



FACULTAD DE CIENCIAS AGRARIAS  
UNIVERSIDAD NACIONAL DE ROSARIO

Paquete de R y aplicación web Shiny para el análisis de datos  
provenientes de ensayos multiambientales

JULIA ANGELINI

TRABAJO FINAL PARA OPTAR AL TÍTULO DE ESPECIALISTA EN  
BIOINFORMÁTICA

---

DIRECTOR: Dr. Gerardo Cervigni  
CO-DIRECTOR: Mgs. Marcos Prunello

AÑO: 2022

---

# Paquete de R y aplicación web Shiny para el análisis de datos provenientes de ensayos multiambientales

Julia Angelini

Licenciada en Estadística – Universidad Nacional de Rosario

Este Trabajo Final es presentado como parte de los requisitos para optar al grado académico de Especialista en **Bioinformática**, de la Universidad Nacional de Rosario y no ha sido previamente presentada para la obtención de otro título en ésta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en **el Centro de Estudios Fotosintéticos y Bioquímicos (CEFOBI)**, durante el período comprendido entre **los años 2017 y 2022**, bajo la dirección del **Dr. Gerardo Cervigni** y la co-dirección del **Mgs. Marcos Prunello**.

Nombre y firma del autor

Lic. Julia Angelini

Nombre y firma del Director

Dr. Gerardo Cervigni

Nombre y firma del Co - Director

Mgs. Marcos Prunello

Defendida: \_\_\_\_\_ de 20\_\_\_\_\_.

---

# Agradecimientos

*En este trabajo final, directa o indirectamente, participaron muchas personas a las que les quiero agradecer.*

*En primer lugar al Dr. Gerardo Cervigni por haberme propuesto realizar la Especialización Bioinformática, compartir su conocimiento y experiencia a lo largo de todo el proceso, contagiando su pasión, entusiasmo y energía.*

*Al Mgs. Marcos Prunello por acompañarme en el desarrollo del trabajo final, por su dedicación, sus consejos y su ejemplo que me incentiva a superarme como profesional. Sin su confianza, apoyo y atención, este trabajo no hubiera sido posible. No sólo me enriquecí en lo académico sino también con la amistad que pudimos forjar.*

*Al Centro Computacional del Centro Científico Tecnológico de Rosario, miembro del Sistema Nacional de Computación de Alto Rendimiento, por la predisposición, asesoramiento e instalación de los recursos adicionales necesarios para este trabajo.*

*Al Dr. Sergio Arciniegas Alarcón por su predisposición en la inclusión de los avances metodológicos realizados por su equipo de investigación en este trabajo.*

*A mis compañeros de la Especialización, por las largas horas de cursos, mates y almuerzos. En especial, a Jor y Lu, por el aliento en todo momento, por compartir excelentes momentos y porque gracias a la ayuda de ambas he podido entender cosas que no habría podido sola.*

*A los docentes de la Especialización en Bioinformática por su dedicación y paciencia para enseñarle a alumnos provenientes de las más diversas áreas esta hermosa combinación entre Biología, Informática y Estadística.*

*A mis padres por el amor y apoyo incondicional y por el esfuerzo de dedicar sus vidas a brindarnos a mi hermano y a mí la posibilidad de construir nuestros futuros. A mi hermano, por su cariño, apoyo, acompañamiento y sentido del humor. A Otto, por su incomparable mezcla de amor y comprensión, por darme fuerzas en los momentos de debilidad y por alentar-me a seguir a pesar de todo. A Segundo, Mia y Kalita, por su hermosa compañía día a día.*

*Por último, pero no menos importante, a Gaby y Euge mis compañeras de CEFOTI, por acompañarme en las partes más empedradas del camino, por compartir las risas y las lágrimas, por su amistad y consejos. No hubiese alcanzado mucho de mis logros sin su ayuda, compañía y aliento en todo momento.*

---

# Abreviaturas

**ACP:** análisis de componentes principales.

**AEC:** coordenada ambiental promedio (siglas en inglés de *Average environment coordination*).

**ANOVA:** análisis de la variancia (siglas en inglés de *analysis of variance*).

**AMMI:** modelo de efectos principales aditivos e interacción multiplicativa (siglas en inglés de *Additive Main effects and Multiplicative Interaction*).

**CCT:** Centro Científico Tecnológico.

**COI:** interacción con cambio de rango (siglas en inglés de *crossover interaction*).

**CONICET:** Consejo Nacional de Investigaciones Científicas y Técnicas.

**CRAN:** *Comprehensive R Archive Network*.

**DVS:** descomposición en valores singulares.

**EM:** maximización de la esperanza (siglas en inglés de *Expectation Maximization*).

**EMA:** ensayos multiambientales.

**G:** efecto genotípico.

**GE:** genotipo-ambiente (siglas en inglés de *Genotype-Environment*).

**GGE:** genotipo más genotipo-ambiente (siglas en inglés de *Genotype plus Genotype-Environment*).

**IGA:** interacción genotipo-ambiente.

**NCOI:** interacción sin cambio de rango (siglas en inglés de *no crossover interaction*).

**QTL:** siglas en inglés de *Quantitative Trait Loci*.

**SREG:** modelo de regresión por sitio (siglas en inglés de *Site Regression model*).

**SVP:** partición de los valores singulares (siglas en inglés de *Singular Value Partition*).

**ui:** interfaz del usuario (siglas en inglés de *user interface*).

---

# Resumen

Las variedades mejoradas de cultivos vegetales son el resultado del trabajo de desarrollo genético llevado a cabo en los programas de fitomejoramiento, los cuales se extienden a lo largo de varios años y requieren cuantiosas inversiones. En etapas avanzadas, los ensayos multiambientales (EMA), que comprenden experimentos en múltiples ambientes, son herramientas fundamentales para incrementar la productividad y rentabilidad de los cultivos. La vigencia comercial de las variedades puede extenderse durante varias décadas, por lo que su elección es crítica para que el productor evite pérdidas económicas por malas campañas y el suministro al mercado sea constante. Consecuentemente, un análisis adecuado de la información de los EMA es indispensable para asegurar el éxito del programa de mejoramiento de cultivos. Actualmente, R es uno de los lenguajes de programación más utilizados para el análisis de datos debido a su distribución como software libre y a la gran variedad de herramientas que ofrece. Sin embargo, los mejoradores que no están familiarizados con la programación tienden a utilizar programas que responden a instrucciones por menú en lugar de escribir líneas de código, a pesar de los costos económicos derivados del pago de sus licencias. Aquellos que sí tienen afinidad con el uso de código para el análisis de datos se enfrentan con dificultades a la hora de identificar las herramientas apropiadas entre el gran número de instrumentos disponibles. Por lo tanto, en este trabajo se presenta el desarrollo de dos herramientas informáticas para asistir en el análisis de datos provenientes de EMA. Por un lado, se creó un nuevo paquete de R que incluye metodología recientemente publicada que no se encuentra disponible en el software y al mismo tiempo reúne todas aquellas de mayor utilidad, de modo que aquellos usuarios que posean un manejo del lenguaje puedan simplificar su tarea. Por otro lado, se confeccionó una interfaz gráfica de usuario mediante una aplicación web Shiny que permite realizar los principales análisis implementados en el paquete sin necesidad de programar y se encuentra publicada en internet para su libre acceso.

**Palabras Clave:** análisis estadístico, ensayos multiambientales, interfaz gráfica, lenguaje R, programación.

---

# Abstract

Crop improvement is the result of genetic development which requires several years and large investments. In advanced stages of breeding programs, multi-environment trials (MET), which consist of evaluating different cultivars in multiple environments, are essential tools to increase crop productivity. Since varieties remain on market for decades, their choice is essential to avoid economic losses due to bad seasons and to ensure a constant supply. Consequently, an adequate analysis of MET data is essential to guarantee the success of a breeding program. Currently, R is one of the most widely used programming language for data analysis due to its distribution as free software and the wide variety of tools it offers. However, breeders who are unfamiliar with programming tend to use other types of programs that respond to menu prompts instead of writing lines of code, despite the financial costs of their licenses. Whereas, those who have an affinity with the use of code for data analysis face difficulties in identifying the right tools from the large number of instruments available. Therefore, in this work two tools are develop for MET data analysis. On one hand, a new R package that includes new methodology not available in the software and at the same time brings together all those most useful created to facilitate the users task. On the other hand, a graphical user interface was created using a Shiny web application that allows the main analyzes implemented in the package to be carried out without the need for programming and is published on internet for free access.

**Keywords:** multi-environment trials, programming, statistical analysis, user interfaz, R language.

---

# Índice general

Capítulos	Página
<b>1. Introducción</b>	<b>1</b>
<b>2. Objetivos</b>	<b>6</b>
2.1. Objetivo general . . . . .	6
2.2. Objetivos específicos . . . . .	6
<b>3. Métodos</b>	<b>7</b>
3.1. Métodos estadísticos . . . . .	7
3.1.1. Modelo AMMI y SREG . . . . .	7
3.1.2. Modelo AMMI robusto . . . . .	10
3.1.3. Métodos de imputación . . . . .	11
3.2. Creación de un paquete de R . . . . .	13
3.2.1. <i>geneticae</i> . . . . .	13
3.2.2. Estructura general del paquete . . . . .	14
3.2.3. Archivo DESCRIPTION . . . . .	15
3.2.4. Archivos de código . . . . .	16
3.2.5. Documentación . . . . .	18
3.2.6. Uso de funciones de otros paquetes . . . . .	19
3.2.7. Testeos . . . . .	20
3.2.8. <i>Datasets</i> . . . . .	21

3.2.9. Archivo README . . . . .	21
3.2.10. Archivo NEWS . . . . .	22
3.2.11. Viñetas . . . . .	23
3.2.12. R CMD check e instalación . . . . .	24
3.2.13. Publicación y difusión . . . . .	24
3.3. Aplicación web Shiny . . . . .	24
3.3.1. Programación de una aplicación web Shiny . . . . .	25
<b>4. Resultados</b>	<b>27</b>
4.1. Paquete de R <i>geneticae</i> . . . . .	27
4.1.1. Conjuntos de datos en <i>geneticae</i> . . . . .	28
4.1.2. Uso del paquete para ajustar el modelo AMMI . . . . .	28
4.1.3. Uso del paquete para ajustar el modelo SREG . . . . .	31
4.1.4. Uso del paquete para imputar matrices de datos incompletas . . . . .	40
4.2. Aplicación web <i>Geneticae</i> . . . . .	41
4.2.1. Lectura de un archivo de datos para el uso de la aplicación web <i>Geneticae</i>	42
4.2.2. Uso de la aplicación web <i>Geneticae</i> para análisis descriptivo . . . . .	44
4.2.3. Uso de la aplicación web <i>Geneticae</i> para ajustar el modelo SREG . .	46
4.2.4. Uso de la aplicación web <i>Geneticae</i> para ajustar el modelo AMMI . .	49
4.2.5. Ayuda . . . . .	50
<b>5. Conclusión</b>	<b>51</b>
<b>Bibliografía</b>	<b>53</b>



---

# Índice de Figuras

Figura 1.1: Representación gráfica de tipos de interacción genotipo - ambiente: (A) <i>crossover</i> , (B) <i>no crossover</i> y (C) sin interacción. . . . .	2
Figura 3.1: Chequeo de disponibilidad del nombre <i>geneticae</i> elegido para el paquete en desarrollo mediante la función <code>available()</code> del paquete <i>available</i> . . . . .	14
Figura 3.2: Creación del paquete <i>geneticae</i> mediante la función <code>create_package()</code> del paquete <i>usethis</i> . . . . .	15
Figura 3.3: Archivo DESCRIPTION de <i>geneticae</i> . . . . .	17
Figura 3.4: Fragmento de la función <code>GGEmodel()</code> del paquete <i>geneticae</i> . . . . .	17
Figura 3.5: Fragmento de los comentarios roxygen de la función <code>GGEmodel()</code> del paquete <i>geneticae</i> . . . . .	19
Figura 3.6: Resultado de correr la función <code>test()</code> del paquete <i>devtools</i> en <i>geneticae</i> . . . . .	20
Figura 3.7: Porcentaje del código de <i>geneticae</i> que es evaluado durante los tests obtenidos mediante la función <code>test_coverage()</code> del paquete <i>covr</i> . . . . .	21
Figura 3.8: Fragmento del archivo README mostrado en el repositorio GitHub del paquete <i>geneticae</i> . . . . .	22
Figura 3.9: Archivo NEWS de <i>geneticae</i> . . . . .	23
Figura 4.1: Biplot GE derivado del modelo AMMI clásico basado en los datos de rendimiento de trigo de invierno obtenidos en Ontario en 1993. El 71,66% de la variabilidad de la IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en letras minúsculas y los ambientes en mayúsculas. . . . .	30

Figura 4.2: Biplot GGE basado en datos de rendimiento de trigo de invierno obtenido de Ontario en 1993. El método de partición de los valores singulares utilizado es el simétrico (opción por defecto). El 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en minúsculas y los entornos en mayúsculas. . . . .	33
Figura 4.3: Ranking de (A) cultivares en el ambiente OA93 y (B) ambientes para cultivar Kat, basado en datos de rendimiento de trigo de invierno obtenido de Ontario en 1993. El método de partición de los valores singulares utilizado es el simétrico (opción por defecto). El 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en minúsculas y los entornos en mayúsculas. . . . .	34
Figura 4.4: Comparación de los cultivares Kat y Cas. El método de partición de los valores singulares utilizado es el simétrico (opción por defecto). El 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en minúsculas y los entornos en mayúsculas. . . . .	35
Figura 4.5: Vista poligonal del biplot GGE, que muestra qué cultivares presentaron mayor rendimiento en cada mega-ambiente. El método de partición de los valores singulares utilizado es el simétrico (opción por defecto). El 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en minúsculas y los entornos en mayúsculas. . . . .	36
Figura 4.6: (A) Evaluación de los cultivares con base en el rendimiento promedio y la estabilidad y (B) clasificación de genotipos con respecto al genotipo ideal, basado en el método de partición de los valores singulares enfocado en los genotipos. . . . .	38
Figura 4.7: (A) Relación entre ambientes y (B) clasificación de ambientes con respecto al ambiente ideal, basado en el escalado centrado en los genotipos. . . . .	40
Figura 4.8: Importar el conjunto de datos <i>yanwinterwheat</i> en la aplicación web <i>Geneticae</i> . . . . .	43

Figura 4.9: Diagrama de caja de (A) genotipos y (B) ambientes para el conjunto de datos <i>yanwinterwheat</i> obtenido con la aplicación web <i>Geneticae</i> . .	44
Figura 4.10: (A) Gráfico y (B) matriz de correlación entre genotipos para el conjunto de datos <i>yanwinterwheat</i> obtenido con la aplicación web <i>Geneticae</i> . .	45
Figura 4.11: Gráfico de interacción para (A) ambientes a través de genotipos y (B) genotipos a través de los ambientes para conjunto de datos de <i>yanwinterwheat</i> obtenido con la aplicación web <i>Geneticae</i> . . . . .	46
Figura 4.12: Vistas del biplot GGE usando la partición simétrica de los valores singulares obtenidos con la aplicación web <i>Geneticae</i> . . . . .	47
Figura 4.13: Vistas del biplot GGE usando la partición de los valores singulares enfocada en los genotipos obtenidos con la aplicación web <i>Geneticae</i> . .	48
Figura 4.14: Vistas del biplot GGE usando la partición de los valores singulares enfocada en los ambientes obtenidas con la aplicación web <i>Geneticae</i> . .	48
Figura 4.15: Biplot GE derivado del modelo AMMI clásico basado en datos de rendimiento de trigo de invierno obtenido de Ontario en 1993 obtenido con la aplicación web <i>Geneticae</i> . . . . .	49

---

# Capítulo 1

## Introducción

A lo largo de la historia de la agricultura, el hombre ha desarrollado el mejoramiento vegetal en forma sistemática y lo ha convertido en un instrumento esencial para incrementar la producción agrícola en términos de cantidad, calidad y diversidad. Las variedades mejoradas son el resultado del trabajo llevado a cabo en los programas de fitomejoramiento, los cuales se extienden a lo largo de varios años y requieren cuantiosas inversiones. En etapas avanzadas de estos programas, comúnmente se llevan a cabo ensayos multiambientales (EMA) de comparación de rendimientos, donde un conjunto de variedades se evalúan en múltiples ambientes. Éstos son esenciales ya que además de los efectos genotípicos y ambientales, se puede detectar un efecto adicional, el proporcionado por la interacción entre ambos (Cruz y Regazzi, 1997). La interacción genotipo ambiente (IGA) es la respuesta diferencial de los genotipos a través de un rango de ambientes y es considerada, casi unánimemente por los fitomejoradores, como el principal factor limitante para la selección de cultivares superiores, disminuyendo la eficiencia de los programas de mejoramiento (Crossa et al., 1990; Cruz Medina, 1992; Kang y Magari, 1996). Cuando los ambientes son muy diferentes, la IGA usualmente gana importancia porque cambia el rango de las líneas de mejoramiento. Gauch y Zobel (1997) explicaron que si no hubiera interacción, una sola variedad o híbrido rendirían al máximo en todo el mundo, además los materiales podrían evaluarse en un solo lugar y proporcionarían resultados universales.

Peto (1982) ha distinguido las interacciones cuantitativas, conocidas también como sin cambio de rango o *no crossover interaction* (NCOI), de las cualitativas, denominadas a su vez como con cambio de rango o *crossover interaction* (COI). Cuando dos genotipos  $G_1$  y  $G_2$  tienen una respuesta diferencial en dos ambientes, se dice que la IGA es del tipo COI si

hay cambios en el orden de los genotipos según su rendimiento (Figura 1.1(A)) y del tipo NCOI si su ordenamiento permanece sin cambios (Figura 1.1(B)). Por otro lado, se dice que la IGA es inexistente cuando los genotipos responden de manera similar en ambos ambientes (Figura 1.1(C)).

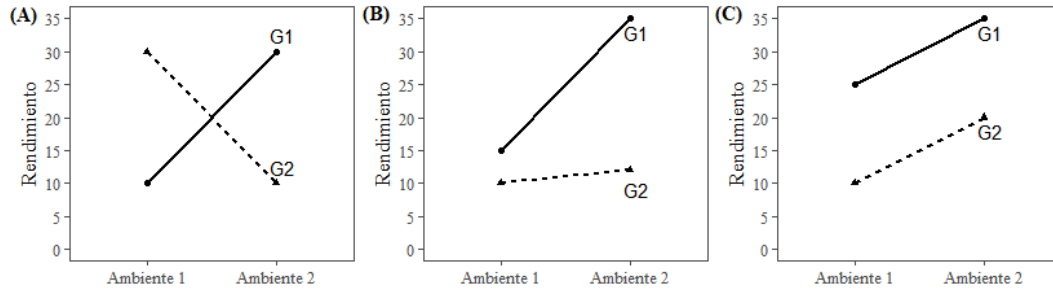


Figura 1.1: Representación gráfica de tipos de interacción genotipo - ambiente: (A) *crossover*, (B) *no crossover* y (C) sin interacción.

Los estudios de adaptabilidad y estabilidad fenotípica tienen como objetivo analizar el comportamiento de los materiales en diferentes ambientes, interesándose en la respuesta diferencial de los genotipos a la variación ambiental (Borém, 2001). En general, se ha aceptado que, a mayor variabilidad genética de una especie/población, mayor será su estabilidad sobre el ambiente. Allard y Bradshaw (1964) indican que una población podría estar conformada por individuos diferentes, cada uno de ellos adaptados a un rango de ambientes, o conformada por individuos semejantes, pero cada uno adaptado a un rango de ambientes. La adaptabilidad se refiere a la habilidad del genotipo de tener buen desempeño (por ejemplo; rendimientos altos) con respecto a determinadas condiciones ambientales. Conceptualmente la adaptabilidad es dividida en amplia y específica. La primera refiere a cultivares adaptados a una amplia red ambiental, y la segunda a una adaptación regional o mega-ambiente. Por otro lado, también la estabilidad es conceptualizada. Según Becker (1981) existen dos tipos de estabilidad. Una denominada estática con sentido homeostático, y otra llamada dinámica o agronómica. La primera, caracteriza al desempeño de materiales que muestran variancia cero entre ambientes y no es deseable por los mejoradores ya que los materiales no responden a la mejora ambiental. La segunda, considera un material estable si su performance es buena respecto al potencial del ambiente. La estabilidad estática puede asociarse a una escasa IGA, en términos de ecovalencia, o falta de interacción, y la estabilidad dinámica con la interacción simple o NCOI. La COI es consecuencia del comportamiento impredecible de los materiales y es, por lo general, una limitante seria en la selección de genotipos.

La IGA es útil si el objetivo del programa es el desarrollo de materiales con adaptación específica. Es decir, la identificación de genotipos que tienen un comportamiento destacado en una determinada región o mega-ambiente, lo que permite señalar nichos ambientales propicios para una mayor productividad y calidad. Por el contrario, la COI complica el proceso de selección ya que disminuye la correlación de los valores fenotípicos entre los ambientes, dificultando la identificación de aquellos genotipos con adaptación amplia.

Un análisis adecuado de la información de los EMA es indispensable para el éxito del programa de mejoramiento genético de los cultivos. El rendimiento medio de los genotipos en los ambientes es un indicador suficiente del desempeño de ellos sólo en ausencia de IGA (Yan y Kang, 2003). Sin embargo, la aparición de IGA es inevitable y no basta con la comparación de las medias de los genotipos, sino que se debe recurrir a una metodología estadística más apropiada. Las más difundidas para analizar los datos provenientes de EMA se basan en modificaciones de los modelos de regresión, análisis de variancia (ANOVA, siglas en inglés de *analysis of variance*) y técnicas de análisis multivariado.

Particularmente, para el estudio de la IGA y los análisis que de ella se derivan, dos modelos multiplicativos han aumentado su popularidad entre los fitomejoradores como herramientas de análisis gráfico: el modelo de los efectos principales aditivos e interacción multiplicativa (AMMI, siglas en inglés de *Additive Main effects and Multiplicative Interaction*) (Gauch, 1988, 1992) y el de regresión por sitio (SREG, siglas en inglés de *Site Regression model*) (Cornelius et al., 1996; Crossa y Cornelius, 1997). Estos modelos se ajustan en dos etapas. Primero, se realiza un ANOVA para obtener estimaciones de los efectos principales aditivos de ambientes y genotipos (G) en AMMI y sólo de los ambientes en SREG. En segundo lugar, los residuos del ANOVA se ordenan en una matriz con genotipos en las filas y ambientes en las columnas y se aplica una descomposición en valores singulares (DVS), representando los patrones de IGA presentes en los residuos en AMMI y de G e IGA conjuntamente en SREG. El resultado de los dos primeros términos multiplicativos de la DVS a menudo se presentan en un biplot llamado GE (genotipo-ambiente, siglas en inglés de *Genotype-Environment*) para el modelo AMMI (Zobel et al., 1988) y GGE (genotipo más genotipo-ambiente, siglas en inglés de *Genotype plus Genotype-Environment*) para SREG (Yan et al., 2000). Sin embargo, estos modelos no siempre son lo suficientemente eficientes para analizar la estructura de datos provenientes de EMA de programas de mejoramiento vegetal (de Oliveira et al., 2016; Jarquín et al., 2016; Hadasch et al., 2018). Por un lado, tienen serias limitaciones frente a información faltante y, a pesar de que los EMA están diseñados para que todos los genotipos se evalúen en todos los ambientes, la presencia de valores perdidos es muy común (Woyann et al., 2017;

Aguate et al., 2019). Esto ocurre, por ejemplo, debido a errores de medición o destrucción de plantas por presencia de animales, inundaciones o problemas durante la cosecha, además de la dinámica propia de las evaluaciones en las que se incorporan y se descartan genotipos debido a su mal desempeño (Hill y Rosenberg, 1985). Numerosas metodologías de imputación se han estado desarrollando en los últimos años para solventar esta limitación (Arciniegas-Alarcón et al., 2010, 2014; Josse y Husson, 2016; Arciniegas-Alarcón et al., 2020). Por otro lado, ambos modelos son sensibles a la presencia de observaciones atípicas, lo cual es una regla más que una excepción cuando se consideran datos reales. Para superar esta fragilidad, recientemente distintas metodologías robustas se han desarrollado para el modelo AMMI (Rodrigues et al., 2016).

En este contexto, el análisis de datos provenientes de EMA requiere metodología estadística cuyas rutinas informáticas no se encuentran disponibles en programas comerciales debido a su reciente desarrollo o bien se deben utilizar varios de ellos para cumplir un único objetivo. Esto último genera el inconveniente de tener que disponer de todos los programas necesarios para los distintos análisis, atender los requerimientos de formatos de datos usados por cada uno de ellos y comprender los diversos tipos de salidas en las que se presentan los resultados obtenidos. Además, los costos de las licencias algunos programas pueden resultar muy elevados.

Ante estas dificultades, una alternativa para el análisis es el empleo de algún lenguaje de programación de distribución libre y gratuita, que le confiera al analista la flexibilidad necesaria para cumplir con su objetivo. En este contexto, R es uno de los lenguajes de programación desarrollados para el análisis de datos de mayor uso en la actualidad. R es un software de uso libre y distribuido bajo los términos de la *General Public Licence*. Este programa se descarga de un repositorio mantenido por *The R Foundation for Statistical Computing* conocido como CRAN (*Comprehensive R Archive Network*), en el cual también se encuentran disponibles miles de paquetes adicionales que consisten en conjuntos de funciones desarrolladas con fines específicos que se distribuyen con un protocolo determinado, garantizando su correcto funcionamiento. Cualquier desarrollador puede producir su propio paquete y publicarlo en CRAN, siempre que cumpla con los requisitos establecidos y pase correctamente por los procedimientos de control. Además, hay paquetes que pueden obtenerse de otros repositorios como GitHub, Bioconductor, rOpenSci, entre otros.

R es propicio para el análisis de datos de EMA puesto que se ha desarrollado metodología específica para este entorno computacional. Algunos de los paquetes desarrollados para tal fin son: *agricolae* (de Mendiburu, 2021), *gge* (Wright y Laffont, 2021), *GGEbiplots* (Dumble,

2022) y *metan* (Olivoto y Lúcio, 2020). Frutos et al. (2013) desarrollaron el paquete *GGEbiplotGUI* que consistía en una implementación computacional interactiva para el análisis de datos EMA que requería un mínimo conocimiento del lenguaje R (instalar, cargar librerías e importar bases de datos). Sin embargo, en 2021 fue archivado de CRAN debido a falta de mantenimiento (sigue disponible en GitHub).

A pesar de las ventajas del uso de R, el análisis de datos de EMA en dicho software presenta algunos desafíos. Por un lado, existen numerosos paquetes con funcionalidad afín que hay que identificar cómo combinar adecuadamente. Por otro lado, el software puede resultar dificultoso para aquellos analistas no familiarizados con la programación. Atendiendo a estas dos necesidades, se crea un paquete que incluya metodología recientemente publicada y reúna las funciones más útiles a fin de solventar la primera de ellas. Para la segunda, se crea una aplicación web Shiny de libre acceso mediante conexión a internet que permita realizar los principales análisis implementados en el paquete sin necesidad de escribir líneas de código.



---

# Capítulo 2

## Objetivos

### 2.1. Objetivo general

Desarrollar un paquete de R para el análisis de datos provenientes de EMA y una interfaz gráfica de usuario para el mismo a través de la aplicación web Shiny.

### 2.2. Objetivos específicos

- Mostrar un flujo de trabajo reproducible para la construcción de paquetes de R.
- Programar e incluir en el paquete metodología para el análisis de datos provenientes de EMA recientemente publicada y no disponible en R.
- Añadir en el paquete de R funciones ya existentes con modificaciones o agregados para favorecer su uso.
- Desarrollar una aplicación web Shiny que sirva como interfaz gráfica de usuario para el paquete.
- Publicar el paquete y la aplicación web para su libre uso.

---

# Capítulo 3

## Métodos

La primera parte de este capítulo presenta la metodología estadística para el análisis de datos provenientes de EMA incluída en el paquete de R y que puede ser empleada interactivamente mediante la aplicación web Shiny. En la segunda y tercera sección, se presenta un flujo de trabajo reproducible para el desarrollo del paquete de R y una descripción de las etapas para la creación de la aplicación web Shiny.

### 3.1. Métodos estadísticos

#### 3.1.1. Modelo AMMI y SREG

Para el estudio de la IGA y los análisis que de ella se derivan, dos modelos multiplicativos han aumentado su popularidad entre los fitomejoradores como herramientas de análisis gráfico: el modelo de los efectos principales aditivos e interacción multiplicativa (AMMI, siglas en inglés de *Additive Main effects and Multiplicative Interaction*) (Gauch, 1988, 1992) y el de regresión por sitio (SREG, siglas en inglés de *Site Regression model*) (Cornelius et al., 1996; Crossa y Cornelius, 1997). Estos modelos se ajustan en dos etapas. Primero, se realiza un ANOVA para obtener estimaciones de los efectos principales aditivos de ambientes y genotipos en AMMI y sólo de los ambientes en SREG. En segundo lugar, los residuos del ANOVA se ordenan en una matriz con genotipos en las filas y ambientes en las columnas y se aplica una DVS, representando los patrones de IGA presentes en los residuos en AMMI y de G e IGA conjuntamente en SREG.

Las ecuaciones de los distintos modelos son:

$$\text{AMMI: } y_{ij} = \mu + G_i + A_j + \sum_{k=1}^K \lambda_k \alpha_{ik} \gamma_{jk}$$

$$\text{SREG: } y_{ij} = \mu + A_j + \sum_{k=1}^K \lambda_k \alpha_{ik} \gamma_{jk}$$

donde

- $y_{ij}$  es el carácter fenotípico evaluado (rendimiento o cualquier otro carácter de interés) del  $i$ -ésimo genotipo en el  $j$ -ésimo ambiente,
- $\mu$  es la media general,
- $G_i$  es el efecto del  $i$ -ésimo genotipo con  $i = 1, \dots, g$ ,
- $A_j$  es el efecto del  $j$ -ésimo ambiente con  $j = 1, \dots, a$ ,
- $\sum_{k=1}^K \lambda_k \alpha_{ik} \gamma_{jk}$  es la sumatoria de términos multiplicativos utilizadas para modelar la IGA en AMMI o de G e IGA conjuntamente en SREG.  $K$  es el número de términos multiplicativos retenidos en el modelo con  $K \leq \min(g - 1, a - 1)$  en AMMI y  $K \leq \min(g, a - 1)$  en SREG;  $\lambda_k$  es el  $k$ -ésimo valor singular y  $\alpha_{ik}$  y  $\gamma_{jk}$  son los elementos de los autovectores asociados con el  $i$ -ésimo genotipo y el  $j$ -ésimo ambiente para el  $k$ -ésimo término multiplicativo, respectivamente. En general, los dos primeros términos multiplicativos ( $K = 2$ ) son suficientes para explicar los patrones de la IGA en AMMI y de G e IGA en SREG; la variabilidad remanente se interpreta como ruido aleatorio.

El resultado de los dos primeros términos multiplicativos de la DVS se presenta a menudo en un biplot llamado GE (genotipo-ambiente, siglas en inglés de *Genotype-Environment*) para el modelo AMMI (Zobel et al., 1988) y GGE (genotipo más genotipo-ambiente, siglas en inglés de *Genotype plus Genotype-Environment*) en SREG (Yan et al., 2000) y representan una aproximación de dos rangos de los efectos multiplicativos. Dado que para seleccionar cultivares el G e IGA deben considerarse simultáneamente, el modelo SREG resulta superior a AMMI para visualizar patrones en datos provenientes de EMA. Un biplot GGE que explica suficiente variabilidad debida a G e IGA de un conjunto de datos provenientes de EMA permite, entre otras cosas, visualizar tres aspectos importantes:

- (i) las relaciones entre los genotipos y ambientes representadas por el patrón “cuál-ganó-donde” (*which-won-where*), que facilitan la investigación de mega-ambientes (Gauch y Zobel, 1997).
- (ii) las interrelaciones entre los ambientes de prueba, que facilitan la identificación de mejores ambientes para la evaluación de cultivares (Cooper et al., 1997) y de aquellos que son redundantes y pueden descartarse (Yan y Rajcan, 2002);
- (iii) las interrelaciones entre genotipos en cada mega-ambiente posibilita la comparación

entre ellos y la clasificación de los mismos comparándolos con un genotipo “ideal” que es aquel con el rendimiento más alto y que es absolutamente estable (Yan et al., 2001). Al compararse el orden de los genotipos en cada mega-ambiente, podría identificarse aquellos con adaptabilidad amplia (el orden del genotipo no varía entre los ambientes), y aquellos con adaptabilidad específica (sólo muestra buen desempeño en uno o pocos ambientes). Sin embargo, la posibilidad de que algunos genotipos sean seleccionados o no, dependerá de la comparación que se realice con los genotipos controles que se incluyen en los ensayos.

Aunque el biplot GGE se usó inicialmente solo para la exploración de los efectos conjuntos de G e IGA, su aplicación se ha extendido a cualquier conjunto de datos que tenga una estructura de doble entrada. En el área de fitomejoramiento en particular, el biplot GGE se ha utilizado para abordar preguntas importantes que es probable que un fitomejorador o investigador plantee. Hasta el momento, ha permitido una visualización holística de la asociación genotipo x carácter, datos de cruces dialélicos, genotipo x marcador, análisis de QTL (siglas en inglés de *Quantitative Trait Loci*) con datos de mapeo, y de interacción planta x patógeno (Yan y Kang, 2003; Singh et al., 2020; Ezequiel-Hernández et al., 2020; Adu et al., 2021).

En un biplot la puntuación del  $i$ -ésimo genotipo en la  $k$ -ésima componente principal se muestra como un punto definido por  $g_{ik} = \lambda_k^s \alpha_{ik}$  y la correspondiente al  $j$ -ésimo ambiente en la  $k$ -ésima componente por  $e_{kj} = \lambda_k^{1-s} \gamma_{jk}$  donde  $k = 1, 2$  para un biplot bidimensional y  $s$  es el factor de partición de los valores singulares. Teóricamente, el factor de partición puede tomar cualquier valor entre 0 y 1. Dentro de este rango, la elección de  $s$  no altera las relaciones o interacciones relativas entre los genotipos y los ambientes, aunque la apariencia del biplot será diferente. Cuando  $s = 1$ ,  $g_{ik} = \lambda_k \alpha_{ik}$  y  $e_{kj} = \gamma_{jk}$ , los valores singulares se dividen por completo en los autovectores de los genotipos. En esta escala la unidad de las puntuaciones de los genotipos ( $g_{ik}$ ) es la unidad original del carácter fenotípico evaluado y las puntuaciones ambientales ( $e_{kj}$ ) están normalizadas (es decir no tienen unidad). Cuando  $s = 0$ ,  $g_{ik} = \alpha_{ik}$  y  $e_{kj} = \lambda_k \gamma_{jk}$ , los valores singulares se dividen por completo en los autovectores de los ambientes. En esta escala las puntuaciones ambientales están en la unidad original del carácter fenotípico evaluado y las de los genotipos no tienen unidad. Cuando  $s = 0,5$ ,  $g_{ik} = \lambda_k^{0,5} \alpha_{ik}$  y  $e_{kj} = \lambda_k^{0,5} \gamma_{jk}$ , la partición es simétrica. En esta escala las puntuaciones de los genotipos y las ambientales tienen la misma unidad que es la raíz cuadrada de la unidad original. El valor de  $s = 0,5$  es empleado en el biplot GE y el más utilizado en GGE, aunque dependiendo de los intereses de la investigación, se pueden construir numerosas

vistas del biplot GGE derivado de SREG. Independientemente del factor de partición en valores singulares utilizado, los biplots GGE revelan el mismo patrón *which-won-where*. Sin embargo, difieren en su precisión al mostrar la interrelación entre ambientes y genotipos. La partición centrada en los genotipos ( $s = 1$ ) muestra la interrelación entre genotipos con mayor precisión que cualquier otro método; la partición enfocada en los ambientes ( $s = 0$ ) es la más informativa sobre las interrelaciones entre los ambientes; y la simétrica ( $s = 0,5$ ) permite visualizar la magnitud relativa tanto de la variación de los genotipos como de los ambientes.

### 3.1.2. Modelo AMMI robusto

El modelo AMMI, en su forma estándar, asume que no hay valores atípicos en el conjunto de datos. Sin embargo, la presencia de *outliers* es más una regla que una excepción cuando se consideran datos agronómicos debido a características inherentes de los genotipos que se evalúan, errores de medición o el efecto inesperado de plagas o enfermedades que pueden afectar el rendimiento de algunos genotipos.

Rodrigues et al. (2016) proponen una generalización robusta del modelo AMMI, que resulta de ajustar un modelo lineal robusto basado en el estimador M-Huber (Huber, 1981) y luego utilizar un procedimiento de DVS o de análisis de componentes principales (ACP) robusto. Para la DVS o el ACP los autores consideraron varios métodos, dando lugar a un total de cinco modelos robustos llamados:

- R-AMMI: utiliza la DVS robusta propuesta por Hawkins et al. (2002) en la cual se reemplaza la norma euclídea por la suma de valores absolutos para calcular una aproximación robusta a la DVS de una matriz rectangular.
- H-AMMI: se basa en el ACP robusto propuesto por Hubert et al. (2005) que combina las ventajas de un ACP basado en una matriz de covarianza robusta y uno basado en la búsqueda de proyección.
- G-AMMI: considera el algoritmo *Grid* robusto propuesto por Croux et al. (2007) que utiliza técnicas de búsqueda de proyección para calcular estimadores del ACP.
- L-AMMI: utiliza ACP esférico robusto propuesto por Locantore et al. (1999) que consiste en realizar ACP clásico en los datos proyectados en una esfera unitaria.
- PP-AMMI: considera una técnica de búsqueda de proyección robusta propuesta por Croux y Ruiz-Gazen (2005) que calcula los autovalores y autovectores secuencialmente sin estimar una matriz de covarianza robusta.

El empleo de la versión robusta del modelo AMMI puede ser extremadamente útil debido a que una mala representación de genotipos y ambientes puede resultar en una mala decisión con respecto a qué genotipos seleccionar para un conjunto dado de ambientes (Gauch y Zobel, 1997; Yan et al., 2000). A su vez, la elección de los genotipos incorrectos pueden provocar grandes pérdidas en términos de rendimiento. Los biplots obtenidos de los modelos robustos mantienen las características e interpretación estándar del modelo AMMI clásico (Rodrigues et al., 2016).

### 3.1.3. Métodos de imputación

Una limitación importante que presentan los modelos multiplicativos descriptos previamente es que requieren que el carácter fenotípico bajo estudio se encuentre registrado para todas las combinaciones entre genotipos y ambientes, es decir no admiten valores perdidos. Aunque los EMA están diseñados para que todos los genotipos se evalúen en todos los ambientes, la presencia de valores faltantes es muy común debido a errores de medición o pérdidas de plantas por animales, inundaciones o problemas durante la cosecha, además de la dinámica propia de las evaluaciones en las que se incorporan o descartan genotipos debido a su mal desempeño (Hill y Rosenberg, 1985).

Para superar el problema de los datos incompletos, existen diversas soluciones: (i) extraer el subconjunto completo de datos eliminando aquellos genotipos o ambientes con valores faltantes, generando una gran pérdida de información; (ii) completar las celdas incompletas utilizando la media ambiental, lo cual no es una buena estrategia, especialmente cuando la cantidad de valores perdidos es grande; (iii) considerar un método más complejo que permita la falta de datos; y (iv) completar las celdas faltantes con valores estimados utilizando métodos de imputación (Gauch y Zobel, 1990; Troyanskaya et al., 2001; Arciniegas-Alarcón et al., 2010; Paderewski, 2013; Arciniegas-Alarcón et al., 2014; Josse y Husson, 2016; Arciniegas-Alarcón et al., 2020). Dado que las dos primeras opciones no resultan apropiadas y la tercera complejiza mucho el análisis, en el paquete se incluyen metodologías de imputación para poder ajustar el modelo AMMI o SREG, aún en presencia de valores perdidos, entre las cuales se encuentran:

- EM-AMMI: Gauch y Zobel (1990) desarrollaron un procedimiento iterativo que utiliza el algoritmo de maximización de la esperanza (EM, siglas en inglés de *Expectation Maximization*) incorporando el modelo AMMI. Para llevar a cabo este procedimiento, en primer lugar los parámetros aditivos del modelo AMMI se establecen mediante

la media general, la de los genotipos y de los ambientes obtenidas a partir de los datos observados. Los residuos de las celdas observadas se calculan como la media de dicha celda menos la media del genotipo menos la media del ambiente más la media general. Los  $K$  parámetros multiplicativos del modelo se obtienen al aplicar la DVS sobre la matriz de residuos, y los valores faltantes se completan con las estimaciones de AMMI. En iteraciones posteriores, el procedimiento habitual de AMMI se aplica a la matriz completa y los valores faltantes se actualizan mediante las estimaciones de AMMI correspondientes. Las iteraciones se detienen cuando el cambio entre los valores imputados para las celdas faltantes de dos pasos sucesivos sea pequeño, por ejemplo 0.01. El código de R de esta metodología fue implementado por Paderewski (2013).

- EM-SVD: Troyanskaya et al. (2001) propusieron un método de imputación que combina el algoritmo EM con la DVS. Este método reemplaza los valores faltantes de una matriz inicialmente por valores arbitrarios para obtener una matriz completa, y luego se calcula iterativamente la DVS de esta matriz. Al final del proceso, cuando las iteraciones alcanzan la estabilidad, se obtiene una matriz que contiene las imputaciones de los valores faltantes.
- EM-PCA: Josse y Husson (2016) propusieron imputar los valores faltantes de un conjunto de datos utilizando el algoritmo ACP iterativo. Este procedimiento consiste primero en imputar valores perdidos con valores iniciales como la media de la variable. El segundo paso es realizar un ACP en el conjunto de datos completo. Luego, imputa los valores faltantes con las fórmulas de reconstrucción de orden  $k$  (la matriz ajustada calculada con una cantidad de componentes “ $k$ ” para los *scores* y cargas). Estos pasos de estimación de los parámetros vía ACP e imputación de los valores faltantes usando la matriz ajustada se iteran hasta la convergencia.
- GabrielEigen fue propuesto por Arciniegas-Alarcón et al. (2010) y luego fue extendido a una versión ponderada llamada WGabriel (Arciniegas-Alarcón et al., 2014). El primero combina una regresión y aproximación de rango inferior utilizando la DVS. Este método reemplaza inicialmente las celdas faltantes por valores arbitrarios. Posteriormente las imputaciones se actualizan a través de un esquema iterativo que define una partición de la matriz para cada valor faltante y utiliza una regresión lineal de columnas (o filas) para obtener la nueva imputación. En esta regresión se aproxima la matriz de diseño por una matriz de menor rango utilizando la DVS. WGabriel es una modificación de GabrielEigen que usa pesos elegidos por validación cruzada para imputar los datos.

## 3.2. Creación de un paquete de R

Un paquete de R es un conjunto de funciones programadas en este lenguaje que comparten fines específicos y se distribuyen con un protocolo estandarizado, garantizando su correcto funcionamiento. Para la creación de un paquete se deben seguir ciertas convenciones referidas a la creación y almacenaje de carpetas y archivo con código de programación, documentación e instrucciones de sistema. La gestión de todos estos documentos puede ser manual, pero existen paquetes de R que asisten en la tarea del desarrollo de nuevos paquetes automatizando ciertas fases del proceso. En este trabajo se usaron los paquetes: *devtools* (Wickham et al., 2021), *usethis* (Wickham y Bryan, 2021), *roxygen2* (Wickham et al., 2020), *testthat* (Wickham, 2011) y *available* (Ganz et al., 2019).

Una vez finalizada esta etapa, dichas carpetas y archivos se compilan y comprimen para su distribución. Si esto se realiza con Windows como sistema operativo, se debe descargar e instalar el software Rtools disponible en CRAN.

Todo este proceso se realizó utilizando Git y GitHub. Git es un sistema de control de versiones, una herramienta que toma inicialmente una versión de un documento y luego registra los cambios que sufre el mismo a lo largo del tiempo. Esto facilita el trabajo colaborativo entre distintas personas ya que si más de una persona trabaja en el mismo documento, el sistema de control de versiones las puede integrar en una nueva. Git es más útil cuando se combina con GitHub. Este último es el servicio de *hosting* que se utiliza para que el proyecto tenga una presencia en la web permitiéndole a otras personas explorar los archivos, su historia, sincronizarse con la versión actual, proponer y realizar cambios, etc. Git y GitHub en conjunto forman el entorno más popular para los desarrolladores de paquetes de R ya que permite a cualquier persona descargar e instalar un paquete e incluso realizar aportes, detectar errores, incluir sugerencias, etc.

A continuación se detallan los distintos pasos que componen la creación de un paquete en R bajo un enfoque de trabajo reproducible, lo cual significa que los mismos pueden usarse de ejemplo para el desarrollo de nuevos paquetes o para imitar la creación del paquete *geneticae* que es objeto de desarrollo de este trabajo.

### 3.2.1. *geneticae*

En primer lugar se debe elegir el nombre del paquete cumpliendo con ciertas reglas: solo puede contener letras, números o puntos; debe tener al menos dos caracteres y empezar con



una letra y no terminar con un punto. Se debe chequear si el nombre elegido está disponible en los repositorios CRAN, Bioconductor y GitHub. Para ello, se utiliza la función `available()` del paquete *available*, que además indicará si el nombre elegido tiene algún significado especial que podemos desconocer (revisa las webs de *Wikipedia*, *Wiktionary* y *Urban Dictionary*). El nombre elegido en este caso fue “geneticae” (Figura 3.1).

```
> library(available)
> available("geneticae")
Urban Dictionary can contain potentially offensive results,
should they be included? [Y]es / [N]o:
1: Y
— geneticae —
Name valid: ✓
Available on CRAN: ✓
Available on Bioconductor: ✓
Available on GitHub: ✓
Abbreviations: http://www.abbreviations.com/geneticae
Wikipedia: https://en.wikipedia.org/wiki/geneticae
Wiktionary: https://en.wiktionary.org/wiki/geneticae
Urban Dictionary:
Not found.
Sentiment:???
```

Figura 3.1: Chequeo de disponibilidad del nombre *geneticae* elegido para el paquete en desarrollo mediante la función `available()` del paquete *available*.

### 3.2.2. Estructura general del paquete

Un paquete de R se construye creando y guardando diversos archivos y carpetas en un directorio cuyo nombre es igual al elegido en el paso anterior. Algunos elementos son de presencia obligatoria, entre ellos:

- Archivo DESCRIPTION: archivo de texto que describe el contenido del paquete, quiénes son sus desarrolladores, establece cómo se va a relacionar con otros, el tipo de licencia con el que se distribuye, los requisitos de sistema, etc.
- Carpeta R: contiene el o los archivos de código (*scripts*) de R con las funciones del paquete.
- Carpeta man: incluye archivos con la documentación del paquete, funciones y *datasets*.
- README: archivo de texto que brinda información sobre el proyecto.
- Archivo NAMESPACE: declara las funciones del paquete que se ponen a disposición de los usuarios y lista las funciones de otros paquetes de las cuales hace uso.

Existen otros elementos cuya inclusión es opcional, por ejemplo:

- Carpeta data: contiene objetos de R con conjuntos de datos.

- Carpeta vignettes: contiene los tutoriales que muestran ejemplos de uso del paquete, generalmente escritos en Rmarkdown, un formato de escritura que facilita la presentación de texto entrelazado con código y resultados.
- Carpeta tests: incluye código que permiten someter al paquete a diversos controles.

Si bien estas carpetas y archivos pueden crearse en forma manual, es conveniente utilizar la función `create_package()` del paquete *usethis* que se encarga de generar automáticamente un directorio con todas las componentes requeridas para el desarrollo del paquete que sirve de plantilla para facilitar la parte inicial de este proceso (Figura 3.2).

```
> # Crear el paquete geneticae
> create_package("/home/julia-fedora/Escritorio/geneticae")
✓ Creating '/home/julia-fedora/Escritorio/geneticae/'
✓ Setting active project to '/home/julia-fedora/Escritorio/geneticae'
✓ Creating 'R/'
✓ Writing 'DESCRIPTION'
Package: geneticae
Title: What the Package Does (One Line, Title Case)
Version: 0.0.0.9000
Authors@R (parsed):
  * First Last <first.last@example.com> [aut, cre] (YOUR-ORCID-ID)
Description: What the package does (one paragraph).
License: 'use_mit_license()', 'use_gpl3_license()' or friends to
  pick a license
Encoding: UTF-8
LazyData: true
Roxygen: list(markdown = TRUE)
RoxygenNote: 7.1.1
✓ Writing 'NAMESPACE'
✓ Writing 'geneticae.Rproj'
✓ Adding '.Rproj.user' to '.gitignore'
✓ Adding '^geneticae\\.Rproj$', '^\\.Rproj\\.user$' to '.Rbuildignore'
✓ Opening '/home/julia-fedora/Escritorio/geneticae/' in new RStudio session
✓ Setting active project to '<no active project>'
> |
```

Figura 3.2: Creación del paquete *geneticae* mediante la función `create_package()` del paquete *usethis*.

### 3.2.3. Archivo DESCRIPTION

El archivo DESCRIPTION provee toda la metadata sobre el paquete presentada a través de campos, algunos de los cuales tienen que estar de forma obligatoria y otros son opcionales. Los campos obligatorios son:

- Package: nombre del paquete.
- Title: título del paquete (hasta 65 caracteres).
- Version: número de la versión actual del paquete, en este caso 0.1.9000.
- Author, Maintainer o Authors@R: autores, contribuyentes y personas a cargo del mantenimiento del paquete.
- Description: descripción del paquete.

- **License:** nombre de la licencia bajo la cual se distribuye el paquete. Si se pretende que cualquiera lo puede usar, entonces se debe recurrir a los tipos más comunes de licencia para código abierto: CC0, MIT o GPL.

Entre los campos opcionales los más importantes son:

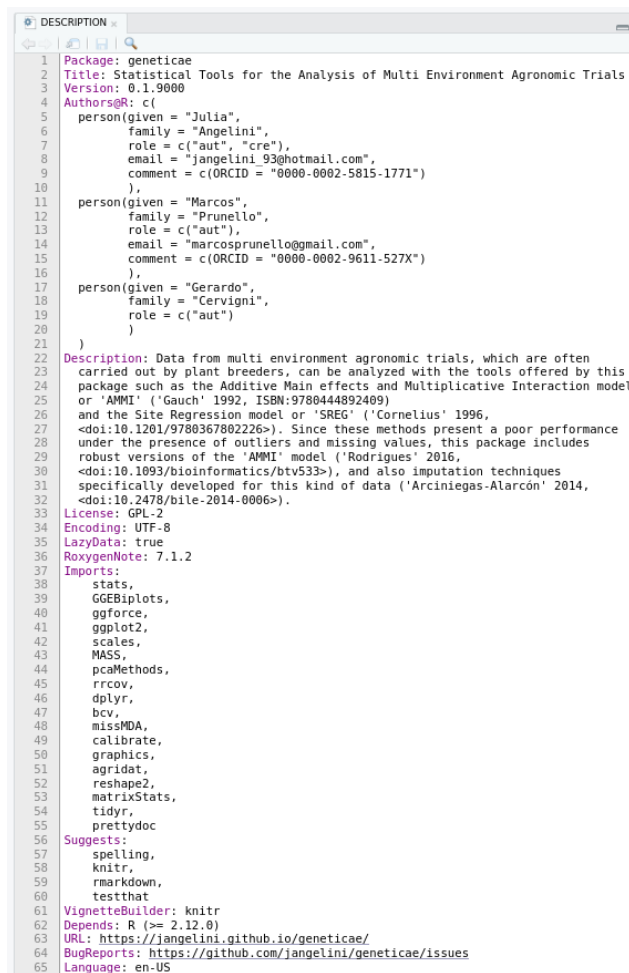
- **Imports:** en el caso de que el código desarrollado haga uso de funciones pertenecientes a otros paquetes, los mismos deben ser listados en este campo.
- **Suggests:** listado de paquetes que no son imprescindibles para el uso del nuevo paquete, pero que pueden ser útiles como herramientas secundarias para el mismo (por ejemplo, para seguir los tutoriales).
- **Depends:** mínima versión de R con la cual el paquete es compatible.
- **URL:** dirección de la página web del paquete.
- **BugReports:** dirección donde los usuarios pueden enviar avisos sobre los problemas que encuentren al utilizar el paquete.

El archivo DESCRIPTION del paquete *geneticae* se muestra en la Figura 3.3.

### 3.2.4. Archivos de código

Una vez creada la estructura del paquete y el archivo DESCRIPTION se deben programar las funciones que el mismo contendrá. Estas deben ser guardadas en *scripts* con extensión .R, en el subdirectorio R/. Los *scripts* pueden contener código para una o más funciones y ser guardadas con cualquier nombre, aunque es recomendable que el mismo esté relacionado con su contenido. La Figura 3.4 muestra un fragmento de la función `GGEmodel()` del paquete *geneticae*.

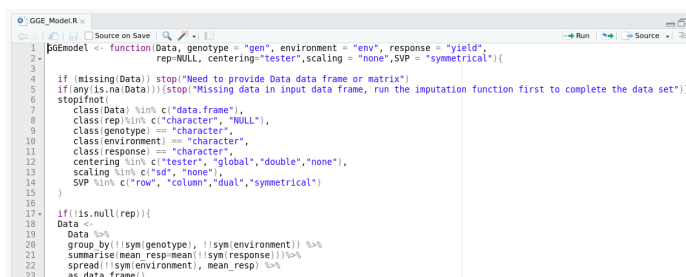
A medida que se desarrolla el paquete se van generando relaciones complejas entre las funciones programadas. Algunas de ellas son de uso interno (son invocadas por otras funciones del paquete para cumplir con alguna tarea específica) y otras son diseñadas para que estén disponibles para los usuarios (es decir, se “exportan”). Además algunas funciones invocan a otras pertenecientes a paquetes escritos por terceros. Esto hace necesario realizar pruebas del funcionamiento del código producido durante todo el proceso de desarrollo para garantizar que realice lo que realmente se desea y para corregir errores en la programación (depuración o *debugging*). Para simular el proceso de construcción, instalación y carga del paquete durante su desarrollo se utiliza la función `load_all()` de *devtools* que permite acceder a las funciones del paquete para su evaluación.



```

1 Package: geneticae
2 Title: Statistical Tools for the Analysis of Multi Environment Agronomic Trials
3 Version: 0.1.9000
4 Authors@R: c(
5   person(given = "Julia",
6     family = "Angelini",
7     role = c("aut", "cre"),
8     email = "jangelini_93@hotmail.com",
9     comment = c(ORCID = "0000-0002-5815-1771")
10  ),
11   person(given = "Marcos",
12     family = "Prunello",
13     role = c("aut"),
14     email = "marcosprunello@gmail.com",
15     comment = c(ORCID = "0000-0002-9611-527X")
16  ),
17   person(given = "Gerardo",
18     family = "Cervigni",
19     role = c("aut")
20  )
21 )
22 Description: Data from multi environment agronomic trials, which are often
23 carried out by plant breeders, can be analyzed with the tools offered by this
24 package such as the Additive Main effects and Multiplicative Interaction model
25 or 'AMMI' ('Gauch' 1992, ISBN:9780444892409)
26 and the Site Regression model or 'SREG' ('Cornelius' 1996,
27 <doi:10.1201/9780367802226>). Since these methods present a poor performance
28 under the presence of outliers and missing values, this package includes
29 robust versions of the 'AMMI' model ('Rodrigues' 2016,
30 <doi:10.1093/bioinformatics/btv533>), and also imputation techniques
31 specifically developed for this kind of data ('Arciniegas-Alarcón' 2014,
32 <doi:10.2478/bile-2014-0006>).
33 License: GPL-2
34 Encoding: UTF-8
35 LazyData: true
36 RoxygenNote: 7.1.2
37 Imports:
38   stats,
39   GGEbiplots,
40   ggforce,
41   ggplot2,
42   scales,
43   MASS,
44   pcaMethods,
45   rrcov,
46   dplyr,
47   bcv,
48   missMDA,
49   calibrate,
50   graphics,
51   agridat,
52   reshape2,
53   matrixStats,
54   tidyrr,
55   prettydoc
56 Suggests:
57   spelling,
58   knitr,
59   rmarkdown,
60   testthat
61 VignetteBuilder: knitr
62 Depends: R (>= 2.12.0)
63 URL: https://jangelini.github.io/geneticae/
64 BugReports: https://github.com/jangelini/geneticae/issues
65 Language: en-US

```

Figura 3.3: Archivo DESCRIPTION de *geneticae*.


```

1 GGEmodel <- function(Data, genotype = "gen", environment = "env", response = "yield",
2   rep = NULL, centering = "tester", scaling = "none", SVP = "symmetrical"){
3
4   if (missing(Data)) stop("Need to provide Data data frame or matrix")
5   if (any(is.na(Data))) stop("Missing data in input data frame, run the imputation function first to complete the data set")
6   stopifnot(
7     class(Data) %in% c("data.frame"),
8     class(rep) %in% c("character", "NULL"),
9     class(genotype) == "character",
10    class(environment) == "character",
11    class(response) == "character",
12    centering %in% c("tester", "global", "double", "none"),
13    scaling %in% c("sd", "none"),
14    SVP %in% c("row", "column", "dual", "symmetrical")
15  )
16
17  if (is.null(rep)){
18    Data <-
19      Data %>%
20      group_by(!sym(genotype), !sym(environment)) %>%
21      summarise(mean_resp = mean(!sym(response))) %>%
22      spread(!sym(environment), mean_resp) %>%
23      as.data.frame()
24  }
25

```

Figura 3.4: Fragmento de la función *GGEmodel()* del paquete *geneticae*.

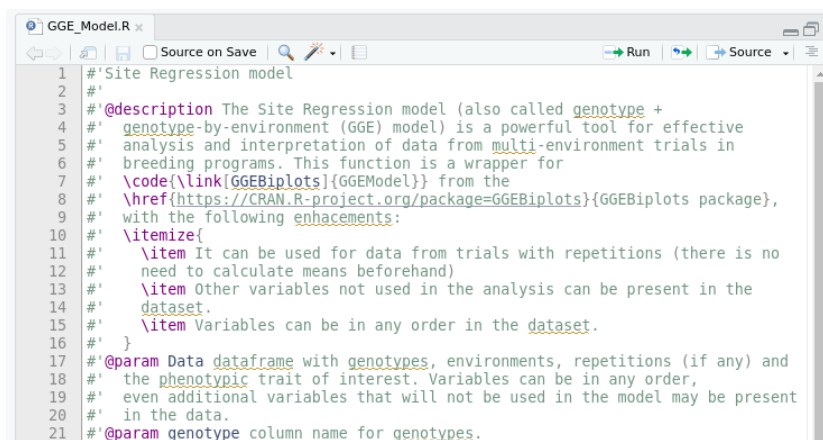
### 3.2.5. Documentación

Uno de los aspectos más importantes de la creación de un paquete es la documentación donde se describe cómo se usa cada función y para qué sirven los argumentos, se aclara qué tipo de resultado devuelve, se proveen ejemplos para el uso, etc. El paquete *roxygen2* provee pautas para incluir todo esto escribiendo comentarios con un formato especial antes de la definición de la función en el mismo archivo de código.

El flujo de trabajo para crear la documentación con el paquete *roxygen2* es el siguiente:

1. Agregar comentarios a los archivos .R. Estos deben comenzar con `#'`, para distinguirlo de los comentarios regulares y preceden a cada función. La primera línea es el título y el párrafo que le sigue es su descripción. Para el resto de los campos de la documentación, se utilizan etiquetas listadas línea tras línea que comienzan con `@`, siendo las más importantes a incluir:
  - `@param`: detalla para qué sirve cada parámetro de la función, qué tipo de objeto es y qué valor toma por defecto (opcional).
  - `@return`: explica qué objeto devuelve la función.
  - `@details`: agrega cualquier aclaración que se considere necesaria.
  - `@examples`: incluye ejemplos de uso de la función.
  - `@export`: indica que la función tiene que estar disponible cuando alguien cargue el paquete con `library()`.
  - `@references`: inserta referencias bibliográficas.
2. Ejecutar `document()` del paquete *devtools* para convertir los comentarios escritos en formato roxygen en los archivos que compondrán el manual de ayuda y que deben ir guardados en la carpeta `man`. Además, esto se encarga de generar el archivo `NAMESPACE`, que tiene como objetivo declarar cuáles son las funciones del nuevo paquete que son para exportar, así como también listar todas las funciones importadas de otros paquetes.

La Figura 3.5 muestra un fragmento de los comentarios roxygen de la función `GGEmodel()` del paquete *geneticae*.



```

1 #' Site Regression model
2 #'
3 #' @description The Site Regression model (also called genotype +
4 #' genotype-by-environment (GGE) model) is a powerful tool for effective
5 #' analysis and interpretation of data from multi-environment trials in
6 #' breeding programs. This function is a wrapper for
7 #' \code{\link[GGEbiplots]{GGEModel}} from the
8 #' \href{https://CRAN.R-project.org/package=GGEbiplots}{GGEbiplots} package,
9 #' with the following enhancements:
10 #'
11 #' \itemize{
12 #'   \item It can be used for data from trials with repetitions (there is no
13 #'     need to calculate means beforehand)
14 #'   \item Other variables not used in the analysis can be present in the
15 #'     dataset.
16 #'   \item Variables can be in any order in the dataset.
17 #' }
18 #' @param Data dataframe with genotypes, environments, repetitions (if any) and
19 #' the phenotypic trait of interest. Variables can be in any order,
20 #' even additional variables that will not be used in the model may be present
21 #' in the data.
22 #' @param genotype column name for genotypes.

```

Figura 3.5: Fragmento de los comentarios roxygen de la función `GGEModel()` del paquete *geneticae*.

### 3.2.6. Uso de funciones de otros paquetes

Cuando el código desarrollado invoca a funciones de otros paquetes, los mismos deben ser listados en el campo Imports del archivo DESCRIPTION como se mencionó en la sección 3.2.3. Esto puede hacerse de manera manual o mediante la función `use_package()` del paquete *devtools* indicando el nombre del paquete.

En el código se debe hacer uso de tales funciones anteponiéndolo a su nombre el del paquete y el operador `::`. Por ejemplo, `dplyr::group_by()` invoca la función `group_by()` del paquete *dplyr*. En el caso de que alguna función se aplicara con mucha frecuencia, se puede prescindir del comando anterior si se agrega el nombre de la función y el paquete de procedencia en la etiqueta `@importFrom` en los comentarios roxygen. Para el ejemplo anterior se debería indicar `@importFrom dplyr group_by`. Esto permite mencionar a la función en el código sin el operador `::`.

Si se utilizan repetidamente muchas funciones de otro paquete, es posible importarlas todas indicando el nombre del mismo en la etiqueta `@import` en los comentarios roxygen. Sin embargo, esta es la solución menos recomendada porque hace que el código sea más difícil de leer y aumenta la posibilidad de que entren en conflicto nombres de funciones pertenecientes a diversos paquetes.

### 3.2.7. Testeos

Probar el código desarrollado, sometiéndolo a casos particulares y a distintos ejemplos, es fundamental en la creación de paquetes ya que permite detectar y corregir errores y asegurarse que el código haga lo que realmente se desea.

Como se mencionó anteriormente, esto puede hacerse de manera dinámica e interactiva durante el desarrollo al instalar y cargar el paquete en gestación con `load_all()`. Sin embargo, si siempre se realizan los mismos controles es posible automatizar este proceso. Para ello, se generan unidades de testeo que ponen a prueba el código corriendo parte del mismo bajo distintas circunstancias y comparando el resultado obtenido con el esperado por el desarrollador. La función `use_test()` del paquete *testthat*, que agrega “testthat” al campo Suggest del archivo DESCRIPTION, crea un directorio `test/testthat` para ubicar los códigos con los testeos y un archivo `testthat.R` que se encarga de la ejecución de los mismos. Una vez escritos estos archivos que establecen cuales son los controles a realizar automáticamente, se pueden evaluar los resultados con la función `test()` del paquete *devtools*. Ante cada error encontrado, se debe corregir el código y repetir este proceso hasta que todas las unidades de testeo pasen la prueba. En la Figura 3.6 se muestra el resultado de evaluar las unidades de testeo creadas para el paquete *geneticae*.

```
> test()
i Loading genetiae
i Testing genetiae
✓ | OK F W S | Context
✓ | 6         | GGE_Model [0.2 s]
✓ | 7         | GGE_Plot  [0.8 s]
✓ | 4         | impute    [0.2 s]
✓ | 3         | r_AMMI    [0.2 s]

== Results ==
Duration: 1.4 s

[ FAIL 0 | WARN 0 | SKIP 0 | PASS 20 ]
```

Figura 3.6: Resultado de correr la función `test()` del paquete *devtools* en *geneticae*.

Una medida de la calidad de un paquete está dada por el porcentaje de código que es evaluado durante los testeos. Esta se puede obtener mediante la función `test_coverage()` del paquete *covr* (Hester, 2020). El paquete *geneticae* tiene un porcentaje total de cobertura de los test igual a 24.75 % (Figura 3.7).

## geneticae coverage - 24.75%

Files	Source						
	File	Lines	Relevant	Covered	Missed	Hits / Line	Coverage
	R/W_GabrielEigen.R	270	131	0	131	0	0.00%
	R/GabrielEigen.R	136	63	0	63	0	0.00%
	R/rAMMI.R	235	126	2	124	0	1.59%
	R/GGE_Plot.R	465	238	57	181	1	23.95%
	R/impute.R	186	49	37	12	1	75.51%
	R/GGE_Model.R	119	27	21	6	4	77.78%
	R/EM_AMMI.R	160	59	49	10	6	83.05%

Figura 3.7: Porcentaje del código de *geneticae* que es evaluado durante los testeos obtenido mediante la función `test.coverage()` del paquete *covr*.

### 3.2.8. Datasets

A menudo es útil incluir conjuntos de datos en un paquete a fin de proporcionar ejemplos de uso de las funciones incluidas. Los *datasets* que se deseen añadir a un paquete deben ser guardados como archivos `.RData` en el directorio `data/`. La función `use.data()` del paquete *usethis* puede ser empleada para automatizar este proceso.

Los objetos de la carpeta `data` siempre se exportan, por lo cual se debe agregar documentación para los mismos. Esto se puede incluir en cualquier archivo de código `.R` dentro del directorio `R/`.

### 3.2.9. Archivo README

Un README es un archivo de texto plano que se utiliza para documentar o brindar información sobre alguna pieza de software o proyecto. Si un directorio contiene un archivo README, se espera que el usuario lo lea antes de explorar el resto del contenido. En el contexto de creación de paquetes de R, el README se suele escribir en RMarkdown y describe brevemente por qué y para qué alguien podría usar el paquete, junto con menciones para su instalación y un ejemplo introductorio. Además, este archivo se muestra en la página del paquete cuando este es publicado en plataformas como GitHub.

Para generar el README se utiliza la función `use.readme.rmd()` del paquete *usethis* que crea un archivo de Rmarkdown con una plantilla donde se completa su contenido.



En el archivo README se suelen incluir insignias (“bagdes”) y el logo del paquete. Algunas funciones del paquete *usethis* permiten agregar las insignias al README. Por otro lado, muchos de los paquetes de R disponen de un logo con forma hexagonal conocido como hexSticker. Esto permite darle identidad al paquete. El logo se puede crear con ayuda del paquete *HexSticker* y ser añadido al README con la función `use_logo()`. La Figura 3.8 presenta un fragmento del archivo README mostrado en el repositorio GitHub del paquete *geneticae*.

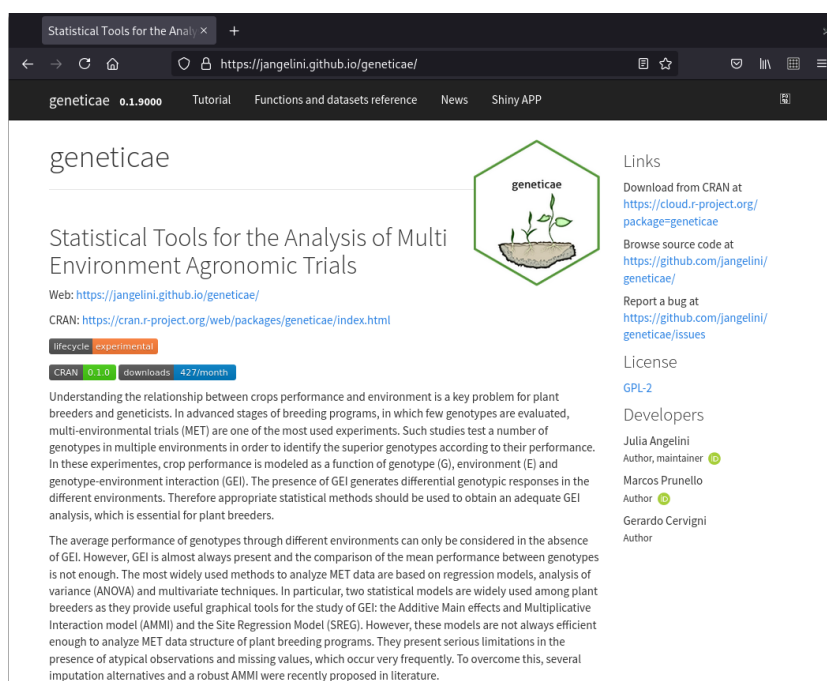


Figura 3.8: Fragmento del archivo README mostrado en el repositorio GitHub del paquete *geneticae*.

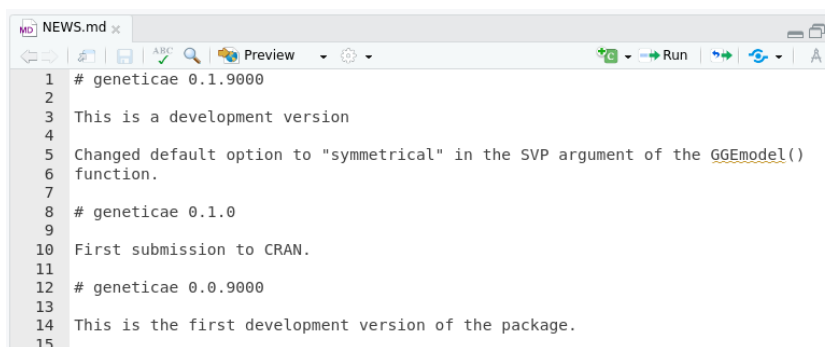
### 3.2.10. Archivo NEWS

El archivo NEWS se encarga de contar los cambios presentes en cada versión nueva del paquete que se publica. Mientras que el README apunta a ser leído por nuevos usuarios, el archivo NEWS es para aquellos que ya usan el paquete.

Se sugiere usar RMarkdown para escribir este archivo y colocar un título principal para cada versión, seguido por títulos secundarios que describen lo realizado (cambios principales, errores arreglados, etc.). Si se trata de cambios impulsados por otras personas, por ejemplo,

a través de sugerencias hechas en GitHub, se los menciona para darles mérito. Una buena práctica es ir escribiendo este archivo cada vez que se realiza algo nuevo en el paquete. La función que permite crear este archivo automáticamente es `use_news_md()` del paquete *usethis*.

La Figura 3.9 presenta el contenido del archivo NEWS del paquete *geneticae*.



```

1 # geneticae 0.1.9000
2
3 This is a development version
4
5 Changed default option to "symmetrical" in the SVP argument of the GGEmodel()
6 function.
7
8 # geneticae 0.1.0
9
10 First submission to CRAN.
11
12 # geneticae 0.0.9000
13
14 This is the first development version of the package.
15

```

Figura 3.9: Archivo NEWS de *geneticae*.

### 3.2.11. Viñetas

Una viñeta es un tipo especial de documentación que puede agregarse al paquete para dar más detalles y ejemplos sobre el uso del mismo. En ella se brinda una descripción del problema que el paquete está diseñado para resolver y se muestra al lector cómo resolverlo. Se diferencian de las páginas de ayuda en que su adición es opcional y no sigue una estructura fija, dándole la libertad al autor de enseñar de la forma que más le guste cómo usar su paquete.

Muchos de los paquetes existentes tienen viñetas a las que se puede acceder utilizando la función `browseVignettes("nombre del paquete")` si el mismo se encuentra instalado o consultando en su página de CRAN. Por ejemplo para el paquete *geneticae*: <https://cran.r-project.org/web/packages/geneticae/vignettes/a-tutorial.html>.

Generalmente, las viñetas son generadas en RMarkdown. Para crear viñetas se emplea la función `use_vignette("nombre del paquete")` del paquete *usethis* que crea un directorio `vignettes/`, agrega las dependencias necesarias a `DESCRIPTION` y genera una plantilla en RMarkdown para redactar la viñeta.

### 3.2.12. R CMD check e instalación

Además de los controles interactivos o automatizados que los desarrolladores realicen, existe un riguroso proceso de control conocido como R CMD check que debe ser superado sin errores, advertencias ni ningún tipo de nota si se desea publicar el paquete en algún repositorio oficial como CRAN. R CMD check esta compuesto por más de 50 chequeos individuales entre los cuales se encuentran: la estructura del paquete, el archivo DESCRIPTION, NAMESPACE, el código de R, los datos, la documentación, entre otros.

Se aconseja realizar verificaciones completas de que todo funciona a medida que se van incorporando funciones para detectar y solucionar problemas de forma temprana. Una vez que se desarrollaron todos los elementos necesarios para el paquete y no se detectan errores, advertencias o notas, se ejecuta la función `install()` del paquete *devtools*, con el objetivo de instalar el paquete en la biblioteca.

### 3.2.13. Publicación y difusión

Por último para que otros usuarios puedan utilizar el paquete desarrollado es necesario publicarlo en algún sitio del cual pueda ser descargado e instalado. Esto se logra subiéndolo a un proyecto público en GitHub o enviándolo a repositorios oficiales como CRAN. Para ser aceptado en CRAN, el paquete además de atravesar con éxito los rigurosos controles de R CMD check debe superar una serie de políticas que son comprobadas manualmente por revisores.

Con el fin de favorecer a la difusión del paquete una vez que el mismo es finalizado, es conveniente publicarlo en una página web. Esta es una tarea relativamente sencilla gracias a dos factores. En primer lugar, la función `build_site()` del paquete *pkgdown* (Wickham y Hesselberth, 2020) toma todo el material creado para el paquete (documentación, README, tutoriales, NEWS, etc.) y crea automáticamente en un sitio web que puede ser posteriormente personalizado. En segundo lugar, GitHub ofrece un servicio gratuito de *web-hosting* que posibilita la publicación del sitio en internet.

## 3.3. Aplicación web Shiny

El paquete *shiny* (Chang et al., 2020) de R permite construir aplicaciones web directamente desde RStudio sin necesidad de conocer en profundidad los lenguajes HTML / CSS

/ JavaScript. Estas aplicaciones constituyen una interfaz gráfica entre el usuario y R, que permiten realizar diversos análisis a través de un navegador web sin necesidad de programar.

### 3.3.1. Programación de una aplicación web Shiny

El desarrollo de una aplicación web Shiny consiste en programar en lenguaje R dos componentes, la interfaz de usuario y las evaluaciones a ejecutar por el servidor que aloje a la aplicación.

La interfaz de usuario (conocida como ui, siglas en inglés de *user interface*) controla el diseño de la aplicación y se encarga de establecer cuáles son los argumentos de entrada (*inputs*) que el usuario provee al hacer uso de la aplicación y las salidas (*outputs*) que el servidor debe mostrar luego de procesar los *inputs*. En general, definir las características de la interfaz de usuario puede no resultar tan sencillo ya que muchas de sus herramientas están vinculadas a otros lenguajes de programación, por ejemplo HTML, CSS o JavaScript. Sin embargo, las funciones del paquete *shiny*, junto con las de otros paquetes auxiliares como *shinyWidgets* (Perrier et al., 2020) *shinythemes* (Chang, 2018) y *shinyhelper* (Mason-Thom, 2019), facilitan la tarea, al proveer código en lenguaje R para administrar el contenido y apariencia de la página web.

El segundo elemento a programar recibe el nombre de *server* e incluye el código de R que le indica a la aplicación qué debe hacer y cómo debe funcionar, incluyendo la lectura y manipulación de datos, el armado de gráficos, el ajuste de modelos, etc. En la programación del *server* se debe prestar especial atención a los *inputs* (datos u opciones elegidas por el usuario a través de la ui) y *outputs* (resultados, tablas, gráficos, mapas, etc.), que se almacenan como objetos de R de tipo lista.

El código para la generación de una aplicación web Shiny debe finalizar con una llamada a la función `shinyApp()` del paquete *shiny* que se encarga de la ejecución y lanzamiento de la aplicación.

Existen diversas formas para que un usuario pueda utilizar una aplicación Shiny:

- En formal local, desde RStudio. Si se dispone del código fuente de forma local, se puede lanzar a correr la aplicación con la función `shinyApp()`. Si el código está disponible en un repositorio público de GitHub, se lo puede correr con la función `runGitHub()` del paquete *shiny*. En ambos casos, se habilita el uso de la aplicación dentro del mismo RStudio o en algún navegador web como *Google Chrome*.

- En forma remota, accediendo a algún servidor donde la aplicación esté alojada. En este caso, se debe contar con conexión a internet para acceder a la web del servidor y poder hacer uso de la aplicación online.

---

# Capítulo 4

## Resultados

### 4.1. Paquete de R *geneticae*

Siguiendo el flujo de trabajo descrito en el capítulo anterior fue posible desarrollar el paquete de R *geneticae* que implementa los métodos estadísticos discutidos para el análisis de datos de EMA.

El paquete *geneticae* fue enviado a CRAN, donde pasó los exhaustivos controles de R CMD check exitosamente, logrando así una rápida aceptación (<https://cran.r-project.org/web/packages/geneticae/index.html>). En el momento de la escritura de este informe, pasaron 3 semanas desde su publicación en el repositorio y cuenta con más de 400 descargas, a pesar de que aún no se ha hecho difusión del paquete. El código fuente se encuentra disponible en GitHub (<https://github.com/jangelini/geneticae>).

Para instalar la versión del paquete publicada en CRAN: `install.packages("geneticae")`, mientras que la versión en desarrollo se debe instalar desde el repositorio de GitHub: `devtools::install_github("jangelini/geneticae")`. Una vez instalado el paquete, se debe cargar en la sesión de R mediante el comando: `library(geneticae)`.

Información detallada sobre las funciones del paquete *geneticae* se puede obtener mediante `help(package = "geneticae")`. La ayuda para una función, por ejemplo `imputation()`, en una sesión R se puede obtener usando `?imputation` o `help(imputation)`. La función `browseVignettes("geneticae")` permite obtener la viñeta del paquete, es decir una descripción del problema que está diseñado para resolver así como ejemplos de aplicación del mismo.

Además, se encuentra disponible una página web que contiene una breve descripción de

la utilidad del paquete, las funciones que se incluyen en él, un tutorial de uso, un enlace de acceso a la aplicación web Shiny, entre otros elementos (<https://jangelini.github.io/geneticae/>).

#### 4.1.1. Conjuntos de datos en *geneticae*

La versión actual (0.3.0) del paquete *geneticae* contiene el conjunto de datos *plrv* (de Mendiburu, 2021) usado para ejemplificar la metodología incluida para analizar los datos provenientes de EMA. Éste contiene información sobre el rendimiento, el peso de planta y de la parcela de 28 genotipos en 6 localidades de Perú. Cada clon fue evaluado tres veces en cada ambiente.

```
data(plrv)
head(plrv)[1:2,]
```

##	Genotype	Locality	Rep	WeightPlant	WeightPlot	Yield
## 1	102.18	Ayac	1	0.5100000	5.10	18.88889
## 2	104.22	Ayac	1	0.3450000	2.76	12.77778

Las versiones anteriores a la mencionada contaban con el conjunto de datos *yan.winterwheat* del paquete *agridat* (Wright, 2020), el cual ha sido ampliamente utilizado para ejemplificar el análisis de los datos provenientes de EMA (Arciniegas-Alarcón et al., 2014, 2020; Yan, 2013), así como también para ilustrar el uso de un software interactivo no comercial *GGEbiplotGUI* (Frutos et al., 2013). Dada la amplia difusión del mismo en el contexto de datos EMA, fue utilizado para mostrar las funciones del paquete y de la aplicación web Shiny desarrollados en este trabajo. Sin embargo, el autor de *agridat* solicitó la eliminación de la dependencia en *geneticae* con el fin que se le simplifique el envío de nuevas versiones de su paquete a CRAN. Por lo que, a partir de la versión 0.3.0, para reproducir los análisis presentados en éste informe y en los tutoriales primero se debe instalar *agridat* y cargar en la sesión actual de R.

#### 4.1.2. Uso del paquete para ajustar el modelo AMMI

Para visualizar el efecto de IGA se utiliza el biplot GE obtenido del modelo AMMI a través de la función `rAMMI()`, que requiere datos en formato largo, es decir, cada fila corresponde a una observación y cada columna a una variable (genotipo, ambiente, fenotipo observado y,

si existe, repetición). Si cada genotipo ha sido evaluado más de una vez en cada ambiente, la media fenotípica para cada combinación de genotipo y ambiente se calcula internamente y luego se estima el modelo. El conjunto de datos puede contener variables adicionales no utilizadas en el análisis (a diferencia de lo que ocurre con muchos softwares y funciones de R disponibles hasta el momento). No se permiten valores perdidos pero se pueden imputar como se indica en la subsección 4.1.4.

El primer argumento de `rAMMI()` es el conjunto de datos de entrada, luego se indican los nombres de las columnas en las cuales se encuentra la información necesaria para aplicar la técnica y por último el biplot que se desea obtener que por defecto es el derivado del modelo AMMI clásico. Opcionalmente, se puede agregar el porcentaje de IGA explicado por el biplot como una nota al pie mediante el argumento *footnote* = *T* y un título con *titles* = *T*.

El biplot clásico para el conjunto de datos *yan.winterwheat* se muestra en la Figura 4.1. En este ejemplo BH93, KE93 y OA93 son los ambientes que más contribuyen a la interacción ya que sus vectores son los de mayor magnitud. Los cultivares m12 y Kat presentan patrones de interacción similares (sus identificadores están próximos entre sí) y son muy diferentes de Ann y Aug, por ejemplo. La cercanía entre el cultivar Dia y el ambiente BH93 indica una fuerte asociación positiva entre ellos, lo que significa que BH93 es un ambiente extremadamente favorable para ese genotipo. Como los identificadores de OA93 y Luc son opuestos, este ambiente es considerablemente desfavorable para ese genotipo. Por último, Cas y Reb están cerca del origen, lo que significa que se adaptan en igual medida a todos los ambientes.





### 4.1.3. Uso del paquete para ajustar el modelo SREG

Con el fin de visualizar conjuntamente el efecto de G e IGA Yan et al. (2000) propusieron el biplot GGE mediante el cual se pueden abordar diversos aspectos relacionados con la evaluación de genotipos y ambientes. Para obtener dicho biplot en primer lugar se debe ajustar el modelo SREG mediante `GGEmodel()` que, del mismo modo que `rAMMI()`, requiere datos en formato largo. Si bien esta función invoca a `GGEmodel()` del paquete *GGEbiplots* (Dumble, 2022), al ser utilizada mediante *geneticae* se permiten repeticiones y variables adicionales en el conjunto de datos. El rasgo fenotípico para cada combinación de genotipo y ambiente debe estar registrado, sino se debe recurrir previamente a alguna técnica de imputación para completar los datos (subsección 4.1.4).

Se presenta a continuación la sentencia utilizada para ajustar el modelo SREG para el conjunto de datos *yan.winterwheat*.

```
GGE1 <- GGEmodel(yan.winterwheat, genotype = "gen", environment = "env",
  response = "yield", rep = NULL, centering = "tester", scaling = "none",
  SVP = "symmetrical")
```

El primer argumento es el nombre del conjunto de datos y en los siguientes se indican los nombres de las columnas que contienen la información de los genotipos, ambientes y del rasgo fenotípico de interés. La función considera que no hay réplicas en el conjunto de datos, sin embargo, si existieran en el parámetro *rep* se debe indicar el nombre de la columna con dicha información. Otros argumentos son el método de centrado, de partición de los valores singulares (SVP, siglas en inglés de *Singular Value Partition*) y escalado. Por defecto los datos se centran utilizando la opción *centering="tester"* lo cual resulta en el modelo SREG. La elección del método de SVP no altera las relaciones o interacciones relativas entre los genotipos y los ambientes, aunque la apariencia del biplot será diferente (Yan, 2002). El método de partición de los valores singulares centrado en los genotipos (*SVP="row"*) muestra la interrelación entre genotipos con mayor precisión, el enfocado a los ambientes (*SVP="column"*) es el más informativo de las interrelaciones entre los ambientes, mientras que el simétrico (*SVP="symmetrical"*) permite visualizar la magnitud relativa tanto de la variación de los genotipos como de los ambientes, por lo que se utiliza por defecto. Por último, se indica que los datos no se deben escalar con el parámetro *scaling="none"*.

La salida de `GGEmodel()` es una lista con los siguientes objetos:

- *coordgenotype*: coordenadas para los genotipos en cada componente.

- *coordenviroment*: coordenadas para los ambientes en cada componente.
- *eigenvalues*: vector de autovalores para cada componente.
- *vartotal*: variancia general.
- *varexpl*: porcentaje de variancia explicado por cada componente.
- *labelgen*: nombres de los genotipos.
- *labelenv*: nombres de los ambientes.
- *axes*: etiquetas de los ejes.
- *Data*: datos escalados y centrados.
- *centering*: método de centrado.
- *scaling*: método de escala.
- *SVP*: método de partición de los valores singulares.

Utilizando la salida de `GGEmodel()`, la función `GGEPlot()` crea numerosas vistas del biplot GGE que permiten dar respuesta a distintos objetivos de los fitomejoradores. En estos gráficos los cultivares se muestran en minúsculas y los ambientes en mayúsculas. El método de centrado, escalado y SVP se muestran en una nota al pie junto con el porcentaje de G e IGA explicado por los dos ejes al agregar el argumento *footnote* = *T* y un título con *titles* = *T*.

### Comparaciones simples utilizando GGE biplot

El biplot básico se obtiene con el parámetro *type* = “*Biplot*” (Figura 4.2). En este ejemplo, el 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los ángulos entre los identificadores de genotipos y entre los vectores ambientales son utilizados para interpretar el gráfico. Así, por ejemplo, Kat tiene un rendimiento por debajo de la media en todos los ambientes debido a su ángulo superior a 90° con todos ellos. Por otro lado, Fun presenta un rendimiento superior a la media en todas las localidades excepto OA93 y KE93, como lo indican los ángulos agudos. La longitud de los vectores ambientales es una medida de la capacidad del ambiente para discriminar entre cultivos.

GGEPlot(GGE1, type = "Biplot", footnote = F, titles = F)

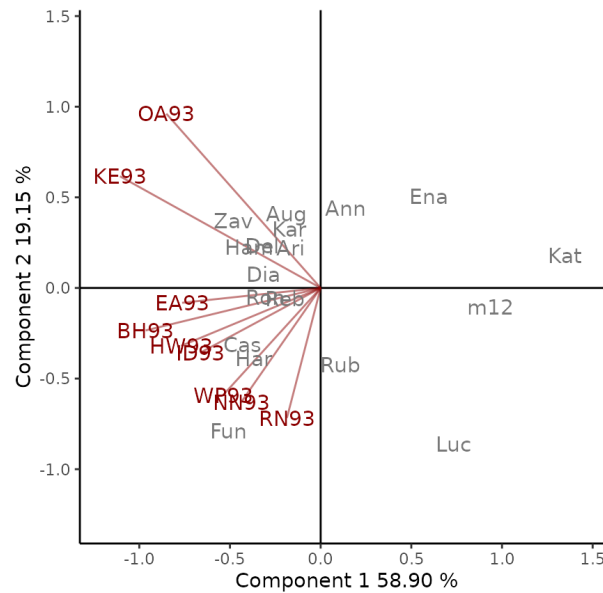


Figura 4.2: Biplot GGE basado en datos de rendimiento de trigo de invierno obtenido de Ontario en 1993. El método de partición de los valores singulares utilizado es el simétrico (opción por defecto). El 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en minúsculas y los entornos en mayúsculas.

Con frecuencia, los mejoradores necesitan identificar los cultivares más adaptados a un ambiente particular, por ejemplo OA93. Para esto Yan y Kang (2003) sugieren construir un eje del ambiente de interés (OA93), trazando una recta que una el identificador del ambiente y el origen de coordenadas, y lo denominan eje OA93. Los genotipos se clasifican en función del rendimiento en dicho ambiente de acuerdo con sus proyecciones, en la dirección indicada por el eje OA93 (Figura 4.3 (A)). Para obtener esta vista del biplot GGE, se indica la opción *Selected Environment* en el argumento *type* de la función y el ambiente a evaluar en el argumento *selectedE*. En este ejemplo, el cultivar de mayor rendimiento fue es Zav seguido por Aug, Ham hasta llegar al genotipo Luc, que es el de menor rendimiento en ese ambiente. El eje perpendicular al del ambiente de interés separa los genotipos con rendimiento mayor al promedio: de Zav a Cas, de aquellos con valores inferior a la media, de Ema a Luc, en OA93.

En forma similar, es posible determinar el ambiente más adecuado para un cultivar graficando una línea que conecte el origen de coordenadas y el identificador del genotipo de interés, por ejemplo Kat, como se muestra en la Figura 4.3 (B) (Yan y Kang, 2003). Los ambientes se clasifican a lo largo del eje del genotipo en la dirección indicada por la flecha. Para obtener este gráfico la opción *Selected Genotype* debe indicarse en el argumento *type* y el genotipo de interés en *selectedG*. El eje perpendicular al del genotipo separa los ambientes en los que el cultivar presentó un rendimiento por debajo y por encima del promedio. En este ejemplo, Kat presentó un desempeño por debajo de la media en todos los ambientes estudiados.

```
# Ranking de cultivares en el ambiente OA93
```

```
GGEPlot(GGE1, type = "Selected Environment", selectedE = "OA93",  
         footnote = F, titles = F)
```

```
# Ranking de ambientes para cultivar Kat
```

```
GGEPlot(GGE1, type = "Selected Genotype", selectedG = "Kat", footnote = F, titles = F)
```

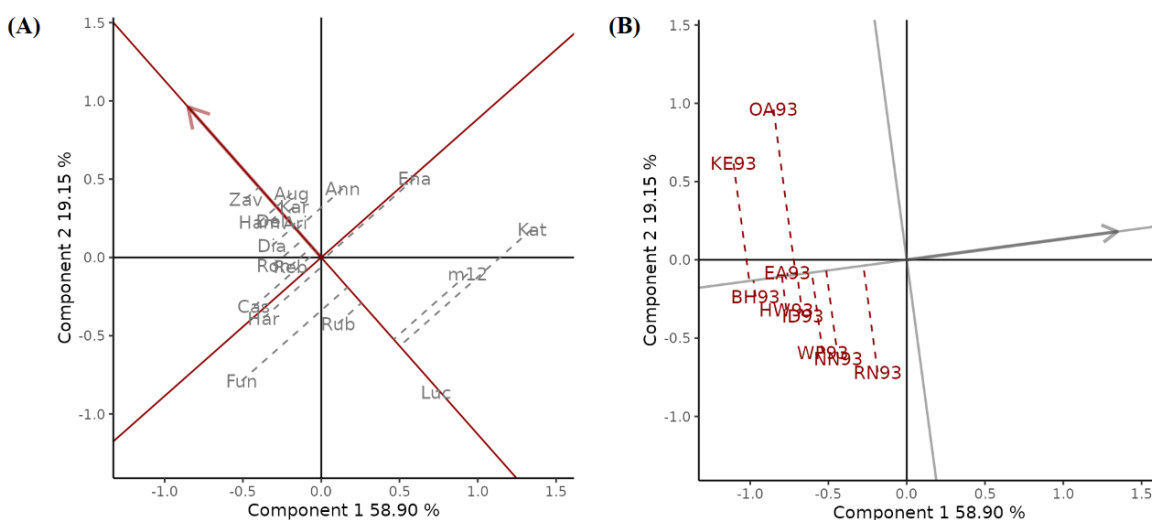
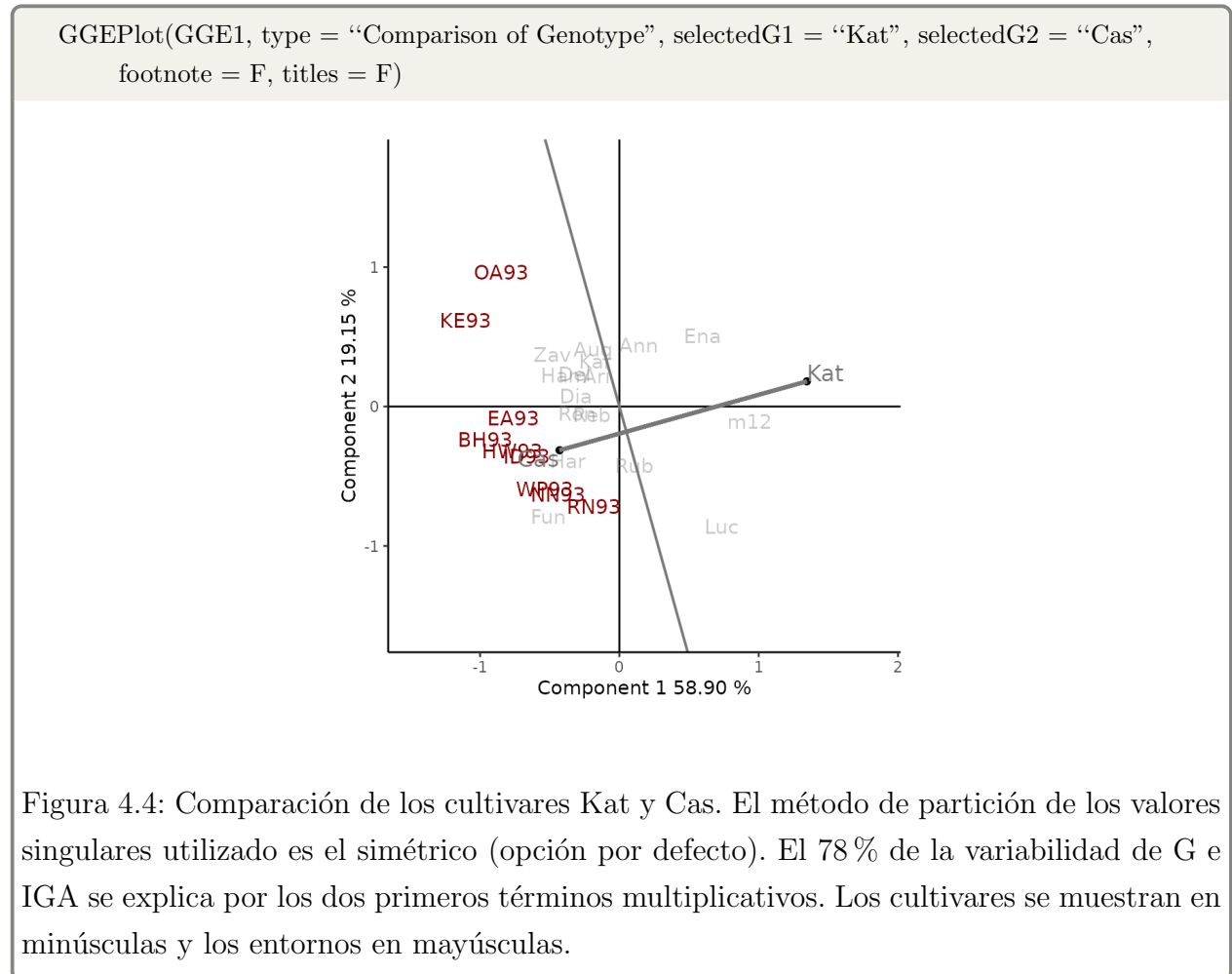


Figura 4.3: Ranking de (A) cultivares en el ambiente OA93 y (B) ambientes para cultivar Kat, basado en datos de rendimiento de trigo de invierno obtenido de Ontario en 1993. El método de partición de los valores singulares utilizado es el simétrico (opción por defecto). El 78 % de la variabilidad de G e IGA se explica por los dos primeros términos multiplicativos. Los cultivares se muestran en minúsculas y los entornos en mayúsculas.

También es posible comparar dos cultivares, por ejemplo Kat y Cas, vinculándolos con una línea y trazando una recta perpendicular a ella (Figura 4.4). Este biplot se obtiene con *Comparison of Genotype* en el argumento *type* y los genotipos a comparar en *selectedG1* y *selectedG2*. Cas fue más rendidor que Kat en todos los ambientes, ya que todos se ubican en el mismo lado de la línea perpendicular que Cas.

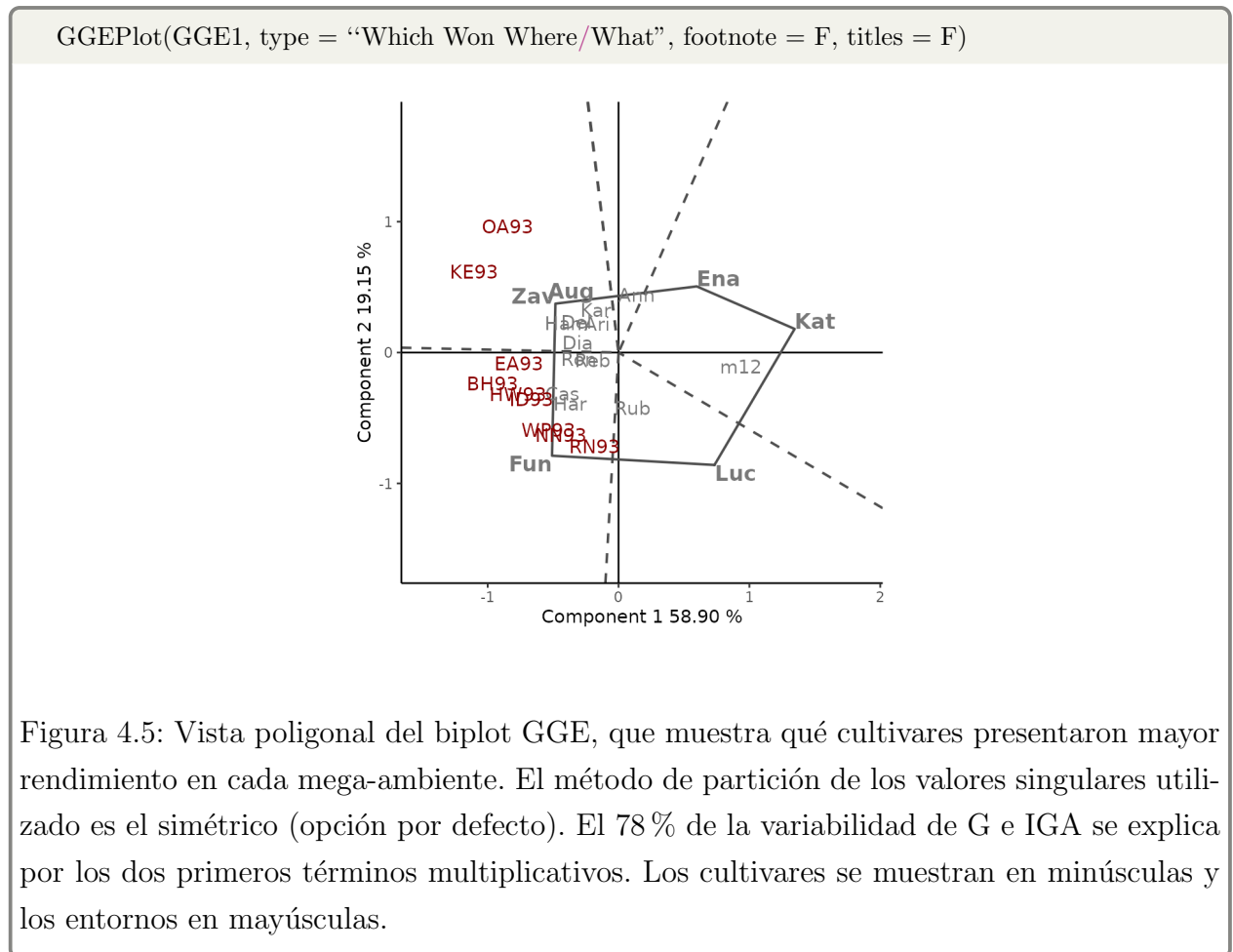


### Identificación de mega-ambientes con GGE biplot

La vista poligonal del biplot GGE, obtenida al indicar *Which Won Where/What* en el argumento *type*, proporciona un medio eficaz de visualización del patrón “cuál-ganó- donde” de un conjunto de datos provenientes de EMA (Figura 4.5). El polígono se obtiene uniendo los cultivares (fun, zav, ena, kat y luc) que se encuentran más alejados del origen de coordenadas, de modo que todos los restantes se encuentren contenidos en el polígono. La distancia de los

cultivares respecto del origen de coordenadas, en sus respectivas direcciones, es una medida de la capacidad de respuesta a los ambientes. Aquellos ubicados en los vértices son los más alejados, por lo tanto son los cultivares que más responden, mientras que los que se encuentran en el origen de coordenadas no responden en absoluto a los ambientes estudiados.

Las rectas perpendiculares a los lados del polígono dividen al biplot en mega-ambientes. El cultivar de mayor rendimiento en todos los ambientes de un mega-ambiente es el que se encuentra en el vértice del polígono. Por un lado, se observa que OA93 y KE93 conforman un mega-ambiente y que Zav es el mejor cultivar. El resto de los ambientes, forman otro mega-ambiente (identificado como ME1) siendo Fun el cultivar que se encuentra en el vértice. En el sector con ena, kat y luc en los vértices del polígono no se observó ningún ambiente, lo cual indica que estos cultivares fueron los menos rendidores en algunos o todos los ambientes considerados.



### Evaluación de los cultivares dentro de un mega-ambientes con GGE biplot

Una vez identificado los mega-ambientes, el siguiente paso es seleccionar cultivares dentro de cada uno de ellos. De acuerdo con la Figura 4.5, zav es el mejor cultivar para los ambientes en uno de los mega-ambiente y fun para el otro. Sin embargo, los fitomejoradores generalmente no seleccionan un único cultivar en cada mega-ambiente, sino que evalúan a todos con el fin de conocer su desempeño (rendimiento y estabilidad).

El biplot GGE, particularmente utilizando el factor de partición de los valores singulares enfocando en los genotipos, es decir utilizando el argumento  $SVP = \text{"row"}$  en la función `GGE-model()`, proporciona un medio superior para visualizar tanto el rendimiento medio como la estabilidad de los genotipos. Esto se debe a que la unidad de ambos ejes para los genotipos es la unidad original de los datos. Además, dado que el interés radica en los genotipos y no en los ambientes, estos son omitidos del gráfico con el argumento  $sizeEnv = 0$ .

La visualización del rendimiento medio y la estabilidad de los genotipos se logra dibujando una coordenada ambiental promedio (AEC, siglas en inglés de *Average environment coordination*). Por ejemplo, la Figura 4.6 (A) muestra el AEC para el mega-ambiente ME1 compuesto por los entornos BH93, EA93, HW93, ID93, NN93, RN93 y WP93. La abscisa representa el efecto de G y la ordenada el de la IGA, que es una medida de la variabilidad o inestabilidad asociada con cada genotipo. Los cultivares se clasifican de acuerdo a su rendimiento medio a lo largo de la abscisa del AEC de acuerdo a la dirección de dicho eje, mientras que una proyección sobre la ordenada AEC muy alejada del origen, independientemente de la dirección, significa mayor inestabilidad. El cultivar de mayor rendimiento promedio en este mega-ambiente fue Fun, seguido por Cas y Har, mientras que Kat fue el de peor rendimiento medio. Rub y Dia son más variables y menos estables que otros cultivares, por el contrario, Cas, Zav, Reb, Del, Ari y Kar, fueron más estables.

La Figura 4.6 (B) compara los cultivares con uno considerado “ideal” por ser el más rendidor y con estabilidad absoluta. Este cultivar ideal se usa como referencia, ya que rara vez existe. La distancia entre los cultivares y el ideal se puede utilizar como medida de conveniencia. Los círculos concéntricos ayudan a visualizar estas distancias. En el ejemplo, para el ME1, Fun es el más cercano al cultivo ideal y por tanto el más deseable, seguido de Cas y Har, mientras que Kat fue el más lejano.



```
ME1 <- yan.winterwheat[yan.winterwheat$env %in% c("BH93", "EA93", "HW93", "ID93",
  "NN93", "RN93", "WP93"), ]

# Modelo SREG enfocando SVD en los genotipos
GGE_Gpartition <- GGEmodel(ME1, genotype = "gen", environment = "env",
  response = "yield", SVP = "row")

# Visualizacion del rendimiento medio y la estabilidad
GGEPlot(GGE_Gpartition, type = "Mean vs. Stability", footnote = F, titles = F, sizeEnv = 0)

# Ranking de los genotipos respecto a uno ideal
GGEPlot(GGE_Gpartition, type = "Ranking Genotypes", footnote = F, titles = F,
  sizeEnv = 0)
```

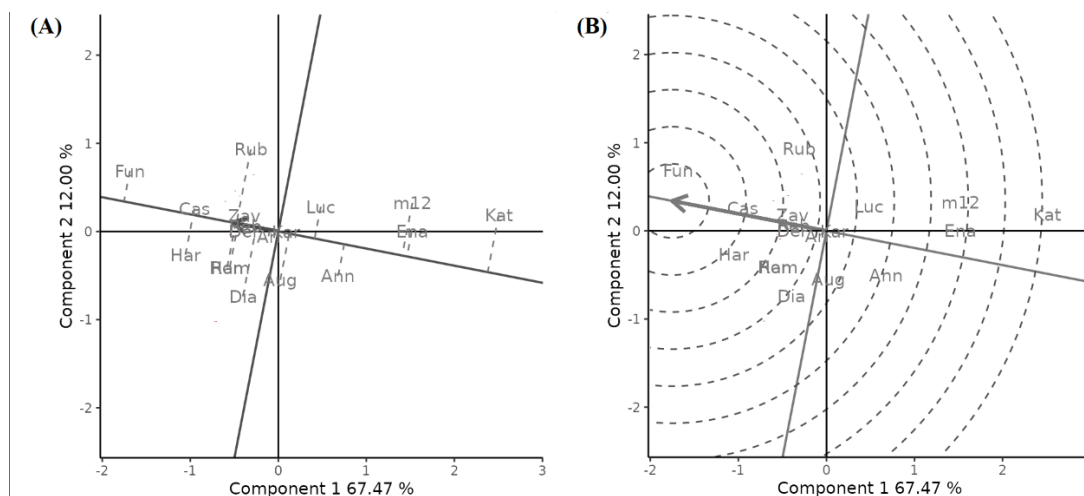


Figura 4.6: (A) Evaluación de los cultivares con base en el rendimiento promedio y la estabilidad y (B) clasificación de genotipos con respecto al genotipo ideal, basado en el método de partición de los valores singulares enfocado en los genotipos.

### Evaluación de los ambientes con GGE biplot

A pesar de que el objetivo principal de los EMA es seleccionar cultivares también es posible evaluar los ambientes. Esto incluye varios aspectos: (i) evaluar si la región objetivo pertenece a uno o más mega-ambientes; (ii) identificar mejores ambientes de prueba; (iii) detectar ambientes redundantes que no brindan información adicional sobre cultivares; y (iv) determinar los ambientes que se pueden utilizar para la selección indirecta. Para ello,

se enfoca la partición de los valores singulares en los ambientes al ajustar el modelo SREG ( $SVP = \text{"column"}$  en la función `GGEmodel()`).

En la Figura 4.7 los ambientes están conectados con el origen de coordenadas a través de vectores, permitiendo comprender las interrelaciones entre ellos. Esta visualización del biplot GGE se obtiene indicando *Relationship Among Environments* en el parámetro *type* (Figura 4.7 (A)). El coeficiente de correlación entre dos ambientes es aproximadamente el coseno del ángulo entre sus vectores. En este ejemplo se considera la relación entre los ambientes de ME1. El ángulo entre los vectores para los entornos NN93 y WP93 es de aproximadamente  $10^\circ$  entre sus vectores; por lo tanto, están estrechamente relacionados; mientras que RN93 y BH93 presentan una correlación negativa débil ya que el ángulo es levemente mayor a  $90^\circ$ . El coseno de los ángulos no se traduce precisamente en coeficientes de correlación, ya que el biplot no explica toda la variabilidad en el conjunto de datos. Sin embargo, son lo suficientemente informativos como para comprender la interrelación entre los entornos de prueba.

Si algunos de los ambientes tienen ángulos pequeños entre sí y, por lo tanto, están altamente correlacionados, la información sobre los genotipos obtenidos de estos ambientes debe ser similar. Si esta similitud se repite a través de los años, estos ambientes son redundantes y la evaluación de uno solo debería ser suficiente, permitiendo reducir costos en la experimentación.

La capacidad de discriminación así como la representatividad respecto del ambiente objetivo, son medidas fundamentales para un ambiente. Si no tiene capacidad de discriminación, no proporciona información sobre los cultivares y, por lo tanto, carece de utilidad. A su vez, si no es representativo no sólo que carece de utilidad sino que también puede proporcionar información sesgada sobre los cultivares evaluados. Para visualizar estas medidas, se define una coordenada ambiental promedio (AEC mencionado anteriormente) y el ambiente ideal como el centro de un conjunto de círculos concéntricos (Figura 4.7 (B)). Para obtener este biplot se debe indicar *Ranking Environments* en el argumento *type* de `GGEPlot()`. El ángulo entre el vector de un ambiente y el eje proporciona una medida de la representatividad. Por lo tanto, EA93 e ID93 son los más representativos, mientras que RN93 y BH93 son los menos representativos del ambiente promedio, cuando se analiza ME1. Por otro lado, para ser discriminativo debe estar cercano al ambiente ideal. HW93 es el ambiente más cercano al ideal y, por lo tanto, es el más deseable del ME1, seguido por EA93 e ID93. Por el contrario, RN93 y BH93 fueron los ambientes de prueba menos deseables de ME1.

```
# Modelo SREG enfocando SVD en los ambientes
```

```
GGE_Epartition <- GGEModel(ME1, genotype="gen", environment="env", response="yield",  
  SVP="column")
```

```
# Relacion entre ambientes
```

```
GGEPlot(GGE_Epartition, type = "Relationship Among Environments", footnote = F,  
  titles = F)
```

```
# Clasificacion de ambientes con respecto al ambiente ideal
```

```
GGEPlot(GGE_Epartition, type = "Ranking Environments", footnote = F, titles = F)
```

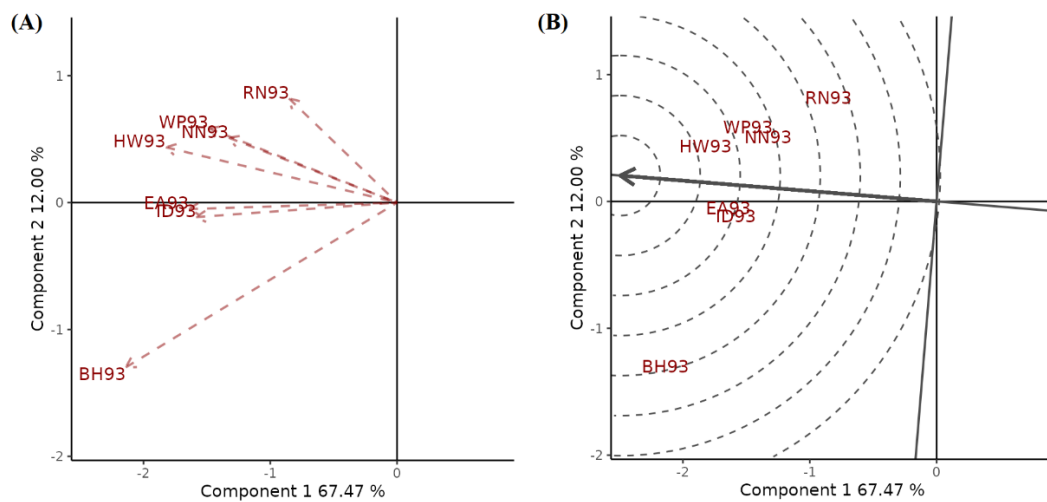


Figura 4.7: (A) Relación entre ambientes y (B) clasificación de ambientes con respecto al ambiente ideal, basado en el escalado centrado en los genotipos.

#### 4.1.4. Uso del paquete para imputar matrices de datos incompletas

Una limitación importante de los modelos presentados anteriormente es que requieren que el conjunto de datos este completo, es decir que todos los genotipos sean evaluados en todos los ambientes. Una forma de superar esta fragilidad es imputando los registros incompletos, para luego poder utilizar la metodología AMMI o SREG. Por lo tanto, en el paquete se incluyen una serie de metodologías de imputación desarrolladas específicamente para datos genotipo-ambiente recientemente publicadas, algunas de las cuales no se encuentran disponible en R, para superar el problema de las observaciones perdidas. Entre los métodos incluidos se encuentran: “EM-AMMI”, “EM-SVD”, “Gabriel”, “WGabriel” y “EM-PCA”, los cuales se

indican en la opción *type* de la función `imputation()`. El formato requerido para el conjunto de datos de entrada es análogo al indicado en las otras funciones incluidas en el paquete.

Para presentar un ejemplo, se eliminan algunas observaciones del conjunto de datos *yan.winterwheat* ya que cuenta con todos los registros completos:

```
# Generando datos faltantes
yan.winterwheat [1,3] <- NA
yan.winterwheat [3,3] <- NA
yan.winterwheat [2,3] <- NA
```

La imputación de valores perdidos con el método “EM-AMMI” se puede realizar de la siguiente manera:

```
imputation(yanwinterwheat, PC.nb = 2, genotype = “gen”, environment = “env”,
           response = “yield”, type = “EM-AMMI”)
```

El resultado es la matriz con datos imputados en aquellas celdas vacías.

## 4.2. Aplicación web *Geneticae*

Siguiendo el procedimiento descrito se desarrolló la aplicación web *Geneticae*, con el objetivo de proporcionar una interfaz gráfica de usuario para el paquete, de modo que pueda ser utilizado por fitomejoradores y analistas sin experiencia previa en programación R. El código fuente de *Geneticae* se encuentra disponible en GitHub (<https://github.com/jangelini/Geneticae-Shiny-Web-APP>).

Es un software interactivo, no comercial y de código abierto, que ofrece una alternativa gratuita al software comercial disponible para analizar datos provenientes de EMA. Momentáneamente se encuentra disponible en un servidor gratuito con una cuota mensual de horas de uso, al cual se puede acceder con el enlace <https://geneticae.shinyapps.io/geneticae-shiny-web-app/> o desde la página web <https://www.cefobi-conicet.gov.ar/bases-de-datos-y-programas/> del Centro de Estudios Fotosintéticos y Bioquímicos del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

A la brevedad la aplicación estará disponible para su uso ilimitado en los servidores del Centro de Cómputos del Centro Científico Tecnológico (CCT) de Rosario, miembro del Sistema Nacional de Computación de Alto Desempeño. Como parte de este trabajo se solicitó

al Centro de Cómputos la instalación de un *Shiny Server* (software que permite publicar este tipo de aplicaciones), habiendo resultado exitosa la publicación de una primera versión de prueba de *Geneticae*. Actualmente, la versión final se encuentra en la cola de trabajo del equipo del CCT para su instalación definitiva y consiguiente apertura al público general.

Como se ha mencionado con anterioridad, Frutos et al. (2013) han propuesto un software interactivo, no comercial y de código abierto llamado *GGEBiplotGUI*, que ofrece una alternativa gratuita a los softwares comercial disponibles hasta el momento. Sin embargo, aunque sigue disponible en GitHub, en 2021 fue archivado de CRAN debido a falta de mantenimiento.

Si bien con la aplicación *GGEBiplotsGUI* se pueden llevar obtener los biplots clásicos incluidos en la aplicación *Geneticae*, tiene ciertas limitaciones. En primer lugar, no dispone de la versión robusta recientemente publicada del modelo AMMI la cual es muy importante dado que los *outliers* son muy frecuentes en datos provenientes de EMA. Además, para importar los datos se debe utilizar usando comandos de R, requiriendo un conocimiento, aunque sea mínimo, del lenguaje. Otra desventaja es que no se admiten repeticiones en el conjunto de datos las cuales son muy común en los experimentos de campo así como tampoco la presencia de variables adicionales registradas que no se usan en éste tipo de análisis. Además, los genotipos deben estar ubicados en las filas y los ambientes en las columnas. Esto implica que se debe realizar un preprocesamiento/configuración de los datos antes del análisis. Por último, no cuenta con opciones para realizar un análisis descriptivo previo a los biplots, por lo que se debe recurrir a otro software tanto para el procesamiento como para la descripción inicial de los datos.

Con la aplicación *Geneticae* se avanza en la construcción de un software interactivo más flexible, no comercial y de código abierto para analizar, visualizar y extraer los resultados desde una página web sin la necesidad de contar con conocimiento específico de un lenguaje de programación y que a su vez incluya los avances metodológicos importantes para campo de estudio. En las subsecciones siguientes se presentará un ejemplo de cómo cargar y analizar datos con la aplicación web *Geneticae*.

#### 4.2.1. Lectura de un archivo de datos para el uso de la aplicación web *Geneticae*

La aplicación web *Geneticae* requiere que los datos de entrada se encuentren en archivos de texto plano, con delimitaciones por comas o punto y coma (formato .CSV) o tabulaciones (formato .TXT). Los nombres de las columnas pueden ubicarse en la primera fila del archivo

(*heading*). Los datos deben estar en formato largo, es decir, cada fila debe corresponder a una observación y cada columna a una variable (genotipo, ambiente, fenotipo observado y, si existe, repetición). Si cada genotipo ha sido evaluado más de una vez en cada ambiente, la media fenotípica requerida por el modelo SREG y AMMI para cada combinación de genotipo y ambiente se calcula internamente antes de ajustar dichos modelos. Las variables adicionales que no se utilizarán en el análisis pueden estar presentes en el conjunto de datos. No se permiten valores perdidos.

Los dos conjuntos de datos *plrv* y *yanwinterwheat* mencionados en la subsección 4.1.1 están disponibles en la pestaña *Data* → *Example datasets* y se pueden descargar en formato .CSV para poder seguir el tutorial de uso de la aplicación. El conjunto de datos *yanwinterwheat* no tiene repeticiones, mientras que *plrv* sí.

En los siguientes ejemplos se trabajará con *yanwinterwheat* para mostrar el uso de la aplicación. Luego de obtener el archivo de datos con el formato indicado, es posible importarlo en la pestaña *Data* → *Upload data*. Por ejemplo, para importar *yanwinterwheat*, se debe cargar el archivo .CSV. Una vez cargado, se debe indicar que está delimitado por comas, que en la primera fila contiene los nombres de cada variable (*heading*) y los nombres de las columnas en las cuales se encuentra la información del genotipo, ambiente y rasgo fenotípico (*gen*, *env* y *yield* en este ejemplo) (Figura 4.8). Si hay repeticiones disponibles, se debe especificar el nombre de la columna con dicha información; de lo contrario, el campo queda vacío.

gen	env	yield
Ann	BH93	4.46
Ari	BH93	4.417
Aug	BH93	4.669
Cas	BH93	4.732
Del	BH93	4.39
Dia	BH93	5.178
Ena	BH93	3.375
Fun	BH93	4.852
Ham	BH93	5.038
Har	BH93	5.195

Figura 4.8: Importar el conjunto de datos *yanwinterwheat* en la aplicación web *Geneticae*.

### 4.2.2. Uso de la aplicación web *Geneticae* para análisis descriptivo

Cualquier estudio debe comenzar con un análisis descriptivo del conjunto de datos. La pestaña *Descriptive Analysis* proporciona algunas herramientas para esto, como *boxplot*, diagrama y matriz de correlación y gráficos de interacción.

Un *boxplot* que compara el rasgo cuantitativo entre ambientes o genotipos puede ser de interés (Figura 4.9). Las medidas de resumen utilizadas para su construcción se muestran de forma interactiva moviendo el *mouse* dentro del panel de la Figura. Además, se puede descargar en formato .PNG o como un archivo interactivo .HTML, haciendo clic en la cámara que aparece en el gráfico o en el botón Descargar, respectivamente. El usuario puede personalizar algunos aspectos del gráfico, como el color de las cajas y los nombres de los ejes.

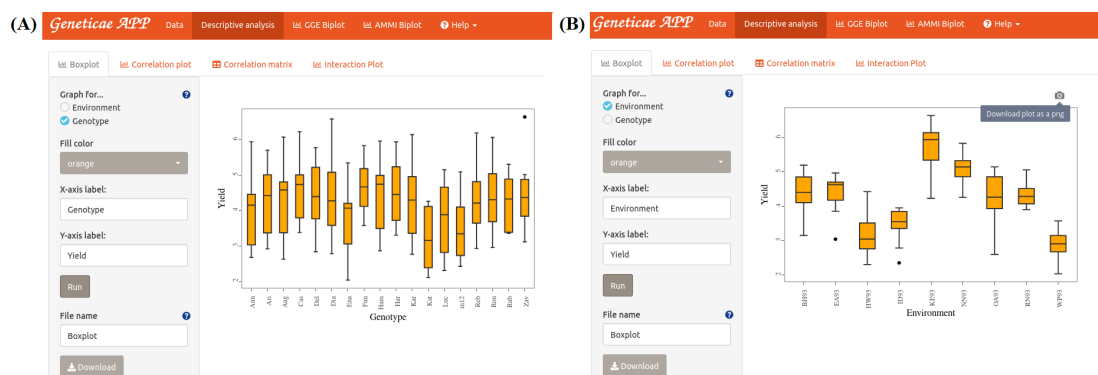


Figura 4.9: Diagrama de caja de (A) genotipos y (B) ambientes para el conjunto de datos *yanwinterwheat* obtenido con la aplicación web *Geneticae*.

Los coeficientes de correlación de Pearson o Spearman entre genotipos o ambientes se pueden mostrar con un gráfico o una matriz (Figura 4.10). Gráficamente, las correlaciones positivas se muestran en azul y las negativas en rojo, mientras que la intensidad del color y el tamaño del círculo son proporcionales a la magnitud de los coeficientes de correlación. La gráfica de correlación se puede descargar en formato .PNG. En el ejemplo, se observan altas correlaciones entre el rendimiento de los genotipos a través de la red ambiental considerada. Esto podría deberse a la falta de IGA, asociada a una estabilidad estática, o bien a la presencia de una IGA NCOI, asociada a una estabilidad dinámica. En el caso de una estabilidad estática los genotipos no son capaces de responder a las variaciones que puedan mostrar los ambientes, mientras que, de existir una estabilidad dinámica, el comportamiento estará cerca del potencial del ambiente.

Por otro lado, si se analizan las correlaciones entre los ambientes, se podrían detectar ambientes similares y contrastantes. Los primeros podrían agruparse en un mega-ambiente, al tiempo que algunos de ellos deberían ser desconsiderados con la finalidad de reducir el costo de los ensayos.

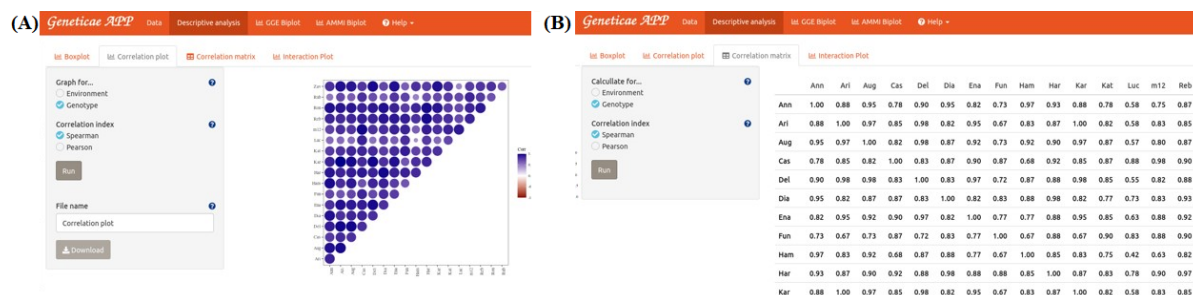


Figura 4.10: (A) Gráfico y (B) matriz de correlación entre genotipos para el conjunto de datos *yanwinterwheat* obtenido con la aplicación web *Geneticae*.

Dado que la IGA genera respuestas genotípicas diferenciales en diferentes ambientes complicando la selección de los mejores cultivares, una gráfico de interacción puede ser de interés (Figura 4.11). El cambio en el efecto genotípico a través de los ambientes se muestra en la Figura 4.11 (A), mientras que el cambio en el efecto ambiental a través de los genotipos en la Figura 4.11 (B). Estos también son gráficos interactivos por lo que es posible descargarla en formatos .HTML o .PNG con el botón Descargar o haciendo clic en la cámara que aparece en el gráfico, respectivamente. Además, el usuario puede personalizar los nombres de los ejes. En este ejemplo se pueden ver inconsistencias en el desempeño de genotipos en diferentes ambientes.

Este análisis grafico puede complementarse con el modelo de regresión lineal propuesto por Eberhart y Russell (1996) ampliamente utilizado por los fitomejoradores para el estudio de la estabilidad y adaptabilidad fenotípica de cultivares. En esta metodología, la estabilidad se expresa en términos de tres parámetros empíricos: el rendimiento medio, el coeficiente de regresión,  $b_i$ , y la suma de cuadrados del error de la regresión,  $(S_{ij}^2)$  (Crossa, 1990; Flores et al., 1998). El  $S_{ij}^2$  está fuertemente relacionado con una parte impredecible de la variabilidad del genotipo y se considera un parámetro de estabilidad. El parámetro  $b_i$  caracteriza la respuesta específica de los genotipos a los efectos ambientales indicando la adaptabilidad del genotipo, o bien, su capacidad de respuesta entre los distintos ambientes (Breese, 1969). Considerando estas tres medidas, el mejorador posee información suficiente para identificar aquellos genotipos de mejor comportamiento de la red de ambiental.



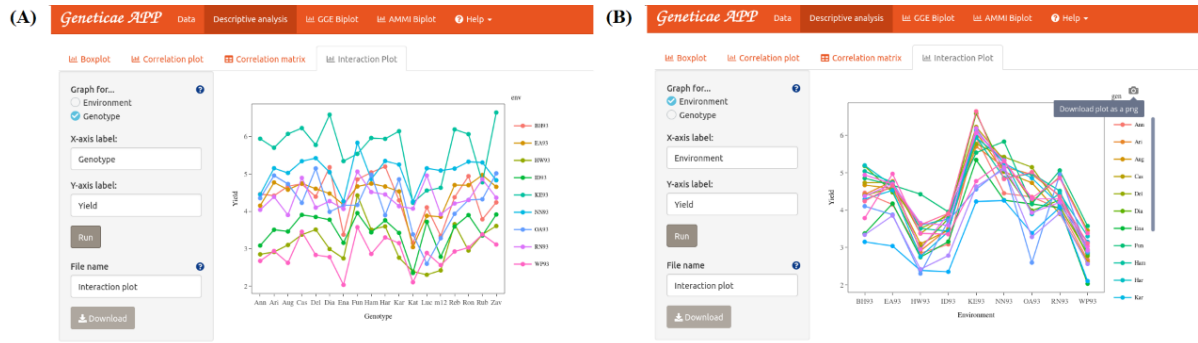
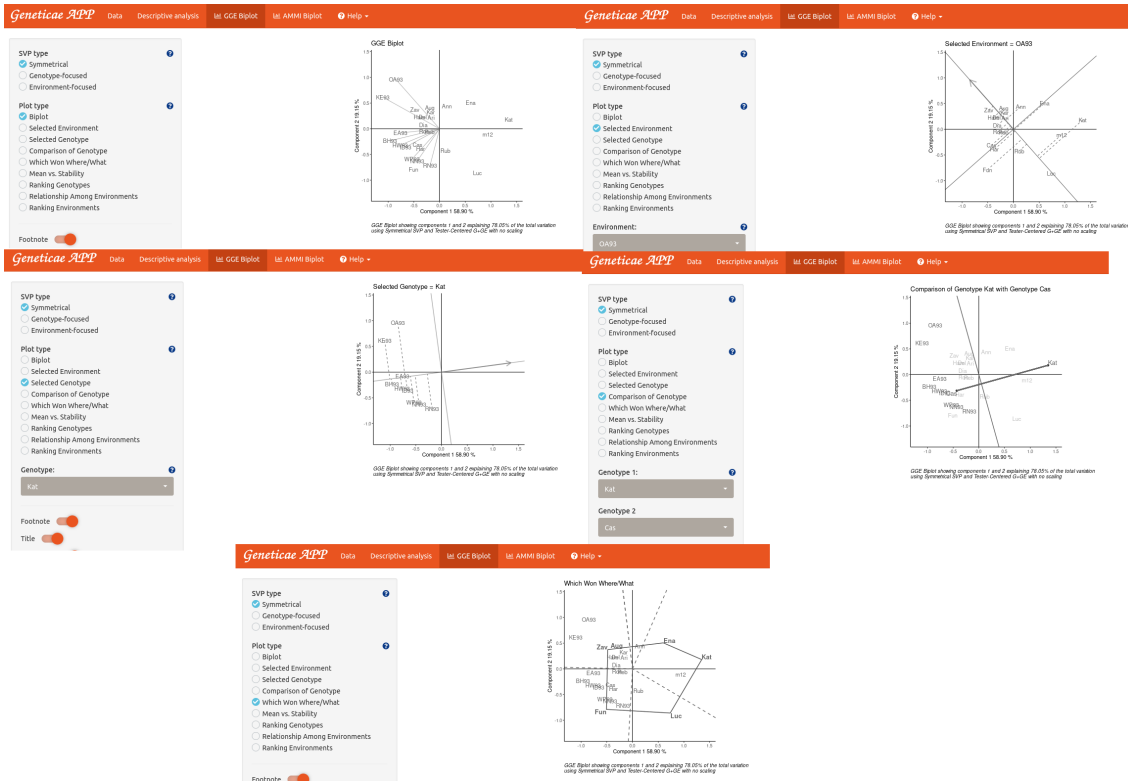


Figura 4.11: Gráfico de interacción para (A) ambientes a través de genotipos y (B) genotipos a través de los ambientes para conjunto de datos de *yanwinterwheat* obtenido con la aplicación web *Geneticae*.

### 4.2.3. Uso de la aplicación web *Geneticae* para ajustar el modelo SREG

La aplicación web *Geneticae* permite generar las vistas del biplot GGE presentados en la subsección 4.1.3 mediante la pestaña *GGE Biplot*. Del mismo modo que en el paquete *geneticae*, los cultivares se presentan en minúsculas y los ambientes en mayúsculas. Dado que el modelo SREG requiere una única observación para cada combinación de genotipo y ambiente, si hay repeticiones, el valor fenotípico promedio se calcula automáticamente antes de ajustar el modelo. No se permiten valores perdidos.

Se debe seleccionar el método de partición de los valores singulares (*SVP type*). Como se mencionó anteriormente esta elección no altera las relaciones o interacciones relativas entre genotipos y ambientes, aunque la apariencia del biplot será diferente (Yan, 2002). La opción simétrica permite la comparación tanto de genotipos como de ambientes (opción por defecto); *Genotype-Focused* muestra la interrelación entre genotipos con mayor precisión que cualquier otro método y *Environment-Focused* es la que más informa sobre las interrelaciones entre ambientes. Una nota a pie del gráfico indica el método de centrado utilizado (para el biplot GGE siempre es *tester-center*), el método SVP elegido por el usuario y que no se aplica ninguna escala a los datos. Además, se puede agregar el porcentaje de variabilidad de G e IGA explicado por las dos PC, mientras que la inclusión de títulos, ejes y etiquetas es opcional. Ciertos atributos estilísticos de dichos gráficos se pueden personalizar como el color y tamaño de los identificadores de los genotipos y de los ambientes. Los gráficos pueden ser descargados.



alguno de estas vistas del biplot GGE se tendrá que señalar cuales son los ambientes que forman el mega-ambiente de interés.

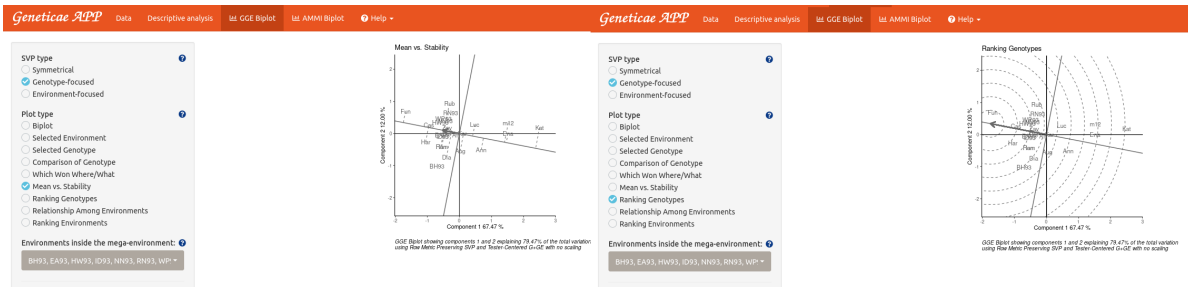


Figura 4.13: Vistas del biplot GGE usando la partición de los valores singulares enfocada en los genotipos obtenidos con la aplicación web *Geneticae*.

Por último, para el análisis de los ambientes de cada mega-ambiente se utiliza el método de partición de los valores singulares centrado en los ambientes (*SVP type*  $\rightarrow$  *environment-focused*). El tipo de gráfico *Relationship Among Environments* permite comprender la interrelación entre los ambientes, mientras que *Ranking Environments* se utiliza para visualizar la capacidad de discriminación y representatividad (Figura 4.14). Como en el caso anterior, al indicar alguna de estas vistas del biplot GGE se tendrá que señalar cuales son los ambientes que forman el mega-ambiente de interés.

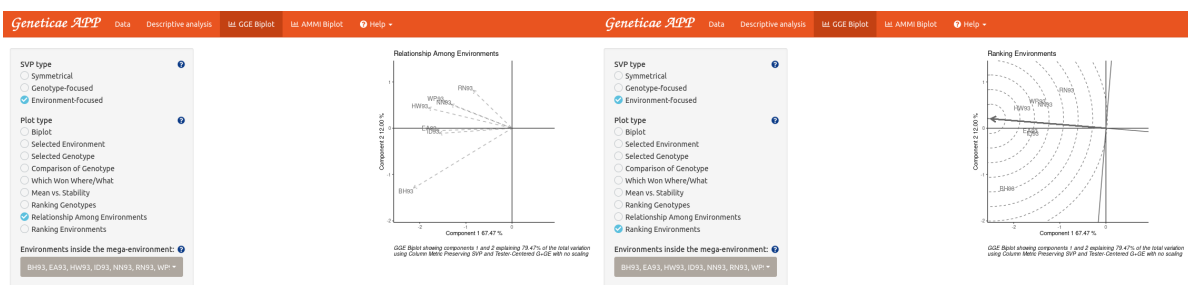


Figura 4.14: Vistas del biplot GGE usando la partición de los valores singulares enfocada en los ambientes obtenidas con la aplicación web *Geneticae*.

#### 4.2.4. Uso de la aplicación web *Geneticae* para ajustar el modelo AMMI

La pestaña *AMMI Biplot* crea el biplot GE. Dado que las alternativas clásica y robustas requieren una única observación para cada combinación de genotipo y ambiente, si hay repeticiones, el valor promedio fenotípico se calcula automáticamente antes de ajustar el modelo. No se permiten valores perdidos. Como con las figuras anteriores, se ofrecen opciones de configuración y de descarga.

Por ejemplo, para obtener el biplot GE derivado del modelo AMMI clásico se debe indicar AMMI en *plot type* (Figura 4.15). En caso de contar con *outliers*, alguna de las alternativas robustas (rAMMI, hAMMI, gAMMI, lAMMI o ppAMMI).



Figura 4.15: Biplot GE derivado del modelo AMMI clásico basado en datos de rendimiento de trigo de invierno obtenido de Ontario en 1993 obtenido con la aplicación web *Geneticae*.

### 4.2.5. Ayuda

En la pestaña *Help* se presenta información general, un tutorial y un video sobre cómo utilizar la aplicación web *Geneticae*.

---

# Capítulo 5

## Conclusión

En el presente trabajo se crearon herramientas informáticas para el análisis para datos provenientes de EMA a fin de facilitar la tarea de los mejoradores genéticos de cultivos. Por un lado, se desarrolló un paquete de R llamado *geneticae* que reúne metodología ampliamente utilizada y recientemente publicada, y que está abierto a que nuevas propuestas para el estudio de la interacción genotipo ambiente sean incorporados. Por otro lado, se confeccionó una interfaz gráfica de usuario que permite analizar, visualizar y extraer los resultados desde una página web sin la necesidad de contar con conocimiento específico de un lenguaje de programación.

Como resultados del presente trabajo fue posible:

**Mostrar un flujo de trabajo reproducible para la construcción de paquetes de R.** El mismo se puede utilizar de ejemplo para el desarrollo de nuevos paquetes o imitar la construcción del paquete *geneticae* objeto de este trabajo.

**Construir un paquete de R llamado *geneticae*** para el análisis de datos provenientes de EMA. A pesar de que aún no se haya hecho difusión del paquete, hasta el momento<sup>1</sup> el paquete cuenta con 2500 descargas. La gran utilidad del mismo se debe a que:

- incluye metodología recientemente publicada para ajustar el modelo AMMI en presencia de *outliers* y para el tratamiento de información faltante que no se encuentra disponible en R así como tampoco en softwares comerciales,
- ofrece mayor flexibilidad en el manejo de la estructura de los conjuntos de datos que

---

<sup>1</sup>30 de marzo 2022

en las herramientas disponibles hasta este momento,

- brinda la posibilidad de generar representaciones gráficas de los biplots de buena calidad y configurables,
- está acompañado por un manual de ayuda completo y por un tutorial (viñeta) para su uso.

**Desarrollar una aplicación web Shiny denominada *Geneticae***, la cual es de suma importancia para aquellos analistas no familiarizados con la programación. Esta es de libre acceso mediante conexión a internet que permite realizar los principales análisis implementados en el paquete sin necesidad de escribir líneas de código.

**Implementar una metodología de desarrollo de software colaborativa y basada en el sistema de control de versiones Git y los servicios web de GitHub**, adhiriendo a los principios de la investigación reproducible y de libre acceso.

Como línea futura de trabajo se plantea continuar con la inclusión de los avances metodológicos que se vayan publicando en el contexto de datos provenientes de EMA tanto en el paquete como en la aplicación web Shiny. Por ejemplo, en el marco de mi tesis doctoral de estadística, he propuesto una alternativa robusta para el modelo SREG (Angelini et al., 2022) que será puesta a disposición de la comunidad científica y de los mejoradores al incluirla en la próxima versión del paquete y en la aplicación web Shiny.

---

# Bibliografía

- Adu, G. B., Badu-Apraku, B., y Akromah, R. Strategies for selecting early maturing maize inbred lines for hybrid production under low soil nitrogen and striga infestation. *Agronomy*, 11:1309, 2021.
- Aguate, F., Crossa, J., y Balzarini, M. Effect of missing values on variance component estimates in multienvironment trials. *Crop Science*, 59:508–517, 2019.
- Allard, R. y Bradshaw, A. Implications of genotype-environmental interactions in applied plant breeding. *Crop Science*, 4:503–508, 1964.
- Angelini, J., Faviere, G., Bortolotto, E., Cervigni, G., y Quaglino, M. Handling outliers in multi-environment trial data analysis: in the direction of robust sreg model. *Journal of Crop Improvement*, págs. 1–25, 2022.
- Arciniegas-Alarcón, S., García-Peña, M., Dias, C., y Krzanowski, W. An alternativemethodology for imputing missing data in trials with genotype-by-environment interaction. *Biometrical Letters*, 47:1–47, 2010.
- Arciniegas-Alarcón, S., García-Peña, M., Krzanowski, W., y Dias, C. An alternativemethodology for imputing missing data in trials with genotype-byenvironment interaction: some new aspects. *Biometrical Letters*, 51:75–88, 2014.
- Arciniegas-Alarcón, S., García-Peña, M., y Rodrigues, P. C. New multiple imputation methods for genotype-by-environment data that combine singular value decomposition and jackknife resampling or weighting schemes. *Computers and Electronics in Agriculture*, 176:105617, 2020.
- Becker, H. Correlations among some statistical measures of phenotypic stability. *Euphytica*, 30:835–840, 1981.



- Borém, A. *Melhoramento de Plantas.*, cap. Interação genótipo x ambiente, adaptabilidade e estabilidade de comportamento, págs. 109–135. UFV, Viçosa, 2001.
- Breese, E. The measurement and significance of genotype–environment interactions in grasses. *Heredity*, 24:27–44, 1969.
- Chang, W. *shinythemes: Themes for Shiny*, 2018. URL <https://CRAN.R-project.org/package=shinythemes>. R package version 1.1.2.
- Chang, W., Cheng, J., Allaire, J., Xie, J., y McPherson, J. *shiny: Web Application Framework for R*, 2020. URL <https://CRAN.R-project.org/package=shiny>. R package version 1.5.0.
- Cooper, R., M. and Stucker, DeLacy, I., y Harch, B. Wheat breeding nurseries, target environments, and indirect selection for grain yield. *Crop Science*, 37:1168–1176, 1997.
- Cornelius, P., Crossa, J., y Seyedsadr, M. *Genotype by Environment Interaction*, cap. Statistical test and estimators of multiplicative models for genotype-by-environment interaction., págs. 199–234. CRC Press, Boca Raton, 1996.
- Crossa, J. Statistical analyses of multilocation trials. *Advances in Agronomy*, 44:55–85, 1990.
- Crossa, J. y Cornelius, P. L. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. *Crop Science*, 37:406–415, 1997.
- Crossa, J., Gauch, H., y Zobel, R. Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. *Crop Science*, 30:493–500, 1990.
- Croux, C. y Ruiz-Gazen, A. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95:206–226, 2005.
- Croux, C., Filzmoser, P., y Oliveira, M. R. Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87:218–225, 2007.
- Cruz, C. y Regazzi, A. *Modelos Biométricos Aplicados Ao Melhoramento Genético.*, cap. Interação genótipo x ambiente., págs. 1–24. UFV, Viçosa, 1997.
- Cruz Medina, R. Some exact conditional tests for the multiplicative models to explain genotype-environment interaction. *Heredity*, 69:128–132, 1992.

- de Mendiburu, F. *agricolae: Statistical Procedures for Agricultural Research*, 2021. URL <https://CRAN.R-project.org/package=agricolae>. R package version 1.3-5.
- de Oliveira, L. A., da Silva, C. P., Nuvunga, J. J., da Silva, A. Q., y Balestre, M. Bayesian gge biplot models applied to maize multi-environments trials. *Genetics and Molecular Research*, 15, 2016.
- Dumble, S. *GGEbiplots: GGE Biplots with 'ggplot2'*, 2022. URL <https://CRAN.R-project.org/package=GGEbiplots>. R package version 0.1.3.
- Eberhart, S. y Russell, W. Stability parameters for comparing varieties. *Crop Science*, 6:36–40, 1996.
- Ezequiel-Hernández, F., Peruzzo, A. M., Pratta, G., y Pioli, R. N. Identificación de diversidad patogénica de phomopsis sp. causal del tizón de tallo y vaina en soya (glycine max) mediante marcadores moleculares. *Agrociencia*, 54:313–326, 2020.
- Flores, F., Moreno, M., y Cubero, J. A comparison of univariate and multivariate method to analyze environments. *Field Crops Research*, 56:271–28, 1998.
- Frutos, E., Galindo, M. P., y Leiva, V. An interactive biplot implementation in r for modeling genotype-by-environment interaction. *Stochastic Environmental Research and Risk Assessment*, 2013.
- Ganz, C., Csárdi, G., Hester, J., Lewis, M., y Tatman, R. *available: Check if the Title of a Package is Available, Appropriate and Interesting*, 2019. URL <https://CRAN.R-project.org/package=available>. R package version 1.0.4.
- Gauch, H. G. Model selection and validation for yield trials. *Theoretical and Applied Genetics*, 80:153–160, 1988.
- Gauch, H. G. *Statistical analysis of regional yield trials:AMMI analysis of factorial designs*. Elsevier, 1992.
- Gauch, H. y Zobel, R. Imputing missing yield trial data. *Theoretical and Applied Genetics*, 79:753–761, 1990.
- Gauch, H. y Zobel, R. Identifying mega-environments and targeting genotypes. *Crop Science*, 37:311–326, 1997.

- Hadasch, S., Forkman, J., Malik, W., y Piepho, H. Weighted estimation of ammi and gge models. *Journal of Agricultural, Biological and Environmental Statistics*, 23:255–275, 2018.
- Hawkins, D. M., Liu, L., y Young, S. S. Robust singular value decomposition technical report number 122 december , 2001 national institute of statistical sciences 19. 2002.
- Hester, J. *covr: Test Coverage for Packages*, 2020. URL <https://CRAN.R-project.org/package=covr>. R package version 3.5.0.
- Hill, J. R. y Rosenberg, J. Models for combining data from germplasm evaluation trials. *Crop Science*, 25:467–470, 1985.
- Huber, P. J. *Robust Statistics*. John Wiley and Sons, 1981.
- Hubert, M., Rousseeuw, P., y Branden, K. Robpca: A new approach to robust principal component analysis. *Technometrics*, 47:64–79, 2005.
- Jarquín, D., Pérez-Elizalde, S., Burgueño, J., y Crossa, J. A hierarchical bayesian estimation model for multi-environment plant breeding trials in successive years. *Crop Science*, 56:2260–2276, 2016.
- Josse, J. y Husson, F. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70:1–31, 2016.
- Kang, M. y Magari, R. *Genotype by Environment Interaction*, cap. New Developments in Selecting for Phenotypic Stability in Crop Breeding., págs. 201–213. Elsevier, New York, 1996.
- Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J.-T., y Cohen, K. Robust principal components for functional data. *Test*, 8:1–28, 1999.
- Mason-Thom, C. *shinyhelper: Easily Add Markdown Help Files to 'shinyApp Elements*, 2019. URL <https://CRAN.R-project.org/package=shinyhelper>. R package version 0.3.2.
- Olivoto, T. y Lúcio, A. D. metan: an r package for multi-environment trial analysis. *Methods in Ecology and Evolution*, 11(6):783–789, 2020. doi:10.1111/2041-210X.13384.
- Paderewski, J. An r function for imputation of missing cells in two-way data sets by em-ammi algorithm. *Communications in Biometry and Crop Science*, 8:60–69, 2013.

- Perrier, V., Meyer, F., y Granjon, D. *shinyWidgets: Custom Inputs Widgets for Shiny*, 2020. URL <https://CRAN.R-project.org/package=shinyWidgets>. R package version 0.5.3.
- Peto, R. c. *Treatment of cancer*, cap. Statistical aspects of cancer trials, págs. 867–871. Chapman and hall, London, 1982.
- Rodrigues, P., Monteiro, A., y Lourenço, V. A robust ammi model for the analysis of genotype-by-environment data. *Bioinformatics*, 32:58–66, 2016.
- Singh, B., Das, A., Parihar, A. K., Bhagawati, B., Singh, D., Pathak, K. N., Dwivedi, K., Das, N., Keshari, N., Midha, R. L., Kumar, R., Pratap, A., Kumar, V., y Gupta, S. Delineation of genotype-by-environment interactions for identification and validation of resistant genotypes in mungbean to root-knot nematode (*meloidogyne incognita*) using gge biplot. *Scientific Reports*, 10:4108, 2020.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., y Altman, R. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 2001.
- Wickham, H. testthat: Get started with testing. *The R Journal*, 3:5–10, 2011.
- Wickham, H. y Bryan, J. *usethis: Automate Package and Project Setup*, 2021. URL <https://CRAN.R-project.org/package=usethis>. R package version 2.0.1.
- Wickham, H., Danenberg, P., Csárdi, G., y Eugster, M. *roxygen2: In-Line Documentation for R*, 2020. URL <https://CRAN.R-project.org/package=roxygen2>. R package version 7.1.1.
- Wickham, H. y Hesselberth, J. *pkgdown: Make Static HTML Documentation for a Package*, 2020. URL <https://CRAN.R-project.org/package=pkgdown>. R package version 1.6.1.
- Wickham, H., Hester, J., y Chang, W. *devtools: Tools to Make Developing R Packages Easier*, 2021. URL <https://CRAN.R-project.org/package=devtools>. R package version 2.4.2.
- Woyann, L. G., Benin, G., Storck, L., Trevizan, D. M., Meneguzzi, C., Marchioro, V. S., Tonnatto, M., y Madureira, A. Estimation of missing values affects important aspects of gge biplot analysis. *Crop Science*, 57:40–52, 2017.
- Wright, K. *agridat: Agricultural Datasets*, 2020. URL <https://CRAN.R-project.org/package=agridat>. R package version 1.17.

- Wright, K. y Laffont, J.-L. *gge: Genotype Plus Genotype-by-Environment Biplots*, 2021. URL <https://CRAN.R-project.org/package=gge>. R package version 1.7.
- Yan, W. Singular-value partitioning in biplot analysis of multienvironment trial data. *Agronomy Journal*, 94:990–996, 2002.
- Yan, W. Biplot analysis of incomplete two-way data. *Crop Science*, 53:48–57, 2013.
- Yan, W., Cornelius, P., Crossa, J., y Hunt, L. Two types of gge biplots for analyzing multi-environment trial data. *Crop Science*, 41:656–663, 2001.
- Yan, W., Hunt, L. A., Sheng, Q., y Szlavnics, Z. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science*, 40:597–605, 2000.
- Yan, W. y Kang, M. *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists*. CRC Press, 2003.
- Yan, W. y Rajcan, I. Biplot evaluation of test sites and traitrelations of soybean in ontario. *Crop Science*, 42:11–20, 2002.
- Zobel, R., Wright, M., y Gauch, H. Statistical analysis of a yield trial. *Agronomy Journal*, 80:88–393, 1988.