IDENTIFICACIÓN DE RELACIONES ENTRE RENDIMIENTOS Y VARIABLES AMBIENTALES VÍA ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN (CART)

SOLEANA ROSALES HEREDIA, CECILIA BRUNO y MÓNICA BALZARINI

RESUMEN

En investigaciones agrícolas son frecuentes los ensayos multiambientales (EMA) para la comparación de rendimientos de varios genotipos en múltiples ambientes. Para caracterizar los ambientes es común registrar covariables meteorológicas y/o de manejo con potencialidad de explicar relaciones entre rendimientos y ambientes. La contribución relativa de las distintas covariables suele analizarse vía técnicas estadísticas univariadas, principalmente prueba t y análisis de regresión. No obstante, los árboles de clasificación y de regresión, algoritmos CART, constituyen una aproximación multivariada alternativa para identificar variables ambientales de más impacto en los rendimientos. Los CART presentan menos restricciones para su implementación que las técnicas de análisis basadas en los modelos lineales clásicos. En el presente trabajo se evalúa el desempeño de algoritmos CART comparado con los procedimientos clásicos de análisis en una base de datos de rendimientos de soja proveniente de un EMA inserto en la región sojera argentina a través de los ambientes. Los resultados muestran que los algoritmos CART constituyen, debido a su bajo error de predicción, una técnica robusta para predecir la variabilidad en los rendimientos como respuesta a variaciones ambientales. El estudio resalta la necesidad de contemplar la multicolinealidad al trabajar con covariables ambientales y modelos clásicos.

La variación de los rendimientos entre ambientes de una misma zona de cultivo suele ser considerable. Por ello, en investigaciones agrícolas son frecuentes los ensayos multiambientales (EMA) con el fin de comparar rendimientos a través de múltiples ambientes. Los EMA son particularmente útiles para explorar estabilidad de rendimientos a través de ambientes y adaptación específica de un cultivo a ambientes particulares, es decir interacción cultivo×ambiente (Balzarini *et al*., 2005). Los diferentes ambientes pueden surgir a partir de la combinación de distintos sitios de ensayo y años, así como de la implementación de distintas prácticas de manejo (fecha de siembra, barbechos, prácticas de labranza). Para caracterizar los ambientes es común registrar variables meteorológicas y/o de manejo, que si bien no son las variables de interés principal, permiten explicar relaciones entre el rendimiento y los ambientes. Numerosos estudios han indagado la importancia relativa de covariables ambientales en la explicación de la variabilidad de los rendimientos de un cultivo en diferentes ambientes (Tittonell *et al*., 2008; Zheng *et al*., 2009).

La contribución relativa de las distintas covariables ambientales en la explicación de los rendimientos suele realizarse con técnicas univariadas, tales como la prueba t, para comparar rendimientos medios en distintos ambientes generados por la combinación de variables regresoras. Así es posible identificar para qué regresora se detectan las mayores diferencias en rendimiento medio. También es común el uso de regresión lineal para ajustar funciones, en las que las covariables ambientales entran como variables regresoras o explicativas, y es posible identificar su contribución en la explicación de variaciones del rendimiento. En ambos casos, es necesario que se cumpla una

PALABRAS CLAVE / Árbol de Decisión / Ensayos Multiambientales / Soja / Variabilidad de Rendimiento /

Recibido: 21/05/2010. Modificado: 18/10/2010. Aceptado: 20/10/2010.

Soleana Rosales Heredia. Ingeniera Agrónoma y Candidata a Doctora en Ciencias Agropecuarias, Universidad Nacional de Córdoba (UNC), Argentina. Becaria del Consejo Nacional de Investigaciones en Ciencia y Tecnología (CONICET), Argentina. Dirección: Facultad de Ciencias Agropecuarias. Av. Valparaíso s/n Ciudad Universitaria, UNC, Córdoba, Argentina. e-mail: srosales@agro.unc.edu.ar

Cecilia Bruno. Ingeniera Agrónoma, M.Sc. y Doctora en Ciencias Agropecuarias, UNC, Argentina. Profesora, UNC, Argentina.

Mónica Balzarini. Ingeniera Agrónoma. Ph.D., Louisiana State University, EEUU. Profesora, UNC, e Investigadora CONICET, Argentina.

serie de supuestos para validar las pruebas estadísticas realizadas para identificar las relaciones rendimiento-ambiente, como linealidad entre las relaciones, normalidad, homogeneidad de varianzas, independencia entre las covariables o ausencia de multicolinealidad (Searle, 1971). Las correlaciones entre variables ambientales y la falta de linealidad de las relaciones entre éstas y el rendimiento, demandan nuevas aproximaciones estadísticas para la identificación de variables potencialmente predictoras de la variabilidad del rendimiento en los ambientes.

Los árboles de clasificación y de regresión, conocidos como algoritmos CART (del inglés classification and regression trees; Breiman *et al*., 1984) constituyen una aproximación multivariada no paramétrica que permite identificar y dimensionar las variables ambientales de mayor impacto en los rendimientos, sin las restricciones que imponen las técnicas de análisis basadas en los modelos lineales clásicos (Lobell *et al*., 2005). Los modelos CART particionan los datos en forma recursiva de modo tal de conformar subconjuntos cada vez más homogéneos en base a criterios de partición de las variables explicativas. Cada árbol se obtiene a partir de la clasificación de un nodo parental o raíz que contiene la totalidad de los datos, mediante un algoritmo de partición especificado en función de un criterio de partición referido al tamaño del nodo formado o a la variabilidad contenida en los datos del nodo (De'Ath, 2002). Luego se separan dos subconjuntos disjuntos cuya unión comprende el nodo parental (Zhang, 1998; De'Ath y Fabricius, 2000). Si estos nodos se ven afectados por una nueva partición, se los llama nodos internos; si por el contrario, los datos del nodo tienen suficiente homogeneidad o bien el tamaño del mismo es suficiente, éste nodo no vuelve a particionarse??? y recibe el nombre de nodo terminal. Dado que los EMA son costosos, es necesario maximizar la información. La eficiencia en el uso de los recursos destinados a esta tarea puede incrementarse a partir del uso de métodos CART (De'Ath, 2002), que permite representar los datos de rendimiento por una serie de nodos hijos (conjuntos disjuntos) desde el nodo parental (Zhang, 1998).

En el presente trabajo se aborda el análisis univariado y multivariado de una base de datos proveniente de un ensayo multiambiental, componente de la Red de Ensayos Comparativos de Soja de INTA (Instituto Nacional de Tecnología Agropecuaria), Argentina. El objetivo es identificar las variables meteorológicas y de manejo que contribuyen a explicar la variación de los rendimientos. Se evalúa el desempeño de algoritmos CART respecto a procedimientos clásicos de análisis.

TABLA I
VALORES MEDIOS, MÍNIMOS, MÁXIMOS Y COEFICIENTES DE VARIACIÓN DEL RENDIMIENTO DE SOJA, VARIABLES AMBIENTALES Y DE MANEJO DE UN ENSAYO CONDUCIDO EN SIETE AMBIENTES DE LA REGIÓN SOJERA ARGENTINA

Variable	Abrev.	Media	Min	Max	CV
Rendimiento (kg·ha^{-1})	Rto	2896,8	487,2	6324,3	40,7
Temperatura promedio entre R5-R7 (ºC)	T 57	22,4	17,8	26,4	8,5
Precipitaciones acumuladas entre R5-R7 (mm)	PP57	199,2	91,8	360,5	34,9
Radiación acumuladas entre R5-R7 (MJ·m^{-2})	Rad57	842,4	354,3	1543,6	27,5
Grupo de madurez (III, IV, V, VI, VII)	GM	4,8	3,0	7,0	21,5
Fecha de siembra (días julianos)	FS	4312,7	4259,0	4355,0	0,6

Materiales y Métodos

Datos

Se usó una base de datos con 105 valores de rendimientos observados en variedades de soja [Glycine max (L.) Merr.] pertenecientes a grupos de madurez III a VII cultivadas en distintas localidades de la región sojera argentina. Los EMA fueron coordinados por el grupo mejoramiento genético de soja del INTA Marcos Juárez (Proyecto PNCER 2341-INTA). A partir de las fechas de siembras y bases de datos meteorológicos del INTA se construyeron variables ambientales para explicar la variabilidad de los rendimientos en los distintos ambientes. Se consideraron los rendimientos en un total de 105 tratamientos, ambientes surgidos a partir de la combinación de 15 variedades en 7 localidades y variables meteorológicas durante el periodo de llenado de granos (R5-R7): temperatura media (T57), precipitaciones acumuladas (PP57) y radiación acumulada (Rad57), cuyos valores medios, extremos y variación se presentan en la Tabla I.

Procedimientos de análisis

Métodos clásicos. Para identificar las variables más importantes en la explicación de las diferencias se realizó una prueba t comparando valores promedios de cada una de las variables ambientales entre los dos grupos definidos por la mediana del rendimiento: A) rendimientos altos, por encima de la mediana y B) rendimientos bajos, menores o iguales a la mediana.

Luego se analizó la variable rendimiento sin categorizar, mediante el siguiente modelo de regresión lineal múltiple (RLM):

$$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\beta_3x_{3i}+\beta_4x_{4i}+\beta_5x_{5i}+\varepsilon_i$$

donde y_i: rendimiento (kg·ha^{-1}) para la i-ésima observación; β_0: ordenada al origen de la función lineal de ajuste; x_{1i}: T57, temperatura promedio (ºC) en el periodo fenológico R5-R7; x_{2i}: R57, radiación solar acumulada (MJ·m^{-2}) en el periodo fenológico R5-R7; x_{3i}: PP57, precipitación acumulada (mm) en el periodo fenológico R5-R7; x_{4i}: FS, fecha de siembra (días julianos); x_{5i}: GM, grupo de madurez (III, IV, V, VI, VII) del cultivar de soja observado, que se utilizó como variable cuantitativa ya que la clasificación puede considerarse como ordinal con niveles aproximadamente equidistantes; β_1, β_2, β_3, β_4 y β_5: coeficientes de regresión asociados a x_1,…, x_5, respectivamente; y ε_i: término de error aleatorio, ε_i~iidN$(0,\sigma^2)$.

Para visualizar la asociación entre la respuesta y cada una de las regresoras ambientales luego de descontar la asociación entre la respuesta y el resto de las regresoras, se realizó un análisis de residuos parciales con diagramas de dispersión (Draper y Smith, 1998). Con base en la información provista por los coeficientes del modelo y por los residuos parciales, se ajustaron nuevos modelos de regresión lineal múltiple adicionando o eliminando términos. Cuando se visualizaron relaciones no-lineales fue necesario incluir polinomios de segundo grado.

Algoritmos CART. Un árbol de regresión o de clasificación consiste en un conjunto de reglas determinadas por un procedimiento de ajuste mediante particiones binarias recursivas, donde un conjunto de datos es sucesivamente particionado???. Esta técnica está relacionada con los conglomerados divisivos, en los que inicialmente

todos los objetos son considerados como pertenecientes al mismo grupo. En cada instancia de separación el algoritmo analiza todas las variables regresoras y selecciona, para realizar la partición binaria de los datos, aquella que permite conformar dos subgrupos o nodos más homogéneos dentro y más heterogéneos entre ellos. Los nodos formados se separarán nuevamente si cumplen una de las siguientes condiciones: si hay suficiente heterogeneidad para producir una partición de observaciones y/o el tamaño del nodo es superior al mínimo establecido para continuar el algoritmo. El proceso se detiene cuando alguna de estas condiciones no se cumple (Balzarini *et al.*, 2008).

Los árboles de clasificación son usados para predecir variables categorizadas, como en este ejemplo. Se aplicó el algoritmo de partición correspondiente a un árbol de clasificación (Breiman *et al.*, 1984) sobre las dos categorías de rendimiento previamente definidas. El algoritmo particiona observaciones según umbrales de las regresoras y evalúa a cada una de ellas en función de una medida de heterogeneidad (en término de deviance) dentro de cada grupo de la partición.

Los árboles de regresión permiten predecir variables respuestas continuas. El algoritmo que los genera particiona??? las observaciones según umbrales de las variables regresoras, considerando la suma de cuadrados de la respuesta como medida de heterogeneidad dentro de cada partición. Como en el árbol de clasificación, la medida de heterogeneidad entre las observaciones que quedan dentro de un nodo debe ser menor que la calculada entre las observaciones de distintos nodos (Breiman *et al.*, 1984). En este trabajo se consideró la variable rendimiento según su naturaleza continua como variable respuesta, y se empleó el algoritmo correspondiente a un árbol de regresión para obtener los umbrales de las variables explicativas del rendimiento y la dispersión del mismo dentro de cada nodo del árbol.

TABLA II
VALORES MEDIOS DE VARIABLES METEOROLÓGICAS Y DE MANEJO SEGÚN CATEGORÍA DE RENDIMIENTO DE SOJA DEFINIDA POR SU MEDIANA (> o ≤3122kg·ha^{-1})

Categorías de rendimiento	T57	PP57	Rad57	FS	GM
A (>3122kg·ha^{-1})	22 a	204 a	900 a	4306 a	5 a
B (≤3122kg·ha^{-1})	22 a	194 a	784 b	4320 b	5 a

T57: temperatura promedio en el periodo R5-R7 (ºC); PP57: precipitaciones acumuladas en el periodo R5-R7 (mm); Rad57: radiaciones acumuladas en el periodo R5-R7 (MJ·m^{-2}); FS: fecha de siembra (días julianos); GM: grupo de madurez (III, IV, V, VI, VII)
Letras distintas indican que existen diferencias estadísticamente significativas según prueba t bilateral para muestras independientes (p≤0,05).

Criterios de evaluación de los resultados del algoritmo CART

Se realizaron pruebas t a los fines de analizar las significancias estadísticas de las diferencias en rendimiento entre los grupos formados (α=0,05). Se confeccionaron tablas de contingencia cruzando la variable respuesta categorizada (A y B) con cada una de las variables ambientales tomadas como criterios de partición para poner de manifiesto la fuerza de la asociación o relación detectada (Yaklich y Vinyard, 2004). Finalmente se puso a prueba la clasificación obtenida con el árbol de clasificación mediante una validación cruzada (Zhao *et al.*, 2009). Esta validación consistió en aplicar la clasificación en 100 muestras aleatorias tomadas de un conjunto de datos de 184 observaciones que incluyó la base de entrenamiento (n=105) con la que se construyó el árbol, y otros 79 datos provenientes del mismo ensayo multiambiental que se dejaron específicamente para la evaluación del algoritmo. Todos los análisis estadísticos fueron realizados con el software estadístico InfoStat (Di Rienzo *et al.*, 2009).

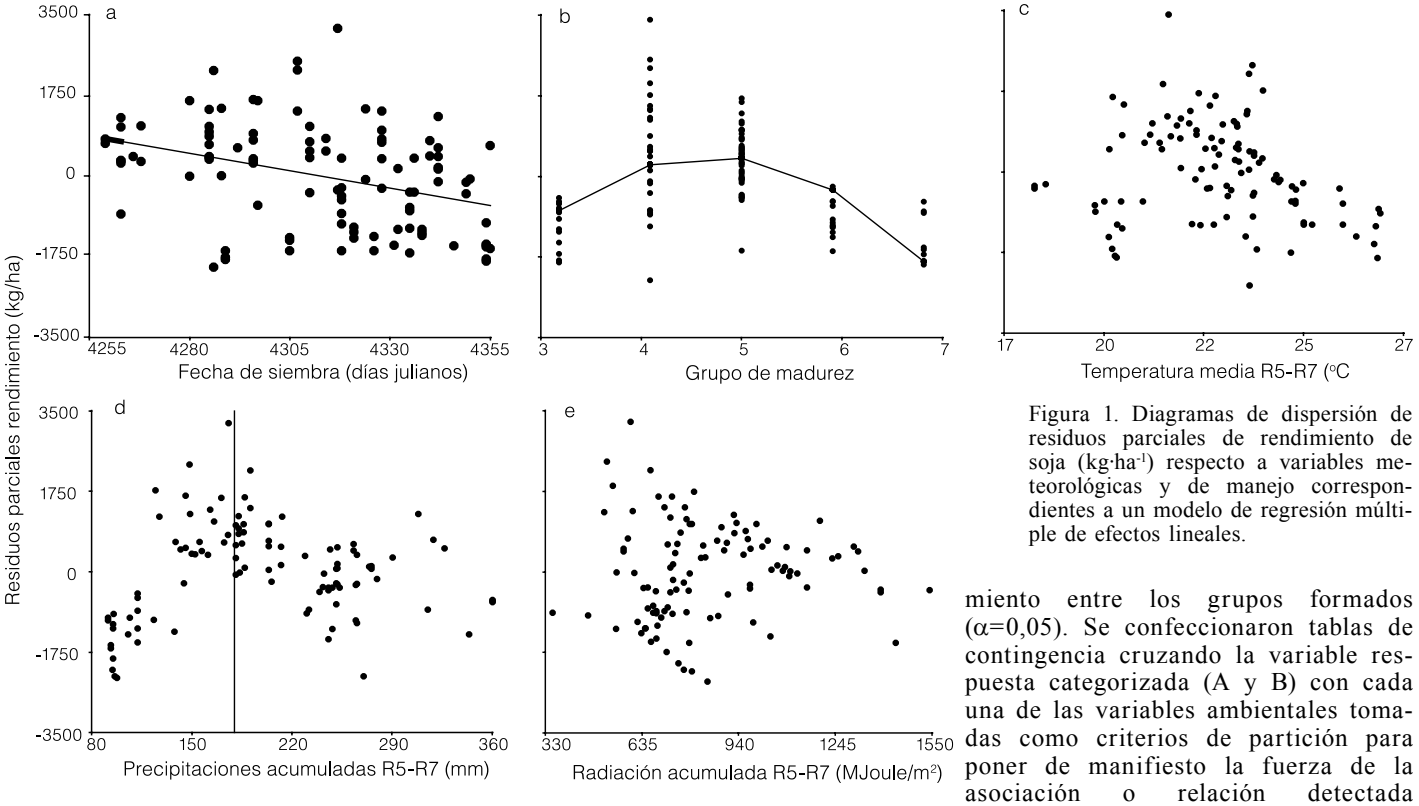

Figura 1. Diagramas de dispersión de residuos parciales de rendimiento de soja (kg·ha^{-1}) respecto a variables meteorológicas y de manejo correspondientes a un modelo de regresión múltiple de efectos lineales.

TABLA III
COEFICIENTES PARA VARIABLES CLIMÁTICAS Y DE MANEJO EXPLICATIVAS DEL RENDIMIENTO DE SOJA EN UN MODELO DE REGRESIÓN MÚLTIPLE DE EFECTOS LINEALES

Variable	Coeficiente estimado	p-valor
T57	-116,4	0,2107
PP57	1,5	0,4649
Rad57	0,1	0,9491
FS	-15,2	0,0389
GM	-169,4	0,3674

T57: temperatura promedio en el periodo R5-R7 (ºC); PP57: precipitaciones acumuladas en el periodo R5-R7 (mm); Rad57: radiaciones acumuladas en el periodo R5-R7 ($MJ \cdot m^{-2}$); FS: fecha de siembra (días julianos); GM: grupo de madurez (III, IV, V, VI, VII).

TABLA IV
COEFICIENTES DE REGRESIÓN PARA VARIABLES METEOROLÓGICAS Y DE MANEJO EXPLICATIVAS DEL RENDIMIENTO DE SOJA EN UN MODELO DE REGRESIÓN MÚLTIPLE QUE INCLUYE UN TÉRMINO CUADRÁTICO PARA GRUPO DE MADUREZ

Variable	Coeficiente estimado	p-valor
T57	-297,4	0,0001
PP57	-0,2	0,8905
Rad57	-1,7	0,0066
FS	-24,8	<0,0001
GM	5831,5	<0,0001
GM^2	-575,4	<0,0001

T57: temperatura promedio en el periodo R5-R7 (ºC); PP57: precipitaciones acumuladas en el periodo R5-R7 (mm); Rad57: radiaciones acumuladas en el periodo R5-R7 ($MJ \cdot m^{-2}$); FS: fecha de siembra (días julianos); GM: grupo de madurez (III, IV, V, VI, VII); GM^2: término cuadrático para la covariable grupo de madurez.

TABLA V
COEFICIENTES PARA VARIABLES CLIMÁTICAS Y DE MANEJO EXPLICATIVAS DEL RENDIMIENTO DE SOJA EN UN MODELO DE REGRESIÓN MÚLTIPLE QUE INCLUYE UN TÉRMINO CUADRÁTICO PARA GRUPO DE MADUREZ Y EXCLUYE EL EFECTO DE FECHA DE SIEMBRA

Variable	Coeficiente estimado	p-valor
T57	-122,0	0,0695
PP57	-0,3	0,8458
Rad57	0,3	0,5554
GM	4850,3	<0,0001
GM^2	-517,6	<0,0001

T57: temperatura promedio en el periodo R5-R7 (ºC); PP57: precipitaciones acumuladas en el periodo R5-R7 (mm); Rad57: radiaciones acumuladas en el periodo R5-R7 ($MJ \cdot m^{-2}$); FS: fecha de siembra (días julianos); GM: grupo de madurez (III, IV, V, VI, VII); GM^2: término cuadrático para la covariable grupo de madurez.

Resultados y Discusión

Prueba t

Mediante la prueba t se compararon las medias de las variables meteorológicas y de manejo para los dos grupos definidos según la mediana del rendimiento (Tabla II). Las variables Rad57 y FS diferencian ambos grupos de rendimiento, mientras que el resto de las variables regresoras no detectan diferencia alguna, y por lo tanto no se identifican como explicativas del rendimiento.

Análisis de regresión lineal múltiple

Los estimadores de las tasas de cambio del rendimiento (coeficientes de regresión) debido a cada regresora sugieren que el efecto de la variable FS es significativamente distinto de cero. Esto guarda consistencia con la dispersión de los residuos parciales de rendimiento respecto a FS, en los que se observa un patrón lineal (Figura 1a). El resto de las variables, en su forma lineal, no genera una tasa de cambio estadísticamente significativa. Los gráficos de residuos parciales muestran patrones que sugieren la falta de efecto sobre el rendimiento de la PP57, T57 y Rad57 (Figura 1). No obstante, todas estas variables han sido reportadas como factores de impacto sobre los rendimientos (Bacigaluppo *et al.*, 2006). El análisis de los residuos parciales de la variable GM sugiere la inclusión de un término cuadrático (Figura 1b). Los residuos parciales respecto a T57 ponen de manifiesto un grupo de outliers que resultan influyentes, generando posible heterogeneidad de varianzas (Figura 1c). Los residuos respecto a PP57 parecieran indicar que en ambientes con bajas precipitaciones en el periodo R5-R7 (<180mm), el cultivo responde positivamente al incremento de precipitaciones, pero en ambientes más húmedos la precipitación parece no afectar al rendimiento (Figura 1d). La dispersión de los residuos parciales de rendimiento respecto a Rad57 no sugiere la existencia de relación (Figura 1e).

Al incluir en el modelo un término cuadrático para la variable GM, además de la FS, también resultaron estadísticamente significativas las variables T57, Rad57, y el término lineal y cuadrático de GM (Tabla IV). La incorporación de este término cuadrático cambió el ajuste de manera considerable. En la Tabla V se muestran los estimadores de los coeficientes de un nuevo modelo que excluye a FS y conserva el término cuadrático de GM. Se observa un efecto estadísticamente significativo de las variables GM y T57. La exclusión de la variable FS afecta las conclusiones sobre los efectos de T57 y Rad57 respecto al modelo anterior. Esta alta sensibilidad de los coeficientes de los modelos lineales es común cuando existe multicolinealidad entre las variables explicativas. La multicolinealidad presente entre regresoras es esperable ya que las temperaturas y las radiaciones se correlacionan y dependen

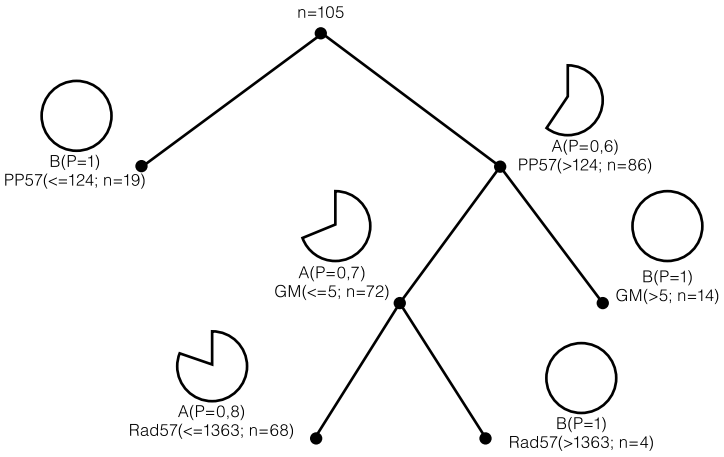

Figura 2. Árbol de clasificación con variable respuesta rendimiento categorizado según su valor de mediana (3122kg·ha^{-1}). Se indica en cada nodo, explícita y gráficamente (gráficos de sectores) la probabilidad de tener rendimientos altos o bajos respecto a 3122kg·ha^{-1}. En cada nodo se indica la variable meteorológica o de manejo que el algoritmo tomó como criterio de partición, el valor umbral a partir del cual se separaron los nodos, y la cantidad de ambientes (n) del nodo formado.

de la FS y el GM usado (Bacigaluppo *et al.*, 2006).

CART

En la evaluación del desempeño de los algoritmos CART se evidencian algunas diferencias con respecto a RLM en la detección de variables que explican variabilidad en los rendimientos. Las particiones binarias de los rendimientos, realizadas mediante un árbol de clasificación (Figura 2) y un árbol de regresión (Figura 3) según las variables meteorológicas y de manejo, identifican la variable PP57 como criterio que permite separar grupos de rendimientos altos y bajos. Cabe destacar que si bien el algoritmo particiona los datos según los criterios de tamaño del nodo o heterogeneidad dentro del nodo, las particiones obtenidas pueden ser anidadas, numerosas y cada vez de menor tamaño (Zhang, 1998). Es por ello importante "podar" las ramificaciones inferiores del árbol con base en la problemática estudiada, ya que solo se desea predecir la relación entre las principales variables y el rendimiento del conjunto de datos, pero no se espera predecir cada dato. Es por ello que se consideran las primeras instancias de partición para poder extender el comportamiento de las variables en otro conjunto de datos.

El árbol de clasificación (Figura 2) sugiere que la variabilidad de los datos se debe en primer lugar a la PP57. Esta variable presenta como umbral el valor de 124mm, a partir del cual se separan dos nodos. Los 19 ambientes cuyo valor de PP57 es menor que el umbral, tienen un rendimiento por debajo de la mediana con una probabilidad de 1. A su vez, para los 86 ambientes observados con valores de PP57 superiores a 124mm, aún existe variabilidad de rendimiento que debe ser explicada; consecuentemente, el algoritmo continúa con la partición de este nodo. En esta instancia de partición la variable que mejor explica esa variabilidad remanente es GM. Los 72 ambientes con GM≤5 tienen un rendimiento superior a la mediana con una probabilidad de 0,70 mientras que los restantes 14, con GM>5 rindieron por debajo del valor de la mediana. Para la región sojera argentina, que incluye los ambientes observados, los grupos intermedios son los más recomendados. Del nodo de 72 ambientes se desprenden dos nodos según la variable Rad57. A partir de un valor de Rad57 <1363MJ·m^{-2} se obtiene un alto rendimiento en el 80% de los casos. Así, este árbol de clasificación logra detectar en los primeros niveles de partición variables importantes en la explicación de los rendimientos (Bacigaluppo *et al.*, 2006). El árbol de regresión (Figura 3) detecta como primer criterio de partición a la variable GM. Los 86 ambientes a los que les corresponde un valor de GM≤5, tienen rendimientos tanto altos como bajos; mientras que los 19 ambientes con GM>5 rinden siempre por debajo de 3122kg·ha^{-1}. Dentro de los 86 ambientes con GM≤5, el algoritmo detectó que se pueden distinguir dos nodos considerando la PP57, también con un valor umbral de 124mm. Se encontraron entonces 14 observaciones que rindieron por debajo de la mediana para valores de PP57 ≤124mm y en el nodo complementario aún se observa variabilidad en el rendimiento. El algoritmo detecta dentro de este último subgrupo rendimientos diferenciados a partir de un umbral del mismo criterio de partición que en la instancia anterior (PP57), con un umbral de 204mm, pero en ambos nodos sigue habiendo variabilidad de rendimientos después del tercer nivel

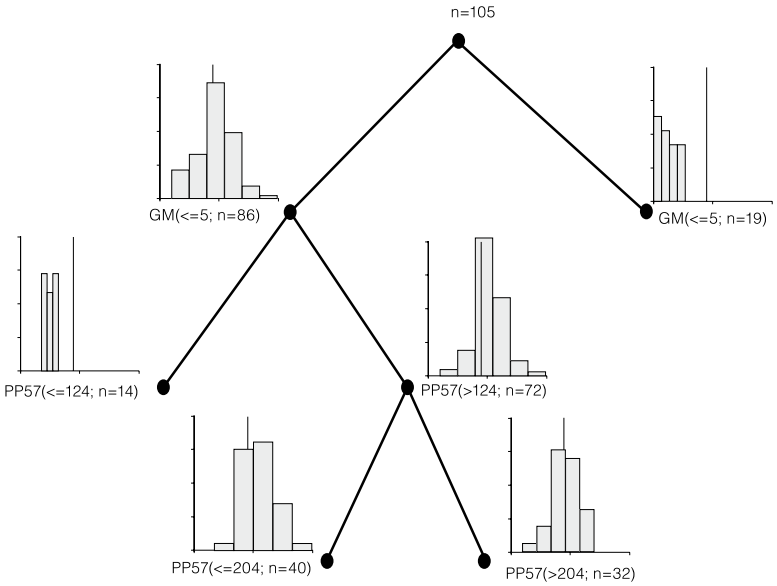

Figura 3. Árbol de regresión con variable respuesta rendimiento en escala continua. Se muestra el histograma de frecuencias de rendimientos en cada nodo y la mediana del rendimiento (línea de referencia sobre abscisas). En cada nodo se indica la variable meteorológica o de manejo que el algoritmo tomó como criterio de partición, el valor umbral a partir del cual se separaron los nodos, y la cantidad de ambientes (n) del nodo formado.

TABLA VI
RENDIMIENTOS PROMEDIO CLASIFICADOS POR VARIABLES METEOROLÓGICAS Y DE MANEJO EN LOS TRES PRIMEROS NIVELES DE PARTICIÓN DEL ÁRBOL DE CLASIFICACIÓN

						Número de ambientes		
Nivel de partición	n	Criterio de partición	Clasificación según umbral	Media[i]	n	A (>3122kg·ha^{-1})	B(≤3122kg·ha^{-1})	Total
1º	105	PP57	{PP57>124}	3150a	86	53	33	86
			{PP57≤124}	1751b	19	0	19	19
						Total 53	52	105
2º	86	GM	{GM>V}	1285b	14	0	14	14
			{GM≤V}	3512a	72	53	19	72
						Total 53	33	86
3º	86	Rad57	{Rad57>1363}	2831b	4	0	4	4
			{Rad57≤1363}	3552a	68	53	15	68
						Total 53	19	72

PP57: precipitaciones acumuladas en el periodo R5-R7 (mm); Rad57: radiaciones acumuladas en el periodo R5-R7 (MJ·m^{-2}); GM: grupo de madurez (III, IV, V, VI, VII).
[i]Letras distintas indican que existen diferencias estadísticamente significativas (p≤0,05).

TABLA VII
RENDIMIENTOS PROMEDIO CLASIFICADOS POR VARIABLES METEOROLÓGICAS Y DE MANEJO EN LOS TRES PRIMEROS NIVELES DE PARTICIÓN DEL ÁRBOL DE REGRESIÓN

Nivel de partición	n	Criterio	Clasificación según umbral	Media[i]	n	Número de ambientes A (>3122kg·ha^{-1})	B (≤3122kg·ha^{-1})	Total
1º	105	GM	{GM>V}	1251b	19	0	19	19
			{GM≤V}	3260a	86	53	33	86
					Total	53	52	105
2º	86	PP57	{PP>124}	3512b	72	53	19	72
			{PP≤124}	1963a	14	0	14	14
					Total	53	33	86
3º	72	PP57[i]	{PP>204}	3216b	32	20	12	32
			{PP≤204}	3749a	40	33	7	40
					Total	53	19	72

PP57: precipitaciones acumuladas en el periodo R5-R7 (mm); GM: grupo de madurez (III, IV, V, VI, VII).[i] Letras distintas indican que existen diferencias estadísticamente significativas (p≤0,05).

de partición. Cada nodo conformado se acompaña de un histograma de los datos que lo integran (Figura 3). Las zonas de producción de soja pueden ser caracterizadas usando las clasificaciones obtenidas de los nodos formados en los tres primeros niveles de partición de los árboles de clasificación y de regresión (Tablas VI y VII).

Cada dato del conjunto de validación fue clasificado según el árbol obtenido, concluyendo sobre su pertenencia al grupo de rendimientos altos o bajos. Se compararon los resultados a los que condujo la clasificación, usando el árbol hasta el tercer nivel de partición, con los datos observados de rendimiento (mayores o menores a 3122kg·ha^{-1}). La validación se repitió 100 veces a partir de muestras de 75 casos tomados bajo un muestreo sin reposición del conjunto total de casos. Se observó un alto porcentaje de coincidencia entre la clasificación sugerida por el algoritmo CART y la clasificación de los rendimientos según el umbral de 3122kg·ha^{-1} (mediana del conjunto de datos de entrenamiento). El 100% de los casos con rendimiento alto fueron clasificados de este modo (clasificación correcta) y el 80% de los casos con bajos rendimientos se clasificaron como bajos (clasificación correcta; Tabla VIII).

TABLA VIII
PORCENTAJE DE CLASIFICACIÓN DE 100 MUESTRAS ALEATORIAS OBTENIDOS POR VALIDACIÓN CRUZADA DEL ÁRBOL DE CLASIFICACIÓN

Clasificación de rendimientos según umbral (3122kg·ha^{-1})	Clasificación de datos de validación Correcta[i]	Incorrecta
Alto	100,0	0,0
Bajo	81,2	18,8
Total	88,0	12,0

[i] Promedio de 100 simulaciones.

Conclusiones

Las pruebas t univariadas resultaron ineficientes para detectar las covariables ambientales que afectan la variabilidad de los rendimientos. Variables importantes como la precipitación y el GM no se detectaron como significativas debido a la alta varianza dentro de los grupos de rendimiento.

La multicolinealidad de las variables ambientales afectó la sensibilidad de la técnica de RLM para identificar variables influyentes en la variabilidad de los rendimientos. La RLM no identificó a la precipitación como una variable de importancia porque ésta evidenciaba, en el dominio de los valores observados, un comportamiento no-lineal.

Los resultados obtenidos con los algoritmos CART no fueron afectados por la multicolinealidad de las covariables ambientales ni por la falta de linealidad en la relación con las precipitaciones. Ambos algoritmos, tanto el árbol de clasificación como el árbol de regresión, identificaron a la precipitación y al grupo de madurez como importante. Los procedimientos aplicados a posteriori para el análisis de nodos de los árboles, permitieron asignar valores de significancia a las asociaciones de estas variables con el rendimiento. El árbol de clasificación resultó más apropiado que el árbol de regresión para el objetivo de identificar variables. El árbol de regresión produjo excesivas particiones de los datos, formando nodos dentro de nodos de la misma variable.

Los análisis estadísticos basados en árboles de clasificación y/o regresión constituyen una alternativa a los modelos lineales clásicos de regresión ya que captan comportamientos no aditivos y sus resultados no se ven afectados por interacciones entre variables regresoras (multicolinealidad).

AGRADECIMIENTOS

Las autoras agradecen a Daniel Collino por la colección de datos meteorológicos y la coordinación del PNCER 2341; a Francisco Fuentes y Beatriz Masiero, coordinador y analista de ensayos multiambientales de soja en INTA (REC-SO); a la Universidad Nacional de Córdoba por brindar su espacio para la investigación; y al Consejo Nacional de Ciencia y Tecnología de Argentina por subsidiar desarrollos biométricos orientados al análisis de ensayos multiambientales.

REFERENCIAS

Bacigaluppo S, Dardanelli J, Gerster G, Quijano A, Balzarini M, Bodrero M, Andriani J, Enrico J (2006) *Variaciones del Rendimiento de Soja en el Sur de Santa Fe. Factores Limitantes de Clima y Suelo*. Para Mejorar la Producción Nº33. EEA INTA Oliveros, 38-42. INPOFOS Informaciones Agronómicas del Cono Sur Nº 32:12-15. www.planetasoja.com y http://e-campo.com

Balzarini M, Bruno C, Arroyo A (2005) *Análisis de Ensayos Agrícolas Multiambientales. Ejemplos en Info-Gen*. Brujas. Córdoba, Argentina. 141 pp.

Balzarini MG, González L, Tablada M, Casanoves F, Di Rienzo JA, Robledo CW (2008) *Manual del Usuario*. Brujas, Córdoba, Argentina. 333 pp.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees*. Wadsworth International Group. Belmont, CA, EEUU.

De'Ath G (2002) Multivariate regression trees: a new technique for modeling species- environment relationships. *Ecology 83*: 1105-1117.

De'Ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology 81*: 3178-3192.

Di Rienzo JA, Casanoves F, Balzarini MG, González L, Tablada M, Robledo CW (2009) *InfoStat versión 2009*. Grupo InfoStat. Universidad Nacional de Córdoba, Argentina.

Draper NR, Smith H (1998) *Applied Regression Analysis*. Wiley. Nueva York, EEUU. 736 pp.

Lobell DB, Ortiz-Monasterio JI, Asner GP, Naylor RL, Falcon WP (2005) Wheat. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. *Agron. J. 97*: 241-249.

Searle SR (1971) *Linear Models*. Publicado en la Wiley Classics Library (1997) Nueva York, EEUU. 532 pp.

Tittonell P, Shepherd KD, Vanlauwe B, Giller KE (2008) Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya- an application of classification and regression tree analysis. *Agric. Ecosyst. Env. 123*: 137-150.

Yaklich R, Vinyard B (2004) Estimating soybean seed protein and oil concentration before harvest. *J. Am. Oil Chem. Soc. 81*: 189-194.

Zhang H (1998) Classification trees for multiple binary responses. *J. Am. Stat. Assoc. 93*: 180-193.

Zhao L, Zheng X, Yan H, Wang S, Zhang K (2009) Construction and application of the decision tree model for agricultural land grading based on MATLAB. 2nd Int. Workshop on Knowledge Discovery and Data Mining. Moscú, Rusia. pp. 155-158.

Zheng H, Chen L, Han X, Zhao X, Ma Y (2009) Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. *Agric. Ecosyst. Env. 132*: 98-105.

IDENTIFYING YIELD AND ENVIRONMENT RELATIONSHIPS USING CLASSIFICATION AND REGRESSION TREES (CART)

Soleana Rosales Heredia, Cecilia Bruno and Mónica Balzarini

SUMMARY

Multi-environmental trials (MET) are common in agricultural research. Meteorological and management covariates which potentially explain yield-environment relationships are recorded to characterize environments. The relative contribution of different covariates is usually done through univariate statistical techniques, particularly regression analysis and t test. However, classification and regression tree algorithms (CART) represent an alternative multivariate approach to identify environmental variables with the greatest impact on yield. The CART model has fewer restrictions to its implementation than classical techniques based on linear models. In this study, the relative performance of CART and traditional linear model-based procedures is evaluated. A database containing soybean yields across a wide environmental range, from a MET conducted in the Argentinean soybean crop region, is used. CART algorithms are a robust technique for identifying environmental covariates predicting yield variability. The study highlights the need to take into account multicolineallity when environmental covariates are used in linear models.

IDENTIFICAÇÃO DE RELAÇÕES ENTRE RENDIMENTOS E VARIÁVEIS AMBIENTAIS VÍA ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO (CART)

Soleana Rosales Heredia, Cecilia Bruno e Mónica Balzarini

RESUMO

Em investigações agrícolas são frequentes os ensaios multiambientais (EMA) para a comparação de rendimentos de vários genótipos em múltiples ambientes. Para caracterizar os ambientes é comúm registrar covariáveis meteorológicas e/ou de manejo com potencialidade de explicar relações entre rendimentos e ambientes. A contribuição relativa das distintas covariáveis costuma ser analisada por técnicas estatísticas univariadas, aprincipalmente prova t e análise de regressão. No entanto, as árvores de classificação e de regressão, algoritmos CART, constituem uma aproximação multivariada alternativa para identificar variáveis ambientais de mais impacto nos rendimentos. Os CART apresentam menos restrições para sua implementação que as técnicas de análises baseadas nos modelos lineares clássicos. No presente trabalho se avalia o desempenho de algoritmos CART comparado com os procedimentos clássicos de análises em uma base de dados de rendimentos de soja proveniente de um EMA incerto na região de soja argentina através dos ambientes. Os resultados mostram que os algoritmos CART constituem, devido a seu baixo erro de predição, uma técnica robusta para predizer a variabilidade nos rendimentos como resposta a variações ambientais. O estudo ressalta a necessidade de contemplar a multicolinearidade ao trabalhar com covariáveis ambientais e modelos clássicos.