parte aérea de la planta, materia orgánica al suelo, la cual se incorporará a la ya existente. Experiencias de la utilización de cultivos de cobertura asociados a especies principales con diversos fines existen en un gran número de especies, tales como olivos (González et al., 2005), frambuesas (Ovalle et al., 2007a), viñedos (Ovalle et al., 2007b), cítricos (Cesco et al., 2006), (Ormeño, 1998) y kiwi (Rombolá et al., 2004). Las utilizaciones de cultivos de cobertura han reportado tanto beneficios en las condiciones físico-químicas de los suelos como en condiciones sanitarias de las plantas, ya que la utilización de estos cultivos asociados a frutales, tales como raps y mostaza, ayudan a disminuir las poblaciones de patógenos como nematodos, dado que las raíces de esas especies generan exudados volátiles, que controlan las poblaciones de estos microorganismos (Aballay e Insunza, 2002). A su vez, la utilización de cultivos de cobertura ha sido probada de manera exitosa para el control de malezas en frutales de carozo y pomáceas, obteniéndose disminuciones de cerca de un 90% dentro de un periodo de dos años, reduciendo de manera importante la aplicación de herbicidas en la hilera de los árboles (Ormeño, 1998).

Como la acumulación de materia seca es un proceso bastante lento, es necesario evaluar las incorporaciones de materia orgánica, mediante cultivos de cobertura, a través de la utilización de diversos cultivos en largos periodos de tiempo. Al mismo tiempo, es necesario tratar de replicar los ensayos en distintos ambientes y/o temporadas, con el fin de seleccionar los cultivos de cobertura que presenten una mejor adaptabilidad a los diversos ambientes evaluados.

Diversas técnicas se han desarrollado para evaluar la interacción que puede existir entre diferentes genotipos utilizados y las diversas localidades en las cuales se han establecidos. El hecho de replicar los genotipos en una gran cantidad de ambientes impacta directamente en el espacio de inferencia, permitiendo, de esta manera, definir los programas de cultivos y recomendaciones de siembra en diversas localidades en función del genotipo mejor adaptado

a tales condiciones. La *interacción genotipo-ambiente* (GxA) se define como el "*desempeño inconsistente de genotipos sobre diferentes ambientes*" siendo éste desempeño particularmente importante tanto en cultivos agrícolas como plantaciones forestales (Yang, 2014). La Figura 1 muestra diferentes escenarios teóricos que pueden suceder al comparar la "performance" de 2 genotipos en dos ambientes diferentes. La Figura 1A refleja una no interacción genotipo ambiente expresando su relación de manera aditiva. Eso significa que independiente de si el genotipo es evaluado en el ambiente 1 o 2 la respuesta será siempre favorable al genotipo b, con un diferencial constante entre ambos rendimientos, observándose gráficamente dos líneas paralelas. En la Figura 1B, 1C y 1D podemos apreciar tres situaciones en las cuales se manifiesta la interacción GxA. La primera (Figura 1B) se conoce como interacción GxA divergente, la segunda (Figura 1C) se conoce como interacción GxA convergente y la tercera (Figura 1D), la podemos considerar como más interesante de estudiar, la cual se conoce como interacción crossover. Esta última condición implica la capacidad de poner a prueba al mejorador el cual deberá considerar el genotipo que mejor se desarrolle en un ambiente determinado. La condición B y C también es conocida como interacción cuantitativa, dado que las diferencias entre los genotipos evaluados sólo varían en magnitud, pero no en dirección (El ranking de genotipos se mantiene inalterable), a diferencia de la condición D, la cual también puede variar en dirección (cambio en el ranking de los genotipos), siendo considerada esta última como una interacción cualitativa (Singh et al., 1999).



env1: ambiente 1, env2: ambiente 2, a: genotipo a, b: genotipo b

Figura 1. Esquema de interacción genotipo ambiente utilizando dos genotipos en dos ambientes diferentes (Malosetti et al. 2013)

Un aspecto adicional que puede ser incluido en la discusión, es considerar que dentro de los esquemas de interacción (cuantitativo-cualitativo), pueden existir cambios en las varianzas dentro de cada ambiente en el cual han sido establecidos los genotipos, lo cual tiene importantes implicancias en programas de mejoramiento (Browman, 1972). Lo anterior significa que la varianza, al momento de estimar un desempeño promedio de un ambiente, considerando todos los genotipos evaluados dentro de él, se mantiene constante al ser evaluados los mismos genotipos en un segundo ambiente. Así, por ejemplo, podemos encontrar interacción crossover sin cambios en varianza, interacción únicamente cuantitativa con cambios en varianza y presencia de interacción crossover con cambios de varianza.

En experimentos agrícolas en los cuales se considera de interés estudiar la interacción genotipo-ambiente, es posible encontrar tres tipos de estrategias de modelación comúnmente utilizadas: modelos lineales, modelos bilineales y modelos lineales-bilineales.

1.2 Modelos lineales

Las primeras investigaciones realizadas para analizar la interacción genotipo ambiente en modelos lineales fueron aproximaciones mediante la utilización de modelos de análisis de regresión linear simple. El modelo consistía en:

$$y_{ij} = \beta_0 + \beta_1 E_j + e_{ij} \qquad (1)$$

Donde y_{ij} es la respuesta del genotipo i en el sitio j, β_0 es la ordenada al origen, β_1 es la pendiente o "sensibilidad genotípica", E_j corresponde al efecto j-ésimo sitio y e_{ij} son los errores experimentales, que se suponen independientes, normalmente distribuidos con media 0 y varianza constante σ_e^2.

La estrategia asumida consiste en evaluar los distintos genotipos en los ambientes definidos previamente, obtener los rendimientos promedios de todos los genotipos por cada ambiente y posteriormente obtener el "efecto sitio", el cual se calcula mediante la diferencia del promedio de todos los genotipos en un ambiente menos el promedio de todos los ambientes. Con esta información se podía analizar, en primera instancia la calidad de los ambientes. Un ambiente de alto potencial tendría valores altos y positivos de E, mientras que ambientes de bajo desempeño tendrían valores negativos grandes (Westcott, 1986). Una vez obtenido el efecto

sitio para cada ambiente, se estima un modelo de regresión para cada genotipo, considerando como variable dependiente los rendimientos del genotipo observados en cada ambiente y como variable independiente los efectos de sitios.

Para comparar distintos genotipos en un "ambiente promedio" se comparan las ordenadas al origen (efecto sitio igual a cero), siendo el genotipo a destacar el que presenta el mayor valor posible. La *sensibilidad genotípica* se deriva de la estimación de la pendiente del modelo de regresión. Un genotipo cuyo rendimiento refleja la calidad del sitio se considera estable en términos dinámicos y su pendiente esperada es 1. Valores de pendiente mayores que 1 son indicativos de genotipos con un alto potencial en ambientes de alta calidad, pero al mismo tiempo un indicador de inestabilidad y rendimientos comparativamente pobres en ambientes de baja calidad. Una ventaja de este modelo es que permite realizar predicciones de rendimientos de genotipos en ambientes no evaluados y no requiere repeticiones de los genotipos dentro de un mismo sitio. Una desventaja es que resulta muy difícil caracterizar un ambiente sólo con una única variable, por lo que una parte importante de la interacción genotipo ambiente permanecerá inexplicada bajo este enfoque (Malosetti et al., 2013). Vale la pena destacar que, a pesar de ser una única variable, el rendimiento resume tanto aspectos genotípicos, ambientales como sus interacciones.

Los estudios realizados por Finlay y Wilkinson (1963), pioneros en el desarrollo de modelos de análisis de datos multiambientales, concluyeron que variedades con altos rendimientos y coeficientes de regresión cercanos a cero eran genotipos con alta adaptación a todos los ambientes. También destacaron en sus análisis que variedades con coeficientes de regresión altos tendían a tener alta sensibilidad, lo cual significa que grandes cambios ambientales se traducen en grandes cambios en los rendimientos. Junto a lo anterior, destacan que el ignorar los resultados obtenidos en ambientes con bajos rendimientos es un grave error, dado que al seleccionar líneas que presenten altos rendimientos, bajo condiciones favorables podrían mostrar grandes caídas en sus rendimientos en la medida que sean sometidos a condiciones adversas.

Una segunda aproximación a los modelos lineales que consideran los genotipos y el ambiente en el cual fueron evaluados, es un modelo a dos vías de clasificación definido de la siguiente manera:

$$y_{ij} = \mu + \tau_i + \delta_j + e_{ij} \quad (2)$$

Donde y_{ij} es la respuesta observada, del *i*-ésimo genotipo en el *j*-ésimo ambiente, μ es la constante del modelo, τ_i corresponde al efecto del *i*-ésimo genotipo, δ_j corresponde al efecto del *j*-ésimo ambiente, y e_{ij} son los errores experimentales, que se suponen independientes, normalmente distribuidos con media 0 y varianza constante σ_e^2. En el modelo propuesto, tanto el efecto genotipo como el ambiente han sido considerados como fijos, esto es, definidos previamente por el investigador ya que sus efectos se suponen reproducibles. Una interpretación directa de este modelo es que la diferencia esperada entre medias de cualquier par de genotipos es independiente del ambiente donde se evalúa. Este modelo, sin embargo, no refleja la experiencia agronómica, donde es frecuente esperar rendimientos de genotipos condicionados al ambiente de prueba.

Un modelo lineal que considere la interacción (GxA) puede ser expresado de la siguiente manera:

$$y_{ijk} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + e_{ijk} \quad (3)$$

Donde y_{ijk} es la respuesta observada, del *i*-ésimo genotipo evaluado en el *j*-ésimo ambiente y repetida una k-ésima vez en el ambiente, μ es la constante del modelo, τ_i corresponde al efecto del *i*-ésimo genotipo, δ_j corresponde al efecto del *j*-ésimo ambiente, $(\tau\delta)_{ij}$ es la interacción (GxA) del *i*-ésimo genotipo con el *j*-ésimo ambiente y e_{ijk} son los errores experimentales, sobre los que hacemos los mismos supuestos descritos para el modelo (2). Desventajas que presenta este modelo es la imposibilidad de separar la interacción genotipo ambiente del error, en el caso de no tener repeticiones por ambiente, además de requerir la estimación de un número grande de parámetros que atenta directamente con la búsqueda de modelos parsimoniosos (Malosetti et al., 2013). En el caso de tener repeticiones, podemos detectar la presencia de interacciones mediante la utilización de la prueba de hipótesis específica. Si la interacción es significativa, seleccionaremos los genotipos que presenten las mayores medias ajustadas en los ambientes evaluados, las cuales pueden ser definidas a través de pruebas de comparaciones múltiples. Un aspecto negativo de este tipo de enfoque es que el esfuerzo experimental para obtener repeticiones atenta contra la exploración de más ambientes y, en consecuencia, con la posibilidad de evaluar adecuadamente la estabilidad en una cantidad apropiada de ambientes.

1.3 Modelos lineales-bilineales

En la actualidad, la utilización de modelos lineales-bilineales es aceptada en la mayor parte de los programas de mejoramiento genético, ayudando a los mejoradores a evaluar tanto la estabilidad de las características evaluadas como predecir el desempeño de nuevos genotipos evaluados bajo condiciones medioambientales variables. El nombre lineal bilineal responde a la existencia de efectos aditivos de genotipo y ambiente (lineal) y componentes multiplicativos que explican los patrones de interacción genotipo ambiente (bilineal). Junto a lo anterior, los experimentos pueden o no ser repetidos dentro de ambientes, lo que le da un valor adicional.

La formulación de un tipo particular de modelo lineal-bilineal consiste en:

$$y_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + e_{ij} \qquad (4)$$

Donde y_{ij} es la respuesta observada, del *i*-ésimo genotipo evaluado en el *j*-ésimo ambiente, μ es la constante del modelo, τ_i corresponde al efecto del *i*-ésimo genotipo, δ_j corresponde al efecto del *j*-ésimo ambiente, λ_k es el valor singular del *k*-ésimo componente bilineal (multiplicativo), α_{ik} son los elementos del *k*-ésimo vector singular izquierdo, representando la sensibilidad genotípica a un determinado ambiente, γ_{jk}, representado por el *k*-ésimo vector singular derecho. Tanto λ_k, α_{ik} y γ_{jk} son producto de la descomposición por valor singular de Z donde $z_{ij} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$ cuando hay balance o generalizando, siendo z_{ij} los residuos de un modelo aditivo para los efectos de genotipo y ambiente (Crossa, 2012). Como se menciona anteriormente, z_{ij} se obtuvo restando al valor esperado (estimado) del genotipo i en el ambiente j las medias ambientales y genotípicas y sumando la media general, por lo que este procedimiento es conocido también como "doblemente centrado" (Gauch, 2006).

El modelo (4) es conocido como "Modelo de efectos principales aditivos e interacción multiplicativa" o AMMI (Additive Main effect and Multiplicative Interaction), por sus siglas en inglés (Gauch, 1988).

Otros modelos lineales bilineales ampliamente utilizados y que tienen en común no utilizar información adicional más que una matriz de doble entrada, compuesta por las medias de rendimientos corresponden a:

a) Modelo de Regresión Genotípica o GREG (Genotypes Regression Model), por sus siglas en inglés, donde el modelo no considera la parte aditiva del componente ambiental.

$$y_{ij} = \mu + \tau_i + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + e_{ij} \quad (5)$$

b) Modelo de Regresión por Ambiente o SREG (Sites Regression Model) por sus siglas en inglés, donde el modelo no considera la parte aditiva del componente genotípico.

$$y_{ij} = \mu + \delta_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + e_{ij} \quad (6)$$

Del modelo anterior es fácil destacar que se está combinando el efecto del genotipo junto con la interacción genotipo ambiente, siendo conocido este modelo como "Modelo ambientalmente centrado" (Gauch, 2006).

c) Modelos Multiplicativos Desplazados o SHMM (Shifted Multiplicative Model) por sus siglas en inglés, donde el modelo no considera en la parte aditiva ni el efecto ambiente ni el efecto genotípico.

$$y_{ij} = \mu + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + e_{ij} \quad (7)$$

El modelo anterior, usado fundamentalmente como una técnica exploratoria, permite generar una subdivisión de los ambientes evaluados en el cual los efectos de genotipos son separados del efecto ambiente. Este modelo fue el primero en identificar grupos de genotipos o ambientes en los cuales los cambios en los rankings de genotipos serían despreciables (Crossa et al., 2006).

d) Modelo Completamente Multiplicativo o COMM (Completely Multiplicative Model) por sus siglas en inglés, no considera en la parte aditiva el efecto ambiente, el efecto genotípico ni la media general del modelo.

$$y_{ij} = \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + e_{ij} \quad (8)$$

En todos los modelos lineales bilineales descritos anteriormente, el parámetro de escala λ_k se obtiene de tal manera que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t \geq 0$. Además α_{ik} y γ_{jk} satisfacen tanto la restricción de ortogonalidad y normalización dada por:

$$\sum_{i=1}^{g} a_{ik} a_{ik'} = \sum_{j=1}^{s} \gamma_{jk} \gamma_{jk'} = 0 \text{ para todo k} \neq k', \text{y } \sum_i a_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$$

Es necesario destacar que los términos multiplicativos de los modelos GREG, SREG, COMM Y SHMM se obtienen a partir de la descomposición por valor singular de la matriz $Z'Z$ donde para los modelos GREG, $\{z_{ij}\} = \bar{y}_{ij} - \bar{y}_{i.}$; para SREG, $\{z_{ij}\} = \bar{y}_{ij} - \bar{y}_{.j}$; para COMM, $\{z_{ij}\} = \bar{y}_{ij}$ y para SHMM $\{z_{ij}\} = \bar{y}_{ij} - \bar{y}_{..}$

Operativamente en los modelos AMMI se considera, en primera instancia, ajustar un modelo lineal aditivo, originalmente un Análisis de la Varianza (ANDEVA) para luego ordenar los residuos obtenidos en la matriz Z de dimensiones GxA y finalmente aplicar la descomposición por valor singular de Z tal que $Z = UDV'$. Las primeras t columnas de U corresponden, en el modelo (4) a los vectores α_k, los primeros t elementos diagonales de D a los escalares λ_k y las t primeras columnas de V a los vectores γ_{jk}. A partir de la descomposición por valor singular de Z, se puede obtener un biplot (Gabriel, 1971), que es una representación gráfica utilizada para visualizar la interacción genotipo ambiente, representando simultáneamente, en un plano, la posición de los ambientes y los genotipos de una forma tal que sus relaciones son simples de interpretar. En el caso de los modelos AMMI, dependiendo del valor de t, la representación gráfica corresponderá a AMMI1 ($t = 1$) o de manera más habitual al modelo AMMI2 ($t = 2$). La única diferencia entre los modelos AMMI y SREG consiste en la forma de obtener los elementos que conforman Z. Esto significa que los parámetros multiplicativos encontrados en las ecuaciones 4 y 6 son diferentes, dado que la descomposición por valor singular es realizada en matrices Z diferentes (Gauch et al., 2008).

Una decisión clave en estos análisis es considerar el número óptimo de términos multiplicativos, para poder explicar los patrones de interacción, a partir de la Hipótesis nula $\lambda_k = 0$. Estas pruebas se realizan a través de un Análisis de la Varianza para modelos AMMI, donde la suma de cuadrados de la interacción representa a los autovalores $\left(\sum_{k=1}^{t} \lambda_k^2\right)$. Lo

anterior puede estar "inflado", debido a la presencia de variación inexplicable por el modelo, por lo que se hace necesario realizar un ajuste de la interacción a través de una descomposición por valor singular de la suma de cuadrados de la interacción. Dependiendo del rango de la matriz se procederá a obtener el número de componentes principales. Cada uno de los componentes en este ANDEVA representa los autovalores para cada componente, los cuales al sumarlos secuencialmente dan una aproximación a las sumas de cuadrados de la interacción, por lo tanto, de manera intuitiva, la selección de componentes se realiza en función del aporte de cada uno de ellos a la suma de cuadrados de la interacción (Hongyu et al., 2014). Es necesario considerar que los valores de las sumas de cuadrados serán decrecientes, dado que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t \geq 0$. Junto con lo anterior existen dos propuestas de construcción de estadísticos y su respectivas distribuciones, con el fin de realizar una prueba de hipótesis formal. La propuesta de Gollob (1968), conocida como modelo FANOVA ("factor ANOVA"), considera una prueba que aproxima a un *F*-test, asumiendo que la expresión $n\hat{\lambda}_m^2/\sigma^2$ se distribuye como una variable Chi-cuadrado, sugiriendo el estadístico $F = n\hat{\lambda}_m^2/f_1 \cdot s^2$ contrastándolo con una distribución F con $f1 = g + e - 1 - (2m)$ y $ge(n-1)$ grados de libertad, o $g(e-1)(n-1)$ grados de libertad si hay presentes bloques en el diseño original, permitiendo probar los m términos multiplicativos del análisis. Estudios posteriores mediante simulaciones han demostrado que lo propuesto por Gollob (1968), tiende a ser bastante liberal (presenta altas tasas de error tipo I) por lo que no es muy recomendada su utilización (Cornelius et al., 1993). Alternativas que protegen contra el error tipo I fueron propuestos por Cornelius et al. (1993) donde $F = (SS\big(GEI - \sum_{k=1}^{m} \hat{\lambda}_k^2\big)/f_2 \cdot s^2)$ presenta una distribución F con $f2 = (g-1-m)(e-1-m)$ y $ge(n-1)$ grados de libertad, o $g(e-1)(n-1)$ grados de libertad en el caso de tener bloques.

Dependiendo de cómo se obtenga Z, definida previamente, se obtienen diferentes biplots que tienen a su vez diferentes interpretaciones. Si se ajusta un modelo SREG, que no incluye el efecto principal de genotipo, los residuos de ese modelo contienen tanto el efecto genotípico como la interacción GxA y el biplot se conoce como GGE biplot. Igualmente, si se utiliza el modelo GREG, el gráfico resultante se conoce como GEE biplot. Cada uno de estos modelos y su correspondiente visualización biplot, han sido presentados para destacar distintos aspectos de la interacción GxA.

1.4 Representaciones gráficas de modelos lineales-bilineales.

La clásica manera de representar gráficamente modelos lineales bilineales ha sido mediante la utilización de gráficos biplot. Según Gabriel (1971) cualquier matriz $\boldsymbol{Y}$ de rango ≥ 2 puede ser representada a través de un biplot, el cual consiste en vectores $\boldsymbol{a_1}, \boldsymbol{a_2}, \dots, \boldsymbol{a_i}$ para las filas de $\boldsymbol{Y}$ y vectores $\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_j}$ para las columnas de **Y** de tal modo que el producto interno aproxime a la matriz $\boldsymbol{Y}$ lo mejor posible.

Siguiendo a Gabriel (1971) una matriz $\boldsymbol{Y}$ de rango r puede ser factorizado de la siguiente manera:

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{B}'$$

Donde $\boldsymbol{Y}$ es una matriz $i \times j$ de rango r, $\boldsymbol{A}$ es una matriz $i \times r$ y $\boldsymbol{B}$ es una matriz $j \times r$, ambas necesariamente de rango r. Una de las formas de factorizar la matriz $\boldsymbol{Y}$ (no la única), consiste en elegir r columnas de $\boldsymbol{A}$ como una base ortonormal del espacio fila de $\boldsymbol{Y}$. La factorización puede ser escrita de la siguiente forma:

$$y_{ij} = \boldsymbol{a_i'}\boldsymbol{b}_i$$

Donde y_{ij} es un elemento de la i-ésima fila y j -ésima columna de $\boldsymbol{Y}$, $\boldsymbol{a_i'}$ corresponde a la i-ésima fila de $\boldsymbol{A}$ y $\boldsymbol{b}_j$ corresponde a la j-ésima fila de $\boldsymbol{B}'$. De esta manera, los vectores $\boldsymbol{a_1}, \boldsymbol{a_2}, \dots, \boldsymbol{a_i}$ son asignados a cada una de las filas de $\boldsymbol{Y}$ y los vectores $\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_j}$ son asignados a cada una de las columnas de $\boldsymbol{Y}$. En una matriz de rango 2, tanto los vectores $\boldsymbol{a_1}, \boldsymbol{a_2}, \dots, \boldsymbol{a_i}$ como los vectores $\boldsymbol{b_1}, \boldsymbol{b_2}, \dots, \boldsymbol{b_j}$ son vectores de orden 2, lo que significa que los $i + j$ vectores pueden ser representados en un plano, permitiendo dar la representación de los $i\ j$ elementos de $\boldsymbol{Y}$ a través del producto interno de ambos vectores. Dado que el gráfico es capaz de representar tanto los efectos fila como columna de manera simultánea, es que este tipo de gráfico se conoce como biplot, siendo representados, al ser una matriz de rango igual a 2, de manera exacta (Gabriel, 1971).

Si la dimensión r de la matriz $\boldsymbol{Y}$ es mayor a 2, sería extremadamente útil representar una matriz $\boldsymbol{Y}$ de rango q donde $q < r$ tratando que la matriz $\boldsymbol{Y}_{(q)}$ represente lo mejor posible a la matriz $\boldsymbol{Y}_{(r)}$ donde:

$$\boldsymbol{Y}_{(q)} = \boldsymbol{A}_{(q)}\boldsymbol{B}'_{(q)}$$

La manera de encontrar $\boldsymbol{Y}_{(q)} = \boldsymbol{A}_{(q)}\boldsymbol{B}'_{(q)}$ que aproxime lo mejor posible a $\boldsymbol{Y}_{(r)}$ consistirá en minimizar la siguiente expresión:

$$\sum_i \sum_j \left(y_{ij} - y_{(q)ij}\right)^2 = traza\left[\left(\boldsymbol{Y} - \boldsymbol{Y}_{(q)}\right)\left(\boldsymbol{Y} - \boldsymbol{Y}_{(q)}\right)'\right]$$

Posteriormente a lo planteado por Gabriel (1971), Gabriel y Samir (1979) propusieron una alternativa a la expresión anterior, al agregar un ponderador w_{ij} y proceder a minimizar la siguiente expresión:

$$\sum_i \sum_j w_{ij}\left(y_{ij} - y_{(q)ij}\right)^2$$

Con respecto a ambas opciones, la alternativa más conocida para resolver la minimización consiste en lo definido por Golub et al. (1987) como "*The Eckart-Young-Mirsky matrix approximation theorem*" planteado en 1937 y conocido actualmente como "descomposición por valor singular de una matriz" (SVD, por sus siglas en ingles), representado de la siguiente forma:

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$$

Donde:

$\boldsymbol{U}$ = Matriz cuyas columnas tienen los vectores propios de $\boldsymbol{Y}\boldsymbol{Y}'$

$\boldsymbol{V}$ = Matriz cuyas columnas tienen los vectores propios de $\boldsymbol{Y}'\boldsymbol{Y}$

$\boldsymbol{D}$ = Matriz diagonal que contiene los valores propios de $\boldsymbol{Y}$

Siendo $\boldsymbol{D} = diag\{\lambda_1, \lambda_2, \ldots \lambda_r\}$ y $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$

Además, se puede verificar que: $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$

Si $q \leq r$ y consideramos que:

$\boldsymbol{U}_{(q)}$ y $\boldsymbol{V}_{(q)}$ son matrices construidas con las primeras columnas de $\boldsymbol{U}$ y $\boldsymbol{V}$ respectivamente y que $\boldsymbol{D}_{(q)}$ es una matriz diagonal de orden q de valores propios. Podemos realizar la aproximación de la matriz $\boldsymbol{Y}$ en rango q de la siguiente forma:

$$\boldsymbol{Y}_{(q)} = \boldsymbol{U}_{(q)}\boldsymbol{D}_{(q)}\boldsymbol{V}'_{(q)} = \sum_{k=1}^{q} \lambda_k u_k v'_k$$

Por lo tanto:

$$\boldsymbol{Y}_{(q)} = \boldsymbol{U}_{(q)}\boldsymbol{D}_{(q)}\boldsymbol{V}'_{(q)} = \boldsymbol{A}_{(q)}\boldsymbol{B}'_{(q)}$$

Siendo:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}^{\gamma} \text{ y } \boldsymbol{B} = \boldsymbol{V}\boldsymbol{D}^{1-\gamma}$$

Dependiendo del valor de γ utilizado, se definirán tres tipos de biplots (Gabriel, 1971):

a) GH biplot (CMP-biplot):

Este biplot considera la utilización de $\gamma = 0$, siendo finalmente:

$$\boldsymbol{A} = \boldsymbol{U} \text{ y } \boldsymbol{B} = \boldsymbol{VD}$$

Esta metodología, al preservar la métrica de las columnas presenta una alta calidad en la representación de éstas (Frutos, 2011).

b) JK biplot (RMP-biplot):

Este biplot considera la utilización de $\gamma = 1$, siendo finalmente:

$$\boldsymbol{A} = \boldsymbol{UD} \text{ y } \boldsymbol{B} = \boldsymbol{V}$$

A diferencia de la metodología GH, la utilización de un biplot JK, al preservar la métrica de las filas, obtiene una alta calidad en la representación de éstas (Frutos, 2011).

c) SQRT biplot:

Este biplot considera la utilización de $\gamma = 1/2$, siendo finalmente:

$$\boldsymbol{A} = \boldsymbol{UD}^{\frac{1}{2}} \text{ y } \boldsymbol{B} = \boldsymbol{VD}^{\frac{1}{2}}$$

Lo anterior considera una participación simétrica tanto de las filas como de las columnas de la matriz original $\boldsymbol{Y}$.

1.5 Estrategias de modelamiento

Al momento de considerar modelos lineales bilineales, en los cuales se utilizarán tanto los biplot y el análisis de componentes principales descritos anteriormente, es necesario establecer que los distintos genotipos pueden tener distintos desempeños y variabilidades en los ambientes en los cuales se están evaluando. Antecedentes en heterogeneidad en la variabilidad genética han sido reportados por Przystalski et al. (2008) en genotipos evaluados en granjas orgánicas y en granjas convencionales, observando que los rendimientos de éstos eran significativamente más altos en granjas convencionales que en las orgánicas, aunque acompañados de una variabilidad residual diferencial asociada al genotipo, por lo que es importante al momento de ajustar los modelos considerar posibles problemas de

heterogeneidad de varianza tanto en los ambientes utilizados como en los genotipos evaluados.

Una opción de modelación de datos con heterogeneidad de varianza, los cuales no pueden ser manejado mediante un modelo lineal clásico, consiste en la utilización de modelos lineales extendidos y mixtos, los cuales permiten introducir varianzas heterogéneas dentro de genotipos y/o ambientes, además de considerar posibles estructuras de correlación (West et al., 2015), junto con incluir efectos aleatorios distintos al error.

Los modelos heteroscedásticos pueden representarse mediante la incorporación de una función de varianza multiplicativa del término del error del modelo. Por ejemplo, el modelo (3) en su versión heteroscedástica tendría la siguiente representación:

$$y_{ijk} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + e_{ijk}g(v_{ijk}) \qquad (9)$$

Donde $g(.)$ es una función con imagen en los reales positivos y v_{ij} es un covariable cualitativa o cuantitativa. Igualmente, es posible en estos modelos incluir una función de correlación para superar el supuesto de falta de independencia de los errores. Modelamientos de varianzas en análisis genotipo-ambiente utilizando diversos genotipos de maíz y generando modelos de estimación más robustos los podemos encontrar en Orellana et al. (2014), mediante un enfoque bayesiano y en So y Edwards (2011), quienes determinaron que los modelos con mejores ajustes fueron los que modelaron las varianzas en ambientes específicos, destacando la importancia de considerar en un modelo la falta de heterogeneidad.

Los programas de mejoramiento han crecido tanto en escala como en complejidad. Esta complejidad se refleja en ensayos con estructura de parcela compleja que requiere métodos más avanzados de modelación. Asimismo, esta complejidad, conlleva problemas de desbalance, esto es, que no todos los genotipos se evalúan en todos los ambientes, generando un desbalance difícil de superar en el contexto de los modelos lineales clásicos. El abordaje moderno del análisis de datos en programas de mejoramiento, que contempla tanto los problemas de heteroscedasticidad, falta de independencia, la estructura compleja de parcela y la presencia de desbalance, son los modelos lineales mixtos (MLM).

Una manera de representar un MLM de forma matricial consiste en:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

Donde $\boldsymbol{y}$ es un vector de respuesta, $\boldsymbol{X}$ y $\boldsymbol{Z}$ corresponden a matrices de incidencia, $\boldsymbol{\beta}$ corresponde al vector de parámetros fijos, $\boldsymbol{u}$ corresponde al vector de efectos aleatorios y $\boldsymbol{e}$ corresponde al vector de los términos del error. Los supuestos distribucionales del modelo consideran:

$$\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G})$$
$$\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{R})$$
$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$$
$$\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$$
$$\text{cov}(\boldsymbol{u}, \boldsymbol{e}) = 0$$

Donde $\boldsymbol{G}$ corresponde a la matriz de varianzas y covarianzas del vector $\boldsymbol{u}$ de efectos aleatorios, $\boldsymbol{R}$ corresponde a la matriz de varianza y covarianzas del vector $\boldsymbol{e}$ de errores, $\boldsymbol{X}$ es la matriz de incidencia de los efectos fijos, $\boldsymbol{Z}$ es la matriz de incidencia de los efectos aleatorios y $\boldsymbol{\beta}$ el vector de parámetros de la parte fija del modelo. Los MLM más sencillos pueden considerar que los efectos aleatorios y el error presenten varianzas homogéneas, lo cual puede ser representado mediante $\mathbf{G} = \sigma_u^2 \boldsymbol{I_{UxU}}$ y $\boldsymbol{R} = \sigma_e^2 \boldsymbol{I_{nxn}}$ respectivamente, donde $\boldsymbol{I}$ representa a una matriz identidad que se diferenciará en ambos casos en función de la dimensión de la matriz $\boldsymbol{G}$ y $\boldsymbol{R}$ respectivamente. Lo anterior no limita a generar modelos en los cuales se consideren estructuras diferentes de varianzas y covarianzas para las matrices $\boldsymbol{G}$ y $\boldsymbol{R}$, lo cual es una ventaja a la hora de modelar datos heteroscedásticos y correlacionados.

Si los ambientes o los genotipos son considerados como aleatorios dentro del MLM, los efectos del genotipo "i" dentro del ambiente "j" deberán ser predichos mediante eBLUPs (empirical Best Linear Unbiased Predictor) por sus siglas en inglés, definidos de la siguiente forma (Henderson, 1984):

$$eBLUP = \widehat{\boldsymbol{G}}\boldsymbol{Z'}\widehat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

Lo anterior ha demostrado ser una alternativa capaz de generar mejores ajustes en la predicción que los obtenidos mediante la utilización de modelos que utilizan de forma exclusiva efectos fijos (Balzarini, 2000). La presente investigación se enfocará en la exploración y propuesta de modelos bajo el enfoque de modelos mixtos, en los cuales se

propondrá a su vez modelar, en la medida que se presenten, diferentes estructuras de varianzas y covarianzas con el fin de obtener modelos adecuados e informativos para un set de datos de obtenidos a partir de la materia seca de diferentes cultivos de cobertura.

2. Hipótesis y Objetivos

2.1 Hipótesis:

Las técnicas de modelación utilizadas para el análisis de la interacción genotipo x ambiente utilizadas en el mejoramiento de cereales y oleaginosas, pueden utilizarse para el análisis de estabilidad y desempeño de cultivos de cobertura. Estas técnicas, combinadas con técnicas de estimación mediante modelos mixtos, tienen la plasticidad de soportar el desbalance originado por la ausencia de algunas coberturas en algunos ambientes y también, la habilidad de manejar problemas de heteroscedasticidad, tanto de ambientes como de genotipos que se conoce, ocurren en ensayos de cultivos de cobertura.

2.2 Objetivos:

2.2.1 Objetivo General:

Hacer recomendaciones de cultivos de cobertura basadas en técnicas modernas para el análisis de la interacción genotipo x ambiente, utilizadas en los programas de mejoramiento de cultivos extensivos, pero aplicadas, en este caso, al análisis del comportamiento de cultivos de cobertura.

2.2.2 Objetivos Específicos:

- Entregar alternativas al modelado de ensayos comparativos de rendimiento de materia seca en especies no emparentadas y en diferentes ambientes considerando las posibles estructuras de correlación presentes en los ensayos, mediante la utilización de MLM.
- Considerar, mediante la utilización de MLM estructuras de varianzas que puedan estar presentes al momento de modelar los datos.

- Proponer modelos AMMI, GGE y de estabilidad a partir de la utilización de eBlups y residuos obtenidos de la interacción genotipo ambiente mediante MLM y las medias corregidas de los genotipos para datos desbalanceados.
- Ilustrar las propuestas mediante la utilización de una base de datos de ensayos de diversos cultivos de cobertura asociados a vid en diferentes ambientes.

3. Materiales y Métodos

Con el fin de determinar cultivos de cobertura óptimos como mejoradores de las condiciones del suelo, en función de su capacidad de generar materia orgánica que sea incorporada, es que durante las temporadas 2006-2009 se realizaron ensayos en distintas localidades de los valles de Copiapó y Vallenar (Atacama, Chile), con el fin de definir las especies que mejor se adaptan a los distintos ambientes (Figura 2).

Figura 2. Cultivos asociados a vid. a) avena, b) nabo forrajero, c) sorgo y d) mostaza en diferentes etapas de desarrollo.

Los ambientes fueron definidos de dos maneras, esto con el fin de cubrir los objetivos propuestos:

3.1 Caso 1: Combinación de campañas y localidades como ambientes.

Los ambientes fueron definidos como la combinación de las campañas en que se evaluaron los ensayos y la ubicación en las distintas zonas de los valles de Copiapó y Vallenar (19 ambientes), esto con el fin de evaluar estabilidad genotípica (Tabla 1 y Tabla 2).

3.2 Caso 2: Campañas utilizadas como repeticiones

Los ambientes fueron definidos como la ubicación en las distintas zonas de los valles de Copiapó y Vallenar, pero en este caso las campañas serán utilizados como repeticiones, esto con el fin de hacer recomendaciones de especies por ambiente o mega-ambiente, según las sugerencias entregadas por los GGE biplot a utilizar (7 ambientes) (Tabla 3 y Tabla 4).

La selección de las especies a evaluar corresponde a especies pertenecientes a la familia *Fabaceae* (leguminosas), *Brassicaseae* (crucíferas) y *Poaceae* (gramíneas). Es necesario destacar que ninguna de las especies utilizadas en los ensayos se siembra de manera comercial en la zona, dado que las condiciones climáticas no permiten un desarrollo eficiente de estos cultivos, por lo que se definió establecer mediciones exclusivamente en función de la producción de materia seca. La Tabla 1 muestra los cultivos de cobertura utilizados.

Tabla 1. Especies evaluadas en los distintos ambientes. Caso 1.

Especie	**Tipo de siembra**	**N° Ambientes en los que fue evaluados**
Avena	Sobre y entre hilera	7
Cebada	Sobre-hilera	6
Chícharo	Sobre-hilera	4
Haba	Sobre-hilera	9
Maíz	Sobre-hilera	4
Mostaza	Sobre y entre hilera	8
Nabo	Sobre-hilera	7
Raps	Sobre-hilera	11
Sorgo	Sobre-hilera	4

Cabe destacar que en los ambientes evaluados, tanto en el caso 1 como en el 2 no se repitieron siempre todas las especies, por lo que los datos presentan un desbalance implícito que deberá ser considerado.

Al momento de montar cada ensayo en un ambiente, estos fueron establecidos mediante un diseño en bloques completos aleatorizados con 5 repeticiones. Una de las variables obtenida de los cultivos correspondió a la producción de materia seca·m^{-2} de superficie, la cual será utilizada como variable a analizar. La Tabla 2 muestra el desbalance que presentan los datos, observándose que existe aproximadamente un 65% de celdas vacías. Los ambientes que evaluaron un mayor porcentaje de especies corresponden a Cneu-Hol9 (88,9% de especies), CNeu-Ho7 y Cneu-Ho8 (66,7% de especies) y Cagro-Gh9 (55,6% de especies).

Tabla 2. Presencia de los cultivos en los ambientes evaluados. Caso 1.

Cultivo/Ambiente	Cagro-Gh7	Cagro-Gh8	Cagro-Gh9	Cneu-Ho6	CNeu-Ho7	Cneu-Ho8	Cneu-Hol9	CPro6	CRio6	CRio7	CRio9	CUni7	CUni8	CUni9	VAlv6	VAlv7	VAlv8	VAlv9	VGae6
Avena	0	1	0	0	1	0	1	0	1	0	1	0	0	0	0	0	1	1	0
Cebada	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Chícharo	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
Haba	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0
Maíz	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Mostaza	1	0	1	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	0
Nabo	0	0	0	0	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0
Raps	0	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0
Sorgo	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

1: Presencia de la especie en el ambiente. 0: Ausencia de la especie en el ambiente.

Tabla 3. Especies evaluadas en los distintos ambientes. Caso 2.

Especie	**Tipo de siembra**	**N° Ambientes en los que fue evaluados**
Avena	Sobre y entre hilera	4
Cebada	Sobre-hilera	3
Chícharo	Sobre-hilera	2
Haba	Sobre-hilera	5
Maíz	Sobre-hilera	2
Mostaza	Sobre y entre hilera	5
Nabo	Sobre-hilera	3
Raps	Sobre-hilera	6
Sorgo	Sobre-hilera	2

Tabla 4. Presencia de los cultivos en los ambientes evaluados (Caso 2)

Especies	Am1	Am2	Am3	Am4	Am5	Am6	Am7
Avena	1	1	0	1	0	1	0
Cebada	1	0	0	1	1	0	0
Chícharo	0	0	1	1	0	0	0
Haba	1	1	0	1	1	1	0
Maíz	0	0	1	1	0	0	0
Mostaza	1	1	0	1	0	1	1
Nabo	0	0	0	1	0	1	1
Raps	1	1	0	1	1	1	1
Sorgo	1	0	0	1	0	0	0
Total	6	4	2	9	3	5	3
% presencia sp.	66,7	44,4	33,3	100,0	33,3	55,6	33,3

1: Presencia de la especie en el ambiente. 0: Ausencia de la especie en el ambiente.

En el Caso 2 los ambientes que presenta el mayor desbalance fueron los ambientes 5 y 7, con un 33,3% de las especies evaluadas.

Lo anterior hace necesario buscar alguna metodología que permita realizar una estimación de los datos ausentes. La propuesta implicó utilizar técnicas de imputación que aprovechen la información de cultivos que sí estuvieron en ambientes determinados, procediendo a probar distintas metodologías de imputación descritas en la literatura. Estas metodologías serán previamente sometidas a una base de datos obtenida de la literatura donde se evaluaron diferentes líneas de una determinada especie en todos los ambientes definidos, procediendo a eliminar progresivamente porcentajes de los datos observados y siendo estimados mediante estas metodologías. El criterio, definido más adelante, involucra estimar los valores ausentes y escoger la metodología que presente la menor discrepancia con el valor observado.

3.3 Análisis estadístico

3.3.1 Caso 1

Se consideró modelar los datos provenientes de los 19 ambientes mediante MLM. Una de las opciones para el modelo fue considerar las especies como efectos fijos y tanto los ambientes como los bloques, los cuales se encuentran anidados a los ambientes, como efectos aleatorios (modelo AMMI), y la otra opción solo incluir los ambientes y los bloques anidados a los ambientes, siendo ambos efectos aleatorios (Modelo GGE). Una vez obtenidos los modelos ajustados, se utilizaron los índices AIC (Akaike Information Criterion) y BIC (Bayesian Information Criterion), ambos criterios de verosimilitud penalizada, utilizados para definir los

modelos más parsimoniosos y que cumplan además con el supuesto de normalidad de los errores y contemplen problemas de heteroscedasticidad. Los criterios se definieron como:

$$\text{AIC} = -2\ln Lik + \ 2n_{par}$$
$$\text{BIC} = -2\ln Lik + \ n_{par}Log(N)$$

Donde Lik es el valor maximizado de la función de verosimilitud, n_{par} es el número de parámetros en el modelo y N el total de observaciones usadas para realizar el ajuste. Bajo esta definición, se consideró que mientras más bajo sea el valor obtenido, y manteniendo sin alterar el número de observaciones y la parte fija del modelo, mejor es el modelo estimado.

La estimación de los modelos se realizó mediante la librería `nlme` (Pinheiro et al., 2018) en el lenguaje de programación R (R core team, 2018).

El modelo teórico propuesto para estos efectos correspondió a:

$$\boldsymbol{y} = \mathbf{X\beta} + \boldsymbol{Z_1}\mathbf{a} + \boldsymbol{Z_2}\mathbf{b} + \boldsymbol{Z_3}\mathbf{c} + \mathbf{e}$$

Donde $\boldsymbol{y}$ es un vector de respuesta, $\boldsymbol{X}$ corresponde a una matriz de incidencia del efecto fijo (cultivo); $\boldsymbol{\beta}$ corresponde al vector de parámetros de los efectos fijos; $\mathbf{Z_1}$ corresponde a la matriz de incidencia de los ambientes (efecto aleatorio); **a,** corresponde al vector de efectos aleatorios de los ambientes; $\mathbf{Z_2}$ corresponde a la matriz de incidencia de los efectos aleatorios de los bloques; $\boldsymbol{b}$ corresponde al vector de efectos aleatorios de los bloques; $\mathbf{Z_3}$ corresponde a la matriz de incidencia de la interacción especie (cultivos) y ambientes; **c** corresponde al vector de efectos aleatorios de la interacción genotipo (cultivo) y ambiente y $\boldsymbol{e}$ corresponde al vector de los términos del error. Los supuestos distribucionales del modelo consideraron: $\boldsymbol{e}\sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$; $\boldsymbol{a}\sim N(\mathbf{0}, \sigma_a^2\boldsymbol{I})$; $\boldsymbol{b}\sim N\left(\mathbf{0}, \sigma_b^2\boldsymbol{I}\right)$; $\boldsymbol{c}\sim N(\mathbf{0}, \sigma_c^2\boldsymbol{I})$ respectivamente. Además, se suponen que todos los efectos aleatorios son independientes.

En la estimación de los componentes de varianza mediante el algoritmo REML, fue posible estimar los componentes asociados al error experimental, efecto bloque, efecto ambiente y la interacción genotipo ambiente, lo que permitió estimar el porcentaje de la variabilidad de la cual es responsable cada componente.

Con respecto a la interacción genotipo (efecto fijo) y ambiente (efecto aleatorio) modelados mediante MLM, esta interacción es un efecto aleatorio, por lo que los eBLUP de la interacción

pueden obtenerse directamente, los cuales fueron utilizados para realizar los biplots, mediante la utilización de la librería `GGEBiplotGUI` (Frutos et al., 2014). Los BLUPs, según lo definido previamente, incorporan los posibles problemas heteroscedásticos de los errores. Esta heteroscedasticidad puede corregirse mediante la utilización de funciones de varianza, lo que permite tener mayor flexibilidad que la tradicional utilización del ANDEVA. Antecedentes de la utilización de BLUPs en modelos del tipo interacción genotipo ambiente pueden encontrarse en Piepho (1994) quien los utilizó como alternativa a la utilización de modelos AMMI obtenidos a partir del ANDEVA, sugiriendo su uso en grandes bases de datos.

3.3.2 Caso 2

En el Caso 2 en el cual los ambientes están definidos como la ubicación en las distintas zonas de los valles de Copiapó y Vallenar, siendo los años utilizados como repeticiones, los datos fueron modelados a través de MLM en el cual se consideró el ambiente, los años y bloques como efectos aleatorios y los cultivos como efectos fijos. Con la estimación del modelo propuesto se obtuvieron los eBLUP, como se hizo previamente en el Caso 1, siendo éstos corregidos por la suma de las medias estimadas según el modelo de cada cultivo menos la media general de los cultivos. Como se menciona previamente y dado que en ambos casos no se podrá obtener una matriz completa de ambientes con sus respectivos genotipos, se hizo necesario imputar los datos faltantes, para poder realizar de forma adecuada la modelación estadística, existiendo la posibilidad de modelar los patrones de interacción incluso en ausencia del 60% de los datos faltantes (Yan, 2013).

3.3.3 Tipos de datos faltantes

Con respecto a la ausencia de datos, Rubin (1976), considera tres tipos de datos faltantes: MCAR, MAR y MNAR. Estos 3 mecanismos establecen la relación de las variables analizadas y la probabilidad de ausencia de datos, lo cual es importante al momento de definir la técnica de imputación. En primer lugar, una variable es MCAR (Missing Completely at Random) si la probabilidad de pérdida de una observación en una variable X para todos los individuos es la misma y no depende de otras variables (Baraldi and Enders, 2009). En segundo lugar, una variable es MAR (Missing at Random) si la observación perdida está relacionada con otras variables medidas en el análisis. Se debe considerar que este no es un mecanismo aleatorio (a pesar de que en su nombre aparece la palabra “aleatorio”), ya que la ausencia de un valor dado está correlacionada con otras variables analizadas en el mismo estudio (Baraldi and

Enders, 2009). Finalmente, una variable es MNAR (Missing Not at Random) si la probabilidad de que la observación perdida se relaciona directamente con el valor perdido (por ejemplo, en una encuesta no hay respuesta por no entender la pregunta).

Considerando lo anterior, se propuso definir una metodología de imputación de datos para los dos casos, dentro de una amplia variedad de opciones que suponen distintos modelos de pérdida de datos.

3.3.4 Estrategias de imputación

Las estrategias de imputación se realizaron de la siguiente manera, y de forma similar a lo planteado por Schmitt et al. (2015):

1) Se utilizó una base de datos en la cual se haya realizado un análisis de interacción genotipo ambiente y que sea evidente la existencia de interacción. Para lo anterior se determinó utilizar una base de datos correspondiente a 18 cultivares de trigo evaluados en 9 localidades en Ontario Canadá (Xu, 2010), siendo esta base de datos ampliamente expuesta en la literatura de modelos bilineales.
2) Se procedió a introducir un 10, 20, 40 y un 60% de datos ausentes en forma aleatoria al set de datos.
3) Se realizó la imputación de los datos a través de 5 métodos de imputación presentes en la literatura.
4) La performance de los resultados se evaluó a través de la raíz cuadrada de error cuadrático medio (RMSE por sus siglas en inglés) de acuerdo a la siguiente fórmula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i^{obs} - X_i^{esp}\right)^2}{n}}$$

Donde:

X_i^{obs} Corresponden a los valores observados en los datos.

X_i^{esp} Corresponden a los valores imputados.

5) Lo anterior se repitió 1000 veces y se promediaron por cada simulación realizada el resultado de RMSE obtenido, por porcentaje de dato ausente.

Dentro de las técnicas de imputación se utilizaron las siguientes:

3.3.4.1 Imputación valor medio

A pesar de no ser una técnica muy recomendada, se consideró esta técnica como la alternativa más básica al momento de realizar una imputación. Simplemente consiste en la imputación del dato ausente a través del valor medio de la variable, en este caso la media del ambiente.

3.3.4.2 MICE (Multiple Imputation by Chained Equations)

La imputación por cadena de ecuaciones, o también conocida como imputación MICE es una técnica ampliamente descrita en la literatura. MICE asume que los datos ausentes son MAR, lo que significa que la probabilidad que un valor se pierda depende sólo de los valores observados, pudiendo ser predicho a través de éstos. Los pasos para la imputación son los siguientes (Azur et al., 2011):

Paso 1: Se imputa simplemente la media del dato faltante por ambiente. El valor se conoce como "Place holder"

Paso 2: El dato imputado, para una variable, vuelve a ser considerado como ausente.

Paso 3: Los valores observados de la variable en el paso 2 son considerados como la variable dependiente, la cual es utilizada en un modelo de regresión, utilizando el resto de las variables como variables independientes.

Paso 4: El valor ausente es reemplazado con la predicción (es lo que corresponde a la imputación).

Paso 5: Los pasos del 2–4 se repiten para cada variable que ha perdido datos. Una iteración para cada variable con dato ausente corresponde a un ciclo.

Paso 6: el paso 2 a 4 puede ser repetido en varios ciclos, en los cuales la imputación será actualizada. La decisión en el número de imputaciones a realizar es definida por el investigador.

3.3.4.3 Imputación mediante Random Forest (Missing Forest):

El algoritmo Random Forest ha demostrado ser una alternativa a la imputacion de datos faltantes principalmente por su flexibilidad de tratar con múltiples tipos de datos ausentes (cuantitativos-cualitativos). Junto a lo anterior, es una técnica no paramétrica que permite considerar interacciones y relaciones no lineales entre las variables analizadas (Stekhoven y Bühlmann, 2012). Esta metodología se basa en la imputación de los datos faltantes mediante la predicción utilizando el algoritmo Random forest.

Los pasos para la imputación son los siguientes (Stekhoven y Bühlmann, 2012):

Paso 1: Para una variable arbitraria (Xs), incluidos los valores ausentes, se separa los datos en 4 partes:

a) Valores observados de la variable Xs ($y_{obs}^{(s)}$)
b) Valores ausentes de la variable Xs ($y_{miss}^{(s)}$)
c) Variables que no sean Xs con observaciones $i_{obs}^{(s)}$ denominado ($x_{obs}^{(s)}$)
d) Variables que no sean Xs con observaciones $i_{miss}^{(s)}$ ($x_{miss}^{(s)}$)

Paso 2: A una matriz con datos faltantes se procede a imputar generalmente la media de la variable u otro medio de imputación.

Paso 3: Se procede a ordenar el número de variables de acuerdo con la cantidad de datos faltantes, partiendo con las variables con las cantidades más bajas.

Paso 4: Para cada variable los valores son ajustados primero a través de la imputación Random Forest, usando como variable respuesta a $y_{obs}^{(s)}$ y como predictores los $x_{obs}^{(s)}$

Paso 5: Predecir valores ausentes $y_{miss}^{(s)}$ utilizando el proceso de entrenamiento Random Forest a los $x_{miss}^{(s)}$

Paso 6: El criterio de detención γ se cumple tan pronto como la diferencia entre los elementos imputados en el paso j y los imputados en el paso $j-1$ aumenta por primera vez con respecto a ambos tipos de variables. La diferencia para el set de matrices se puede definir como:

$$\Delta N = \frac{\sum_{j \in N} \left(X_{New}^{imp} - X_{old}^{imp}\right)^2}{\sum_{j \in N} \left(X_{New}^{imp}\right)^2}$$

3.3.4.4 SVD imputation algorithm

Los pasos para definir los datos faltantes mediante esta metodología se pueden definir de la siguiente forma (Troyanskaya et al., 2001):

Paso 1: La descomposición por valor singular (SVD) debe ser realizada previamente con una matriz completa, por lo tanto las filas ausentes son sustituidas al inicio por los valores promedios de las columnas, o ceros si no hay presencia.

Paso 2: Se realiza la descomposición por valor singular de la matriz completa, en las matrices **U**, **D** y **V** tal que **UDV´** es la matriz original.

Paso 3: Los valores ausentes son predichos en la matriz reconstruida utilizando los 2 primeras columnas de **U** y **V** y los dos primeros elementos diagonales de **D**.

Paso 4: Los datos predichos son utilizados para llenar nuevamente la matriz de datos originales.

Paso 5: El proceso se repite desde el Paso 2 hasta que la discrepancia entre valores sucesivamente imputados es despreciable. El criterio para detener la iteración consiste en minimizar la siguiente expresión:

$$d = \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - x_i^{´}\right)^2\right]^{1/2}$$

Donde x corresponde a la última imputación, $x^{´}$ a la penúltima imputación y n al total de datos ausentes. Un cálculo adicional corresponde a la gran media de la matriz de doble entrada, utilizada como alternativa a la utilización de una media simple, definida de acuerdo a la siguiente fórmula:

$$\bar{y} = \left[\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}^2\right]^{1/2}$$

Donde y_{ij} corresponde al valor en la i fila, j a la columna de los datos presentes y N corresponde al número de datos existentes. El criterio para finalizar la iteración de dos sucesivas estimaciones se da cuando $d/\bar{y} < 0{,}01$ (Yan, 2013).

3.3.4.5 K Nearest Neighbors (KNN)

Un método no paramétrico aplicable a las tres condiciones de datos ausentes descritos anteriormente (MCAR, MAR y MNAR.) es conocido como imputación basada en el vecino más cercano. Se basa en encontrar los k vecinos de la observación que presenta algún atributo ausente, donde se consideran como información para su estimación los atributos que si están presentes en los vecinos. Este dato ausente, dentro de la observación, es reemplazado con el dato más frecuente, si este es categórico o con la media de sus vecinos si el dato es continuo (Zhang, 2012). Este tipo de imputación se conoce también como método “Hot Desk” donde la imputación de datos faltantes de una matriz se realiza a partir de datos presentes en la misma

matriz (Joenssen and Bankhofer, 2012), asumiendo que un valor dado puede ser aproximadamente el valor de los puntos cercanos a éste, basados en el resto de las variables analizadas. La utilización de esta metodología depende de:

1. La métrica de distancia para definir cercanía, siendo la más utilizada la métrica Minkouski donde la distancia entre dos puntos es definida como:

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Cuando $p = 1$ corresponde a la distancia Manhattan y para $p = 2$ corresponde a la distancia Euclídea.

2. El número de vecinos a utilizar (k). Se considera que a menor k mayor será el ruido captado y menos generalizable será el resultado, y a mayor k se tienden a dispersar los efectos locales.
3. El método de agregación (media, mediana, moda, etc.).
4. Transformación de los datos. Algunas sugerencias de transformación de datos, tales como normalización o log-transformación se han sugerido para atenuar el efecto de datos outliers.

Una vez realizada las imputaciones de los datos faltantes utilizando las metodologías anteriormente descritas, se procedió a graficar el desempeño de éstas, considerando que a menor discrepancia entre los valores observados y los esperados (RMSE) mejor será la técnica de imputación definida, y por lo tanto la utilizada para la base de datos de cultivos.

4. Resultados y discusión

4.1 Estimación de los modelos

4.1.1 Caso 1

El modelo propuesto consideró como variable respuesta el rendimiento de materia seca por metro cuadrado de superficie y como criterios de clasificación los cultivos, los ambientes y los bloques en donde se realizaron los ensayos. La parte fija consistió en los cultivos y como efectos aleatorios los ambientes, los bloques anidados a los ambientes y la interacción de los

cultivos con los ambientes. Con el fin de explorar la interacción genotipo x ambiente, se obtuvieron los BLUPs de la interacción. El modelo propuesto corresponde a:

$$y_{ijk} = \mu + \tau_i + a_j + (\tau a)_{ij} + b_{k(j)} + e_{ijk} g(v_{ijk})$$

Donde:

$y_{ijk} =$ es la respuesta observada (rendimiento materia seca), del *i*-ésimo cultivo evaluado en el *j*-ésimo ambiente y en el k-ésimo bloque, μ es la constante del modelo, τ_i corresponde al efecto del *i*-ésimo cultivo, a_j corresponde al efecto del *j*-ésimo ambiente, $(\tau a)_{ij}$ corresponde a a la interacción dej i-ésimo cultivo con el j-ésimo ambiente, $b_{k(j)}$ corresponde al k-ésimo bloque, anidado al j-ésimo ambiente, e_{ijk} corresponde a los errores, $g(v_{ijk})$ es una función con imagen en los reales positivos y v_{ijk} es un covariable cualitativa o cuantitativa.

Dentro de la estrategia de modelación, se consideraron modelos homoscedásticos, verificándolos a través de herramientas gráficas, como gráficos de residuos versus predichos y qq-plot normal. En primera instancia no se pudo generar alguna alternativa razonable de un modelo que cumpliera los supuestos mencionados anteriormente, por lo que se procedió a modelar los datos a través de una transformación log_{10} observándose un muy buen ajuste sobre los datos modelados.

La Tabla 5 muestra los resultados obtenidos en función de 5 modelos propuestos, partiendo de un modelo homoscedástico (modelo 1) y cuatro modelos heteroscedásticos (modelo 2-5) en función del índice AIC y el índice BIC. Ambos índices consideran la bondad de ajuste como un factor positivo y la complejidad (número de parámetros), como un factor negativo. Ambos índices tienen en consideración una penalización directamente relacionada con el número de parámetros estimados, esto con el fin de evitar un posible sobreajuste del modelo. La interpretación de estos índices, cuando el método de estimación es REML, es mediante la comparación de modelos que mantengan inalterada su parte fija. Se considerará un modelo parsimonioso mientras más bajo sean ambos índices. Lo anterior es válido incluso para valores negativos que podrían obtenerse.

Con el fin de verificar los supuestos de homogeneidad de varianza y normalidad de los errores, se realizaron gráficos de residuos versus predichos para el primer supuesto y un

gráfico qq-plot normal para el segundo. Aparentemente, observando el primer gráfico (residuos versus predichos) se consideró, en primera instancia una ligera violación a este supuesto, proponiéndose 4 modelos heteroscedásticos, construidos a través de funciones de varianza.

Con respecto a lo anterior, Pinheiro y Bates (2000) definen una función general de varianza en un modelo linear mixto extendido de la siguiente forma:

$$Var(e_{ij}|a_j) = \sigma^2 g^2(\mu_{ij}, v_{ij}, \delta)$$

Donde:
$g(\cdot)$ es una función con imagen en los reales positivos, conocida como función de varianza.
$\mu_{ij} = E[y_{ij}|a_j]$ es la esperanza condicional al efecto aleatorio, en este caso, será el ambiente.
v_{ij} es un vector de covariables.
δ es un vector de parámetros de g(·).

En un experimento multinivel (efecto ambiente y efecto bloque, tal como se plantea en el caso 1) lo anterior puede generalizarse de la siguiente manera:

$$Var = (e_{ijk}|a_j, b_{k(j)}) = \sigma^2 g^2(\mu_{ijk}, v_{ijk}, \delta)$$

Donde:
$\mu_{ijk} = E[y_{ijk}|a_j, b_{k(j)}]$ lo que corresponde a la esperanza condicional a ambos efectos aleatorios (ambiente y bloque anidado a ambiente).

Lo anterior supone que, al incluir efectos aleatorios al modelo, los errores intra grupo y los efectos aleatorios ya no pueden ser asumidos como independiente, generando dificultades computacionales en su estimación (Pinheiro y Bates, 2000). Lo anterior se soluciona integrando los efectos aleatorios a la función de varianza, reemplazando los valores esperados μ_{ij} por sus respectivos valores predichos:

$$\hat{\mu}_{ijk} = x_i^T \beta + z_j^T \hat{a}_j + z_k^T \hat{b}_k$$

Finalmente:

$$Var(e_{ijk}) \cong \sigma^2 g^2(\hat{\mu}_{ijk}, v_{ijk}, \delta)$$

Bajo esta propuesta, se puede asumir que los errores intragrupo se comportan independientes de los efectos aleatorios, posibilitando la utilización de la expresión planteada.

Siguiendo a Pinheiro y Bates (2000) las alternativas que plantean estas funciones de varianzas $g(\cdot)$ son bastante flexibles. Un modelo que considere que la variabilidad residual aumenta a través de la potencia del valor absoluto de una covariable tiene la siguiente forma:

$$Var(e_{ijk}) = \sigma^2 |v_{ijk}|^{2\delta}$$

Donde la función de varianza corresponde a:

$$g(v_{ijk}, \delta) = |v_{ijk}|^{\delta}$$

Una segunda alternativa, que considere que la variabilidad residual aumenta de manera exponencial se expresa de la siguiente manera:

$$Var(e_{ijk}) = \sigma^2 \exp(2\delta v_{ijk})$$

Donde la función de varianza corresponde a:

$$g(v_{ijk}, \delta) = \exp(\delta v_{ijk})$$

El parámetro, al no tener restricciones puede considerar modelos que la varianza se incremente o disminuya con la covariable utilizada.

Una tercera opción corresponde a una función capaz de estimar tantas varianzas como estratos se hayan definido previamente. Sea el número de estratos definidos desde $\{1,2, \dots, S\}$, luego la varianza se define como:

$$Var(e_{ijk}) = \sigma^2 \delta^2_{Sijk}$$

Donde la función de varianza corresponde a:

$$g(S_{ijk}, \delta) = \delta_{ij}k$$

Esta propuesta estima un total de S varianzas residuales.

Dadas las propuestas existentes sobre modelos heteroscedásticos, se procedió a estimar 5 modelos, siendo el primero un modelo homoscedástico y cuatro alternativas en los cuales se consideraron modelos que introducen funciones de varianzas (Tabla 5). La decisión de las cuatro alternativas heteroscedásticos se basaron principalmente en los gráficos de residuos versus predichos.

Tabla 5. Comparación de modelos propuestos.

Modelos	AIC	BIC
1) Homoscedástico	63,98	113,68
2) Heteroscedástico. Potencia	65,88	119,40
3) Heteroscedástico. Por ambiente	124,32	174,02
4) Heteroscedástico. Por ambiente (2)	110,64	160,35
5) Heteroscedástico. Exponencial	130,35	180,05

AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion. Valores menores implica mejor.

Con el fin de evaluar los dos modelos con menores valores de AIC y BIC (modelo 1 y 2), se realizó una prueba de cociente de verosimilitud para verificar la significancia del modelo heteroscedástico, según la siguiente construcción del estadístico LR:

$$LR = 2logLik_{modelo\ 2} - 2logLik_{modelo\ 1}$$

El cual se distribuye como una $\chi^2_{gl\ modelo\ 2 - gl\ modelo\ 1}$

La Tabla 6 muestra que no existen diferencias significativas (p-value = 0,7458) del modelo 2 (Heteroscedástico) con respecto al modelo 1 (homoscedástico). Lo anterior permite seleccionar a este último, siendo el definitivo sobre el cual se realizarán los análisis.

Tabla 6. Test de cociente de verosimilitud entre el modelo 1 (homoscedástico) y el modelo 2 (heteroscedástico).

Modelo	gl	AIC	BIC	logLik	Test	LR	p-value
1)	13	63,98	113,68	-18,9908			
2)	14	65,87	119,40	-18.9382	1 vs 2	0,105132	0,7458

gl: grados de libertad. AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion. logLik: logaritmo de la verosimilitud del modelo; LR; Cociente de verosimilitud.

Con respecto al factor fijo, se procedió a realizar un test de Wald, esto con el fin de verificar si existen diferencias entre los cultivos evaluados, existiendo diferencias significativas entre ellos ($\chi^2 = 49{,}896; df = 8;\ p - value < 0{,}001$). A su vez la estimación REML de los componentes de varianza de ambiente, bloque, interacción y error correspondieron a $\sigma_{amb.} = 0{,}302$ $\sigma_b = 0{,}0000134$, $\sigma_{amb \times cult} = 0{,}191$ y $\sigma_e = 0{,}204$ respectivamente.

Al modelo homoscedástico definido previamente, se procedió a la predicción de los BLUPs correspondientes a la interacción GxA. Como se mencionó anteriormente, el conjunto de datos presenta desbalance, el cual se manifiesta al momento de calcular los BLUPs (la matriz de eBLUPs[2] predicha entrega valores iguales a cero al utilizar la librería `nlme` en la combinación de cultivo x ambiente donde no se obtuvieron datos experimentales) por lo cual se hizo fundamental plantear alguna estrategia que considere imputar los datos ausentes sobre dicha matriz.

4.1.1.1 Imputación de datos faltantes

Con el fin de determinar una alternativa a la presencia de datos faltantes mediante alguna técnica de imputación ya existente, es que se procedió a utilizar una base de datos sobre la cual se realizó una simulación de datos faltantes. Esta base de datos fue obtenida a partir de los resultados de Xu (2010) en un experimento realizado sobre diferentes genotipos de trigo en Ontario, Canadá en diferentes localidades. La variable evaluada correspondió al rendimiento del cultivo. Los métodos de imputación definidos y descritos previamente correspondieron a: 1) Media, 2) Multiple Imputation by Chained Equations (MICE), 3) Random Forest, 4) SVD (impute algorithm using singular value decomposition) y 5) k-Nearest Neighbors (KNN). De forma aleatoria fueron eliminados 10, 20, 40 y 60% del total de los datos de rendimiento y luego imputados mediante las 5 metodologías, repitiéndose durante ciclos de 1000 veces. Para medir la discrepancia entre los valores imputados y los eliminados, se estimó la raíz cuadrática media del error (RMSE), esperando que la metodología propuesta nos entregue los menores valores posibles en los 4 escenarios planteados.

La Figura 3 muestra los resultados de las respectivas imputaciones. Random Forest correspondió a la metodología que presentó una menor discrepancia entre los valores

[2] De aquí en adelante se llamarán BLUP.

observados y los predichos, para todos los escenarios de datos ausentes (10, 20, 40 y 60%), siendo esta metodología utilizada en la imputación de los BLUPs ausentes.

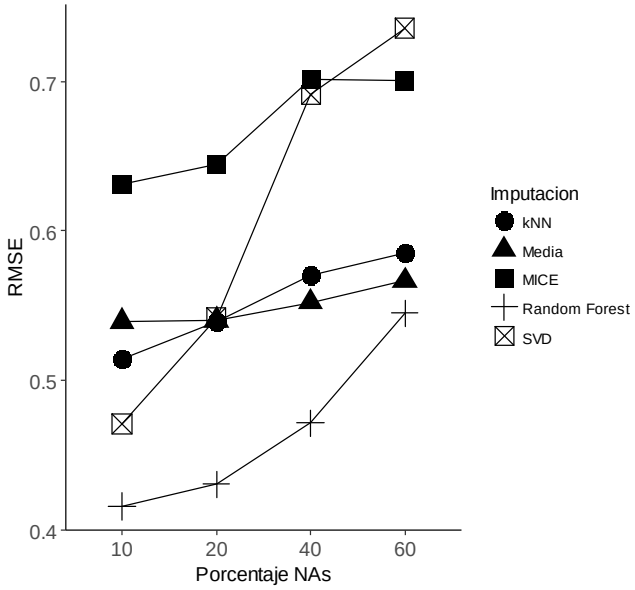


Figura 3. Media de la raíz del error cuadrático medio de predicción obtenido a partir de la simulación de 1000 ciclos con 10, 20, 40 y 60% de datos ausentes (NAs). kNN: imputación *K Nearest Neighbors*, Media: Imputación mediante el valor medio de la columna, MICE: Imputación mediante *Multiple Imputation by Chained Equations*, Random forest: Imputación mediante *Random forest* (Missing Forest), SVD: Imputación mediante *singular value descomposition* (impute algorithm). RMSE: Media de la raíz del error cuadrático medio de predicción.

Observando el desempeño de las imputaciones, Random Forest fue la metodología que mantuvo siempre los menores valores, independiente del porcentaje de datos faltantes. Resulta interesante destacar el gran aumento en la discrepancia entre los valores observados e imputados en la metodología SVD, la cual teniendo la segunda menor discrepancia al ser evaluada con un 10% de datos ausentes (la menor discrepancia con ese valor fue Random Forest) termina con la mayor discrepancia al 60% de datos ausentes.

La Tabla 7 presenta la estadística descriptiva para las diferentes metodologías empleadas, en función de los porcentajes de datos eliminados.

Tabla 7. Estadística descriptiva para RMSE por metodología.

		Porcentaje de valores eliminados			
Método	Estadístico	10%	20%	40%	60%
kNN	Media	0,51	0,54	0,57	0,59
	DE	0,13	0,08	0,06	0,05
	Mediana	0,50	0,53	0,57	0,58
	Q1	0,42	0,48	0,53	0,56
	Q3	0,59	0,59	0,61	0,61
Media	Media	0,54	0,54	0,55	0,57
	DE	0,10	0,07	0,04	0,03
	Mediana	0,54	0,54	0,55	0,57
	Q1	0,46	0,49	0,52	0,54
	Q3	0,61	0,59	0,58	0,59
MICE	Media	0,63	0,64	0,70	0,70
	DE	0,11	0,08	0,06	0,05
	Mediana	0,63	0,64	0,69	0,69
	Q1	0,54	0,59	0,65	0,67
	Q3	0,69	0,67	0,77	0,73
Random F.	Media	0,42	0,43	0,47	0,54
	DE	0,08	0,06	0,06	0,06
	Mediana	0,41	0,42	0,47	0,54
	Q1	0,36	0,38	0,43	0,50
	Q3	0,46	0,47	0,51	0,58
SVD	Media	0,47	0,54	0,69	0,74
	DE	0,15	0,18	0,17	0,12
	Mediana	0,45	0,50	0,66	0,73
	Q1	0,37	0,43	0,56	0,66
	Q3	0,53	0,57	0,79	0,80

Experiencias en imputación de datos faltantes sobre tablas de doble entrada del tipo genotipo ambiente las podemos encontrar en Yan (2013) quien generando una simulación de imputación de datos desde un 5% de datos ausentes hasta un 60% de datos ausentes, y utilizando una única imputación, mediante descomposición por valor singular (SVD), similar a la presentada en esta simulación, comparó dos bases de datos (18 genotipos × 9 ambientes y otra de 15 genotipos × 40ambientes). Para ambas destaca que es posible rescatar patrones de interacción al ser graficados y observados mediante biplots, tanto para patrones de ausencia de un 40 y un 60% de datos ausentes respectivamente.

La imputación definida finalmente para los datos fue la técnica Random Forest, la que se aplicó directamente sobre la matriz de BLUPs para la interacción genotipo x ambiente. Según Stekhoven y Bühlmann (2012) el potencial de las imputaciones mediante Random Forest radica en que no es necesario asumir ninguna distribución de los datos (metodología no paramétrica), puede imputar tanto datos cuantitativos y cualitativos, puede ser utilizado cuando existen más variables que observaciones y además esta metodología expresa todo sus atributos cuando los datos incluyen tanto interacciones complejas como relaciones no lineales, lo cual es clave cuando se realizan imputaciones de rendimientos de genotipos en diferentes ambientes.

4.1.1.2 Modelo AMMI

Una vez obtenida la imputación, se realizó la descomposición por valor singular de la matriz imputada previamente mediante Random Forest. La Figura 4 muestra el modelo AMMI (2) (utilización de dos componentes principales para obtener la típica representación de la interacción en el plano) en los cuales la representación de los ambientes es a través de segmentos de recta que parten del origen del sistema de coordenadas y la representación de los cultivos en los puntos identificados con sus nombres. Esta representación es un biplot de componentes principales (CMP-biplot). Las dos primeras componentes resumen un 71% de la variabilidad de la Tabla de BLUPs, (la primera CP corresponde al 41 % de la variabilidad y la segunda componente representa el 30%) lo cual es suficiente para expresar los patrones de interacción. Observando la primera componente (eje X) los ambientes Valv9 Cagro-Gh8 y Valv8, al poseer la mayor distancia desde el origen, son los ambientes más extremos en esa componente, lo cual significa que ellos explican en mayor porcentaje los patrones de interacción y los cambios más importantes que se puedan observar en genotipos son precisamente en esos ambientes. Por su parte, al observar la segunda componente (eje y) los ambientes más extremos correspondieron a CNeu-Ho7 y Cagro-Gh9. Un grupo importante de los ambientes se encontraron cerca del origen (VGae6, CRio6, CRio9, CUni7, CUni8, CUni9, etc.) siendo esos ambientes los que menos aportan a la interacción.

Con respecto a los cultivos evaluados, y observando la primera componente se destaca la mostaza, la cual desarrolla un mayor desempeño en el ambiente ValV9 y en menor medida en el ambiente Cneu-Ho8. A su vez, la avena desarrolla su mejor desempeño en el ambienteVAlv8 y en menor medida en CRIo6, siendo su peor desempeño en el ambiente VAlv9. Para la

segunda componente, el sorgo en el ambiente CNEu-Ho7 presentó su mayor efecto y para las habas esto se manifestó en el ambiente Cagro-Gh9. Para el resto de los cultivos (nabo, raps, maíz, chícharo y cebada) las producciones de materia seca por metro cuadrado fueron más estables.

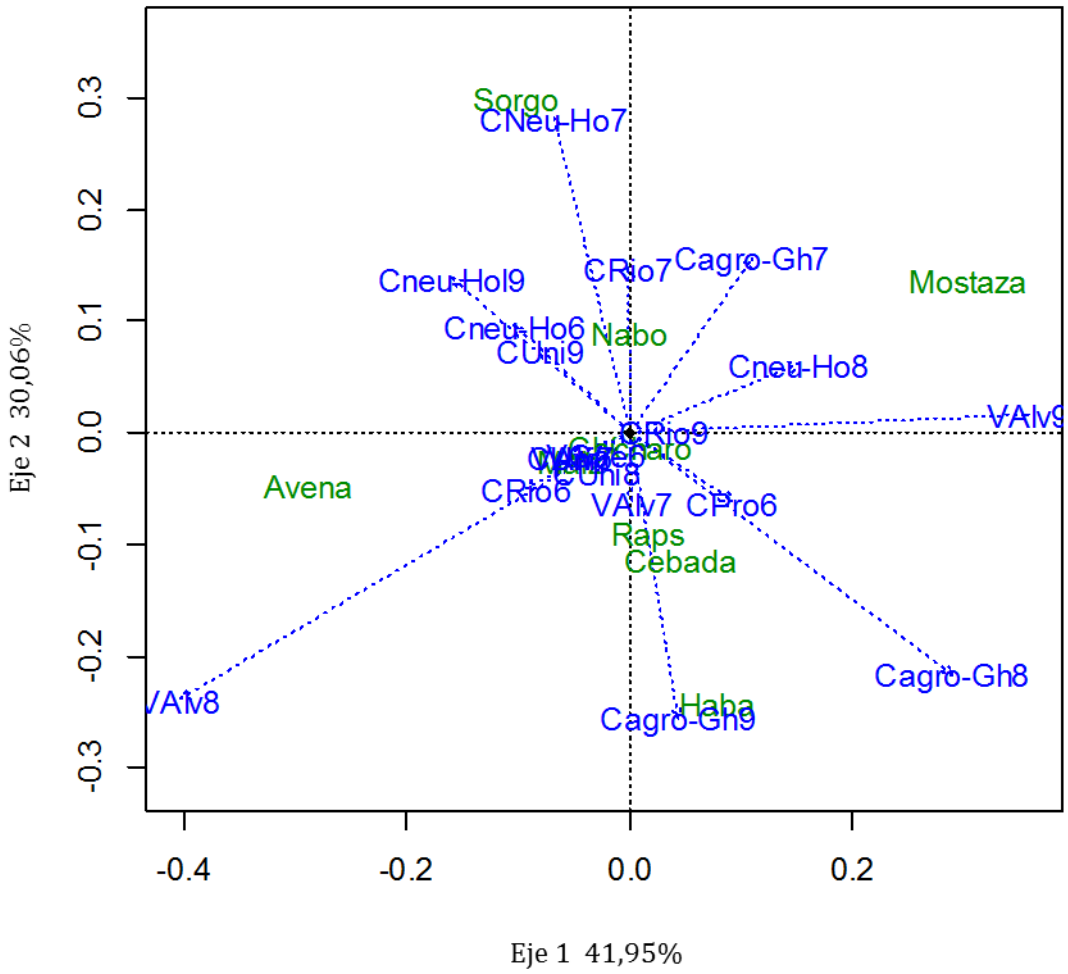


Figura 4. Biplot construido a través del modelo AMMI 2. 19 ambientes y 9 cultivos.

La Tabla 8 muestra la primera componente obtenida a partir de la combinación lineal de las variables originales en la tabla de doble entrada obtenida a partir de los BLUPs predichos (en este caso los ambientes fueron las variables utilizadas para construir las componentes). En la medida que los cultivos estén más cerca de valores en el eje de la primera componente principal cercanos a cero, más estable son. En función a lo anterior, se construyó un ranking en función de los cultivos más estables, evaluando la magnitud de su correspondiente valor de componente, siendo el chícharo, nabo, raps y la cebada los que presentaron los niveles más

estables en los ambientes evaluados. Ratificando lo visto en el modelo AMMI(2) observado en el biplot, la mostaza, avena, sorgo y haba fueron las que presentaron los mayores aportes a la interacción.

Tabla 8. Estabilidad de los diferentes cultivos evaluados.

Cultivo	kg/mt^2	CP1	Diferencia	Ranking
Avena	0,481	-0,454	0,454	8
Cebada	0,639	0,068	0,068	4
Chícharo	0,369	0,000	0,000	1
Haba	0,294	0,118	0,118	6
Maíz	1,834	-0,082	0,082	5
Mostaza	0,689	0,477	0,477	9
Nabo	0,823	0,002	0,002	2
Raps	0,438	0,023	0,023	3
Sorgo	1,922	-0,153	0,153	7

4.1.2 Caso 2

4.1.2.1 Modelos SREG-GGE Biplot

El modelo propuesto para los datos consideró como variable respuesta el rendimiento de materia seca por metro cuadrado de superficie y como criterios de clasificación los ambientes, los bloques y los cultivos, A su vez las campañas en los cuales se realizaron los experimentos se utilizaron como repeticiones. Los ambientes y los bloques (anidados a los ambientes) también fueron considerados en la modelación. Este análisis consideró un modelo de efectos fijos y aleatorios, en la cual la parte aleatoria correspondió a campaña ambiente y bloque en estructura de anidamiento y los cultivos como parte fija. Una vez estimados los respectivos BLUP e imputados los datos ausentes mediante Random Forest, se corrigieron sumando la diferencia, en cada una de las celdas de la media del cultivo estimada según el modelo menos la media general.

Dentro de las primeras estrategias de modelación, se consideró un modelo homoscedástico con distribución normal de los errores, verificándose los supuestos de homogeneidad de varianza y normalidad a través de gráficos de residuos versus predichos y gráficos qq-plot

normal. El modelo no cumplió con los supuestos (principalmente el de normalidad de los errores) por lo que se propuso como alternativa modelar los datos a través de una transformación log_{10} observándose, de manera similar a los modelos AMMI un muy buen ajuste sobre los datos modelados posterior a esta transformación (se corrige la normalidad de los errores).

De manera similar a lo visto en los modelos AMMI, se procedió a considerar diferentes propuestas de modelos (1 homoscedástico y 4 heteroscedásticos) dado que se sospechaba que era posible además ver síntomas de heterogeneidad de varianza en gráficos de residuos versus predichos. La construcción de estos modelos (Tabla 9) fue mediante la utilización de funciones de varianzas, tal como sugiere Pinheiro y Bates (2000).

Tabla 9. Comparación de modelos propuestos.

Modelos	AIC	BIC
1) Homocedástico	109,72	167,71
2) Heterocedástico Exponencial	110,34	167,72
3) Heterocedástico Potencia	111,48	168,83
4) Heterocedástico Ambiente	105,36	166,53
5) Heterocedástico Cultivo	91,58	148,92

AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion. Valores menores implica mejor.

Obtenidos los modelos, se consideraron los dos que presentaban los valores más bajos de AIC y BIC (modelo 4 y 5) procediendo luego a realizar una prueba de cociente verosimilitud para ver si existían diferencias significativas entre ellos, las cuales fueron observadas (Tabla 10). Finalmente, la sección del modelo a utilizar correspondió al modelo heteroscedástico Cultivo (Modelo 5).

Tabla 10. Test de cociente de verosimilitud entre el modelo 1 (Homoscedástico) y el modelo 5 (Heteroscedástico Cultivo).

Modelo	gl	AIC	BIC	logLik	Test	LR	p-value
1)	16	105,36	166,53	-36,68			
5)	15	91,58	148,92	-30,79	1 vs 2	11,78	0,0006

gl: grados de libertad. AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion. logLik: logaritmo de la verosimilitud del modelo; LR; Cuociente de verosimilitud.

Una vez definido el modelo, se estimaron los valores ausentes mediante Random Forest, tal como se hizo la imputación en los modelos AMMI. Una vez obtenida la matriz, se procedió a la obtención del primero de los gráficos GGE, el cual expresará su respuesta en función del genotipo más el genotipo x ambiente. Según Yan y Tinker (2006) los GGE biplots consisten en una serie de interpretaciones de gráficos, en los cuales pueden ser visualizadas evaluaciones de genotipos y pruebas de evaluaciones de ambientes. Junto con lo anterior es posible determinar mega-ambientes, evaluación y comparación de genotipos, evaluación de ambientes, pruebas de asociación y pruebas de perfiles (Yan et al., 2007). Los GGE construidos fueron representados mediante gráficos de *escores* de los genotipos y ambientes del primer componente, junto con los escores del segundo componente (Yan et al., 2007).

La Figura 5 muestra la representación de un GGE biplot (CMP-Biplot) con los datos de los diversos cultivos evaluados en los 7 ambientes definidos. En él están representados un 97,6% de la variabilidad total de los datos (CP1 = 91,6% y CP2 = 6%). Este gráfico permite estudiar de forma directa el rendimiento del genotipo dado en un determinado ambiente. Frutos (2011) considera las siguientes reglas de interpretación en este tipo de gráficos:

1) El rendimiento de un genotipo en un ambiente es mayor que la media de los genotipos si el ángulo entre el vector genotipo y el vector ambiente es menor de 90° y es peor que la media si el ángulo es mayor a 90°. Lo anterior implica que el ángulo generado en este gráfico determinará la dirección de la interacción.
2) La magnitud de la interacción se determina mediante el coseno del ángulo y la longitud del vector (dada la propiedad del producto interno del biplot). Esto nos permite utilizar esa información para ordenar los cultivos en un ambiente (en función de su rendimiento) u ordenar los ambientes en función de los rendimientos que obtenga en este caso un cultivo.
3) El origen del biplot corresponde a la media para cada ambiente, y es igual a cero para la media de los genotipos. Lo anterior implica que los cultivos que se encuentra más alejados del origen presentarán los mayores desvíos con respecto a la media del ambiente.

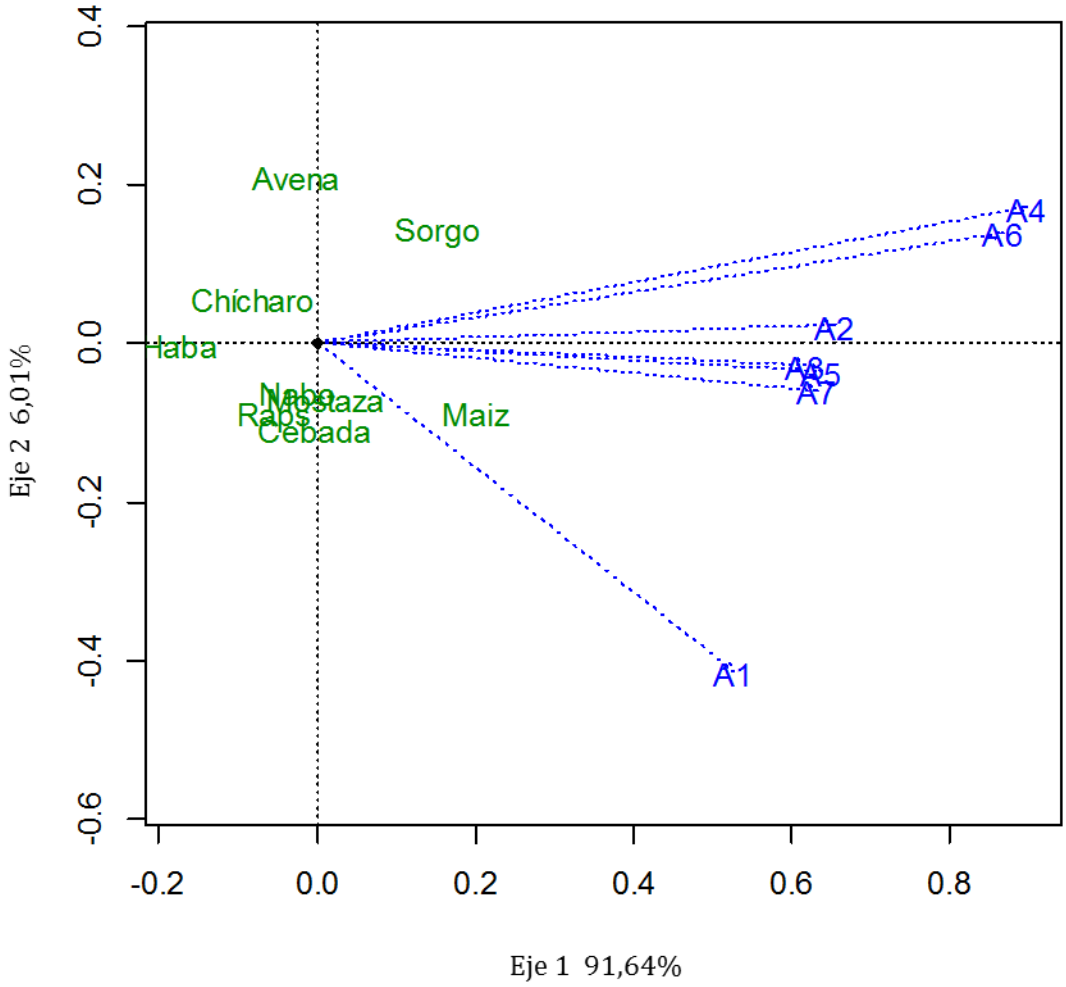


Figura 5. Biplot construido a través del modelo SREG (GGE).

Observando la Figura 5 los cultivos que permanecen en valores más cercanos al origen correspondieron a chícharo, nabo, raps, cebada y mostaza, los cuales coincidieron con las apreciaciones visuales observadas en los ensayos evaluados. La adaptación de la mostaza a los climas semidesérticos fue bastante buena, a pesar de no generar materia seca suficiente para competir de mejor manera con los de mayor rendimiento. Es muy probable que la condición semi-desértica en la cual se establecieron el resto de los cultivos no permitiera su pleno desarrollo, presentando este tipo de resultados. Los dos cultivos que destacan y se alejan del origen, presentando una dirección que coincide con los ambientes evaluados correspondieron al sorgo y maíz. Ambos cultivos fueron los que presentaron rendimientos en materia seca claramente superiores al resto, principalmente el maíz, el cual se adaptó bastante al clima presente y a las condiciones sobre las cuales fue establecido, estando sobre la media en todos

los ambientes evaluados. Con respecto al haba, esta se encuentra por debajo de la media en todos los ambientes, dado su ángulo obtuso con respecto a los ambientes. Lo anterior significó que el ángulo presente entre los genotipos y los ambientes determinó la dirección de la interacción.

4.1.2.2 Which –Won-Where

Otro de los GGE biplot de interés corresponde al biplot denominado "Which-Won-Were". Yan y Thinker (2006) mencionan que esta propuesta consiste en construir, una vez obtenido el biplot, un polígono irregular producto de la unión de los genotipos más extremos. Una vez construido este polígono, una serie de líneas desde el origen serán dibujadas de tal manera que intersecten de forma perpendicular a los lados del polígono construido previamente. Lo anterior generará una división en sectores del polígono, cada uno de ellos teniendo un cultivar ganador en los ambientes que queden agrupados dentro los límites de la división del polígono, el cual se ubicará en el vértice de éste. Estos cultivos tienen los vectores más largos, lo cual mide directamente su capacidad de respuesta en el ambiente determinado.

Un primer escenario teórico podría ocurrir si todos los ambientes caen dentro de un sector del polígono. Esto reflejaría que un único genotipo (el ubicado en el vértice) tiene el mayor rendimiento en todos los ambientes determinados. Un segundo escenario considera que no necesariamente todos los ambientes caen en el mismo sector del polígono. Si éste fuese el caso, el genotipo ubicado en cada sector, con los respectivos ambientes dentro de ese sector, sería el genotipo con mayores rendimientos. Lo anterior no solo permite identificar que cultivo es mejor en un ambiente determinado, sino que es capaz de dividir todos los ambientes en "grupos de ambientes". Estos grupos de ambientes son conocidos como "mega-ambientes".

Una definición de mega-ambiente entregada por CIMMYT[3] corresponde a "amplias áreas, no necesariamente contiguas, usualmente internacionales y frecuentemente transcontinental, definido por similares condiciones bióticas y abióticas, de similares requerimientos de sistemas de cultivo y preferencias de consumo". Según Gauch y Zobel (1997) esta definición engloba tanto aspectos medioambientales, genotípicos, geográficos e incluso aspectos

[3] International Maize and Wheat Improvement Center. Web page official.

económicos, por lo que es necesario plantear 4 criterios para la correcta identificación de mega-ambientes:

1) Enfoque selectivo en la variación del rendimiento, que es relevante para identificar mega-ambientes.
2) Relevancia en los agrónomos y mejoradores, enfocada principalmente en responder la pregunta "Quién gana y dónde" (Which-Won-Where).
3) Análisis dual de los genotipos y ambientes.
4) Flexibilidad en el manejo de datos.

Lo descrito anteriormente permite enfocar el análisis en la variación relevante de los datos e ignorar la variabilidad que es irrelevante, con el fin de que el mejorador considere las pruebas de variedades a nivel regional o desde el agrónomo, cuyo objetivo principal es ofrecer recomendaciones de genotipos para ambientes determinados (Gauch y Zobel, 1997).

En la construcción e interpretación a través de un gráfico GGE biplot de mega-ambientes, según lo definido por Yan y Tinker (2006), es posible considerar, para la Figura 6, que existe un único mega-ambiente, donde el cultivo vértice correspondió al maíz, lo cual nos sitúa en el primer escenario teórico descrito previamente, donde al caer todos los ambientes dentro de un sector del polígono, el cultivo vértice, en este caso el maíz, será el de mejor desempeño en función de su rendimiento. El ambiente A4 y A6 se encuentran en el límite del polígono de cultivos, pero a pesar de lo anterior no son un mega ambiente para el sorgo.

Es posible que al trabajar con especies muy diferentes en producción de materia seca y no con líneas de alguna especie en particular, como se hace habitualmente, haya quedado de manifiesto la gran diferencia que existe entre los rendimientos en función de la materia seca entre el maíz (planta C4) y el resto de las especies evaluadas (plantas C3, a excepción del sorgo la cual es C4) lo cual se manifiesta en que todos los ambientes evaluados se consideren un mega ambiente, con el cultivo de maíz como cultivo vértice y a su vez con un alto porcentaje de variabilidad debida al fuerte componente genotípico que esta especie demuestra.

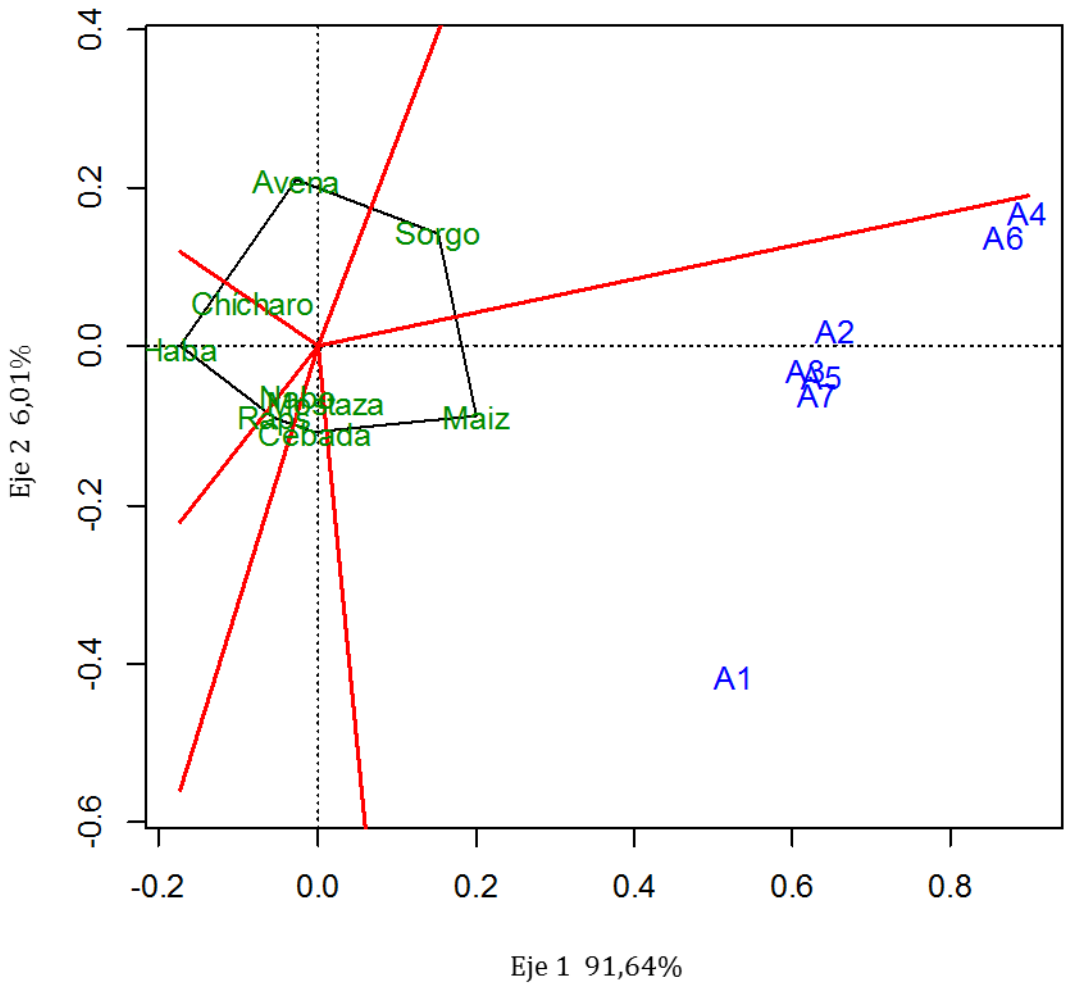


Figura 6. Biplot construido a través del modelo SREG (GGE). Which-Won-Where.

A modo de prueba, la Figura 7 y la Figura 8 muestran gráficos GGE Which-Won-Were, donde se eliminó, en el primer caso el cultivo de maíz y en el segundo caso el cultivo de maíz y sorgo de manera simultánea, procediéndose en ambos casos a la construcción de nuevos modelos sin esos cultivos y la posterior estimación de los datos ausentes. En la primera eliminación, la del maíz, es posible diferenciar 2 mega ambientes, destacándose en uno de ellos, el mega ambiente con mayor cantidad de ambientes (6 en total), con el cultivo vértice el sorgo (segundo cultivo con valores medios más altos) y para el modelo en el cual se elimina el maíz junto al sorgo (Figura 12), el cultivo vértice del mega ambiente que agrupa más ambientes correspondió a la mostaza. En ambos casos, el porcentaje de variabilidad del primer componente cae (componente genotípico), lo cual se puede explicar por la eliminación de los cultivos con altos rendimientos, dando paso a un aumento en los porcentajes de variabilidad

del segundo componente (componente GxA). Esos cambios ocurren de manera simultánea, lo cual podría ser un indicativo de la gran importancia que tuvieron los cultivos altamente productivos de materia seca, tales como el sorgo y maíz. El resto de los cultivos con bajos rendimientos de materia seca no modificaron en general su comportamiento, tal como se observa en las Figuras 7 y 8.

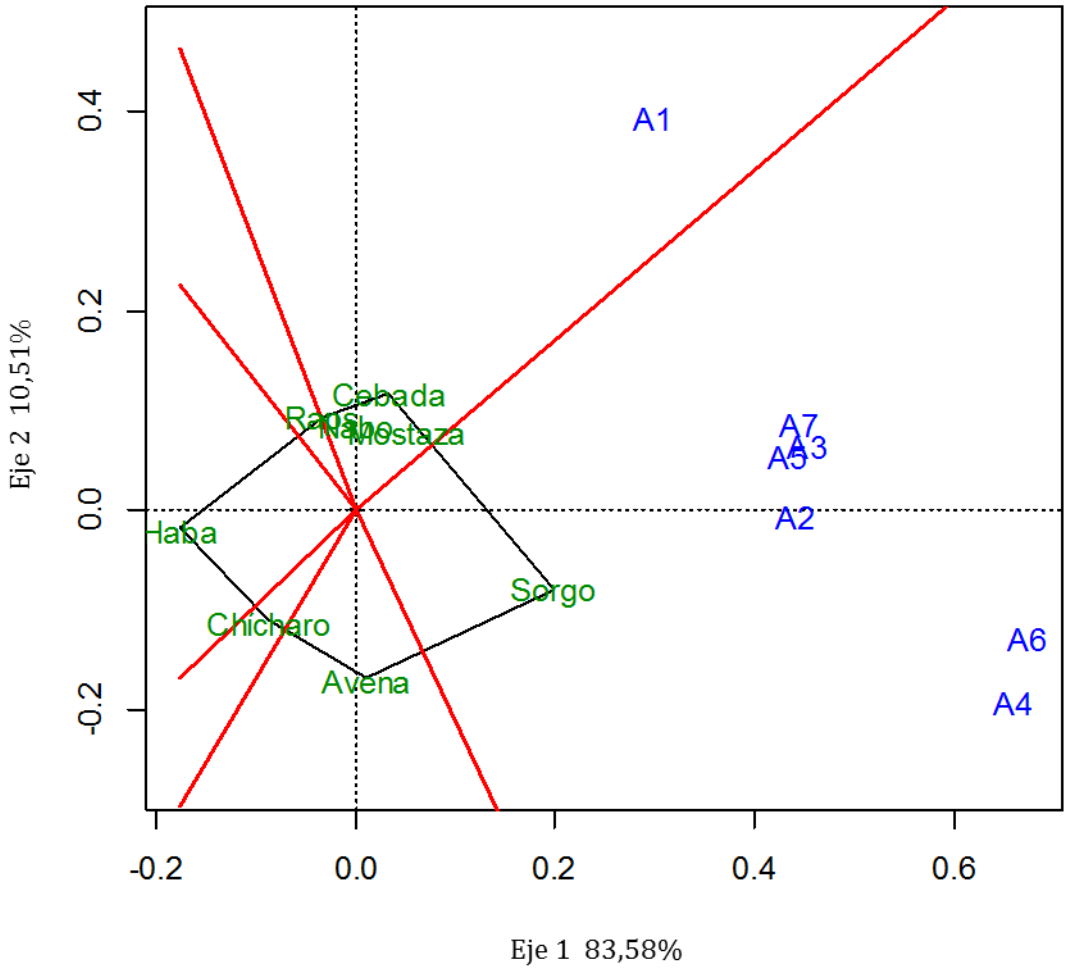


Figura 7. Biplot construido a través del modelo SREG (GGE). Which-Won-Where. Sin maíz.

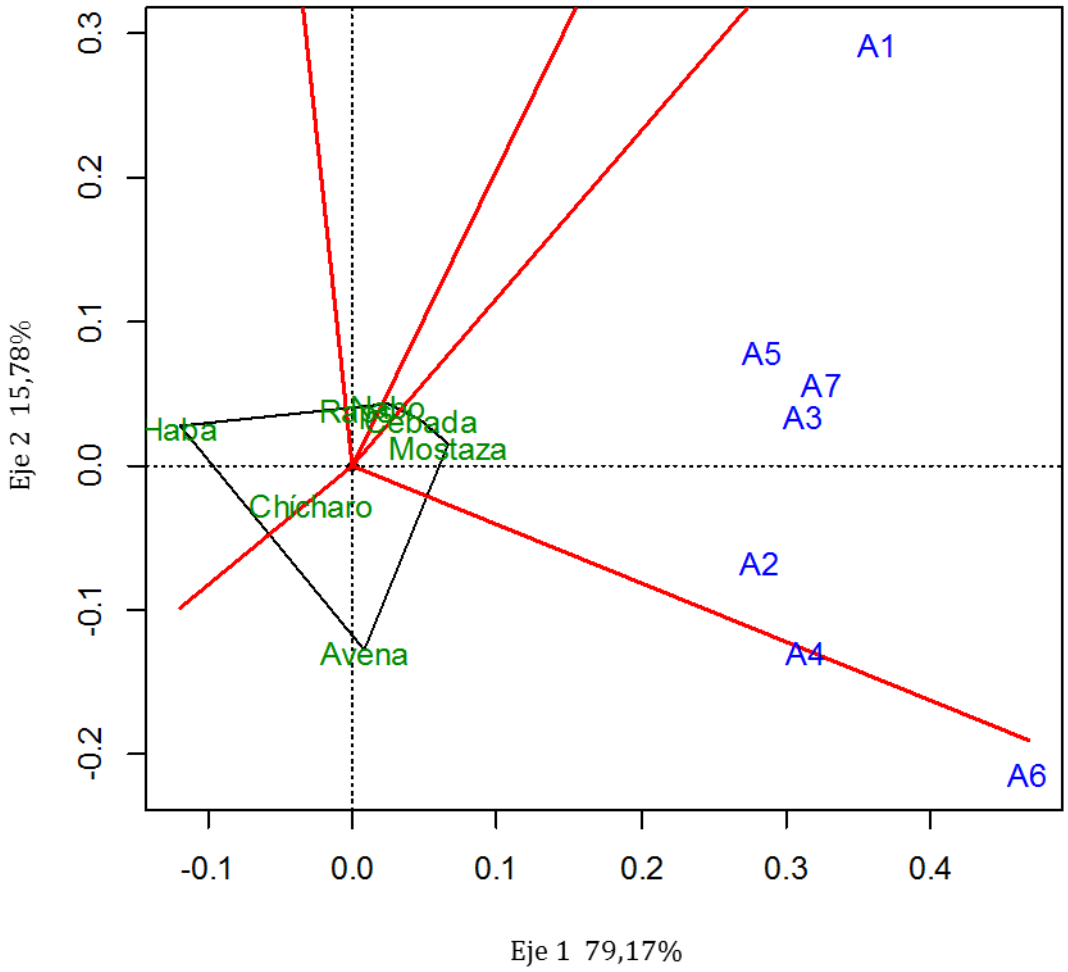


Figura 8. Biplot construido a través del modelo SREG (GGE). Which-Won-Where. Sin maíz y sin sorgo.

4.1.2.3 Estabilidad de los cultivos. Rendimiento medio

Para estudiar la estabilidad genotípica y el rendimiento medio, se realizó un AEC (Average environment coordinate biplot), utilizando en su construcción un biplot del tipo JK, con el fin de preservar la métrica de las filas (Figura 9). El círculo presente en el gráfico representa el ambiente "promedio" obtenido a partir de la media de las coordenadas del ambiente. Las rectas que pasan por el origen del biplot y sobre el ambiente promedio se denominan "ejes de abscisa del ambiente promedio" y todas las proyecciones de los cultivos sobre esos ejes aproximarán al rendimiento medio (Frutos, 2011) (Este gráfico tiene sentido cuando hay un único mega ambiente). La abscisa representa en este caso al efecto G, y la ordenada al efecto GxA asociado al rendimiento. Lo anterior, y observando la Figura 9, significa que las

proyecciones sobre la abscisa, con mayores rendimientos medios sobre los mega-ambientes correspondieron al sorgo y maíz, siendo este último el cultivo vértice del único mega-ambiente definido. Los cultivos que tuvieron mayor proyección sobre el eje de la ordenada, fueron el sorgo, y la avena, los cuales fueron los más inestables. A su vez, el maíz y chícharo fueron los cultivos más estables. Lo anterior implica que el maíz, es un cultivo superior al rendimiento medio, y a su vez estable. El chícharo por su parte fue un cultivo estable, pero de bajo rendimiento.

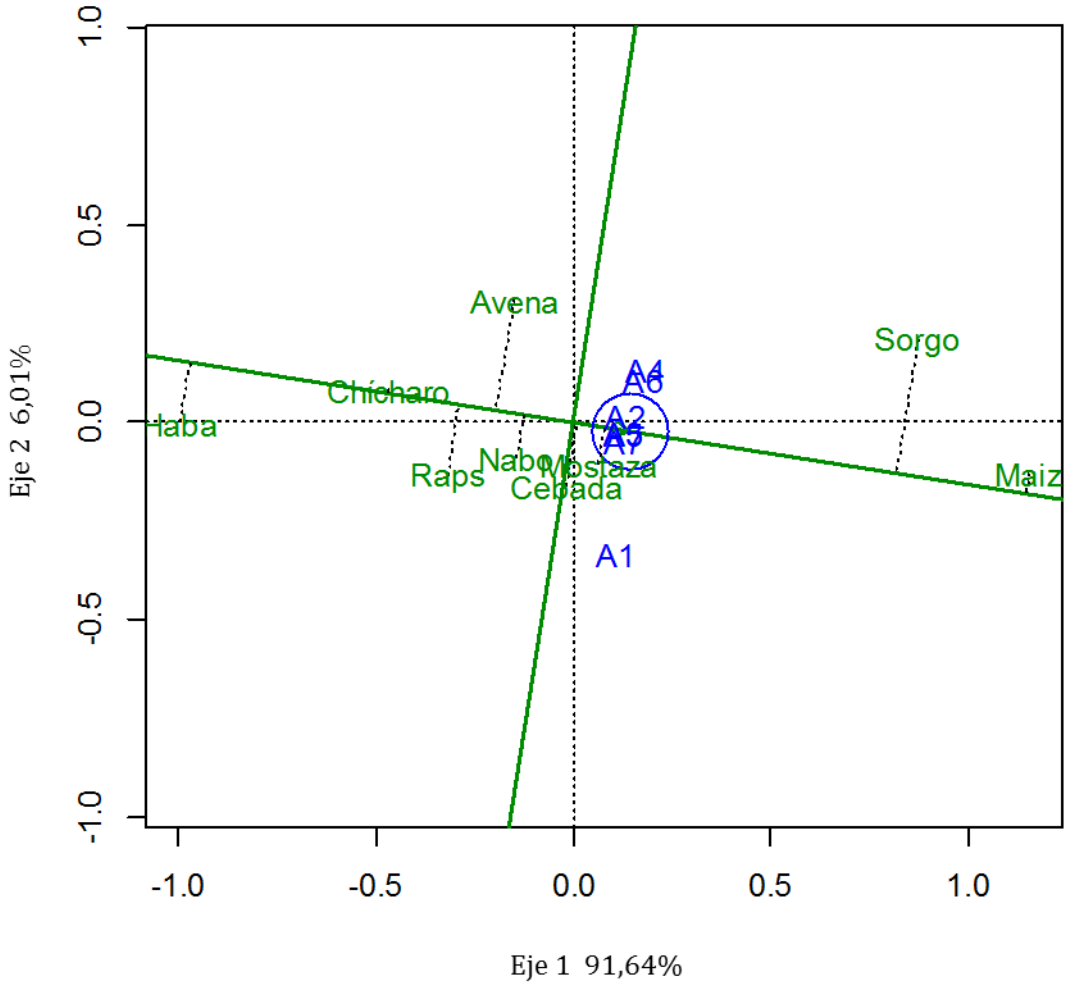


Figura 9. Biplot construido a través del modelo SREG (GGE). Rendimiento medio y ranking de los genotipos.

4.1.2.4 Cultivos evaluados versus el "cultivo ideal"

Al momento de hacer una selección en un programa de mejoramiento, o una sugerencia de cultivo que un agrónomo pretende realizar, no es suficiente que el genotipo recomendado

tenga un rendimiento superior a la media de los genotipos evaluado, sino que también este genotipo sea estable en los ambientes en los cuales fue evaluado. Los círculos concéntricos observados en la Figura 10 muestran la distancia que tienen los cultivos evaluados con respecto a un cultivo ideal teórico (mayor rendimiento y mayor estabilidad) donde la selección o recomendación se debería acercar. Lo anterior sugiere que el cultivo de maíz es el que más se acerca a la condición optima al momento de realizar una selección, seguido del sorgo y un grupo de cultivos como avena, chícharo, nabo, mostaza, cebada, y raps. El más lejano correspondería al cultivo de haba, siendo el menos recomendable, a pesar de su alta estabilidad (estable con bajo rendimiento de materia seca, muy alejado del anillo del cultivo "teórico ideal").

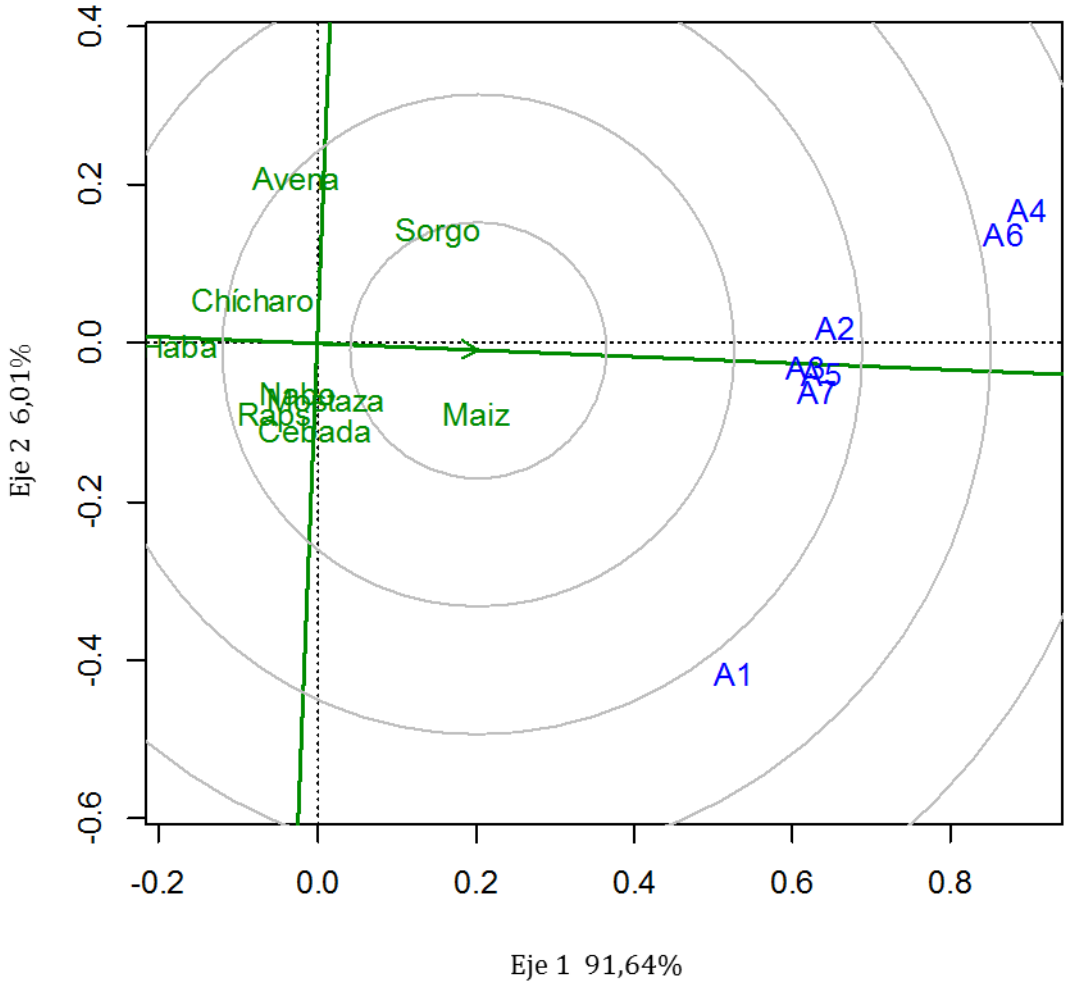


Figura 10. Biplot construido a través del modelo SREG (GGE). Identificación del cultivo ideal y su relación con el resto de los cultivos.

Una relación interesante es la que se observa tanto en las Figuras 9 y 10 en las cuales el maíz es el cultivo más estable y de mayor rendimiento y a su vez es el genotipo más cercano al genotipo teóricamente ideal. Esto ocurre principalmente ya que la magnitud de G (genotipo) es mucho mayor que GxA. Lo anterior es posible de determinar al observar el porcentaje de variabilidad que resume la primera componente (91,6%), la cual engloba principalmente el efecto Genotipo, versus la segunda componente (6%) que es la que abarca la interacción Genotipo Ambiente. Una posible razón a tales resultados es la gran diferencia de rendimiento de materia seca que posee el maíz versus el resto de los cultivos evaluados, lo cual se traduce en arrastrar toda la variabilidad al primer componente. Esto se ve reflejado en la escasa interacción GxA (6% de la variabilidad) en la medida que aparece un cultivo de alto rendimiento que desafía al resto de los cultivos.

4.1.2.5 Relación entre ambientes

La Figura 11 muestra la proyección que tienen los ambientes con el origen del biplot, construyendo lo que se denominan "vectores ambientes". En la interpretación de un biplot de componentes principales, el coseno generado entre los vectores ambientes es una buena aproximación a la correlación que puede ocurrir entre los ambientes, en la medida que el porcentaje de representación del biplot construido sea idealmente lo más cercano a uno (Peña, 2002). Lo anterior implica que ángulos agudos entre vectores ambientes son ambientes correlacionados de manera positiva, ángulos obtusos significa que los ambientes se correlacionan de forma negativa y ángulos cercanos a valores de 90° indican ausencia de correlación. A su vez, la distancia observada entre ambientes puede ser considerada una medida de disimilitud entre ellos. El alto porcentaje observado en estos biplot asegura una buena interpretación de las correlaciones existentes.

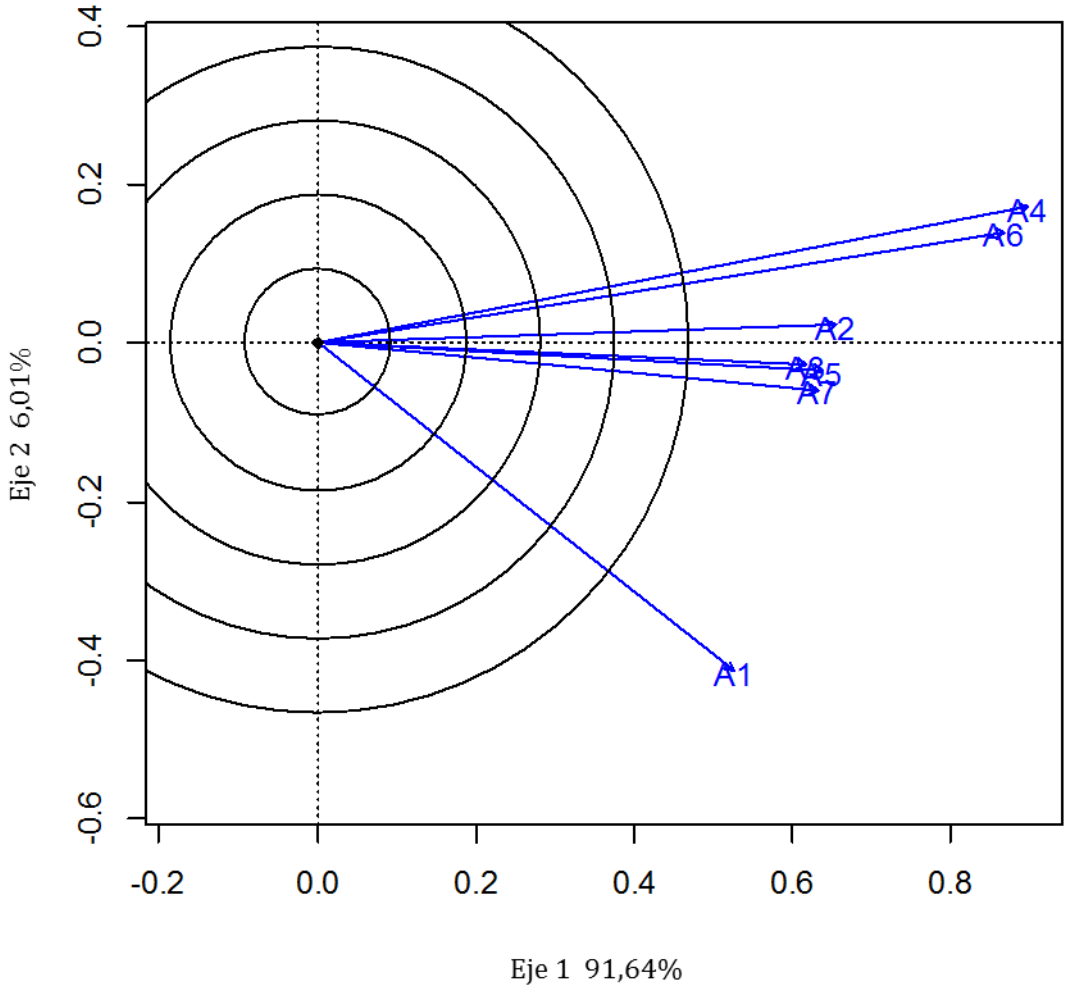


Figura 11. Biplot construido a través del modelo SREG (GGE). Relación entre los ambientes evaluados.

Según lo anterior es posible observar que:

1) Los ambientes, desde ahora pertenecientes al grupo 1, A4 y A6 se encuentran fuertemente correlacionados de manera positiva.
2) Los ambientes, desde ahora pertenecientes al grupo 2, A2, A3 A5, A6 y A7 se encuentran fuertemente correlacionados de manera positiva.
3) Los grupos de ambientes 1 y 2 están correlacionados de manera positiva, no tan fuerte como los ambientes dentro de cada grupo.
4) Debe existir una baja correlación entre los ambientes pertenecientes al grupo 1
5) Existe una baja correlación entre los ambientes A3-A4 y los ambientes A1-A5-A6 y A2-A7.

6) Las distancias generadas entre los grupos propuestos hacen suponer la existencia de 3 grupos, el primero corresponde a los ambientes A4 y A6, el segundo a los ambientes A2, A3 A5, A6 y A7 y el tercero al ambiente A1.
7) La correlación que pueda existir entre ambientes permitirá, si fuese necesario seleccionar menos ambientes de prueba para estos cultivos, reduciendo futuras pruebas de selección (uno por cada grupo).

El gráfico anterior es complementario a lo observado en los gráficos correspondientes al "Which-Won-Where" donde se definen los mega ambientes utilizados en este estudio, donde todos los ambientes están en una única dirección, dada la inercia con la cual los arrastra el maíz.

4.1.2.6 Ambiente medio

La Figura 12 representa un biplot GGE en la cual la línea azul observada atraviesa el origen del biplot corresponde a la proyección de un eje ATC. Este eje (Average tester coordinate) pasa por el centro del biplot y por el "promedio de los ambientes", el cual se define como el promedio de los escores de las componentes principales (CP1 y CP2) sobre todos los ambientes, siendo este el ATC-x (Yan 2001) El ATC-y corresponde a un eje que también pasa por el centro del biplot, pero es perpendicular al eje ATC-x. El centro de los círculos concéntricos en el eje ATC es donde debería estar el teórico ambiente ideal. La interpretación de estos datos sugiere que mientras más cerca estén los ambientes a este ambiente teórico mejor serán como ambientes de prueba. Lo anterior sugiere que los ambientes A4 y A6 son los mejores ambientes, seguido de los ambientes A2 A3, A5, A6 y A7 y finalmente el ambiente A1. Es posible que este resultado sea más robusto que el que se proporciona habitualmente, dado que se cuentan con repeticiones de los ensayos en distintas campañas, que es lo que se recomienda tener a la hora de interpretar este tipo de gráficos.

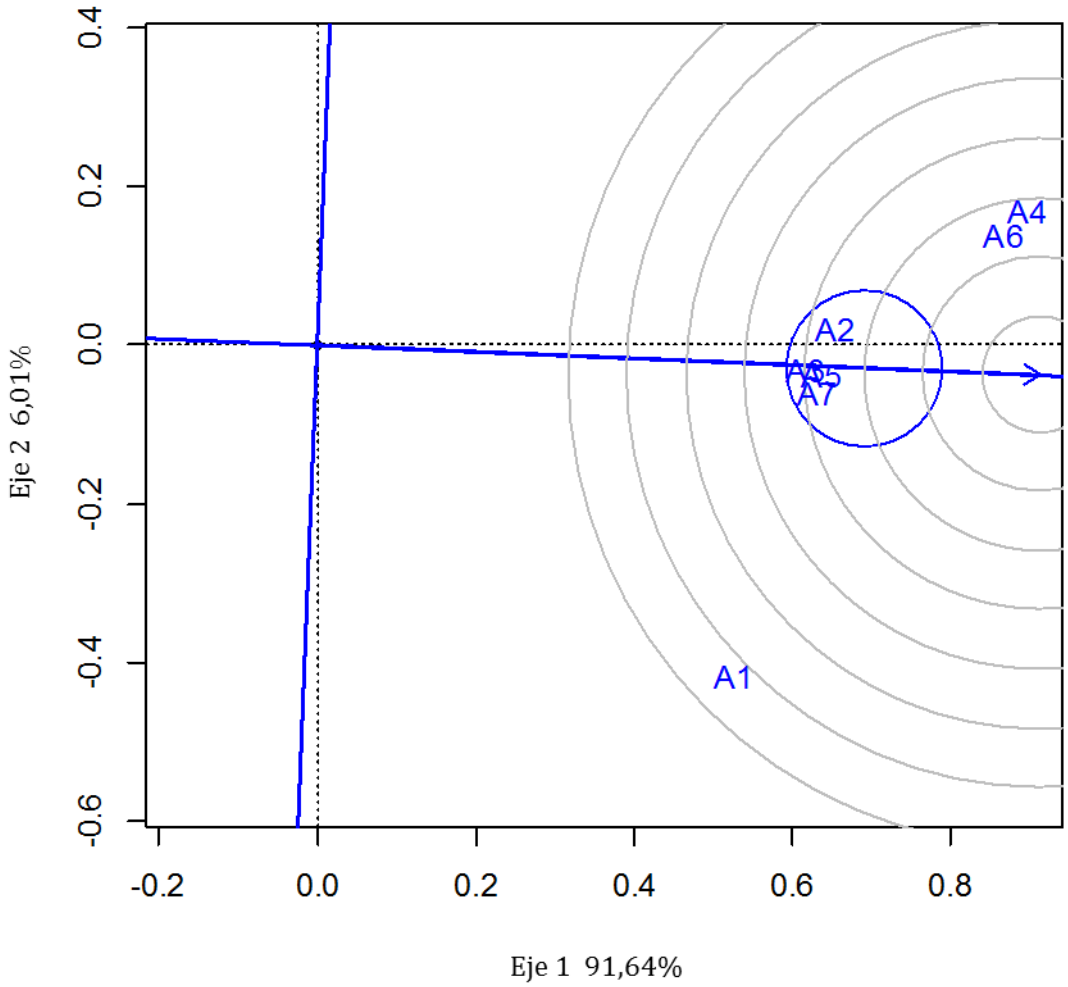


Figura 12. Biplot construido a través del modelo SREG (GGE). Ranking de ambientes en base al poder discriminativo de la media.

4.2 Discusión

4.2.1 Limitaciones consideradas en el uso de modelos AMMI y GGE

Yang et al. (2009) plantean algunas interrogantes que deben ser consideradas al momento de utilizar modelos lineales-bilineales. La primera de ellas referente al número de componentes principales al momento de presentar los gráficos biplot. De manera habitual, se presentan tanto modelos AMMI2 como GGE2, dada su facilidad de interpretación (a través de 2 dimensiones). Estos modelos pueden ser respaldados por el porcentaje de variabilidad total que puede ser representada por esos dos componentes. Idealmente si la suma de los primeros 2 componentes (tanto para AMMI2 como para GGE2), sobre la suma total representa un

porcentaje alto, idealmente superior al 70% es probable que representen de forma óptima la variabilidad total. Otra opción es la planteada a través del modelo FANOVA (Gollob, 1968) teniendo alternativas de ser modelados en caso de ser análisis obtenidos de diseños completamente aleatorizados o bajo diseños en bloques completamente aleatorizados, siendo estos analizados bajo el enfoque de análisis de la varianza. En caso de requerir un mayor número de componentes principales, es posible considerar lo sugerido por Yan y Tinker (2006) quienes proponen la utilización de un 3D biplot, el cual será suficiente para mostrar los principales patrones de los datos. La dificultad de presentar un 3D biplot radica en que es necesario utilizar algún software gráfico con el fin de ir rotando los ejes y observar los patrones, además que un 2D biplot será siempre más informativo para los dos primeros CP que un biplot estático en 3D.

Los resultados observados tanto en los biplot AMMI como en los modelos GGE presentados en este trabajo superaron el 70% de la variabilidad resumida en los gráficos biplot 2D presentados, siendo de un 72% para los modelos AMMI y de un 96% para los modelos GGE, lo cual hace que estos gráficos sean bastante informativos, no siendo necesario en ambos casos explorar más dimensiones. Llama la atención el alto porcentaje de variabilidad resumida en las distintas representaciones, principalmente para los gráficos GGE, en los cuales el porcentaje de variabilidad, únicamente en la primera componente, correspondió al 91,6%, lo que indica que el peso de los análisis los tiene principalmente la parte genotípica (representada por la primera componente) y no la parte que agrupa la parte genotípica más la interacción genotipo x ambiente (representada por la segunda componente). La sobresaliente capacidad de producir materia seca del maíz en los ambientes en los cuales fue evaluado es la causa más probable de estos resultados, manifestándose en ser considerado el genotipo más estable, de mayor rendimiento y a su vez, por su inercia arrastrar a todos los ambientes evaluados a ser considerados como un mega ambiente en el cual, el maíz es el cultivo dominante.

4.2.2 Presencia de datos ausentes

Las alternativas que existen, al momento de enfrentarse a datos ausentes, ya sea por acciones planificadas o no, se podrían resumir en 3: 1) Eliminar las combinaciones de GxA no evaluados, 2) proceder a reemplazar los datos ausentes mediante la media ambiental o 3) utilizar herramientas estadísticas de imputación de datos que utilicen la información presente

en la base de datos (Yan, 2013; Woyann et al., 2017). Junto a lo anterior vale la pena recordar que la presentación del biplot como resultado de la modelación de la parte bilinear es exclusivamente de manera descriptiva (Yang et al., 2009), siendo principalmente de interés la representación visual de los patrones de comportamiento de los genotipos en ambientes determinados. Lo anterior resulta interesante al momento de modelar a través de biplot matrices de doble entrada que tengan datos ausentes, los cuales, si es posible imputar no deben necesariamente reflejar una imputación del dato con baja variabilidad, pero idealmente que esta imputación sea lo más insesgada posible. Lo anterior se puede deducir por lo presentado por Yan (2013), quien realizando imputaciones de datos faltantes a matrices de doble entrada, obtenidas de ensayos GxA observó que los patrones en biplot GGE cuando se simulaban datos ausentes en hasta un 40% para matrices pequeñas (18×9) y hasta un 60% en matrices grandes (15×40) eran perfectamente reproducibles, observando los gráficos, al realizar imputaciones de datos mediante descomposición por valor singular (SVD), lo cual permite evaluar de forma razonable genotipos no evaluados en determinados ambientes y reproducir estos patrones en biplot confiables. Esto fue tomado en consideración durante esta investigación, dado que no todos los cultivos fueron evaluados en todos los ambientes, por lo que fue necesario imputar esos datos, con el fin de no eliminar las combinaciones genotipos ambientes inexistentes. Con respecto a la utilización de la metodología Random Forest para realizar dicha imputación, la literatura menciona que los datos imputados, a través de esta metodología, son capaces de asimilar comportamientos no lineales e interacciones complejas. En el caso de este estudio, la utilización de diferentes especies, lo cual involucra una susceptibilidad única a condiciones fitosanitarias particulares en cada especie, la respuesta de los cultivos al manejo agronómico propio de las vides, más la interacción propia de una especie a un ambiente determinado, hacen extremadamente difícil considerar que los datos imputados sean capaces de resumir toda la información antes mencionada. Es probable que este tipo de imputaciones sea más robusta al momento de realizar imputaciones de datos faltantes en la medida que sea una única especie (como se hace habitualmente en los planes de mejoramiento) en vez de múltiples especies (como lo realizado en este estudio).

Según Woyann et al. (2017), diferentes porcentajes de valores ausentes podrían tener diferentes impactos en el porcentaje de variabilidad resumida por los componentes. En su estudio encontraron que para una base de datos con 0% de datos ausentes, el porcentaje de variabilidad resumido por 2 CPs en un GGE correspondió a un 42,3%. Al eliminar

aleatoriamente el 30% de los datos en la matriz y estimar los datos ausentes, el porcentaje de variabilidad aumenta a valores entre 58 y 66% y para la estimación de un 60% la cantidad varía entre 83,6 y 86,4%. En el caso de modelos AMMI, el porcentaje de variabilidad, para ausencia del 60%, correspondió a valores que oscilaron entre 81,1 y 92,6%. Lo importante a destacar en este estudio, es que a partir de un 60% de imputación de datos se pierden los patrones en gráficos GGE Which-Won-Where, no replicándose los mega ambientes ni los cultivos vértices, indicativo de cultivares ganadores, lo cual debe tomarse como precaución a la hora de interpretar esos gráficos. Con respecto al aumento en los porcentajes de variabilidad de los componentes en la medida que aumentan los datos ausentes, ellos consideran que estos cambios son producto de la reducción de la complejidad de la interacción GxA, lo cual repercute directamente en la performance que pueda observarse tanto en los genotipos como ambientes evaluados. Su recomendación, para los estudios de GGE Which-Won-Where, es imputar como máximo un 30% de datos ausentes para conservar las tendencias en los respectivos gráficos.

Con respecto a los datos analizados en este estudio, el caso 1 presenta una ausencia de un 64,5% y el caso 2 presenta una ausencia de un 49% de los datos, lo cual podría ser motivo de alerta al momento de presentar los resultados, principalmente para el caso 1. A pesar de lo anterior, y dada la gran ausencia de datos, los altos rendimientos de maíz, los cuales arrastran los porcentajes de variabilidad hacia la primera componente hace que sea bastante difícil modificar los patrones de comportamiento en la medida que se evalúen cultivos que presentan valores de rendimiento muy diferentes entre sí.

4.2.3 Uso de modelos mixtos en ensayos multiambientales

La manera tradicional en la cual se realizan los ajustes, mediante estimaciones por mínimos cuadrados (enfoque ANDEVA) resulta restrictiva en ensayos multiambientales, donde la presunción de independencia de las observaciones no satisface los supuestos del modelo lineal, en contraste a la utilización de modelos lineales mixtos, donde es posible incorporar estructuras de covarianza sobre las observaciones (Balzarini, 2002). De manera adicional, los procesos de estimación de parámetros a través de modelos lineales mixtos tienen la capacidad de evitar los problemas que se puedan generar por ausencia o desbalance de datos.

La utilización de modelos mixtos implica previamente definir factores fijos y factores aleatorios. Factores fijos en un modelo corresponderán a variables de clasificación en las cuales el investigador ha incluido todos los niveles sobre los cuales tiene interés inferir en el respectivo estudio, siendo estos niveles reproducibles y posibles de ser contrastables. Factores aleatorios corresponden a variables de clasificación, en las cuales sus niveles pueden ser pensados como seleccionados de manera aleatoria de una población de estudio (West et al. 2015). La utilización de los genotipos y los ambientes, bajo el enfoque de modelos lineales mixtos, puede ser considerado tanto como factores fijos como aleatorios. Lo más habitual es considerar que los genotipos, al inicio de un programa de mejoramiento, pueden ser una muestra aleatoria tomada desde una población, lo cual consideraría directamente a los genotipos, o en nuestro caso a los cultivos como efectos aleatorios (Piepho, 1998). A su vez, modelos que consideran efectos aleatorios, el espacio de inferencia es posible centrarlo en los componentes de varianza o cualquier índice que pueda ser construido a partir de él (Coeficientes de correlación intraclase, heredabilidad, etc). En experimentos agrícolas, efectos que son considerados habitualmente fijos, por el interés de ser evaluadas las medias de sus respectivos niveles, es posible, en la medida que la distribución de sus efectos sea razonablemente simétrica, utilizar la predicción de los BLUPs, en vez de la estimación de los BLUE (best linear unbiased estimator, para efecto dijo), lo que significa considerar el anterior efecto fijo como efecto aleatorio (Stroup and Mulitze, 1991). En los procesos avanzados de selección de genotipos, pocos elementos son seleccionados para continuar con los procesos tardíos de selección. En estas etapas tardías, es más frecuente considerar a los genotipos como efectos fijos, dado que la investigación se centrará en esos genotipos seleccionados, teniendo siempre en consideración que lo más apropiado, a pesar de definir como fijo al genotipo, es que la interacción genotipo ambiente se analice como efecto aleatorio, si este último ha sido considerado como tal (Balzarini, 2002). Ensayos en etapas tempranas de selección, al ser incorporados muchos genotipos que serán evaluados posiblemente una única vez, de manera conjunta con variedades “check”, es deseable considerar a los genotipos como efectos aleatorios (Federer, 1998). De todas las alternativas de modelamiento presente, se consideró para este trabajo analizar los cultivos como efectos fijos y los ambientes como efectos aleatorios (además de las temporadas y los bloques), dando como resultado una interacción de ambos efectos como un efecto aleatorio. Esta interacción fue la utilizada principalmente en la construcción de los modelos AMMI y GGE presentes en este estudio.

5. Consideraciones finales

En el presente trabajo se abordó el análisis de una base de datos proveniente de ensayos realizados en diferentes temporadas en diferentes localidades ubicadas en la Tercera Región de Chile. Los cultivos utilizados presentaban un desbalance importante, por lo que la propuesta de análisis consideró esta dificultad. La forma de analizar tradicionalmente ensayos en los cuales es de interés explorar la interacción entre los genotipos evaluados en diferentes ambientes es a través de una combinación de técnicas estadísticas que incluyen: análisis de la varianza, descomposición por valor singular y biplot. Al momento de la presentación de los resultados, se debe considerar que los supuestos para el análisis de la varianza, que corresponden a independencia, homogeneidad de varianza y normalidad de los errores se deben cumplir de forma simultánea.

La propuesta de análisis de este trabajo involucró la imputación de datos ausentes, el modelamiento y presentación de los resultados mediante modelos lineales mixtos, descomposición por valor singular de matrices de BLUPS y biplot, todo con el objetivo de flexibilizar los supuestos que involucran tanto los modelos AMMI como los modelos GGE, escasamente cuestionados en la presente literatura. Es destacado que la estimación REML utilizada en la estimación de componentes de varianza de los efectos aleatorios coincide con la estimación utilizada en el tradicional análisis de la varianza en la medida que existan datos balanceados (Corbel and Searle, 1976), junto con la deseable propiedad de estimaciones insesgadas de varianzas. En el caso de desbalance de datos, la estimación REML, utilizada en el presente estudio, presenta ventajas sobre las estimaciones basadas en el análisis de la varianza, ya que sus estimaciones son únicas (Balzarini, 2000) y las estimaciones de varianzas son no negativas, como es posible encontrar en algunos casos particulares de estimaciones ML (Corbel and Searle, 1976).

Junto con lo anterior, es necesario explorar el impacto que tienen la utilización de modelos lineales- bilineales en casos en los cuales no se cumplen los supuestos de homogeneidad de varianza. Alternativas que consideren la utilización de modelos lineales mixtos junto con la modelación de varianzas, pueden ser encontradas en los estudios de So y Edwards (2011), quienes mencionan que en la modelación de estructuras de varianzas y covarianzas de datos con evidente heterogeneidad, las predicciones en performance de híbridos de maíz mejoran en un 63% de la data analizada al realizar validaciones cruzadas. Modelos AMMI que

consideran la heterogeneidad han sido evaluados por Rodríguez et al. (2014) quienes desarrollaron los denominados modelos W-AMMI (Weighted-AMMI) los cuales al momento de construir las matrices de doble entrada y una vez detectada la variabilidad de la variable respuesta a través de los ambientes, pondera diferencialmente las celdas de esa matriz, obteniendo a su vez diferentes residuos al cuadrado ponderados. Eso significa que, de acuerdo a esta estrategia, la posibilidad de encontrar variabilidades diferentes entre ambientes, debido al diseño experimental utilizado o la variabilidad espacial puede ser considerada en la interacción, ajustando en primer lugar los efectos principales mediante lo que denominan mínimos cuadrados ponderados y posteriormente una descomposición por valor singular ponderada.

Es de esperar que futuros estudios incluyan alternativas de modelamiento en los casos que sea necesario considerar la heterogeneidad de varianza y/o estructuras de correlación, ya sea a través de modelos W-AMMI, W-GGE (Hadasch et al., 2018) o modelos que incluyan funciones de varianzas al momento de ajustar modelos lineales.

Con respecto a la presencia de datos ausentes, la propuesta de este estudio involucró su respectiva imputación. Diversas metodologías fueron evaluadas a partir de simulaciones en bases de datos completas, procediéndose a utilizar la más indicada de acuerdo a este estudio. La literatura presente advierte, en un caso, que sólo debe hacerse hasta un 30% de datos ausentes, siendo otro caso, realizar imputaciones de hasta un 60% de los datos ausentes, sin alterar los patrones observados en los gráficos biplot para modelos AMMI y GGE. Cabe destacar que lo anterior corresponden a estudios de ensayos multiambientales en los cuales se evalúa exclusivamente una única especie (con sus respectivas lineas) y no diferentes especies como se hizo en esta investigación. Es muy probable que ausencias mayores de datos puedan ser estimadas sin problemas, no modificando los patrones en los gráficos biplots en la medida que los genotipos sean muy diferentes entre sí, ya que al construir los gráficos biplot GGE, el porcentaje de variabilidad del primer componente no se verá mayormente alterado, principalmente por la fuerza genotípica expresada en la primera componente.

Para finalizar, y en un aspecto netamente agronómico, vale la pena destacar que la fertilización, riego y manejo de los cultivos evaluados se realizó a través de los manejos habituales del cultivo principal del agricultor (vid), por lo que el manejo de los cultivos de

cobertura no requirió de costos adicionales más que su establecimiento, por lo cual su siembra se puede perfilar como una técnica rentable a futuro para el agricultor, la cual con una correcta recomendación de un determinado cultivo, permitirá expresar todos los beneficios mediambientales descritos al inicio de este trabajo. Con respecto a la utilización de cultivos de cobertura y en función de la productividad de materia seca, el maíz es el cultivo más apropiado de ser utilizado en los ambientes acá descritos.

6. Literatura citada

Aballay, E. e Insunza, V. 2002. Evaluación de algunos cultivos de cobertura sobre el control de Xiphinema index en vid de mesa, cultivar Thompson Seedless en la zona central de Chile. Agricultura Técnica, pp. 357-365.

Azur, M., Stuart, E., Frangakis, C., and Leaf, P. 2011. Multiple Imputation by Chained Equations: What is it and how does it work? J. Methods Psychiatr. Res. 20(1): 40–49.

Baginsky, C., Seguel, O. y Contreras, A. 2010. Impacto de la utilización de cultivos y enmiendas orgánicas sobre la funcionalidad del suelo. Libro Serie Ciencias Agronómicas Nº 17. 143 p.

Balzarini, M. 2000. Biometrical models for predicting future perfomance in plant breeding. Thesis Ph. D. Lousiana State University, Baton Rouge, Luisiana, USA, 268 p.

Balzarini, M. 2002. Applications of mixed models in plant breeding. p. 353–363. In M.S. Kang (ed.) Quantitative genetics, genomics, and plant breeding. CAB Int., Wallingford, UK. 400p.

Baraldi, A. and Enders, C. 2009. An introduction to modern missing data analyses. Journal of School Psychology, 48: 5–37.

Bohn, H., McNeal, B. and Connor, G. 1993. Química del Suelo. Ciudad de México. Limusa. 370p.

Bowman, J. 1972. Genotype x environment interaction. Annales de génétique et de sélection animale 4(1): 117-123.

Callejas, R., Kania, E., Contreras, A., Peppi, C. y Morales, L. 2013. Evaluación de un método no destructivo para estimar las concentraciones de clorofila en hojas de variedades de uva de mesa. Idesia, 31(4): 19-26.

Cesco, S., Rombolà, A. D., Tagliavini, M., Varanini, Z. and Pinton, R. 2006. Phytosiderophores released by graminaceous species promote 59Fe-uptake in citrus. Plant and Soil, 287: 223–233.

Corbeil, R. and Searle, S. 1976. Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. Technometrics, 18(1): 31-38.

Cornelius, P. L., Crossa, J. and Seyedsadr, M. 1993. Tests and estimators of multiplicative models for variety trials. Annual Conference on Applied Statistics in Agriculture, Kansas State University.

Crossa, J. Burgeño, J., Cornelius, P., McLaren, G., Trethowan, R. and Krishnamachari, A. 2006. Modeling Genotype x Environment Interaction Using Additive Genetic Covariances of Relatives for Predicting Breeding Values of Wheat Genotypes. Crop Science, 46: 1722–1733.

Crossa, J. 2012. From genotype × environment interaction to gene × environment interaction. Current Genomics, 13(3): 225–44.

Federer, W. 1998. Recovery of Interblock, Intergradient, and Intervariety Information in Incomplete Block and Lattice Rectangle Designed Experiments. Biometrics, 54(2): 471-481.

Finlay, K.W. and G.N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Aust. J. Agric. Res. 14: 742-754.

Frutos, E. (2011). Interacción Genotipo Ambiente: GGE Biplot y Modelos AMMI (Tesis de Magister), Universidad de Salamanca. Disponible en: https://gredos.usal.es/jspui/bitstream/.../TFM_MAADM_Frutos_Bernal_Maria_Elisa.pdf

Frutos, E., Galindo, M. P. and Leiva, V. 2014. An interactive biplot implementation in R for modeling genotype-by-environment interaction. Stochastic Environmental Research and Risk Assessment, 28(7): 1629–1641.

Gabriel, K. R. 1971. The biplot-graphical display of matrices with applications to principal components analysis. Biometrika, 58: 453–467.

Gabriel, K. R. and Samir, S. 1979. Lower Rank approximation of matrices by least square with any choices of weight. Technometrics, 4: 489-498.

Gauch, H. 1988. Model selection and validation for yield trials with interaction. Biometrics, 44(3): 705-715.

Gauch, H. and Zobel, R. 1997. Identifying Mega-Environments and Targeting Genotypes Hugh. Crop Science, 37: 311-326.

Gauch, H. 2006. Statistical Analysis of Yield Trials by AMMI and GGE. Crop Science, 48(3): 866-899.

Gauch, H., Piepho, H. and Annichiarico, P. 2008. Statistical analysis of yield trials by AMMI and GGE: Further considerations. Crop Science, 48(3): 866-889.

Gollob, H. 1968. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika, 33(1): 73-115.

Golub, G.H., Hoffman, A. and Stewart, G.W. 1987. A Generalization of the Eckart-Young-Mirsky Matrix Approximation Theorem. Linear algebra and its applications, 89: 317-327.

González, E., Antonio, H. and Antonio, R. 2005. Las cubiertas vegetales en el olivar: Un beneficio para el olivicultor y para el medioambiente. Revista de la asociación Española de Agricultura de Conservación/Suelos Vivos, p. 8.

Hadasch, S., Forkman, J., Malik, W. and Piepho, H. 2018. Weighted Estimation of AMMI and GGE Models. Journal of Agricultural, Biological and Environmental Statistics, 23(2): 225-275.

Henderson, C. 1984. Application of linear models in animal breeding. Tercera ed. Ontario. University of Guelph. 462p.

Hongyu, K., Garcia-Peña, M., Borges de Araujo, L. and Dos Santos Dias, C. 2014. Statistical analysis of yield trials by AMMI analysis of genotype x environment interaction. Biometrical Letters, 51(2): 89-102.

Horn, R. and Baumgartl, T. 1999. Dynamics properties of soils. In: Handbook of soil science. New York: CRC Press, p. 2148.

Joenssen, D. and Bankhofer, U. 2012. Hot Deck Methods for Imputing Missing Data. The Effects of Limiting Donor Usage *in.* Petra Perner (Ed). Machine Learning and Data Mining in Pattern recognition, pp (63-75). Springer-Verlag Berlin, Germany.

Malosetti, M., Ribaut, J. M. and van Eeuwijk, F. A. 2013. The statistical analysis of multi-environment data: Modelling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology, 4: 1–17.

ODEPA. 2018. Distribución de la superficie frutal por provincias. Recuperado de: https://reportes.odepa.gob.cl/#/catastro-superficie-fruticola-regional.

Orellana, M., Edwards, J. and Carriquiry, A. 2014. Heterogeneous variances in multi-environment yield trials for corn hybrids. Crop Science, 54: 1048-1056.

Ormeño, J. 1998. El centeno (Secale cereale) como cubierta vegetal y mulch para un manejo alternativo de malezas durante el establecimiento de árboles frutales. Chile Agrícola, Noviembre, p. 273.

Ovalle, C., González A, M. I., del Pozo L, A., Hirzel C, J. y Hernaiz, V. (2007a). Cubiertas Vegetales en Producción Orgánica de Frambuesa: Efectos sobre el Contenido de Nutrientes del Suelo y en el Crecimiento y Producción de las Plantas. Agricultura Técnica, 67(3): 271–280.

Ovalle, C., Del Pozo, A., Lavín, A., y Hirzel, J. 2007b. Cubiertas vegetales en viñedos: Comportamiento de mezclas de leguminosas forrajeras anuales y efectos sobre la fertilidad del suelo. Agricultura Técnica, 67(4): 384–392.

Peña, D. 2002. Análisis de datos multivariante. McGraw Hill. Madrid, España. 539p.

Piepho, H. 1994. Best Linear Unbiased Prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. Theoretical and Applied Genetics, 89: 647-654.

Piepho, H. P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoretical and Applied Genetics, 97(1–2): 195–201.

Pinheiro J. and Bates D. 2000. Mixed effects models in S and S plus. Springer-Verlag, New York, USA. 537p.

Pinheiro J., Bates D., DebRoy S., Sarkar D. and R Core Team. (2018). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131. URL: https://CRAN.R-project.org/package=nlme>.

Przystalski, M., Osman, A., Thiemt, E. M., Rolland, B., Ericson, L., Østergård, H. and Krajewski, P. 2008. Comparing the performance of cereal varieties in organic and non-organic cropping systems in different European countries. Euphytica, 163(3): 417–433.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rawls, W. J., Pachepsky, Y. A., Ritchie, J. C., Sobecki, T. M. and Bloodworth, H. (2003). Effect of soil organic carbon on soil water retention. Geoderma, 116(1–2): 61–76.

Rodríguez, P., Malosetti, M., Gauch, H. and Fred A. Van Eeuwijk, F. 2014. A Weighted AMMI Algorithm to Study Genotype-by-Environment Interaction and QTL-by-Environment Interaction. Crop Science, 54: 1555–1570.

Rombolá A., Baldi E., Franceschi A., Ueno D., Marangoni B., Ma J. and Tagliavini M. 2004. Prevention of iron chlorosis in kiwifruit (Actinidia deliciosa) through cultivation in a mixed cropping system with graminaceous species. Abstract, XII International Symposium on Iron Nutrition and Interactions in Plants, p. 30.

Rubin, D. 1976. Inference and Missing Data. Biometrika, 63(3): 581-592.

Schmitt, P., Mandel, J. and Guedj, M. 2015. A comparison of six methods for missing data imputation. Journal of Biometrics and Biostatistics. 6(1) :1-6.

Singh, M., Ceccarelli, S. and Grando, S. 1999. Genotype x environment interaction of crossover type: Detecting its presence and estimating the crossover point. Theoretical and Applied Genetics, 99:988–995.

Stekhoven, D. and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1): 112–118.

Stroup, W. and Mulitze, D. 1991. Nearest Neighbor Adjusted Best Linear Unbiased Prediction. The American Statistician, 45(3): 194-200.

So, S. and Edwards, J. 2011. Predictive ability assessment of linear mixed models in multienvironment trials in corn. Crop Science, 51: 542-552.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6): 520-525.

Varnero, M. 1992. El suelo como sistema biológico. En: Suelos, una visión actualizada del recurso. Santiago: Facultad de Ciencias Agronómicas y Forestales de la Universidad de Chile, p. 345.

West, B., Welch, K. and Galecki, A. 2015. Linear mixed model, a practical guide using statistical software. CRC Press Boca Ratón, Florida, USA. 405p.

Westcott, B. 1986. Some methods of analysing genotype—environment interaction. Heredity, 56(2): 243–253.

Woyann, L., Benin, G., Storck, L., Trevizan, D., Meneguzzi, C. Marchioro, V., Tonatto, M. and Madureira, A. 2016. Estimation of missing values affects important aspects of GGE biplot analysis. Crop Science, 57: 1–13.

Xu, Y. 2010. Molecular plant breeding. Cab International.México DF. México. 734p.

Yan, W. 2001. GGEbiplot—A Windows Application for Graphical Analysis of Multienvironment Trial Data and Other Types of Two-Way Data. Agronomy Journal, 93: 1111–1118.

Yan, W. and Tinker, N.A. 2006. Biplot analysis of multi-environment trial data: Principles and applications. Can. J. Plant Sci. 86: 623–645.

Yan, W., Kang, M., Ma, B., Wood, S. and Cornelius, P. 2007. GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. Crop Science, 47: 641-653.

Yan, W. 2013. Biplot analysis of incomplete two-way data. Crop Science, 53(1): 48–57.

Yang, R., Crossa, J., Cornelius, P. and Burgueño, J. 2009. Biplot Analysis of Genotype × Environment Interaction: Proceed with Caution. Crop Science, 49: 1564–1576.

Yang, R. 2014. Analysis of linear and non-linear genotype× environment interaction. Frontiers in Genetics, 5: 1–7.

Zhang, S. 2012. Nearest neighbor selection for iteratively kNN imputation. The Journal of Systems and Software, 85: 2541–2552.